A Content Analysis of YouTube Trending Video Contents Using Web Content Mining Approach and Sentiment Analysis

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: - Nowadays, among the many social media platforms, the number of individuals earning a livelihood through the YouTube social media platform is rapidly increasing. In this situation, YouTube also has many channels and categories, and what is trending can vary depending on the country and year. In making a living with this platform, choosing content that is likely to be trending can help you become a successful YouTuber in a short time. It is proposed with the aim of having a significant impact on content analysts. This system is implemented by utilizing a blend of trend analysis and sentiment analysis to select the appropriate career path, with more accurate and better results. The trending analysis involves developing a new trending algorithm inspired by YouTube's trending system and analyzing the content of trending YouTube videos. Sentiment analysis is implemented using a combination of TF-IDF and Multinomial Naive Bayes models. In terms of performance, the accuracy of both analyses is good, and the results are accurate. For the trending analysis, linear regression and logistic regression were used as two approaches to calculate performance, and an accuracy value of up to 91% was achieved. The performance classification result for sentiment analysis was calculated at up to 89.54%, and it was compared with the ground truth to demonstrate its accuracy. This system uses India YouTube trending videos from Kaggle and will be very helpful and beneficial for YouTube career seekers, content analysts, and professional YouTubers.

Key-Words: - YouTube Trending Analysis; Sentiment Analysis; Statistical Count; Linear Regression; Logistics Regression; Multinomial Naïve Baye; TF-IDF, Pearson's correlation Analysis

Received: June 11, 2025. Revised: July 13, 2025. Accepted: August 16, 2025. Published: December 31, 2025.

1 Introduction

YouTube is a global platform that enables users to upload, watch, and share videos freely, generating billions of views daily. In India, the first YouTube video was uploaded on April 23, 2005, and the platform officially launched there on December 15, 2006. Its popularity surged in the late 2000s alongside the rapid growth of internet users. In India, YouTube serves various purposes, including skill development, social connection, and sharing interests. Content ID technology allows rights holders to scan videos for copyright enforcement. Notable technological advancements include the introduction of native 360-degree video in 2015, live streaming of 360° videos since April 2016, and the VR 180 stereoscopic format for virtual reality viewing. The number of YouTube users in India has grown steadily over recent years: 337.66 million in 2020, 424.69 million in 2021, 497.93 million in 2022, approximately 574.28 million in 2023, and an estimated 637.28 million in 2024, making India the leading country in YouTube user count. In 2023, analysis of trending videos reflected diverse interests across the Indian user base.

As the number of YouTube users continues to grow, so does the number of individuals who earn a living on the platform. For aspiring professionals, identifying trending careers by country is challenging without proper analysis. Making the right career choice can open opportunities to succeed as a professional YouTuber. This is the motivation behind proposing this system. According to the analysis, the top 5 trending categories in 2021 are Entertainment, People & Blogs, Music, Gaming, and Comedy as top 1 to top 5 respectively. In 2022, there was no change from top 1 to top 3, but the top 4 Gaming fell slightly to the top 5, and the top 5 Comedy rose to the top 4. In 2023, there is no change in the top 1 to top 5 categories as in 2022. Currently, with each update, 1 time for 15 minutes is not always consistent; videos on the trending list may trend up or down [1]. It is often difficult to identify which channels and categories are trending on YouTube. This challenge makes it hard for users seeking to earn income and make career decisions on the platform. To address this, research was conducted to improve the trending algorithm and to show how audience comments can reveal hidden dislike counts. The approach involved two main steps: first, creating a new algorithm based on YouTube's trending system; second, developing a method to incorporate negative sentiment analysis results into the original data as a feature. This paper is organized into the seven major portions, such as Introduction, Literature Review, Research Methodology, System Design and Implementation, Experimental Results, Evaluation, Conclusion, and References.

2 Literature Review

The study [2] analyzes YouTube trending videos using categorization, association, and clustering techniques. The study found that YouTube's algorithm for identifying popular videos is greatly affected by likes and views. It also examined how opinions, likes, dislikes, and comments influence the algorithm. The number of views was identified as the most significant factor in making a video trend, followed by likes, clicks, the number of words in the title, and keywords. According to the [3], Engagement and metadata, including title length and keywords, were found to be the two factors influencing YouTube's algorithm. The study also included Pearson's Correlation method for prediction and classification analysis. Combining Naïve Bayes and Support Vector Machine methods yields better accuracy and stronger performance, with a precision of 91%, a 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 [4]. And TF-IDF weighting techniques are used before classifying using multinomial Naive Bayes. When the data ratio is 80:20, ngram_range= (1,2), alpha = 0.1, the accuracy reached 90.65%. According to the research [5], sentiment analysis and machine learning techniques are used to identify trending patterns in YouTube gaming channels. They collected nearly 1000 popular gaming videos through the YouTube API and used preprocessing, metadata conversion, and sentiment analysis of viewer comments. They used Support Vector Machine (SVM), Naive Bayes, and Logistic Regression for classification. The study used line graphs to identify frequently used words in comment sections and bar and histogram charts to display posting time and title length. The research also reveals that video publishing time, title length, views, likes, and audio quality are essential factors in making a video trending on YouTube. The study recommends the most trending and accurate results based on YouTube statistical counts and viewer opinions.

The research [6] analyzes YouTube data and predicts views using tags and video titles. It introduces a new trending algorithm for YouTube trending videos, enhancing analysis. The system compares actual and predicted results using a linear regression model. According to research [7] compared three classifiers (SVM, Naive Bayes, and Logistic Regression) were compared for analyzing sentiments in Movie Reviews, Election Opinions, and Food Reviews. The Logistic Regression classifier performed better overall, with F1 scores for all classifiers. The study [8] suggests future research should focus on real-world scenarios, such as not seeing the dislike count, to better understand these classifiers' performance. The study [9] compares decision trees, XG Boost, and Tuned XGBoost algorithms for predicting YouTube video popularity. Tuned XGBoost outperforms the decision tree in accuracy. Factors like video quality, duration, sound features, and region-specific language popularity should be considered. The study [10] analyzed YouTube trending videos from 2007 to 2016, finding that viewers prefer short, HD-quality videos, Science & Technology, and People & Blogs categories. Viewer engagement is influenced by views. However, limitations include a 2007-2016 dataset and a focus on trending videos, which may not reflect changing viewer preferences.

The study [11] analyzed the attributes of trending YouTube videos. The study outlined five main tasks: analyzing the correlation between likes and views, comparing the average time it takes for different categories to become trending, identifying the most commonly used tags, determining the optimal title length, and finding the best day for a video to trend. It utilized data from Kaggle collected between September 2020 and January 2022, covering countries such as the USA, Canada, Britain, Japan, Germany, France, Russia, Brazil, Mexico, and South Korea. The findings offer valuable insights for content creators and users around the world. The study [12] uses a nonlinear regression technique to predict internet video recognition based on views, specifically on YouTube victimization datasets. The technique uses deep neural networks or scene dynamics metrics for quality prediction. The study [13] shows a 71% accuracy for high-peak quality videos. To improve performance, the study suggests adding channel-specific data, trending data, and scraping normal video data. According to a study [14], text mining was identified as the most effective technique for feature extraction and interpretation within classification models. The merging approach utilizes Multinomial Naïve Bayes to vectorize textual data, generating feature representations suitable for classification with Support Vector Machines. The research [15], Web scraping involves extracting useful information from web pages, predominantly through automated methods that create a Document Object Model (DOM) tree for data access. This traditional method, however, often overlooks time efficiency. The study introduces UzunExt, a novel approach that bypasses DOM tree construction, significantly increasing extraction speed by employing string methods that search for patterns, count closing HTML tags, and facilitate content retrieval. UzunExt is approximately 60 times faster than DOM-based methods and offers further time improvements by utilizing additional information gathered during crawling. This approach not only enhances extraction efficiency but can also be integrated into existing DOM-based studies to optimize performance. UzunExt, available as an open-source project, addresses limitations in traditional scraping solutions, particularly regarding Ajax requests, with future plans to automate dataset generation and improve element detection without DOM reliance. However, businesses restrict access to their data to commercialize their content, leading to concerns about computational social science becoming exclusive to large corporations, governmental organizations, and select university researchers. Researchers may need access to cloud servers and public domain computing environments for quantitative social science. The research [16] identifies the most popular YouTube categories and channels in the US in 2022 based on eight key criteria: views, likes, comments, and more. It also reveals the highest volume of video advertising. Pearson's Correlation technique was used to study trends based on various standards. The study highlights the crucial role views, likes, and comments play in a video's potential to trend. The research aims to improve YouTubers' algorithms for trending video analysis by comparing the difference principle with and without a negative sentiment feature.

In the research finding [17], 8 classification models for YouTube career analysis Multinomial Naïve Bayes (90.65%), Bernoulli Naïve Bayes (80.32%), Gaussian Naïve Bayes (89.04), Logistic Regression (91.01%), KNN (72.88%), Random Forest (51.67%), XGBRF (59.12%), and LightGBM (69.29%) were tested and obtained accuracy results respectively. Among them, Logistic Regression, which has the highest accuracy result, and Multinomial Naïve Bayes, which has the second highest accuracy, were used for classification analysis. The study [18] YouTube serves as an effective platform for promoting films through trailers, with audience comments helping producers gauge public reaction. This study performs a sentiment analysis on Indonesian movie trailer comments, categorizing them into four genres: action, romance, comedy, and horror. Using the Delta TF-IDF weighting method and various classification algorithms, it was found that Logistic Regression and Naïve Bayes excel in sentiment classification for specific genres, while the SVM model performs well generally. Specifically, Logistic Regression led in action and comedy genres, whereas Naïve Bayes was most effective for romance films. For horror, the SVM with a linear kernel achieved the best overall performance. The study recommends that producers develop personalized sentiment analysis models based on genre. Future research should focus on incorporating more data and addressing neutral sentiment, sarcasm, and irony. According to the research [19], sentiment analysis is a crucial aspect of natural language processing (NLP) due to its wide range of applications. Initially, it used traditional machine learning techniques like preprocessing and normalization. However, with advancements, researchers have shifted focus to deep learning techniques like GloVe, word2vec, and fastText, which capture text patterns and are fed into deep learning models like CNN, LSTM, and GRU. Future research should focus on fine-grained sentiment analysis and sentiment quantification for strategic decision-making. The study [20], Online video streaming services, particularly YouTube, are gaining immense popularity as they provide content creators the opportunity to share their ideas and content widely. YouTube features a trending section showcasing popular videos, while the remaining content's visibility is often unpredictable. Data analysis and mining play vital roles in enhancing business operations, especially on social media platforms. This paper aims to explore the data behind YouTube's trending videos using metrics such as Views, Comments, Likes, and Dislikes. Various classification algorithms, including Linear Regression and Decision Trees, can be employed through Python libraries like pandas and matplotlib for effective data classification and analysis.

3 System Design and Implementation

This system is designed for the composition of seven major roles: data scraping, information extraction, data preprocessing, feature extraction / combination, data analysis, performance scaling, and visualization, as shown in Figure 1. There are various methods to employ in each role. This research utilized YouTube data to analyze YouTube careers. The data was collected from YouTube's search results through the YouTube Data API v3, which provides four HTTP methods for data retrieval. The system extracted data from YouTube videos using JSON and CSV format types, extracting features based on video_id, title, publishedAt, channelId, channelTitle, categoryId, trending_date, tags, views, likes, dislikes, comments, commentDisabled, ratingDisabled, containsCapitalzed, and titleLength. The data analysis involved sentiment and trending analysis using supervised learning, machine learning techniques, and other methods. Performance scaling was performed using R_Squared, MSE, MAE for prediction, and a confusion matrix for classification. The trending results were visualized and compared to the ground truth. The system uses web content mining, machine learning, trending analysis, sentiment analysis, and YouTube career analysis to create a hybrid approach. It uses statistical analysis and term weighting techniques, with detailed explanations in 15 steps, as shown in Figure 2.

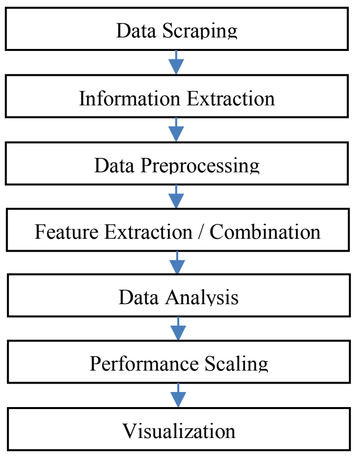

Fig. 1: Research State, Source: created by the authors

Step 1: Data acquired from the YouTube Social Media Platform

Step 2: Web Scraping using YouTube APIs Key

Step 3: Data Preprocessing: which include six steps: data profiling, data cleansing, data reduction, data transformation, data enrichment, and data validation

Step 4: Store the preprocessed data in the dataset (trending dataset)

Step 5: To create a sentiment dataset, first extract the comments feature, which is statistical count in the trending dataset (Step 4) and then convert it from a statistical value to a categorical value, which is called data conversion, and then store that data as a sentiment dataset

Step 6: Translate the comments: In India, most of the comments given by the audience as opinions on a video are only given in their native language, so they have to be translated into the same English language.

Step 7: Data preprocessing to make sentiment analysis: which consist of various preprocessing concepts: change lowercasing, tokenization, removing punctuation and URLs, stop removal, lemmatization or stemming, specially handling negations (“not good” or “didn’t like”), handling intensifiers (“very”, “extremely”, or “highly”), handling emojis 😚😇 , convert abbreviations into text (“asap” : “as soon as possible”, “afk”: “away from keyboard”), special characters and vectorization (“$” : “dollar”, “€” : “euro”).

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 sentiment dataset at Step5

Step 10: To do feature combination, extract and test the features

if (sentiment result = = -1)

negative; (use the negative sentiment to do trending analysis)

else

positive; (store the positive sentiment into the sentiment dataset at step 5)

Step 11: Combine the features view, like, and comment in the trending dataset at Step 4 and the negative sentiment 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 (prediction analysis for the statistical input: views, likes, comments)

(a) predict the outcome of observed data using R-Squared (R2), which is a statistical measurement

(b) evaluate the prediction result using loss function: Mean Squared Error (MSE) and Mean Absolute Error (MAE) to solve overfitting problem

Step 15: Create Logistic Regression Model, including training 80% and testing 20% with random state 50 (classification analysis for categorical and boolean input: trending_days, negative_sentiment)

(a) evaluate the model's classification capabilities using the confusion matrix: accuracy, precision, and recall

Step 16: Display the model's performance results

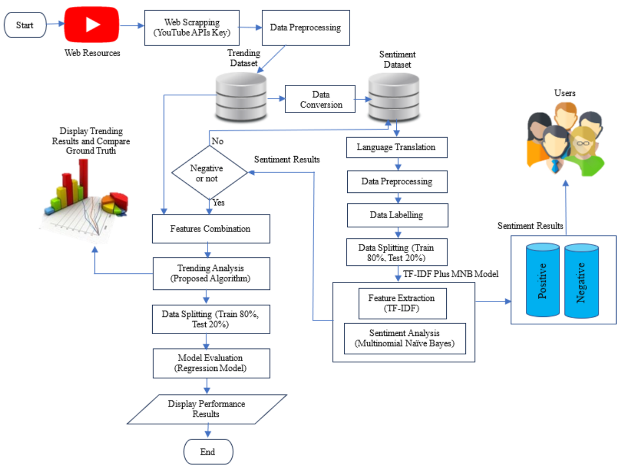

Fig. 2: The Proposed System Design, Source: created by the authors

3.1 Datasets

This system extracts streaming YouTube data from the YouTube API, scraping it to create a dataset for YouTube Career analysis. The dataset includes trending daily updated videos from various channels and categories for India, addressing the challenges faced by academics accessing raw data from major sources. The analysis of YouTube videos involves extracting the statistical count using YouTube API_KEYS to establish parameters for trending analysis. Raw data is pre-processed into usable data, followed by feature extraction to identify important attributes like views, likes, and comments. YouTube also considers "where views are coming from, including outside of YouTube" and tagCount as a system for trending analysis. The collected data include video hashtags, titles, descriptions, video IDs, channel IDs, view counts, likes, comments, and other relevant metrics. Web data extraction involves collecting specific information from online sources and transferring it into a local database or spreadsheet for subsequent analysis or retrieval. In this study, the focus is on identifying the most trending channels, categories, and videos, as well as analyzing user opinions on video content for the period from January 1, 2023, to December 15, 2023, using the assembled dataset. There are 16 columns and 68040 rows for the trending dataset shown in Table 1. Moreover, when extracting comments from YouTube videos using the official YouTube API, we cannot access all comments fully; there might be limitations like only retrieving a certain number of comments per request, potential delays in data updates, and restrictions on accessing replies to comments depending on how we structure API calls. However, for most public comments on a video, we can retrieve a significant portion of them through the API. That's why we also used a YouTube comment downloader to generate it in Python, but it doesn't say that we got it completely. Because there are also trending videos with comments disabled, and comments are prohibited based on age. So that there are 6 columns and 23443 rows for the sentiment dataset for the same period seen in Table 2. Data scraping and how to obtain a dataset can be studied in detail in [15], which I have already conducted and published.

Table 1. Trending Dataset, Source: created by the authors

video_id	title	publishedAt	channelId	channelTitl	categoryId	trending_date	tags	view_count	likes	dislike:	comment_count	thumbnail	comments_dis:	ratings_dis:	description
jnBZboK17	Thunivu (	2022-12-31T	UC3jMepk	Zee Studios	1	2023-01-01T0(	[None]	23707221	1126597	0	77117	https://i.y	FALSE	FALSE	Presenting
irDToH4Y)	ZOMBIE -	2022-12-31T	UCt4atlEx	Round2hel	24	2023-01-01T0(	zombie|z	10137419	1581988	0	132829	https://i.y	FALSE	FALSE	#skkilygam
lYoJIZxczC	I STOLE M	2022-12-30T	UCX8pnu3	Techno Ga	20	2023-01-01T0(	GTA|GTA	10444053	955897	0	53629	https://i.y	FALSE	FALSE	I STOLE MC
iF2x1IPLDI	Every Ind	2022-12-31T	UC0QKe0N	The Paayal	23	2023-01-01T0(	the paaya	1578147	74736	0	4212	https://i.y	FALSE	FALSE	Hey GuysTI
K_pLuxQF	PM Modi	2022-12-30T	UC7wXt18	Republic Bl	25	2023-01-01T0(	pm modi r	5924122	0	0	7547	https://i.y	FALSE	TRUE	PM Modi's
X2D9abM:	New Yeai	2022-12-31T	UC-ahrcxV	Easy & Sim	26	2023-01-01T0(	kolam|rar	712290	3797	0	35	https://i.y	FALSE	FALSE	New Year M
cjTKGXz5V	THUNIVU	2022-12-31T	UCZVcQLq	tamilcinem	1	2023-01-01T0(	Tamil Cine	387422	16092	0	1330	https://i.y	FALSE	FALSE	Call to +91
Her-XWTv	EMIWAY	2022-12-30T	UCHmk8iN	Emiway Ba	10	2023-01-01T0(	drill|emiw	1851153	288852	0	40699	https://i.y	FALSE	FALSE	Artist/lyrics
gMr1HSNI	PM Modi	2022-12-30T	UCtFQDgA	ANI News	25	2023-01-01T0(	modi mot	3967960	37921	0	5133	https://i.y	FALSE	FALSE	Prime Mini
51XUamSS	Hidden D	2022-12-31T	UCD3LV6q	SS Music	24	2023-01-01T0(	thunivu tr	310798	17756	0	843	https://i.y	FALSE	FALSE	Hidden Det
Oq8s3paq	Backbenc	2022-12-29T	UCLcEgVo	Tej India	24	2023-01-01T0(	backbencl	2334013	191274	0	9324	https://i.y	FALSE	FALSE	Backbench
cjINGeO9k	ALONE O	2022-12-31T	UC_J44ky(	Aashirvad (	24	2023-01-01T0(	Alone|alo	678411	35414	0	2632	https://i.y	FALSE	FALSE	Official Trai
e-_H34R6(	New Yeai	2022-12-31T	UC-OZOSC	Bindu rang	26	2023-01-01T0(	New Year	658470	5935	0	43	https://i.y	FALSE	FALSE	#Bindurang
bE51DIQh	Mashura	2022-12-29T	UCNyo-TZ	Basheer Ba	22	2023-01-01T0(	[None]	1755842	34862	0	4191	https://i.y	FALSE	FALSE	Antique Go
WpDRXjIn	PM Modi	2022-12-30T	UC1NF71E	Narendra N	25	2023-01-01T0(	Narendra	5508240	56991	0	4362	https://i.y	FALSE	FALSE	PM Modi te
5Z8C1v4n(	YES!!! SH	2022-12-30T	UCrgSOCta	Lakshmi Na	22	2023-01-01T0(	skinquiz|f	465830	0	0	3673	https://i.y	FALSE	TRUE	#lakshmina
iT0_qao7F	HAPPY NI	2022-12-31T	UC0H6I4rI	Ruchi and I	24	2023-01-01T0(	ruchi and	718938	15110	0	1717	https://i.y	FALSE	FALSE	2022 ko aui

Table 2. Sentiment Dataset, Source: created by the authors

video_id	comment_text	like	category	category_i	translated_text
-5QoGTm	[illegible]	375	Comedy	23	I can't go back and change the past but I can go forward and change my future.
-5QoGTm	[illegible]	212	Comedy	23	A time comes for every hardworking actor,,,, Prasad, now your time has started
-5QoGTm	After a long time,had so much	59	Comedy	23	After a
-5QoGTm	Dear Team, I was pretty	82	Comedy	23	Dear
-5QoGTm	[illegible]	491	Comedy	23	You have no equal in making you laugh......Maa Divorce Siris laughed out loud...Prasad garu Hatsoff.....
-5QoGTm	Entire team [illegible]	30	Comedy	23	Entire team did very well. Wish you a very good future Prasad. Script writing, dialogue delivery and timing
-5QoGTm	I love these series,I laugh a lot aft	54	Comedy	23	I love these series,I laugh a lot after so many days [illegible]
-5QoGTm	I forget to laugh heartfully for a l	16	Comedy	23	I forget to laugh heartfully for a long time, but after watching this series I laughed a lot. Thank you brothe
-5QoGTm	[illegible] charac	58	Comedy	23	Rao's character was killed in the episode by Prasad Bhaiya [illegible]
-5QoGTm	Good vibes on script & dialogue	17	Comedy	23	Good
-5QoGTm	Hats off to everyone especially di	5	Comedy	23	Hats off to everyone especially director. Best healthy comedy. Love you team [illegible]
-5QoGTm	[illegible]	37	Comedy	23	When you are not three years old, you will die of laughter after watching this video. Super bro
-5QoGTm	Insta reels chusi occhi full episod	27	Comedy	23	Insta reels chusi occhi full episode chustunna lots of fun.... [illegible]
-5QoGTm	I really laugh [illegible], a lot af	46	Comedy	23	I really laugh [illegible], a lot after long time really amazing series [illegible]
-5QoGTm	Waiting NXT episode [illegible]	12	Comedy	23	Waiting
-5QoGTm	[illegible]	89	Comedy	23	Prasad
-5QoGTm	Asalu prasad anna , neeraja gari	18	Comedy	23	Asalu

3.2 Comparison of the Two Algorithms

This section displays the comparison of the two algorithms: one, a YouTube trending algorithm, and an enhancement algorithm based on that algorithm to enhance the quality and completeness of video reviews reasonably. According to the YouTube Help Center, YouTube decides whether a video is featured on the trending list by balancing various factors. To do this, the trending algorithm evaluates numerous signals, including those outlined in YouTube's trending video algorithm.

- View count: The number of views a video has
- View velocity: How quickly the video is generating views
- View sources: Where views are coming from, including outside of YouTube
- View age: How old the video is

- Channel Performance: How the video performs compared to other recent uploads from the same channel

By looking at the above factors, YouTube determines whether a video is trending. It is generally considered to be based on view count alone among the many features. The more views a video receives, the more it appears trending. However, this often means that the audience is not actually watching the video content. Because YouTube increases the view count by 1 after a 30-seconds playback. This system combines trending concepts with audience preferences, likes, and dislikes, and users' opinions to understand what they don't watch. According to the research findings, views, likes, comments, and negative are strongly correlated and should be considered. YouTube career seekers must choose a category that audiences stick to long-term, as recent ephemera are short-lived. So, the following enhanced YouTube's video trending algorithm, including 11 steps, aims to benefit many content creators by identifying long-term trends and focusing on long-term audiences. And also consider choosing a category that aligns with passions and expertise – some of the most popular and potentially lucrative categories – remember to pick a niche within the chosen category to stand out and attract a dedicated audience.

Step1: Create new features view_count_start and view_count_end from view and then create function min(views), max(views)

Step2: Create new features likes_start and likes_end from likes and then create function min(likes), max (likes)

Step3: Create new features comment_count_start and comment_count_end from comment and then create function min(comments), max(comments)

Step4: Create new feature tags_count from tags and then create function tags ()

Step5: Create new features trending_date_start and trending_date_end from trending_date and then create function min(trending_date), max(trending_date)

Step6: Create a new feature hoursTakenToTrend from publishedAt and trending_date_start and then create a function hoursTakenToTrend()

Step7: Create a new feature trendingDaysDuration from publishedAt and trending_date_start, and then create function trendingDaysDuration()

Step8: Create new feature negative obtained from sentiment analysis and then create a function negative () to get the opinions of YouTube users and test whether negative (-1) or not, if negative will be used, if not, store the data into the sentiment dataset

Step9: 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 most trending categories and channels

Step10: Visualize the trending result compared to Ground Truth

Step 11: Calculate the overall accuracy

While YouTube's trending video algorithm can only display whether each video is trending or not, it cannot show what the most trending channel / category among the numerous video content. Therefore, as a career seeker, they can only know which channels and categories are trending in their country by doing an analysis. It is not easy to do that, so this analysis is done by them. The proposed new YouTube trending algorithm's workings have been thoroughly explained, and its many advantages are listed below:

- Offer comprehensive support for the YouTube career seeker, considering both trending trends and the audience's preferences
- Allow users to track the most frequently uploaded video content through timely and daily updates
- Indicate the time it takes for a video content to trend and the number of days it takes for the content to be trending
- Explore the trending days' duration of a video content
- Determine the most trending channels and categories
- Support what kind of YouTube video content you will create
- Support YouTube video content research skills

3.3 Sentiment Analysis

This paper presents the results of trending YouTube videos analyzed using features such as views, likes, comments, negative comments, trending days, and tag counts. The YouTube trending algorithm is not solely based on view count, as it is impossible to determine who is watching the video until the end. Additionally, negative comments play an important role in understanding a video's overall reception. YouTube has stopped showing dislike counts to the public since December 14, 2021, making it difficult to know the dislike count [4]. To know the dislike count, the system uses negative sentiment results obtained through comments. Sentiment analysis is employed to capture genuine audience opinions, extending beyond mere view counts. Decisions are based on a combination of metrics such as views, likes, comments, negative comments, trending days, and tag counts. When evaluating trending videos, viewers should consider whether the video is also among the most watched. If the recommended categories and channels align with the most trending, most viewed, most liked, and most commented videos, this system can assist users in making more informed career decisions. The insights from these analyses are integrated into a hybrid model, specifically tailored for India, to help individuals choose the most suitable career path on the platform.

YouTube trending analysis uses negative sentiment results from audience comments to calculate dislike counts. Sentiment analysis encompasses essential system components such as language translation, preprocessing, term weighting techniques, labeling, and classification. The Term Frequency-Inverse Document Frequency (TF-IDF) model and the Multinomial Naive Bayes algorithm are employed to analyze textual data. Text documents are converted into TF-IDF feature matrices, which quantify the relevance of words. These texts are then labeled with positive or negative sentiment annotations and divided into training and testing datasets for model development and evaluation. The Multinomial Naive Bayes model is applied to the TF-IDF features, and models are trained using the labelled data [17se]. Performance metrics, including accuracy, are calculated to assess predictions' quality.

	video_id	comment_text	like	category	category_id	translated_text	clean_text
0	-5QoGTmcSUE	[illegible] ...	375	Comedy	23	I can't go back and change the past but I can ...	back change past forward change future
1	-5QoGTmcSUE	[illegible] actor ki [illegible] ,,, Pr...	212	Comedy	23	A time comes for every hardworking actor,,, P...	time come every hardworking actor prasad time ...
2	-5QoGTmcSUE	After a long time,had so much laughter Clean...	59	Comedy	23	After a long time,had so much laughter Clean...	long timehad much laughter clean comedy prasad...
3	-5QoGTmcSUE	Dear Team, I was pretty impressed. It's very d...	82	Comedy	23	Dear Team, I was pretty impressed. It's very d...	dear team pretty impressed decent watch family...
4	-5QoGTmcSUE	[illegible]......	491	Comedy	23	You have no equal in making you laugh.....Maa...	equal making laughmaa divorce siris laughed lo...
...	...	...	...	...	...	...	...
95	-EAUNDgIVnQ	16:08 jetha in action	16	Comedy	23	16:08 jetha in action	1608 jetha action
96	-EAUNDgIVnQ	Old episode best comedy super	10	Comedy	23	Old episode best comedy super	episode best comedy super
97	-EAUNDgIVnQ	I respect to everyone who were involved in thi...	64	Comedy	23	I respect to everyone who were involved in thi...	respect everyone involved seriously best piece...
98	-EAUNDgIVnQ	Jethalal Singham style ekdam faadu	5	Comedy	23	Jethalal Singham style ekdam faadu ???	jethalal singham style ekdam faadu
99	-EAUNDgIVnQ	Old is gold jo purane episodes mein maza ati...	19	Comedy	23	Old is gold jo old episodes mein maza ati hai ...	gold episode mein maza alag replace

Fig. 3: Getting the Results of comment_text, translated_text, and clean_text, Source: created by the authors

video_id	like	translated_text	cleann_text	Sentiment
..L_wDHUaxY	16K	Guys we crossed "14M subscribers" Than	crossed subwcriber than you much love support	positive
..L_wDHUaxY	1.4K	Who is the realfan of Techno Gamerz	real techno gamerz	positive
..L_wDHUaxY	82	Ujjwal proofed that he can entertain us wit	ujjwal proofed entertain without using meme tecl	positive
..L_wDHUaxY	40	Fact: Ujj never fail us to laugh	fact ujjwal never faile laugh	negative
..L_wDHUaxY	6	Congratulations Bhai For 14M	congratulation bhai million	positive
..L_wDHUaxY	133	How many will your mother live???. Ignor	many mother live 1 ignor year subscription year	negative
..L_wDHUaxY	45	He is the only guy who entertains us witho	entertains without meme gaalies	positive
..L_wDHUaxY	33	Congratulations in advance for 20M subscr	congratulation advance subscribers	positive

Fig. 4: Labelling Results Using Valence Aware Dictionary and Sentiment Reasoner (VADER), Source: created by the authors

Number of Terms

1	add	android	bank	bin	car	charneys	china	dog	dov	reedeeming	regard	roll	story	tricked	tv	uncertain	wearable	white	yuan	...
2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.420669	0.42067	0.420669	0.00000	0.00000	0.00000	0.00000	0.00000	...
3	0.00000	0.00000	0.00000	0.00000	0.00000	0.50000	0.00000	0.00000	0.50000	0.50000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	...
4	0.37796	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.377964	0.00000	0.00000	0.377964	0.00000	0.00000	0.00000	0.00000	0.377964	0.00000	0.377964	0.00000	...
5	0.00000	0.463693	0.00000	0.00000	0.463693	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.4637	0.00000	0.463693	0.00000	0.00000	...
6	0.00000	0.00000	0.37796	0.37796	0.00000	0.00000	0.377964	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.377964	...
7	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Number of Documents

Fig. 5: Terms Weighting Results Using TF-IDF, Source: created by the authors

Youtube Comments Sentiment Analysis

	video_id	like	translated_text	clean_text	sentiment	category	category_id
1	--L_wDHUaxY	16K	Guys we crossed "14M subscribers" Than	crossed subscriber thankyou much love support	positive	Gaming	20
2	--L_wDHUaxY	1.4K	Who is the real fan of Techno Gamerz	real techno gamerz	positive	Gaming	20
3	--L_wDHUaxY	82	Ujwal proofed that he can entertain us wi	ujwal proofed entertain without using meme tec	positive	Gaming	20
4	--L_wDHUaxY	8	Congratulations for 36.6 million subs	congratulation million subscriber	positive	Gaming	20
5	--L_wDHUaxY	33	Congratulations in advance for 20M subsc	congratulation advance subscriber	positive	Gaming	20
6	--L_wDHUaxY	40	Fact Ujwal never fail us to laugh	fact ujwal never fail laugh	negative	Gaming	20
7	--L_wDHUaxY	45	He is the only guy who entertains us wit	entertains without meme gaalies	positive	Gaming	20
8	--L_wDHUaxY	6	Congratulations Bhai For 14 Million	congratulation bhai million	positive	Gaming	20
9	--L_wDHUaxY	99	He provw that he can entertain us withou	provw entertain without using meme abuseive la	positive	Gaming	20
10	--L_wDHUaxY	0	Congratulations for 14M	congratulation	positive	Gaming	20
11	--L_wDHUaxY	32	Techno bro entertain us with out memes an	techno entertain meme watch video full concentr	positive	Gaming	20
12	--L_wDHUaxY	1	Congratulations for 36 million subscribers	congratulation million subscriber	positive	Gaming	20
13	--L_wDHUaxY	64	Only legends know that first he played gam	legend know first played game offline knew play	positive	Gaming	20
14	--L_wDHUaxY	46	Congratulations for your 14 million subscr	congratulation million subscriber intro getting ev	positive	Gaming	20
15	--L_wDHUaxY	114	"Ujwal proved that he can make us laugh w	ujwal proved make laugh skill editing memeste	positive	Gaming	20
16	--L_wDHUaxY	5	Congratulations for 14 million subscribers	congratulation million subscriber game play sto	positive	Gaming	20
17	--L_wDHUaxY	9	Congratulations for 14 Million Sir Keep it	congratulation million keep	positive	Gaming	20
18	--L_wDHUaxY	15	Legends will watch it after 2 years	legend watch year	positive	Gaming	20
19	--L_wDHUaxY	28	Congratulations for 14M	congratulation	positive	Gaming	20
20	--L_wDHUaxY	6	Congratulations for 14 Million Bro	congratulation million	positive	Gaming	20
21	--L_wDHUaxY	133	How many will your mother live???1). Igno	many mother live1 ignore year subscription year	negative	Gaming	20
22	--L_wDHUaxY	88	Congratulations bro ek heart dedo	congratulation heart dedo	positive	Gaming	20
23	--L_wDHUaxY	51	Congratulations	congratulation	positive	Gaming	20

Fig. 6: Sentiment Classification Results of India YouTube trending videos, Source: created by the authors

Figure 3 displays the results of each step, including comment_text, translated_text, and clean_text, which are then translated and pre-processed after receiving the comments. Figure 4 shows the pre-processed comments labeled using the Valence Aware Dictionary and Sentiment Reasoner (VADER), categorized into two classes: positive and negative. Figure 5 illustrates the application of the Term Frequency-Inverse Document Frequency (TF-IDF) method to weight terms within the documents. The results of sentiment classification are presented in Figure 6, which includes details such as video ID, sentiment-labeled comments, likes, translated text, cleaned text, category ID, and category name. This system effectively assesses viewers' sentiments toward the content, providing valuable insights not only for individuals aiming to start a career on YouTube but also for current YouTubers in their professional journey.

During analysis, we observed that some comments contain slang and misspellings, which pose challenges for accurate classification. Despite these issues and the resulting noise in the data, the Multinomial Naive Bayes model demonstrated stability, achieving the highest accuracy of nearly 90%, as shown in Figure 7. Eight classification models for YouTube career analysis, Multinomial Naïve Bayes (90.65%), Bernoulli Naïve Bayes (80.32%), Gaussian Naïve Bayes (89.04), KNN (72.88%), Random Forest (51.67%), XGBRF (59.12%), and LightGBM (69.29%) were tested and obtained accuracy results respectively. Among them, Multinomial Naïve Bayes, which has the highest accuracy, was used for sentiment analysis. MNB is primarily used to design for discrete data, particularly when dealing with counts of frequencies, like the number of times a word appears in a document. The ratio for the MNB model is (80: 20) 80% training and 20% testing, with random 50. TF-IDF techniques are also used to enhance the accuracy of the MNB model, so the result is more accurate, and the execution time is saved because the low IDF score will effectively downweigh them, essentially filtering them out when analyzing the document's key terms. Finally, we got the sentiment distribution implemented as seen in Figure 8, including the result of sentiment classification results negative (22.58%) and positive (77.42%) out of the 23443 comments. Then, negative sentiment results were combined according to video_id as a feature in the original trending dataset. Data validation was necessary to ensure accuracy. The sentiment analysis only took negative sentiment results, while the performance evaluation part was conducted. The performance evaluation results were not used or combined. Figure 8. (left) displays the negative sentiment percentages of each video content and also shows negative sentiment percentages by category at Figure 8. (right). When the negative sentiment results are combined as a feature, that more or less negative sentiment percentage is included by category. 30.07% in News & Politics, 14.28% in Nonprofits & Activism, 8.72% in Gaming, 8.71% in Comedy, 8.64% in Film & Animation, etc., respectively. In Autos & Vehicles, the lowest negative sentiment content is found at 0.52%. After that, the completed dataset that has been validated several times is obtained, and the trending analysis is continued.

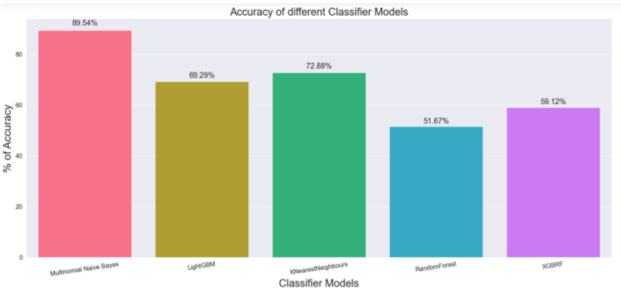

Fig. 7: Performance Classification Results of Sentiment Analysis, Source: created by the authors

	video_id	clean_text	neg sentiment %
0	-5QoGTmcSUE	[back change past forward change future, time ...	20.0
1	-Aod3YCLKDs	[extra show faltu kharcha pure proposal family...	33.333333
2	-CnLhGzVimg	[fact everyone waiting himlands, myth said zom...	33.333333
3	-EAUNDgIVnQ	[whose scooter roaming around sits home, chara...	16.666667
4	-JkWALLHYNI	[comment likes, game abhyudaya giving correct ...	20.0
...	...	...	...
780	zduyUKUTlzs	[want saurabh bhais family well soon, love sup...	13.333333
781	zg2bw9W2gtE	[khesari song , belief sings touch heart, hind...	13.333333
782	zmFygS2oGkw	[drop everything subscribe geos youtube channe...	16.666667
783	zmOpWEedTHk	[another masterpice kannada industry, kannada ...	20.0
784	zu48C7VJhh0	[happy birthday preman babu, happy back, belat...	0.0

```
News & Politics            30.070922
Nonprofits & Activism      14.285714
Gaming                      8.724280
Comedy                      8.717949
Film & Animation            8.645533
Education                   8.510638
Music                       6.168934
Entertainment               5.647654
Sports                      4.113924
People & Blogs              3.810695
Pets & Animals              3.260870
Science & Technology        3.225806
Howto & Style               1.261261
Travel & Events             0.934579
Autos & Vehicles            0.526316
Name: categoryId, dtype: float64
```

Fig. 8: Negative Sentiments Distributions by video_id and Category, Source: created by the authors

3.4 Trending Analysis

After the dataset was reconstructed, trending analysis was performed on the complete dataset. Following the feature combination, we created additional features based on the characteristics identified in the proposed YouTube trending video algorithm. For example, from the time a video content is posted, we created additional features to determine the number of view counts recorded between the start and end times of the view count measurement. Similarly, features such as view_count_start, view_count_end, like_count_start, like_count_end, comment_count_start, comment_count_end, hourTakenToTrend, trendingDaysDuration, and sentiment_scores, which were not in the original dataset, were created again. Finally, an enhanced new Data Frame with 23 columns was successfully created, which can be seen at Figure 9 (left). To assess the correlation between features in the dataset, we first performed statistical analysis using Pearson's correlation method. This method measures the relationship between different video contents based on their features.

```
Out [67]: < class 'pandas. core. frame. DataFrame'>
RangeIndex: 23443 entries, 0 to 23442
Data columns (total 23 columns):
#    Column                  Dtype
0    video_id                object
1    title                   object
2    channelTitle            object
3    categoryId              object
4    tags                    object
5    comments_disabled       bool
6    ratings_diabled         bool
7    description             object
8    likes_start             int64
9    likes_end               int64
10   view_count_start        int64
11   view_count_end          int64
12   tagCount                int64
13   hoursTakenToTrend       float64
14   trendingDaysDuration    int64
15   comment_count_start     int64
16   comment_count_end       int64
17   dayofweek               object
18   Year                    int64
19   Month                   int64
20   negative_sentiment      int64
21   videoDuration           float64
22   madeForKids             bool
Dtype: object (7). bool (3). int64 (11). float64 (2)
```

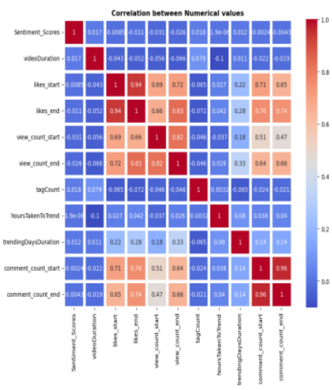

Fig. 9: The New Data Frame and the Results of Statistical Analysis, Source: created by the authors

The confusion matrix on the right side of Figure 9 shows the result based on the final enhanced data frame. By looking at it, one can easily see that views, likes, and comments are strongly correlated 0.94, 0.83, 0.82, 0.72, 0.69, 0.66 as respectively. We were able to conduct detailed and accurate research on the time a video starts getting view counts, the time the view count ends, the time the like count starts, the time the like count ends, the time the comment count starts, and the time the comment count ends after a video is uploaded. In addition, since the sentiment score is a dislike-like feature, we were able to reveal the correlation between the unreleased dislike feature and other features. Furthermore, we can know the importance of the tagCount feature for a video to become a trend, depending on its correlation with other features. The tagCount feature is also used to calculate where the increase in video count comes from outside YouTube. Contributing a feature called hoursTakenToTrend also allowed us to study how long it takes for a video to become a trend, how it relates to other features to become a trend, how to become a trend quickly, what to focus on to become a trend, and when to upload. The trendingDaysDuration feature also allows us to know how long our uploaded video content stays in the trending list, and we found that features such as views, likes, and comments are related to 0.33, 0.18, and 0.14, respectively to stay trending longer. The correlation values in the figure also prove how much each enhanced feature is related to each other. Therefore, the system irrefutably proves that the accurate trending results are due to these enhanced features.

Correlation analysis is a technique used to assess the relationship between two variables. It helps identify trends, patterns, and interactions between datasets or features. A high correlation suggests a strong relationship between the variables, while a low correlation indicates a weaker connection. Positive correlations occur when one feature increases and the other do too, while negative correlations indicate a negative relationship. The correlation coefficient ranges from +1 to -1, with 1 indicating high correlation and 0 indicating no correlation. The research reveals that it typically takes 10 -18 hours (q1: 10.5, trace 0) to (q3: 18, trace 0) in 24 hours for a video content to become trending in India, as shown in the Box Plot, Figure 10. The time is displayed on the X-axis of the histogram, while the number of movies is displayed on the Y-axis. Some videos become trending quickly, while others take days. The Histogram of the image shows in Figure 10 that 2710 of the 3068 videos took 10 hours and 9 minutes, while 358 took more than 18 hours. The analysis of the hourly range of videos showed that 139 videos took 1 - 2 hours, 198 took 3 - 4 hours, 562 took 5 - 6 hours, 1560 took 7 - 8 hours, 2710 took 9 - 10 hours, 2832 took 11 - 12 hours, 2001 took 13 - 14 hours, 2318 took 15 - 16 hours, 3068 took 17 - 18 hours, 2282 took 19 - 20 hours, 752 took 21 - 22 hours, and 167 took 23 - 24 hours.

Fig. 10: The Number of Hours Taken by a Video to Get into the Trending List, Source: created by the authors

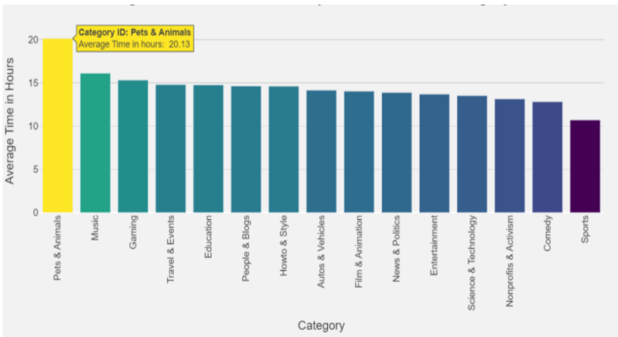

Fig. 11: The Analysis of Video Times Taken to Trend by Category, Source: created by the authors

We conducted an experiment to determine the average time it takes for each category related to career choice to become trending, as shown in Figure 11. Our findings indicate that, among the 15 categories analyzed, the "Pets & Animals" category remains trending the longest, with an average duration of 20 hours and 13 minutes. The trending time for various categories is varying, with “Music” taking over 16 hours, “Gaming” taking 15:32, “Travel & Events” taking 14:7, “Education”, “People & Blogs” taking 14 hours, “Howto & Style”, “Pets & Animation”, “News & Politics”, “Entertainment”, “Science & Technology” taking 13 hours, “Nonprofits & Activism” taking 13 hours, “Comedy” taking 12:8 hours, and “Sports” taking 10:7 minutes. The study reveals that the trending days duration results for each category are shown in Figure 12. More than 3 days in "Music", "Comedy", "Film & Animation", "Pets & Animals" categories, almost 3 days. "People & Blogs", "Howto & Style", "Autos & Vehicles", "Gaming" for almost 2 and a half days. "Travel & Event", "Sports", "Entertainment", and "Nonprofits & Activism" have been trending for almost 2 days. "Science & Technology", "Education", and "News & Politics" are trending for more than 1 day.

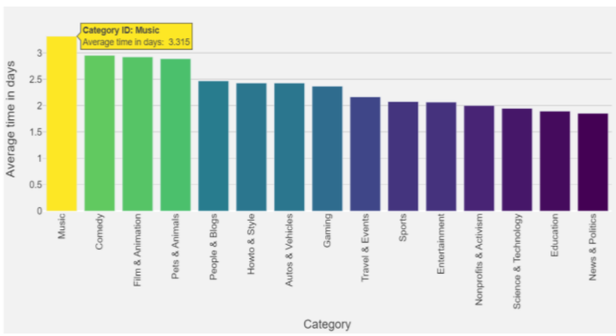

Fig. 12: The Analysis Trending Days Duration by Category, Source: created by the authors

In addition, in this study, based on the view, both likers and commenters were able to analyze together by category. The category with the most significant likes on view is Gaming. "Science & Technology", "Comedy", "Travel & Events", "Education", and "Music" did not have much difference, but in "News & Politics", it was found to be only 0.025% least. In another comparison, the most commented category among viewers was found to be "Music".

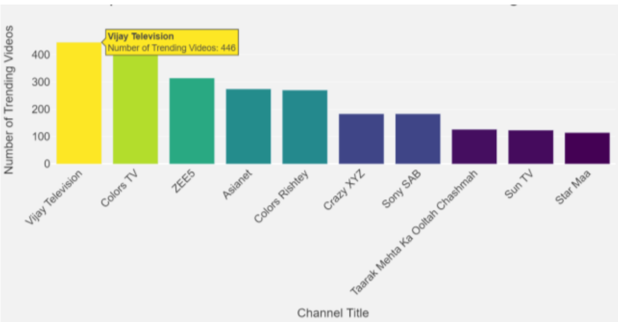

Fig. 13: The Most Trending Channels in India, Source: created by the authors

The Top 10 most trending channels in India include "Vijay Television" with 446K videos, followed by Colors TV (408), ZEE5 (315), Asianet (275), and Color Rishtey (271). Crazy XYZ and Sony SAB channels also have 184 trends. Taarak mehta Ka ooltah Chashmah, Sun TV, and Star Maa also have 127 and 124 videos, respectively, seen in Figure 13. According to this study, the most trending categories were also successfully implemented, seen in Figure 14. The Entertainment category is the most trending, with the number of trending videos 8042. It obviously stands in the Top 1 category than other categories. "People & Blogs" is the second most trending, with 2477 trending videos. Music, which is the third most trending, has 2311 trending videos, which is not too different from it. Comedy, Gaming, and Science & Technology are trending with 1336, 818, and 727 trending videos, respectively. The system-generated results will provide users with accurate and reliable information due to the inclusion of key features.

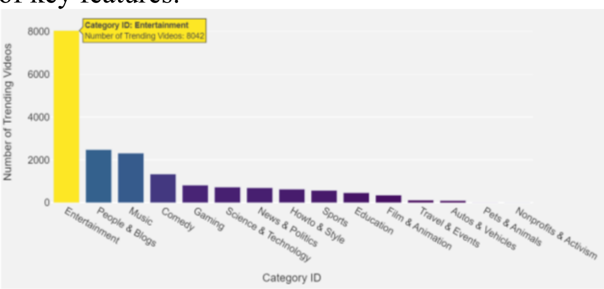

Fig. 14: The Most Trending Category in India, Source: created by the authors

According to the research findings, in Figure 15, which appears at a time, the variety of those who upload videos dramatically increases from 13PM, 12PM, and 05AM, which are the most video upload times of 20,000 videos in India. There were barely any video uploads around 21PM, and there were under 5,000 at 17PM. The days of video uploading are also not different from one day to another throughout the week in India. But the best video upload day is Friday. On that day, nearly 35,000 video uploads were achieved. Despite being the least active day of the week, Sunday had about 30,000 video uploads in India.

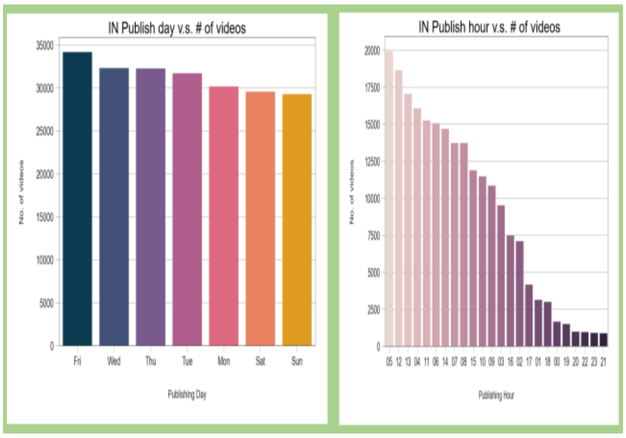

Fig. 15: The Analyses on the Best Day and Time to Upload videos on YouTube for India, Source: created by the authors

According to the research findings, the maximum common title “Kundali Bhagya” 55 times was revealed, with channelId "UCvrhwpnp2DHYQ1CbXby9ypQ", channelTitle "Vijay Television", 3035 times, and categoryName "Entertainment" and categoryId "24", 89380 times seen in Figure 16. In order to prove this, this study revealed the most frequent title as well as the most trending and watched video, and found that the top 1 video is “Kundali Bhagya” and it is correct and consistent according to Figure 17.

	index	title	channelId	channelTitle	category	category_	tags	thumbnail_link	descriptio	country
1	count	230244	230244	230243	230244	230244	230244	230244	230244	230244
2	unique	72685	4403	4645	15	15	45409	71680	61218	1
3	top	Kundali Bhagya	UCvrhwpnp2DHYQ1CbXby9ypQ	Vijay Television	Entertainment	24	[None]	https://i.ytimg.com/vi/y_RHQ4mBDKA/defau		IN
4	freq	55	3035	3035	89380	89380	41173	19	17806	230244

Fig. 16: The Most Frequent Title Among Trending Titles in India, Source: created by the authors

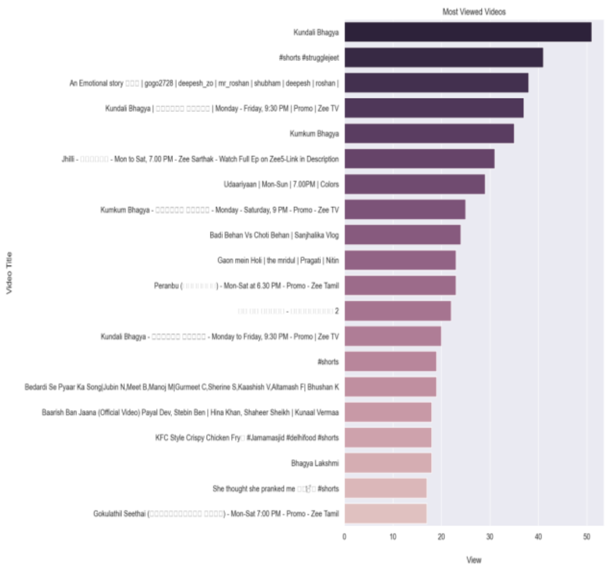

Fig. 17: The Most Trending and Watched Video in India, Source: created by the authors

3.5 Challenges

The dataset is also crucial to the success of this system. The data is unstable when it is updated daily since it is a private dataset that is scraped every day, and the results remain unstable until the entire year of analysis is finished. This is only one of the many challenges. Having encountered several challenges because there are restrictions on getting audience feedback while conducting sentiment analysis. Ultimately, we created it with the YouTube comment downloader, and it was successful. Google had a daily restriction, which made the language translation process take a long time. After sentiment analysis, negative sentiment results were obtained, but the feature combination part was the most challenging. Even though we tried using various techniques, the data did not match. Finally, only after grouping the categories and adding them according to video_id, the comments given by the audiences according to the video match, and the data validation part is successful. Since each feature's input value differs, it is impossible to assess them all using a single model in the evaluation stage. Because the inputs are boolean (0,1), categorical (text), and statistical (number), we had to transform and then evaluate them. Regression prediction and classification were the two components of the performance evaluation section; thus, we had to employ distinct methods to measure each component's successful outcome. Since it just contains raw data from Kaggle, it is difficult to compare with ground truth. As a result, it must be contrasted with the outcome that was obtained after the trending result was initially produced. Three different types of trending outcomes are clearly compared in the evaluation section: the trending result achieved using the ground truth dataset, the enhanced trending result with negative features, and the trending result without negative features using the own dataset. The challenges that arose throughout the research are highlighted.

4 Performance Evaluation

The study implemented both linear regression and logistic regression models in machine learning.

4.1 Evaluation Results of Linear Regression Model

Linear regression is a statistical technique that is not suitable for classification tasks, where the goal is to predict the category or class to which an observation belongs. Logistic regression is a classification technique used in machine learning to identify the best model that uses probability to classify the value of linear regression into a particular class. Therefore, it is used to classify the days of the week (trending_dates), and negative_sentiment (negative_feature) in trending analysis. The logistic regression model converts the continuous output of a linear function into a categorical value by applying a sigmoid function. This function maps any real-valued input from the independent variables into a probability between 0 and 1, representing a binary outcome. Figure 18 illustrates the structure of the regression model, which includes the transformation and preprocessing of numerical, categorical, and boolean features.

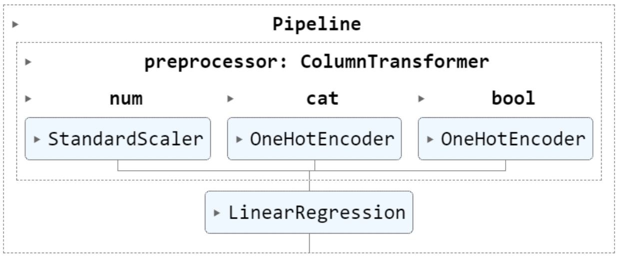

Fig. 18: Regression Model Building, Source: created by the authors

The equation Y = b0 + b1x1 + b2x2 + b3x3 + ... + bnxn is the linear part of the logistic regression model. To transform this into a probability, the logistic function is applied, which involves calculating the odds as Y/ (1 - Y). Taking the natural logarithm of the odds yields the log-odds (or logit):

$$\log\left(\frac{Y}{1-Y}\right) = b0+b1x1+b2x2+b3x3+...+bnxn$$

This log-odds value ranges from negative to positive infinity, corresponding to probabilities between 0 and 1.

According to the research findings, this system reveals the performant prediction results of view, like, and comment features over a trending video seen in Figure 19. The Y-axis in this figure shows the features, while the X-axis shows the statistical count of the YouTube trending video. The model's forecast result is 2059511, despite the fact that the actual count for the View feature is 2185108. The performance measurement showed that the model was 94.25% accurate, with 125597 incorrect predictions. The performance evaluation showed that the Like feature was 94.47% accurate, with the actual count of 1527731 being 1617142 and 89411 being incorrectly predicted. The performance measurement showed that the comment function was 98.77% accurate, with the actual count of 799198 being forecasted as 809125 and 9927 being incorrectly predicted. Based on the regression model's performance matrix for the entire system, Table. 3 displays the evaluation of the test and train scores. The test score in the R2_Score view is 85%, while the train score is 88%. Like received a test score of 85%, but the train score was 92%. While Comment's test score was 91%, the train score was 96%.

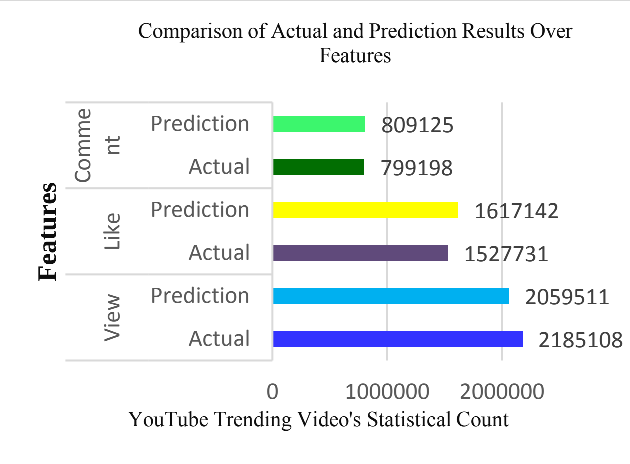

Fig. 19: Comparison of Actual and Predicted Results Over the Features of a Video Content, Source: created by the authors

Table 3. Performance Classification Results Using Linear Regression Model, Source: created by the authors

Country		R2_Score		Mean Squared Error		Mean Absolute Error	
		Train Score	Test Score	Train Score	Test Score	Train Score	Test Score
India	Views	88%	85%	1.20E+13	1.48E+13	1.11E+06	1.18E+06
	Likes	92%	85%	2.24E+10	2.43E+10	3.93E+04	4.15E+04
	Comments	96%	91%	3.40E+08	2.45E+08	1.78E+03	1.85E+03

4.2 Evaluation Results of Logistic Regression Model

Logistic regression model evaluation involves assessing predictor significance, evaluating model fit, and measuring predictive performance using metrics like accuracy, precision, recall, F1-score, and ROC/AUC. The choice of binary classification threshold influences metrics, and a confusion matrix visualizes performance on correctly and incorrectly classified instances. This system also provides performance prediction results for the trending_days and negative_sentiment features of each trending YouTube video in India. The trending_days feature is categorical, classified based on the number of days a video remains trending. The negative_sentiment feature is a boolean indicator that determines whether the sentiment is negative. In India, the model achieves an accuracy of 94% on the training set and 88% on the test set for trending_days. For negative_sentiment, the accuracy is 97% on training data and 91% on testing data, as shown in Table 4.

Table 4. Performance Prediction Results Using Logistic Regression Model, Source: created by the authors

Country	Features	Accuracy		Precision		Recall	
		Train Score	Test Score	Train Score	Test Score	Train Score	Test Score
India	trending_days	94%	88%	73%	68%	91%	82%
	negative_sentiment	97%	91%	85%	83%	100%	87%

4.3 Comparing and Illustrating the Ground Truth Results

This section provides a detailed explanation of a histogram that compares the ground truth trending result and the enhanced trending result, as seen in Figure 20. In that figure, the X-axis shows the category names along with colors differentiated by India's top levels. The Y-axis represents the YouTube Level in India. When comparing the enhanced trending result with the ground truth result, it was found that 10 out of 15 YouTube trending categories were identical, while 5 categories showed slight changes. The top 1, 2, 3, 8, 9, 10, 12, 13, 14, and 15 categories were found to be identical to the ground truth result. The top 4, 5, and 6 categories showed slight changes. However, the top 7, Film & Animation category, showed a significant decrease in trending level from top 7 to top 11. Additionally, the Top 11, Science and Technology category, showed a significant increase in trend level from top 11 to top 6 after enhancement. According to research findings, this could be due to current trends and circumstances, or it could depend on the audience's interests, likes, and dislikes. Therefore, since the trending accuracy of this system depends on the increase or decrease of the changed index when enhanced based on the ground truth index, the accuracy was calculated based on Mean Reciprocal Rank, or the Reciprocal function, or the Ranking Matrix, resulting in 87.6%. Based on the research findings, the proposed trending algorithm is highly valuable for researchers in generating accurate and timely rankings. The results indicate that this capability can be further enhanced and effectively applied within specific research domains in India. Table 5. shows the experimental findings for the YouTube trending videos in India, 2023. As the table illustrates, this system proves that the execution time is good and quick, and the accuracy is correct and accurate. Furthermore, the detailed calculation results of many other characteristics were also found to be clearly evident.

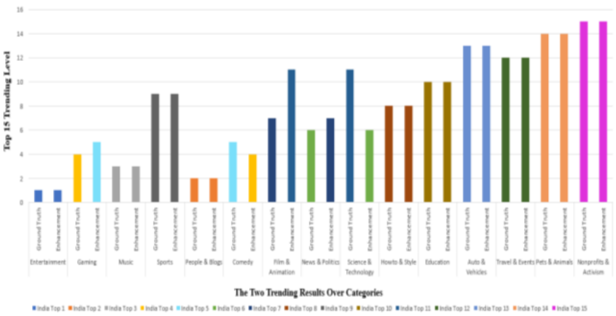

Fig. 20: The Comparison of Enhanced and Ground Truth Trending Results, Source: created by the authors

Table 5. The Experimental Results for the India YouTube trending videos, 2023, Source: created by the authors

Characteristics	2023 Enhanced Trending Analysis	India
Dataset	YouTube Videos Trending Data is daily updated and obtained from the YouTube API by Scrapping. It consists of a total 208180 rows and 16 columns (original) or 28 columns (enhanced) from 01 January 2023 to 15 December 2023.	68040
Video Title Length	Title Length between 30 and 60 characters	
Most Trending Category	Top 1, 2, 3, 4, 5	1.Entertainment 2.People & Blogs 3.Music 4.Comedy 5.Gaming
Most Trending Channel	Top 1, 2, 3, 4, 5	1.Vijay Television 2.Colors TV 3.ZEE5 4.Asianet 5.Colors Rishtey
Common Words	Video Titles	#shorts, Episode, Song, New, 2023, Official, Video
Views, Likes, Comments	YouTube Trending Video content	Trending Videos of 87.44% has views less than 5 million views. Trending videos of 99.21% has likes less than 2 million likes. Trending videos of 98.79% has comments less than 100000 comments.
Published Day and Time	The Best published Day and Time	Friday, Wednesday, Thursday 05:00, 12:00, 13:00
Video appeared on the Trending List	On Trending Videos take at least 1 day.	Between 10 hours and 18 days
The Most Frequently Times	The Videos are the most frequent times in the trending videos list	“Kundali Bhagya” 55 Times
Correlation of Trending Videos	Views, Likes, and Comments	Views and Likes = 0.72 Likes and Comments = 0.66 Views and Comments = 0.45
Overall Accuracy	Sentiment Analysis	97% (train), 91% (test)
	Trending Analysis	87.6%
Execution Time	Sentiment Analysis	00:01:11
	Trending Analysis	00:00:58

4.4 Limitations and Further Extensions

This system relies on data provided by YouTube. If data for a specific country is available, it can be easily analyzed by simply replacing the dataset. Currently, the system has conducted YouTube career analysis only for India, but it can be extended to other countries if YouTube data is accessible. Most countries, especially in the US, primarily use English, so comments are not translated, although translation could be implemented. In India, comments are translated from native languages into an international language before categorizing opinions as positive or negative. If there is no need to translate comments and analysis is conducted in the country's native language, the translation step can be skipped, which may be more beneficial for countries using native languages.

For sentiment analysis, the system employs the Multinomial Naive Bayes classifier with TF-IDF features. However, other machine learning classification methods can also be applied. Similarly, for trending analysis, alternative prediction and classification algorithms can replace the LG model used for performance measurement. While web content mining has been utilized, advanced AI techniques such as deep learning could also be incorporated. Additionally, tools like Big Data frameworks, Hadoop Streaming API, and XGBoost can be used for data processing and analysis.

5 Conclusion

The dataset is also crucial to the success of this system. The data is unstable when it is updated daily since it is a private dataset that is scraped every day, and the results remain unstable until the entire year of analysis is finished. This is only one of the many challenges. The data was preprocessed twice, first by scraping it and transforming it into useful data through profiling, cleansing, reduction, transformation, enrichment, and validation. The second time, it was preprocessed for sentiment analysis, including changes in lowercasing, tokenization, punctuation, URL, stopword removal, lemmatization, negations, intensifiers, emojis, abbreviations, special characters, and vectorization. The preprocessing process was able to emphasize a lot, resulting in good accuracy in the results. The process was crucial for achieving accurate results. Having encountered several challenges because there are restrictions on getting audience feedback while conducting sentiment analysis. Ultimately, we created it with the YouTube comment downloader, and it was successful. Google had a daily restriction, which made the language translation process take a long time. After sentiment analysis, negative sentiment results were obtained, but the feature combination part was the most challenging. Even though we tried using various techniques, the data did not match. Finally, only after grouping the categories and adding them according to video_id, the comments given by the audiences according to the video match, and the data validation part is successful. Since each feature's input value differs, it is impossible to assess them all using a single model in the evaluation stage. Because the inputs are boolean (0,1), categorical (text), and statistical (number), we had to transform and then evaluate them. Regression prediction and classification were the two components of the performance evaluation section; thus, we had to employ distinct methods to measure each component's successful outcome. Moreover, ground truth data is taken from Kaggle, which is just raw data. As a result, it must be contrasted with the outcome that was obtained after the trending result was initially produced. Three different types of trending outcomes are clearly compared in the evaluation section: the trending result achieved using the ground truth dataset, the enhanced trending result with negative features, and the trending result without negative features using the own dataset. The challenges that arose throughout the research are highlighted and successfully implemented.

References:

[1] Sowmiya Ka,1, Supriya Sb & R.Subhashinic, (2021) “Scraping and Analysing YouTube Trending Videos for BI,” Smart Intelligent Computing and Communication Technology, Vol: 38, pp. 542 – 547, India, DOI: 10.3233/APC210099

[2] Johanes Fernandes Andry, Stefan Azriel Reynaldo, & Kevin Christianto, (2021) “Alogrithm of Trending Videos on YouTube Analysis using Classification, Association and Clustering”, 2021 International Conference on Data and Software Engineering (ICoDSE), IEEE, pp. 25 – 30, Indonesia, DOI:10.1109/ ICoDSE53690 .2021. 9648486.

[3] S. Amudha[1] , V R. Niveditha[2] , Dr. P.S. Raja Kumar[3] , & M.Revathi[4], (2020) Dr.S.Radha Rammohan 5, “Youtube Trending Video Metadata Analysis Using Machine Learning,”, International Journal of Advanced Science and Technology, Vol.29, No.7s, May 2020, pp. 3028 – 3037, ISSN 2005-4238, India.

[4] [1st]Muhammad Alkaff, [2nd]Andreyan Rizky Baskara, & [3rd]Yohanes Hendro Wicaksono, (2020) “Sentiment Analysis of Indonesian Movie Trailer on YouTube Using Delta TF-IDF and SVM”, 2020 Fifth International Conference on Informatics and Computing (ICIC) 24 December 2020 by IEEE Xplore, Indonesia, DOI:10.1109/ICIC50835.2020.9288579.

[5] G.M.H.C.Gajanayake1 , & T.C.Sandanayake2, (2020) “Trending Pattern Identification of YouTube Gaming Channels Using Sentiment Analysis”, 20th International Conference on Advances in ICT for Emerging Regions (ICTer 2020): IEEE, pp. 149 – 154, University of Colombo School of Computing (UCSC) in Sri Lanka,
DOI: 10.1109/ICTer51097.2020.9325476.

[6] Eliganti Ramalakshmi[1] , A Bindhu Sree Reddy[2] , & Sharvani G [3], (2022) “YouTube Data Analysis and Prediction of Views and Categories,”

International Journal for Research in Applied Science & Engineering Technology, 2022, Issue: VI, Volume:10, pp. 568 – 573, India, DOI: 10.22214/ijraset.2022.43636.

[7] Aiman Abdullah Anees1st, Harsh Prakash Gupta2nd, Aditya Prashant Dalvi3rd, Suhas Gopinath4th and Biju R Mohan5th, "Performance Analysis of Multiple Classifiers using different Term Weighting Schemes for Sentiment Analysis", Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE Xplore, (5) (6) (2020), pp. 637 – 641, India DOI: 10.1109/ICCS45141.2019.9065895.

[8] Swati Gayakwad[1], Rajas Patankar[2], & Dashrath Mane[3], (2020) "Analysis on YouTube Trending Videos", International Research Journal of Engineering and Technology, 2020, (e-ISSN: 2395-0056), Volume: 07, Issue: 08, pp. 4247 – 4253, Online (www.irjet.net) proceeding.

[9] Meher UN Nisa[1] , Danish Mahmood [1,*] , Ghufran Ahmed[2] , Suleman Khan [3,4] , Mazin Abed Mohammed [5] & Robertas Damaševiˇcius [6,*], (2021) "Optimizing Prediction of YouTube Video Popularity Using XGBoost", Advanced Application of Machine Learning and Meta-Learning in Image and Text Analysis, Volume 10, Issue 23, Electronics, 28 November 2021, 10, 2962.
https://doi.org/10.3390/electronics10232962, Electronics | An Open Access Journal from MDPI

[10] Jiaxi Lu, (2023) "Trending Videos on YouTube", Journal of Education, Humanities and Social Sciences, MEHSS, Volume 7, January (2023), pp. 84 – 91, United Kingdom, DOI: https://doi.org/10.54097/ehss.v7i.4016.

[11] MD SAKIBUR HASAN, Bishal Sarker, Diksha Shrestha, Roshan Shrestha, & Sajal N. Shrestha, (2023) "Trending YouTube Video Analysis", Research Square, February 7th, 2023, pp. 1 – 6, ISSN 2693-5015 (online proceeding) DOI: 10.21203/rs.3.rs-2548456/v1.

[12] Lau Tian Rui[1] , Zehan Afizah Afif[2] , R. D. Rohmat Saedudin[3] , Aida Mustapha[4] , & Nazim Razali[5], (2019) "A regression approach for prediction of Youtube views", Bulletin of Electrical Engineering and Informatics, Volume 8, No.4, December 2019, pp. 1502 – 1506, Malaysia, DOI: 10.11591/eei.v8i4.1630.

[13] Punam Wake1st, Nita Borkar2nd, Swati Dakhore3rd, Mr. Vijay R. Wadhankar4th, Mr. Abhishek Kumar5th, "Predictive analysis of YouTube Using Machine Learning", International Research Journal of Modernization in Engineering Technology and Science, Volume 4, Issue 11, (2022), pp.2397 – 2404, India, DOI: 10.56726/IRJMETS31771.

[14] Abbi Nizar Muhammad1st, Saiful Bukhori 2nd, Priza Pandunata 3rd, "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, (16) (10) (2019), pp.199 – 205, Indonesia,
DOI: 10.1109/ICOMITEE53461.2021.9650239.

[15] Erdinc Uzun, "A Novel Web Scrapping Approach Using the Additional Informartion Obtained From Web Pages", IEEE Access, Volume: 8, March 31, 2020, pp. 61726 – 61740, DOI: 10.1109/ACCESS.2020.2984503.

[16] 1st Khin Than Nyunt, 2nd Naw Thiri Wai Khin, "Customized Criteria Based Trending Analysis for YouTube Social Media Platform", Proceedings of The 2023 IEEE Conference on Computer Applications (ICCA), IEEE Xplore, (27-28) (2) (2023), pp. 33 – 38, Myanmar, DOI: 10.1109/ICCA51723.2023.10181458.

[17] 1st Khin Than Nyunt, 2nd Naw Thiri Wai Khin, "Performance Analysis of Logistic Regression Using Term Frequency – Inverse Document Frequency for Sentiment Analysis", 2023 International Conference on Communication and Computer Research (ICCR), Korea, (23 - 24) (10)(2023).

[18] 1st Muhammad Alkaff, 2nd Andreyan Rizky Baskara, 3rd Yohanes Hendro Wicaksono, "Sentiment Analysis of Indonesian Movie Trailer on YouTube Using Delta TF-IDF and SVM",Fifth International Conference on Informatics and Computing (ICIC), May 14, 2020, IEEE Xplore, Indonesia, DOI: 10.1109/ICIC50844.2020.9288732.

[19] Kian Long Tan, Chin Poo Lee1st, and Kian Ming Lim2nd, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research", Applied Sciences, MDPI, Switzerland, (4) (13) (2023), pp. 1– 22, DOI: 10.3390/app13074550

[20] Arushi Pathik1st, Saumya Patni2nd, Vaibhav Patel3rd, Jash Patel4th, Artika Singh5th, "YouTube Trend Analysis", 2022 IEEE Pune Section International Conference (PuneCon), IEEE Xplore, January 19, 2023, India , DOI: 10.1109/PuneCon55413.2022.10014717.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

The authors equally contributed to the present research, at all stages from the formulation of the problem to the final findings and solution.

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

No funding was received for conducting this study.

Conflict of Interest

The authors declare that they have no conflicts of interest relevant to the content of this article.