VidMine: Digging Deep into YouTube Insights

Amulya Ambati
amam1261@colorado.edu
University of Colorado, Boulder
USA

Prasanna Poloru
lapo4503@colorado.edu
University of Colorado, Boulder
USA

Sneha Sasanapuri
snsa4676@colorado.edu
University of Colorado, Boulder
USA

Figure 1: Youtube Analytics

ABSTRACT

With the exponential growth of digital content consumption, understanding user behavior and content trends on platforms like YouTube has become paramount for content creators, marketers, and businesses. "VidMine" presents a comprehensive data mining project aimed at unearthing valuable insights from YouTube's vast repository of data. Leveraging the YouTube Data API and advanced data mining techniques, this project delves deep into YouTube's ecosystem to extract, analyze, and interpret key metrics, trends, and patterns. The project's primary objectives include data extraction, analysis, insight generation, visualization, and recommendations. Through a systematic data collection, "VidMine" uncovers actionable insights pertaining to user engagement, content performance, audience demographics, and emerging trends. Natural Language Processing (NLP) techniques are employed to analyze textual data such as comments and descriptions, providing valuable insights into audience sentiment and engagement.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Amulya Ambati, Prasanna Poloru, and Sneha Sasanapuri. 2024. VidMine: Digging Deep into YouTube Insights. In . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the digital age, where online video consumption is at its peak, platforms like YouTube have emerged as ubiquitous sources of entertainment, education, and information. With billions of users and an ever-expanding library of content, YouTube offers a treasure trove of data waiting to be explored. Understanding the dynamics of this vast ecosystem is essential for content creators, marketers, and businesses striving to thrive in the digital landscape [2].

Through advanced data mining techniques, we embark on a quest to dig deep into the wealth of insights hidden within YouTube's vast repository of content, engagement metrics, and user interactions. In this report, we present the culmination of our efforts—a comprehensive exploration of YouTube's data terrain. From data extraction to analysis, visualization, and actionable insights, "VidMine" offers a holistic view of YouTube's ecosystem and its implications for content creators, marketers, and stakeholders.

2 STRUCTURE OF THE REPORT

This report is structured to provide a comprehensive overview of the "VidMine" project, encompassing its methodology, findings, insights, and recommendations. We begin by outlining the methodology employed in data extraction, analysis, and visualization. Subsequently, we delve into the insights derived from the analysis,

shedding light on user behavior, content trends, and audience engagement metrics. Finally, we conclude with actionable recommendations for leveraging these insights to enhance content strategy, optimize audience engagement, and drive growth on YouTube

3 UNDERSTANDING YOUTUBE LANDSCAPE

This paper researches the influence of venue dynamics on cricket match outcomes and prompts the investigation of whether playing at home holds a substantial impact on a team's likelihood of securing victory, with a consideration of potential exceptions. It also investigates to see if there is an increased likelihood of winning the match for teams that win the toss. Additionally, we aim to discern patterns surrounding teams that demonstrate exceptional prowess or face challenges when competing in away matches.

4 RELATED WORK

Chalkias et al. (2023) [1] investigates the sentiments expressed by users in comments on educational YouTube videos and identifies video attributes influencing viewer experience through topic clustering. The study analyzed 167,987 comments using the YouTube Data API and employed sentiment analysis tools like VADER and TextBlob, alongside Latent Dirichlet Allocation (LDA) for topic clustering.The sentiment analysis revealed that most comments were neutral, followed by positive, and few negative sentiments. VADER and TextBlob showed comparable results, though TextBlob identified slightly more extreme sentiments due to its methodology. The LDA analysis identified key themes related to video content attributes such as animation and music, which potentially impact viewer engagement and learning.

The findings contribute to understanding the educational potential of YouTube, suggesting that educational content creators can enhance learning experiences by focusing on elements that resonate with viewers. In similar lines we have also used VADAR for understanding sentiment of comments.

5 DATA

5.1 Introduction to Data Source

In order to analyze the impact and engagement surrounding the topic of keywords entered in search bar on YouTube, we utilized the YouTube Data API to access various resources such as activities, videos, search, and comments. Through these API endpoints, we were able to gather a comprehensive dataset containing valuable information pertaining to videos related to the specified keyword. For our project, we have taken keywords as "neural networks" and fetched data of top 50 channels.

5.1.1 *Data Retrieval from YouTube Data API.*

Upon querying the YouTube Data API with the search term "neural networks," we retrieved a multitude of videos matching the search criteria. From these search results, we extracted pertinent information to construct our dataset. The YouTube Data API provides resources to extract data under different module. It is organised effectively to easily identify the different types of resources that can be retrieved using the API.

A "search" resource result contains information about a YouTube video, channel, or playlist that matches the search parameters specified in an API request. An "activity" resource contains information about an action that a particular channel, or user, has taken on YouTube. The actions reported in activity feeds include rating a video, sharing a video, marking a video as a favorite, uploading a video, and so forth. A "channel" resource contains information about a YouTube channel. A video resource represents a YouTube video and is uniquely identified by an id. A "commentThread" [3] resource contains information about a YouTube comment thread, which comprises a top-level comment and replies, if any exist, to that comment. This resource can represent comments about either a video or a channel based on the id supplied.

5.2 Attributes for analysis

The key features extracted include:

- Number of Views: The total number of views garnered by each video, indicating its popularity and reach among viewers.
- Number of Comments: The count of comments posted on each video, providing insights into the level of engagement and interaction generated by the content.
- Date of Published: The timestamp indicating the date and time when each video was published on the platform, enabling temporal analysis and trend identification.
- Number of Likes: The quantity of likes received by each video, reflecting the audience's positive reception and appreciation towards the content.
- Captions Availability: A boolean indicator denoting whether each video has captions or not, facilitating accessibility and catering to diverse audience needs.
- Description: The textual description accompanying each video, offering context and additional information about the content.
- Title: The title of each video, serving as a concise representation of its subject matter and attracting viewer attention.
- Comments: Extracted comments from viewers, providing valuable feedback, opinions, and discussions surrounding the topic of neural networks.
- Time of comment: The extraction of comment timestamps provides valuable insights into the temporal dynamics of user engagement and interaction with YouTube content related to the topic of interest.
  The Youtube API returns response in JSON format. Below is the snippet of data before processing.

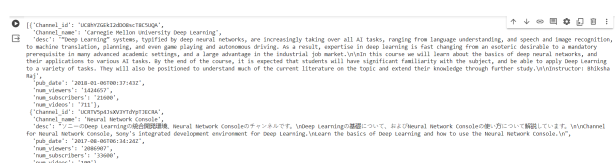


Figure 2: Snapshot of data before processing

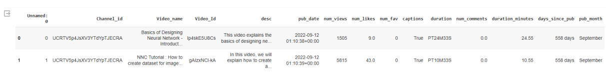
6 METHODS

6.1 Data Cleaning

6.1.1 Dealing with Memory issues. We have encountered memory issues while reading data of comments into working space. This has made us to read comments data into small chunks of 1000 size at a time and then looping through them in order to convert into dataframe chunk by chunk. As we have pulled data from API, there is very minimal chances of missing values. Null values are present only in 'desc' column which is categorical in nature and doesn't need any imputation method.

6.1.2 Parsing Date Column. The date columns in the dataset were subjected to parsing and standardization to ensure consistency and facilitate meaningful temporal analysis. This process involved converting the date information from its original format into a standardized date format that is easily interpretable and compatible with analytical tools.

After all steps of data cleaning and processing, the dataframe looks like as follows:



	Unnamed: 0	Channel_id	Video_name	Video_Id	desc	pub_date	num_views	num_likes	num_fav	captions	duration	num_comments	duration_minutes	days_since_pub	pub_month
0	0	UCRTV5p4JsXV3YTdYpTJECRA	Basics of Designing Neural Network - Introduct...	Ip4skE5U8Cs	This video explains the basics of designing ne...	2022-09-12 01:10:38+00:00	1505	9.0	0	True	PT24M33S	0.0	24.55	558 days	September
1	1	UCRTV5p4JsXV3YTdYpTJECRA	NNC Tutorial : How to create dataset for image...	gAIzxNCl-kA	In this video, we will explain how to create a...	2022-09-12 01:10:39+00:00	5815	43.0	0	True	PT10M33S	0.0	10.55	558 days	September

Figure 3: Snapshot of data after processing

7 EXPLORATORY DATA ANALYSIS(EDA)

Our goal is to unearth meaningful stories embedded in the data and set the stage for more targeted analyses. We aim not only to uncover the inherent characteristics of the dataset but also to lay the groundwork for informed decision-making in subsequent phases of our data exploration.

Using the data insights acquired through the analysis of different features, the creation of various info graphics, and the incorporation of our existing knowledge of how Youtube gives search results and what criteria makes it possible.

7.1 Data Visualisations

7.1.1 Top channels based on views. In order to identify the most viewed channels within the realm of content related to our topic of interest [keyword (neural networks)], we conducted an analysis of the total number of views garnered by individual channels. This analysis provides valuable insights into the popularity and reach of content creators within our target domain. **3Blue1Brown** channel emerged as the front runner with a remarkable total view count, reflecting its broad appeal and engaging content offerings. Following closely, **Sebastain Lague** channel and **StatQuest with Josh Starmer** channel demonstrated strong viewer engagement and resonance, securing prominent positions in the ranking. **Deep Learning AI** and **Neural Networks & Deep Learning** channels rounded out the top five with consistent performance and audience affinity, highlighting their relevance and influence within the niche content landscape.

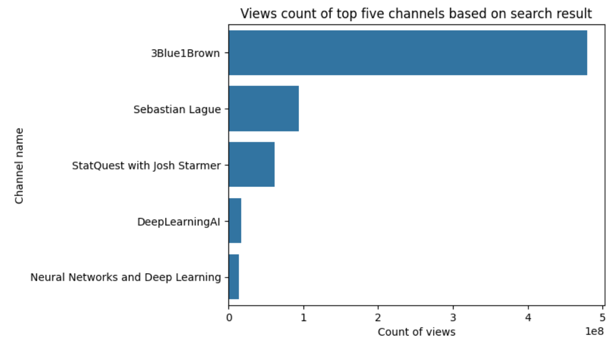


Figure 4: Top Five Channels Based on Number of Views: Insights

7.1.2 Insights into the Distribution of Content Creation Time Among Channels. In this pie chart, we have visualized the top 10 channels by distribution of their content creation time. This visualization provides a clear overview of the proportion of content creation efforts contributed by these channels. This analysis enhances our understanding of the content creation dynamics within our target domain and informs strategic decision-making regarding content partnerships, audience targeting, and engagement initiatives. This pie chart reveals **Alexander Amini** channel as the leading contributor, with 19% of the total content creation minutes, followed closely by **Killian Weinberger** channel with 13%. Remaining channels collectively contribute varying shares, ranging from 5% to 12%.

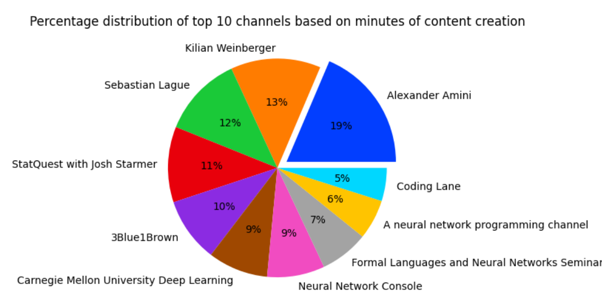


Figure 5: Breakdown of Top 10 Channels by Share of Content Creation Time

7.1.3 Analyzing the relationship between Video Duration and View Count. We utilized Seaborn's regplot function to generate a scatter plot examining the relationship between the duration of videos and the number of views they accrued. The data was filtered to include only videos with a view count of 100,000 or fewer views, ensuring a focused analysis on relatively less-viewed content. From this, we can infer that there is no preference for shorter videos among users which is proved by an almost straight line in the plot. It indicates very weak or no relationship between duration of video length and number of views.

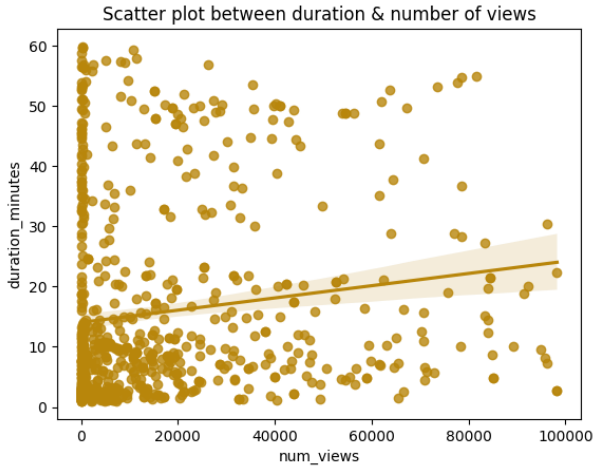


Figure 6: Linear regression between video length & views

7.1.4 Exploring videos durations Across Years: A Box Plot Analysis. In our analysis of the year-wise distribution of video duration, we utilized a box plot visualization to explore how the duration of videos has evolved over time. From the visualization, we observe variations in video duration across different years, with some years exhibiting a wider range of durations than others. It is evident that there is more range in the year **2018** spanning to larger length videos. This analysis provides valuable insights into trends and patterns in video duration over time, informing content creators and stakeholders about audience preferences and content consumption habits across different years.

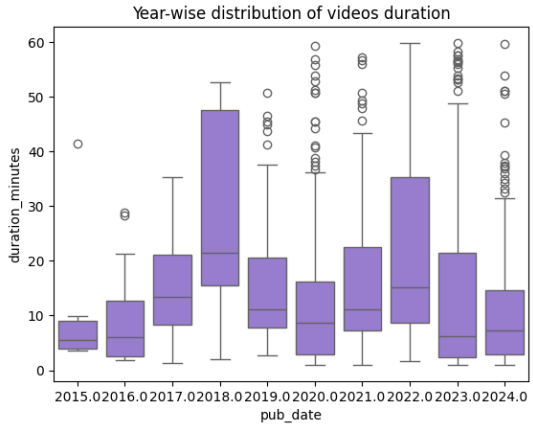


Figure 7: Year-wise Distribution of Video Duration

7.1.5 Distribution of Video Generation on a Monthly Basis. The histogram visualization depicts the month-wise distribution of the number of videos generated within our dataset. We can clearly see that there are more number of videos generated in early months of year compared to mid year months and again picking up pace in later months of the year.

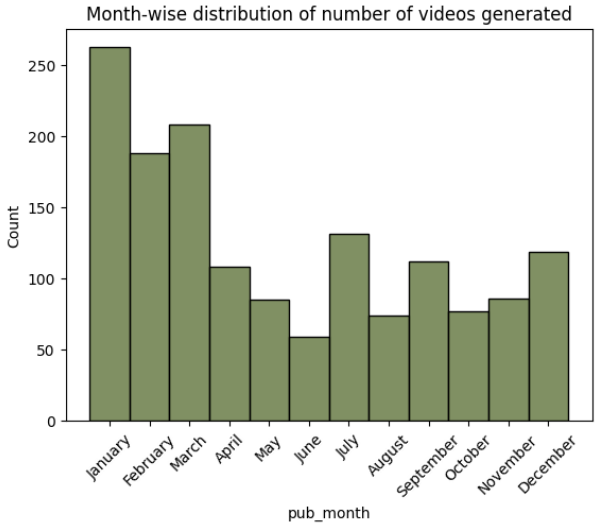


Figure 8: Month-wise distribution of number of videos generated

7.1.6 Exploring the Relationship Between Likes and Comments: Insights into Viewer Engagement Patterns. In our analysis, we utilized a scatter plot visualization as a tool to investigate the correlation between two key engagement metrics: the number of likes and the number of comments garnered by each video within our dataset. While it might seem intuitive that individuals would comment on a video they have watched or liked, the visualization allowed us to delve deeper into the underlying patterns of human behavior. Specifically, we sought to understand the propensity for individuals to engage with content by leaving comments, particularly in instances where they have expressed a positive or negative sentiment through liking or disliking the video. This observation underscores a fundamental aspect of human nature – the tendency to express opinions or reactions when experiencing an emotional response. Consequently, our analysis suggests a symbiotic relationship between the number of comments and likes, where heightened engagement in one metric often corresponds to increased activity in the other.

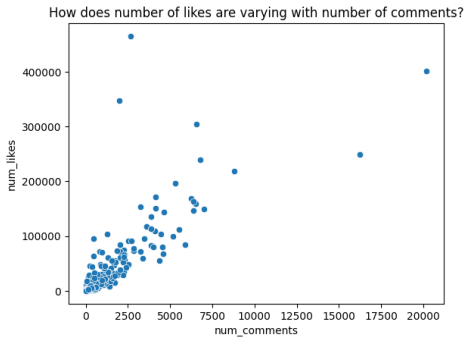


Figure 9: Scatterplot between comments and likes count

7.1.7 Word Cloud of videos descriptions. Upon analyzing the word cloud generated from the descriptions of videos related to neural networks, several key terms emerged prominently, providing valuable insights into the prevalent themes and topics within this domain. Notably, words such as "machine learning," "deep learning," "beginner," "topic," "AI," and "specialization" recurred frequently, indicating their significance and prevalence in the descriptions of these videos. The prominence of these terms suggests that the videos likely cover introductory concepts and discussions surrounding machine learning and deep learning, catering to individuals with varying levels of expertise, including beginners seeking to explore the topic further. Additionally, the mention of "AI" underscores the intersectionality of neural networks with artificial intelligence, reflecting the broader context in which these videos are situated. Overall, the word cloud provides valuable insights into the dominant themes and focal points of the videos, guiding viewers and content creators alike in navigating the expansive landscape of neural network-related content on the platform.

Figure 10: Lowest Winning margins: By Runs

7.1.8 Analyzing Viewer Engagement Based on Captions Presence in Videos. The strip plot visualization showcases the distribution of views for videos categorized based on the presence or absence of captions. The x-axis represents the captions status, with "Yes" indicating videos with captions and "No" indicating videos without captions, while the y-axis depicts the corresponding count of views. The plot provides a comparative analysis, allowing us to discern any potential differences in viewership between videos with and without captions. The title, "Count of views for videos with captions Vs without captions," succinctly summarizes the main focus of the visualization, facilitating clear interpretation. This visualization serves as a valuable tool for exploring the impact of captions on viewer engagement and accessibility within our dataset of videos, offering insights that can inform content creators and stakeholders in optimizing their video content strategies. We can see that videos having captions have more number of views.

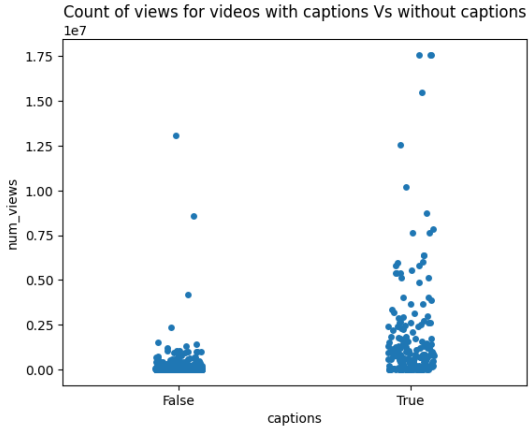


Figure 11: Plot of number of views against caption

7.1.9 Exploring Video Duration Distribution Over Time with Captions Analysis. This catplot visualization employs violin plots to explore the distribution of video durations across different years, with a distinction made between videos with and without captions. These plots display the distribution of video durations within each year, with the width indicating the density of videos at different duration values. The hue parameter separates the violin plots based on the presence or absence of captions, allowing for a comparative analysis of video durations between the two categories. The mean of duration of videos with captions is relatively higher than those videos without captions. By utilizing violin plots, this visualization effectively illustrates the variability and distribution of video durations over time, while also highlighting any discernible patterns or differences in duration between videos with and without captions.

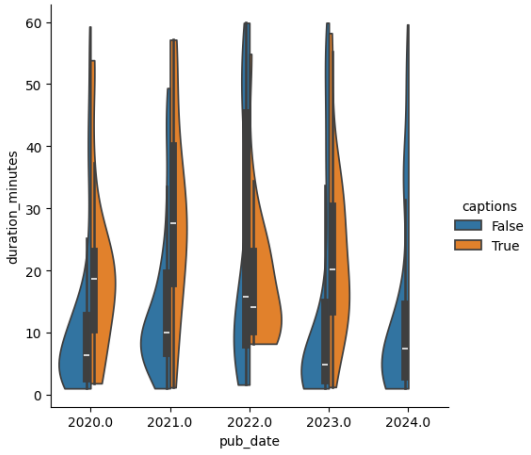


Figure 12: Plot of number of views against caption

7.1.10 Determining association between Duration and Number of Likes. This jointplot reveals that there is no clear conclusion to draw but this plot helps in understanding independence between these two features. Also, it is not very clear to interpret the relationship

between these features. This visualization serves as a valuable tool for understanding how video duration influences viewer engagement, as measured by the number of likes, while also considering the presence of captions as a potential factor influencing viewer interaction.

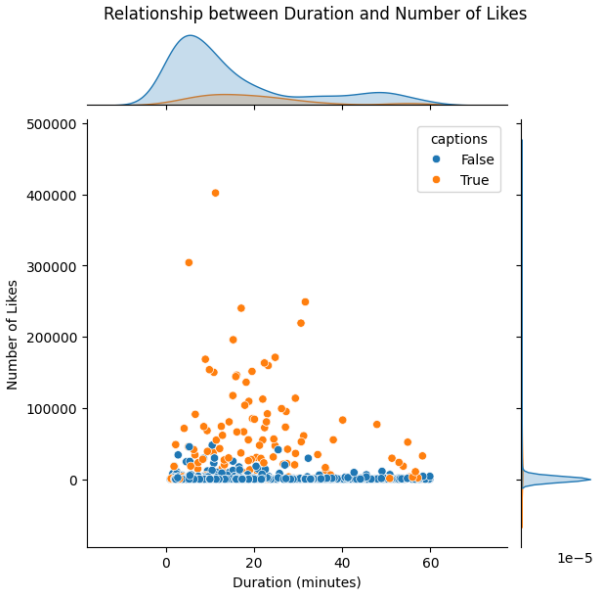


Figure 13: Jointplot of number of likes and duration of videos

7.1.11 Exploring Univariate Distribution of Video Durations. The swarm plot visualization provides a univariate distribution of the duration of videos within our dataset. Each data point on the plot represents an individual video, with the x-axis depicting the duration of the videos in minutes. From the visualization, we observe a wide range of durations, spanning from shorter to longer videos. Additionally, the swarm plot highlights the density of videos at lower duration values, with clusters of data points indicating areas of higher prevalence. Overall, this visualization offers valuable insights into the distribution and variability of video durations within our dataset, informing our understanding of the content landscape and helping to identify potential trends or patterns in video duration.

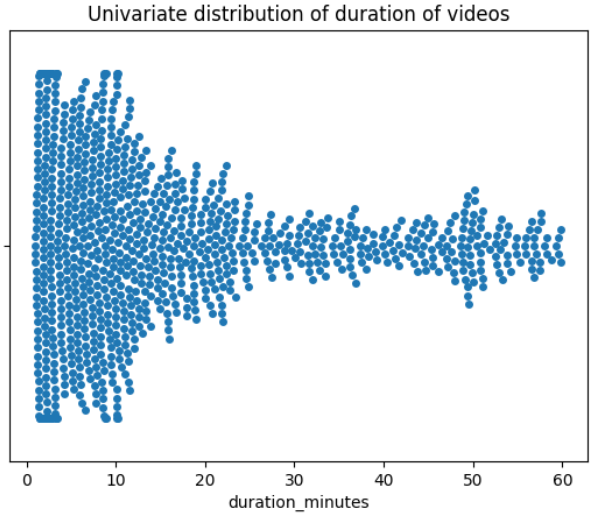


Figure 14: Swarmplot of number of likes and duration of videos

8 IMPLEMENTATION OF MODELS

In this section, we present the implementation of machine learning (ML) models on our YouTube dataset, aimed at uncovering insights into video performance and audience engagement. YouTube, as one of the largest platforms for video content consumption, presents a rich source of data encompassing a myriad of metrics such as views, likes, comments, and subscriber counts. Leveraging this dataset, our analysis seeks to harness the power of ML techniques to predict key performance indicators and understand the factors influencing video success. This sets the stage for our exploration of ML-driven analysis on YouTube data, offering a glimpse into the potential of data-driven approaches in unlocking the platform's dynamics and driving informed decision-making.

8.1 Sentiment Analysis of comments

In conducting sentiment analysis on the comments within our YouTube dataset, we employed natural language processing (NLP) techniques to extract insights into the sentiments expressed by viewers [1]. Initially, we preprocessed the text data by removing comments which are not in English and also the duplicated comments. Subsequently, we utilized VADER (Valence Aware Dictionary and sEntiment Reasoner) library to compute polarity scores for each comment. These polarity scores quantified the sentiment of each comment on a numerical scale, typically ranging from -1 (indicating extremely negative sentiment) to +1 (indicating extremely positive sentiment). We categorized the polarity values as 'positive' if greater than 0.05, 'negative' if less than -0.05, and 'neutral' otherwise. By aggregating these polarity scores across all comments for each video, we gained a comprehensive understanding of the overall sentiment towards the content. This approach to sentiment analysis provided valuable insights into the audience's emotional response to the videos on the YouTube platform, facilitating informed decision-making for content creators and marketers. Now, let's

have a look at snapshot of the dataset before and after transformation illustrating the preprocessing steps undertaken to prepare the data for analysis. The reason for conducting this sentiment analysis is to visualise whether sentiment associated with comments has impact on outreach of videos, in turn, affecting number of views.

	Unnamed: 0	channel_id	video_id	comment_text	comment_time
0	0	UCRTV5p4JsXV3YTdYpTJECRA	tFmQj7W4qlk	Can you provide a simple example calculation o...	2023-06-01T11:53:06Z
1	1	UCRTV5p4JsXV3YTdYpTJECRA	tFmQj7W4qlk	who chose this music :/	2022-12-27T16:17:47Z
2	2	UCRTV5p4JsXV3YTdYpTJECRA	tFmQj7W4qlk	Nice work. Just that the background music is s...	2022-12-02T06:19:11Z
3	3	UCRTV5p4JsXV3YTdYpTJECRA	tFmQj7W4qlk	For anyone interested for the music in the bac...	2022-11-09T08:45:29Z
4	4	UCRTV5p4JsXV3YTdYpTJECRA	tFmQj7W4qlk	Can you provide a simple example calculation o...	2023-06-01T11:53:06Z

Figure 15: Snapshot of data before analysis

Before transformation, the raw dataset consists of a collection of YouTube video data, containing features such as video titles, descriptions, comments, likes, views, number of favorites, no. of days since published and other relevant metrics. Each row represents a unique video entry, with associated attributes reflecting various aspects of video content and viewer engagement.

	Channel_id	num_viewers	num_likes	num_fav	captions	duration_minutes	days_since_published	negative	neutral	positive
0	UCRTV5p4JsXV3YTdYpTJECRA	4812	74.0	0	True	10.116667	589	1.0	1.0	2.0
1	UCRTV5p4JsXV3YTdYpTJECRA	7764	41.0	0	False	20.350000	589	3.0	1.0	1.0
2	UCRTV5p4JsXV3YTdYpTJECRA	8290	109.0	0	False	20.083333	1068	NaN	NaN	1.0
3	UCRTV5p4JsXV3YTdYpTJECRA	8966	109.0	0	False	22.600000	1076	NaN	NaN	1.0
4	UCRTV5p4JsXV3YTdYpTJECRA	6403	117.0	0	False	3.233333	1130	NaN	1.0	NaN

Figure 16: Snapshot of data after polarity scores

Following transformation, the dataset undergoes several steps to enhance its suitability for analysis. This includes text preprocessing techniques such as removing special characters, stopwords, and tokenization.

8.2 Random Forest for predicting number of likes

In our analysis of the YouTube dataset using random forest regression, we aimed to predict the number of likes for various videos. Our evaluation metrics provide insights into the model's performance. The R-squared (R^2) score of 0.95 indicates that our model explains approximately 95% of the variability in the number of likes, suggesting a strong ability to capture trends and patterns within the data. Despite this, there are opportunities for improvement to reduce prediction errors and enhance the model's accuracy. Future iterations could focus on refining the model by incorporating additional features or exploring alternative regression techniques. Additionally, delving deeper into the dataset and considering factors such as viewer engagement metrics or video content characteristics may offer valuable insights for optimizing strategies to maximize viewership on the YouTube platform. Below is the glimpse of dataset before predicting. This is the same input format which we have used in all our regression models. We have transformed all all categorical columns to one-hot encoding. Features were scaled down to same scale in order to avoid bias of model towards any large-scaled features. Below are the snippets of dataset before and after regression. After regression, the dataset is enriched with predicted values for the number of views, providing insights into the expected viewership for each video based on the learned patterns and relationships within the data. The results of this ML model are tabulated as below.We have obtained best tree in Random Forest Regressor at depth=20. Below is the graph which shows features with highest information gain. Feature importance in Random Forest refers to a technique used to understand the significance of input variables (features) in predicting the output variable (target) using a Random Forest model. The Random Forest algorithm provides a straightforward approach for evaluating feature importance, which is based on how much each feature decreases the impurity of the split (often measured using Gini impurity or entropy in classification tasks, and variance reduction in regression).

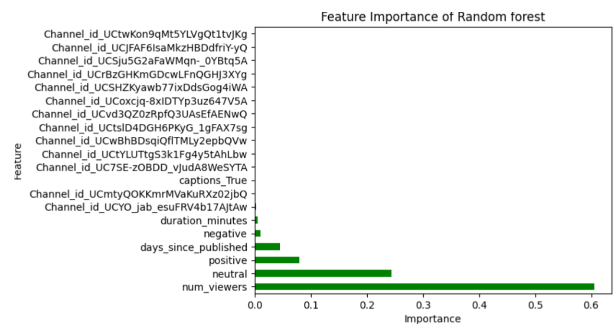


Figure 17: Understanding the Features Influencing Predictions

	num_viewers	num_fav	duration_minutes	days_since_published	negative	neutral	positive	captions_True	Channel_id_UC4vAON9PsmKoyDGqftXWS-A	Channel_id_UC6-2NodomKW-ZFNs9Vuj8Mw	...	Channel_id_UCqA6
42	157950	0	8.433333	2564	10.0	28.0	148.0	0.0	0.0	0.0	...	
215	15407	0	5.100000	753	6.0	9.0	20.0	0.0	0.0	0.0	...	
193	389693	0	11.633333	2450	14.0	31.0	101.0	0.0	0.0	0.0	...	
453	89262	0	10.116667	2897	7.0	10.0	24.0	0.0	0.0	0.0	...	
576	17651	0	20.383333	2292	5.0	3.0	21.0	0.0	0.0	0.0	...	

Figure 18: Snapshot of data before regression

Metric	Value
MAE	1128
R-Sq	0.95

Below is the snippet of how output table looks like alongwith true values of number of likes.

	Num_likes	Predicted_Num_likes
589	2271.0	5108.956334
720	1081.0	1146.438030
374	516.0	568.052235
478	401.0	442.932694
574	1692.0	1878.830750

Figure 19: Snapshot of Random Forest predicted values

8.3 LGBM Regression of number of likes

In this study, we employed the LightGBM (Light Gradient Boosting Machine) regression model to predict the number of likes a video would receive on the YouTube platform. Our predictive model utilized various parameters extracted from the video metadata and user interactions, including the number of comments, presence of captions, number of likes, number of favorites, polarity score of comments, video duration, and the number of days since the video was published. By leveraging these features, we aimed to understand the factors influencing viewership and provide insights into predicting video performance on YouTube. We have used GridSearch to find best hyperparamater by varying max depth, number of leafs etc., by optimising mean absolute error. We Below is the glimpse of the dataset before and after transformation for this regression analysis. The results of evaluation metrics are tabulated as follows.

Metric	Value
MAE	1333
R-Sq	0.94

	Num_likes	Predicted_Num_likes
589	2271.0	2811.401852
720	1081.0	1196.680802
374	516.0	587.699274
478	401.0	519.800037
574	1692.0	1490.263637

Figure 20: Snapshot of LGBM predictions

The table above presents the predicted number of likes generated by our LightGBM regression model alongside the actual number of likes observed for a sample of five videos from the dataset.

8.4 XGBM Regression of number of likes

Here, we have deployed XGBoost (Extreme Gradient Boosting) regression methodology to predict the number of likes of YouTube videos. Leveraging a comprehensive array of parameters sourced from video metadata and user engagements, our predictive framework delved into key indicators such as comment count, presence of captions, likes, favorites, sentiment analysis scores, video duration, and publication age. By harnessing the robust capabilities of XGBoost, we endeavored to unearth the nuanced dynamics driving viewership patterns, aiming to provide actionable insights for enhancing video performance and engagement on the YouTube platform. We have used GridSearch to find best hyperparamater by varying max depth, nestimators etc., by optimising mean absolute error.

Metric	Value
MAE	1124
R-Sq	0.95

	Num_likes	Predicted_Num_likes
589	2271.0	4322.191895
720	1081.0	1157.703979
374	516.0	571.750977
478	401.0	442.534393
574	1692.0	1679.945557

Figure 21: Snapshot of XGBoost predicted values

The table above presents the predicted number of likes generated by our XGboost regression model alongside the actual number of likes observed for a sample of five videos from the dataset.

8.5 Comparison of Model performance

8.5.1 Interpretation of Metrics. Mean Absolute Error (MAE): This metric measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences are weighted equally. A lower MAE value indicates a better fit of the model to the data.

R-squared (R-sq): This is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. The higher the R-sq, the better the model fits your data. An R-sq of 0.95 means that 95% of the variance in the dependent variable is predictable from the independent variables.

LightGBM shows a slightly lower R-sq of 0.94, indicating a somewhat less accurate fit to the variance in the data compared to Random Forest and XGBoost. Both Xgboost and Randomforest show very similar performance in terms of R-sq, each achieving a score of 0.95, which is excellent.

The MAE is very close as well, with XGBoost performing slightly better (1124.83) compared to Random Forest (1128.83). This suggests that, on average, XGBoost's predictions are closer to the actual values. However, we can prefer Random Forest over XGBoost for its straightforward interpretability regarding feature importance.

9 CONCLUSION

In conclusion, our YouTube analytics project has provided valuable insights into the performance of various machine learning models in predicting video engagement metrics. Mean Absolute Error (MAE) served as a critical metric for evaluating the accuracy of predictions, measuring the average magnitude of errors without considering

their direction. A lower MAE signifies a better fit of the model to the data. Additionally, R-squared (R-sq) offered a statistical measure of how well the data align with the fitted regression line, with higher values indicating a superior fit of the model to the data.

In summary, while XGBoost may have a slight edge in prediction accuracy, the interpretability of Random Forest makes it a favorable choice for our YouTube analytics project, enabling clearer insights into the factors influencing video engagement metrics.

Our aim of the project always revolved around answering the questions we had before kick-starting the analysis. Below are the findings from our project:

(1) The 3Blue1Brown channel stood out with the highest total views, showing its wide appeal and interesting content. Sebastain Lague's channel and StatQuest with Josh Starmer's channel also had strong engagement, ranking closely behind. Deep Learning AI and Neural Networks & Deep Learning channels completed the top five with steady performance and a loyal audience, indicating their significance in the niche content scene.
(2) Alexander Amini emerged as the top contributor with 19% of content creation minutes, followed closely by Killian Weinberger, while the rest of the channels collectively contribute shares ranging from 5% to 12%.
(3) There's no user preference for shorter videos, as seen in the nearly straight line on the plot, indicating little to no connection between video length and views.
(4) There are more videos created in the beginning and end of the year compared to the middle months.
(5) Our analysis indicates that more comments often go hand in hand with more likes, suggesting a close relationship between the two engagement metrics.
(6) The word cloud analysis of neural network-related video descriptions reveals prevalent themes such as "machine learning," "deep learning," and "AI," indicating a focus on introductory concepts catering to viewers with varying expertise levels, while also highlighting the intersectionality of neural networks with artificial intelligence, guiding both viewers and content creators through the expansive landscape of content on the platform.
(7) The analysis revealed that there is importance of captions in enhancing viewer engagement and accessibility, indicating that videos with captions tend to garner more views.
(8) There is no clear evidence emerged out for understanding the influence of video duration on viewer engagement, while also considering captions as a potential factor affecting viewer interaction. So, we can see them as independent factors.
(9) There are more number of shorter videos compared to lengthier ones in the range of 2-13 minutes.

9.1 Significanve of findings

The findings of this project are instrumental in advancing our understanding of viewer engagement dynamics and optimizing content strategies within the realm of online video analytics. By identifying top-performing channels, contributors, and prevalent themes, the analysis provides actionable insights for content creators, marketers, and platform administrators. These insights enable stakeholders to emulate successful practices, forge strategic partnerships, and allocate resources effectively to maximize engagement and viewership. Moreover, the analysis of engagement metrics, including likes, views, comments, and video duration, sheds light on the factors influencing viewer interaction and content consumption patterns. Understanding these dynamics empowers content creators to tailor their content to meet audience preferences, optimize content packaging strategies, and enhance viewer engagement.

Furthermore, the analysis of video metadata, such as titles, descriptions, and captions, underscores the importance of effective content presentation and accessibility features in driving viewer engagement and retention. By leveraging these insights, content creators can enhance the accessibility, inclusivity, and discoverability of their content, thereby expanding their reach and impact on the platform. In summary, the findings from this project serve as a valuable resource for content creators, marketers, and platform administrators, offering actionable insights and best practices for optimizing content strategies, enhancing viewer engagement, and achieving success in the dynamic landscape of online video content.

9.2 Improvement Suggestions

- Further Exploration of Viewer Preferences: Conducting surveys or collecting additional data from viewers can provide deeper insights into their preferences, such as preferred video formats, topics of interest, and viewing habits. This data can complement the quantitative analysis of engagement metrics and offer qualitative perspectives on what resonates with the audience
- Analysis of Engagement Metrics Across Different Topics: Segmenting the data based on video categories or topics can reveal patterns and trends specific to different content genres. Understanding how engagement metrics vary across categories can help content creators tailor their content to specific audience interests and preferences
- Investigation into the Impact of Video Metadata: Analyzing the influence of video metadata, including titles, descriptions, tags, and thumbnails, on viewer engagement can provide insights into effective content packaging strategies. A/B testing different metadata elements can help identify which attributes contribute most to attracting viewers and encouraging interaction
- Continuous Monitoring of Trends and Methodologies: Video analytics is a dynamic field, with new trends, technologies, and methodologies emerging regularly. Continuous monitoring of industry developments and updates in analytics tools and techniques ensures that analysis methods remain up-to-date and relevant. This ongoing learning process allows for the refinement and improvement of analytical approaches over time

9.3 Potential Use Cases

- Informing Content Creation Strategies: Insights from video analytics can guide content creators in developing content

strategies that resonate with their target audience. By understanding what types of content perform best in terms of engagement metrics, creators can focus their efforts on producing content that is more likely to attract and retain viewers.

- Guiding Marketing and Promotion Efforts: Video analytics data can inform marketing and promotion strategies by identifying the most effective channels, platforms, and promotional tactics for reaching and engaging the target audience. This data-driven approach allows marketers to allocate resources more efficiently and maximize the impact of their promotional campaigns.
- Enhancing Accessibility and Inclusivity: Understanding the impact of features like captions on viewer engagement can help content creators prioritize accessibility and inclusivity in their content. By providing captions and other accessibility features, creators can ensure that their content is accessible to a wider audience, including viewers with disabilities.
- Identifying Collaboration Opportunities: Analyzing common themes and interests among channels can help identify opportunities for collaboration or content partnerships. By partnering with channels that share similar audiences or content themes, creators can expand their reach and attract new viewers who may be interested in their content.

Overall, the suggestions and use cases outlined above highlight the potential of video analytics to inform and optimize content creation, marketing, and promotion strategies, ultimately leading to more engaging and impactful content experiences for viewers.

REFERENCES

[1] Ilias Chalkias, Katerina Tzafilkou, Dimitrios Karapiperis, and Christos Tjortjis. 2023. Learning Analytics on YouTube Educational Videos: Exploring Sentiment Analysis Methods and Topic Clustering. *Electronics* 12, 18 (2023). https://doi.org/10.3390/electronics12183949

[2] M. Laeeq Khan and Aqdas Malik. 2022. *Research YouTube: Methods, Tools, Analytics.* 651–663.

[3] Samant Saurabh and Sanjana Gautam. 2019. Modelling and statistical analysis of YouTube's educational videos: A channel Owner's perspective. *Computers Education* 128 (2019), 145–158. https://doi.org/10.1016/j.compedu.2018.09.003

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009