Analyzing and Predicting YouTube Video Popularity Using Machine Learning

Submitted By: Piyush Raj (1NT21CS036), Shivansh Raj (1NT21CS166),

Naim (1NT21CS105), Rachit (1NT21CS138)

Department of Computer Science

Nitte Meenakshi Institute of Technology

Abstract – The realm of data science is increasingly pivotal in shaping digital content strategies, with YouTube serving as a prime platform for content creators and marketers alike. This project undertakes an extensive analysis and predictive modelling of YouTube trending videos to uncover patterns and factors influencing video popularity. Utilizing a comprehensive dataset sourced from YouTube, encompassing metrics such as views, likes, dislikes, comments, and video characteristics, this study employs various machine learning algorithms including Linear Regression, Decision Tree, Random Forest, and Gradient Boosting. Each algorithm's performance is evaluated and compared to identify the most effective model for predicting video engagement metrics. The primary objective is to equip content creators, digital marketers, and YouTube stakeholders with actionable insights derived from predictive modelling. By understanding the predictors of video success, stakeholders can optimize content creation strategies, enhance viewer engagement, and improve overall performance on the platform. Through this project, we aim to bridge the gap between data science methodologies and digital content strategies, fostering informed decision-making and strategic content optimization in the dynamic landscape of online video platforms.

I. INTRODUCTION

Online video content has become essential to marketing, entertainment, and communication initiatives in the current digital era. The way people and businesses communicate with audiences across the world has been completely transformed by platforms like YouTube, thus it is essential for marketers and content creators to comprehend the factors that influence viewer engagement and video popularity. The ability to forecast video engagement metrics, including likes, comments, and views, is essential for improving the efficacy of digital marketing campaigns and content initiatives. Algorithms for machine learning (ML) provide strong tools to analyse big datasets and find trends that affect YouTube video performance. This project uses machine learning (ML) approaches to create prediction models that anticipate a video's likelihood of receiving likes based on various criteria. These features include a wide range of topics, including the number of views, comments, upload time, category, and length of interaction for videos. This study's main goals are to determine the variables that have a major influence on YouTube audience engagement and to investigate the effectiveness of several machine learning algorithms in predicting video likes. We want to evaluate the effectiveness of algorithms such as Random Forest, Support Vector Machine (SVM), and Gradient Boosting in this forecasting job by utilizing an extensive dataset of YouTube videos covering various content categories. We want to identify the algorithm that provides the best reliable performance in this situation by doing a thorough evaluation utilizing measures like accuracy, precision, recall, and F1-score. The present study advances our comprehension of the complex processes behind YouTube video engagement and highlights the usefulness of the findings for marketers, digital strategists, and content creators. This study aims to provide stakeholders with practical information to optimize content strategies, improve audience engagement, and ultimately achieve better success in the highly competitive online video marketing industry by bridging the gap between machine learning approaches and digital media analytics.

II. LITERATURE REVIEW

In recent research, J. S. Immaculate, A. S. Janet, and K. J. C. Angel[1] conducted a comprehensive study on social media analytics, highlighting the technological innovations, sentiment analysis, and the role of social networking platforms. They explored the statistical approaches and the significance, advantages, and disadvantages of social media analytics (SMA). Social networks have revolutionized the world, allowing users to display personal information and leave digital footprints on the Internet. Analyzing this information can assist companies in conducting interviews, providing employers with a comprehensive view of a person's personality and social behavior. A study by A. M. Khasanova and M. O. Pasechnik [2] explored effective working group formation and identifying deviant behavior through personal profile analysis on the social network VKontakte. This study involved data collection, pre-processing, and analysis, followed by organizing users into groups using machine learning and deep learning techniques. They employed neural networks and other machine learning methods, utilizing the K-means clustering algorithm to group users by interests. A wide range of frameworks are used in the generation of enormous amounts of information in various data formats like text, content, sound, and video. Social media accounts are major platforms where people share their perspectives or sentiments in the above-mentioned formats. To investigate the field of sentiment analysis, social media platforms are crucial areas to examine. The increment and acquirement of applications that work on social media provide both opportunities and challenges to researchers in this field. The vast amount of information produced by people utilizing social media platforms results from the combination of their experiences and daily actions. These social platforms generate information that can be highly useful for any field. This paper[4] discusses the investigation of social media as the best place for finding opinions, emotions, and sentiments. We focus on sentiment analysis over social media, aiming to inspire researchers to delve into social media analysis and its major issues. In a broad sense, social media refers to a conversational, distributed mode of content generation, dissemination, and communication among communities. Unlike broadcast-based traditional and industrial media, social media has torn down the boundaries between authorship and readership. The process of information consumption and dissemination is becoming intrinsically intertwined with the process of generating and sharing information. [5] This special issue samples the state of the art in social media analytics and intelligence research, which has direct relevance to the AI subfield from either a methodological or domain perspective . While there has been some scholarly activity on social media analytics, few publications [7] explain YouTube's analytics in detail and how educators can use them to improve their content. Baron's (2023) article critically explains and analyzes 14 of YouTube's social media analytics, assisting educators in enhancing their impact. The article reports on the success of various engineering tutorial videos published in 2020 and 2021, which have accrued over 1 million views on YouTube. It aims to provide readers with practical tools to improve their offerings on public networks such as YouTube. Given that up to 60% of all YouTube views come from the platform's recommendations rather than direct search queries, understanding YouTube's recommender system is crucial and is presented in this article. During the COVID-19 lockdowns in South Africa, undergraduate laboratory sessions were forbidden, prompting the use of video-based tutorials as a tentative solution to address the lack of in-person practical demonstration sessions. Five videos on electrical engineering topics were filmed, uploaded, and publicly shared on YouTube. [6] An investigation was conducted to determine whether these videos could be useful for teaching practical engineering content in a university context. This article reports on the findings of using YouTube as a platform for sharing and evaluating engineering educational practical tutorial videos. The goal of this article is to introduce YouTube's social media analytics as a tool for educators to evaluate their educational videos. The findings suggest that educators may consider using social media analytics to evaluate their videos, but these analytics should be reviewed critically and comprise several metrics measured temporally. Additionally, understanding YouTube's recommender system and its influence on the platform is an important factor in evaluating video content. In the dynamic realm of YouTube, addressing the diverse needs of its extensive audience while empowering content creators is crucial. A study introduces a Chrome extension designed to enhance user engagement and support IT-related content creators on YouTube. Utilizing advanced Natural Language Processing (NLP) and Machine Learning (ML) techniques, the extension offers four key functionalities: user engagement analysis, user requirement extraction, controversial topic identification, and keyword and title recommendations. The study [8] employed algorithms such as Random Forest, VADER Lexicon, Multinomial Naive Bayes (MNB), Latent Dirichlet Allocation (LDA), Bag of Words (BOW), and GPT-3. The system achieved an overall accuracy of 89.72%, with individual components achieving accuracies of 83.04%, 92.83%, 94%, and 89%, respectively. These methodologies can be adapted for other platforms and categories beyond IT content, such as entertainment, gaming, cooking, and travel. This research enhances YouTube analytics and provides tools to optimize engagement and foster deeper online dialogues. Machine learning has been pivotal in data classification across social media platforms like Twitter, Facebook, and WhatsApp for over two decades. Recent research aims to enhance accuracy in sentiment analysis and opinion mining of user comments. By improving classification models, this work q seeks to provide deeper insights into user sentiments and opinions, addressing previous limitations for more precise analysis on social media.

III. DESIGN AND METHODOLOGY

Dataset Used

The dataset utilized in this YouTube like prediction project consists of information on trending videos in the United States. It encompasses a wide range of attributes crucial for understanding video engagement and trends. Key features include metrics such as views, likes, dislikes, and comment counts, which provide insights into audience interaction and popularity. Temporal data such as publishing time and trending date allow analysis of video trends over time and optimal publishing periods. Categorical information such as category IDs and corresponding category names categorize videos into genres like Music, Entertainment, and Sports, aiding in genre-specific analysis. Additionally, textual features like video titles and descriptions offer textual insights into content themes and topics. This comprehensive dataset not only facilitates predictive modelling of video likes but also enables exploratory analysis of factors influencing video popularity on YouTube. The dataset's richness and diversity make it well-suited for machine learning applications aimed at understanding and predicting audience engagement dynamics in online video content. The dataset used for this project's research includes a wide range of variables, including textual elements, temporal data, video metrics, and category information. The data pretreatment method included handling null values, properly formatting date-time attributes, and feature engineering to give additional metrics like trending days difference. Exploratory Data Analysis (EDA) was utilized to visualize trends across time, peak posting times, and distributions of engagement metrics (likes and views) among different video categories. The selection and appraisal of features was based on the identification of factors that impact viewers' interaction with videos. To find the most reliable and accurate predicted performance, a number of machine learning algorithms were evaluated as part of the model selection process. The models evaluated were:

i. Logistic Regression
ii. Decision Tree
iii. Random Forest
iv. Gradient Boosting
v. Support Vector Classifier
vi. Gaussian Naive Bayes

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.857751	0.860173	0.857751	0.857456
Decision Tree	0.948152	0.948152	0.948152	0.948152
Random Forest	0.963308	0.963331	0.963308	0.963306
Gradient Boosting	0.922095	0.922178	0.922095	0.922096
Support Vector Classifier	0.845254	0.845253	0.845254	0.845253
Gaussian Naive Bayes	0.733050	0.797383	0.733050	0.717093

Model Selection

The Random Forest method was chosen from the assessment findings because to its better performance in all measures. During training, the ensemble learning technique Random Forest builds many decision trees. Every tree is trained using distinct subsets of the dataset that are chosen at random for their features and data points (bootstrap samples) (feature bagging). Because each tree will capture distinct features of the data due to this randomization, overfitting will be lessened and the model's capacity to generalize to new data will be enhanced. The Random Forest's capacity to ascertain feature relevance is essential to its efficacy. The algorithm determines which parameters are most significant by calculating the relative contributions of each feature (e.g., views, category, and posting time) to the prediction of video likes over all trees. In addition to improving comprehension of the underlying patterns influencing video interaction, this feature importance analysis helps content producers make strategic decisions that will maximize the performance of their videos. Random Forests are also a good choice for applications that need accurate forecasts based on complicated datasets because of their reputation for high accuracy in classification tasks, such as binary outcome prediction (likes or not).

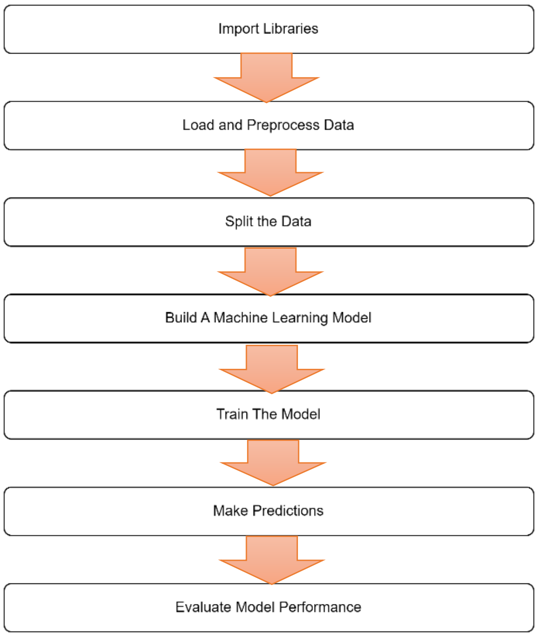

Fig.1 Flowchart of Chatbot

Algorithms Used – Random Forest

In this YouTube like prediction project, the Random Forest algorithm was employed to forecast the number of likes videos would receive based on various attributes and features. Random Forest operates through ensemble learning, where it constructs multiple decision trees during training. Each tree is trained on different subsets of the dataset, randomly selected both in terms of data points (bootstrap samples) and features (feature bagging). This randomness ensures that each tree captures unique aspects of the data, thereby reducing overfitting and improving the model's ability to generalize to new data. During the prediction phase, the algorithm aggregates the predictions from all trees in the forest to produce a final prediction for each video's likes.

Key to the Random Forest's effectiveness is its ability to determine feature importance. By analysing how much each feature (such as views, category, and publishing time) contributes to predicting video likes across all trees, the algorithm identifies which factors are most influential. This feature importance analysis not only enhances understanding of the underlying patterns driving video engagement but also guides strategic decisions for content creators seeking to optimize their video performance. Moreover, Random Forests are known for their high accuracy in regression tasks like predicting numerical values, making them well-suited for applications requiring precise forecasts based on complex datasets. In this project, the Random Forest model was trained, evaluated using metrics like accuracy and F1-score, and used to derive actionable insights that enhance decision-making in content strategy and audience engagement on YouTube.

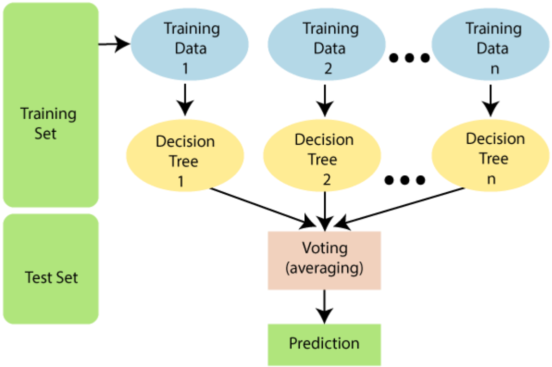

Fig.2 Random Forest Algorithm

IV. CONCLUSION AND FUTURE SCOPE

In conclusion, this project has successfully demonstrated the efficacy of the Random Forest algorithm in predicting video likes on YouTube using a comprehensive dataset of trending videos in the United States. Through ensemble learning, Random Forests effectively captured intricate patterns and relationships among various attributes such as views, likes, dislikes, comment counts, category IDs, and temporal data like publishing time and trending dates. The model's performance was evaluated using metrics such as accuracy, precision, recall, and F1-score, showcasing its robustness in predicting video engagement metrics. Key insights gleaned from the analysis include the significant impact of factors like viewer engagement metrics and content categorization on video popularity. These findings provide actionable insights for content creators to optimize their content strategies and maximize viewer engagement on the platform.

Looking forward, there are several avenues for further exploration and enhancement of this project. First, additional feature engineering could be explored, such as sentiment analysis of comments or advanced text processing techniques on titles and descriptions, to capture more nuanced aspects of viewer interaction. Second, optimizing the Random Forest model through hyperparameter tuning and exploring other ensemble methods like Gradient Boosting Machines (GBM) or Neural Networks could potentially improve predictive accuracy. Implementing real-time data pipelines and developing user-friendly interfaces for content creators to access predictive insights in real-time are also promising areas for future development. Lastly, evaluating the model's performance on datasets from different regions or platforms would provide insights into its generalizability and applicability beyond the specific dataset used in this project. These advancements aim to further advance the understanding and prediction of video engagement metrics on YouTube, offering valuable tools and strategies for content creators and marketers to enhance their reach and impact on the platform.

V. REFERENCES

1. J. S. IMMACULATE, A. S. JANET and K. J. C. ANGEL, "A Study of Social Media Analytics," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-13, doi: 10.1109/ICRITO51393.2021.9596247.

2. A. M. Khasanova and M. O. Pasechnik, "Social Media Analysis with Machine Learning," 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), St. Petersburg, Moscow, Russia, 2021, pp. 32-35, doi: 10.1109/ElConRus51938.2021.9396713.

3. R. Singh and P. Sharma, "An Overview of Social Media and Sentiment Analysis," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-4, doi: 10.1109/ISCON52037.2021.9702359.

4. S. Kuamri and C. N. Babu, "Real time analysis of social media data to understand people emotions towards national parties," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 2017, pp. 1-6, doi: 10.1109/ICCCNT.2017.8204059.

5. D. Zeng, H. Chen, R. Lusch and S. -H. Li, "Social Media Analytics and Intelligence," in IEEE Intelligent Systems, vol. 25, no. 6, pp. 13-16, Nov.-Dec. 2010, doi: 10.1109/MIS.2010.151.

6. P. Baron, "YouTube's Social Media Analytics as an Evaluation of Educational Teaching Videos," 2022 IEEE IFEES World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC), Cape Town, South Africa, 2022, pp. 1-8, doi: 10.1109/WEEF-GEDC54384.2022.9996202.

7. P. Baron, "Using YouTube's Social Media Analytics for Engineering Educators," 2023 IEEE Global Engineering Education Conference (EDUCON), Kuwait, Kuwait, 2023, pp. 1-10, doi: 10.1109/EDUCON54358.2023.10125146.

8. P. Peiris, T. Herath, R. Dissanayaka, K. Saranga, S. Thelijjagoda and I. Weerathunge, "Comprehensive Browser Extension for Analysing YouTube User Engagement, Controversy, User Requirements, and Trending Keywords," 2023 33rd International Telecommunication Networks and Applications Conference, Melbourne, Australia, 2023, pp. 134-139, doi: 10.1109/ITNAC59571.2023.10368503.

9. H. Singh, S. Ahamad, G. T. Naidu, V. Arangi, A. Koujalagi and D. Dhabliya, "Application of Machine Learning in the Classification of Data over Social Media Platform," 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, Himachal Pradesh, India, 2022, pp. 669-674, doi: 10.1109/PDGC56933.2022.10053121.

10. H. Singh, S. Ahamad, G. T. Naidu, V. Arangi, A. Koujalagi and D. Dhabliya, "Application of Machine Learning in the Classification of Data over Social Media Platform," 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, Himachal Pradesh, India, 2022, pp. 669-674, doi: 10.1109/PDGC56933.2022.10053121.