A Deep Learning Approach for Video Metadata Generation and Classification

Dr. T. Raghunadha Reddy[1], P. Sreekari[2], J. Nikhil Kumar Reddy[3], and V. Jyothsna[4]
[1]*Associate Professor, Department of CSE, Matrusri Engineering College, Hyderabad*
[2, 3, 4] *Student, Department of CSE, Matrusri Engineering College, Hyderabad*

***Abstract*— Video information is one of the most emerging and easy ways to learn and know about anything that is going on around the world. On the internet, video material has grown in popularity influencing many parts of our life including education, entertainment, and communication. Video content is one of the most attractive ways where each and every individual attracts to the pictographic and visualization of the content which helps in easy understanding and gain of knowledge. YouTube is the prime source for generating and classifying the videos and is considered as of the most entertaining media where worldwide information is present. The main objective of this article is to generate and classify video content into different categories. We consider videos from YouTube which has subtitles. The primary goal is to extract and categorize information from videos. The procedure entails using Natural Language Processing (NLP) to extract text that may contain unwanted characters or symbols, necessitating text cleaning. The NLP is basically for analysing the relevant information. To extract important information from text, certain text pre-processing techniques such as tokenization and stemming need to be applied on the text. In this work, The YouTube URL is copied and uploaded to the front-end web page. After the URL is uploaded, the NLP process generates the subtitles of text by using the dataset which is considered for this work. The dataset contains a CSV file which contains the records where the data is pre-processed and keywords are generated. Once the pre-processing is done, the summary is generated from the retrieved text and categorized based on keywords and synonyms. The entire process is sent to the LSTM model to train and test the model for accurate output. Users can provide URLs, and the system will create a summary that is categorized appropriately. To give an interactive web-based output, this article is incorporated into the Flask framework.**

***Index Terms*— LSTM, Flask Framework, YouTube Videos, Natural Language Processing.**

I. INTRODUCTION

Information from video footage and categorising it based on specified criteria such as title, description, creator, creation date, duration, and tags is one important task. This is often accomplished by analysing and extracting data from video using automated techniques such as machine learning algorithms and computer vision technologies. After that, the retrieved metadata is saved in a structured format for simple access and classification. Video classification is the process of categorising video footage based on criteria such as genre, subject, audience, or language. It can be done manually or automatically using machine learning algorithms that analyse the video content and extract pertinent characteristics for classification. Overall, these activities are critical for effective video management and user experience enhancement.

In this article, we proposed a method to extract the subtitles from which we generate the keywords. The text extraction is taking place through NLP process. The major goal is to categorise data gathered from videos. To do this, we use NLP to extract text by obtaining the YouTube subtitle file using the transcript function. We next clean up the text by deleting any undesired symbols or letters. To extract relevant information, we may need to analyse the text further by doing tasks like tokenization, stemming, or entity recognition, depending on the unique use case.

The text has been extracted from the video subtitles, it is classified by categorizing the generated summary. The classification is performed based on keywords and synonyms derived from the summary. The entire project is integrated to present the interactive web-based output through Flask framework. Where the URL is uploaded, the summary is generated and classified.

This article is planned in 6 sections. Section 2 discuss about the existing works proposed for video classification. Section 3 explains the characteristics of dataset. The proposed method and components used in the proposed method are described in section 4. The experimental results are presented and explained in section 5. The section 6 mentions the conclusions of this article and possible future enhancements.

II. LITERATURE SURVEY

This section presents a review of the current research on video metadata generation and classification, which employs a range of machine learning and deep learning techniques. The research draws upon various data sources, including pre-existing datasets available on platforms like Kaggle, in order to provide a comprehensive analysis of the field.

Shweta Bhardwaj et al., explained [1] the fewer frames concept. Where, the entire video is divided into each frame with a specific time slot. Where, two datasets are compared by different algorithms and mathematical calculations to reduce the time for each frame. The paper introduces a unique approach that uses the notion of distillation to minimize the computing time necessary for video categorization. Training a teacher network, which constructs a video representation using all frames of the movie, is followed by training a student network, which analyses just a certain number of frames (k). Different loss function combinations are used to ensure that the student network's final representation and output probability distributions are similar to those of the teacher network. The suggested models are compared to a strong baseline and a skyline, with the results demonstrating that the proposed technique outperforms the baseline and offers a considerable decrease in computing time and cost when compared to the skyline. The method's success is illustrated on the YouTube-8M dataset, where the computationally less costly student network may cut calculation time by 30% while outperforming the instructor network.

"Metadata extraction and classification of YouTube videos using sentiment analysis", the research paper explains [2] about the classification of video metadata can also be done by extracting and analysing from a video followed by the decision-making of the content. Categorizing the video data into different titles which include the subject, URL of the video, Time limit of the video, video type, description, caption, etc., is an important task. The information is segregated into different datasets to predict the accuracy of positive, negative, and neutral videos which thus calculates the rating of the video. The paper describes a method for evaluating YouTube video URLs that entails using sentiment analysis to determine the polarity of the video objects. The correctness of the entire process, however, is dependent on the Python inbuilt dictionary Corpus, which consists of a collection of terms and their accompanying scores. The authors personally added new terms and their scores to the Corpus to evaluate its enhancement. As the dataset expands, automation will be required to make this process more efficient. Manually updating the Corpus dictionary would be time-consuming and inefficient. We suggest leveraging Machine Learning principles such as Neural Networks, Genetic Algorithms, SVMs, and Bayesian Learning to automate the process.

"Large-scale Video Classification with Convolutional Neural Networks" [3], the current study looks at the effectiveness of convolutional neural networks (CNNs) in large-scale video classification. According to the findings, CNN architectures may successfully learn potent features from poorly labeled data, outperforming feature-based approaches in terms of performance. Deep convolutional neural networks (CNNs) were first used for video categorization in this landmark article. CNNs may be trained to recognize complicated visual patterns in films, such as object motion and scene changes, according to the authors. They also demonstrate that their model outperforms hand-crafted feature extraction approaches. Surprisingly the intricacies of the architecture's connectedness in time have no influence on this enhancement. A qualitative study of network outputs and confusion matrices reveals identifiable faults. Furthermore, the study demonstrates that a Slow Fusion model consistently outperforms early and late fusion alternatives. Furthermore, the study shows that a single-frame model can already achieve high performance, implying that local motion cues may not be necessary, even for dynamic datasets like Sports. Transfer learning tests on UCF-101 show that the taught traits are general and generic features, investigate approaches that consider camera motion, and explore the use of recurrent neural networks to

combine clip-level predictions into global video-level predictions.

The article "Automated Metadata Generation for Video Content" discusses [4] a novel approach to automated video metadata generation. The proposed method involves a multi-task deep learning framework that combines different sources of information, including visual and textual features, to generate metadata for video content. The authors conducted experiments on a large-scale dataset and compared the performance of their approach with several state-of-the-art methods. The results demonstrate that the proposed method outperforms existing methods in terms of accuracy, efficiency, and scalability. The study provides insights into the potential of using deep learning techniques for automating the process of video metadata generation, which can help content creators and providers to organize, search, and discover video content more effectively. To automatically extract information from films and provide appropriate metadata, the authors use a combination of technologies such as computer vision, natural language processing, and machine learning. They employ object identification algorithms to recognize things in the video, voice recognition to transcribe audio information, and natural language processing to extract keywords and subjects. The extracted data is then utilized to build metadata for the movie, such as the title, description, and tags. The suggested technique is assessed and compared to current methods on a dataset of movies, yielding encouraging results in terms of accuracy and efficiency.

"Sentiment Analysis on YouTube: A Brief Survey" [5], Researchers are still struggling with the categorization of general events and identifying the sentiment polarity of user comments on YouTube. Despite tremendous progress, there is still a long way to go in tackling this issue. This study outlines numerous challenges that must be addressed in order to assess the polarity of comments submitted by YouTube users, such as the limits of current sentiment dictionaries, users' informal language patterns, and sentiment estimates for community-created terminology. Proper event labeling and obtaining acceptable classification performance have also been cited as important problems in social media sentiment analysis. The study explores numerous strategies for determining the polarity of comments, such as User Sentiment Detection, Event Classification, and Predicting YouTube Comments. The authors suggest that future studies should focus on enhancing the social lexicon and statistically verifying it, as well as correct event categorization, to improve the effectiveness of forecasting comment ratings.

"Video description: A comprehensive survey of deep learning approaches " [6], this survey paper provides a comprehensive overview of deep learning-based video classification techniques. The authors discuss various deep learning architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms. They also review recent advances in video classification, such as temporal modeling and multi-modal fusion. This is a survey work that covers and analyses several deep learning-based video classification approaches. The study discusses several deep learning models for video categorization, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and modifications such as 3D CNNs, temporal CNNs, and attention-based models. The study also examines several pre-processing approaches that may be used to improve the performance of video classification models, such as data augmentation, feature extraction, and data normalization. It also goes through the various datasets and assessment criteria that are utilized for video classification jobs. Furthermore, the paper discusses transfer learning techniques and how they can be used to improve video classification performance by leveraging pre-trained models. It also explores how to enhance video categorization results by using ensemble learning and active learning approaches. Overall, the study provides a thorough assessment of deep learning-based video categorization approaches as well as insights into the field's current state-of-the-art methodologies.

"Video Classification: A Literature Survey" [7], at present, so many videos are available from many resources. But viewers want videos of their interest. So for users to find a video of interest work has started for video classification. Video Classification literature is presented in this paper. There are mainly three approaches by which the process of video classification can be done. For video classification, features are derived from three different modalities: Audio, Text, and Visual. From these features, classification has been done. The video categorization literature was evaluated, and a variety of techniques,

including audio-based, text-based, and visual-based approaches, were discovered. Each technique has unique characteristics that were explored in the study. Researchers employed these methodologies separately and in varied combinations based on the application needs, resulting in improved classification accuracy.

"Video Captioning Using Deep Learning Approach-A Comprehensive Survey" [8], the process of creating natural language sentences to reflect the contents of a video is known as video captioning. Deep learning-based approaches are frequently employed for this purpose, and research in this area has made tremendous progress in recent years. This survey study offers a thorough evaluation of current research, benchmarking approaches, and training datasets. It also goes through the most popular neural network variations for feature extraction and language synthesis. ResNet and VGG are common visual feature extractors, according to the report, whereas 3D convolutional neural networks are extensively employed for spatiotemporal feature extraction. The most prevalent language model is LSTM. For video description evaluation, metrics like as BLEU, ROUGE, METEOR, CIDEr, SPICE, and WMD are often utilized.

L. N. Abdullah et al., proposed [9] a research idea to categorize online videos based on their information. Web video metadata was extracted and stored in a database for classification to accomplish this. For categorizing the films, the Random Tree and J48 classification algorithms were used. When the results of both models were examined, it was discovered that the Random Tree classification model was more successful at classifying online movies based on metadata. The Random Tree classification findings were subjected to a category-wise cost/benefit analysis, and the lowest and maximum cost/benefit values were obtained using classification accuracy. However, due to insufficient web metadata, the classification process encountered difficulties, resulting in only 79% of tuples being classified, with 21% being ignored by both the Random Tree and J48 classification models. Furthermore, the J48 classification model was found to be less efficient in partitioning web video categories based on numeric independent attributes.

"Metadata generation process for video action detection" [10], the goal of this research is to develop a model for creating multidimensional metadata in order to recognize human actions in films. To recognize and categorize activities across many modalities such as audio, motion, colour, and edge, the model employs a semantic approach. The model filters the data and extracts just the relevant and helpful aspects for the framework's succeeding modules. A representation of an action's multimodal behaviour is created by merging input from each scenario. The research focuses on the metadata creation process, which includes video capture, feature extraction, activity detection, inference, and presentation level. The suggested approach's performance is assessed using accuracy and recall criteria.

III. DATASET CHARACTERISTICS

The dataset is divided into 3 labels headlines, article and category. The entire dataset consists of 4,816 records [11]. The dataset is taken from the pre-existing website form the Kaggle. The dataset 1st column is the headline of the youtube we can consider as the title of the video. The 2nd column consists of article where the entire subtitles summary is present where further it is used for pre-processing. The 3rd column consists of the main classification the category section, after the summary is generated based upon the keywords the summary is classified into different categories. The categories include automobile content, news, sports, entertainment, technology, politics, world, science.

IV. PROPOSED METHODOLOGY

Video metadata generation and classification play a critical role in the structured generation of text and videos. The foundational technique for generating text using NLP involves analysing text to extract relevant information. To employ NLP for text extraction, it is first necessary to identify the text source, such as a website, document, or social media platform, and collect the data in a format suitable for NLP algorithm processing. This data collection may involve web scraping, using APIs, or manual data entry. After data collection, preprocessing is required to remove noise and irrelevant information such as stop words, punctuation, and HTML tags. Tokenization of the text into individual words or phrases and normalization of the text's case and format is then possible. The extraction of features from the text involves the identification of relevant information such as named

entities, keywords, topics, and sentiment. Lemmatization is a technique for obtaining a word's basic or dictionary form, known as its lemma, while keeping its intended meaning. It entails analysing a word's part of speech and doing morphological analysis to determine its root form. The produced lemma is the standard form of the term and may be used in a variety of natural language processing activities such as machine translation and text analysis. Lemmatization, as opposed to stemming, which cuts word ends to determine the root form, provides a more exact and informed strategy for word normalisation based on linguistic considerations. The pre-processing of a dataset using text extraction techniques enables the extraction of relevant information, followed by the classification of subcategories. The pre-processed data is then input into an LSTM model for prediction and accuracy detection. After prediction, the generated summary is classified into different categories. The architecture of proposed method is represented in Figure 1.

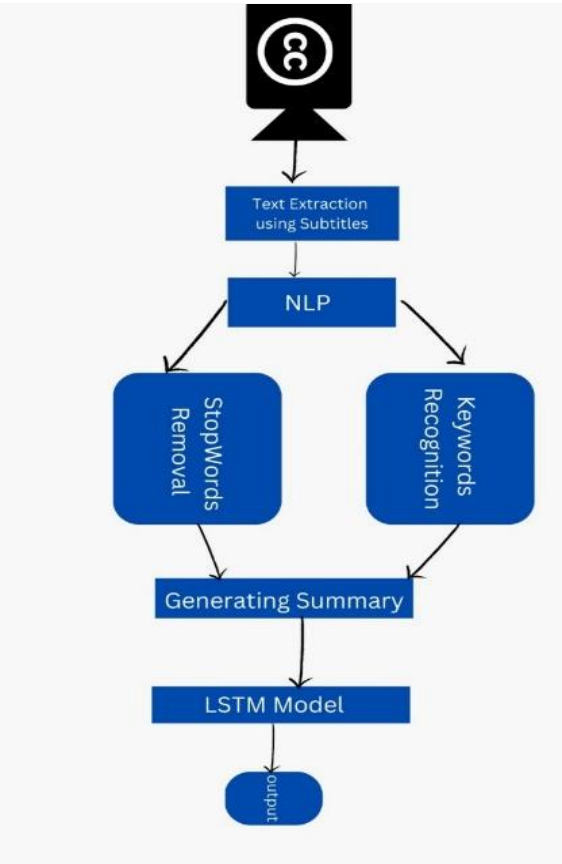

Fig. 1 The proposed method architecture

The basic architecture explains about where the video is uploaded the text extraction using subtitles is generated using NLP algorithm. Using NLP we pre-process a dataset with text extraction techniques lemmatization, stop words removal, keyword recognition entails collecting useful information and identifying subcategories. This pre-processed data is then loaded into an LSTM model to predict and identify accuracy. Following the prediction, the resulting summary is categorised into several groups. Once the data is pre-processed sent into LSTM model the output is generated in the format of web application. The connection of user and server is done with the help of Flask framework. The front end application is done with the UX/UI design, web applications where the URL is uploaded and paste is the search box where summary is generated and classification is done.

A. LSTM Model

In the field of deep learning, the long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture that is widely used. What sets LSTM apart from standard RNNs is the presence of "memory cells" that can store information over extended periods of time. Additionally, LSTM features three gates that regulate the flow of information into and out of the memory cells: the input gate, the forget gate, and the output gate. LSTM networks have been employed for various tasks, such as speech recognition, language modelling, and machine translation. In more recent times, they have also been utilized for general sequence learning tasks, including activity recognition and music transcription. To ensure the accuracy of a machine learning model, relevant information is extracted and used for training. The initial step involves preparing the text data for input into the LSTM model, which includes text cleaning and tokenization. The LSTM model is then trained on a labeled dataset, where each input sequence is associated with a label or category. Once trained, the LSTM model can be utilized to generate accurate predictions on new text data. To assess the model's performance, it is evaluated on a test dataset using metrics such as accuracy.

V. EXPERIMENTAL RESULTS

Input design is a vital aspect of information system development that establishes a critical connection between the user and the system. This involves creating specifications and procedures for transaction data preparation, either manually or through automated methods for reading data from written or printed documents. The objective of input design is to minimize input requirements, reduce errors and delays, simplify workflows, and eliminate unnecessary stages, while also prioritizing security, usability, and privacy preservation. To ensure data accuracy, input design incorporates data verification procedures, such as the use of screens and error messages to guide the data entry process and prevent user confusion. The ultimate goal is to create an intuitive input layout that is easy for users to understand, promoting data accuracy and minimizing the risk of errors.

Output design is a critical process of presenting processed information in an understandable and useful manner that meets the needs of the end-users. It involves determining how information will be displayed for immediate use and hard copy output. The main objective of output design is to enhance the system's interaction with the user and aid in decision-making. The process includes planning and designing the computer-generated information presented to users or other systems. The design must ensure that the information is clear, relevant, and structured in a user-friendly manner to facilitate the decision-making process. Therefore, output design requires careful consideration of specific requirements to effectively communicate the desired information. Our experiment attained an accuracy of 97.92% for video metadata generation and classification. The Figure 2 shows the accuracy of this proposed method.

Fig. 2 Output of Proposed Method

The Figure 3 shows the flask framework screen to request for URL of input video.

Fig. 3 The Input screen

Figure 4 shows the output screen which shows the metadata related to given input video and the category of video.

Fig. 4 The output screen

VI. CONCLUSIONS AND FUTURE SCOPE

In this work, we proposed a deep learning technique based method for video metadata generation and classification by using a technique of LSTM model. In this work, we used flask framework for creating input and output screens for better visualization to users. The proposed method attained an accuracy of 97.92% for video metadata generation and classification.

In the future, we plan to expand the testing of our model to include other video processing tasks, such as generating summaries, answering queries, and creating captions. Additionally, we aim to develop a training approach involving a group of instructors working with a student, preferably from diverse backgrounds. The future work involves extracting of information directly from uploading the video and also the text is generated through audio.

REFERENCE

[1] Efficient Video Classification Using Fewer Frames", Shweta Bhardwaj, Mukundhan Srinivasan- NVIDIA Bangalore Mitesh M. Khapra,https://doi.org/10.48550/arXiv.1902.10640

[2] Metadata extraction and classification of YouTube videos using sentiment analysis, October 2016, Conference: 2016 International Carnahan Conference on Security Technology (ICCST).
[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1725-1732, doi: 10.1109/CVPR.2014.223.
[4] Automated Metadata Generation for Video Content, X. Wang, L. Xie, and W. Zhang (2021)
[5] Sentiment Analysis on YouTube: A Brief Survey, Muhammad Zubair Asghar, Shakeel Ahmad, Afsana Marwat, Fazal Masud Kundi, 2015, https://doi.org/10.48550/arXiv.1511.09142
[6] Video description: A comprehensive survey of deep learning approaches, Ghazala Rafiq, Muhammad Rafiq & Gyu Sang Choi, Artificial Intelligence Review (2023)
[7] Pravina Baraiya, Asst. Prof. Disha Sanghani, "Video Classification: A Literature Survey". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 6, no. 3, Mar. 2018, pp. 01-05, doi:10.17762/ijritcc.v6i3.1447.
[8] Jacob, J., Devassia, V.P. (2023). Video Captioning Using Deep Learning Approach-A Comprehensive Survey. In: Sharma, H., Saha, A.K., Prasad, M. (eds) Proceedings of International Conference on Intelligent Vision and Computing (ICIVC 2022). ICIVC 2022. Proceedings in Adaptation, Learning and Optimization, vol 17. Springer, Cham. https://doi.org/10.1007/978-3-031-31164-2_7
[9] Prashant Bhat, "Metadata Based Classification and Analysis of Large Scale Web Videos", International Journal of Emerging Trends & Technology in Computer Science, June 2015
[10] L. N. Abdullah and S. A. M. Noah, "Metadata generation process for video action detection," 2008 International Symposium on Information Technology, Kuala Lumpur, Malaysia, 2008, pp. 1-5, doi: 10.1109/ITSIM.2008.4631660.
[11] https://www.kaggle.com/datasets/shashichander009/inshorts-news-data