iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.43636

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

YouTube Data Analysis and Prediction of Views andCategories

Eliganti Ramalakshmi[1], A Bindhu Sree Reddy[2], Sharvani G[3]
[1]*Assistant Professor Department of Information, Technology, Chaitanya BharathiInstitute of Technology Hyderabad, India*
[2, 3]*Student Department of Information, Technology, Chaitanya BharathiInstitute of Technology Hyderabad, India*

***Abstract:** There's enormous growth and fashionability of YouTube. It has all implicit to move billion of lives encyclopedically as the number of observers is growing day by day. Nearly billions of vids are watched on YouTube every single day, generating a huge quantum of data daily.*
YouTube data is actually in unshaped form, so there's great demand to store the data, process the data and assaying the data. This analysis will help in discovering how people are performing on YouTube, one can fluently identify what contentworks best on YouTube. The primary purpose ofthis design is to find how real time data can be anatomized to get the rearmost analysis and trends in YouTube.
The analysis is done using stoner features similar as views, commentary, markers, likes, and dislikes. Analysis can be performed using algorithms like direct retrogression, bracket and other machine literacy models and python libraries like pandas, matplotlib library to classify the YouTube data and gain useful information.
***Keywords:** YouTube, titles, Prediction ofcategories, Data analysis.*

I. INTRODUCTION

YouTube is the world's most renowned electronic video site, with customers watching 4 billion hoursof video consistently, and moving 72 hours of video reliably (YouTube, 2013). YouTube began in February 2005 and was set up by Chad Hurley, SteveChen, and Jawed Karim who named it"YouTube.com". Through the YouTube stage, people started to make a video-sharing site on which customers could move, deal, and view accounts. From here on out, YouTube has obtained a horde of individuals of billions of customers including educators and scientists. Data Analysis and Mining are becoming fundamental to receive the significant examples in return. Regardless, an enormous pieceof the Data made is generally in Huge Size and comes in unstructured course of action. Enormous

Data can't be examined by ordinary informational collection structures and cycles. To decide this issue,numerous new gadgets that execute ParallelProcessing are being sent in these affiliations. We utilized informational collection from Kaggle of a particular day and performed Data Analysis on the data to comprehension into latest examples and customer responsibility in YouTube concerning Classes. Data Investigation and Visualization was done using Google Collaboratory. Examination of coordinated data has seen colossal achievement previously. In any case, assessment of huge extension unstructured data as video configuration remains a troublesome area.

YouTube, a Google association, has north of a billion customers and produces billions of points of view. Since YouTube data is getting made in an especially colossal aggregate and with a correspondingly extraordinary speed, there is a gigantic interest to store, process and circumspectly focus on this tremendousproportion of data to make it usable.

II. EXISTING SYSTEM

Existing system predicts YouTube views with a certain amount of accuracy. During the analysis of previous papers, we choose certain algorithms which have a greater accuracy rate and also included certain parameters which improves the accuracy and prediction. This gives rise to proposed system.

III. PROPOSED SYSTEM

The proposed system mainly focuses on YouTube views and category prediction. By Considering the tags of the video, we predict the YouTube views using Random Forest Regressor. And also, we predict the categories for the given titles of YouTubevideos. We analyze the data and bring more insights for the content creators so that would be helpful for them.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

IV. IMPLEMENTATION

This Paper "YouTube data analysis & Prediction of views and categories" focuses on analyzing the YouTube data to infer some important information which is useful for content creators and predicting the number of views of a video by considering the YouTube tags & predicting YouTube categories by taking the YouTube title into consideration.

We have used dataset from Kaggle and imported the required packages. The data set contains 82430rows &16columns. Apart from our dataset we have a JSON file which has categories of videos but categories names is not included in taken dataset, so we have extracted and appended an extra column in the data set & given the categories by linking the category id for each row We used the category column to get important information from theYouTube dataset & visualized the given data to showsome important results.

We have analyzed the YouTube data & inferred some of the insights from the data so that would be helpful for new content creators. We have observed that the same video can be on trending various times for consecutive days. We can see the difference inthe number of views, likes, dislikes, comment count as the time progresses. While working with the information we've seen that numerous video pattern on different occasions. We check which channelstrended the most times. We get to know which videos trended most of the times. We have also inferred the top ten videos which are filtered using number of views count so that we can analyze which type of videos is being watched most by the users. We have also extracted the total number of videos belonging to each category so that we can get know which type of videos are being uploaded by content creators in order to grow their channel.

A. Analysis On Channels And Their Trending Videos

Just by checking channels & trending videos count, we can get which channels produce most trending videos. This should be possible utilizing the functions from panda's libraries on information outline containing the number of times the channel was on trending section. Amount of count for each channel is determined & afterward plotted in diagram with the assistance of python's matplotlib.

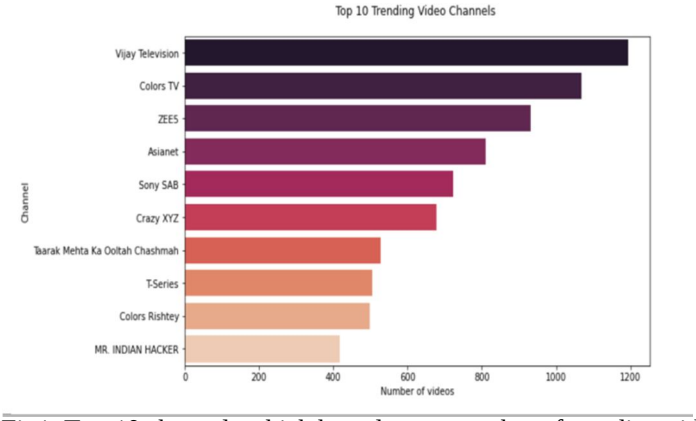


Fig1. Top 10 channels which have largest number of trending videos

B. Analysis On Categories And Their Trending Videos

We can use classification algorithms to classify the categories of videos with their number of trending videos count to understand which category has more trending videos. Shows the channels which have biggest number of trending videos

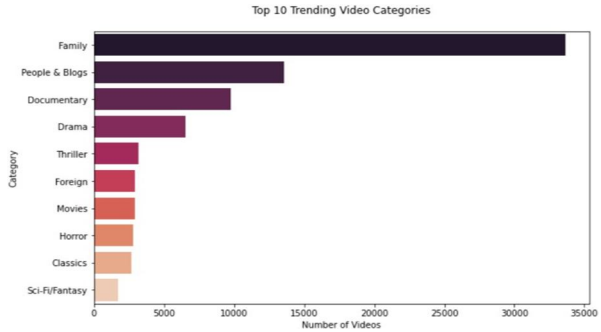


Fig.2 Top 10 categories which have largestnumber of trending videos

C. Analysis On Title Keywords

With the assistance of python and pandas library we can get the most widely recognized words in video titles subsequent to eliminating the standardorganizing words, for example, articles, relational words and consolidating words, for example, a, an, the, yet and so forth

Overlooking words like "the" and "of", we can see that the words Punjabi, tune, Taarak, Mehta showed up additional.

Beneath graph addresses a word cloud to envision most normal words in titles of our Trending Videos. The more normal the words is, will be greater in text dimension. This should be possible utilizing python libraries, for example, matplotlib and word count library.

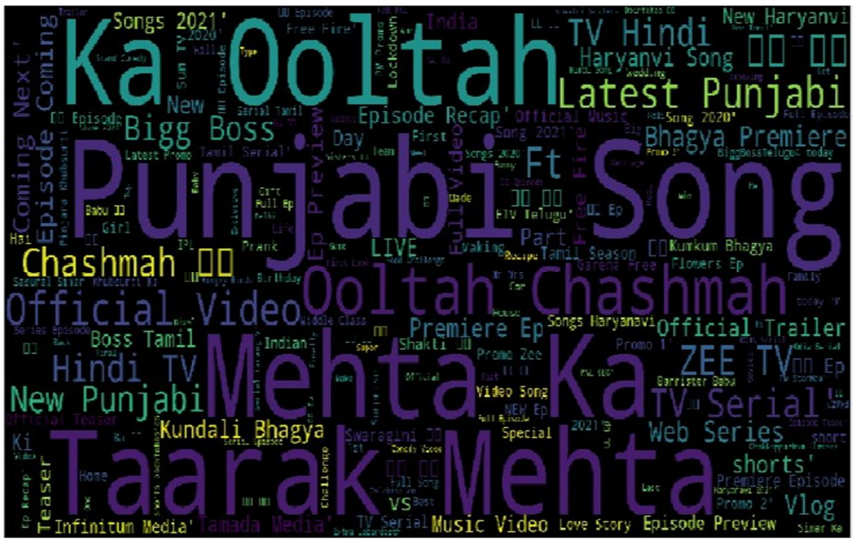


Fig.3 most used keywords in titles of trending videos

D. Analysis On Best Time To Upload Videos

By removing distributing time for each video with recordings count for each time allotment we can see best an ideal opportunity for a moving video to get transferred and distributed. Hence by using algorithm like regression we can predict that a video published in a given time slot having certain set of features can trend ornot.

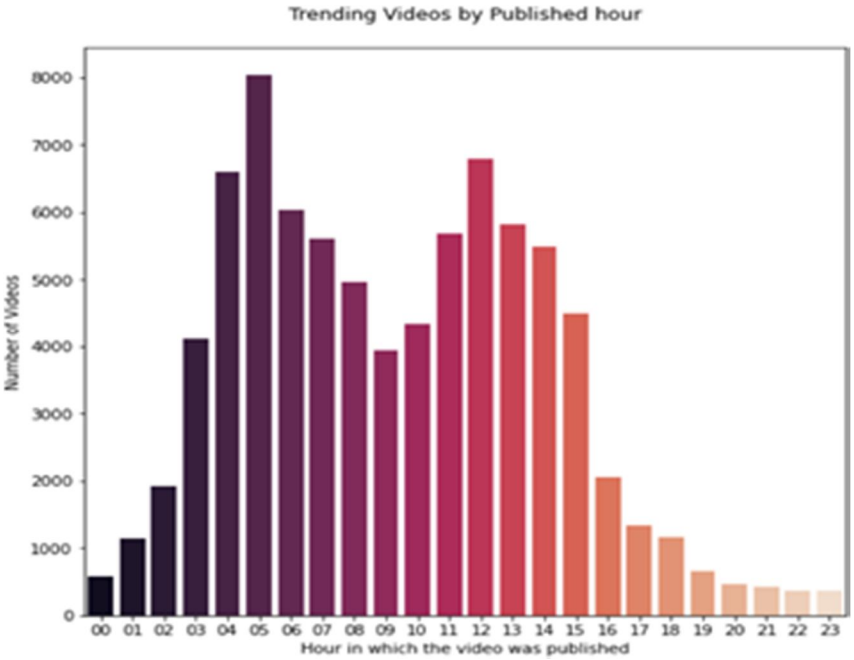


Fig.4 Number of Videos published per hour

E. Comments, Likes And Ratings For Top10 Trending Channels

Correlation represents as value between +1, 0 and - 1 where +1 denotes highly positive correlation, -1 denotes highly negative correlation and 0 denotes Nocorrelation.

F. Analysis On Video Tags

At the point when you transfer a video to YouTube, you can add labels to the video. Labels can't show onthe video page yet you can see labels by review the source code of the page or utilizing programexpansions

Indeed, by utilizing word count library and matplotlib we can make a picture comprising of the best number of words utilized in theselabels.

Fig.5 Common words in Trending Video Tags

G. Correlation Matrix

Using correlation matrix, we can determine the relation between one column and other column i.e., we get the relationship between categoryId, view_count, likes, dislikes, comment_count, comments_disabled, ratings_disabled. We can observe in fig.

Fig.6 correlation matrixREGRESSION PLOT

seaborn.regplot() - This function is used to plot data and a linear regression model fit. There are various totally unrelated choices for assessing the regression.

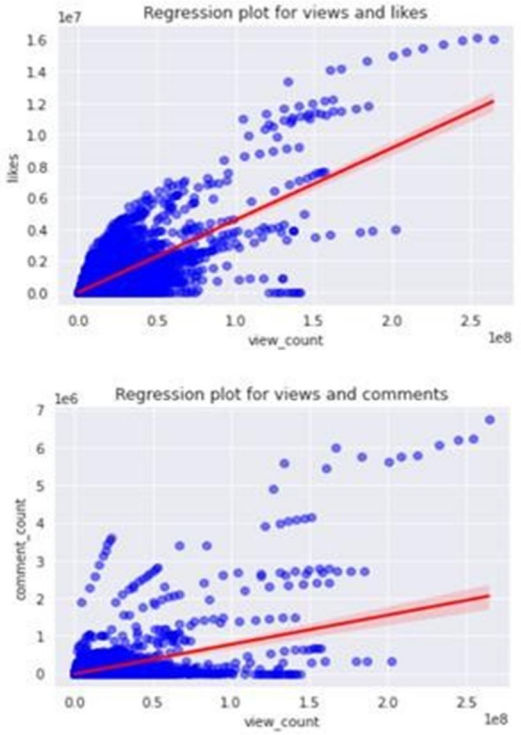


Fig.7 Regression Plot

Guileless Bayes calculation is a managed learning calculation, which depends on Bayes hypothesis and utilized for taking care of order issues.

It is chiefly utilized in text arrangement that incorporates a high-layered preparing dataset.

Naive Bayes Classifier is one of the basic and best Classification calculations which helps in building the quick AI models that can make fast expectations.

We have constructed another naive bayes model to anticipate the classification for a specific title of a video via preparing the model to decide the classification of a video by fitting the counts and target esteems where the counts contain the series of words utilizing count vectorizer and targets is the count of classification id esteems. The Naive Bayes Algorithm scored an exactness of 85% and afterwardwe give a few theoretical titles to foresee for, we embed the titles into naive bayes model and anticipate the categories of those titles.

V. RESULTS

A. Model

We developed the AI model utilizing the Random Forest Regressor by taking the labels of the video foranticipating the you tube sees.

We have pre-handled the labels by eliminating the images and prevent words from the labels so wehave just the fundamental words and can stay away from the un-vital words. we have utilized the Count Vectorizer work which gives a straightforward approach to both tokenize an assortment of text reports and construct a jargon of known words, yet in addition to encode new archives utilizing that jargon.

So first, we instate the "Count Vectorizer" object, which is scikit-learn's sack of words instrument.

Count Vectorizer accompanies its own choices tonaturally do preprocessing, tokenization, and stop word expulsion. fit_transform () completes two capacities: First, it fits the model and learns the jargon; second, it changes our preparation information into highlight vectors. The contribution to fit_transform ought to be a rundown of strings. Furthermore, we changed over the outcome into a NumPy exhibit so the computations would besimple.

Then, at that point, Fit the woodland to the preparation set, involving the sack of words as elements and the perspectives names as the reaction variable and anticipate the quantity of perspectivesby looking at the words in youtube labels.

```
In [29]: output
Out[29]:
```

	video_id	views	views_preds
3633	xGuGjvIfof8	168468	2.624696e+05
2554	b-znn2eQL08	380526	5.836062e+05
6241	pSMCIdcVGgA	16268	3.375678e+04
6486	abeF5zQbQEM	393046	4.145169e+05
1153	zcqZHYo7ONs	994795	6.343440e+05
...	...	...	...
1774	T-IXvsQuogs	233903	4.689145e+05
1757	RkBEEiG-wd4	1735326	1.524076e+06
7235	bvev1Bjx6oA	5438764	5.847633e+06
1237	_pc1kON-doU	4763854	4.328367e+06
1815	dUOUo02v4XE	323563	4.804421e+06

1502 rows × 3 columns

```
In [30]: mse = np.mean((output['views'] - output['views_preds'])**2)
In [31]: print (mse)
2979321354165.125
```

Fig.8 Prediction of views using RandomForest Regressor

	Predicted Category	Hypothetical Video Title
0	Entertainment	Indian Cricketers With HOT Girlfriends & Wives
1	People & Blogs	Sudigaali Sudheer Performance Extra Jabardasth...
2	Sports	Olympics opening ceremony highlights
3	Entertainment	Nagarjuna Funny Comments on Samantha Dress Nag...
4	News & Politics	CNN world news on donald trump
5	News & Politics	Police Chase in Hollywood
6	News & Politics	Elephant hit man in Facebook live
7	Comedy	how to do eyeshadow

Fig.9 Prediction of category using Naïve Bayes

VI. CONCLUSION

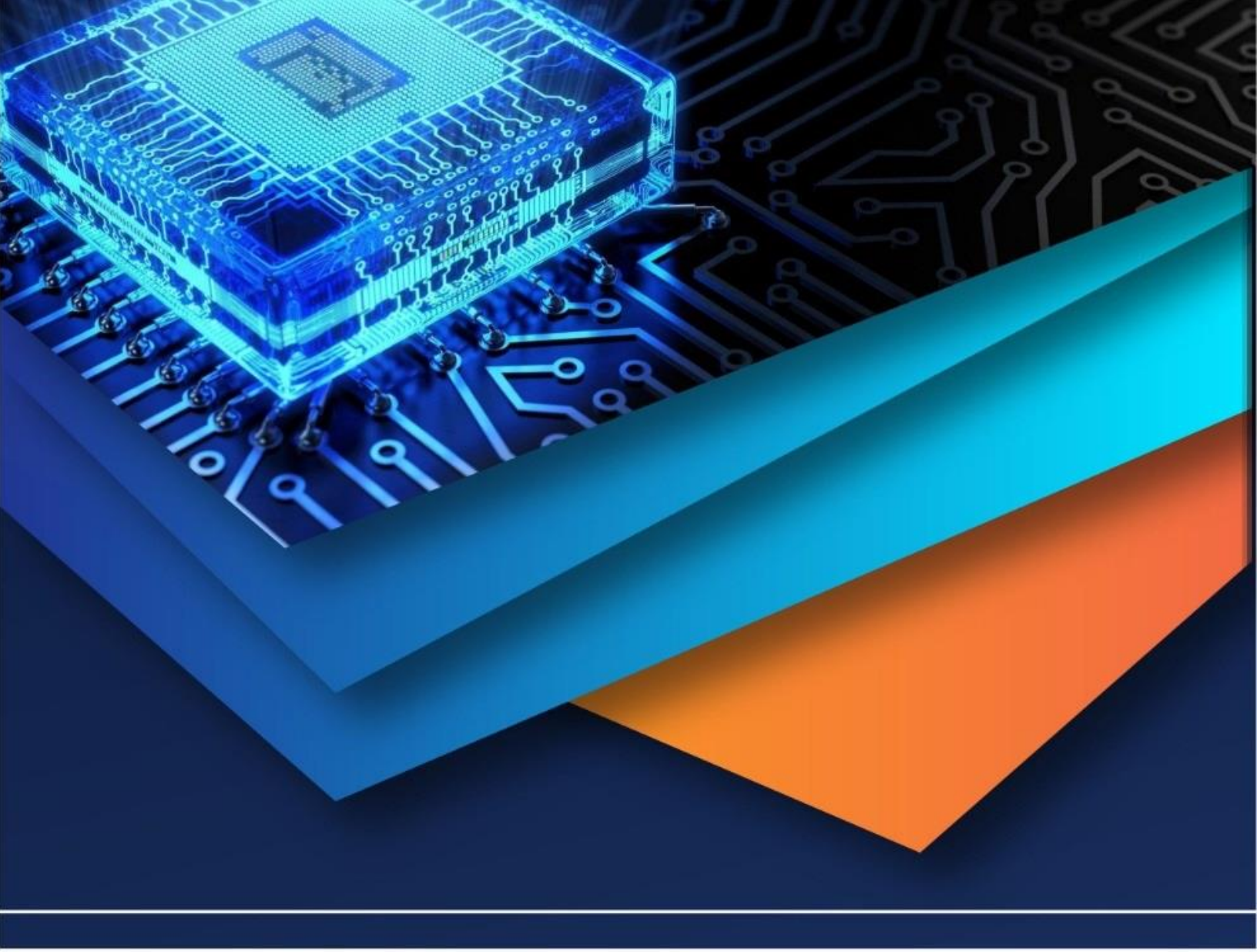
This paper focuses on the analysis of YouTube data and prediction of YouTube views by considering the tags for a particular video and also classifying the category of a video by taking the title of video. Random forest regressor was used for the YouTube views prediction. By analyzing this YouTube data, content creators can get to know the right time at which they can upload the video and also likewise they can get more inferences from the analysis we have made so that would be helpful for them in orderto grow their channel.

VII. FUTURE WORK

we can integrate our model and implement a website so that would be easily accessible to everyone. Recommendation system can be built for giving title for videos based on categories so that the videos can go trending.

REFERENCES

[1] Rui, Lau & Afif, Zehan & Saedudin, Rd & Mustapha, Aida & Razali, Nazim. (2019). A regression approach for prediction of Youtube views. Bulletin of Electrical Engineering and Informatics. 8. 10.11591/eei.v8i4.1630.
[2] Pinto, Henrique & Almeida, Jussara & Gonçalves, Marcos. (2013). Using early view patterns to predict the popularity of YouTube videos. WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining 10.1145/2433396.2433443.
[3] Bärtl M. YouTube channels, uploads and views: A statistical analysis of the past 10 years. Convergence. 2018;24(1):16-32. doi:10.1177/1354856517736979
[4] Kousha, Kayvan & Thelwall, Mike & Abdoli, Mahshid. (2012). The role of online videos in research communication: A content analysis of YouTube videos cited in academic publications. Journal of the American Society for Information Science and Technology. 63. 1710-1727. 10.1002/asi.22717.
[5] Borghol, Y, Ardon, S, Carlsson, N. (2012) The Untold Story of the Clones: Content-Agnostic Factors that Impact Youtube Video Popularity. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012, pp 1186–1194. New York: ACM.
[6] Che, X, Ip, B, Lin, L (2015) A Survey of Current YouTube Video Characteristics. IEEE Multimedia22(2): 56–63.
[7] Cheng, X (2013) Understanding the characteristics of internet short video sharing: a youtube-based measurement study. Transactions on Multimedia 15(3): 1184–1194.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY