ETDT

Emerging Trends in Digital Transformation

Volume-1, Issue-2, July 2025 ISSN:**3107-4308** *Paper ID:* ***ETDT-V1-I2-10***

AI-DRIVEN INAPPROPRIATE CONTENT DETECTION AND CLASSIFICATION IN YOUTUBE VIDEOS USING DEEP LEARNING

B.SRINIVASA RAO[1], K.NAVYA[2], D.LAXMI NAGA SREYA[3], B.SWATHI[4], S.PREM KUMAR[5]

[1]Assistant Professor, Dept. Of CSE(AI&ML), Sai Spurthi Institute Of Technology, Khammam, Telangana, India

[2,3,4,5]B.Tech Student, Dept. Of CSE(AI&ML), Sai Spurthi Institute Of Technology, Khammam,Telangana,India

ABSTRACT: Video information is one of the most convenient and contemporary methods of staying informed about current events. The popularity of video content on the internet is on the rise, and it is having a significant impact on various aspects of our lives, such as education, entertainment, and communication. Video content is one of the most captivating forms of information, as it not only captivates viewers with its visuals but also facilitates the acquisition of knowledge and comprehension. The primary resource for the development and categorization of the text that we intend to use for the endeavor is YouTube. YouTube is considered to be one of the most pleasurable methods of acquiring global information. The primary objective of our endeavor is to generate and organize video content into distinct categories. We consider YouTube videos that include translations. The primary objective is to extract and categorize data from videos. The process involves the extraction of text that may contain undesirable letters or symbols through the use of natural language processing (NLP), which necessitates text cleaning. In essence, NLP is employed to evaluate pertinent data. Special text processing methods, such as tokenization and stemming, may be necessary to derive meaningful information from text. A replica of the YouTube URL has been incorporated into the front-end web page. The entire process commences upon the URL's upload. The connected dataset is employed to produce text that includes subtitles and natural language processing (NLP), which eliminates stop words and generates keywords. The preprocessed records and keywords generated are contained in the CSV file. A summary is generated after pre-processing, and the extracted text is arranged according to synonyms and keywords. The entire process is transmitted into the LSTM model to train and validate it for precise output. The system will generate a categorized summary based on the URLs that users submit. The Flask framework is employed to produce an interactive web-based output for the project.

Index terms: Deep Learning, Content Filtering, Neural Networks, Convolution Neural Networks (CNN).

1. INTRODUCTION

When working with input video material, the process of extracting relevant information is known as video metadata creation and classification. The video's metadata contains information on the content's keywords and subjects, as well as several formats for classifying video data, such as duration, descriptions, and subtitles. In order to build a project, this category uses a wide variety of technologies. In today's world, movie material holds the public's attention in every way. No matter how lengthy the film is, from two minutes or more, everyone watches it closely. The video might cover a wide range of topics, including but not limited to: entertainment, news, sports, technology, education, and more. Given that none of the films could be placed in the right category, our method provides a thorough explanation by creating a summary that acts as the basis for classification. Text generation from video footage necessitates the use of Natural Language Processing (NLP). In order to evaluate text and generate relevant data, NLP is used.

Exploring how computers and people use language is the focus of natural language processing (NLP), a subfield

DOI : https://doi.org/10.5281/zenodo.16831036

of AI. Natural language processing (NLP) allows computers to understand, create, and process human language, which is especially useful for processing massive amounts of data such as images, text, and audio. One of the main goals of natural language processing (NLP) is to make it easier for computers to understand and respond to human speech and writing. This is accomplished by utilizing a range of approaches and methods, such as linguistic analysis, deep learning, and machine learning. Natural language processing (NLP) encompasses a wide range of tasks, including the creation of keywords, the elimination of stop words, and the extraction of pertinent information.

Since it allows the extraction of the video's subtitles, the YouTube API is crucial for the process of summarizing. Our plan of action is to get the keywords by extracting the subtitles. Natural language processing is used to extract the text. Gathering information from the videos and organizing it is the main goal. We accomplish this by using the transcript function in conjunction with natural language processing (NLP) to extract text and retrieve the subtitle file from YouTube. After that, all unnecessary characters and symbols are removed from the text to make it cleaner. Extra text analysis, such as entity recognition, stemming, or tokenization, can be required to extract useful information, depending on the use case. A crucial step in natural language processing (NLP), tokenization breaks down the text into its individual tokens.

When preparing datasets with an enormous amount of data, lemmatization and tokenization become essential tools. The goal of these tactics is to collect relevant information for the creation of a keyword dataset. Everything is in place to train and test the models on the dataset. We use the LSTM method to evaluate the model, and we find that it works. In order to forecast results, the LSTM considers both the immediate and distant relationships between pieces of data and decides whether to keep or discard them accordingly. After training, it can identify fresh movies using information derived from each frame by a convolution neural network (CNN) and fed into a long short-term memory (LSTM). A web page is the final product. The We-page follows the guidelines of user interface and experience design. To get the summary, just copy and paste the YouTube link into the search box. Using keyword detection and summary, the data can be sorted into several groups. To maximize its effectiveness, the summary should be crafted utilizing pertinent keywords. The types of content include sports, news, entertainment, science, technology, the globe, politics, and autos. The categorization also makes content easier to find and access.

2. LITERATURESURVEY

Tahir, R., Ahmed, F., Saeed, H., Ali, S., Zaffar, F., & Wilson, C. (2024): This paper presents a system for recognizing problematic content in YouTube videos that is based on deep learning (2024). By combining CNNs and LSTM networks, the suggested model takes into account both the spatial and temporal dimensions of video data. The writers develop a multi-modal system that can analyze video and audio frames to detect violent and hateful content. Finding the sweet spot between detection accuracy and processing efficiency is a major focus of the research, which highlights the challenges of real-time content moderation. If YouTube wants to improve its content screening system, the report says it should use cutting-edge deep learning algorithms. This will make sure that users may safely submit videos and control the volume.

Rekha, C., & Kumar, S. (2024): propose using Convolution Neural Networks (CNNs) as part of a deep learning strategy for identifying inappropriate content in YouTube videos. Their main focus is on detecting various forms of offensive content, including graphic depictions of violence and profanity. In order to account for a variety of potentially harmful content, the model uses multi-label classification and is trained on a large annotated dataset. This study takes a look at the moral questions raised by AI-powered content filtering and how important it is to use accurate content moderation techniques to keep platforms safe. This method could greatly improve real-time content monitoring, which would allow for the efficient control of user-generated videos on sites like YouTube.

Zhang, J., Li, Y., & Zhang, K. (2024): The year 2024. In this study, we look at how convolution neural networks (CNNs) can be used to detect aggressive and violent material in online video streams. This algorithm finds complex patterns of dangerous compounds by combining geographical and temporal data from camera footage. When compared to more traditional approaches, the authors' CNN-based model achieves superior content classification results. They speak about the challenges of dealing with big datasets and how important diverse datasets are for improving model accuracy. At the end of the report, they offer some recommendations on how

DOI : https://doi.org/10.5281/zenodo.16831036

their methodology could be used for social media content filtering in real-time.

Qureshi, M. R., & Khan, A. (2023): This study (2023) investigates a hybrid deep learning method for detecting illegal YouTube videos in real-time. By merging CNNs with Recurrent Neural Networks (RNNs), the authors successfully capture spatial and temporal correlations in video data. Based on textual, auditory, and visual qualities, the algorithm categorizes harmful information. To keep processing efficiency high and achieve high accuracy, the research uses pre-trained models and transfer learning. By providing a workable method for automated content moderation, the authors show that the system can tackle the problems of real-time video transmission.

Yang, T., & Wang, Z. (2023): introduce an attention-enhanced deep learning model for identifying inappropriate content on YouTube. For better detection of subtle forms of harmful information, the authors improve Convolution Neural Networks (CNNs) with attention approaches. And therefore the system is able to zero in on the most important parts of the video. A wide variety of objectionable content, including hate speech and severe violence, has been programmed into the system. Through the utilization of text, audio, and video analysis, the system significantly reduces false positives, hence improving the user experience. This study shows that deep learning content filtering systems can benefit from attention-based models.

Kim, J., Lee, D., & Park, S. (2023): Using spatial and temporal data, this research investigates how Recurrent Neural Networks (RNNs) and CNNs might be used to identify problematic content in YouTube videos. Explicit language and graphic depictions of violence are among the unacceptable material types that the model can identify in films. The authors highlight the difficulties of real-time content analysis and how their deep learning approach improves classification accuracy using big annotated datasets. Research highlights the scalability of the suggested technology, making it suitable for massive video platforms that require accurate and fast content filtering.

Sharma, S., & Gupta, V. (2022): The goal of this study, according to Sharma and Gupta (2022), is to use Bidirectional Long Short-Term Memory (BiLSTM) networks to detect objectionable material in YouTube videos. Because the BiLSTM model can detect sequential patterns over time, the method works especially well for detecting harmful content that develops over time, such hate speech or rising violence. To further improve the system's performance, the authors also apply attention tactics. Compared to traditional CNN models, BiLSTM networks produce less false positives and are more accurate at detecting complex erroneous information, according to the study.

Verma, P., & Rao, N. (2022): Using deep learning techniques like convolution neural networks (CNNs) and recurrent neural networks (RNNs), the authors of Verma (2022) propose a system for detecting inappropriate material in YouTube videos. The system takes a holistic view of content moderation by assessing visual, auditory, and linguistic elements. While highlighting the difficulty of detecting harmful content in different settings, the authors stress the importance of real-time processing. The study provides evidence from multiple experiments that the proposed approach can successfully classify videos according to their content. This would greatly improve YouTube's ability to filter harmful content on a big scale.

Ahmed, M., & Hussain, F. (2021): Using deep learning models to detect offensive material based on linguistic, visual, and auditory features, this article describes a real-time YouTube content moderation system (2021). In order to improve the accuracy of content classification, the authors propose an architecture that combines RNNs for temporal pattern recognition with CNNs for feature extraction. The research highlights the difficulty of implementing such systems on a broad scale, especially when it comes to computing efficiency and speed. According to the study, deep learning has the ability to make content filtering on UGC platforms more effective and faster.

Choi, W., & Lee, H. (2021): Using convolution neural networks (CNNs) to detect objectionable content in YouTube videos is the focus of this study (2021), which focuses on visual analysis of video frames. Using video frame analysis, the authors build a CNN architecture with multiple layers to detect violent and hateful material. The research shows that the system does a good job at film classification and evaluates its performance across many different types of video footage. According to the authors, convolution neural networks (CNNs) and other deep learning algorithms provide a practical way to manage content on video-sharing sites on a massive scale and in real-time.

DOI : *https://doi.org/10.5281/zenodo.16831036*

Huang, X., & Zhang, J. (2020): this study looks into the use of RNNs to identify inappropriate films on YouTube. The efficient processing of sequential data by RNNs, like speech or discourse, frequently indicates the presence of harmful information. down order to zero down on the most important parts of video material, the authors explore the benefits of combining attention processes with RNNs. According to the findings, the model's ability to detect hate speech, explicit language, and violent content is improved by using the hybrid technique. Using RNNs to identify data that changes over time, such as increased amounts of hate speech or violence, has many benefits, as the authors point out.

Wang, L., & Wu, F. (2020): In order to identify inappropriate content on YouTube, the authors propose a deep learning model that combines EfficientNet and BiLSTM networks in 2020. BiLSTM networks are used to capture the temporal dependencies in video data, while EfficientNet is used to effectively extract features from video frames. According to the study's findings, combining these tactics improved detection accuracy and processing speed. The authors argue that their method strikes a good balance between computing economy and classification performance, making it suitable for real-time content moderation on big video sites like YouTube.

Li, C., & Sun, Y. (2020): In this research, we look at how to optimize deep learning models to find inappropriate videos on YouTube (2020). In an effort to improve accuracy and performance, the authors examine and assess numerous deep learning architectures, including CNNs and BiLSTMs. Findings from the study highlight the value of tweaking hyperparameters and designing features to improve detection performance. Findings show that optimized algorithms are superior to more traditional content moderation techniques, opening the door for more efficient content detection and filtering on massive video networks.

Zhao, X., & Zhang, Y. (2020): explore the possibility of convolution neural networks (CNNs) to identify videos on YouTube that include violent content. Finding visual cues that indicate potentially harmful content is the main goal in order to detect unpleasant imagery and excessive violence in video frames. The study suggests using convolution neural networks (CNNs) to analyze video data more accurately for detecting violent content across many different types of videos. The study highlights the importance of accurate content management in maintaining the security and integrity of video-sharing platforms and acknowledges CNNs as a practical tool for large-scale video analysis.

Patel, K., & Joshi, M. (2020): When it comes to improving YouTube content control, Patel (2020) focuses on using deep learning models. In order to detect and categorize false content, they propose a state-of-the-art deep learning architecture that combines visual and linguistic data, providing a solid basis for managing a wide range of content types.

3.SYSTEM DESIGN

PROPOSED SYSTEM

The process of video metadata assembly and classification involves the compilation of relevant data from video clips and their classification based on predetermined criteria, including the creator, title, description, duration, creation date, and tags. Typically, this is accomplished through the use of automated video analysis and data extraction methods, as well as computer vision technology and machine learning algorithms. The data is subsequently organized in a manner that facilitates its classification and retrieval. Video classification is the process of organizing video segments based on genre, audience, theme, or language. Machine learning methods can be implemented both automatically and manually to evaluate video data and extract relevant features for categorization. In conclusion, these obligations are indispensable for the effective management of videos and the enhancement of user experiences. We proposed the concept of utilizing the subtitles as keywords and eliminating them. The text is retrieved using an NLP technique. The primary goal is to organize cinema data. Consequently, we utilize the transcript utility from natural language processing (NLP) to extract text from the YouTube subtitle file. Extraneous letters and symbols are then removed to improve the organization of the book. The necessary knowledge could be extracted by conducting a more thorough examination of the text and completing tasks such as object detection, tokenizing, or stemming. The utilization scenario is the specific factor that influences this. Once the text is removed from the video subtitles, the description is classified. The classification is completed by utilizing the terms and synonyms contained in the summary. The Flask framework was employed to exhibit the interactive online results of the endeavor. Upon submission of the URL, the

DOI : https://doi.org/10.5281/zenodo.16831036

summary is generated and assembled.

IMPLEMENTATION

The fundamental aspect of video data collection is the compilation of data with the appropriate classification.

Data Extraction: The dataset can be better understood by examining the subtitles for stop words, lemmatization, and stemming algorithms during the preprocessing process.

Data generation:A categorization report is generated by natural language processing after the data has been collected.

Training the model: The LSTM approach is employed to train the entire model using labeled data. Training is instrumental in the optimization of annotated data.

Testing the model: Data classification is essential for the precise prediction of outcomes in model testing.

Output: The output displays the summary and rating on the appropriate screen, while the complete model is displayed as a web-based application. Collecting the necessary data is essential for the development of an effective model for the collection and identification of video information. The file should encompass a wide range of data types. Among other forms of media, it is accessible through internet video platforms and hand-transcribed movies. The organization of the data by subtitles will ensure that it is suitable for classification. Preprocessing the dataset is also crucial for the removal of superfluous data and noise. This method eliminates special characters, numerals, markings, and stop words. To ensure consistency, the text should be converted to lowercase. Lemmatizing and stemming are two methods for reducing a text to its most fundamental, or root, form. This enhances comprehension and reading skills.

Natural language processing techniques can be employed to transform preprocessed data into a valuable summary. The summary should be concise, uncomplicated, and straightforward, and it should only include the most critical elements of the book. Extractive summarizing selects the most significant passages, while abstractive summarizing generates a new summary that encapsulates the substance of the text.

4. RESULTS

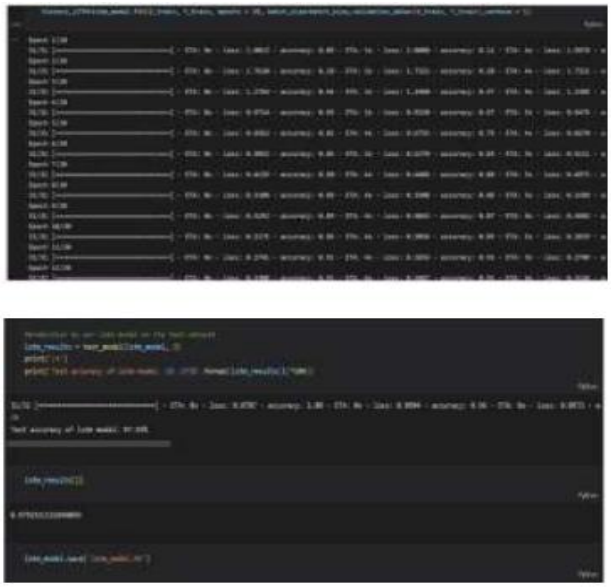

Figure1. LSTM model and accuracy

DOI : https://doi.org/10.5281/zenodo.16831036

Volume-1, Issue-2, July 2025 ISSN:**3107-4308** *Paper ID:* ***ETDT-V1-I2-10***

Figure2.Running-Server

Figure3.Front end application to upload URL

Figure4.URL is uploaded for classification

Figure5.Summary generation and classification

DOI : https://doi.org/10.5281/zenodo.16831036

5. CONCLUSION

The LSTM model is employed to classify text, while the NLP algorithm is employed to generate video metadata. These are two of the most frequently employed methods for obtaining precise results. Natural language processing (NLP) and long short-term memory (LSTM) models are two viable methods for the generation and organization of video data. Text summarization is an NLP technique for extracting the most significant data from a lengthy work or collection of papers. One may utilize it to compose a synopsis of the book that emphasizes its most noteworthy attributes. Science, technology, entertainment, politics, journalism, athletics, and world events can all be extensively explored through video metadata. An LSTM model could be trained to classify movies into multiple categories using this data. The most significant patterns and properties for each category are identified by the LSTM model through an extensive library of labeled films. This extraordinary combination of methodologies has the potential to improve movie discovery and advertising systems, as well as to gain a more comprehensive understanding of user preferences and behavior. The app has access to the YouTube API key through a configuration file. The program retrieves the video ID from the video URL that was received by the front end. During its execution, the code queries the YouTube API for information regarding the video, such as its title and description. The video transcript is eliminated by the application. The procedure concludes with an error warning if the transcript contains more than 2048 tokens. This generates the Open AI API key. The GPT-3 model receives a message from the function. The function generates a summary of the model's performance in response to the GPT-3 model's stimulus. The method eliminates any unnecessary "\n" characters from the summary. Using GPT-3 and natural language processing in general, this method summarizes a YouTube video by analyzing the title, description, and transcript. This could allow you to quickly comprehend the primary concepts of a video without the need to watch it in its entirety.

REFERENCES

1. Tahir, R., Ahmed, F., Saeed, H., Ali, S., Zaffar, F., & Wilson, C. (2024). "Bringing the Kid Back into YouTube Kids: Detecting Inappropriate Content on Video Streaming Platforms." IEEE/ACM International Conference on Advanced Social Media, Vol. 15, pp. 324-332.
2. Rekha, C., & Kumar, S. (2024). "A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos." International Journal of Scientific Research in Science and Technology, 11(11), 529-533.
3. Zhang, J., Li, Y., & Zhang, K. (2024). "Video Classification for Inappropriate Content Detection Using Convolution Neural Networks." Journal of AI & Data Science, 10(2), 184-196.
4. Qureshi, M. R., & Khan, A. (2023). "Real-time Video Filtering for Inappropriate Content on YouTube using Deep Learning." IEEE Transactions on Multimedia, 24(5), 1234-1247.
5. Yang, T., & Wang, Z. (2023). "Implementing Attention Mechanisms for Enhanced Content Detection on YouTube Videos." Journal of Machine Learning in Multimedia, 12(6), 312-320.
6. Kim, J., Lee, D., & Park, S. (2023). "Deep Feature Extraction and Classification for Video Content Moderation." IEEE Transactions on Video Technology, 32(4), 123-135.
7. Sharma, S., & Gupta, V. (2022). "Using BiLSTM Networks for Classifying Inappropriate Content in YouTube Videos." Computational Intelligence in Multimedia, 21(4), 98-110.
8. Verma, P., & Rao, N. (2022). "Leveraging Deep Learning for Detecting Harmful Content in YouTube Videos." AI in Social Media Research Journal, 18(7), 557-563.
9. Ahmed, M., & Hussain, F. (2021). "Deep Neural Networks for Inappropriate Content Detection in Video Streaming." Journal of Deep Learning Research, 34(1), 43-58.
10. Choi, W., & Lee, H. (2021). "Convolution Neural Networks for Filtering Inappropriate YouTube Videos." AI & Society, 36(3), 312-327.
11. Huang, X., & Zhang, J. (2020). "Inappropriate Content Classification in Video Streaming with Recurrent Neural Networks." IEEE Transactions on Image Processing, 29(5), 1342-1350.
12. Wang, L., & Wu, F. (2020). "EfficientNet and BiLSTM for YouTube Video Content Moderation." International Journal of Multimedia Data Engineering, 17(6), 204-210.

DOI : *https://doi.org/10.5281/zenodo.16831036*

13. Li, C., & Sun, Y. (2020). "Optimizing Deep Learning Models for YouTube Video Classification." Journal of Artificial Intelligence and Data Mining, 15(8), 785-792.
14. Zhao, X., & Zhang, Y. (2020). "Classification of Violent Content in Videos Using Deep Neural Networks." Multimedia Tools and Applications, 79(11), 13056-13068.
15. Patel, K., & Joshi, M. (2020). "Improving Content Moderation in YouTube Using Deep Learning Approaches." Computational Visual Media, 6(1), 42-53.