



Research article



An improved deep convolutional neural network-based YouTube video classification using textual features

Ali Raza^a, Faizan Younas^b, Hafeez Ur Rehman Siddiqui^c, Furqan Rustam^d,
Monica Gracia Villar^{e,f,g}, Eduardo Silva Alvarado^{e,h,i}, Imran Ashraf^{j,*}

^a Department of Software Engineering, The University of Lahore, Lahore, Pakistan

^b Department of Computer Science and Information Technology, The University of Lahore, Lahore, Pakistan

^c Faculty of Computer Science and Information Technology, Khawaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan

^d School of Computer Science, University College Dublin, D04 V1W8 Dublin, Ireland

^e Universidad Europea del Atlántico, Isabel Torres 21, 39011 Santander, Spain

^f Universidad Internacional Iberoamericana, Arecibo, PR 00613, USA

^g Universidade Internacional do Cuanza, Cuito, Bié, Angola

^h Universidad Internacional Iberoamericana, Campeche 24560, Mexico

ⁱ Universidad de La Romana, La Romana, Dominican Republic

^j Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Korea

ARTICLE INFO

Dataset link: <https://www.kaggle.com/datasets/aashishbidap/youtube-api-data-for-text-categorization>

Keywords:

YouTube video categorization
Convolutional neural network
Text categorization
Text features

ABSTRACT

Video content on the web platform has increased explosively during the past decade, thanks to the open access to Facebook, YouTube, etc. YouTube is the second-largest social media platform nowadays containing more than 37 million YouTube channels. YouTube revealed at a recent press event that 30,000 new content videos per hour and 720,000 per day are posted. There is a need for an advanced deep learning-based approach to categorize the huge database of YouTube videos. This study aims to develop an artificial intelligence-based approach to categorize YouTube videos. This study analyzes the textual information related to videos like titles, descriptions, user tags, etc. using YouTube exploratory data analysis (YEDA) and shows that such information can be potentially used to categorize videos. A deep convolutional neural network (DCNN) is designed to categorize YouTube videos with efficiency and high accuracy. In addition, recurrent neural network (RNN), and gated recurrent unit (GRU) are also employed for performance comparison. Moreover, logistic regression, support vector machines, decision trees, and random forest models are also used. A large dataset with 9 classes is used for experiments. Experimental findings indicate that the proposed DCNN achieves the highest receiver operating characteristics (ROC) area under the curve (AUC) score of 99% in the context of YouTube video categorization and 96% accuracy which is better than existing approaches. The proposed approach can be used to help YouTube users suggest relevant videos and sort them by video category.

* Corresponding author.

E-mail addresses: ali.raza.scholarly@gmail.com (A. Raza), younasfaizan97@gmail.com (F. Younas), hafeez@kfu.edu.pk (H.U.R. Siddiqui), furqan.rustam1@gmail.com (F. Rustam), monica.gracia@uneatlantico.es (M.G. Villar), eduardo.silva@uneatlantico.es (E.S. Alvarado), ashrafimran@live.com (I. Ashraf).

<https://doi.org/10.1016/j.heliyon.2024.e35812>

Received 21 September 2023; Received in revised form 4 August 2024; Accepted 5 August 2024

Available online 10 August 2024

2405-8440/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Video content on the web platform has increased explosively during the past decade, thanks to the open access to Facebook, YouTube, etc. YouTube, founded in early 2005, has expanded rapidly to become the second-largest video-sharing platform with approximately 37 million channels [1]. YouTube is a video-sharing website that allows users to watch, share, enjoy, comment on, and upload their videos. This implies that individuals are continually seeking information on YouTube and discovering videos about different topics of interest. The YouTube platform serves as an easy way for users to share and save videos online. Every minute, individuals all around the world post approximately 300 hours of video on YouTube. YouTube videos cover a large variety of subjects that someone chooses to advertise a video on. These videos are easy to distribute via social media platforms, websites, and email, and they may be even embedded in other websites. YouTube video content categories manage YouTube channels and videos. The finest-suited category for video aids the viewers in easily finding the video. To gain better visibility and rank, videos must be categorized and different tags must be used. The availability of such a large amount of video content requires automatic categorization to make them easily searchable and useful for the users [1].

Deep learning models have been extensively used for tasks related to image processing including video classification. Deep learning techniques have produced impressive achievements in Natural language processing (NLP) tasks [2]. Most of the NLP work with deep learning approaches has included learning word-to-vector representations using neural language techniques and doing classification configuration over the learned word-to-vectors. Convolutional neural networks (CNN) make use of layers that contain convolving filters that are applied to local features. CNN techniques were originally developed for computer vision, but they have now been demonstrated to be successful for NLP, with outstanding results in semantic parsing, sentence modeling, search query retrieval, and other standard NLP tasks. For video classification, video, and textual features can be used where the former contains signals-related features while the former relies on the textual information provided by the user like title, description, etc. Text-based video categorization requires less computing capacity and is less complicated as compared to video or image processing, however, it is limited when the information provided for the video is missing or short. Also, the amount of labeled data for text-based categorization is rather small.

Despite the above-mentioned challenges, text-based video categorization has potential applications and can be leveraged to achieve better results with less computational complexity. The process of categorizing text materials into two or more groups is known as text classification [3]. The most prevalent method is binary classification, which assigns the content in the corpus to one of two groups. Text categorization is frequently the first stage in selecting a group of documents for subsequent processing, and it can sometimes be the sole step in text processing. The purpose of text categorization is not to extract information from a text other than the document's category. The fundamental approach to text classification is to create a collection of characteristics to characterize a document and then use an algorithm intended to analyze and use these features to determine the proper category for a given document. This study leverages the text-based video categorization and focuses on the following functions in essence:

- YouTube exploratory data analysis (YEDA) is performed to get useful insights related to YouTube videos. The daily posted content analysis and videos by category are analyzed to find out their relevant domain. For this purpose, the YouTube dataset scraped using the YouTube application programming interface (API) is utilized.
- The study proposes an improved deep CNN for video categorization with high accuracy. The architectural analysis of the proposed approach is analyzed in terms of the layers stack involved in model building. In addition, two advanced deep learning approaches are selected for video categorization including recurrent neural network (RNN), and gated recurrent units (GRU) for making the performance appraisal. Hyperparameter tuning of all the employed approaches is carried out to get more accurate results.
- The comparative performance analysis of the employed approaches is done in terms of training accuracy, testing accuracy, recall, precision, F1 score, and receiver operating characteristic (ROC) area under curve (AUC) accuracy score. Moreover, k-fold cross-validation and comparative analysis with state-of-the-art approaches are performed to show the supremacy of the proposed approach. In addition, logistic regression (LR), decision tree (DT), support vector machine (SVM), and random forest (RF) models are implemented for performance comparison.

The remainder of this study is organized into four sections. In Section 2, the literature work related to our research study is examined. The methodological analysis of the study, along with the implemented deep learning models is conducted in Section 3. Results and evaluation of the proposed technique are examined in Section 4. The study is wrapped up in Section 5.

2. Literature review

This section analyzes past research studies related to YouTube video classification. A machine learning technique was employed in [4] for video classification. An improved RF classifier was used to fine-tune different parameters to classify videos into 6 categories. Using the bag of words (BoW) approach with several preprocessing steps, the study achieved promising results, however, the models were not investigated extensively regarding different hyperparameters. A text-based tool for YouTube video content categorization using descriptions, titles, and comments was proposed in [5]. The text features were extracted using lexical, content-specific, and syntactic features. Three machine learning models were employed including Naïve Bayes (NB), SVM, and DT for experiments. A self-collected dataset was used in the study to obtain 87.2% accuracy using SVM. Despite investigating several machine learning models, the reported accuracy was not suitable for YouTube video categorization and should further be improved.

A deep learning approach was presented in [6] for classifying YouTube nudity content. The NPDI cartoon dataset of YouTube videos was employed for classification. The dataset contains 111,156 manually annotated cartoon clips related to safety, sexual nudity, and fantasy violence. A deep learning-based EfficientNetB7 with CNN was utilized in this regard. Results showed better performance than existing studies with 93% precision, 92% recall, and 92% F1 score, however, the proposed model had a higher computational complexity.

Besides classifying the videos, the analysis of metadata and posted views is an important research field to increase the visibility of the videos. For example, the study [7] proposed a technique to extract and analyze metadata from YouTube videos. The proposed technique utilized a sentiment analysis approach for classification and finding the polarity of the YouTube videos into positive, neutral, and negative. Genetic algorithms, neural networks, Bayesian learning, and SVM were applied to classify YouTube video content. Results are promising, however, due to the small size of the dataset, the results cannot be generalized. In addition, the accuracy of the employed models is influenced by the type of text words.

Along the same lines, [8] classified YouTube video comments into positive or negative categories based on the comment sentiment. Five machine learning-based classification models were utilized with parameter optimization to obtain better results. Linear SVM and logistic regression obtained superior results each with an 86% F1 score. YouTube video comments classification was done by a deep neural network-based sentiment classification model by [9]. Word embedding was implemented to convert the text into tensors. The proposed model contained two convolutional layers to extract prominent features and to make the classification a fully connected dense layer. The performance evaluation results demonstrated that the technique can accurately classify 84% of videos.

The automatic classification of YouTube video comments was discussed in [10]. The video's comments were classified into four categories: irrelevant, relevant, negative, and positive. For the classification task, the YouTube video description content was associated with the comments. The association list and BoW techniques were applied to classify the content of the video. The assigning of relevant categories such as cooking, sports, and dogs to YouTube videos based on a text-based approach was proposed in [11]. A text-based weakly supervised classifier was trained on the YouTube video content. The dataset content was based on YouTube video titles, user tags, descriptions, and comments. The proposed Video2Text classifier achieved an 85% accuracy score.

The study [12] introduced a model for identifying the authors of literary texts by applying deep learning techniques structured into three distinct stages: initial text processing, the extraction of features, and the final classification process. Initially, the text data was processed using Word2Vec to create word vectors. Subsequently, an enhanced method for text feature extraction, utilizing both CNN and Attention mechanisms, was employed to identify significant text features. These features were then fed into the convolution layer of a CNN. The classification phase employed LSTM and Softmax algorithms to finalize the author identification. Despite achieving an accuracy of 78.98%, the results indicated that the model's performance fell short compared to leading methodologies, highlighting the necessity for further advancements and improvements in machine learning.

In [13], the researchers evaluated various pre-trained language models for carbon emissions, time, and accuracy in multi-label text classification by applying AutoML methodologies. They employed a dataset comprising 250,000 Turkish language entries to evaluate these models. The findings demonstrated that the BERTurk language model delivered superior accuracy, requiring only 66 minutes for training while maintaining minimal carbon emissions. The model attained an accuracy rate of 88.09%, which, although impressive, poor when compared to more recent approaches. The authors suggested adopting more advanced machine-learning techniques could enhance these performance metrics.

The study [14] explored sentiment analysis within a multilingual framework, comparing traditional machine learning techniques and advanced hybrid deep learning approaches. They implemented a Conv1D-LSTM model for categorizing sentiments. The findings highlighted that the support vector machine model outperformed all other evaluated models in terms of accuracy, recording a rate of 82.56% for English and 86.43% for Bangla sentiment analysis, with the assistance of the Porter stemming algorithm. In deep learning models, the BiLSTM-based approach was identified as the most effective, achieving accuracies of 78.10% for English and 83.72% for Bangla texts, again utilizing Porter stemming. However, compared to the leading-edge methods for text classification, these performance metrics were considered subpar.

In the research [15], the focus was on categorizing events related to food safety through the application of a layered transformer model. The method began by feeding text segments into the BERT model. Following this, the outputs generated by the BERT model were merged with key sentences extracted from the text, a Transformer model for transforming features and processes. The concluding step involved employing a classifier specifically for categorizing food safety news. The effectiveness of the model was demonstrated by its accuracy rate of 0.8402 in classifying newsworthy events, although there was an acknowledgment that these promising results could be enhanced further.

Table 1 provides an analytical summary of the discussed research works. Although several machine learning and deep learning approaches already exist for YouTube video classification, they have several limitations. First, often the number of classes used for evaluating the performance of models is rather small. Second, text-based video classification approaches are not studied very well. Third, the provided accuracy for YouTube content categorization is low. This study aims to resolve such issues by proposing a fine-tuned deep learning model. The proposed research study is based on an advanced deep CNN model which is employed to classify YouTube videos into 9 categories.

3. Methodology

This study uses a deep learning-based approach for YouTube video categorization, as shown in Fig. 1. The YouTube dataset used in this study was created by scraping the YouTube video title and description using the YouTube API. To make the YouTube dataset completely preprocessed, text noise cleaning was applied to the title and description of each YouTube video data. YEDA was

Table 1
The literature research limitations analysis.

Ref	Approach for video classifications	Advantages of research	Disadvantages & Limitations
[4]	Random Forest classifier	Research helps to categorize relevant YouTube videos based on user relevance.	The machine learning models were not hyperparameter-tuned. In addition, deep learning models are not investigated in this regard.
[8]	Linear Support Vector classifier	Helps to classify YouTube video comments into positive or negative categories.	Lower performance is reported necessitating further research.
[6]	EfficientNetB7	Research helps to classify videos as safe, sexual nudity, and fantasy violence.	Computational complexity, no validation.
[7]	Genetic Algorithms	Helps to classify YouTube video comments into positive or negative categories.	A smaller dataset is used for the experiment consisting of only 15 videos.
[5]	Naïve Bayes	YouTube video classification results to identify implicit cyber communities	Low performance scores were achieved.
[9]	Convolutional Neural Network	Helps to classify YouTube video comments into positive or negative categories.	Low-performance scores were achieved.
[10]	Association list and bag of words-based techniques	Youtube video's comments were classified into four categories labels irrelevant, relevant, negative, and positive.	Classical machine learning models are used only, performance comparison with state-of-the-art is missing.
[11]	Video2Text classifier	The relevant categories such as cooking, sports, and dogs to YouTube videos.	Low-performance scores were achieved.

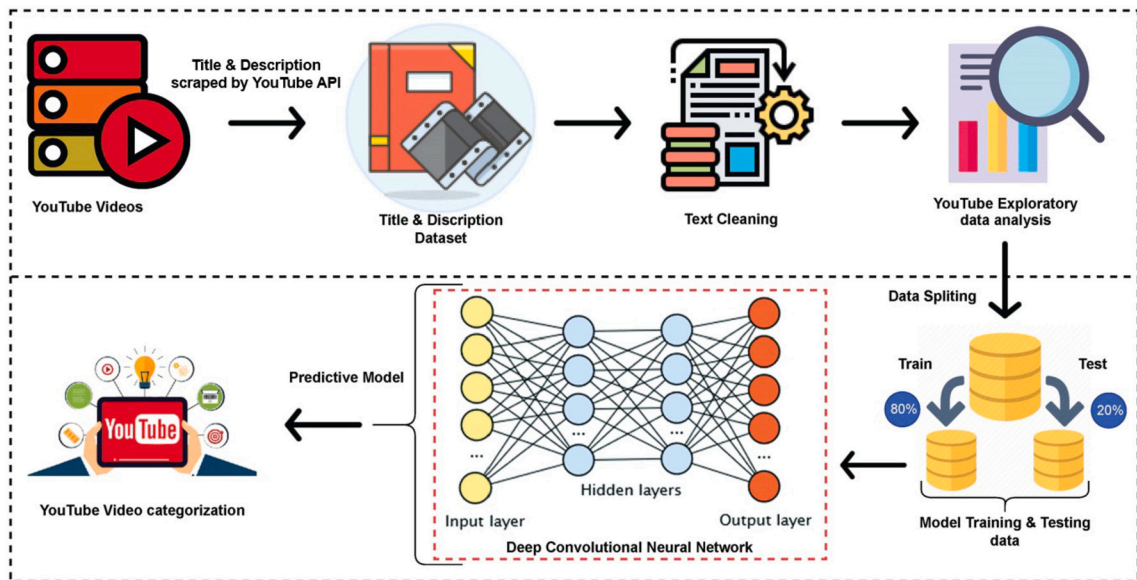


Fig. 1. The methodological architecture of the proposed approach.

performed to get useful insights related to various video categories. For training and testing purposes, the dataset was split into test and train datasets. Three deep learning-based models were employed including the proposed deep CNN and two other deep learning models including GRU and RNN.

3.1. YouTube dataset

This study uses a novel dataset scraped and created by using YouTube API [16]. The YouTube dataset is publicly available for research purposes at Kaggle [17]. The dataset includes video ID, video title, video description, and category features. The dataset is labeled using the YouTube API which includes a YouTube video description and the respective category of the video. The dataset contains 20000 videos falling under nice categories including adventure, art and music, food, history, manufacturing, nature, science and technology, sports, and travel. A few sample records from the research dataset are demonstrated in Table 2.

3.2. Text noise cleaning

YouTube video title and description text have unnecessary noise which was removed to make the proposed model more accurate for category prediction. The preprocessing pipeline followed several steps [18]. In the first step, the text was transformed into lowercase. Then the unnecessary numbers were removed. In the next step of text cleaning, punctuations were removed from the title and

Table 2
Sample information related to YouTube dataset.




Thumbnail	Title	Description	Video ID	Category
	Welcome to Bali Travel Vlog Priscilla Lee	DISCLAIMER* Please do not ride elephants when visiting any country. At the time I didn't know (yes, I was dumb) so it is shown in the video, but I do not support	i9E_Blai8vk	travel
	Under The Gravestones Time Team (Archaeology Documentary) Timeline	The team attempted to uncover an ornate mosaic floor in the burial grounds of St Kyneburgha church. Get 3 months of History Hit access for \$3 using code 'timeline'	Oyx5slLn-MrQ	history
	VIRAL TIKTOK RECIPE DELICIOUS FOOD HACK 2020	Are you bored? Looking for something to do or something to eat? TRY THIS VIRAL FOOD RECIPE from TikTok. This is delicious Easy to make Simple	OIE_ulrdRgM	food

Table 3
Actual and preprocessed YouTube video contents.

Title	Description	Preprocessed Title	Preprocessed Description
Welcome to Bali Travel Vlog Priscilla Lee	DISCLAIMER* Please do not ride elephants when visiting any country. At the time I didn't know (yes, I was dumb) so it is shown in the video, but I do not support	welcome bali travel vlog priscilla lee	disclaimer please ride elephants visiting country time do not know yes was dumb so shown video, but do not support
Under The Gravestones Time Team (Archaeology Documentary) Timeline	The team attempted to uncover an ornate mosaic floor in the burial grounds of St Kyneburgha church. Get 3 months of History Hit access for \$3 using code 'timeline'	under gravestones time team archaeological documentary timeline	team attempted uncover ornate mosaic floor burial grounds stkyneburgha church get months history hit access using code timeline
VIRAL TIKTOK RECIPE DELICIOUS FOOD HACK 2020	Are you bored? Looking for something to do or something to eat? TRY THIS VIRAL FOOD RECIPE from TikTok. This is delicious Easy to make Simple	viral tiktok recipe delicious food hack	you bored looking something do something eat try viral food recipe tiktok delicious easy make simple

description of each YouTube video. Now the extra white spaces were removed. Tokenization was applied to convert text to smaller units or words. In the next step, the non-alphabetic tokens were removed, in addition to removing the stop words. In the last step of text cleaning, word lemmatization was applied to convert all tokens into their root words, as shown in Table 3.

3.3. YouTube exploratory data analysis

We performed analyses on the dataset and got the exploratory statistics of the YouTube dataset. The extracted statistics are in two forms: observational level and corpus level. At the observational level, there are 100 average characters per the description of the video. There are 13 average words per description. We also calculated the average vocabulary size [19] of every description and found that the average vocabulary size per description is 112. We also calculated the lexical richness [20] of the description column. The quality of the vocabulary in a language sample is referred to as lexical richness. Some studies associate it with the variety of lexis, whereas others see it as a multidimensional term [20]. The average lexical richness per description is 0.93. When we explored the statistics on the corpus level, we found that the average quantity of words in the corpus description column is 394,497, and vocabulary size is 43,971, the total lexical richness is 0.11 and the average number of characters per word is 6.31.

Afterward, we find 50 common words which are used in the description column of the corpus. The most common word used in the description is 'video'. It is repeated in the description more than four thousand times, as shown in Fig. 2 which shows the count of each of the top 50 words used in the corpus. Fig. 3, on the other hand, shows the word cloud of the 50 most common words used

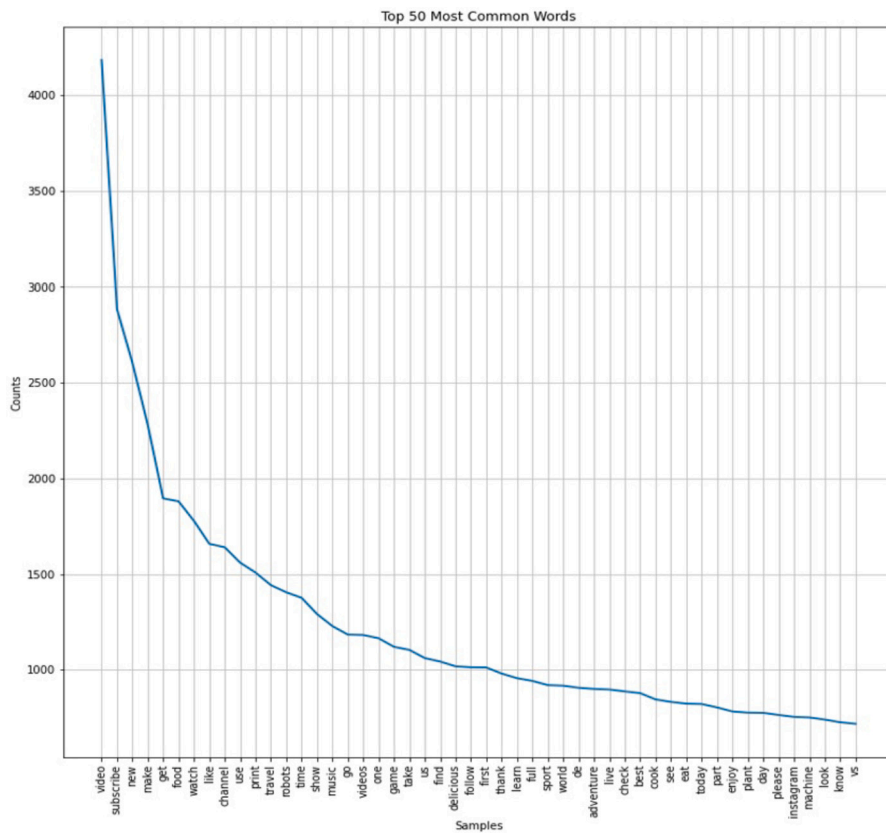


Fig. 2. The most 50 common words bar analysis.



Fig. 3. The most 50 common words cloud analysis.

in the description column. It provides an easy and illustrative way to find the most common words. The second most common word that is repeated in the description is ‘subscribe’ which is repeated 3500 times.

YouTube exploratory data analysis was conducted on the dataset. YouTube video category bar chart distribution analysis is shown in Fig. 4 which shows that the science and technology videos category has the highest number of samples among all categories, followed by the food and travel category. The adventure category has the lowest 2091 samples for the video dataset.

The ratio-wise distribution of the video categories is shown in Fig. 5 which indicates that science and technology, food, and travel categories each have a 14% share of all the categories.

Word clouds have evolved as a simple and visually engaging way of text representation. They are utilized in a variety of tasks to provide a summary by reducing text down to the words that appear the most frequently. This is typically done statistically as simple text summarization [21]. We make the word cloud on the description column and show the most dominant words used in the description regarding each category. The word clouds of the most common words found in each category are shown in Fig. 6.

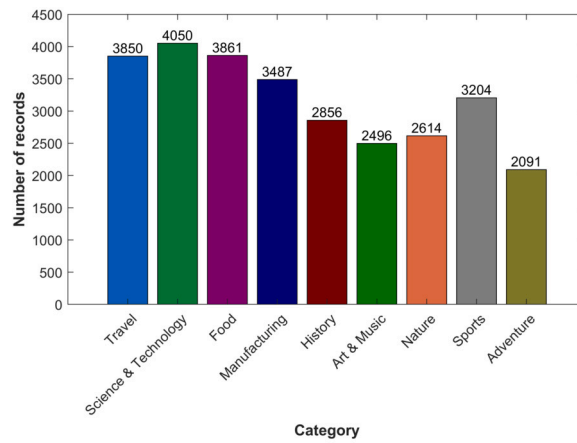


Fig. 4. The YouTube video category bar chart distribution analysis.

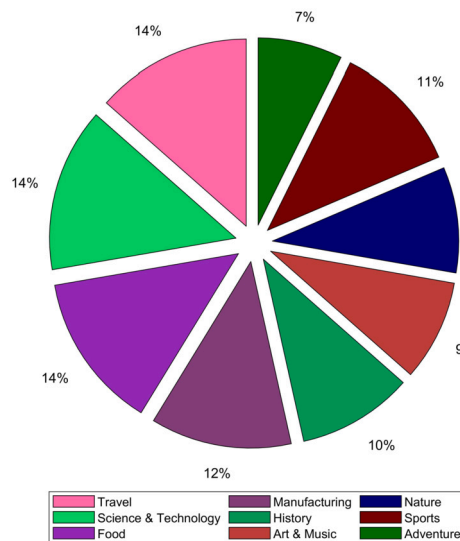


Fig. 5. The YouTube video category group pie chart distribution analysis.

3.4. YouTube data splitting

The dataset was split into training and test sets in the ratios of 0.8 to 0.2 to train the models and evaluate their performance, respectively.

3.5. One hot encoding

The one-hot encoding is applied with the employed deep neural networks in this study. The one-hot encoding is characterized as the crucial process of changing categorical data variables which improves the classification and prediction accuracy of a model. This form of encoding generates a new binary feature form for every category and assigns the feature of every sample that belongs to its actual category a value of 1. The vocabulary size of one-hot encoding for our proposed approach is 5000. Then the sentences are padded to the length of 100.

3.6. Deep neural networks used in study

The RNN is an artificial neural network that is used to analyze sequential input [22]. Since RNN contains memory, they are particularly successful for sequence tasks and NLP tasks. RNNs are referred to as recurrent because they do a similar task for each element of a sequence, with the result dependent on past calculations. They can receive inputs $(x)_t$ one at a time and remember certain information via hidden layer activations transferred from one step to the next. This enables RNN to use past information to process subsequent inputs. The RNN converts an input data sequence x value to an output data sequence value (o) . The difference



Fig. 6. The top 50 words cloud analysis of YouTube video categories, (a) Travel, (b) Science, (c) Food, (d) Manufacturing, (e) History, (f) Art, (g) Natural, (h) Sports, and (i) Adventure.

between the expected output o and the actual output y is measured by a loss function L . The RNN additionally features input-to-hidden connections, hidden-to-hidden connections, and hidden-to-output connections that are parametrized by a weight matrix U , W , and V , respectively. Then from time step $t = 1$ to $t = n$ the Equations (1) to (4) are applied:

$$a^{(t)} = b + Wh^{(t-1)} + Ux^{(t)} \quad (1)$$

$$h^{(t)} = \tanh(a^{(t)}) \quad (2)$$

$$o^{(t)} = c + Vh^{(t)} \quad (3)$$

$$\hat{y}^{(t)} = \text{softmax}(o^{(t)}) \quad (4)$$

where $x^{(t)}$ is the input vector at time t , $h^{(t)}$ is the hidden state, U , W , and V are weight matrices for input-to-hidden, hidden-to-hidden and hidden-to-output connections, respectively.

The bi-directional GRU is a step forward from the ordinary RNN [23]. GRUs are a gating technique in artificial recurrent neural networks that were discovered to be comparable to long short-term memory (LSTM) techniques [24]. A GRU is a form of RNN that tackles the problem of long-term connections, which can result in disappearing gradients in bigger vanilla RNNs. The GRUs overcome this issue by retaining memory from the prior time point to assist the network in making future predictions. The GRU network is a version of the LSTM that tries to lower the computation cost of the network. The GRU model is expressed in Equations (5) and (6)

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (5)$$

Table 4
The architecture configurations of GRU, and RNN.

Model layers	Unit	Activation	Output shape	Parameters
Recurrent Neural Network				
Feature Embedding layers.	50000	N/A	(None, 100, 64)	3,200,000
Gated recurrent unit layers.	20	N/A	(None, 20)	5,160
Dropout layers	0.5	N/A	(None, 20)	0
Dense layers.	10	RELU	(None, 10)	210
Dropout layers	0.5	N/A	(None, 10)	0
Dense layers.	9	SOFTMAX	(None, 9)	99
Gated Recurrent Unit				
Feature Embedding layers.	50000	N/A	(None, 100, 64)	3,200,000
Recurrent Neural Network layer	32	N/A	(None, 32)	3,104
Dense layers.	16	RELU	(None, 16)	528
Dense layers.	9	SOFTMAX	(None, 9)	153

$$\tilde{h}_t = g(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \quad (6)$$

where h_t shows the output vector, x_t is the input vector, z_t is the update gate vector, \odot shows the Hadamard product, W_h , U_h , and b are parameters learned during training.

These models are optimized regarding the layered structure, dropout layers, and the number of neurons, and a detailed specification of GRU and RNN is provided in Table 4.

3.7. Proposed deep convolutional neural network

This study proposes a deep CNN which is mostly utilized in computer vision tasks. However, they have also been used for a variety of NLP tasks and provided promising results. It was identified that CNN [25] also has an outstanding capacity for sequential data analysis tasks such as NLP. The numerous shapes and sizes series of filters involved in CNN convolve the actual sentence matrix to reduce it into matrices of low dimension. The pooling and convolution are the two main operations of CNN. The spatial information can be preserved in convolution operation by extracting the feature map from the dataset using multiple filters. The dimensionality reduction of feature maps from the convolution operation is done by using the pooling operation, also known as subsampling. The activation function rectified linear unit (ReLU) is the common choice to transfer gradient in the training process by backpropagation in CNN. The convolution formula is expressed in Equation (7)

$$(f * g)(t) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau \quad (7)$$

In-text classification CNN is being applied to distributed and discrete word embedding. The downsampling technique used in the CNN is L2 regularization. CNN utilizes an activation function that helps it run in kernel high dimensional space for neural processing. When a special pattern is detected in data, the result of each convolution will be kept. The expressions such as ‘very good’, and ‘I hate’ could be Patterns. The CNNs identify the patterns in the sentence instead of their position. The architecture stack of the proposed CNN is shown in Fig. 7. The embedding layer, pooling layer, convolution layer, and fully connected layer are the four types of layers in the architecture of CNN. Our proposed CNN is utilized for multi-label categorization tasks. The nine classes are used for the multi-label categorization task in our research problem.

3.7.1. Proposed deep learning technique hyperparameter tuning

The hyperparameter tuning of the proposed DCNN is examined in this section. The tuning parameter is the model layers used, the number of neuron units, the activation function, the total trainable parameters, and the output shape of each layer utilized by the architecture stack layers. The layers architecture of each model is based upon the feature embedding layer, model layers, and the dense layers family. Table 5 shows the details regarding the layers, activation function, and parameters used for training.

3.7.2. Optimized deep CNN layers architectural analysis

The proposed DCNN approach is examined in the context of its architectural layers analysis, as visualized in Fig. 7. The first layer is the input layer of the model. The output of the input layer is given to the next embedding layer in the architecture. Furthermore, a family of conventional layers is based on two stacked layers of 1D conventional and 1D max pooling. Then a flattened layer is implemented in the architecture to make data in a one-dimensional sequence. In conclusion, a family of dense layers is employed at the end of the architecture. It contains the model output layer.

The proposed DCNN architecture is based on a unique concept of two stacks of conventional layers. The proposed DCNN achieved high-performance scores for YouTube video classification as it is highly optimized layer-wise and hyperparameter-tuned. These high-performance scores are achieved due to the rich text information learned by the proposed DCNN architecture, as visualized in Fig. 7.

Table 5
Deep CNN hyperparameter configuration.

Model layers	Unit	Activation	Output shape	Parameters
Feature Embedding layers.	50000	N/A	(None, 100, 64)	3,200,000
Convolutional layers.	64	ReLU	(None, 93, 64)	32,832
Global max pooling layers.	0	N/A	(None, 46, 64)	0
Convolutional layers.	32	ReLU	(None, 39, 32)	16,416
Global max pooling layers.	0	N/A	(None, 19, 32)	0
Flatten layers.	0	N/A	(None, 608)	0
Dense layers.	16	ReLU	(None, 16)	9,744
Dense layers.	9	Softmax	(None, 9)	153

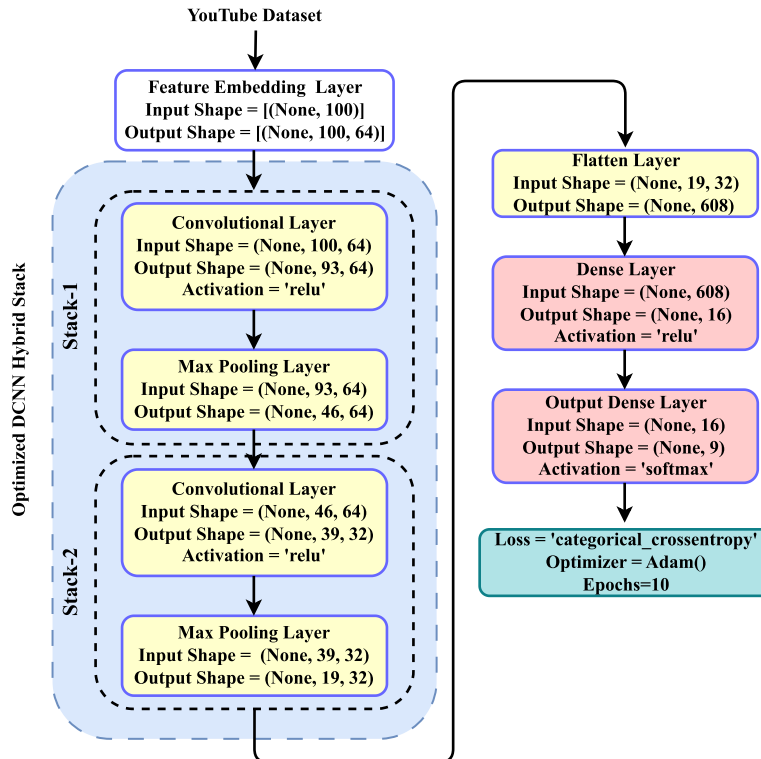


Fig. 7. The architecture of the proposed DCNN.

4. Results and discussions

4.1. Experiment setup

All the research experiments were done on a computer with a CPU of an Intel I5-8265U, random access memory (RAM) of 15GB, and a graphic card NVIDIA Tesla K80 GPU. TensorFlow 2.8.0 platform was used to implement the deep learning models with the Python 3.7.12 version. The deep learning models were utilized to categorize the videos. The accuracy, recall, precision, ROC curve, and F1 score were used as model evaluation metrics.

4.2. Results of DCNN, GRU, and RNN

Experimental results regarding epochs, training time, training loss, validation loss, training accuracy, and validation accuracy of all models are shown in Table 6. Each model was trained for 10 epochs. In comparison with other applied models, the proposed DCNN has less training loss and validation loss values. The training accuracy score of DCNN is 97%, while RNN and GRU each have an accuracy score of 95%. This training analysis shows that the proposed DCNN has achieved higher training accuracy and best fit trained on the YouTube dataset.

The classification report of the DCNN regarding the testing accuracy is presented in Table 7 which includes category-wise precision, recall, and F1 score. Results vary regarding each category as the number of samples, text, and description vary for each category. The

Table 6
The Deep Convolutional Neural Network training accuracy and validation effects.

Epoch	Train time/step	Training loss	Training accuracy	Validation loss	Validation accuracy
Deep Convolutional Neural Network					
1	115 s 201 ms	0.8748	0.6953	0.3332	0.9156
2	111 s 195 ms	0.1825	0.9502	0.2690	0.9290
3	111 s 195 ms	0.0877	0.9750	0.2853	0.9301
4	113 s 198 ms	0.0589	0.9800	0.3223	0.9263
5	112 s 196 ms	0.0518	0.9824	0.3708	0.9191
6	112 s 196 ms	0.0448	0.9824	0.3590	0.9239
7	112 s 196 ms	0.0431	0.9834	0.3742	0.9220
8	113 s 198 ms	0.0389	0.9838	0.3682	0.9283
9	112 s 197 ms	0.0370	0.9841	0.4023	0.9228
10	113 s 198 ms	0.0326	0.9847	0.4611	0.9169
Gated Recurrent Unit					
1	45 s 85 ms	1.9900	0.2333	1.5228	0.4179
2	42 s 84 ms	1.4492	0.4382	0.9862	0.6943
3	42 s 84 ms	1.1729	0.5569	0.7846	0.7380
4	42 s 84 ms	1.0471	0.5935	0.6854	0.7826
5	42 s 85 ms	0.9846	0.6216	0.6142	0.8653
6	42 s 84 ms	0.9320	0.6401	0.6434	0.8488
7	42 s 84 ms	0.8962	0.6688	0.5080	0.9069
8	42 s 84 ms	0.8515	0.6827	0.4743	0.9091
9	41 s 83 ms	0.8277	0.6875	0.4487	0.9185
10	42 s 84 ms	0.7993	0.7034	0.4351	0.9158
Recurrent Neural Network					
1	33 s 55 ms	1.3775	0.4647	0.8695	0.6662
2	32 s 55 ms	0.4961	0.8384	0.4575	0.8551
3	33 s 57 ms	0.2009	0.9419	0.4357	0.8762
4	31 s 54 ms	0.1130	0.9673	0.4233	0.8900
5	32 s 56 ms	0.0726	0.9780	0.3979	0.9014
6	31 s 55 ms	0.0756	0.9760	0.6007	0.8731
7	32 s 56 ms	0.0664	0.9763	0.4803	0.8930
8	31 s 55 ms	0.0449	0.9827	0.4732	0.8972
9	32 s 55 ms	0.0517	0.9796	0.6095	0.8683
10	32 s 56 ms	0.0380	0.9832	0.6222	0.8680

Table 7
Performance results for the proposed DCNN.

Category	Precision	Recall	F1 score
Adventure	0.89	0.89	0.89
Art and music	0.94	0.91	0.92
Food	0.95	0.95	0.95
History	0.99	0.97	0.98
Manufacturing	0.93	0.96	0.95
Nature	0.95	0.95	0.95
Science and technology	0.96	0.96	0.96
Sports	0.97	0.90	0.93
Travel	0.90	0.90	0.90
Accuracy	0.96		

proposed DCNN model has the highest precision score of 0.99 for the history category of videos while the lowest precision of 0.89 is for the adventure class. The values of the F1 score and recall follow the same trend for the given categories.

Table 8 shows the results for the GRU model indicating a 1.00 precision for the sports category while the adventure class has the lowest precision of 0.79. GRU obtains higher values for recall and F1 score as well which are 0.98 and 0.97, respectively for the history category. GRU shows the highest accuracy score of 0.98 for the history class. However, the average accuracy for all classes is 0.91 which is substantially lower than DCNN which obtained an average accuracy score of 0.96.

Results for RNN are provided in Table 9 which indicates the values for accuracy, recall, precision, and F1 score are lower than DCNN and the GRU model. DCNN shows better accuracy of 0.96 compared to GRU and RNN which have 0.91 and 0.87 accuracy scores, respectively. The highest precision of RNN is 0.96 which is for the history category while the lowest precision of 0.65 belongs to the adventure category. On average, the performance of RNN is inferior to DCNN regarding all evaluation metrics.

The ROC AUC accuracy curve analysis for all employed approaches is examined in Fig. 8. Figs. 8a, 8b, and 8c shows the ROC curve for DCNN, GRU, and RNN, respectively. The analysis demonstrates that the AUC accuracy of each target label is at different threshold values. The proposed DCNN achieved a higher accuracy of 99%. This analysis demonstrates that the higher the ROC accuracy score the better the proposed model classification accuracy.

Table 8
Results of GRU for video categorization.

Category	Precision	Recall	F1 score
Adventure	0.79	0.79	0.79
Art and music	0.89	0.89	0.89
Food	0.98	0.93	0.96
History	0.95	0.98	0.97
Manufacturing	0.97	0.92	0.95
Nature	0.89	0.89	0.89
Science and technology	0.93	0.97	0.95
Sports	1.00	0.02	0.03
Travel	0.84	0.87	0.85
Accuracy	0.91		

Table 9
The RNN classification report.

Category	Precision	Recall	F1 score
Adventure	0.65	0.64	0.65
Art and music	0.83	0.80	0.82
Food	0.91	0.94	0.93
History	0.96	0.92	0.94
Manufacturing	0.90	0.89	0.90
Nature	0.89	0.83	0.86
Science and technology	0.90	0.89	0.89
Sports	0.93	0.85	0.89
Travel	0.80	0.86	0.83
Accuracy	0.87		

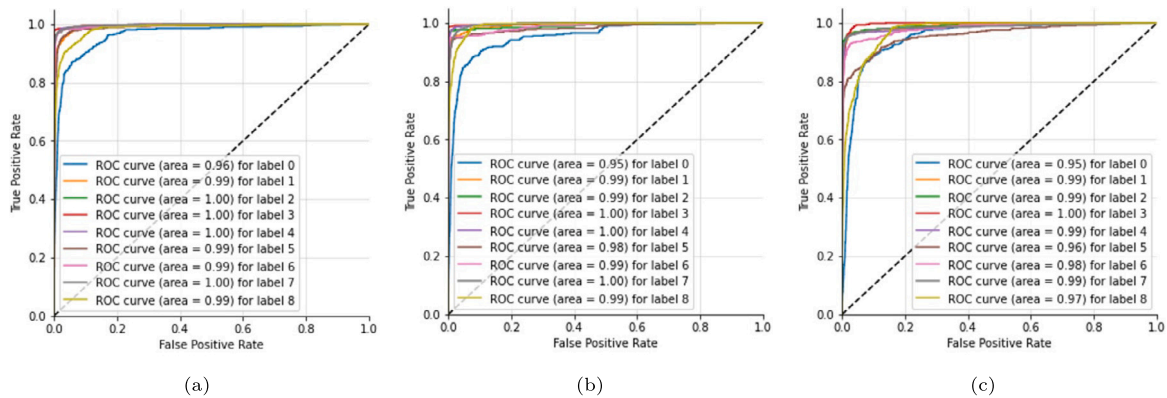


Fig. 8. Receiver operating characteristics for all models, (a) Deep convolutional neural network, (b) Gated recurrent unit, and (c) Recurrent neural network.

Table 10
Performance comparative analysis of all employed deep learning approaches.

Technique	Training accuracy%	Testing accuracy%	Precision%	Recall%	F1 score%	ROC (micro)%
DCNN	99	96	96	95	96	99
GRU	95	91	94	88	90	98
RNN	95	87	88	86	87	98

Performance analysis of all employed models is given in Table 10 which indicates that the proposed DCNN shows better performance in comparison to GRU and RNN regarding testing accuracy, training accuracy, average recall, precision, F1 score, and ROC.

The confusion matrix for the proposed DCNN is shown in Fig. 9. It provides the number of wrong and correct predictions for each of the 9 categories used for experiments. The matrix determines the actual categories and the predicted categories for the test dataset. It indicates that the highest ratio of correct predictions belongs to the history class where only 8 predictions are wrong out of a total of 550 predictions.

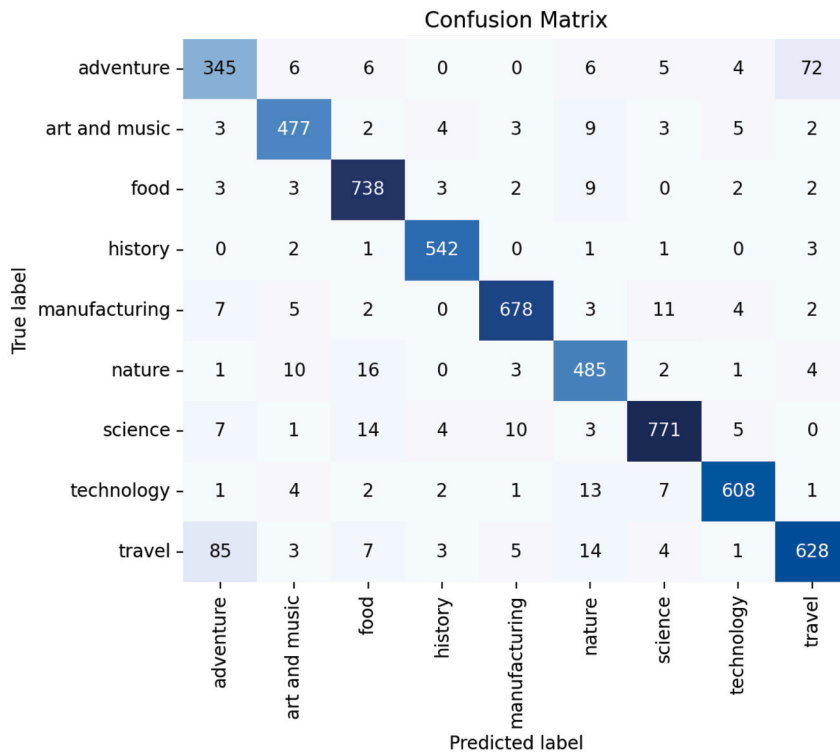


Fig. 9. Confusion matrix for the proposed DCNN.

Table 11
Correct and wrong predictions of the proposed DCNN model.

Category	Correct predictions	Wrong predictions	Total
Adventure	345	99	444
Art and Music	477	31	508
Food	738	24	762
History	542	8	550
Manufacturing	678	34	712
Nature	485	37	522
Science	771	44	815
Technology	608	31	639
Travel	628	132	760
Total	5,272	440	5712

Class-wise correct and wrong predictions are shown in Table 11. The DCNN model makes a total of 5,272 correct predictions out of a total of 5,712 predictions. A total of 440 samples, for all classes, are predicted incorrectly indicating the false positives and false negatives. These statistics are much better than other models used in this study.

4.3. Results of machine learning models

For a fair comparison, machine learning models are also employed in this study including LR, SVM, DT, and RF. These models are selected regarding their performance reported in previous studies [26–28]. Unlike deep learning models, machine learning models require handcrafted features using feature extraction approaches like Word2Vec, term frequency, etc. This study adopts the term frequency-inverse document frequency (TF-IDF) and a bag of words (BoW) features due to wide use and performance as reported in different studies [29,30]. Also, machine learning models are optimized regarding several hyperparameters. Table 12 shows the performance of machine learning models for YouTube video categorization. Results indicate the good performance of all models with BoW and TF-IDF, however, models perform better when using TF-IDF features with LR achieving the highest accuracy of 0.95 with both features.

Table 12

The performance evaluation of employed machine learning classifiers with Bag of Words and TF-IDF features.

Classifier	BoW accuracy	TF-IDF accuracy
LR	0.95	0.95
SVM	0.94	0.95
DT	0.93	0.94
RF	0.94	0.95

Table 13

K-Fold cross-validation comparative analysis of all employed deep learning approaches.

Fold	Accuracy (%)		
	DCNN	GRU	RNN
1	92	85	90
2	95	94	92
3	96	96	95
4	97	96	92
5	97	94	91
6	97	96	95
7	97	96	89
8	96	97	87
9	97	97	87
10	97	97	92
Mean accuracy	96	94	91
Standard deviation	± 0.51	± 3.43	± 2.17

Table 14

The comparative analysis of our proposed approach with other state-of-the-art studies.

Ref.	Year	Framework tool	Learning type	Proposed technique	Accuracy (%)	ROC Accuracy (%)
[4]	2019	Python	Machine learning	Random Forest	94	-
[8]	2021	Python	Machine learning	Logistic Regression	94	-
[31]	2019	Python	Machine learning	Extra Tree	93	-
Proposed	2022	Python	Deep learning	Improved DCNN	96	99

4.4. Results of cross-validation

This study carried out k-fold cross-validation to validate the performance of the proposed approach in comparison to GRU and RNN deep learning models. Table 13 shows the results of a 10-fold cross-validation for DCNN, GRU, and RNN indicating that the highest accuracy score is obtained using the proposed DCNN with the least standard deviation of ± 0.51 .

4.5. Comparison with existing approaches

The comparative analysis of our proposed approach with other state-of-the-art studies is provided in Table 14. The problem of YouTube videos classification into 6 categories using an RF is carried out in [4]. The study [8] proposed the sentiment classification of YouTube video comments using the LR model. Similarly, [31] employed the extra tree classifier for text classification. These state-of-the-art studies are built, deployed, and tested for classification on our YouTube videos dataset. The analysis demonstrates that the employed approach achieves the highest accuracy score among the compared studies even with 9 categories while other studies experiment with 6 categories.

4.6. Research implications

The economic implications of the proposed work are crucial. While the current research primarily focuses on enhancing video classification accuracy using textual features, the broader implications for various stakeholders are given as

- **Content Creators and Marketers:** Improved video categorization can lead to more accurate recommendations, thereby enhancing user engagement. For content creators and marketers, this means better targeting of audiences and potentially increased monetization opportunities.
- **YouTube Platform Owners:** Enhanced video classification can improve user satisfaction, retention, and ad revenue. Platforms can attract more advertisers and maintain a healthy ecosystem by providing relevant content to the users.

- **Advertisers:** Accurate video categorization allows advertisers to place their ads contextually, reaching the right audience. This alignment can lead to better conversion rates and return on investment.
- **Researchers and Developers:** The proposed approach contributes to the field of deep learning and natural language processing. It opens avenues for further research, innovation, and the development of more efficient algorithms.

In summary, the proposed research has the potential to impact various economic aspects within the digital content ecosystem. We encourage further exploration and collaboration to understand and leverage these implications fully.

4.7. Study limitations

Although the proposed approach contributes significantly to the existing literature, it may be limited by several challenges and limitations

- **High Training Time:** The training time for the applied deep learning models is high, which can be reduced by optimizing the layered architecture further. So, a trade-off may be required between high accuracy and increased training time.
- **Moderate Recall Results:** The recall results of the proposed method are moderate, but we are actively working to improve them for YouTube video categorization.
- **More Genres for Dataset:** Although this study included 9 categories of YouTube videos, with the increased number of uploaded videos, the categories are expanding. With increased video categories come the challenges of obtaining higher accuracy for multi-class problems which requires further research.

5. Conclusion

With the explosive growth of video-sharing platforms, video classification has appeared as an essential research area focusing on devising approaches to obtain high accuracy and efficiency. This study proposes a deep learning-based customized deep convolutional neural network for better performance regarding the categorization of YouTube videos using textual features of videos like title, description, etc. Performance analysis against the gated recurrent unit, recurrent neural network, and machine learning models was carried out involving 9 categories of YouTube videos. The proposed DCNN model achieved a 0.96 accuracy score and a high ROC AUC value of 0.99 among all the models. Performance appraisal with existing studies also showed superior results which further corroborate the supremacy of the model. The proposed research benefits various stakeholders, including content creators and marketers. Improved video categorization can lead to more accurate recommendations, enhancing user engagement. For YouTube platform owners, enhanced video classification can improve user satisfaction, retention, and ad revenue. Advertisers benefit from accurate video categorization, allowing them to place their ads contextually and reach the right audience.

In the future, we intend to enhance the YouTube video categories for more complex datasets. Advanced features engineering approaches will also be applied. Efforts are also planned to improve the computational complexity of the proposed deep learning model. Additionally, we plan to build an API for real-time categorization of YouTube videos based on their textual features, which will help users find videos closely aligned with their interests.

CRedit authorship contribution statement

Ali Raza: Writing – original draft, Data curation, Conceptualization. **Faizan Younas:** Writing – original draft, Formal analysis, Conceptualization. **Hafeez Ur Rehman Siddiqui:** Methodology, Formal analysis, Data curation. **Furqan Rustam:** Visualization, Software, Methodology. **Monica Gracia Villar:** Visualization, Investigation, Funding acquisition. **Eduardo Silva Alvarado:** Software, Project administration, Investigation. **Imran Ashraf:** Writing – original draft, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset used in this study is publicly available at: <https://www.kaggle.com/datasets/aashishbidap/youtube-api-data-for-text-categorization>

Acknowledgements

This study is funded by the European University of Atlantic.

References

- [1] F. Li, J.W. Chung, M. Claypool, Three-year trends in YouTube video content and encoding, in: 18th International Conference on Signal Processing and Multimedia Applications (SIGMAP 2021), 2021, pp. 15–22.
- [2] D.W. Otter, J.R. Medina, J.K. Kalita, A survey of the usages of deep learning for natural language processing, *IEEE Trans. Neural Netw. Learn. Syst.* 32 (2) (2020) 604–624, <https://doi.org/10.1109/TNNLS.2020.2979670>.
- [3] D. Zhang, M. Dong, Improved deep learning model text classification, in: 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), IEEE, 2021, pp. 217–220.
- [4] G.S. Kalra, R.S. Kathuria, A. Kumar, YouTube video classification based on title and description text, in: 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), IEEE, 2019, pp. 74–79.
- [5] C. Huang, T. Fu, H. Chen, Text-based video content classification for online video-sharing sites, *J. Am. Soc. Inf. Sci. Technol.* 61 (5) (2010) 891–906, <https://doi.org/10.1002/asi.21291>.
- [6] K. Yousaf, T. Nawaz, A deep learning-based approach for inappropriate content detection and classification of YouTube videos, *IEEE Access* 10 (2022) 16283–16298, <https://doi.org/10.1109/ACCESS.2022.3147519>.
- [7] S. Rangaswamy, S. Ghosh, S. Jha, S. Ramalingam, Metadata extraction and classification of YouTube videos using sentiment analysis, in: 2016 IEEE International Carnahan Conference on Security Technology (ICCST), IEEE, 2016, pp. 1–2.
- [8] R. Pokharel, D. Bhatta, Classifying YouTube comments based on sentiment and type of sentence, arXiv preprint, arXiv:2111.01908, 2021, <https://doi.org/10.48550/arXiv.2111.01908>.
- [9] A.A.L. Cunha, M.C. Costa, M.A.C. Pacheco, Sentiment analysis of YouTube video comments using deep neural networks, in: *International Conference on Artificial Intelligence and Soft Computing*, Springer, 2019, pp. 561–570.
- [10] K. Kavitha, A. Shetty, B. Abreo, A. D'Souza, A. Kondana, Analysis and classification of user comments on YouTube videos, *Proc. Comput. Sci.* 177 (2020) 593–598, <https://doi.org/10.1016/j.procs.2020.10.084>.
- [11] K. Filippova, K.B. Hall, Improved video categorization from text metadata and user comments, in: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011, pp. 835–842.
- [12] X. Tang, Author identification of literary works based on text analysis and deep learning, *Heliyon* 10 (3) (2024), <https://doi.org/10.1016/j.heliyon.2024.e25464>.
- [13] P. Savci, B. Das, Comparison of pre-trained language models in terms of carbon emissions, time and accuracy in multi-label text classification using AutoML, *Heliyon* 9 (5) (2023), <https://doi.org/10.1016/j.heliyon.2023.e15670>.
- [14] R.K. Das, M. Islam, M.M. Hasan, S. Razia, M. Hassan, S.A. Khushbu, Sentiment analysis in multilingual context: comparative analysis of machine learning and hybrid deep learning models, *Heliyon* 9 (9) (2023), <https://doi.org/10.1016/j.heliyon.2023.e20281>.
- [15] S. Xiong, W. Tian, V. Batra, X. Fan, L. Xi, H. Liu, L. Liu, Food safety news events classification via a hierarchical transformer model, *Heliyon* 9 (7) (2023), <https://doi.org/10.1016/j.heliyon.2023.e17806>.
- [16] Google Developers, YouTube data API, 2020.
- [17] A. Bidap, YouTube API data for text categorization, 2020.
- [18] D. Ji-Zhaxi, C. Zhi-Jie, C. Rang-Zhuoma, S. Maocuo, B. Mabao, A corpus preprocessing method for syllable-level Tibetan text classification, in: 2021 3rd International Conference on Natural Language Processing (ICNLP), IEEE, 2021, pp. 33–36.
- [19] B. Laufer, P. Nation, Vocabulary size and use: lexical richness in l2 written production, *Appl. Linguist.* 16 (3) (1995) 307–322, <https://doi.org/10.1093/applin/16.3.307>.
- [20] D. Malvern, B. Richards, Measures of lexical richness, the encyclopedia of applied linguistics, <https://doi.org/10.1002/9781405198431.wbeal0755>, 2012.
- [21] F. Heimerl, S. Lohmann, S. Lange, T. Ertl, Word cloud explorer: text analytics based on word clouds, in: 2014 47th Hawaii International Conference on System Sciences, IEEE, 2014, pp. 1833–1842.
- [22] H. Hu, M. Liao, C. Zhang, Y. Jing, Text classification based recurrent neural network, in: 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), IEEE, 2020, pp. 652–655.
- [23] Y. Liu, J. Ma, Y. Tao, L. Shi, L. Wei, L. Li, Hybrid neural network text classification combining tcn and gru, in: 2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE), IEEE, 2020, pp. 30–35.
- [24] Y. Zhang, Z. Rao, n-bilstm: bilstm with n-gram features for text classification, in: 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), IEEE, 2020, pp. 1056–1059.
- [25] G. Xuyang, Y. Junyang, X. Shuwei, Text classification study based on graph convolutional neural networks, in: 2021 International Conference on Internet, Education and Information Technology (IEIT), IEEE, 2021, pp. 102–105.
- [26] E. Saad, S. Din, R. Jamil, F. Rustam, A. Mehmood, I. Ashraf, G.S. Choi, Determining the efficiency of drugs under special conditions from users' reviews on healthcare web forums, *IEEE Access* 9 (2021) 85721–85737, <https://doi.org/10.1109/ACCESS.2021.3088838>.
- [27] R. Jamil, I. Ashraf, F. Rustam, E. Saad, A. Mehmood, G.S. Choi, Detecting sarcasm in multi-domain datasets using convolutional neural networks and long short term memory network model, *PeerJ Comput. Sci.* 7 (2021) e645, <https://doi.org/10.7717/peerj-cs.645>.
- [28] M. Khalid, I. Ashraf, A. Mehmood, S. Ullah, M. Ahmad, G.S. Choi, Gbsvm: sentiment classification from unstructured reviews using ensemble classifier, *Appl. Sci.* 10 (8) (2020) 2788, <https://doi.org/10.3390/app10082788>.
- [29] V. Rupapara, F. Rustam, H.F. Shahzad, A. Mehmood, I. Ashraf, G.S. Choi, Impact of smote on imbalanced text features for toxic comments classification using rvc model, *IEEE Access* 9 (2021) 78621–78634, <https://doi.org/10.1109/ACCESS.2021.3083638>.
- [30] V. Rupapara, F. Rustam, A. Amaar, P.B. Washington, E. Lee, I. Ashraf, Deepfake tweets classification using stacked bi-lstm and words embedding, *PeerJ Comput. Sci.* 7 (2021) e745, <https://doi.org/10.7717/peerj-cs.745>.
- [31] M.S. Rehan, F. Rustam, S. Ullah, S. Hussain, A. Mehmood, G.S. Choi, Employees reviews classification and evaluation (ERCE) model using supervised machine learning approaches, *J. Ambient Intell. Humaniz. Comput.* (2021), <https://doi.org/10.1007/s12652-021-03149-1> 1–18.