The Enhancement of Trending Analysis for YouTube Social Media Platform

KHIN THAN NYUNT
Department of Information Science,
Naypyitaw State Polytechnic University,
Kayae Hostel, Building No. (105), Zabuthiri Township, Nay Pyi Taw,
MYANMAR

ORCiD: 0009-0009-2704-1164

Abstract: - Developments in technology have led to a rise in YouTube users, but choosing the right career can be challenging due to the numerous categories and channels available. YouTube trend analysis can help users make informed decisions about their professional lives, especially those seeking to use YouTube as a source of income. A hybrid system combining trending and sentiment analysis using the linear regression method of prediction for trending and the multinomial Naive Bayes classification model for sentiment analysis is proposed for the country United States (US), 2023. This system provides recommendations for YouTube career choices and identifies the most trending categories and channels, providing valuable insights for those struggling to make a living on the platform.

Key-Words: - YouTube Trending Analysis, Sentiment Analysis, Statistical Count, Linear Regression, Multinomial Naïve Baye, TF-IDF.

Received: June 27, 2024. Revised: November 5, 2024. Accepted: November 27, 2024. Published: December 13, 2024.

1 Introduction

Today, YouTube allows users to post, watch, and share videos for free. Every day, all YouTube users receive billions of views and hundreds of millions of hours of video. YouTube content creators submitted 10,000 as they can earn money by creating a channel on the site. YouTube is a source of income for many people, and therefore a video's popularity ultimately becomes the top priority for sustaining a steady income, provided that the popularity of videos remains the highest. To choose the right career in YouTube, a particular approach is to collect all the data regarding popular videos, such as the number of views each video has had on YouTube and the duration of its trend. The YouTube API, [1], which the YouTube developer has made public, may be used to collect the data. With the aid of these data, which show the dislike count of any user to the public, one may better understand the fundamental requirements for a video to be listed in YouTube's trending category, [2]. Since December 14, 2021, YouTube has ceased making various content analytics, such as the number of dislikes, publicly available due to privacy concerns. Rather than displaying a personalized video, trending presents an equivalent selection of popular videos to a large number of people nationwide. According to the YouTube Help Center inquiries, YouTube determines if a video is ranked on trending to balance all of these considerations. To achieve this, trending considers many signals on YouTube, including the following (but not limited to):

- View count
- How quickly the video is generating views
- Where views are coming from, including outside of YouTube
- The age of the video
- How the video performs compared to other recent uploads from the same channel

On YouTube, with each update 1 time for 15 minutes, videos on the trending list may trend up or down, [2], [3]. Therefore, it is difficult to determine which channels and categories are trending. Because of this, it has become quite challenging for those who wish to use YouTube as a source of income and encounter multiple challenges when it comes to making a career decision. By looking at these factors, YouTube determines whether a video is trending. It is considered to be based on view count alone. It must be said that this is general. In this paper, a video to trend, the trending result is specifically correct, as it is reviewed based on 4 features, such as views, likes, comments, and

negatives. YouTube is trending; A review based on view count alone may be correct to determine whether or not it is. That is really general. Every time some audience views a video, it cannot be said that they watch it until the end of the video. Only the audience can know who is actually watching a video until the end. Good comments are also important to consider. Although a high view count may be common on YouTube, based on audiences' likes, I don't know why I don't like it. Audiences like and dislike YouTube because it has stopped showing the dislike count to the public since December 14, 2021, so it is not easy to know and only the like count is seen. Therefore, to know the dislike count, we have to use the negative sentiment results obtained by conducting sentiment analysis through comments. The new contribution of this system is to add negative sentiment results as a feature and generate more accurate and correct results based on the as four features such as view, like, comment, negative, etc. Therefore, sentiment analysis is done to get the real opinion from the audience, and the feature is not based on the view count alone; the decision is made based on the views count, like count, comment count, and negative count. For every video that is really trending, the audience has to think about whether it is the most watched or not. If the categories and channels to be recommended to the user match both the most trending and the most-watched, this system will help the user make a more correct career choice. The results from these analyses are combined as a hybrid and then recommended for the US to help them decide the most suitable career choice for individuals who plan to earn a living from that platform. Therefore, this system is proposed for that reason.

2 Research Methodology

This system is a hybrid implementation of trending analysis and sentiment analysis combined seen in Figure 1. This system is mainly designed with the concepts of data scraping, information extraction, feature extraction, and data analysis, and performance scaling.

In trending analysis, Pearson's correlation method was used to determine whether features correlate with one another. It is used to measure how one video's content is related to another based on the following features: views count, likes count, comment count, etc.: Moreover, the linear regression model is mostly used in trending analysis to anticipate YouTube trends.

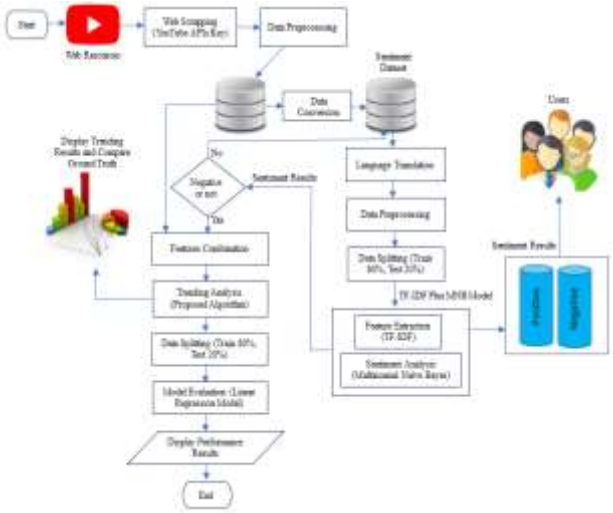


Fig. 1: The Proposed System Design

A statistical model called linear regression makes use of known data values to forecast the value of unknown data. In order to forecast the outcome of future occurrences, it offers a linear connection between an independent variable and a dependent variable. In sentiment analysis, the Multinomial Naive Bayes model and the Term Frequency-Inverse Document Frequency (TF-IDF) model are used to analyse text data. To create a classification model that predicts sentiment, text documents must be converted into a matrix of TF-IDF features. Word relevance/significance in a collection of documents is quantified using the TF-IDF model. It is employed to transform the text into a TF-IDF feature matrix. The next section displays the results of the classification. Figure 1 displays the proposed system design by using web content mining over the YouTube social media platform.

3 Research State

3.1 Trending Analysis

In trending analysis, the statistical count of every YouTube video is extracted from the YouTube video content by using YouTube API_KEYS to establish the parameters for the YouTube trending analysis. The data includes the period of January 1, 2023, to December 15, 2023. There are 16 columns and 244387 rows for the United States (US). After the parameters / features are established, the raw data is preprocessed into data that can be used. After that, the feature extraction step is done. Feature extraction is a process that identifies important features or attributes of the data. To accomplish this, among those features, views, likes, and comments will be used to determine whether it is trending or

not. In addition, YouTube takes into account "Where views are coming from, including outside of YouTube, as mentioned in section I, so tagCount is also considered as part of this system. In the YouTube trending analysis, we discussed in section I, Introduction, YouTube and previous research based just on views count, but in this study, the system-generated result will offer accuracy and correctness to users because of the use of various features. The algorithm proposed as a new contribution to this system has 12 steps in the following, and this part is the trending analysis part.

Step 1: Data acquisition from the YouTube Social Media Platform
Step 2: Web Scraping using YouTube APIs Key
Step 3: Data Preprocessing
Step 4: Store the preprocessed data in the dataset
Step 5: Extract comments and convert from statistical value to categorical value and then store that data in sentiment dataset
Step 6: Translate the comments of the three countries: the US, India, and Japan, as the English Language
Step 7: Data preprocessing to make sentiment analysis
Step 8: Split data into training 80% and testing 20%; random state 50
Step 9: Create a Model with the combination of TF-IDF and Multinomial Naïve Bayes
(a) vectorize as TF-IDF features
(b) classify the sentiment data using the Multinomial Naïve Bayes Model
(c) display the sentiment results and their performance to users
(d) store the sentiment results in the sentiment dataset at Step5
Step 10: Extract and test the features

```
if (sentiment result = = -1)
    negative;
else
    positive;
```

Step 11: Combine the features view, like, and comment in the dataset at Step 4 with the negative feature obtained from Step 10
Step 12: Make a trending analysis using the proposed trending algorithm
Step 13: Display the trending results and compare the ground truth
Step 14: Create a Linear Regression Model, including training 80% and testing 20% with random state 50
(a) predict the outcome of observed data using R-squared (R^2) that is a statistical measurement
(b) evaluate the prediction result using the loss function: Mean Squared Error (MSE) and Mean Absolute Error (MAE) to solve the overfitting problem
(c) evaluate model's predictive capabilities: accuracy, precision, and recall overview like, comment, negative, and trending
Step 15: Display the model's performance results

Primarily based on research observations, the number of views for each trending video was calculated together with the number of views, likes, and comments in the bar plot shown in Figure 2 (Appendix). More than 200,000 trending videos have more than 98.29% of views below 20,000,000 million. More than 99.28% of the users who liked over 7,000 trending videos were found to be under 2 million in number. It has been observed that nearly 98.76% of the 160,000 trending videos have less than 100,000 comments.

Pearson's Correlation method is used to measure how one video content is related to another, based on the features: view_count_start, like_count_start, comment_count_start, tagCount, hoursTakenToTrend, and negative_sentiment score shown in Figure 3 using a heatmap. In this Figure, there may be a strong and wonderful correlation value of 0.79% between view_count_start and like_count_start, in addition to a study and wonderful correlation value of 0.69% between like_count_start and comment_count_start. The number of views_count_start and comment_count_start has a useful fine correlation with the value of 0.53%. The correlation values between hoursTakenToTrend, which is the time it takes for a video to become trending, and view_count_start, like_count_start, and comment_count_start is 0.052%, 0.045%, and 0.047%, respectively. According to this analysis, If the view count increases from that analysis, the like count will likely increase. Similarly, assume that if the number of likes increases, it is likely that the number of comments will also increase. Therefore, correlation analysis is used to determine how much one variable has changed as a result of another's change. A high correlation denotes a robust relationship between two features, whereas a low correlation denotes a weak relationship. Finding correlations, trends, patterns, and important interactions between two qualities or datasets is accomplished with the use of this technique. When one feature increases and the other does too, there is a positive relationship between the two. The value of the correlation coefficient is between +1 and -1. A score of 1 indicates a high degree of correlation

between the two features, whereas 0 denotes no correlation at all.

3.2 The Proposed Trending Algorithm

While YouTube bases its decision on whether a video is trending or not based on five factors discussed in the Introduction section, this approach is reasonable in light of the fact that a more useful and accurate trending algorithm has been proposed as a new contribution.

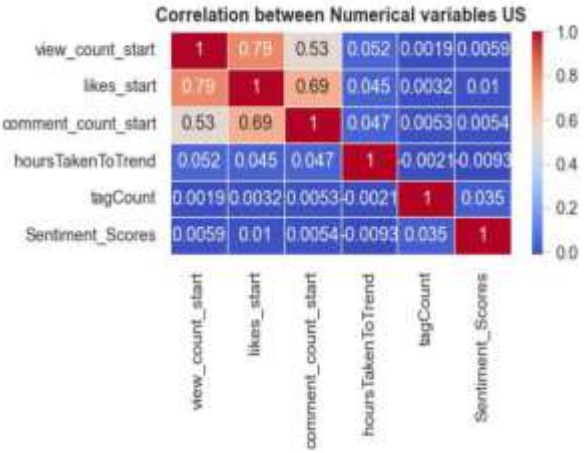


Fig. 3: The Correlation Analysis between One Feature and Another

Step 1: Create new features view_count_start and view_count_end from view and then create the function min(views), max(views)
Step 2: Create new features likes_start and likes_end from likes and then create the functions min(likes), max (likes)
Step 3: Create new features comment_count_start and comment_count_end from comment and then create function min(comments), max(comments)
Step 4: Create new feature tagsCount from tags and then create function tags ( )
Step 5: Create new features trending_date_start and trending_date_end from trending_date and then create the functions min(trending_date), max(trending_date)
Step 6: Create new feature hoursTakenToTrend from publishedAt and trending_date_start and then create function hoursTakenToTrend()
Step 7: Create a new feature trendingDaysDuration from publishedAt and trending_date_start and then create function trendingDaysDuration()
Step 8: Create a new feature negative obtained from sentiment analysis and then create a negative ( )
Step 9: Extract the published time of a day, published day of a week, published month of each video, published year of each video, and display the trending categories and channels, the most trending and most watching

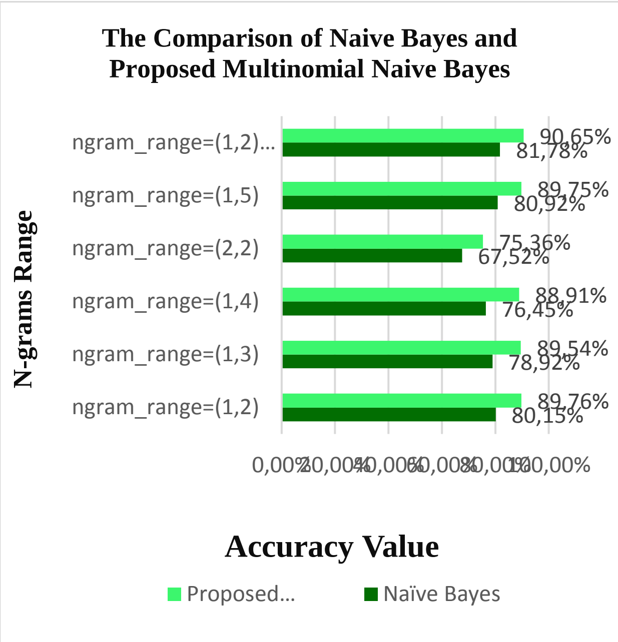
3.3 Sentiment Analysis

In an effort to precisely calculate the YouTube trending analysis: the negative sentiment result is inserted into a dislike count obtained from the sentiment results acquired from the audience's comments. It also intends to know how the audience feels about the content by analyzing the comments on a video. Sentiment analysis covers the key components of the overall system, such as the language translation process, preprocessing process, term weighting techniques, labelling process, and classification process, [4]. The next contribution is to contribute the combination of YouTube trending results and sentiment results as a hybrid system to accomplish the YouTube Career analysis. Sentiment analysis applies the Term Frequency-Inverse Document Frequency (TF-IDF) model and the Multinomial Naive Bayes method to examine textual data. To develop a sentiment-predictive classification model, text documents need to be transformed into a matrix of TF-IDF features. Using the TF-IDF model, word relevance/significance in a set of documents is measured. It is used to convert the text to a feature matrix in the TF-IDF model, [5]. The text data in this system is first labelled with positive or negative sentiment labels and then separated into training and test datasets to analyze the model. The Multinomial Naive Bayes model is applied to the TF-IDF features of the text data. The models are trained using the labeled data, and their performance is evaluated on the test dataset. Model shows the sentiment classification results and performance metrics. Including accuracy, calculated to assess the quality of predictions. Figure 4 illustrates how a percentage of each video is correctly combined when a negative feature is added, taking into account both the category_id and each video's video_id.

We discovered that certain comments on YouTube contain slang and misspellings, making them difficult to classify. Although it produced inaccurate results due to the noise in the data, the Multinomial Naïve Bayes model remains stable due to its best prediction, which achieved the highest accuracy of nearly 90%, as shown in Figure 5. The sentiment description is implemented as seen in Figure 4. The left-side figure shows the result of sentiment classification results and also displays the percentage of negative sentiment results in each category of the right-side figure. This system also well implements how the audience is feeling about

the content by analyzing the sentiment of comments on a video. This is very helpful not only for those who want to start a Career with YouTube but also for current professional YouTubers. The accuracy values varied depending on the n-gram range, with n-gram (1,2), alpha (0.1) having the highest accuracy value of 90.65% as shown in Figure 5. The research findings highlight the importance of the n-gram range in sentiment analysis.

Fig. 4: Adding Negative Sentiment by video_id and category for the US




```
Accuracy: 89.8
Detail:
              precision    recall  f1-score   support

           0       0.86      0.88      0.87      3440
           1       0.93      0.96      0.94      4854
           2       0.92      0.81      0.86      1458
           3       0.87      0.87      0.87      3355

    accuracy                           0.90     13107
   macro avg       0.90      0.88      0.89     13107
weighted avg       0.90      0.90      0.90     13107
```

Fig. 5: The Performance Results of Sentiment Analysis

4 Problem Statement and Solution

The research problem and solutions are illustrated in Figure 6 and Figure 7 to make them more visible. This system is proposed to solve these problems and to enhance YouTube videos trending algorithm.

Fig. 6: Problem Statement

Fig. 7: Problem Solution

5 Experimental Results

Linear regression model is one of the easiest and most popular machine learning approach, [6]. It is a statistical method that is used for trending analysis. It makes predictions for continuous / real or numeric variables such as view, like, comment in this system. Linear regression is a statistical technique used to understand relationship between two continuous variables by fitting a straight line to the data points, [7]. However, it's not suitable for classification takes where the goal is to predict which category or class an observation belongs to. Therefore, logistic regression is used to classify the days of weeks (trendingDate), and sentiment(negative_features) in trending analysis. This section details the trending results obtained using the enhanced trending algorithm. According to the reference [8], these papers analyzed YouTube trending videos before 2020. In those years, the dislike count can still be easily obtained, and it is not difficult to find it. This system was implemented to solve this problem.

5.1 Trending Analyses on the Channels and Categories

We quantify how many videos are trending on each channel to determine which ones are the trendiest. The capabilities of Panda's libraries should be usable on data outlines that include the frequency at which the channel has become a trending location. The dependencies for each channel are calculated and shown using matplotlib, a Python tool. According to Figure 8 and Figure 9, the top 10 channels and 15 categories can be analyzed to determine which channel and category have the most popular content. There are a lot of trending videos every day among the many videos that are uploaded on the US version of the YouTube social media platform. With 3645 videos, the "Gaming" genre is the most popular among them. Figure 9(a) makes it very evident that the "Music" category is in the top 3 with 3161 trending videos, the "Entertainment" category is in the top 2 with 3559 trending videos, and the "Gaming" category is in the top 1 with 3645 trending videos. Figure 9(a) shows the trending results obtained after negative feature integration and Figure 9(b) shows the trending results obtained before negative feature integration. Therefore, Figure 9 is a comparison of trending result differences before and after negative feature integration, according to research findings. Although there is no change from Top 1 to Top 5, a slight change from Top 6 onwards can be clearly seen in the result image. In (b) before negative feature integration, Comedy was in the Top 7, but after negative feature integration, it reached the Top 6. This is also reasonable. Because the negative percentage of Comedy (8.717949%) is higher than Film & Animation (8.645533%). Therefore, it can be said that this system is more reasonable and accurate. Comparably, with 212 videos, the "NBA" channel is the most popular trending channel, followed by the "NFL" channel with 181. Figure 8 makes it rather evident.

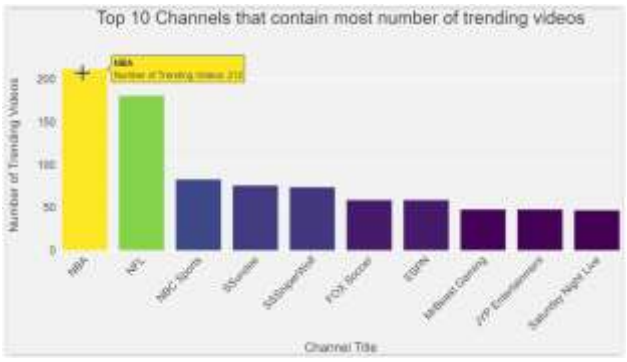
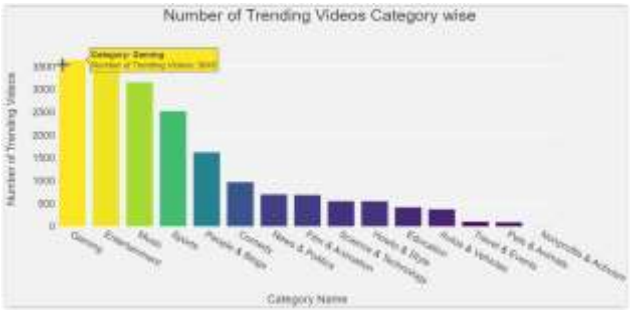
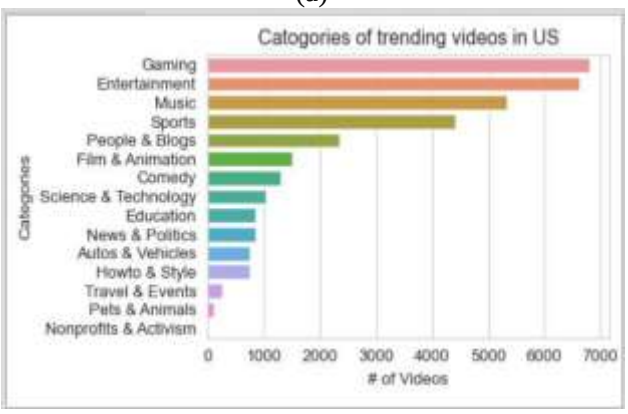


Fig. 8: The Most Trending Channels in US, 2023



(a)



(b)

Fig. 9: Analyzing and Contrasting The Trending Results

It took an average of at least 4 to 13 and a half hours from the time it was published to trend, as can be seen from (q1: 4.6, trace 0) to (q3: 13.3, trace 0) in Figure 10. Conducted a study to find out how long it takes for a video to become trendy. It is best to take most of the videos at (medium: 8, trace: 0). The time is displayed on the X-axis of the histogram, while the number of movies is displayed on the Y-axis. As can be observed, 2185 videos took longer than 4 hours, 2807 videos took 8 hours, and 3109 videos took 10 hours on average. 1315 videos took longer than 23 hours, and 262 videos took longer than 25 hours, according to my research findings.

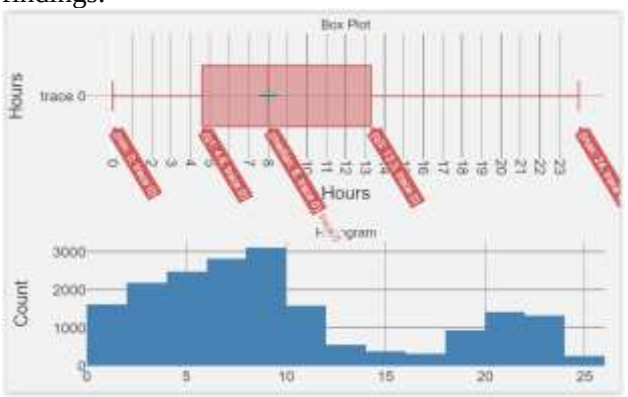


Fig. 10: The Number of Hours Taken by a Video to Get into Trending List

Also, according to [9], the analysis was done from the period of January 1, 2022, to November 18, 2022, analyzing YouTube trending videos in the US. At the time of the analysis, since 14 December 2021, YouTube has not made the dislike count public, so the analysis had to be done without the dislike count. This point became a new contribution for me, so this system was proposed and implemented to enhance this system. According to [10] found text mining to be the most effective method for interpreting words in a classification model. The merging method uses Multinomial Naïve Bayes to convert text into a feature vector for the Support Vector Machine. The Naïve Bayes-Support Vector Machine method achieves optimal classification results when data training is varied. Combining Naïve Bayes and Support Vector Machine methods yields better accuracy and stronger performance with 7:3 scale data with precision of 91%, recall of 83%, and f1 score of 87%. In my research study, data preprocessing step includes change lowercasing, tokenization, removing punctuation and URL, stopword removal, lemmatization or stemming, handling negations ("not good" or "didn't like"), handling intensifiers ("very", " extremely", or "highly"), handling emojis 😳😨 , converting abbreviations into text, special characters, and vectorization. And TF-IDF weighting techniques are used before classifying using multinomial Naive Bayes. When data ratio 80:20, ngram_range= (1,2), alpha = 0.1, the accuracy reached 90.65%.

5.2 The Analyses of Taken Time and Day to Trend by a Video

We experimented with the purpose of knowing the average times and days to be trending by category for career choice. According to my research findings, out of 15 categories, the "Nonprofits & Activism" category takes the longest time to trend with 13 hours and 46 minutes, as can be seen in Figure 11. Similarly, in Figure 12, analyzing the average days taken by category, it was observed that the "Nonprofits & Activism" category also took the most days for more than 7 trending days duration. According to the research findings, this system reveals the performant evaluation results of trending_days and negative_sentiment features using a Logistic regression model. In trending_days, the input is a categorial value and is classified based on the category according to the relevant days. The negative_sentiment feature is a boolean input value that classifies whether it is negative or not. The trending_days accuracy score is train 82% and test 76%, while negative_sentiment has train 95% and test 88%. The Nonprofits & Activism category took 13:46 hours to trend, so it took the longest time to trend compared to other categories. According to the Analysis of Videos that Days to Trend by Category, it was observed that the duration of the Nonprofits & Activism category increased by almost 8 days.

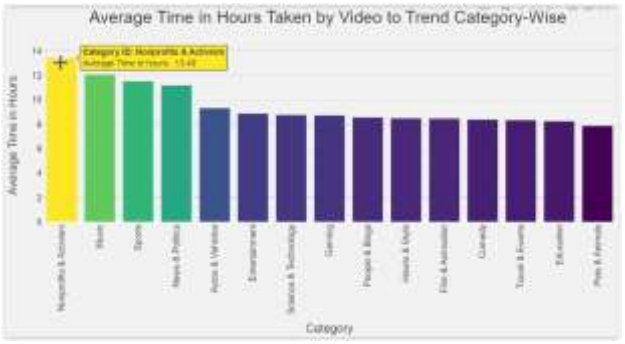


Fig. 11: The Analysis of a Video Take Times to Trend By Category

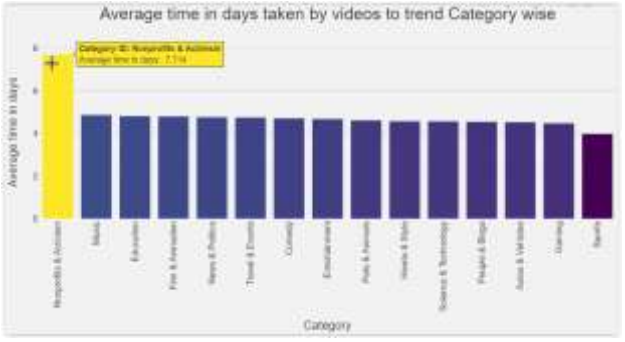


Fig. 12: The Analysis of a Video Take Days to Trend By Category

5.3 The Analyses of the Most View, Like, and Comments

Following the experiment, the most-watched views, likes, and comments are from research by category, the findings are evidenced in Figure 13, Figure 14 and Figure 15. In Figure 13, the Nonprofits & Activism and Pets & Animations categories have the least number of views, while the other categories do not have much difference in the number of views, but the Music category has the highest number of views.

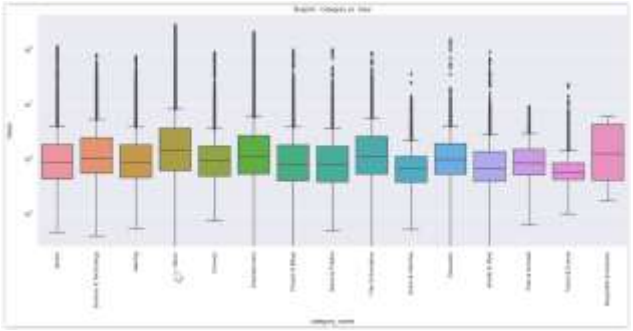


Fig. 13: The Analysis of Most Views by Category

According to the most-like evaluation, the category with the most likes is "Music". Despite their similarities, the categories of Entertainment, Education, Gaming, Sports, People & Blogs, and Film & Animation are distinct from one another. The least amount of likes have been observed in the categories of Pets & Animations and Nonprofits & Activism.

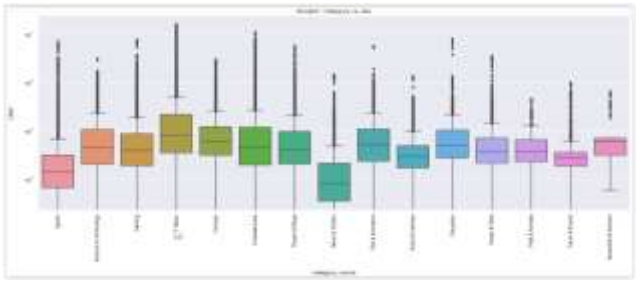


Fig. 14: The Analysis of Most Like By Category

When implementing the most comment, the category that is the most significant among all the categories is "Music". The second and third most are in the "Gaming" and "Entertainment" categories, respectively. People & Blogs, Howto & Style, News & Politics, and Film & Animation categories are not far from each other. It was found that the most views, likes, and comments were all the same in the Music category. By looking at it, it is clear that view, like, and comment are strongly correlated. A category with a high view count is likely to have a high like count. The comment count can also be high, as illustrated by the box model in Figure 15.

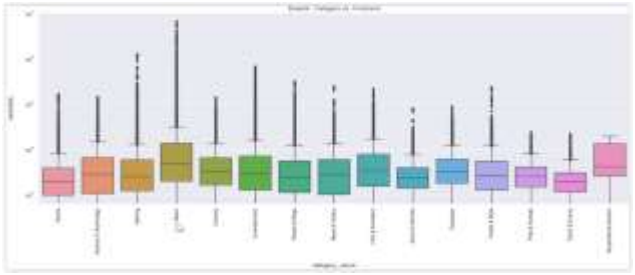


Fig. 15: The Analysis of Most Comments by Category

In the US, the title Starlink Mission by SpaceX, published on 2020 August 18 with the video-id name jTMJK7wb0rM, started trending on 2020 August 19. SpaceX designs manufactures and launches the world's most advanced rockets and spacecraft. As of December 2023, it had over 6.64 million subscribers and 526 trending videos. It can be clearly seen in Figure 16 that Starlink Mission has stood as the most-viewed and most-watched video for 4 years, with the most frequent times until today. "We broke up" and "Most Oddly Satisfying Video to Watch Before Sleep" videos were in the top 2 and top 3 in 2022, respectively, but in 2023, "Every Country On Earth Fights For $250,000!" and "7 Days Standardized At Sea" videos have found their place.

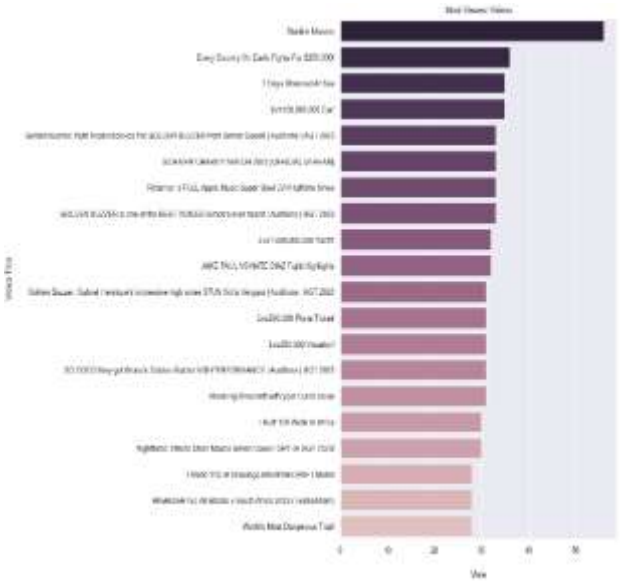


Fig. 16: The Analysis of the Most Trending and Most Watched Videos

5.4 Results Discussion

The purpose of this research was to provide support for those who will make a living on YouTube social media platforms in the US, to make the right choice when choosing a YouTube career. In doing so, career choice is also important. Trending analysis has to be done for this. Because among the many careers on YouTube, it's It is usually not easy to know which careers are trending the most. You can indeed become a successful YouTuber only if you can choose the trending, the most-watched career in your country. To help them, since December 14, 2021, when the dislike count is not shown to the public, it is not known whether a video is disliked. Sentiment analysis was done to identify it, and when doing trending analysis, it was enhanced by taking it into account. That's why we implemented a hybrid system, combining the negative results obtained from sentiment analysis with views, likes, comments, tagCount, hoursTakenToTrend, trendingDaysDuration, etc. to enhance the trending analysis. In addition to the 5 factors considered by YouTube for trending, an enhanced trending analysis was made by combining these 7 features. In the US, 16, 17, and 15 PM are the most frequently uploaded videos, and by Day, it was observed that Fridays have significantly more videos uploaded than other days. Comparing the conditions before

and after the enhancement in 2023, there was no change in the top 1 to the top 5 trending categories, but there was a change in the categories after that. Films & Animations, which was in the top 6, dropped to the top 8. In Comedy, which was in the top 7 places, we saw a slight rise to the top 6 places. Sciences and Technology, which was in the top 8, dropped slightly to the top 9. Education, which was in the top 9, fell to the top 11. News & Politics, which is in the top 10, has risen to the top 7 places. Auto & Vehicles, which was in the top 11, dropped to the top 12. Howto & Style, which was in the top 12 places, has been promoted to the top 10 places. The Travel & Events, Pets & Animals, and Nonprofits & Activism categories, which are in the Top 13, 14, and 15 places, have seen no change. The results obtained now are not analyzed based on views count alone, but based on views, likes, comments, negative, hoursTakenToTrend, trendingDaysDuration, and tagCount, so it is more accurate. The evidence proves that it is reasonable. It cannot be denied that News & Politics, which rose to 3 places, Howto & Style, which rose to 2 places, and Comedy categories, which rose to 1 place, are trending today. This system was able to display these points correctly. A low negative percentage in a category does not affect trending or not, but it is directly proportional to the likes and dislikes of the audience, so it is very useful for YouTubers who are having a hard time choosing a career. Based on this system, analysis can be done for other countries as well, so it has helped us a lot. Since it is the base language of the US is English; we did not find any difficulty in language translation when doing the sentiment, but for other countries, we would like to suggest that it may be necessary to change the language from the native language to English according to the relevant country. It has been observed that the part of combining negative features is the most difficult and important part of this system. Therefore, it was proposed as a new contribution and was successfully implemented.

For example: In the US, a video titled "Starlink Mission", channel ID "UCDVYQ4Zhbm3S2dlz7P1GBDg", channel Title "NFL", category name "Entertainment" is a trending video published on August 18, 2020. 16:07:55 and published on 2020-08-19 00 It was found to be trending at: 00:00:00 and it took 8 hours to trend. This video has 238 trending frequencies until 2023. 1578 times as channelId; 48868 times as category name; 39747 times as tags; thumbnail_link was found to be trending 37 times. The same title name "Starlink Mission" but a different channel ID "UCtI0Hodo5o5dUb67FeUjDeA" was published on 2020-08-18 16:07:55, but only trended on 2020-08-22 00:00:00, so it took 80 hours to trend. It was found that it increased for 3.3 days. The video_id "3ryID_SwU5E", title "$1 Vs. $100,000,000 House!", channel ID "UCX6OQ3DkcsbYNE6H8uQQuVA", channel title "MrBeast" published on 2023-10-14 16:00:00 2023-10-15 00:00:00. It was found that it was trending. It was found that it took almost 8 hours to become trending, that is almost taken time to trend. The video_id "qS6ozdhzSVQ", title "HYAENA", channelId "UCf_gP4AMRSgAfyzbkeS9k4g", channel title "Travis Scott - Topic" published on 2023-07-28 at 04:03:50 and trending on 2023-07-29 at 00:00:00 It was found that it took almost 20 hours to become trending. The video_id "1niAIYz6JZg", title "How Tupac Helped Master P At The Start Of His Career | No Limit Chronicles Ep 2", channelId "UCcVqCJ_9owb1zM43vqswMNQ", channel title "BETNetworks" published on 2020-08-18 20:31:11 that it was trending at 00:00:00 on 2020-08-19. It was found that it only took more than 3 hours to become trending. This was found to be minimal. It is still trending.

According to the analysis, it was observed that the "Nonprofit & Antivism" category took the longest time to trend, with 13:46 hours. The "Music" category took 12:09 hours, "Sports" took 11:54 hours, "News & Politics" took 11:02 hours, "Auto & Vehicle" took 09:32 hours, and "Entertainment" took 8:08 hours. It was found that it took time to become trending. It was found that most of the categories, like Sciences & Technology, Gaming, People & Blogs, Howto & Style, Film & Animation, Comedy, Travel & Events, and Education categories took around 8 hours. This research finding proves that the category that takes less than 7 hours is Pets & Animals.

6 Evaluation

The study has performed a linear regression and logistic regression model of Machine Learning. It is a statistical model that describes the relationship between a dependent variable and one or more independent variables, meaning that as one variable changes, the other changes proportionally by fitting a trend line to a dataset. Linear Regression measures the relationship between two variables: X and Y. X is the independent variable and Y is the dependent variable. The features such as view, like, comment, negative_sentiment, view_count_start, view_count_end, like_count_start, like_count_end, comment_count_start, comment_count_end,

tags_count, trending_date_start, trending_date_end, hoursTakenToTrend, trending_days_duration, and trending_date, and tagCount have been the variables for the analysis: YouTube trending. This analysis is performed using Python version 3.10.9, which comes with all the required libraries, Afterwards, partition the data into training (80%), testing (20%) sets, and random_state 50. The linear regression model is fitted to the training set, and its accuracy and generalizability are assessed on the testing set. Create a regression equation in the structure of Y = mX + b + e that depicts the relationship between the dependent and independent variables after fitting the linear regression model using the training set. The slope (m) and intercept (b) show the direction and strength of the relationship between the dependent and independent variables. The error term (e) captures the variability and uncertainty in the data. Lastly, assess the linear regression model's fit and predictive accuracy using metrics such as R-squared, mean squared error, and mean absolute error. R-squared measures how much of the variation in the dependent variable is explained. Mean Squared Error (MSE) and Mean Absolute Error (MAE) are loss functions that are used to utilize regression problems. MSE measures the average of the squared differences between predicted and actual values. MAE also takes the absolute value of the differences between predicted and actual values instead of squaring the error terms, which makes it inherently robust to outliers. Therefore, MAE treats all errors equally, minimizing the impact of outliers on the loss function.

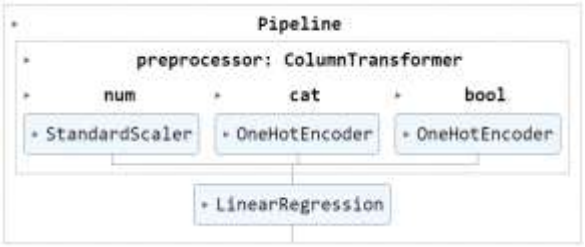


Fig. 17: Linear Regression Model Building

	Actual Result	Predicted Result	Accuracy
Views_Count	18002795	19885498.27	98.16%
Likes_Count	2408661	2562473.38	99.45%
Comments_Count	6235	6187.44	99.44%
Negative_Count	576	508.25	88.19%

Fig. 18: The Predicted Analysis Using Linear Regression Model

Figure 17 designs the structure of the Linear Regression model. It includes numerical, categorical, and boolean feature transformation and preprocessing.

Figure 19(a, b, c, d) displays the outcome of evaluating the Linear Regression model's performance following the experiment. The performance outcome for view count, which is determined by dividing the train and test, is displayed in figure 'a'. R_Squared, Mean Absolute Error, and Mean Squared Error are all measured in this estimation. There are various methods for assessing and determining a regression's goodness-of-fit to estimate the best-fit regression line for every given set of input data. A popular metric for assessing the regression's accuracy is the Standard Error of the estimation. A big Standard Error denotes a poor fit since it shows a lot of residual variation. Conversely, when the standard error is low, and there is little residual variance, indicating a strong fit. Therefore, the estimation's standard error aids in assessing the predictive accuracy of the model. Consequently, a more accurate assessment of the goodness-of-fit of the model can be determined using the coefficient of determination, or R-squared. To avoid generating a numerical error term, R-squared generates a percentage error term. The percentage shows how the regression explains the variation in the Y variable. Therefore, I am confident in my experiment's capacity to forecast because the R-squared values of 94% for a, 97% for b, and 97% for c, and the summary table for d provide an excellent fit. The actual value of each feature and the result predicted by the model are shown in Figure 18, along with the accuracy value of each feature. In view_count, the difference between actual and predicted is 1882703.27, so the model's accuracy is 98.16%, and in like_count, 153812.38 is wrongly predicted, so the model's accuracy is 99.45%. Comment_count has a small gap of 50.56%, so it has the highest accuracy with 99.44%, and negative_count has only 67.75% wrong predictions, so it is a good_fit with an accuracy of 88.19%.

In this system, views, likes, and comments are numerical / statistical values. By measuring their performance using a linear regression model, prediction results are obtained. For trending_days and negative_sentiment features, a logistic regression model is used to classify them. This is because trending_days is a categorical value and negative_sentiment is a boolean value, so it is more suitable for classification than for prediction. Therefore, Figure 19 shows the prediction results, and Figure 20 shows the classification results. This

system uses a linear regression model to compare the actual results and prediction results on the features of YouTube trending videos based on web content mining. The authors [11] analyzed trending YouTube videos from September 2020 to January 2022 based on only 2 features, views and likes. In that paper, only correlation analysis is done and only visualization is included, but there is no performance analysis part. This system is best implemented to meet these needs.

	Train	Test
R2_score	9.368426e-01	8.595938e-01
Mean Squared Error	1.049452e+13	2.625532e+13
Mean Absolute Error	7.999906e+05	8.360066e+05

(a)

	Train	Test
R2_score	9.678508e-01	9.637785e-01
Mean Squared Error	2.245183e+08	2.882497e+08
Mean Absolute Error	1.450290e+03	1.475520e+03

(b)

	Train	Test
R2_score	9.668745e-01	9.657066e-01
Mean Squared Error	9.350204e+09	9.970974e+09
Mean Absolute Error	2.056027e+04	2.154942e+04

(c)

Country	Features	R2_Score		Mean Squared Error		Mean Absolute Error	
		Train Score	Test Score	Train Score	Test Score	Train Score	Test Score
US	Views	93%	86%	1.05E+13	2.63E+13	8.00E+05	8.36E+05
	Likes	97%	96%	9.35E+09	9.97E+09	2.06E+04	2.15E+04
	Comments	97%	96%	2.25E+08	2.88E+08	1.45E+03	1.48E+03

(d)

Fig. 19: The Prediction Results of the Linear Regression Model

Country	Features	Accuracy		Precision		Recall	
		Train Score	Test Score	Train Score	Test Score	Train Score	Test Score
US	trending_days	82%	70%	75%	62%	64%	60%
	negative_sentiment	95%	88%	82%	74%	75%	61%

Fig. 20: The Classification Results of Logistic Regression Model

7 Conclusion

To sum up, this study had a lot of difficulties. However, it was able to recommend users who had difficulty choosing a career on YouTube with impressive results. Nowadays, YouTube is a social media platform used for, trending analysis, sentiment analysis, content analysis, web content mining, and so on. However, this system is a hybrid system that proposes YouTube career analysis, which is different from other analyses. US YouTube trending videos from January 1, 2023, to December 15, 2023, are used to accomplish this system. Following the experiment, sentiment analysis highlighted that the opinions of the audience from the comments of those videos can be obtained by doing sentiment analysis for attributes such as dislike that don't show a statistical count to the public. The negative result obtained from the sentiment analysis will be replaced by the dislike count. It was quite difficult to combine the negative results obtained from sentiment analysis with features such as view, like, and comment as a main contribution. YouTube's comment_count statistical values were transformed into categorical values for sentiment analysis. The transformed data is created as a dataset, and then sentiment analysis is done. When the created dataset is added back to the original dataset as a negative feature, the two datasets must match. Once the dataset is consistent and correct, the trending analysis should continue. After that, trending analysis is implemented based on mainly six features, such as view, like, negative, comment, trending_days, and tagCount, and then the obtained results are recommended for the career choice on the YouTube social media platform. Therefore, this system is very helpful for those who are choosing a career to make a living on the YouTube social media platform for the US, as well as content creators and YouTubers who are making a living as a professional life. In terms of how a video is trending, it is not decided based on views alone, but by considering views, likes, comments, and negative sentiment results, trnding_days, and tagCount so that the results of this model are reasonable and specifically correct with evidence. According to this analysis, having a negative sentiment feature has an effect on whether a video is trending or not, and considering it is reasonable, and this system proves that it has been examined from all aspects to make it more perfect for career seekers.

Acknowledgement:
The author would like to thank Dr. Soe Linn Aung, Rector of Naypyitaw Technological University, and Dr. Thet Paing Phyoe, Professor and Head of Electronic Engineering, Naypyitaw Technological University, special thanks to all the people who deserve to be thanked.

References:

[1] Sowmiya Ka, Supriya Sb and R.Subhashinic, "Scraping and Analysing YouTube Trending Videos for BI," *Smart Intelligent Computing and Communication Technology*, India, October 2021, pp: 542-547, DOI: 10.3233/APC210099.

[2] Johanes Fernandes Andry, Stefan Azriel Reynaldo, Kevin Christianto, "Alogrithm of Trending Videos on YouTube Analysis using Classification, Association and Clustering", *2021 International Conference on Data and Software Engineering, IEEE*, Indonesia, December 2021, pp: 25-30, DOI: 10.1109/ICoDSE53690.2021.9648486.

[3] S. Amudha, V R. Niveditha, P. S. Raja Kumar, M. Revathi, S. Radha Rammohan, "Youtube Trending Video Metadata Analysis Using Machine Learning,", *International Journal of Advanced Science and Technology*, Australia, Vol.29, No.7s, 18 June 2020, pp: 3028-3037.

[4] Muhammad Alkaff, Andreyan Rizky Baskara, Yohanes Hendro Wicaksono, "Sentiment Analysis of Indonesian Movie Trailer on YouTube Using Delta TF-IDF and SVM", *2020 Fifth International Conference on Informatics and Computing (ICIC)*, Indonesia, 3-4, November 2020, IEEE , pp.1-5, https://doi.org/10.1109/ICIC50835.2020.9288579.

[5] G. M. H. C. Gajanayake, T.C.Sandanayake, "Trending Pattern Identification of YouTube Gaming Channels Using Sentiment Analysis", *20th International Conference on Advances in ICT for Emerging Regions (ICTer 2020): IEEE*, 4-7 November 2020, Colombo, Sri Lanka, pp: 149-154, https://doi.org/10.1109/ICTer51097.2020.9325476.

[6] Eliganti Ramalakshmi, A Bindhu Sree Reddy, Sharvani G., "YouTube Data Analysis and Prediction of Views and Categories," *International Journal for Research in Applied Science & Engineering Technology*, India, Vol. 10, Issue VI, June 2022, pp.568-573. https://doi.org/10.22214/ijraset.2022.43636.

[7] Lau Tian Rui, Zehan Afizah Afif, R. D. Rohmat Saedudin, Aida Mustapha[4], Nazim Razali[5] "A regression approach for prediction of Youtube views", *Bulletin of Electrical Engineering and Informatics*, Malaysia, 14 January 2021, pp: 1502-1506, DOI: 10.11591/eei.v8i4.1630.

[8] Mohammed Shahid Irshad, Adarsh Anand, Mangey Ram, "Trending or not? Predictive analysis for YouTube videos", *International Journal of Systems Assurance Engineering and Management*, Springer, Sweden, August 2023, Last accepted date: 14 July 2023, pp.1-13, http://dx.doi.org/10.1007/s13198-023-02034-8.

[9] Khin Than Nyunt, Naw Thiri Wai Khin, "Customized Criteria Based Trending Analysis for YouTube Social Media Platform", *IEEE 20th International Conference on Computer Applications (ICCA)*, Myanmar, 28 February, 2023. https://doi.org/10.1109/ICCA51723.2023.10181458.

[10] Abbi Nizar Muhammad, Saiful Bukhori, Priza Pandunata, "Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier", *International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), IEEE Xplore*, Indonesia 16 October 2019, pp.199-205. https://doi.org/10.1109/ICOMITEE.2019.8920923.

[11] Md Sakibur Hasan, Bishal Sarker, Diksha Shrestha, Roshan Shrestha, Sajal N. Shrestha, "Trending YouTube Video Analysis", Research Square, Berlin, Germany, February 7th, 2023, pp,1-16 https://doi.org/10.21203/rs.3.rs-2548456/v1.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The autghor has implemented the entire system using Python.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.

APPENDIX

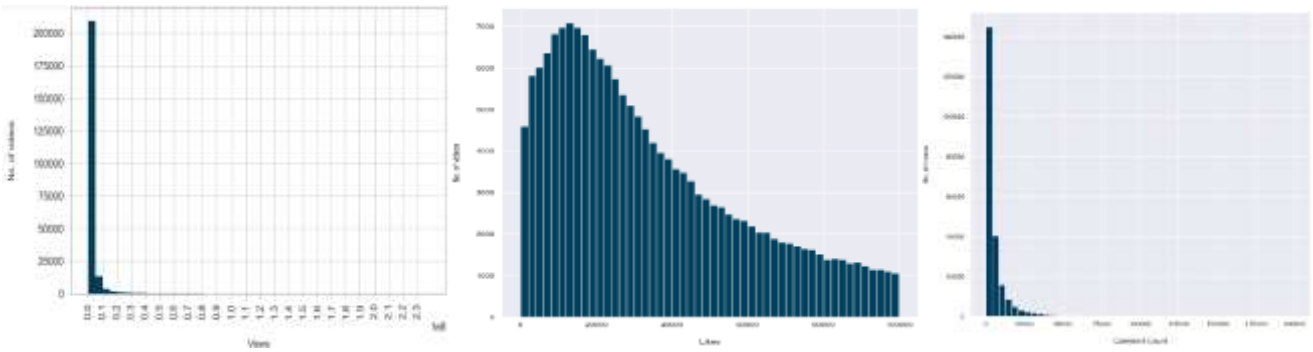


Fig. 2: The Statistical Analysis of View, Like, and Comment over Trending Videos