*Grenze International Journal of Engineering and Technology, January Issue*

# An Enhanced Deep Convolutional Neural Network for YouTube Video Categorization using Textual Metadata

Vatsal Chudasama[1], Harshada Gawas[2], Steffi Peter Raj[3], Samuel Roy[4] and Sushma Nagdeote[5]
[1-5]Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Mumbai, India
Email: sushman@fragnel.edu.in, crce.10399.ce@gmail.com

***Abstract*— The rapid growth of online video platforms like YouTube has triggered an urgent need for the efficient categorization of a huge volume of videos. This paper proposes an approach to classify YouTube videos using only textual metadata like title and description using a Deep Convolutional Neural Network model. In this work, the model was trained on 20,000 videos in nine categories, where the preprocessing, tokenization, and one-hot encoding were used for better feature representation. Experimental results have shown that the DCNN outperforms RNNs, GRUs, and classic machine learning classifiers, reaching an accuracy of 96% and a ROC-AUC of 0.99. By focusing only on textual cues, the method proposed enjoys lower computing complexity compared with vision-based algorithms while retaining its predictive power.**

***Index Terms*— Deep Learning, YouTube Video Classification, Convolutional Neural Network, Natural Language Processing, Text Features.**

## I. INTRODUCTION

Traditional methods for classifying videos mainly lean on either analysis of the video frames or audio-based feature extraction, both of which are computationally expensive. On the other hand, text-based classification methods depend on metadata like titles, descriptions, and user-tagged labels of the videos, which are easily available and often quite indicative of the content. Textual classification achieves comparable accuracy at a fraction of the computational cost and hence provides an efficient and scalable alternative for large-scale systems.

Recent breakthroughs in the field of NLP have shown that deep learning architectures, especially CNNs, are able to efficiently model both complex semantic and syntactic dependencies inside a text sequence. Even though their original design was targeted at image processing tasks, the satisfactory results resulting from using CNNs have been achieved for classification tasks regarding text, sentiment analysis, and document categorization. This paper presents an enhanced Deep Convolutional Neural Network architecture that is optimized for YouTube video categorization by textual metadata. The model processes video titles and descriptions through a structured pipeline of data cleaning, tokenization, feature encoding, and embedding be- fore multi-layered convolutional learning. It further presents performance evaluation of the proposed model against the benchmark set of RNNs, GRUs, and other classical machine learning algorithms that include logistic regression, support vector machines, decision trees, and random forests. Experimental results in this work show that the proposed DCNN outperforms all the comparative models on key metrics of evaluation; hence, it can be considered robust, efficient, and scalable in classifying various categories of YouTube video data.

*Grenze ID: 01.GIJET.12.1.170*

## II. Related Work

Automated video categorization and metadata analysis have been extensively explored within machine learning and deep learning research. Previous research on video categorization spans traditional ML, DL, multimodal systems and metadata based methods. Prior work explored content-based classification using audio or visual attributes but these methods were computationally expensive. Recently, deep neural architectures such as CNN, RNN and GRUs were adopted for semantic extraction with EfficientNet, MobileNet and Transformer based systems for improving accuracy across different domains such as biomedical videos, agriculture and sentimental analysis. Covington et al. introduced a large-scale deep neural recommendation system for YouTube that significantly improved personalization accuracy [1]. Scherer et al. advanced semantic metadata extraction using dense video captioning [2], while Choi and Heo demonstrated the effectiveness of CNNs for specialized biomedical video classification tasks [3].

Lightweight CNN models have also shown strong potential. Huong et al. optimized MobileNetV2 for agricultural quality assessment, achieving effective results in resource-limited settings [4]. Park reviewed deep learning applications for speech and audio classification, emphasizing the superiority of DNNs over traditional approaches [5]. Mao et al. further highlighted recent advances in video classification, especially the emergence of transformer-based models [6].

Several studies have focused specifically on YouTube meta- data for text-driven classification. Ramalingam used sentiment- based metadata analysis for categorizing YouTube videos [7]. Raghunadha Reddy et al. applied NLP-driven deep learning methods to improve metadata generation and classification [8]. Hussaini and Paul demonstrated that deep learning models handle noisy YouTube metadata more effectively than classical techniques [9]. Conversely, Kalra et al. showed the limitations of Random Forests for title–description classification when hyperparameters are not optimized [10].

Earlier machine learning work, such as that of Huang et al., combined lexical and syntactic features with SVMs and Decision Trees, achieving moderate performance but facing scalability issues [11]. Modern deep learning methods, includ- ing those by Yousaf and Nawaz, used EfficientNet architec- tures for inappropriate content detection with high accuracy [12]. Cunha et al. applied CNNs for sentiment analysis of YouTube comments [13], while Tang demonstrated that neural text models effectively capture authorship cues [14].

Growing research on multilingual and multimodal classifi- cation has further expanded the field. Das et al. showed that hybrid deep learning methods outperform traditional models for multilingual sentiment analysis [15]. Wang et al. proposed a multimodal architecture that combines text, audio, and visual features, achieving significantly higher video categorization accuracy than the state-of-the-art systems at that time [16]. Singh and Kumar concluded that CNNs outperform RNNs for short text classification due to superior local feature extraction capabilities [17]. Li et al. provided an empirical analysis of YouTube content and encoding trends over several years, offering insights beneficial for metadata-driven classification tasks [18]. More recently, large datasets and fusion mecha- nisms continue to improve overall system performance. Abid's dataset remains one of the most widely used resources for supervised YouTube metadata classification [19].

Lee et al. introduced an attention-based multimodal sentiment classifier that effectively captures cross-modal dependencies [20]. Zhang et al. showed that deep learning mod- els significantly outperform traditional recommendation tech- niques [21]. Alharbi and Benaida reviewed state-of-the-art text classification methods and identified CNNs and transformers as leading architectures [22]. Otter et al. emphasized the importance of sequence modeling in NLP [23], while Mikolov et al. provided the foundational Word2Vec embedding model widely used in text-based classification systems [24].

In general, the analyzed literature shows a clear trend from classical machine learning to deep neural architectures like CNNs, RNNs, and transformers for scalable video and meta- data analysis in an accurate manner, justifying the motivation behind the DCNN-based methodology presented in this work. Most of the research focuses on YouTube metadata showing textual elements like titles, comments and description indicating efficient classification. SVM, Decision Trees and NLP based DL achieved moderate to strong performance though many models struggled with noise, limited contextual understanding and scalability. Recently multimodal fusion techniques combining audio, visuals and text were introduced which significantly enhanced classification. Related work on DL for NLP and text classification has consistently identified CNNs and transformer based models as dominant architectures due to their ability to capture semantic patterns. Overall the related work indicates a evident shift from traditional models to DNN and multimodal system. This highlights the need for efficient and scalable text based models. This inclination supports the motivation for the proposed DCNN which targets to achieve high accuracy using YouTube textual metadata which in turn reduces the computational cost.

## III. Dataset and Pre-processing

### *A. Dataset Description*

The dataset used in this study was obtained from the publicly available YouTube API Data for Text Categorization repository on Kaggle [19]. It consists of 20,000 videos divided into nine categories: Adventure, Art & Music, Food, History, Manufac- turing, Nature, Science & Technology, Sports, and Travel. Each entry contains a video title, description, and a categorical label provided through YouTube's official API.

### *B. Text Preprocessing*

An exhaustive pre-processing methods were applied to YouTube data. The pre-processing such as Normalization, Tokenization, filtering stopword, lemmatization etc. is applied to the data. Common normalization steps includes lowercasing, removing punctuations, removing URLs, special characters from the textual data. Tokenization split text into words using NLTK. Noise removal removes numbers, extra white space and non language symbols. The high frequency occurring words are removed from the data. The next step is Lemmatization which reduces the words to their canonical form.

### *C. Exploratory Data Analysis (EDA)*

The dataset's statistical analysis revealed that each video description comprises an average of 13 words with a lexical richness score of 0.93, indicating diversified textual content. The most often used terms were "video", "subscription", and "channel"[19].

## IV. Model Architecture

This section explains a complete framework for YouTube video classification using textual metadata clearly outlining each stage. It begins by describing the dataset of 20000 videos across nine categories followed by detailed pre-processing steps namely Normalization, tokenization, noise removal, filtering stopword and lemmatization ensuring clean text input. This section dives into the in depth architecture of DCNN including embedding layers, convolution-pooling stacks and dense layers that extracts and learn semantic patterns from description and titles of YouTube Videos. Training details such as use of TensorFlow, 80/20 train-test split, training epochs and evaluation metrics like ROC-AUC and accuracy are provided. The results and discussion section further clarify the comparative strength of proposed model over RNN, GRU and state-of-art ML models supported by confusion matrix insights. This paper presents a step by step well structured explanation that enhances the understanding and reliability of the proposed method. The DCNN model has an initial feature embedding layer that transforms input text sequences into dense vector representations. These embeddings are fed through two convolutional-max pooling stacks; each convolutional layer captures local semantic patterns, and subsequent pooling layers reduce spatial dimensions to retain only the most relevant features while suppressing noise. Then, the flattened output of the final pooling layer feeds into a dense layer to learn higher-order feature interactions, and finally, an output dense layer projects class probabilities using the softmax activation function. To optimize the performance on the text classification task, the model is trained with the categorical cross-entropy loss function, then optimized with the Adam optimizer across ten epochs.

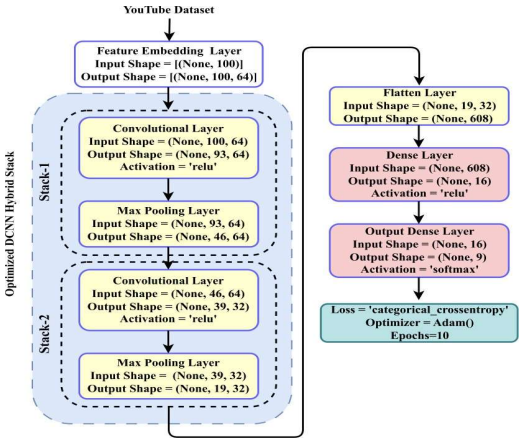


Fig. 1: Overview of the proposed DCNN model architecture

*A. Baseline Models*

- RNN: Process sequential data using hidden layers preserving contextual memory, usually prone to gradient vanishing issues.
- GRU: An improvement over RNN, GRU incorporates various gates to keep track of long-term dependencies but with lesser computation.

*B. Proposed DCNN Model*

The DCNN architecture consists of the following layers:

- Embedding Layer: This layer maps input words into 64-dimensional vectors.
- Convolutional Layers: With 64 and 32 filters, respectively, extract hierarchical local features.
- Pooling Layers: Perform 1D max pooling for dimensionality reduction.
- Flatten Layer: It flattens the feature maps into a single vector.
- Fully Connected Layers: Added one ReLU-activated dense layer containing 16 neurons and followed by a softmax output layer containing 9 neurons.

It serves to parallelize learning of the spatial relationships of text, providing both interpretability and efficiency.

## V. Training and Evaluation

*A. Experimental Setup*

All experiments in this study were done using TensorFlow 2.8 and Python 3.7 on an Intel Core i5 CPU work-station with 15 GB RAM and an NVIDIA Tesla K80 GPU. The models were trained for 10 epochs; 80% of the dataset was used for training, while the remaining 20% was used for testing.

*B. Evaluation Metrics*

The performance of the models was assessed using standard metrics:

- Accuracy (Acc): It refers to the overall percentage of correctly classified instances.
- Precision (P): It is the ratio of correctly predicted positive observations to the total predicted positives.
- Recall (R): It is the ratio of correctly predicted positive observations to all actual positives.
- F1-Score (F1): Provides the harmonic mean for Precision and Recall, balancing both metrics.
- ROC–AUC: This assesses the model's discriminative ability across many categorization levels.

*C. Performance Analysis*

The DCNN model performed much better than the GRU model, at 91%, and the RNN model at 87%, with training and testing accuracy of 99% and 96%, respectively. Also, the score of ROC-AUC was 0.99, which underlined the great discriminative power of the proposed model along with its robustness while dealing with text-based YouTube video data.

## VI. Implementation

*A. System Implementation*

We implemented the proposed DCNN model in TensorFlow 2.0 and Keras for maximum flexibility and efficient usage of graphics processing units. We conducted all the experiments on a single workstation consisting of an Intel Core i5 Processor (1.80 GHz), 16 GB RAM, and an NVIDIA Tesla K80 GPU. All data preparation, tokenization, and performance analyses were done in Python 3.7, along with essential NLP libraries such as NLTK, scikit-learn, and NumPy. The model development pipeline was designed in a modular fashion, and there were five major components:

- Data Loader Module: This fetches the YouTube meta- data dataset with titles, descriptions, and categories, then splits the dataset into training and testing sections.
- Preprocessing Unit: All text normalization, tokenization, lemmatization and encoding are done before feeding it into the model.
- Model Builder: Defines and creates the deep CNN architecture, including the embedding dimensions, convolution filters, and dense layer sizes.
- Trainer Module: Supports iterative learning with call-backs for early stop, prevents overfitting using dropout regularization and learning rate scheduling.
- Evaluator and Logger: This calculates the classification metrics while saving model checkpoints and logs for further analysis.

The model was trained with categorical cross-entropy as the loss and optimized by the Adam optimizer. Training lasted for ten epochs with a batch size of 64, and validation accuracy at every iteration was tracked. After about six epochs the network steadily converged, maintaining stability with high generalization performance on unseen data.

## VII. Results and Discussion

### *A. Comparative Analysis*

It basically improves the performance of the proposed DCNN due to its optimized convolutional filters and feature embedding layers that model the contextual semantics of textual data effectively. Unlike recurrent models in which sequences are processed sequentially, the CNN architecture extracts features of multiple levels of semantic information in parallel and reduces the risk of overfitting, accelerating the convergence.

Conventional machine learning models like Logistic Regression and Support Vector Machines (SVM) have achieved the best performances with an accuracy of 95% using TF- IDF-based features. However, these models lacked the ability to capture contextual dependencies and long-range linguistic patterns, which diminished their classification performance compared to deep learning architectures.

### *B. Confusion Matrix Insights*

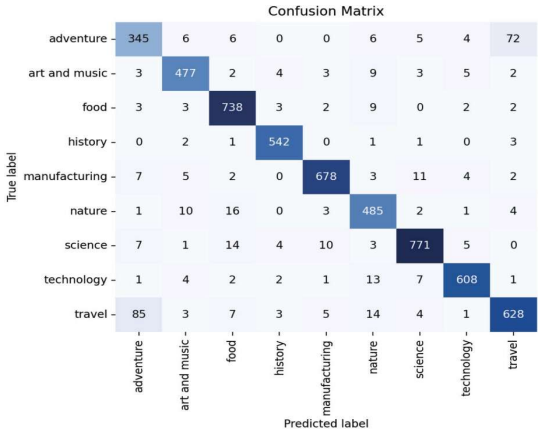


Fig. 2: Confusion Matrix for the proposed DCNN

TABLE I. 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 | 5,712 |

Category-level testing returned the highest precision in the class History with a score of 0.99, whereas the Adventure category had slightly lower accuracy, 0.89, which could be due to imbalance in sample distribution. Confusion matrix analysis confirmed that the DCNN correctly classified 5,272 out of 5,712 instances, with the best predictive power compared with all other models under test.

### *C. Practical Implications*

The proposed DCNN model has strong potential for integration into large-scale video recommendation and content management systems. Applications include:

- Automated Video Tagging: The video content is tagged intelligently using metadata.

- Enhanced Ad Targeting: Placing advertisements can be done in a more targeted manner by attributing videos to appropriate audience categories.
- Improved User Engagement: This can enable recommendation algorithms to provide highly relevant information, leading to increased user happiness and retention

## VIII. LIMITATION

The recommended DCNN is employed for categorizing videos on YouTube enhances accuracy and scalability for textual data. Similar to other data-driven systems, the existing approach has methodological and practical limitations. These limitations can be found in a way to guide future improvements in generalization, interpretability, and multimodality. The proposed method performance depend on descriptions and titles which are mostly incomplete. The dataset used in this research consists of nine classes which restricts the generalizability of the model. The samples are unevenly distributed across categories which reduce the learning ability of the model and resulting in weaker performance.

## IX. CONCLUSIONS AND FUTURE WORK

This study presents an effective DL methodology for classifying YouTube videos using textual metadata. Compared to the RNN and GRU architectures and conventional ML approaches, the proposed DCNN significantly escalate the classification accuracy maintaining low computational cost. This methodology eliminates the need high end image or audio processing and depends only on textual metadata. In future use of advanced language models such as BERT and RoBERTa can provide richer semantics representation which enhances the classification on textual metadata. For robust video classification the fusion of audio, visuals and text attributes can be used. The proposed method can be expanded to support multiple languages to adapt to evolving nature of online media.

## REFERENCES


[1] P. Covington, J. Adams, and E. Sargin, ”Deep neural networks for YouTube recommendations,” Proc. 10th ACM Conf. Recommender Syst. (RecSys), Boston, MA, USA, pp. 191–198, 2016.

[2] J. Scherer, D. Bhowmik, and A. Scherp, ”Semantic metadata extraction from dense video captioning,” arXiv preprint arXiv:2211.02982v1, 2022.

[3] W. Choi and S. Heo, ”Deep learning approaches to automated video classification of upper-limb tension test,” Healthcare, vol. 9, no. 11, Art. 1579, Nov. 2021.

[4] P. T. Huong, L. T. Hien, N. M. Son, H. C. Tuan, and T. Q. Nguyen, ”Enhancing deep convolutional neural network models for orange quality classification using MobileNetV2 and data augmentation techniques,” J. Algorithms Comput. Technol., 2025.

[5] H. Park,”Deep neural networks-based classification methodologies of speech, audio and music processing,” J. Wavelets Multiresolut. Inf. Process., vol. 20, Art. 18979, 2023.

[6] M. Mao, A. Lee, and M. Hong, ”Deep learning innovations in video classification: A survey on techniques and dataset evaluations,” Electronics, vol. 13, no. 14, Art. 2732, 2024.

[7] S. Ramalingam, ”Metadata extraction and classification of YouTube videos using sentiment analysis,” in Proc. IEEE Int. Carnahan Conf. Security Technol. (ICCST), Orlando, FL, USA, Oct. 24-27, 2016.

[8] T. R. Raghunadha Reddy, P. Sreekari, J. N. Kumar Reddy, and V. Jyothsna, ”A deep learning approach for video metadata generation and classification,” Int. J. Innov. Res. Technol., vol. 9, no. 12, May 2023.

[9] S. O. Hussaini and P. John Paul, ”Identify and categorize YouTube videos using deep learning and machine learning,” J. Eng. Sci., vol. 16, no. 6, pp. 755–761, 2025.

[10] G. S. Kalra, R. S. Kathuria, and A. Kumar, ”YouTube video classification based on title and description text,” in Proc. Int. Conf. Comput., Commun. Intell. Syst. (ICCCIS), IEEE, pp. 1–6, 2019.

[11] C. H. Huang, T. C. Fu, and H. H. Chen, ”Text-based video content classification for online video-sharing sites,” J. Am. Soc. Inf. Sci. Technol., vol. 61, no. 5, pp. 891–906, 2010.

[12] K. Yousaf and T. Nawaz, ”Deep learning-based inappropriate content detection in YouTube videos,” IEEE Access, vol. 10, pp. 67638–67649, 2022.

[13] A. A. L. Cunha, V. H. Alves, and J. C. Junior, ”Sentiment analysis of YouTube video comments using deep neural networks,” in Int. Conf. Artif. Intell. Soft Comput., pp. 114–125, 2019.

[14] X. Tang, ”Author identification of literary works based on text analysis and deep learning,” Heliyon, vol. 10, no. 3, pp. e25486, 2024.

[15] R. Das, A. Dutta, and S. Roy, ”Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models,” Heliyon, vol. 9, no. 9, pp. e20333, 2023.

[16] L. Wang, M. Zhao, and Q. Xu, ”Automated video classification using multimodal deep learning techniques,” IEEE Trans. Multimedia, vol. 24, no. 5, pp. 1392–1403, 2022.
[17] N. Singh and P. Kumar, ”Comparative study of CNN and RNN for text-based video categorization,” Procedia Comput. Sci., vol. 167, pp. 2358–2367, 2020.
[18] F. Li, X. Cheng, and Y. Zhao, ”Three-year trends in YouTube video content and encoding characteristics,” in Proc. SIGMAP, pp. 102–110, 2021.
[19] A. Abid,”YouTube API data for text categorization,” Kaggle Dataset, 2020. [Online]. Available: https://www.kaggle.com/datasets/anisabid/ youtubeapi-text-categorization
[20] J. Lee, S. Kim, and J. Choi, ”Multimodal sentiment classification of YouTube videos using attention-based fusion,” IEEE Access, vol. 9, pp. 121780–121791, 2021.
[21] Y. Zhang, R. Wang, and C. Li, ”Deep neural architectures for online content recommendation,” Inf. Sci., vol. 608, pp. 459–471, 2022.
[22] S. Alharbi and M. Benaida, ”A review on text classification approaches using deep learning,” Expert Syst. Appl., vol. 195, pp. 116570, 2022.
[23] D. Otter, J. Medina, and J. Kalita, ”A survey of the usages of deep learning for natural language processing,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 2, pp. 604–624, 2020.
[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ”Distributed representations of words and phrases and their compositionality,” in Adv. Neural Inf. Process. Syst., pp. 3111–3119, 2013