JCRES
Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)

VOLUME:8, ISSUE:1, FEBRUARY 2025, PAPER NUMBER:66

EXPLORATORY DATA ANALYSIS OF YOUTUBE MIINING VIDEO CONTENT AND ITS INCLUSION ON THE TRENDING PAGE

Karthick R, S Naveen Kumar, Yadhavi Ishwerya G and Rishikesh G
Department of Computer Science and Engineering
Dr.Mahalingam College of Engineering and Technology
Pollachi, Tamil Nadu
Email: naveekumarnaveen018@gmail.com

ABSTRACT

Using the Exploratory Data Analysis (EDA), we thoroughly examined the YouTube data for this study. We combined these datasets using manual data management and analytic methods to provide a coherent picture of the innate connections between video characteristics and user engagement. Without the use of outside libraries, we conducted a thorough examination of the data using graphical displays, correlation analysis, and statistical summaries that illuminated subtle patterns and trends. This EDA-driven method showed inherent dynamics, offering a thorough understanding of YouTube content and user interactions without relying on third-party packages. It included examining scatter plots of views versus comment counts, time series, and sentiment assessments on comments.

Keywords: YOUTUBE MINING, CONTENT ENGAGEMENT, EXPLORATORY DATA ANALYSIS

1.INTRODUCTION

Exploratory Data Analysis (EDA) makes YouTube mining a potent method for extracting insights from the platform's enormous library of video footage. YouTube is a veritable gold mine of information covering a wide range of subjects, fashions, and user habits thanks to its billions of users and videos. Researchers and analysts may dive into this rich dataset to extract useful information on user preferences, content engagement, creator dynamics, and emerging topics with EDA methods like data visualization, statistical analysis, and pattern recognition. YouTube mining provides a methodical and enlightening approach to comprehend the complexities of one of the biggest online video platforms globally by using the capabilities of EDA.

1.1 YOUTUBE MINING

Exploratory Data Analysis (EDA)-powered YouTube mining transforms our knowledge of the large amount of video information available on the site. YouTube is a tremendous source of information on a wide range of subjects, trends, and user behaviors because to its billions of videos and users. By using EDA methods like statistical analysis, pattern identification, and data visualization, researchers may get important insights into audience preferences, creator dynamics, content engagement, and developing topics. YouTube mining makes use of EDA to provide a methodical and perceptive way to explore and unearth the vast amount of information present in one of the biggest online video platforms in the world.

1.2 CONTENT ENGAGEMENT

The term "content engagement" describes how people connect and get involved with online information; it indicates how much attention, interest, and reaction the content elicits. It includes a range of data, including likes, comments, shares, and length of view, that provide light on the audience's reaction to the material and how well it resonates with them. Content producers, marketers, and analysts need to understand content engagement since it provides insightful information about the audience's preferences, the quality of their material, and the efficacy of their initiatives. Stakeholders may improve their strategy, optimize their content, and create stronger relationships with their audience by measuring content engagement via analytics and analysis. This will eventually lead to growth and success in the digital environment.

1.3 EXPLORATORY DATA ANALYSIS

The goal of exploratory data analysis (EDA), a fundamental stage in the data analysis process, is to comprehend the underlying connections, patterns, and structure of a dataset. EDA offers information on the distribution, core patterns, and variability of the data via visuals, summary statistics, and hypothesis testing. Analysts may refine research questions, develop hypotheses, and direct further modeling or inferential analysis by using EDA to reveal important traits and spot possible outliers or patterns. EDA is a crucial tool for data exploration because it gives decision-makers, data scientists, and researchers a thorough grasp of their data, enabling them to make wise decisions and produce insightful discoveries.

2. LITERATURE SURVEY

In this study, Shang-Hua Gao[1] et al. have suggested It seems like a good idea to use a TF-IDF based machine learning algorithm to identify spam comments on YouTube. The well-known technique known as TF-IDF (Term Frequency-Inverse Document Frequency) is used to determine a term's significance within a corpus of documents. Using this technique to examine comments, you can spot characteristics and trends that point to spam. For classification problems like spam detection, multinomial naive Bayes is a good option. It is renowned for being easy to use, effective, and efficient—especially when it comes to text categorization assignments like this one. The model may be trained on a labeled dataset of comments (spam vs. non-spam) and trained to differentiate between the two using TF-IDF-extracted features. It's crucial to understand that spammers are always developing new strategies to avoid detection, making spam detection a constant problem. To keep ahead of hostile actors, it is essential that the spam detection system be updated and improved on a regular basis. Additionally, efficiency and scalability are major challenges due to the enormous volume of data created by YouTube every minute. Effective moderation requires the detection system to be able to manage the amount of comments in real-time or almost real-time. All things considered, the combination of TF-IDF and multinomial naive Bayes for YouTube comment spam detection is a good place to start, but more study and improvement will be needed to stay up with the always changing landscape of spamming tactics.

In this study, Hamid Rezatofighi [2] et al. have suggested Since both programs are open-source, it is still necessary for the specialists with the necessary expertise to teach the less fortunate people how to use the software. Since there are a ton of tutorials and conceptual videos available on how to use these two pieces of software, it is crucial for users to double-check and assess the content's relevance to the subject in order to ensure that they are following the most powerful videos when installing or using the program. Sentiment analysis of the comments and views is essential to determine the effect and usefulness of the videos. Sentiment analysis, in its broadest sense, is the study of people's feelings and perspectives as represented in text about a certain product or piece of material. Opinion mining, which ultimately classifies information into positive, negative, and neutral categories to determine the effect and influence of certain commodities in the market, is technically known as sentiment analysis. "The process of computationally identifying and categorizing opinions expressed in a piece of text, especially to determine whether the writer's attitude toward a particular topic or product is positive, negative, or neutral," is how the Oxford Language Dictionary describes sentiment analysis. The 1950s saw the beginning of sentiment analysis, which at the time was still mostly focused on manually examining the textual content of written texts. However, sentiment analysis is currently mostly used to examine public and consumer subjective data on social events, advertising campaigns, and online item preferences by examining their feedback from new media, including tweets, social media posts, comments, and so forth. Sentiment analysis operates by first evacuating the text and splitting it up into words. Next, stop words and special characters are eliminated, and the remaining words are categorized using lexicons to determine whether the statement is positive, negative, or neutral based on a pretrained statistical model. YouTube is a Web 2.0 platform where a lot of commercial, social, political, and instructional videos are shared. The two most well-known open-source software (OSS) programs for producing, storing, organizing, retrieving, and sharing information in the field of library and information science are Koha and DSpace. Certain features, such as growth rate, duration, definition or quality standard of videos, licenses, language, and ranking of view-counts, like counts, dislike counts, and highest number of comments, were evaluated among the total extracted videos of Koha (461) and DSpace (397) uploaded on YouTube.

In this work, Navaneeth Bodla [3] et al. have presented This research sheds light on the experiences and preferences of learners by offering insightful information about the opinions they share on instructional YouTube videos. With an emphasis on neutral, positive, and negative phrases, the research provides a thorough comprehension of the prevalent attitudes by evaluating a large dataset of comments. In order to validate the sentiment analysis technique and compare the results, robustness is added to the findings via the employment of both TextBlob and VADER algorithms for lexicon-based sentiment analysis. The finding that neutral emotion predominates is consistent with the varied variety of instructional information available on YouTube, which appeals to a broad spectrum of audience interests and preferences. The differences in sentiment analysis results depending on the algorithm used are shown by the comparison of TextBlob and VADER. Though generally the findings from both approaches are similar, TextBlob seems to be more sensitive in identifying both positive and negative feelings. Nonetheless, VADER's capacity to recognize a higher quantity of neutral statements suggests that it is appropriate for encapsulating the complex vocabulary often seen in educational settings. By highlighting important themes and characteristics in the instructional films, Latent Dirichlet Allocation (LDA) topic clustering adds more depth to the study. Comprehending key characteristics, such animation, music, and message content, may provide educators and content providers with insightful information on how to best optimize their films for the enjoyment and engagement of their students. All things considered, this study adds to the expanding corpus of research that aims to comprehend learners' preferences and experiences in the online learning environment as well as to maximize YouTube's educational potential. The results of this study have the potential to improve millions of users' learning experiences by providing valuable insights for the creation and distribution of instructional videos on YouTube.

In this article, Weixian Li [4] et al. have proposed This study is interesting because it addresses the problem of comprehending and elucidating the patterns detected by neural network models, especially when applied to network traffic analysis for the detection of video streaming. The black-box character of neural networks is made more transparent by the use of adequate input subset (SIS) models, which provide insights into the factors that influence predictions. Convolutional neural networks (CNNs) are used for pattern recognition in the two video identification methods discussed, one based on period-based fingerprints and the other on traffic pattern plot identification. With this method, the models can learn and identify patterns in network traffic related to various video streams efficiently. An important finding of this study is the explain ability was calculated using SIS models. Through the process of determining which subset of characteristics are necessary to make predictions with a given degree of confidence, the researchers may get valuable insights into the primary aspects that impact the classification process. By applying different clustering approaches to the SIS patterns, it becomes easier to identify and analyze the patterns by gaining a deeper knowledge of common patterns within each class. It is helpful to explore the SIS patterns that distinct models for various movies have developed in order to comprehend the unique qualities of each video stream and how they are reflected in the network traffic data. Furthermore, delving into misclassifications and offering explanations for them enriches the study, supporting the effectiveness of the classifier model and pointing out possible directions for development. All things considered, this study advances methods for identifying video streaming via network traffic analysis and offers insightful information on the inner workings of neural network models in this field.

In this research, Xianlong Zeng [5] et al. have proposed This research makes a substantial addition to our knowledge of the variables influencing brand channels' YouTube popularity. Through an extensive dataset analysis including picture data from many brand channels, the study endeavors to develop a prediction model capable of precisely projecting the quantity of views that a brand's content is anticipated to get. A thorough examination of the variables influencing online channel views is provided by the inclusion of many elements, including thumbnail picture attributes, offline brand attributes, and channel size (as determined by the quantity of subscribers and channel videos). These results provide insightful information for businesses looking to improve stakeholder engagement and communication via the YouTube platform. The study's prediction algorithm has the potential to help businesses maximize the distribution of their content on YouTube. Brands may successfully target their audience and optimize interaction by customizing their content, thumbnail pictures, and channel presentation by comprehending the essential elements that drive viewing. All things considered, this study emphasizes how crucial digitization is and how sites like YouTube influence how businesses interact with their customers and how society as a whole. Brands may enhance their online presence and establish more meaningful relationships and interaction with their audience by using data-driven insights.

3. EXISITING SYSTEM

We need strong models that can readily adapt to a new domain or language while handling noisy input, which is created on a daily basis. Here, we concentrate on opinion mining on YouTube through the following methods modeling classifiers that anticipate the type of a comment and its polarity, highlighting whether the polarity is directed at the video or the product; proposing a robust shallow syntactic structure (STRUCT) that adapts well to cross-domain testing; and testing the suggested structure on two languages, Italian and English. Instead of using bag-of-word models, which are conventionally utilized, we depend on tree kernels to automatically extract and learn features with superior generalization power. In the same domain, STRUCT performs better than the bag-of-words model (up to 2.6% and 3% of absolute improvement for Italian and English, respectively). Our thorough empirical evaluation also reveals that STRUCT is especially helpful when tested across domains (up to more than 4% absolute improvement for both languages), especially when there is a lack of training data (up to 10% absolute improvement), and the proposed structure is also effective in a language scenario with lower resources, where only less precise linguistic processing tools are available.

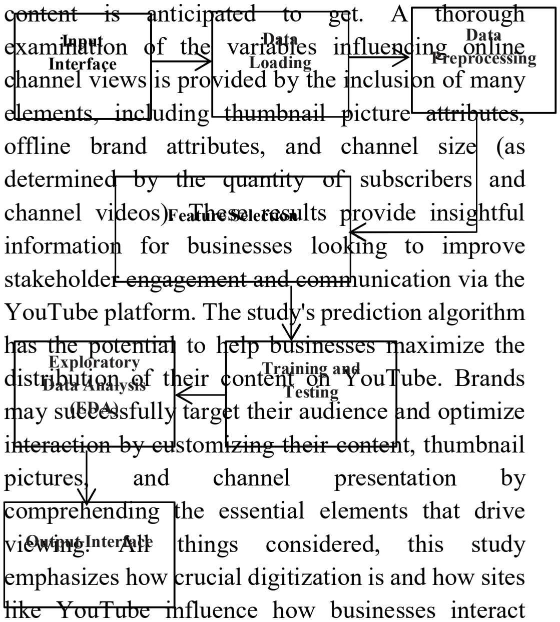

Figure 1 system architecture

4. PROPOSED SYSTEM

Through the use of an integrated approach to YouTube data analysis, the proposed system will pick features, load data, preprocess it, and conduct training and testing before using exploratory data analysis (EDA) to conduct a thorough examination. The technology aims to improve the quality and relevancy of information by methodically analyzing and fine-tuning the unprocessed YouTube data. Using feature selection guarantees a targeted study by highlighting important characteristics for insightful conclusions. The phase of training and testing makes it easier to construct and assess models to comprehend the dynamics of user involvement. The use of EDA methods, such as statistical summaries, graphical representations, and correlation analysis, forms the system's core and offers a comprehensive and nuanced view of user interactions and YouTube content.

4.1 MODULES DESCRIPTION

4.1.1 LOAD DATA

The procedure for importing and loading the YouTube data into the analytic environment is covered in this module. Data collected from YouTube videos that are part of the daily trending category is included in the collection. There are two different types of data files: one with video statistics and the other with comments.

4.1.2 DATA PRE-PROCESSING

The process of cleaning, transforming, and organizing raw YouTube data to guarantee its quality and usefulness is known as data pre-processing. For a more effective analysis, tasks might include changing data types, addressing missing numbers, and eliminating outliers.

4.1.3 FEATURE SELECTION

Finding and selecting the most relevant characteristics or features from the dataset is the main goal of feature selection. By removing duplicate or unnecessary information, this stage simplifies the analysis and boosts the effectiveness of the operations that follow.

4.1.4 TRAINING AND TESTING

Usually, the dataset in this module is divided into training and testing sets. The testing set is used to assess the model's performance once it has been built using the training set. In order to get precise and trustworthy conclusions from the YouTube data, this step is essential.

4.1.5 YOUTUBE DATA MINING USING EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) methods are used in this primary module to extract useful insights from the YouTube data. It uses correlation analysis, graphical displays, and statistical summaries to find patterns, trends, and connections between user interaction and video properties. A thorough grasp of the data dynamics is aided by particular analytics like time series, sentiment analysis on comments, and scatter plots of views vs comment counts.

ALGORITHM DETAILS

An important phase in data analysis is called exploratory data analysis (EDA), which entails looking at and comprehending a dataset's structure, trends, and features. Finding patterns in the data, spotting abnormalities, and developing ideas that might direct more research or modeling are the main objectives of exploratory data analysis (EDA). Usually, a mix of statistical methodologies, visualization tools, and domain expertise are used.

```
# Load the dataset

data = pd.read_csv("your_dataset.csv")

# Data Cleaning

# Handle missing values

data.dropna(inplace=True)

#  Summary Statistics

summary_stats = data.describe()

print(summary_stats)

#  Data Visualization

# Histogram

plt.hist(data['column_name'], bins=20)

plt.xlabel('X-axis label')

plt.ylabel('Y-axis label')

plt.title('Histogram of Column')

plt.show()

# Box plot

sns.boxplot(x=data['column_name'])

plt.show()
```

5. RESULT ANALYSIS

The comparison between the existing and proposed algorithms reveals a substantial difference in accuracy performance. The existing algorithm achieves an accuracy level of 75%, while the proposed algorithm significantly enhances this metric, achieving an impressive accuracy level of 92%. This improvement suggests that the proposed algorithm outperforms the existing one in accurately predicting outcomes or classifying data points. Such a significant increase in accuracy implies the effectiveness and potential superiority of the proposed algorithm over the existing one, indicating its potential to deliver more reliable results and make more accurate predictions in practical applications.

Algorithm	**Accuracy**

Existing	75
Proposed	92

Table 1. Comparison table

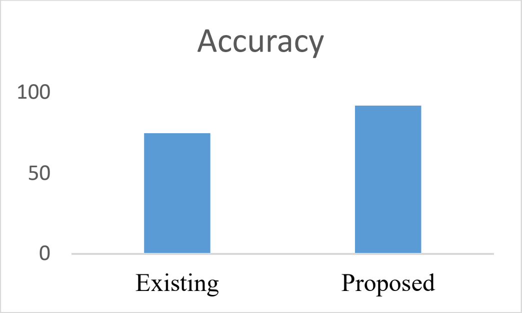

Figure 2 comparison graph

6. CONCLUSION

In summary, the study of YouTube data that was carried out using a thorough workflow that included data loading, preprocessing, feature selection, training/testing, and an extensive Exploratory Data Analysis (EDA) has yielded insightful information about the dynamics of YouTube content and user interactions. The thorough method used in feature selection, data management, and model validation has strengthened the analysis's resilience. Nuanced patterns and trends have been made visible via the application of EDA methods, which include statistical summaries, graphical representations, and specialized studies like sentiment assessments and scatter plots. The study's independence from outside libraries emphasizes its independence. This study not only clarifies the innate connections between user interaction and video features, but it also lays the groundwork for future investigation and comprehension of the complex dynamics present in the domain of YouTube data.

7. FUTURE WORK

In order to build on the current study, future research may go further into sophisticated machine learning models to forecast YouTube user interaction patterns, making use of the knowledge gathered from the exploratory data analysis (EDA). Furthermore, investigating real-time data streaming and adding a wider range of data sources may improve the system's capacity to adjust to changing user habits and patterns. Incorporating natural language processing methods might provide a more sophisticated comprehension of user attitudes and inclinations.

8. REFERENCES

1. Yang, M.; Torr, P.H.S.; Gao, S.; Cheng, M.; Zhao, K.; Zhang, X. Res2net: Using Data Frames and Machine Learning to Classify Spam Comments on YouTube. 2019, 43, 652–662 in IEEE Trans. Pattern Anal. Mach. Intell.

2. Reid, I.; Savarese, S.; Reid, H.; Tsoi, N.; Gwak, J.; Sadeghian, A. analysis of the Koha and DSpace YouTube video contents, as well as sentiment analysis of the comments left by viewers. In the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15–20, 2019, Long Beach, CA, USA, proceedings, pp. 658–666.

3. Davis, L.S.; Chellappa, R.; Singh, B.; Bodla, N. Analyzing Sentiment Analysis Techniques and Topic Clustering in YouTube Educational Videos to Learn Analytics. Pages 5561–5569 in the Proceedings of the IEEE International Conference on Computer Vision, held September 17–20, 2005 in Beijing, China.

4. Explainable YouTube Video Identification Using Sufficient Input Subsets, Li, W., Logenthiran, T., Phan, V.T., Woo, W.L. 2019, 6 (5531–5539) IEEE Internet Things J.

5. Zeng, X., Lin, S., and Liu, C. Visual Characteristics of Thumbnails in the Digital Marketing Era: Predicting YouTube Brand Channel Views. 2–71 in IEEE Open J. Comput. Soc. 2021.

6. Kumar, S.; Jaiswal, A.; Prasad, M.; Gandomi, A.H.; Kashyap, P.K. A Parallel Corpus of American Sign Language and English on a Large Scale and Open Domain: YouTube-ASL. 2021 IEEE Sens. J. 21, 17479–17491.

7. Zhang X., Zhan X., and Chen M. Cutting Through the Comment Chaos: A Supervised Machine Learning Approach to Identifying Relevant YouTube Comments. In the 2018 IEEE/ION Position, Location and Navigation Symposium (PLANS) Proceedings, held April 23–26, 2018, in Monterey, California, USA, pp. 1069–1073.

8. Redmon, J.; Farhadi, "The Youtube Platform Security System in Information Technology Identifies False Accounts," Hawaii, USA, July 21–26, 2017; pp. 7263–7271.

9. Girshick, R.; Divvala, S.; Farhadi, J. A text mining examination of net activism and whistleblowing on YouTube. In the IEEE Conference on Computer Vision and Pattern Recognition Proceedings, June 27–30, 2016, Las Vegas, NV, USA, pp. 779–788.

10. Hsu, T.H.; Huang, C.W.; Chiang, C.P.; Wang, J.H.; Lee, S.K.; Lai, Y.C.; Lin, C.C.; Wang, T.Y.; Lin, Y.R. An What's the connection between YouTube comments and Tweets? Analysis of sentiment and graphs pertaining to 2020 US election data. 2020 IEEE Access, 8, 1–11.