Journal of Theoretical and Applied Information Technology

30th September 2023. Vol.101. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BIG DATA-DRIVEN SUPPORT SYSTEM FOR YOUTUBE CHANNEL IMPROVEMENT

K. SUBHA[1], N. BHARATHI[2]

[1] Department of Computer Science and Engineering, College of Engineering and Technology, SRM Institute of Science and Technology, Vadapalani Campus, No.1, Jawaharlal Nehru Road, Vadapalani, TN, India.

[2]Associate Professor, Department of Computer Science and Engineering, College of Engineering and Technology, SRM Institute of Science and Technology, Vadapalani Campus, No.1, Jawaharlal Nehru Road, Vadapalani, TN, India.

E-mail: [1]sk3114@srmist.edu.in, [2]bharathn2@srmist.edu.in

ABSTRACT

YouTube is gaining a lot of traction and popularity. It has the potential to affect billions of people around the world, as the number of YouTube users continues to rise. YouTube is a video streaming platform owned by Google, with billions of subscribers and 400 hours of video posted every minute. Large volumes with complex data are called big data. The social network YouTube is one of the sources for generating such a high volume of data called big data. YouTube data is not structured data. It is a big challenge to store, process, and analyze such big data in real-time. YouTubers can check their channel performance with YouTube Analytics. The problem with YouTube Analytics is that it's impossible to check another competitor's channel. The proposed system will analyze real-time YouTube data from the list of channels. It will assist the YouTuber in finding out the Competitor's channel and how the competitors are doing well on social media platforms such as YouTube. The proposed work analyzes the YouTube data, finding competitors' channels using novel algorithms, and the results are represented in graphical form, which can be utilized by the person or any organization for their decision to improve their revenue.

Keywords: *YouTube, Data analysis, Big Data, Channel statistics, Real-time data.*

1. INTRODUCTION

Data plays an essential role in any application. The Internet, social media (YouTube, Facebook, Twitter, and WhatsApp), mobile phones, and IoT devices are the primary sources of data generation. According to the recent Statista, the generated data will reach above 180 zettabytes in 2025 [15]. A large amount of data is called big data, which can be in different formats, such as structured, unstructured, and semi-structured data. The big data has the characteristics of Volume, Variety, and Velocity, shown in Figure 1. Volume defines the size of the data; Velocity is the data generation speed; and Variety denotes different kinds of data. Big data can be used to gain better business insights, uncover hidden trends, and obtain other relevant information.

YouTube is a video streaming platform owned by Google, with billions of subscribers and 400 hours of video posted every minute. Every day, people watch different kinds of videos, such as education, kids learning, cooking, tourism, news, fun videos, etc., based on their interests. There are nearly billions of videos on YouTube, generating massive amounts of data called big data. YouTube has grown into a global learning and teaching platform. However, users frequently create and upload the content to the digital network. During the pandemic situation, most people were busy with online networks.

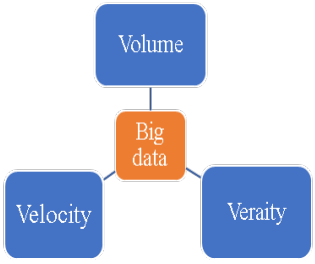

Figure 1: Big Data Characteristics

YouTube channels publish different kinds of videos, such as education, news, tourism, personal vlogs, political videos, healthcare, etc. YouTube political channels post many videos, especially during the election period [10]. Data analysis is the most crucial phase of data science. The first level of the data analysis of any YouTube channel is to find the number of uploaded videos, the total number of subscribers, the number of people who liked and disliked posted videos, and monthly uploaded videos. [13]

The rest of the paper is laid out as follows. Related work is discussed in Section 2. The problem statement and research question are explained in Section 3. Section 4 explains the different application areas using YouTube data. Section 5 contains the details of the proposed system and system architecture components. The experimental results are presented in Section 6. Section 7 contains the conclusion.

2. RELATED WORK

[1] The author explained the importance of YouTube channels during the pandemic situation, the education department such as tourism affected to visit the places for their study purpose to avoid such a problem, the author uses the online platform YouTube to see and explain the study and the producers' perspective to provide a complete examination of this YouTube channel over four years. The benefits and drawbacks of using YouTube in tourism education are discussed. There are suggestions and recommendations for tourism academics who want to become YouTube creators.

[2] The author examines the mediatization of the Islamic religion through an analysis of Islamic videos that have recently been uploaded to YouTube. The goal of this study is to investigate three linked questions. What are the evolving trends in Islamic videos on the internet? What are the most common ways viewers connect with online Islamic videos? What is the relationship between the indicators of interactions? The interaction variable's correlation is tested and tabulated.

[3] Political incivility is challenging task in online video-sharing platforms such as YouTube. Political incivility is created through false political advertising on YouTube. The author analyzed the YouTube channel data related to political incivility about former U.S. President Donald Trump. The proposed work explored three queries using dynamic network analysis and exponential random graph modeling. This work has some limitation that the intra group communication does not consider for identification of political incivility advertising.

[4] With the introduction of the internet, young people are using digital networks such as YouTube, Facebook, and Twitter. The main issue is how helpful the platform is for young people; it may contain all information. The author investigated what content is most popular among teenagers and how much time they spend on YouTube. The purpose was to see which websites they spent the most time on, which social networks they used, what kind of content they viewed, and whether they had been subjected to ill-treatment because of their digital network usage. Most of the kids spend their time watching fun videos and gaming. The work was limited to younger students.

[5] YouTube has evolved into a global educational platform for formal and casual learning. In different kinds of learning videos, finding the best source is a complicated process. The design work uses the aggregated ranking algorithm. The author uses a qualitative and quantitative analysis of more than 190 lists obtained from more than 100 websites to highlight critical elements and rank the educational relevant YouTube channels. The aggregated lists were then compared to track key elements such as the channel's lifetime, views of visitors, the total number of uploaded videos, and the count of subscribers. This study's limitation is the algorithm's complexity, so it’s tough to use for ordinary people.

[6] Blind or visually challenged people are special people who show their talent in different fields. Online platforms such as YouTube are sources to share the activity or lifestyle of such a particular person. The author analyzes the YouTube content posted by the visually challenged person,

and the designed work supports the visually impaired vlogger on social media networks.

[7] Machine learning is used to predict the outcomes of different applications. Social media consumer behavior is predicted using big data machine learning algorithms. Big data is a large volume of data. The author analyzes big data, such as YouTube data, to predict social media consumer behavior. The sellers need to know the behavior of the consumer. In this study, data is collected from different social media networks such as Facebook, Twitter, YouTube, Instagram, Snapchat, and the collected data is analyzed using a different machine learning algorithm. Finally, it predicts consumer behavior.

[8] The customer grouping concept is essential for E-commerce enterprises or other companies to improve their revenue. Customer grouping is the segment of users having similar interests or ideas about a product. Customer segmentation may lead to a better understanding of client preferences, needs, and wants. Organizations can better engage with customers, audiences, or users based on the customer cluster information. The author analyzes the YouTube channel data through the non-negative matrix factorization to segment the user into behavioral and Demographic clusters. Based on the result, the organization can reach its targeted customer.

[9] YouTube comments are used by most researchers in different applications. A personality disorder is a mental health-related disease. The researcher analyzed 1197 comments on the personality disorder video, which are posted by online networks such as YouTube, and the algorithm used for the data analysis is thematic. The study results are to improve the mentally affected person and get the mental health and the treatment procedure. But the source of the work is limited to medical websites and patient experience videos.

[11] On YouTube, any person can create their channel and upload videos. The viewers can comment on such personal thoughts, and positive or negative ideas about the content of the visited videos. The challenge phase is to classify the comments. The author designed work to classify the user comments into positive and negative.

3. PROBLEM STATEMENT AND RESEARCH QUESTION

With growing platforms such as YouTube that allow content producers to communicate with a large audience, digital content creation has been a big shift in the digital world. However, as the volume of data and user involvement grows, content creators find it difficult to fine-tune their content strategy, engagement techniques, and overall channel success. Traditional analytics methodologies are insufficient for managing the large amounts of data created by YouTube channels, which are constantly changing. As a result, there is an urgent need to develop a complete and effective support system based on big data analytics. The goal of this system is to deliver helpful information to content creators for them to advertise their YouTube channels, make decisions, and increase the channel's efficiency.

Research Question

"How can a big data-driven support system be designed and implemented to effectively enhance YouTube channel performance and provide content creators with actionable insights for improvement?"

This research question will serve as the foundation for investigating the design, development, and deployment of a support system that uses big data analytics to address the challenges faced by content creators in optimizing their YouTube channels. By focusing on this question, the study aims to contribute to the advancement of knowledge in the realm of digital content creation, data analytics, and platform optimization.

4. APPLICATIONS OF YOUTUBE DATA

YouTube is a popular video streaming platform. This is one of the primary sources of big data. During the COVID-19 outbreak, many of the people started their YouTube channel to upload their videos after filming them with a different application. Many application areas such as Tourism, Political, Education, Medical, E-commerce, etc. Table 1 below explains the other authors' work on various applications using YouTube channel data. The comments on the videos play an important role in sentiment analysis to explore positive and negative thoughts. Most of the researchers spent their time analyzing YouTube

data such as channel statistics and the comments of the videos, which were collected from YouTube for text analysis.

Table 1: Application Of YouTube Data

S.NO	AUTHORS	FINDINGS	APPLICATION AREA	PUBLISHED YEAR
1	Tolkach, Denis, and Stephen Pratt.	The authors explained the importance of the youtube channel during the pandemic situation.	Tourism	2021
2	Chen, Yingying, and Luping Wang	Predicted the false political advertising on YouTube	Political	2022
3	Kopecký, Kamil, et al	Analyzed what kind of content is most popular among teenagers.	Education	2020
4	Tadbier, Abdul Wadood, and Abdulhadi Shoufan	Ranking the education channels on YouTube	Education	2020
5	Chaudhary,Kiran,et al	Recommendation system based on user interest and the historical data.	E-commerce	2021
6	An, Jisun, et al	Customer segmentation.	E-commerce	2018
7	Kavitha, K. M., Asha Shetty, Bryan Abreo, Adline D'Souza, and Akarsha	Analyzed and classified the user comments on YouTube videos	Sentiment analysis	2020

5. PROPOSED SYSTEM AND SYSTEM ARCHITECTURE

YouTube analytics has a variety of insights that can be used for YouTube analytics. Competitors cannot check other competitors' channel analytics in the YouTube studio. The proposed system aids in the discovery of new information such as channel analytics and video analytics of their own channel and their competitor's channel. The system predicts the most popular channels in terms of uploaded videos, subscriber count, and view count, and it reveals the most popular video of the competitor channel. The implemented system demonstrates how real-time data from the YouTube channel may be collected through the API and mined for all different types of television relevant data analysis on YouTube data. After the data collection part, the descriptive data analysis is done to get critical key features of many numbers of channels. The analyzed data is forwarded to the novel algorithm to predict the post popular channel. From this result, the user can find the best competitor, and the popular channel is further analyzed to reveal the most post popular videos which are mostly viewed by the people. The analyzed data is fed into the report format, such as a graph or chart, to get a pictorial representation of the analyzed data. The report shows that YouTubers get good value to improve their business. The model workflow is explained in Figure 2. The model workflow consists of data collection, data analysis, finding competitor channels, and top video analysis. Data is collected through Application Programming Interface (API): A client is a user who wants to access YouTube data. The user requests the API key, Based on the request, the API provider provides the key. The

generated key is passed through the API gateway to reach the requested client. The user can access the Real Time YouTube data and perform the data analysis.

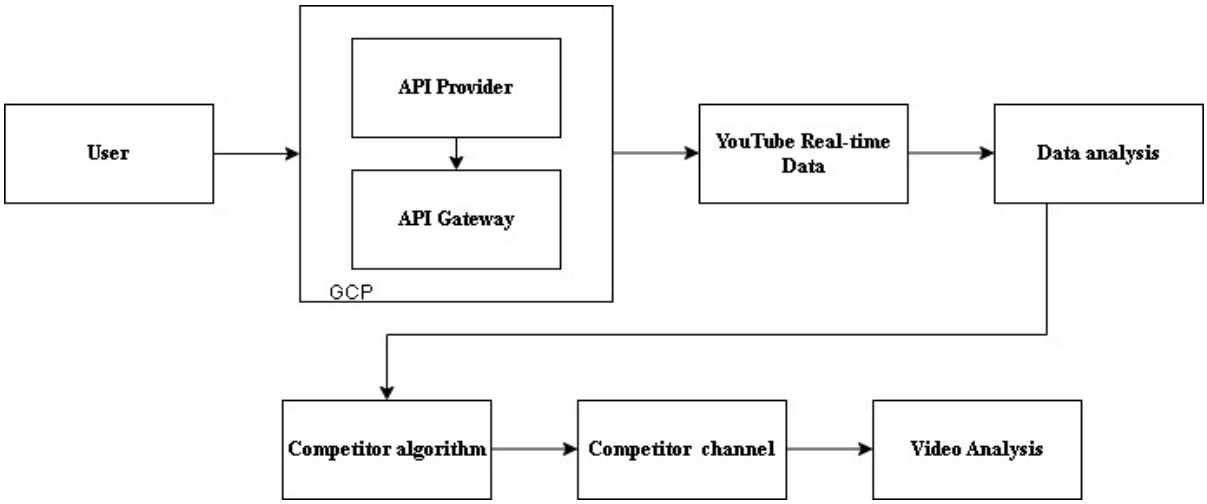

Figure 2: System Architecture

5.1 Data Collection

YouTube is the global most popular and widely used video site, with a massive amount of data regularly updated by the channel owners and viewed by the users. The dataset comprises a list of videos uploaded in India. The dataset is the real-time data set accessed through the API key. A Google account, such as Gmail, is required to utilize the API. Then, go to the Google Cloud platform (https://console.cloud.google.com/) and create a new project. A project name, project ID, and project number are the required fields to create a new project. A client ID, a client secret token is required to use API functions for getting channel details. The dataset contains the channel name, channel id, subscribers, views, likes, and comments. This dataset is used to explore analysis of their own and other channels to improve their business strategies.

5.2 Channel Data analysis

Channel statistics are one of the methods implemented in the proposed work. Channel statistics provide the details of a list of channels which are channel name, subscriber count, number of views, video playlist id, and total number of uploaded videos**.** The standard channel statistics are shown in Figure 3, and Figure 4. The most crucial metric in channel analytics is subscriber count, and viewers count. To improve channel quality, track the number of people subscribed to your channel. In this module, the subscriber count, number of views, and total videos uploaded in the playlist are displayed along with the list of channels. To these statistics, even Zee Tamil uploaded the highest number of videos though the Vijay TV beats with viewers and subscribers.

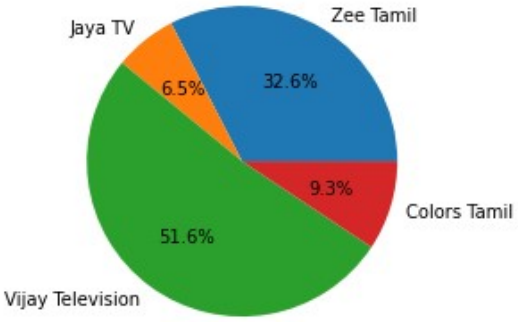

Figure 3: Total No Of Subscribers

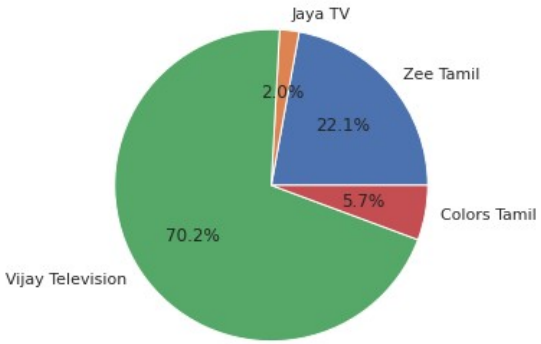

Figure 4: Total No Of Views

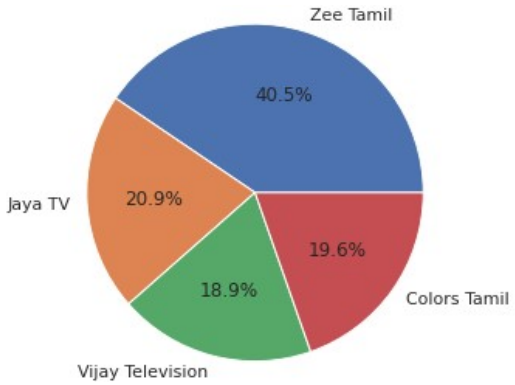

Figure 5: Total No Of Uploaded Videos

5.3 Method for finding competitor channel

In the business world, finding a competitor channel is a challenging task for any user. The proposed work is used to solve this difficult task and provide a list of competitor channels as the output by using an algorithm with different parameters. The parameters are subscriber count, viewers count, and uploaded videos count. The proposed algorithm is tested with N channels and provides good results to the user to improve their channel performance. This algorithm can be executed with different parameters. The general flow diagram for finding competitor channels is explained in Figure 4.

Algorithm for finding competitor channel.

Input: source channel
Output: competitor channel
Begin
channel-list <- name of N channel
subscriber-list <- Subscribers counts of N channel.
Source = First channel from the channel list
Calculate D using equation 1
If D<0
No competitor channel
Else
Add to competitor channel
Display the competitor channel
End

$$D = Sp1 - Cp1 \quad (1)$$

Notes:
D is the difference between source subscriber count and competitor subscriber count
Sp1 = Source subscriber count
Cp1 = Competitor subscriber count
P1 = Parameter1

Parameter = [subscribers, Viewers, Uploaded videos]

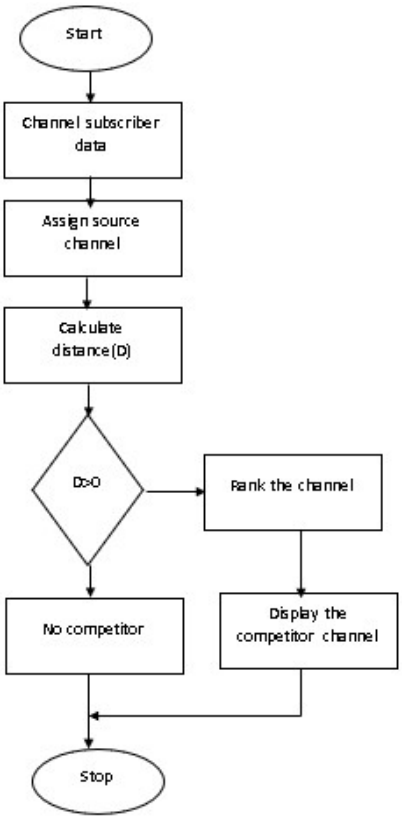

Figure 4: Flowchart For Finding Competitor Channel

5.4 Video Analysis

Video analysis uses YouTube channel data to display the title, published date, and video likes and comments. Like is a qualitative way to judge how

well your video content is being received. The month-wise uploaded videos and top ten videos are listed, are analyzed to improve the business of the YouTuber. They are displayed in Figure 6 and Figure 7. From the results, the YouTubers can gain insights like the month of April has the highest number of uploaded videos.

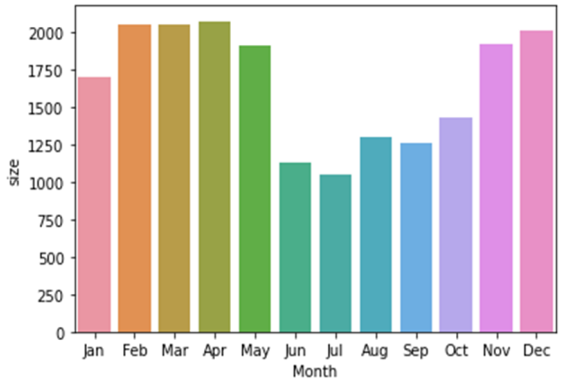

Figure 6: Monthly Uploaded Videos

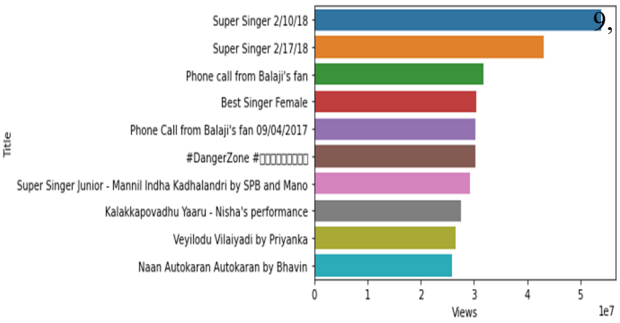

Figure 7: Top 10 Videos

6. IMPLEMENTATION AND RESULT

The Google API key is generated to access the YouTube channel. The Google Cloud Platform (GCP) is one of the most user-convenient cloud platforms. The user receives the requested API key. After the creation of the API key, it is a unique key for users to access the real-time YouTube channel data. In this experiment, Vijay TV, Jaya TV, Zee TV, and Colors TV channel ids are used to perform channel statistics, and the Vijay TV channel id is used for video analytics. Channel Statics and video analytics are performed in the proposed work. Figure 3 and Figure 4 show that Vijay television has the highest number of subscribers and viewers, respectively. Figure 5 shows that the Zee Tamil video has the highest number of uploads. Vijay television channel's monthly uploaded videos and the top 10 trending videos are shown in Figure 6 and Figure 7 using the video analytics model. The experimental channel details are displayed in Figure 8. The competitor of the Color Tamils is Zee Tamil, Jaya TV and Vijay TV displayed in Figure 9, Figure 10, and Figure 11.

	Channel_name	Subscribers	Views	Total_videos	playlist_id
0	Zee Tamil	11100000	6581105316	60233	UU_wlGmvdyAQLtl-U2nHV9rg
1	Jaya TV	2220000	594708683	31113	UUK-VqZSMAUhgkMnZADVBGaA
2	Vijay Television	17600000	20854279673	28058	UUvrhwpnp2DHYQ1CbXby9ypQ
3	Colors Tamil	3180000	1687772564	29168	UUWW46MxidmSoM5WvgGo21IA

Figure 8: Channel statistics

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Channel_name_x	Channel_name_y	vid_diff
Colors Tamil	Zee Tamil	31065.0
Colors Tamil	Jaya TV	1945.0

Figure 9: *Competitor of Colors Tamil with uploaded videos*

Channel_name_x	Channel_name_y	sub_diff
Colors Tamil	Vijay Television	14420000.0
Colors Tamil	Zee Tamil	7920000.0

Figure 10: *Competitor of Colors Tamil with subscriber*

Channel_name_x	Channel_name_y	view_diff
Colors Tamil	Vijay Television	1.916651e+10
Colors Tamil	Zee Tamil	4.893333e+09

Figure 11: *Competitor of Colors Tamil with viewers*

7. CONCLUSION

The proposed approach uses an API key to analyze real-time data from a YouTube channel. The programmer can check their channel in YouTube Studio and cannot be able to match competitors. The proposed work can run its channel analytics and examine competitors' channels. YouTube channel statistics are performed with different parameters. The proposed work analyzes the list of YouTube channel data based on viewer subscriber count, list of media, and the count of visited videos. The determined results are the total number of subscribers, the total number of viewers, the total number of uploaded videos, the top ten trending videos of a specific channel, and monthly uploaded videos. These analytical data can also be represented in demographic form by the system. Individuals and businesses will gain from the system design in terms of revenue, productivity, and profitability.

REFERENCES

[1] Tolkach, Denis, and Stephen Pratt. "Travel Professors: A YouTube channel about tourism education & research." Journal of Hospitality, Leisure, Sport & Tourism Education 28 (2021): 100307.

[2] B.N. Singh, Bhim Singh, Ambrish Chandra, and Kamal Al-Haddad, "Digital Implementation of an Advanced Static VAR Compensator for Voltage Profile Improvement, Power Factor Correction and Balancing of Unbalanced Reactive Loads", Electric Power Energy Research, Vol. 54, No. 2, 2000, pp. 101-111.

[3] Chen, Yingying, and Luping Wang. "Misleading political advertising fuels incivility online: A social network analysis of 2020 US presidential election campaign video comments on YouTube." Computers in Human Behavior (2022): 107202

[4] Kopecký, Kamil, et al. "Behaviour of young Czechs on the digital network with a special focus on YouTube. An analytical study." Children and Youth Services Review 116 (2020): 105191

[5] Tadbier, Abdul Wadood, and Abdulhadi Shoufan. "Ranking educational channels on YouTube: Aspects and issues." Education and Information Technologies 26.3 (2021): 3077-3096.

[6] Seo, Woosuk, and Hyunggu Jung. "Understanding the community of blind or visually impaired vloggers on YouTube." Universal Access in the Information Society 20.1 (2021): 31-44.

[7] Chaudhary, Kiran, et al. "Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics." Journal of Big Data 8.1 (2021): 1-20.

[8] An, Jisun, et al. "Customer segmentation using online platforms: isolating behavioral and demographic segments for personal creation via aggregated user data." Social Network Analysis and Mining 8.1 (2018): 1-19.

[9] King, Clare M., and Darragh McCashin. "Commenting and connecting: A thematic analysis of responses to YouTube vlogs about borderline personality disorder." Internet Interventions (2022): 100540.

[10] Ryoo, Yuhosua, Heeseung Yu, and Eunkyoung Han. "Political YouTube Channel Reputation (PYCR): Development and validation of a multidimensional scale." Telematics and Informatics 61 (2021): 101606.

[11] Kavitha, K. M., et al. "Analysis and classification of user comments on YouTube videos." Procedia Computer Science 177 (2020): 593-598

[12] Oh, Hayoung. "A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model." IEEE Access 9 (2021): 144121-144128.

[13] Vilenchik, Dan. "Simple statistics are sometime too simple: A case study in social media data." IEEE Transactions on Knowledge and Data Engineering 32.2 (2019): 402-408

[14] Shaikh, Farzana, et al. "YouTube data analysis using MapReduce on Hadoop." 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE, 2018.

[15] Subha, K., and N. Bharathi. "Apache Spark based analysis on word count application in Big Data." 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM). Vol. 2. IEEE, 2022

[16] https://developers.google.com/youtube/v3

[17]https://developers.google.com/youtube/analytics

.