IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007019|

Predicting the Fame of YouTube Videos

K. Srikanth [1], Aditi Srishyla Desikachar [2], Akhila Akkunuri [3], K. Sai Prasad [4], Yandapally Harika[5]

Assistant Professor, Department of Computer Science and Engineering, JB Institute of Engineering & Technology, Yenkapally, Moinabad Mandal, Telangana, India[1]

B.Tech Student, Department of Computer Science and Engineering, JB Institute of Engineering & Technology, Yenkapally, Moinabad Mandal, Telangana, India[2,3,4,5]

ABSTRACT*:* YouTube, a video sharing website, is employed in our project to research and anticipate the popularity of a specific video. Being a YouTuber has evolved as a new type of employment in recent decades, as YouTube has become one of the most popular video-sharing platforms. YouTubers make money via advertising on their videos, company sponsorships, gear sales, and donations from their fans.

Consuming and watching videos in YouTube is an integral part of our daily lives. Predicting videos' popularity is of great importance for many services and applications ranging from support design and evaluation, include advertising to earn money. As a result, a YouTuber's prime priority is video popularity.

Two datasets are investigated, each of which is subjected to multiple processing and pre-processing functions. New features such as rateOfViews and age are calculated then a single data frame is generated along with these new features. Other features such as likes, categories, title, Tags, comment_count and more features are also available. Linear Support Vector, a feature selection algorithm is utilised to identify the most relevant features, and machine learning algorithms such as Logistic Regression and K-Nearest Neighbours are employed to forecast the success of a YouTube video. As a result, the algorithm's cost is reduced.

KEYWORDS: Video Popularity, Linear Support Vector, Machine Learning, Logistic Regression, K-Nearest Neighbours.

I. INTRODUCTION

We are frequently urged to comprehend the fundamental notion of popularity growth on the internet, given the vast online global access to data and the simplicity with which inline material may be produced. It is critical for a wide range of services, including building an effective caching model, viral marketing techniques, cost estimation, ad campaigns, and general content optimization. Our method entails using YouTube as a case study and forecasting the popularity of a specific video, or, in other words, the occurrence of a video on the trending page. "Trending Videos" are videos that have become popular as a result of being embedded on some of the web's most popular websites and a large number of people watching the video outside of youtube.com.

YouTube is one of the top three most popular social media networks in terms of active users as of January 2021. YouTube has become a platform for firms to advertise their products due to its vas user base. Being a YouTuber has also become a career in recent times. As a result, a YouTuber's key objective is video popularity in order to sustain a consistent revenue.

II. RELATED WORK

Several attempts have been made to analyse the popularity of internet material. Previous research on popularity prediction includes Cha et Al's [1] study of the YouTube video popularity cycle. According to the statistics, the most popular videos are those that were uploaded recently, but on average, roughly 80% of the videos watched on a given day are older than a month.

Figueiredo et al. [2] investigated the impact of different types of referrers on the popularity progression patterns of YouTube videos. Rodrigues et al discovered that including copies of the same videos resulted in varied popularities.

Yin et al. [3] suggested a model based on user feedback. This can be seen in the form of positive and negative responses, which are represented as likes and dislikes, respectively. Users can demonstrate two personas, according to

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007019|

empirical studies. These personalities are Conformers and Mavericks. Conformists voted for the majority, while mavericks voted against it. It was also discovered that each user possesses both personas to varying degrees. One drawback of this technique is that it necessitates complete information of users' votes in order to construct user profiles. Lee et al. [4] devised a methodology to predict if a discussion forum thread will remain popular in the future. The employed survival analysis approaches influenced by biology, which take into account specific "risk variables". The overall number of comments and the period between the thread's establishment and the first comment are two examples of such criteria. Their model is based on the premise that there are no measures of actual popularity, which may be too restricted for popularity prediction in such systems. This is unrealistic on YouTube in particular, because the system makes such information publicly available.

Szabo and Huberman [5], in contrast, looked at the popularity of YouTube videos and Digg stories, finding that the long-term popularity of online material is substantially connected with its early popularity on a logarithmic scale. They suggested a fairy basic linear popularity prediction model based on that fact. It yielded encouraging results in a variety of applications.

III. METHODOLOGY

The problem of popularity prediction is portrayed as a regression task in this paper. To predict the popularity of YouTube videos, we are implementing Logistic Regression and K-Nearest Neighbours machine learning algorithms. The collected dataset is subjected to multiple user-defined pre-processing and processing functions for the removal of null values and missing values and for the purpose of data formatting. Linear Support Vector Classifier is applied to select important features. After splitting the dataset and training the classifier, the accuracy of the model is tested. We considered three cases with regard to both classifiers:

Case 1: Before applying Linear SVC.

Case 2: After applying Linear SVC.

Case 3: Removing selected few features.

Logistic Regression

- We fit an "S" shaped logistic function (Sigmoid function) in logistic regression to predict two maximum values (0 or 1).
- It can convert any real-valued number to a number between 0 and 1.
- It uses the following formula:

$$\mathbf{Y=1/(1+e^{-x})}$$

Where,

e is Euler's constant = 1.27

X is any real-valued function.

K-Nearest Neighbours

- The K-Nearest Neighbours (K-NN) algorithm assumes that the new case/data and existing cases are comparable, and it places the new case in the category that is most similar to the existing categories.
- The real straight-line distance between two locations in Euclidean space is measured by the Euclidean distance and is given as follows:

$$\mathbf{d(p,q)} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{q_i - p_i})^2}$$

Where,

p, q are two points in the Euclidean n-space.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007019|

$p_i, q_i$are Euclidean vectors, starting from the initial point.

N is the n-space.

Existing System

- A feature was created to determine whether or not the video was popular.
- The popularity function was used in conjunction with logistic regression to predict popularity.
- Uploader, age, category, duration, number of views, rate of video, rating of video, and number of comments are some of the video factors taken into consideration.

Disadvantages

- Rate and rating criteria are no longer available on YouTube.
- It only has a few features.
- The present system may not be compatible with the most recent version of YouTube.

Proposed System

- Using the Logistic Regression algorithm, this project aims to predict the performance of the videos.
- To compare the accuracy outcomes of the logistic regression, the K-Nearest Neighbours Algorithm is employed.
- User-defined functions are used to calculate new characteristics such as RateofViews, age, and RateofLikes.

Advantages

- We've built our model to include the additional data that YouTube currently provides, such as likes, dislikes, views, comments, published date, category ID, video ID, and so on.
- For the training of our model, a significant number of parameters were examined.
- It can efficiently split the data into popular and non-popular locations.

System Architecture

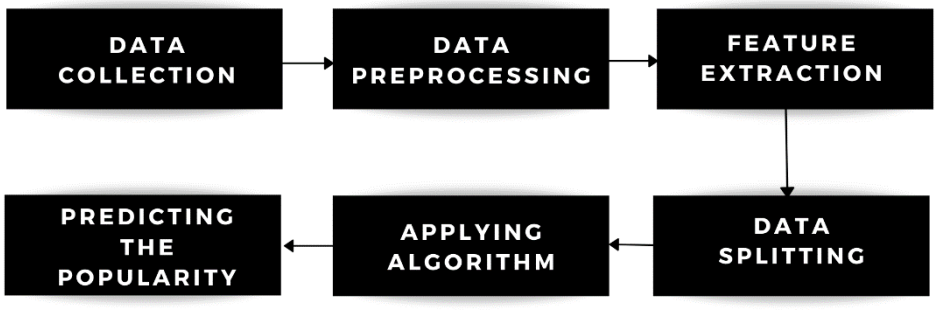

Fig. 3.1 System Architecture

There are four system modules:

1. **Data Collection:**
   The dataset includes information on YouTube videos such as id, title, published date, trending date, likes, dislikes, comments, description, tags, and the uploader's channel name, among other things. Understanding the data to analyse patterns and trends that aid in prediction and evaluation are all part of this subject.
2. **Data Pre-processing:**
   The purpose of data pre-processing is to eliminate any superfluous values. If the dataset contains missing values, one must impute those values to obtain more accurate findings.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007019|

All of the acquired data, both trending and non-trending, is transformed into a comparable format to generate a single file including all of the data.

3. **Feature Extraction:**
   The structured data is transformed into a form that may be utilised to train the model during feature extraction. Two or more features are combined to produce new features, and essential features are chosen using Linear SVC with a penalty value of 0.01. By assessing its significance, Linear SVC determines the most important characteristics from a set of provided features.
4. **Data Splitting:**
   The data is split into training and testing data where, 70% of the data is considered as the training data and 30% of the data is considered as the testing data.
5. **Applying the algorithm:**
   Logistic Regression and K-Nearest Neighbours algorithms are implemented.

IV. EXPERIMENTAL RESULTS

Experimental results are shown in the below figures. Figure 1 depicts the features used for training and testing the logistic regression model, along with the achieved accuracy and classification reports. Figures 2 to 4 depict the ROC curves for each of the three cases, i.e., before applying linear SVC for feature selection, after applying linear SVC for feature selection and dropping a few features manually, respectively. Similarly, figures 5 to 7 show the results for the three cases using K-Nearest Neighbours algorithm, respectively. Figures 8 and 9 are the observed results based on our logistic regression model. The relation between trending videos and their age and rate of likes can be perceived visually with the help of these graphs.

```
Features Used Are:
 ['title', 'channel_title', 'tags', 'views', 'likes', 'dislikes', 'comment_count', 'trending', 'age',
'rateOfViews', 'rateOfLikes', 'category_Autos & Vehicles', 'category_Comedy', 'category_Education',
'category_Entertainment', 'category_Film & Animation', 'category_Gaming', 'category_Howto & Style',
'category_Music', 'category_News & Politics', 'category_Nonprofits & Activism', 'category_People & Bl
ogs', 'category_Pets & Animals', 'category_Science & Technology', 'category_Shows', 'category_Sport
s', 'category_Travel & Events']
Total: 27

Achieved Accuracy is: 84.46620959843291

Classfication Report:
              precision    recall  f1-score   support

           0       0.80      0.91      0.85      2531
           1       0.90      0.78      0.84      2574

    accuracy                           0.84      5105
   macro avg       0.85      0.85      0.84      5105
weighted avg       0.85      0.84      0.84      5105
```

Fig. 1. Accuracy of Logistic Regression before applying LSVC

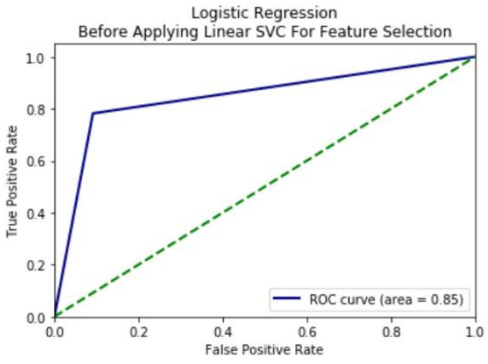

Fig. 2. ROC curve for Logistic Regression before applying LSVC

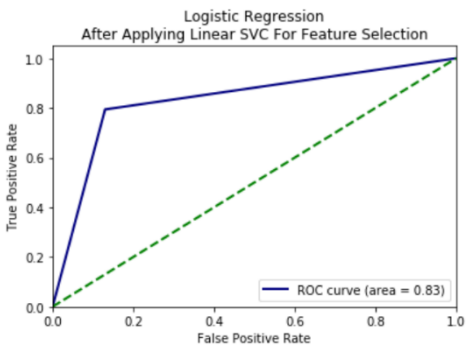

Fig. 3. ROC curve for Logistic Regression after applying LSVC

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007019|

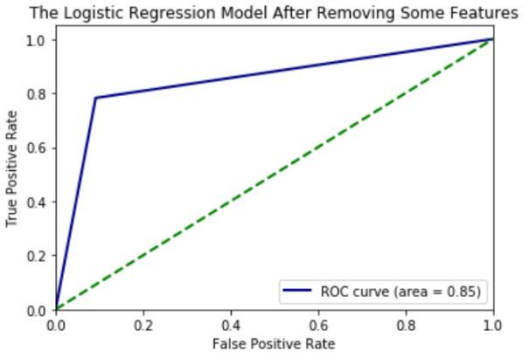

Fig. 4. ROC curve for Logistic Regression after removing few features

```
Features Used Are:
 ['title', 'channel_title', 'tags', 'views', 'likes', 'dislikes', 'comment_count', 'trending', 'age',
'rateOfViews', 'rateOfLikes', 'category_Autos & Vehicles', 'category_Comedy', 'category_Education',
'category_Entertainment', 'category_Film & Animation', 'category_Gaming', 'category_Howto & Style',
'category_Music', 'category_News & Politics', 'category_Nonprofits & Activism', 'category_People & Bl
ogs', 'category_Pets & Animals', 'category_Science & Technology', 'category_Shows', 'category_Sport
s', 'category_Travel & Events']
Total: 27
Value of K          Accuracy
5               90.36238981390792
10              91.00881488736533
15              90.63663075416258
25              90.46033300685602
40              90.73457394711068
50              91.00881488736533
100             90.85210577864838
200             89.83349657198825
500             88.42311459353574
```

Fig. 5. Accuracy of K-Nearest Neighbours before applying LSVC

```
Features Selected Are:
 ['title', 'channel_title', 'tags', 'likes', 'dislikes', 'comment_count', 'age', 'rateOfViews', 'rate
OfLikes', 'category_Autos & Vehicles', 'category_Education', 'category_Entertainment', 'category_Gami
ng', 'category_Howto & Style', 'category_Music', 'category_News & Politics', 'category_Nonprofits & A
ctivism', 'category_People & Blogs', 'category_Pets & Animals', 'category_Science & Technology', 'cat
egory_Sports', 'category_Travel & Events']
Total: 22
Value of K          Accuracy
5               87.09108716944172
Confusion Matrix
 [[2176  387]
 [ 272 2270]]
10              88.28599412340841
Confusion Matrix
 [[2264  303]
 [ 295 2243]]
15              88.54064642507346
Confusion Matrix
 [[2217  353]
 [ 232 2303]]
25              87.83545543584721
Confusion Matrix
 [[2173  371]
 [ 250 2311]]
40              86.01371204701273
Confusion Matrix
 [[2114  445]
```

Fig. 6. Accuracy of KNN after feature selection using LSVC

```
Value of K          Accuracy
5               87.03232125367288
Confusion Matrix
 [[2139  396]
 [ 266 2304]]
10              88.2664054848188
Confusion Matrix
 [[2264  302]
 [ 297 2242]]
15              87.44368266405485
Confusion Matrix
 [[2140  371]
 [ 270 2324]]
25              87.77668952007835
Confusion Matrix
 [[2174  384]
 [ 240 2307]]
40              86.91478942213516
Confusion Matrix
 [[2177  399]
 [ 269 2260]]
50              87.20861900097944
Confusion Matrix
 [[2115  429]
 [ 224 2337]]
100             85.77864838393732
Confusion Matrix
 [[2024  494]
 [ 232 2355]]
200             84.7796278158668
```

Fig.7. Accuracy of KNN after removing few features

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007019|

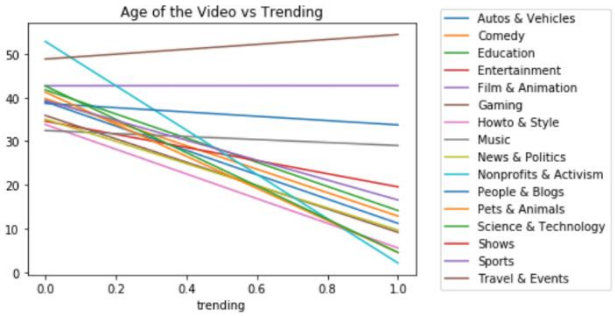

Fig. 8. Age of the Video against Trending

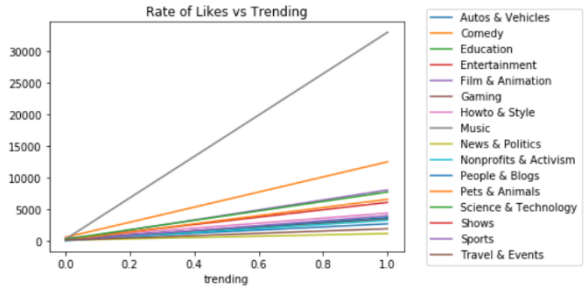

Fig. 9. Rate of Likes against Trending

V. CONCLUSION

Our model uses Logistic Regression analysis to predict the popularity of YouTube videos based on several parameters such as the video ID, trending date, title, channel title, channel ID, published time, tags, views, likes, dislikes, comments, and description of videos popular in Germany, the United States, France, the United Kingdom, and Canada. In our model, the channel title is determined by calculating the frequency with which it appears in the dataset and rating it accordingly. We also utilised the K-NN classifier to compare the accuracy of the Logistic Regression with the accuracy of the K-NN classifier to compare the accuracy of the Logistic Regression with the accuracy of the K-NN classifier. For the Logistic Regression model, the model yields roughly 85% accuracy.

VI. FUTURE ENHANCEMENTS

This model's future work will focus on enhancing the accuracy of Logistic regression. The model may be tested against several other classification algorithms. In our approach, the channel title is determined by calculating the frequency with which it appears in the dataset and rating it appropriately. To increase the accuracy, we can additionally manage channel titles by ranking them depending on the number of subscribers they have. Tags might potentially be managed by gathering Google's daily trend data and then rating the tags. We only evaluated a dataset of a few nations, but it may be expanded to include a global dataset.

Additional video property such as “series of pictures” can also be supplied as input for the features. This feature depicts the video's content in great detail. Because the subtitle may implicitly indicate the length of the video while simultaneously representing the content of the video in depth, certain features such as ‘duration’ may be substituted by other features such as ‘subtitle’.

REFERENCES

[1] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. Y., & Moon, S. (2009). Analysing the video popularity characteristics of large-scale user generated content systems. IEE/ACM Transactions on Networking (TON), 17(5), 1375-1370.
[2] Szabo, G., & Huberman B. A. (2010). Predicting the popularity of online content. Communications of the ACM, 53(8), 80-88.
[3] Yin, P., Wang, M., & Lee, W.-C. (2012). A straw shows which way the wind blows: Ranking potentially popular items from early votes.
[4] Crane, R., & Sornette, D. (2008). Viral, quality, and junk videos on YouTube: Separating content from noise in an information-rich environment. In AAAI Spring Symposium: Social Information Processing, 18-20.
[5] A Lifetime Aware Regression Model for Predicting YouTube Video Popularity by Changsha, Zhisheng and Chang CIKM ‘17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, November 2017.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details