electronics

MDPI

Article

Optimizing Prediction of YouTube Video Popularity Using XGBoost

Meher UN Nisa [1], Danish Mahmood [1,*], Ghufran Ahmed [2], Suleman Khan [3,4], Mazin Abed Mohammed [5] and Robertas Damaševičius [6,*]

1 Department of Computer Science, SZABIST Islamabad, Islamabad 44001, Pakistan; da.meher98@gmail.com
2 School of Computing, National University of Computer and Emerging Sciences (FAST-NUCES), Karachi 75030, Pakistan; ghufran.ahmed@nu.edu.pk
3 Department of Computer and information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK; skhan92@uclan.ac.uk
4 School of Psychology and Computer Science, University of Central Lancashire, Preston PR1 2HE, UK
5 Information Systems Department, College of Computer Science and Information Technology, University of Anbar, Ramadi 31001, Iraq; mazinalshujeary@uoanbar.edu.iq
6 Faculty of Applied Mathematics, Silesian University of Technology, 44-100 Gliwice, Poland
* Correspondence: dr.danish@szabist-isb.edu.pk (D.M.); robertas.damasevicius@polsl.pl (R.D.)

Abstract: YouTube is a source of income for many people, and therefore a video's popularity ultimately becomes the top priority for sustaining a steady income, provided that the popularity of videos remains the highest. Analysts and researchers use different algorithms and models to predict the maximum viewership of popular videos. This study predicts the popularity of such videos using the XGBoost model, considering features selection, fusion, min-max normalization and some precision parameters such as gamma, eta, learning_rate etc. The XGBoost gives 86% accuracy and 64% precision. Moreover, the Tuned XGboost also shows enhanced accuracy and precision. We have also analyzed the classification of unpopular videos for a comparison with our results. Finally, cross-validation methods are also used to evaluate certain combination of parameter's values to validate our claims. Based on the obtained results, it can be said that our proposed models and techniques are very useful and can precisely and accurately predict the popularity of YouTube videos.

Keywords: YouTube videos; feature fusion; video popularity prediction; social networks

Citation: Nisa, M.U.; Mahmood, D.; Ahmed, G.; Khan, S.; Mohammed, M.A.; Damaševičius, R. Optimizing Prediction of YouTube Video Popularity Using XGBoost. *Electronics* **2021**, *10*, 2962. https://doi.org/10.3390/electronics10232962

Academic Editor: Amir Mosavi

Received: 27 October 2021
Accepted: 25 November 2021
Published: 28 November 2021

1. Introduction

Advancements in web technologies have revolutionized the world. Today, there are numerous platforms from which anybody can share his/her content. If that content is appealing to massive, certain rewards can be achieved in return [1]. Social media platforms are playing their part in orchestrating mindsets and educating people about their basic rights (that is, freedom of speech). Through the conventional web, a channel is a possible source for users to reach large audiences. Some services made it possible to share content between producers and consumers around the world. While content is posted by the producers, it is rated (liked or disliked) and discussed by the users. Social networks are becoming a wide source of information for people through the use of internet and with the help of the internet. Users share data through different platforms such as YouTube, Facebook, Instagram, Twitter, etc. Among such platforms, YouTube has become the largest broadcasting source of online videos, wherein anyone can share views or content on social networking sites to improve customer experiences [1].

The motivation behind this study was the notion that certain financial benefits are afforded to YouTube video producers as their viewership increases [2,3]. That is the reason why today, being a YouTuber can involve transitioning from money-making hobby into a proper profession. Numerous YouTuber videos have gone viral, resulting in an exponential increase in fame and wealth for YouTube produces. As a result, many artists

(mostly actors and models) have started their careers through their YouTube channels. Therefore, the importance of recognizing the popularity of online videos and forecasting the success of videos or vice versa is an undeniable fact [4–6]. Each social media user has not contributed equally to the generation of these data. It is of interest to website owners and business analysts to predict which content will be popular. For business and other purposes, YouTube is the most popular and largest video sharing platform. Content distribution and selection are a great source for attracting users to channels [7].

Table 1 contains all the abbreviations used in this paper. The study of companies that host social networks and their users also requires identifying which new platforms may become popular in the near future. Knowing the dynamics of video popularity helps to understand what makes one platform more popular than another. The basics of the business model of YouTube, self-marketing, is also linked with the click-through of related advertisements [8]. Furthermore, popularity prediction can be helpful in proactively allocating resources based on content popularity [9]. However, viral videos on the internet have a remarkable effect on society such as politics and online marketing, as already discussed above. YouTube is a common way to advertise one's services or products. Among the huge variety of uploaded videos, some go viral and grab the attention of millions of viewers overnight. On the other hand, there may be many meaningful videos that never get such a response. The difference in response is not fully understood, as many different factors such as demographics, time of posting, the use of colors, etc. [10] may affect it.

Table 1. List of Abbreviations.

Abbreviation	Meaning
KNN	K-Nearest Neighbor
SVM	Support Vector Machine
IMDB	Internet Movie Database
NLP	Natural Language Processing
MCTCPP-CP	Matrix completion technology based on content popularity prediction
CNN	Convolutional Neural Network
LARM	Lazy Associative Rule Mining
NN	Neural Network
SVR	Support Vector Regression
MRSE	Mean Root Square Error
MAE	Mean Absolute Error
HD	High Definition
SD	Standard Definition
nlikes	New Likes
olikes	Old Likes

Hence, predicting which video will be popular with the masses is a major question. YouTube's internal mechanism and growth pattern are very important in attracting users to videos. When it comes to the analysis of videos, these also evolve in terms of prediction. The deep social effect of viral videos is that they also attract the attention of representatives from various industries and academic researchers. Prediction of videos is difficult, especially early on; although its popularity varies, there are more chances that a video will become viral if the title it contains is popular on other media as well. It is also observed that viral videos have shorter names and lengths [11]. However, there are still no permanent or known variables that can clearly state the main reasons for the popularity of YouTube videos.

This study performs the prediction of YouTube video popularity by:

- Extraction of the new features and addition of them to the existing data set.
- Fusing the corresponding features to improve model performance and reduce computational time.

- Train the over-designed features of the learning model to predict the popularity of videos.

State-of-the-art works such as [12] carried out popularity prediction based on historic data using video data contains features such as video category, likes, dislikes, comment count, and views. However, to increase quality and accuracy (important for precise prediction) other features must be incorporated (e.g., video definition and video duration). More video categories are required to be used for precise and accurate popularity prediction. Hence, in this work, more features/categories are added. For dimensionality reduction, a fusion technique is also utilized.

The popularity of videos in this study is predicted using the XGBoost model, which includes features selection, fusion, Min-Max normalizing, and accuracy parameters such as gamma, eta, learning rate, and others. The XGBoost has an accuracy of 86% and precision of 64%. Furthermore, the Tuned XGboost has improved accuracy and precision. In addition, as a comparison, we looked at the classification of popular and unpopular videos.

The remainder of the paper is organized as follows. Section 2 provides a review of the literature together with a comprehensive critical analysis of existing state-of-the-art work in the domain. Section 3 describes the proposed model that overcomes the limitations of existing works. Section 4 presents the experimental setup and results. Finally, the conclusions of the study are presented in Section 5.

2. Literature Review

In [12], future prediction of online video popularity is considered. Here, the author predicts the future success of online videos based on various elements and successful classification techniques. The decision tree approach is slightly worse than random forest and is preceded by KNN and SVM. Micro-precision and macro- recall are low only for the SVM clasifier. In this domain, extensive work has been conducted. The authors of [13] proposed a hybrid model to predict the popularity of online content. The view count of videos is considered for prediction. In [14], the author proposed a procedure to predict the notoriety of online content. They used endurance analysis to break down the normal endurance time of any occasion or article. The model predicts the popularity of video in a specific period. The data set from dpreview.com and myspace.com is used. In [15], the authors presented an in-depth examination of IMDb and predicted IMDb scores. The database contains categorical and statistical data such as IMDb score, director, gross revenue, and finances, among other things. Instead of listening to critics, this study provides a method for predicting a film's success before it hits the theaters. The authors in [16] predict the popularity of videos using sentimental analysis. Support vector machine (SVM) and natural language processing (NLP) methods are utilized for prediction. The study [17] successfully implements and compares 11 models and finds that the gradient boost model performs best with 79.7% accuracy. They obtained the highest accuracy of 73% with random forest after testing five classification models, while the study achieved the accuracy of 78.1% using gradient boosting.

In [18] the author forecasts the popularity of the video that is supported by social networks. The social forecast algorithm works well without having prior knowledge of popularity evaluation. Youku, a popular video provider in China, has been shown to outperform a previously used model with a 1.58% reduction in relative prediction errors. The author proposed a model [19] that predicts the popularity of online videos. The multivariate linear regression model based on evolution patterns, bursts, and lifetime prediction is used to predict video popularity using the metadata of videos and counts. In [20], a machine learning approach is used to predict the popularity of the box office. The features used to predict popularity do not include the viewers, although viewers can play a vital role in predicting the movie's success.

The research in [21] analyzes the dynamics of online video popularity, considering the popularity evolution patterns on the linear correlation between early popularity and future popularity. Popularity is analyzed from four different aspects: the overall pop-

ularity distribution, the individual popularity distribution, the pattern of evolution of the popularity, and the relationship between the early and the future popularity. A new approach to early prediction of video popularity based on data available at video upload start time is proposed in [22]. The proposed model integrates the results of naive Bayes, SVM, logistic regression, neural network, and random forest. The results were calculated using all 36 features and show that the ensemble method outperforms compared to the 5 methods. The models were compared with various K options, including the top three videos, the top five videos, the top seven videos, and the top nine videos. The degree of popularity is considered as grouping all the view counts over the same period.

In [23], the authors allowed for external popularity forecasts, provided an approach, and considered user comments instead of using view counts. The data set is taken from YouTube. The author proposed a bipartite user-item ranking algorithm (BUIR) that is used to capture the complementary manners between the user and the objects. The author predicts the popularity of the video [24] supported by social networks. The author presents a systematic and relevant approach to predict the success of social media videos for the first time. Specifically, in terms of prediction incentives, the social prediction algorithm exceeds current studies by more than 30%. The author predicts the popularity of online content using only the title [25]. Popularity is predicting using bidirectional recurrent neural network. The main contribution of the paper is title-based prediction and uses pre-trained word vectors in the embedding layer. [26] proposed a model that uses the number of perspectives to anticipate video popularity. To assess the prevalence of online recordings, propose a vector regression reversion strategy with Gaussian radial basis functions. The author in [27] uses visual sentimental and content characteristics to predict the popularity of the Web content. The prediction models for different popularity trend models by adding visual content for prediction. Both the content and the visual sentiments of the video are used for prediction. Grid search is used to optimize the parameters that are defined in each model.

In [28] delivering and viewing YouTube content is a significant part of our regular daily existence. The motivation for this paper is to utilize AI strategies to conjecture the achievement of YouTube videos. Logistics regression and KNN are utilized in this trial research, their proposed way to deal with prediction is to use KNN. The study [29] aims to close this gap in the literature by identifying three factors that can influence the value of sponsored content and attitudes toward YouTube influencers. A conceptual model based on the advertising value model was created for this purpose. By collecting data from 411 university students who frequently use the Internet, the redesigned model was put to the test using structural Eq. modeling. The researcher addressed [30] this issue for MEC, and a joint cache placement approach based on MCTCPP-CP is proposed. The MCTCPP-CP technique is the first to predict content popularity using matrix completion technology.

The authors of [31] provide a brief analysis of the current site content dimensions and also the correct prediction methods to predict the popularity of the content. The second aspect is the features that are symptomatic according to the publication time of the content. Before publication, the content maker is considered; on time of publication the content itself matters; and after the content is published many factors such as views, comments, and sharing evolve in prediction. The author divides the overall prediction into three categories. For content popularity, prediction falls into two categories. One is early prediction and the other is future prediction. The age of the account, and some content features which are easy to extract such as tags, publication time, and location are also relevant to the content.

In [32], the author treats the popularity of online videos as time series over the specified periods and suggests a new time series model for the prediction of popularity. The researchers in [33] give a complete analysis of the removed YouTube videos. The author analyzed more than 73,000 recent YouTube videos over a week and identified those that were deleted or removed. Using three standard media impacts theories [34] (cultivation theory, social cognitive theory, and social comparison theory), a psychological mechanism through which the frequency and interest in cosmetic tutorials affect young women's

postfeminist beliefs. In South Korea, a two-wave longitudinal survey was conducted to predict the popularity of the content.

Researchers [35] present a deep learning model called visual-social convolutional neural network (VSCNN) that predicts the popularity of a posted image by mixing multiple types of visual and social data into a unified network model, which is motivated by multimodal learning, which incorporates input from many modalities and the current success of CNNs in numerous disciplines. In [36] the work conducted concerns what types of background should be use, whether and how to interpret contexts as a whole, and how to use prediction contexts effectively for FM music, Movie Lense, and Amazon Book. The authors also proposed a model to predict popularity which is based on LSTM. Comprehensive studies with three real-world datasets indicate the efficiency of the proposed model relative to other competitive baselines.

In [37], the goal is to easily predict the long-term success of videos on complex YouTube networks. There are two defined problems that the author resolved. LARM is the first research to exploit content life to account for the inadequacy of historical data without network-based assumptions. LARM divides the videos into many subsets and each of these subsets trains a specific model. In [38] the fuzzy-based approach is proposed to predict the popularity of videos. The video is classified as popular or unpopular based on the score. Support vector regression (SVR) shows even results. In [39] a Bayesian learning approach is used for feature space to make accurate predictions, identify important features, and offer confidence levels with each prediction, which can guide developers for successful mashup development.

In [40], deep fusion is proposed a new predictive architecture that uses deep neural networks to combine cross-platform features obtained from Youku and Douban. In [41], the proposed model uses a context-driven approach to estimate the quality of experience, that is, the number of views on video and the participation of the users. The researcher concentrates on popularity metrics and user interaction. In [42], the popularity of short video networks is predicted using a convolutional neural graph-based video popularity prediction algorithm.

Researcher in [43] using telecom data, a cyber physical social system is used to analyze high communication traffic areas. The suggested model creates a graph and analyses it using social network analysis. Following the hotspot extraction process, social network analysis is carried out, which involves measuring the value of each hotspot using network metrics. These figures help determine the value of each hotspot in a telecom data network.

State-of-the-art works are elaborated in Table 2. These state-of-the-art works carried out popularity prediction based on historical data using video data contain features of video category, likes, dislikes, comment count, and views. However, to increase quality and accuracy (which are important for precise prediction) it includes other features that must be incorporated (i.e., video definition and video duration). More video categories are required to be used for a precise and accurate popularity prediction. Hence, in this work, more features/categories are added, and for dimensionality reduction, the fusion technique is utilized. As in [43], the data set used for prediction is performed on data of long time period and its performance can be improved by shortening the prediction time. In our proposed model, predictions are performed on data of short time period as compared to [43].

Table 2. Summary of the state-of-the-art work on video popularity prediction.

Ref.	Problem	Techniques	Performance Metrics	Data Set	Limitations
[12]	Online content popularity prediction	LSTM.	Accuracy	News articles and news videos.	Image and video features are not considered
[13]	Video Popularity Prediction	KNN	Accuracy	Real time data from YouTube	Emotional analysis is not considered

Table 2. *Cont.*

Ref.	Problem	Techniques	Performance Metrics	Data Set	Limitations
[14]	Predicts Popularity at Upload Time	Ensemble classification models (SVM, Naïve Bayes)	Accuracy	YouTube music videos data set is used	Only Music category video is considered
[22]	Predicts the popularity using video views	Szabo- Huberman (SH) Model, Multivariate Linear (ML) Model, MRBF Model	Accuracy	Mashable dataset is used	Feature Optimization is not considered
[16]	Sentiment analysis is performed to predict the popularity	SVM and NLP	Accuracy	MOSI dataset	Problem can be better resolved as regression problem
[17]	News articles popularity prediction	RF, SVM	Accuracy	UCI dataset	Video lifetime is considered of long period which could be shorten to improve accuracy.
[18]	Online content popularity is predicted (Twitter)	Transfer learning algorithm	Accuracy	YouTube Live data is used	Sequel movie is discussed but not considered for prediction
[19]	Social media videos popularity prediction	cumulative distribution function, ordinary least squares, multivariate linear regression model	Accuracy	Use the crawler to collect the Facebook Graph API3	More classification performance metrics could be used
[20]	Movie popularity prediction	SVM, NN, NLP	Accuracy	IMDb, Rotten Tomatoes, Box Office Mojo and Metacritic	Viewers of movie can play a vital role in movie success; viewers data is not taken into consideration.
[23]	Early popularity prediction of videos	Time series methods, Bipartite User-Item Ranking (BUIR)	Hypothetical Studies, compare three different hypothesis	Youku data set	The comments data is considered but sentimental aspect is not considered
[37]	Long Term video popularity is predicted	LARM	Mean Square Error	Ren Ren raw data is used	More features could be considered for better prediction
[38]	Video popularity is predicted by views	SVR, Gaussian Radial Basis Function.	Spearman correlation	YouTube and Facebook data set	Semantic cues are missing which could be used for better prediction.
[25]	Social media videos popularity prediction	Long-term Recurrent Convolutional Network	Accuracy	Facebook videos data set	Based on the views, the prediction problem can be solved as a regression problem.
[32]	Predicting the most popular videos	View Counts Dynamic Model (VCDM)	RMSE	IPTV VOD	More time concerned attempt could help to improve results
[28]	Predicts the future view count of video	Both classification and regression techniques applied.	RMSE	Youku.	More classification performance metrics could be used
[27]	Predicts the popularity based on both sentimental and content features	MRBF model	F1 score, RMSE	YouTube Data set	The researcher considers the correlated features, fusion technique on those features could be applied.
[24]	Social media content popularity prediction	Social-Forecast algorithm	Accuracy	RenRen data set	More features can be added for better results.

Table 2. *Cont.*

Ref.	Problem	Techniques	Performance Metrics	Data Set	Limitations
[26]	YouTube video popularity prediction	SVR, Gaussian Radial Basis Function.	R-square F-test	YouTube video data set	Co-evolution of the popularity metrics, in time. Effect of standard feeds on long term video popularity is examined.
[32]	Predict popular life cycle of videos	VCDM model	Accuracy	YouTube, and Daum UCC data of two categories Entertainment' and 'Science & Technology' is	External video features are not considered
[33]	"Think before you upload": an in-depth analysis of unavailable videos on YouTube	Recurrent Neural Networks (RNN)	Accuracy	Three real-world YouTube datasets are used to evaluate the model.	Only linked categories are used for prediction, only 10 weeks' data is used
[34]	YouTube makeup tutorials reinforce postfeminist beliefs through social comparison.	Long-term Recurrent Convolutional Network	RMSE	YouTube data is used	User's aspects which relate to the quality of videos should be considered.
[35]	Multimodal Deep Learning Framework for Image Popularity Prediction on social media	CNN	RMSE	YouTube data of different categories is used.	Time gap should be more concisely considered

3. System Model

In this section, feature fusion prediction is discussed, which contains various functional components, such as feature extraction, feature fusion, and prediction of popularity.

There are three steps in the proposed framework to predict video popularity:

- Extract new features and add them to the existing data set.
- Fuse features to improve model performance and reduce computational time.
- Train the XGBoost model to predict video popularity using fused features.

3.1. Data Set

YouTube is one of the most famous websites which maintains and represents videos and saves a record of them as well. YouTube videos have multiple features, as shown in (Table 3). The video popularity is influenced by different factors including video likes, dislikes, comment count, and views. Instinctively, a video with more views is more likely to be popular in the future. The data set used for video popularity prediction is publicly available and two more features are added to the existing data set 'video definition', 'video duration'. To access the mentioned features, the YouTube API V3 is used to extract the mentioned features. The features are extracted and combined to predict the popularity of the video. The features of the data set used are (trending date, title, channel title, category id, publication time, tags, number of views, likes, dislikes, comments, thumbnail link, and description). It contains different features that are linked and impact the popularity of videos.

Table 3. Feature Description.

Features	Description
category_id	Category_id tells that the video belongs to which category
video_definition	Either the video is high definition or standard definition
duration	How long the video is (either 4 min, 10 min, 20 or more)
score	The feature obtained through the fusion of three features
views	The final class on the basis of which we have to predict the video

3.2. Data Pre-Processing

The YouTube dataset is commonly collected through APIs, and every video has the same features that could be considered from the videos. The dataset used for this research contains 15 features in total. The missing instances in the dataset are replaced by the mean and mode preprocessing technique to impute missing data. In testing data, the missing values can be imputed in the same way for the finalized model.

3.3. Feature Selection

Feature selection is important to select the correlated features from a dataset in order to reduce the dimensions, which helps the model to avoid overfitting. Correlated features are selected, which play a key role in predicting the popularity of the video.

3.4. Feature Extraction

Feature extraction is performed to add new features to the existing dataset, which affects popularity prediction. Features are added by using the YouTube API. The API key on Google Developer is generated. The YouTube API V3 is used to extract the features of the video. Each video id is read through loop process from a CSV file. A single object is created for each row to store the extracted features against each id. If a video ID contains no information, it will return 0 as a result.

The features extracted using Google API, video definition either the video is high-definition HD videos provide superior video quality and a more enjoyable viewing experience or standard definition (SD), while standard definition (SD) movies do not offer the same quality as high definition as shown in (Table 4). Label encoding is used to convert the feature video definition into numeric (0.1 for SD and 1 for HD). The second feature extracted by YouTube API is the duration of the video, which is converted in minutes. Both features help to predict popularity with more accuracy. The fusion technique is applied to convert the three correlated features into one feature. This feature is considered as the weighted score in the final data set. The non-linear min_max normalization technique is applied using Python to normalize the feature between 0–100.

Table 4. Feature extraction.

Video_ID	**Extracted Features**		
	v-definition	duration	dimensions
n1WpP7iowLc	HD	3	2d
0dBIkQ4Mz1M	HD	5	2d
5qpjK5DgCt4	SD	10	2d
d380meD0W0M	HD	39	2d

3.5. Feature Fusion

According to the correlation of the features, they have fused accordingly. The following is the list of features used in our prediction model. The three features are fused by using the non-linear Min-Max normalization technique. Equation (1) represents the linear technique of normalization, notation such as 'olikes' old videos likes, 'nlikes' new likes, 'min' min value and 'max' max value.

$$l'' = nlikes - min(olikes) \div max(olikes) \times newmax(nlikes) - newmin(nlikes) + olikes - min(olikes) \quad (1)$$

By applying the linear normalization technique, the overfitting and underfitting is faced to overcome that problem, and the non- linear normalization is applied to normalize the score as shown in Equation (1).

$$l'' = 1/2((nlikes + ndislikes) + 1/(1 + fl) + |nlike - ndislikes| \quad (2)$$

The goal is to predict the l'' presenting the popularity score where we use the new likes and dislikes by applying Equation (2). The γ is set to 0.5 in this case, it represents the weight. The features with respective scores which are fed to the proposed model are described in (Table 5).

Table 5. The results of feature fusion.

Likes	Dislikes	Comments_Count	Score
787,425	43,420	125,882	28.32
127,794	1688	13,030	4.55
146,035	5339	8181	5.2

3.6. Encoding and Train Test Split

One hot encoding is an active technique for converting categorical variables into numeric data (binaries 0 and 1). In this case, 0.1 is assigned to SD and 1 to HD. As in one-hot encoding, one bit can be true at a time, so it is an effective technique for handling data that has two categories. In this study, 80% of the data is allocated for the training and 20% of the testing. The data which are part of the train split is used to teach the model, and later it is tested on unknown data, which is 20% of the whole data, kept for testing. The test data are used to check the model performance and generalization of the model. As discussed earlier, Python has different libraries that are meant to perform different tasks. Similarly, the library divides data into tests and training. The Sklearn library is used to split the data into train and test, which uses random state parameters to instruct the size of the split. In case of a small data set, data are divided into 30, 70 and 20:80 in the case of a large dataset. To generate the same set of train and test data points, the random state used in this study is 42. In training, data instances are 3500 and 1500 instances for the testing data. In the x train, there are 3500 instances in which 5 are independent features, (one of which is dependent).

4. Conceptual Workflow and Experimental Setup

An enormous amount of work is conducted for video popularity prediction. All existing studies predict the estimated popularity of videos at different stages. Therefore, different features are considered to predict a video's fame. We proposed a model that calculates the general fame score. The model makes the prediction of overall popularity simple and vibrant. Videos are analyzed on the basis of the views of videos at different time. The proposed model gives the popularity score, which tells whether the video is going to be viral or not. Two features, video quality and video duration, are extracted using YouTube API. The data set contains a total of 5 features after feature selection and fusion and 5000 records. The missing data in the data set is handled through pre-processing.

Pre-processing imputes missing values and converts character or text data to numeric by label encoding. NLP is taken into account to handle text or string data In the data set, text features are title, tags, and description, which always play a role in the prediction of video popularity. The title, tags, and description are fused to check all possible keywords against each video category. YouTube has predefined categories of video such as ID 10 being always allocated to music. Categories are converted to numeric units using the label encoder. Feature engineering is applied to the data set to select the features that are important for prediction. Likes, dislikes, and comment count features are combined to calculate the fame score. The popularity score is calculated using min-max normalization. The data set is divided into 20:80 ratio of training and testing. Using machine learning algorithms, we obtain different results. Precision parameter is used to assess the results. The experiment was performed using Python on an Anaconda environment. The proposed model is elaborated in Figure 1.

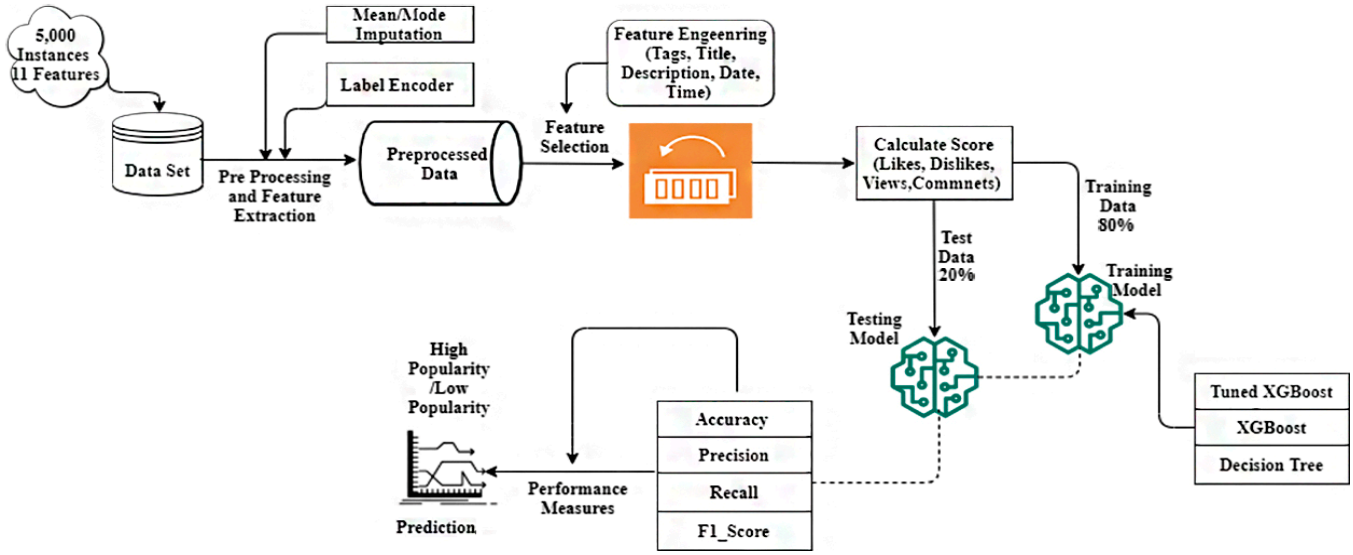

Figure 1. Conceptual workflow optimizing YouTube video popularity.

4.1. Decision Tree

The decision tree results acquired the 80% accuracy, 64% precision, 63% recall, and 63% F1 score. After tuning, accuracy improves from 80% to 83%. The decision tree for prediction is an effective approach that learns from simple decision rules from training data and applies those rules to test data. It is applied to a classification problem to decide whether the video gets popular or not.

4.2. XGBoost

The reason is that XGBoost performs well is that it converts weak learners into strong learners and trains the model well. The well-trained model can perform better in test data. XGBoost gives 86% precision, 84% precision, 63% recall, and 72% of F1_Score.

4.3. Tuned XGBoost

The tuned XGBoost is applied to enhance the results of the XGBoost algorithm. The parameters used for tuning are gamma, eta, learning rate, and n estimators. We adjusted the model's parameters by conducting a 5-fold cross validation method 10 times to evaluate any combination of parameter values to enhance the model's prediction performance. We evaluated 50 and 100 random trees, max depth of trees values (tree complexity) of 2 and 3, learning rate values (the model's resistance to overfitting) of 0.3 and 0.4, and the proportion of observations to create trees of 0.5 and 0.8. We left other model parameters unchanged, such as gamma = 0, the proportion of all predictors to sample for each new tree = 0.8, and the minimum sum of weight for splitting point = 1.

The model acquires an accuracy of 88% after tuning the algorithm.

5. Results

A confusion matrix summarizes the real values to the values predicted by the machine learning model of YouTube videos in (Table 6). This gives us a holistic view of how well the proposed model works as compared to base model [26]. The false negatives are thought to be more tolerant than false positives in predicting YouTube video popularity. On this premise, we claim that the proposed model has performed well than the base model [26].

Table 6. Confusion Matrix of Base Paper.

	Positive	**Negative**
Positive	TP—669	FP—79
Negative	FN—84	TN—168

AS expressed in (Table 5) the model predicts 669 TP and 168 TN out of 917 instances and (Table 6) the model predicts 725 TP and 155 TN out of 1000 instances. Predicting the video popularity, the performance measure that considered is the true positive (TP) rate. YouTube video prediction proposed model performance is better than in the existing study.

The comparison of different performance measures for the popularity prediction is presented in this section in (Table 7). Among all the existing prediction methods, we adapt the XGBoost method to predict the popularity. The comparison of decision tree and XGBoost shows that XGBoost performs well on YouTube data set to predict video fame.

Table 7. Confusion matrix of the classification results.

	Positive	**Negative**
Positive	TP—725	FP—29
Negative	FN—91	TN—155

The classification of the unpopular video has 0.89, 0.96, 0.92, 754 precisions, recall, f1 score, and support, respectively, in (Table 8). Class 1 known as the popular video has 0.84, 0.63, 0.72, 246 precision, recall, f1 score, and support, respectively, with an accuracy of 0.88 and support 1000 for all the classes.

Table 8. Classification Report.

Class	**Precision**	**Recall**	**F1_Score**	**Support**
0	0.89	0.96	0.92	754
1	0.84	0.63	0.72	246
		Accuracy 0.88		

5.1. Precision

The precision is the FP cases of prediction labeled as positive incorrectly shown in Figure 2.

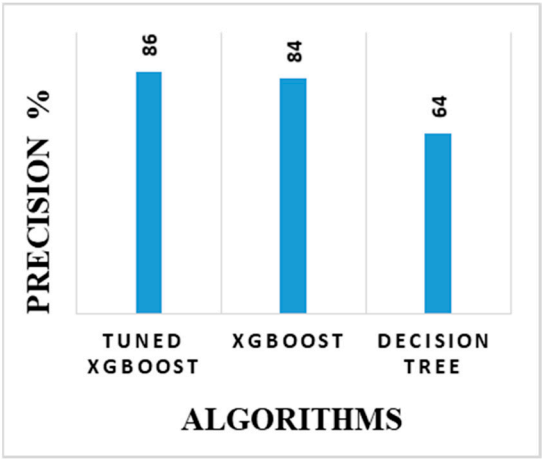

Figure 2. Comparison of precision.

In our case, the model classifies videos as popular that are not actually popular. XGBoost is a boosting technique that has gained a reputation for its speed of execution

and model performance, and is increasingly being used as the default boosting method for predictions rather than decision tree.

5.2. Recall

Recall is the positive rate through the proposed model, the recall acquired through the model is 0.67, which means the model correctly predicts the popularity of videos, that are popular 67% of the time. The percentage of the total relevant results accurately classified by the algorithm is called the recall (represented in Figure 3).

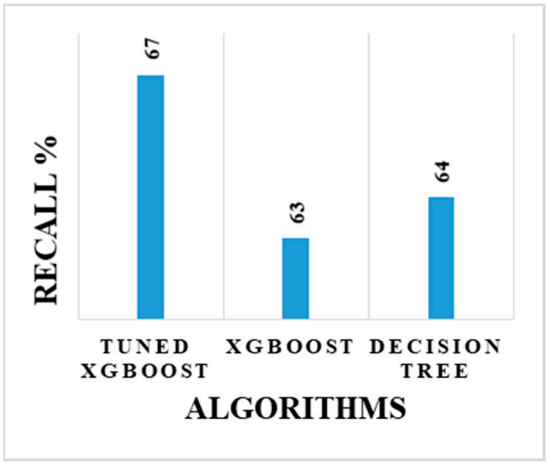

Figure 3. Comparison of recall results.

By looking into the base model or algorithm, we end up with the result that the boosting algorithm XGBoost is good for predictions and is comparable to adaptive boosting (which is implemented in this method).

5.3. F1_Score

In case to check all the actual popular videos, F1 score is considered. F1 score is the harmonic mean of Precision and Recall. The model achieves an F1 score of 65% (as shown in Figure 4).

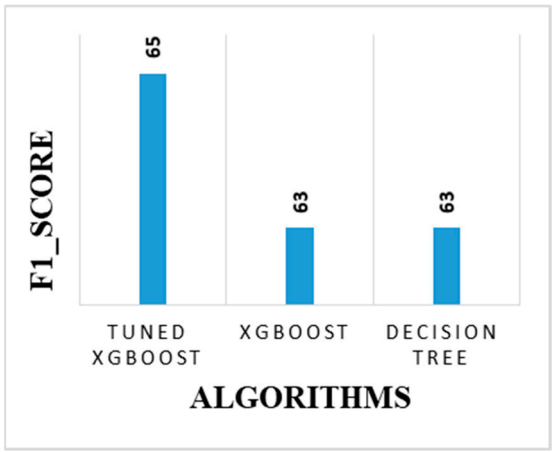

Figure 4. Comparison of F1_Score results.

This study is conducted to compare the old predictors and XGBoost is applied rateher than decision tree with respect to classification errors.

5.4. Accuracy

Greater penalties on the minor class resulted in an increase of overall accuracy as measured by the f1-score since more false positives were present. Accuracy is the percentage of correct predictions for the test data in (Figure 5). Accuracy is the measure of all correctly identified videos. In our case, both classes are equally important (either the video gets viral or not), so computing accuracy is the best measure.

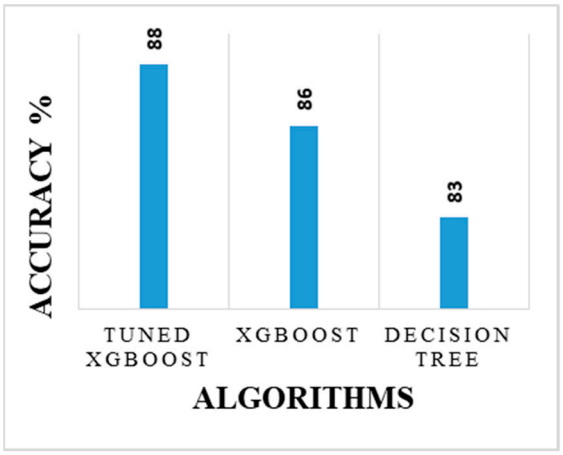

Figure 5. Comparison of accuracy results.

In our case, the number of FP is significant. While, predicting the YouTube videos popularity FP is when in actual video is not going to be viral but the model predicts it as popular. FN is when a video is going to be viral, but the model predicts it as popular. The FP rate is measured in video popularity prediction case, which is why precision is important measuring parameter in video popularity prediction. The FP rate is considered in precision when we are looking for correctly predictions made by the model.

5.5. Computational Time

In the graphical representation, the computational time of the algorithm reduced after feature fusion is shown in (Figure 6). The computational time with three separate features is 35.0 s, and the execution time with fused features (Score) is 27.0 s.

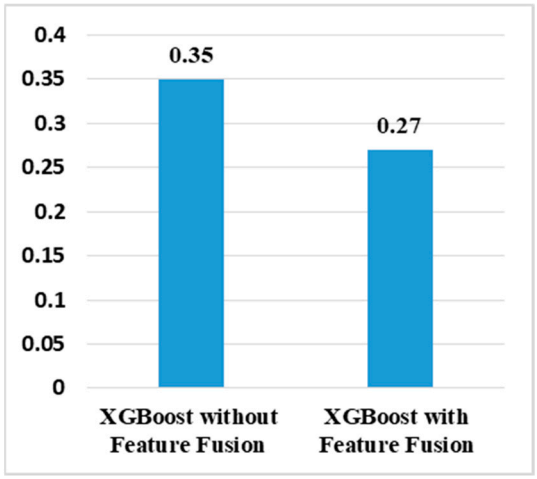

Figure 6. Summary of computational time.

5.6. Discussion

Two different algorithms are applied to the decision tree model and XGBoost. The decision tree (Table 9) shows 80% accuracy, 64% precision, 63% recall, and 63% F1 score.

After tuning, the accuracy improves from 80.0% to 83.0%. The decision tree is used for prediction since it is an effective approach that learns from simple decision rules from training data and applies those rules to test data. Secondly, XGBoost is applied to predict popularity. The XGBoost is tuned by multiple parameters (i.e., gamma, eta, learning rate, and n estimators) to improve the results.

Table 9. Results.

	Decision Tree	XGBoost	Tuned XGBoost
Accuuracy	80%	86%	88%
Precison	64%	84%	84%
Recall	63%	63%	63%
F1_Score	63%	63%	72%

Tuned XGBoost performs well since it converts weak learners into strong learners and trains the model well. The well-trained model can perform better in test data. The tuned XGBoost algorithm acquired 88% accuracy, 84% precision, 63% recall, and 72% F1_Score.

6. Conclusions

Video popularity does not depend only on a few features. Video quality and video duration play a vital role for the video to become viral. The number of views on a video indicates whether a video is popular. Popularity is predicted using the tuned XGBoost algorithm. The decision tree and tuned XGBoost algorithms are compared, and the results show that tuned XGBoost performs better in terms of accuracy.

However, the work discussed in this study provides some limitations and reflects potential future ideas. In future, the work can be extended by considering video sound features; more classifiers can be used for prediction. Moreover, region-wise popularity can be considered, supposing that the researchers can consider in which region the language specified videos got popular. The other factor which can be analyzed is the areas which are content specified. Regarding the capacity of the proposed predictive model, certain approaches are helpful for the researchers to produce accurate and timely rankings. The results show that this capacity can be improved and utilized in specified area of research.

Author Contributions: Conceptualization, M.U.N. and D.M.; methodology, D.M. software, M.U.N.; validation, G.A. and S.K.; formal analysis, M.U.N..; investigation, D.M.; resources, M.A.M. and R.D.; data curation, Mehr Un Nisa; writing—original draft preparation, M.U.N.; writing—review and editing, D.M., S.K. and G.A.; visualization, M.U.N. and D.M.; supervision, D.M.; project administration, G.A. and S.K.; funding acquisition, R.D. and M.A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data is available from the first corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chung, Y.J.; Kim, E. Predicting Consumer Avoidance of Native Advertising on Social Networking Sites: A Survey of Facebook Users. *J. Promot. Manag.* **2020**, *27*, 1–26. [CrossRef]
2. Bielski, A.; Trzcinski, T. Pay attention to virality: Understanding popularity of social media videos with the attention mechanism. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Nashville, TN, USA, 19–25 June 2018; pp. 2398–2400. [CrossRef]
3. Fan, Y.; Yang, B.; Hu, D.; Yuan, X.; Xu, X. Social- and Content-Aware Prediction for Video Content Delivery. *IEEE Access* **2020**, *8*, 29219–29227. [CrossRef]

4. Nguyen, M.; Nakajima, T.; Yoshimi, M.; Thoai, N. Analyzing and predicting the popularity of online contents. In Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services, Munich, Germany, 2–4 December 2019. [CrossRef]
5. Su, Y.; Li, Y.; Bai, X.; Jing, P. Predicting the popularity of micro-videos via a feature-discrimination transductive model. *Multimed. Syst.* **2020**, *26*, 519–534. [CrossRef]
6. Trzcinski, T.; Rokita, P. Predicting Popularity of Online Videos Using Support Vector Regression. *IEEE Trans. Multimed.* **2017**, *19*, 2561–2570. [CrossRef]
7. Vaiciukynaite, E.; Zailskaite-Jakste, L.; Damasevicius, R.; Gatautis, R. Does hedonic content of brand posts affect consumer sociability behaviour on facebook? In Proceedings of the 5th European Conference on Social Media, ECSM, Limerick, Ireland, 21–22 June 2018; pp. 325–331.
8. Liaudanskaitė, G.; Saulytė, G.; Jakutavičius, J.; Vaičiukynaitė, E.; Zailskaitė-Jakštė, L.; Damaševičius, R. Analysis of affective and gender factors in image comprehension of visual advertisement. In *Artificial Intelligence and Algorithms in Intelligent Systems, Proceedings of 7th Computer Science On-line Conference 2018, Zlin, Czech, 25–28 April 2018*; Springer: Cham, Switzerland, 2018; pp. 1–11. [CrossRef]
9. Zailskaite-Jakste, L.; Ostreika, A.; Jakstas, A.; Staneviciene, E.; Damasevicius, R. Brand communication in social media: The use of image colours in popular posts. In Proceedings of the 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2017, Opatija, Croatia, 22–26 May 2017; pp. 1373–1378. [CrossRef]
10. Zailskaite-Jakšte, L.; Damaševičius, R. Gender-related differences in brand-related social media content: An empirical investigation. In Proceedings of the 13th International Computer Engineering Conference: Boundless Smart Societies, ICENCO 2017, Cairo, Egypt, 27–28 December 2017; pp. 118–123. [CrossRef]
11. Jiang, L.; Miao, Y.; Yang, Y.; Lan, Z.; Hauptmann, A.G. Viral video style: A closer look at viral videos on youtube. In Proceedings of the International Conference on Multimedia Retrieval, Glasgow, UK, 1–4 April 2014; pp. 193–200.
12. Zhou, Y.; Wu, Z.; Zhou, Y.; Hu, M.; Yang, C.; Qin, J. Exploring Popularity Predictability of Online Videos with Fourier Transform. *IEEE Access* **2019**, *7*, 41823–41834. [CrossRef]
13. Jeon, H.; Seo, W.; Park, E.; Choi, S. Hybrid machine learning approach for popularity prediction of newly released contents of online video streaming services. *Technol. Forecast. Soc. Chang.* **2020**, *161*, 120303. [CrossRef]
14. Chen, Y.-L.; Chang, C.-L. Early prediction of the future popularity of uploaded videos. *Expert Syst. Appl.* **2019**, *133*, 59–74. [CrossRef]
15. Jog, M.S.; Siras, M.B.; Fender, M.A.; Mandurkar, M.P.; Nikalje, M.Y.; Chhabria, S. Video Popularity Prediction Using Machine Learning. *Int. Res. J. Mod. Eng. Technol. Sci.* **2021**, *3*, 778–783.
16. Gajanayake, G.M.; Sandanayake, T.C. Trending Pattern Identification of YouTube Gaming Channels Using Sentiment Analysis. In Proceedings of the 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 5–6 November 2020; pp. 149–154.
17. Khan, A.; Worah, G.; Kothari, M.; Jadhav, Y.; Nimkar, A.V. News Popularity Prediction with Ensemble Methods of Classification. In Proceedings of the 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bengaluru, India, 10–12 July 2018; pp. 1–6.
18. Bielski, A.; Trzcinski, T. Understanding Multimodal Popularity Prediction of Social Media Videos with Self-Attention. *IEEE Access* **2018**, *6*, 74277–74287. [CrossRef]
19. SU, B.; Wang, Y.; Liu, Y. A new popularity prediction model based on lifetime forecast of online videos. In Proceedings of the 2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC), Beijing, China, 23–25 September 2016; pp. 376–380.
20. Quader, N.; Gani, M.O.; Chaki, D.; Ali, M.H. A machine learning approach to predict movie box-office success. In Proceedings of the 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 22–24 December 2017; pp. 1–7.
21. Shuxin, O.; Chenyu, L.; Xueming, L. Analyzing the dynamics of online video popularity. *J. China Univ. Posts Telecommun.* **2017**, *24*, 58–69. [CrossRef]
22. Pinto, H.; Almeida, J.M.; Gonçalves, M.A. Using early view patterns to predict the popularity of YouTube videos. In Proceedings of the 6th ACM International Conference on Web Search and Data Mining, Rome, Italy, 4–8 February 2013; pp. 365–374.
23. He, X.; Gao, M.; Kan, M.Y.; Liu, Y.; Sugiyama, K. Predicting the popularity of web 2.0 items based on user comments. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, Gold Coast, Australia, 6–11 July 2014; pp. 233–242.
24. Xu, J.; Van Der Schaar, M.; Liu, J.; Li, H. Forecasting popularity of videos using social media. *IEEE J. Sel. Top. Signal Process.* **2014**, *9*, 330–343. [CrossRef]
25. Stokowiec, W.; Trzciński, T.; Wołk, K.; Marasek, K.; Rokita, P.; Kryszkiewicz, M.; Appice, A.; Ślęzak, D.; Rybinski, H.; Skowron, A.; et al. Shallow reading with deep learning: Predicting popularity of online content using only its title. In Proceedings of the International Symposium on Methodologies for Intelligent Systems, Warsaw, Poland, 26–29 June 2017; Springer: Cham, Switzerland, 2017; pp. 136–145.

26. Trzciński, T.; Andruszkiewicz, P.; Bocheński, T.; Rokita, P. Recurrent neural networks for online video popularity prediction. In Proceedings of the International Symposium on Methodologies for Intelligent Systems, Warsaw, Poland, 26–29 June 2017; Springer: Cham, Switzerland; pp. 146–153.
27. Fontanini, G.; Bertini, M.; Del Bimbo, A. Web video popularity prediction using sentiment and content visual features. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, New York, NY, USA, 6–9 June 2016; pp. 289–292.
28. Ouyang, S.; Li, C.; Li, X. A Peek into the Future: Predicting the Popularity of Online Videos. *IEEE Access* **2016**, *30*, 26–33. [CrossRef]
29. Acikgoz, F.; Burnaz, S. The influence of influencer marketing on YouTube influencers. *Int. J. Internet Mark. Advert.* **2021**, *15*, 201–219. [CrossRef]
30. Tan, J.; Liu, W.; Wang, T.; Zhao, M.; Liu, A.; Zhang, S. A high accurate content popularity prediction computational modeling for mobile edge computing using matrix completion technology. *Trans. Emerg. Telecommun. Technol.* **2020**, *11*, e3871. [CrossRef]
31. Yao, Y.; Tong, H.; Xu, F.; Lu, J. On the measurement and prediction of web content utility: A review. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 1–2. [CrossRef]
32. Tan, Z.; Wang, Y.; Zhang, Y.; Zhou, J. A Novel Time Series Approach for Predicting the Long-Term Popularity of Online Videos. *IEEE Trans. Broadcast.* **2016**, *62*, 436–445. [CrossRef]
33. Kurdi, M.; Albadi, N.; Mishra, S. "Think before you upload": An in-depth analysis of unavailable videos on YouTube. *Soc. Netw. Anal. Min.* **2021**, *11*, 1–21. [CrossRef]
34. Chae, J. YouTube makeup tutorials reinforce postfeminist beliefs through social comparison. *Media Psychol.* **2021**, *24*, 167–189. [CrossRef]
35. Abousaleh, F.S.; Cheng, W.-H.; Yu, N.-H.; Tsao, Y. Multimodal Deep Learning Framework for Image Popularity Prediction on Social Media. *IEEE Trans. Cogn. Dev. Syst.* **2020**, *13*, 679–692. [CrossRef]
36. Dou, H.; Zhao, W.X.; Zhao, Y.; Dong, D.; Wen, J.R.; Chang, E.Y. Predicting the popularity of online content with knowledge-enhanced neural networks. In Proceedings of the ACM KDD Conference on Knowledge Discovery and Data Mining, London, UK, 19–23 August 2018.
37. Ma, C.; Yan, Z.; Chen, C.W. Larm: A lifetime aware regression model for predicting youtube video popularity. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 467–476.
38. Sangwan, N.; Bhatnagar, V. Video popularity prediction based on fuzzy inference system. *J. Stat. Manag. Sys.* **2020**, *23*, 1173–1185. [CrossRef]
39. Alshangiti, M.; Shi, W.; Liu, X.; Yu, Q. A Bayesian learning model for design-phase service mashup popularity prediction. *Expert Syst. Appl.* **2020**, *149*, 113231. [CrossRef]
40. Bai, W.; Zhang, Y.; Huang, W.; Zhou, Y.; Wu, D.; Liu, G.; Xiao, L. DeepFusion: Predicting movie popularity via cross-platform feature fusion. *Multimed. Tools Appl.* **2020**, *19*, 1–8. [CrossRef]
41. Laiche, F.; Letaifa, A.B.; Elloumi, I.; Aguili, T. When Machine Learning Algorithms Meet User Engagement Parameters to Predict Video QoE. *Wirel. Pers. Commun.* **2021**, *116*, 2723–2741. [CrossRef]
42. Zhang, Y.; Li, P.; Zhang, Z.; Zhang, C.; Wang, W.; Ning, Y.; Lian, B. GraphInf: A GCN-based Popularity Prediction System for Short Video Networks. In Proceedings of the International Conference on Web Services 2020, Beijing, China, 20–24 July 2020; Springer: Cham, Switzerland, 2020; pp. 61–76.
43. Amin, F.; Choi, G.S. Hotspots Analysis Using Cyber-Physical-Social System for a Smart City. *IEEE Access* **2020**, *8*, 122197–122209. [CrossRef]