ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 03 March 2026

Beyond Detection: An Integrated Machine Learning System for Prediction of YouTube Virality

Maulshree Garg

Alliance University, Bangalore, India
Email: maulshreegarg@gmail.com

Abstract*— The rapid development of social media platforms has changed the way content is being created, propagated and advertised. The launch of short form content such as Tik Toks has completely shifted the market and changed the way it functions. Earlier the same market was dominated by long form content and the way to get engagement and virality was completely different as to what it is now. 12 million shorts are uploaded to YouTube every single day and approximately 1 to 5 percent of them reach virality. This project, "Beyond Detection: An Integrated Machine Learning System for Predicting YouTube Virality," focuses on developing a smart system that helps predict what goes viral in the YouTube short space. The primary objective of the project is quantifying virality and engagement using features already known to the user. The aim is to build a system which can input the features commonly associated with YouTube shorts and use that information as well as the trend of historical data to predict how well a certain video will perform on the platform. The dataset used in this project has been preprocessed and cleaned to prepare for the necessary operations. I used a dataset collected by me using scraping to make it as accurate to my aim as possible. The data set consists of half the data acquired from YouTube shorts with the help of python libraries such as scrapy, Beautiful Soup etc. and half the data acquired from Instagram using a scraping software called Apify. Then the data is combined based on video title and preprocessed in Power BI.

Index Terms— Virality Prediction, Data Scraping, XGBoost, LightGBM, Engagement Calculation

I. INTRODUCTION

In today's social media age nearly all the people in the world depend on short form content as an entertainment source. Social media has become one of the leading sources of income for people and people of all ages and backgrounds use it daily. In such a social climate it is one of the most important new areas in the world. So a tool to predict virality on such platforms is becoming more necessary.

1.1 Rise of YouTube and Its Impact

Through the last few years YouTube has emerged as one of the fastest growing social media platforms on the world. The rapid expansion of YouTube and short form content has changed how content in general is being perceived in the world and a lot of people depend on YouTube as a source of income.

Despite millions of uploads, only a certain number of videos see virality in the current market and for creators sometimes it feels like there are no set metrics to virality and getting engagement. Even the most low budget, badly produced videos can go viral and there is no set metric for it.

1.2 Machine Learning and Its Impacts

In such cases where it seems like there are no rhyme or reasons to virality my project aims to capture exactly what and how a video can go viral on YouTube. Machine learning has changed the world of prediction by using real time as well as historical data to be able to capture trends that might not be visible to a human. There are normal metrics as well as human intuition that are being used by creators daily to gauge how well a certain content idea will do but this is a way of making the process automated and done with the least possible margin of error rather than doing it manually. This method uses features such as video duration, number of subscribers, type of video, keywords in title etc. comments to predict the likes, views, and comments as well as engagement and virality score of a certain video idea. Feature engineering has been used to create features that will make the prediction as accurate as possible.

II. PREDICTION APPROACH

This paper introduces a new approach for predicting engagement and virality in YouTube using two models that have been tested and used to predict virality-

1. A LightGBM model also known as Light Gradient Boost Machine is used as it builds the gradient slowly and sequentially. It produces fast, accurate and effective predictions in machine learning.
2. An XGBoost model also known as Extreme Gradient Boosting is also used as it can handle complex data types and predicts results efficiently. The data I collected had some missing values and because it was a smaller dataset XGboost was the best option for my dataset.

In this project, boundation variable plays a really major role. A boundation variable basically acts as a barrier and a set of limits and conditions under which a machine learning project operates. This variable Acts as an input for the XGBoost as well as LightGBM models and basically effects the output value to make sure they are within the terms and conditions of the program.

My approach was to first scrape the data from the respective sites, clean the data and then combine it to make a dataset for the project. The dataset has around 430 rows and 15 columns. Then I used Power BI to see what the analytics

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 03 March 2026

of the data looked like to decide how to proceed with the dataset.

Then the models were researched and XGBoost and Lightgbm were picked.

The rest of the paper discusses how I collected the data, cleaned it and combined it, visualized it as well as researched the different model choices and picked up the correct models to get the best results possible. I have tried to make a system as useful as possible for the prediction of how a certain video will do on YouTube based on certain features.

III. METHODS AND METHODOLOGY

The implementation of the project for predicting virality using YouTube analytics is a bunch of interconnected modules that take in an input and based on that input transform the existing data into a suitable output. The modular structure means that each layer has its own important function it performs and it goes through the input in stages to perform its functions. The first part begins with input of the dataset, then cleaning the data and preprocessing it, then analyzing the data, then training and testing the model and then ends with UI implementation of the model.

The machine has a simple way of predicting the virality of the video. First as a user you input the features of your new video that you want to upload on YouTube. The features are the type of video, the title, keywords, duration and subscriber count based on all these features the machine basically looks over the general trend in the historical data and uses that as well as gradient boosting algorithms to predict what the trend in your specific data will be. It uses all this information to output its prediction as the number of views, likes and comments the video will get. It is the Data preprocessing module in which the data is collected, cleaned and preprocessed for further analysis. In this module the first thing I did was go over YouTube ad well as Instagram to see which type of videos were doing well on which platform and were there any similar videos on both the platforms as well as if the same video were being uploaded with the same name on both the platforms. The second stage was finding out which content creators I wanted to get the data from for my combined dataset of YouTube and Instagram videos. In this stage I selected eleven creators from 4 different specialties such as education, food, news and entertainment to work on and get the data from. Then I scraped all the common video titles from YouTube with the help of Beautiful Soup and Scrapy and the same titles from Instagram using an apify scraper designed for this specific function.

Then both the datasets which were collected were combined by using Apify software to match them based on the specific video title. Thus, a dataset having 400 plus rows and 15 columns was created. It was an extremely tedious job to make sure that no mistakes were made during the combining of the two datasets. There were a lot of missing values as well as incorrect values that were hand corrected by me.

In the data preprocessing module, the most important task of any system is to convert raw, unusable data into preprocessed, usable data. Our dataset currently has a lot of missing values, incorrect values and values that are categorical, i.e. wordy in nature, which cannot be directly used in a system for any sort of prediction. The categorical variables were converted into numerical variables which can be directly used for prediction. To do this variable encoding was used. The variables which were selected for processing were separated from the rest of the variables and termed as feature variables.

The dataset had a lot of missing values that were null in nature and were filled as 0 in the fields and a lot of incorrect values such as negatives which is not even possible in the field of YouTube virality which were searched and corrected. The data was honed into shape by the code. Another thing that was done was checking the data ranges to see if the values are all within range and none of the values are too unrealistic. Sometimes it is possible for the value to be unrealistically high as they can be the values of a viral video so a lot of double checking went into the data preprocessing.

The data was continuously cleaned in terms of removal of Fibonacci sequence numbers, zero values, etc.

The preprocessing code was very important, since it cleaned up and transformed the dataset into a machine learning readable shape. The collective research proves that there is a big shift coming in the market for social media platforms and some people would say its already here and we have to make ourselves ready for the changes and the challenges that it is going to bring. It is very important to learn how to quantify virality in such a social climate. The process of checking the various methods used by the researchers in quantifying virality have been linked to the RMSE values and the r squared values.

The result was that the dataset is preprocessed and cleaned into a stable and clean form that can be directly used for model prediction and model testing.

The data analysis phase is the second and maybe the most important phase of the entire implementation. In this phase the module basically analyses the dataset that has been collected to look for patterns and relationships inside the data. It does so to see how patterns and relationships affect YouTube virality and how each module contributes to model prediction.

The main purpose of this module is to find meaningful relationships in the data before we go onto applying prediction models on the data. The dataset consists of fields such as Channel location, channel name, total views, comments, dates, duration, likes and views both on YouTube as well as Instagram. It was loaded on excel as a csv file. Then I performed Exploratory Data Analysis on the data to find out the statistical as well as visual trends in the data. Things such as median, mode, maximum and minimum values were calculated for the various numerical features. The variations and trends in the data were identified by statistical methods as well as plots.

There was also the correlation matrix between the different variables to understand the correlation between the different

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 03 March 2026

features better. Visualization techniques such as pie, bar and donut plots were used to visually represent the data better.

In the visualization of the plots on different plots we found things such as the best duration for a short video that holds the attention of the viewer the best was around 120 seconds, and videos with high view count and videos from people with high subscriber count generally do the best on YouTube. It was also seen that food related videos do the best on YouTube as compared to Instagram.

This concludes the data analysis phase where analytics were generated of the given dataset to see the relationship between the different features in the dataset to see how well the data does on both Instagram as well as YouTube.

The module is the main component of the YouTube and Instagram virality system. In this module the preprocessed and data analytics-based system is converted into a system for prediction that has the capability to predict engagement and virality. The modules before complete data cleaning, preprocessing and data analysis and then move out to machine learning based prediction. The main aim of this module is not only to build a system for prediction but also look at various machine learning algorithms to pick out the best one for prediction.

The dataset consists of 400 rows and 15 columns. They contain attributes such as content type, duration, date, time of upload, etc. The dataset contains numerical attributes as well as categorical ones. The categorical attributes are features engineered into new features suitable for analysis. The dataset has both feature attributes that will be used to predict the target attributes and target attributes that will be predicted by the feature attributes. The feature attributes are number of subscribers, duration, title etc. while the target attributes are likes, comments and views etc.

From the dataset that is being used for prediction two parts emerge. One is known as the training set while the other is known as the testing set. In the training set the data is used to teach the model the relationship between the predicted vs prediction attributes, the testing set is the set of values used to evaluate how well the training set performed in training the model. This is done to ensure the model doesn't simply learn and copy the training set results anytime a value is entered but properly predict the values each time based on gradient and trend.

For the first model linear regression was selected. Linear regression is a machine learning module that maps a linear relationship between the targeted and the feature attributes. It is a very simple and common model used to see what sort of relationship exists between the features. The model managed to give coefficients that made sense, at least in a basic way. It could predict stuff okay, but not well.

I think the issue comes from the dataset. There are these non-linear interactions between the features, which linear regression struggles with. It just does not capture those complexities right, so the performance ends up limited. The result of the prediction was around 80 percent accurate and that is way below the acceptable limit of the acceptance percentage. The error matrix showed that a different model might be better for this project.

After linear regression model a decision tree model was implemented. The decision tree models can understand and map nonlinear relationships between different variables better. They work by repeatedly splitting based on feature thresholds. This makes it easier for a decision tree model to capture nonlinear relationships between data better. The decision tree model was better at understanding the complex relationships between the data better and dividing the data based on categories. However, it showed some signs of being overfitted. Overfitting means it performed well on test data but not that well with test data due to it being more tailored to the training data.

To improve performance, this advanced gradient boosting algorithms were used. Boosting is this thing where you train these weak models one after the other. Each one tries to fix what the last one got wrong. It seems like that sequential approach makes a big difference. For structured data, like what we have in the project, these methods do pretty well. The errors get corrected over time, which is helpful.

The boosting gradients used were LightGBM and XGBoost. XG boost was the first model used. XGBoost is known for its strong predictive skills, simple structure and strong foundation. It is a model that prevents overfitting easily. It is also good at capturing nonlinear relationships between the dataset which makes it ideal for our dataset. LightGBM was also used which is a simple and efficient boosting algorithm like XGBoost but has the advantage of being more memory efficient and really good at handling missing as well as null values in the dataset.

Both of these algorithms performed well. Throughout the process hyper parameters were adjusted and the results we got were better than ever. R squared as well as RMSE values were calculated to find out how well a model was doing.

IV. LITERATURE REVIEW

There has been growing research in the fields of social media as the use of social media increases. Wide research has been done in the fields of YouTube, Tik Tok as well as Instagram. The following review combines research done in the areas of marketing on social media, engagement and how it works on various platforms, technical indicators of virality, and the evolving consumption of gen z on social media.

The rapid growth of social media platforms has changed the world as we know it. The first paper explores how YouTube shorts can grow more than they already is by strategic policy formation. Strategies such as variation, value and versatility have been already implemented by YouTube and are being used by it to gain a competitive edge in the current market [1]. Following the launch of YouTube Shorts in July 2021, YouTube's second-quarter revenue reached **$7 billion**, representing a nearly **84% year-on-year increase.** There is the use of 4V marketing strategy for the growth and development of YouTube.

In the second paper the posting format is used to decipher how content will do on Instagram. The study sees the impact

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 03 March 2026

of different posting formats such as reels, videos and photos to see which ones do the best on the platform [2]. 330 total posts are collected from 22 small jewellery business accounts to see which posting format works the best. Weighted engagement is used which is then also used by me in my project the formula of which is

(Number of Likes + (2 × Number of Comments)) / Number of Followers × 100

The third paper explores how digital marketing works on platforms such as Tik Tok and how digital marketing has been revolutionized on such platforms. Unlike other social media, Tik Tok is more concerned with how an individual piece of content does than the follower count of the account. This has changed the entire marketing on SFVS. [In paper 4 we study the specific content of tik Tok videos that lead to their popularity. It is dependent on three things the recommendation algorithm, the creators' profile and the content of the video. Amongst the three creators' popularity is the most important feature that affects the popularity of the video the most. Accounts with over 10,000 followers accounted for 90% of the total viral videos. For the research the people collected 103,587 videos from 63 trending hashtags. [4]

In paper 5 content consumption trends amongst gen z were noted through Instagram. A study was conducted in India by Indian students using the survey method where over 900 students were surveyed from the ages of 18 to 25 to see what their social media consumption habits looked like. In the survey it was noted that approx. 78 percent of the people in the space use Instagram daily for multiple hours. The most common reasons for using reels are content discovery as well as entertainment [5].

The five researched papers explain pretty clearly what kind of research has been done in the already present papers over the internet. Research has been done about the marketing strategies, the expansion of the platform, the type of content that does the best on the platforms and how to successfully market the content for a generation of users. There is little to no scientific research that has already been done in the following fields, and it is a completely new area of research that the paper is trying to explore.

In some of the papers methods are being used to quantify the metrics such as engagement and virality and that is where I've drawn inspiration for my project. The global expansion of short form content has really altered the market as we know it to something completely different. According to Ruoni Xiong YouTube revenue grew from 6.01 million to 7 million as soon as there was the launch of YouTube shorts. That's an increase of 86 percent. The research shows that YouTube basically established and utilized a specific budget for YT shorts to make it grow much faster as compared to other platforms. [6]

V. RESULTS AND DISCUSSION

The findings from modules 3 and 4 are included over here. The test results from the various prediction algorithms as well as how well the algorithms perform are all captured over here. Various models were tested as a part of those modules.

To determine whether a prediction is accurate or not we've used a variety of models to be as unbiased on our predictions as possible. There are so many different results achieved from using the same model but different hyper parameters.

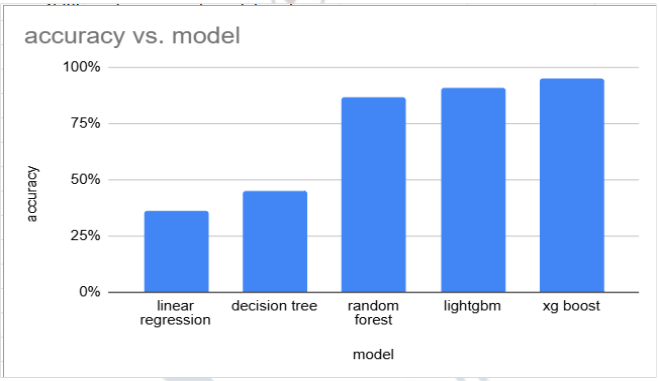

Figure 1: Results

Out of all the algorithms that were tested xg boost did the best. It performed with over 95 percent accuracy which makes it the strongest algorithm and the only one that crosses over the research threshold.

The training trajectory of XG boost model:

n_estimators	max_depth	n_jobs	randoms state	accuracy
100	7	-1	42	91%
150	14	-1	49	97%
180	8	-1	45	98%
100	6	-1	42	90%
180	8	-1	46	96%

VI. CONCLUSIONS

The project was mainly rooted in finding the prediction for YouTube virality based on certain categorical as well as numerical attributes. the main objective of the project was the analysis of various fields related to the video and using those and a generic analysis to find out the engagement and virality score of the given videos.

Proposed models used for this analysis are XG Boost and Light GBM models which are both gradient boosting algorithms in nature. There were a bunch of other machine learning algorithms that were tested out to see which one fits the best with the available data. All the various models were tested against each other, and their accuracy scores were compared.

The accuracy score of XG Boost and LightGBM were the only ones that were above 95 percent and hence in the ideal range for these sorts of projects. In these two specific models the hyperparameters were changed and based on the parameter changes the accuracy was calculated to see which one is giving the best results.

The first step of the project was the data collection where python libraries such as scrapy as well as beautiful soup and software's such as Apify were used to collect the data from YouTube as well as Instagram. The data was cleaned, preprocessed and made ready for further analysis in python and Power BI.

Then the data was analyzed by using Power BI to see what the relationship between the features is and to capture the trends and patterns in the given data. Analysis was done by using statistical as well as graphs and visualization techniques. A lot of information was gleamed from these graphs as they're the perfect way to get information.

Overall, this project captures the practical aspects of Data analysis as well as machine learning and captures how data can be used to predict how a certain YouTube video will do on the platform.

REFERENCES

[1] R. Xiong, "Analyzing the Reasons of Stable Development of YouTube Short Video Market," in *Proceedings of the Decoupling Corporate Finance Implications of Firm Climate Action - ICEMGD 2024*, 2024. doi: 10.54254/2754-1169/102/2024ED0063

[2] S. Liang and J. Wolfe, "Getting a Feel of Instagram Reels: The Effects of Posting Format on Online Engagement," *Journal of Student Research*, vol. 11, no. 4, 2022

[3] S. B. M, A. Iqbal, S. Saha, C. Koneti, L. N. Eni, I. Premkumar, and S. Suman, "Content Marketing in the Era of Short-Form Video: TikTok and Beyond," *Journal of Informatics Education and Research*, vol. 4, no. 2, 2024

[4] C. Ling, J. Blackburn, E. D. Cristofaro, and G. Stringhini, "Slapping Cats, Bopping Heads, and Oreo Shakes: Understanding Indicators of Virality in TikTok Short Videos," 2021

[5] G. Doloi, "The Influence of Instagram Reels on Content Consumption Trends among Gen Z," *Journal of Social Responsibility, Tourism and Hospitality*, vol. 4, no. 6, 2024, pp. 21–31. doi: 10.55529/jsrth.46.21.31

[6] R. Xiong, "Analyzing the Reasons of Stable Development of YouTube Short Video Market," in *Proceedings of the Decoupling Corporate Finance Implications of Firm Climate Action - ICEMGD 2024*, 2024. doi: 10.54254/2754-1169/102/2024ED0063

[7] S. Liang and J. Wolfe, "Getting a Feel of Instagram Reels: The Effects of Posting Format on Online Engagement," *Journal of Student Research*, vol. 11, no. 4, 2022

[8] S. B. M, A. Iqbal, S. Saha, C. Koneti, L. N. Eni, I. Premkumar, and S. Suman, "Content Marketing in the Era of Short-Form Video: TikTok and Beyond," *Journal of Informatics Education and Research*, vol. 4, no.2