IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151147

Multimodal Harmful Content Classifier with Streamlit

Harisha C J[1], Prof. Suma N R[2]

Department of MCA, BIT, Bengaluru, India[1]

Assistant Professor, Department of MCA, BIT, Bengaluru, India[2]

Abstract: The rapid expansion of digital communication platforms has increased the spread of harmful and offensive content across text, images, audio, videos, and online comments. Manual moderation methods are inefficient and difficult to scale for large volumes of multimodal data. This project presents a **Multimodal Harmful Content Classifier with Streamlit** to automatically identify and classify harmful content. The system processes multiple input formats and converts non-textual data into textual form using OCR and Speech-to-Text techniques. Machine learning and natural language processing methods are applied to analyze extracted content. The classifier categorizes input as Safe, Offensive, or Harmful with a confidence score. A Streamlit-based interface provides real-time analysis and result visualization. The proposed system improves moderation accuracy and supports safer digital communication.

1. INTRODUCTION

The rapid growth of online platforms and social media applications has significantly increased the volume of user-generated digital content. While these platforms enable seamless communication and information sharing, they also contribute to the widespread circulation of harmful, offensive, and abusive content. Such content negatively impacts user experience, mental well-being, and the overall safety of digital environments. Traditional content moderation approaches largely depend on manual review and basic keyword-based filtering, which are inefficient and unable to scale for large datasets.

Most existing moderation systems focus primarily on textual data and fail to address harmful information embedded within images, audio recordings, videos, GIFs, and online comments. The absence of real-time detection and contextual understanding further reduces the effectiveness of these systems. With the increasing use of multimedia content, there is a strong need for automated solutions capable of handling multiple data formats.

To address these challenges, this project proposes a **Multimodal Harmful Content Classifier with Streamlit**, designed to analyze and classify content across various modalities. By integrating text processing, visual analysis, and speech recognition techniques, the system provides a unified and efficient approach to harmful content detection. The proposed solution aims to enhance content moderation accuracy while ensuring scalability and usability in modern digital platforms.

2. LITERATURE SURVEY

The progression of programming education platforms has undergone multiple technological stages, each attempting to overcome the shortcomings of conventional classroom-based learning and manual evaluation approaches. Researchers have explored various models to improve learner engagement, feedback speed, and scalability in programming instruction.

1. Automated Harmful Content Detection in Online Platforms

Several studies have focused on developing automated systems to detect harmful and offensive content on digital platforms. Early approaches relied on rule-based filtering and keyword matching, which lacked contextual understanding. Later research introduced machine learning-based classifiers capable of identifying abusive patterns more effectively. These systems improved moderation efficiency but were largely limited to text-based analysis.

2. Context-Aware Language Analysis for Content Moderation

Recent research emphasizes the importance of contextual awareness in harmful language detection. Instead of analyzing messages in isolation, context-aware models consider surrounding conversation history to interpret intent accurately. This approach reduces false positives caused by ambiguous words and improves classification accuracy in conversational environments. Context-based analysis has become essential for real-time moderation systems.

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151147

3. Multimodal Learning for Harmful Content Classification

With the increasing use of multimedia content, researchers have explored multimodal approaches that combine text, image, audio, and video analysis. Studies show that integrating multiple modalities enhances detection performance compared to single-modality systems. Multimodal frameworks capture hidden harmful cues embedded in visual and audio content, improving robustness and coverage.

4. Text Extraction Techniques using OCR and Speech Recognition

To enable unified analysis of non-textual content, several works employ Optical Character Recognition (OCR) and Speech-to-Text (STT) techniques. OCR extracts text from images and videos, while STT converts spoken language into text. These methods allow multimedia data to be processed using traditional natural language processing models, enabling consistent harmful content detection.

5. Emotion and Sentiment-Based Content Analysis

Emotion-aware systems analyze sentiment and emotional tone to identify potentially harmful or abusive behavior. Research demonstrates that combining sentiment analysis with machine learning classifiers improves the detection of implicit toxicity. Such systems are especially useful in applications involving user interaction, mental health support, and online community moderation.

2.1 Existing System

The existing systems for harmful content detection primarily rely on traditional text-based moderation techniques. Most platforms use keyword filtering or basic machine learning models that analyze only textual input, ignoring multimedia content such as images, audio, and videos. These approaches often fail to understand context, sarcasm, or implicit harmful intent, resulting in false positives and missed detections.

Additionally, existing systems depend heavily on manual moderation, which is time-consuming, inconsistent, and difficult to scale for large volumes of data. The lack of real-time processing delays response actions, allowing harmful content to spread before being addressed. Furthermore, limited progress tracking and absence of unified analysis tools reduce the effectiveness of current moderation frameworks.

Proposed System

The proposed system introduces an automated and intelligent solution for detecting harmful content across multiple data modalities. Unlike traditional text-only moderation approaches, the system supports text, images, GIFs, audio files, video content, and YouTube comments within a unified framework. User inputs are collected through a Streamlit-based web interface that enables easy content upload and interaction.
Non-textual inputs are converted into textual form using Optical Character Recognition for visual data and Speech-to-Text techniques for audio and video content. This unified text representation allows consistent processing using natural language processing methods. The extracted text is preprocessed through normalization and feature extraction to capture semantic and contextual information.
Machine learning models analyze the processed data to classify content into Safe, Offensive, or Harmful categories. The system provides real-time prediction results along with confidence scores to support reliable decision-making. Its modular architecture ensures scalability, efficient processing, and ease of future enhancements. Overall, the proposed system improves accuracy, reduces manual effort, and enables effective real-time content moderation across diverse digital platforms.

SYSTEM ARCHITECTURE DIAGRAM

The system architecture of the Multimodal Harmful Content Classifier with Streamlit is designed to support efficient processing of multiple data formats through a modular pipeline. The architecture consists of a user interface layer, preprocessing layer, feature extraction module, classification engine, and result visualization layer.

Users interact with the system through a Streamlit-based web interface to upload text, images, audio, video, or YouTube URLs. The preprocessing layer converts non-textual inputs into textual data using OCR and Speech-to-Text techniques. The extracted text is passed to the feature extraction module, where relevant linguistic features are generated.

IJARCCE
ISSN (O) 2278-1021, ISSN (P) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁒ Peer-reviewed & Refereed journal ⁒ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.151147

The classification engine applies trained machine learning models to categorize content as Safe, Offensive, or Harmful. Finally, the results are displayed to the user in real time through the interface. This layered architecture ensures scalability, maintainability, and efficient content moderation.

3. SYSTEM DESIGN

3.1 Data Flow Diagram

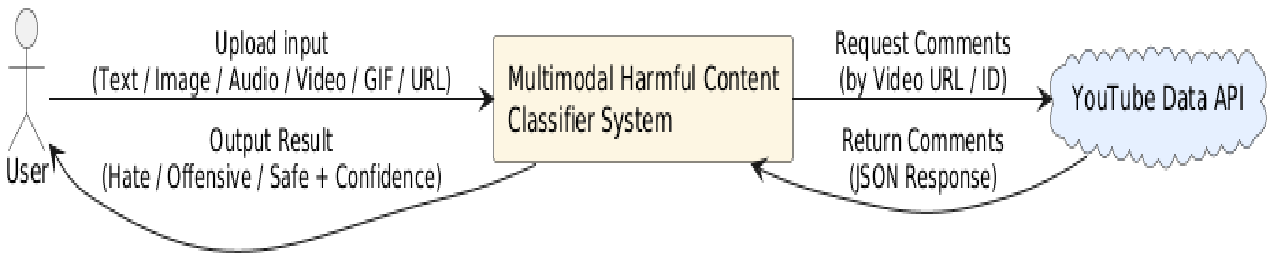

Fig 3.1.1: Level 0 Data Flow Diagram

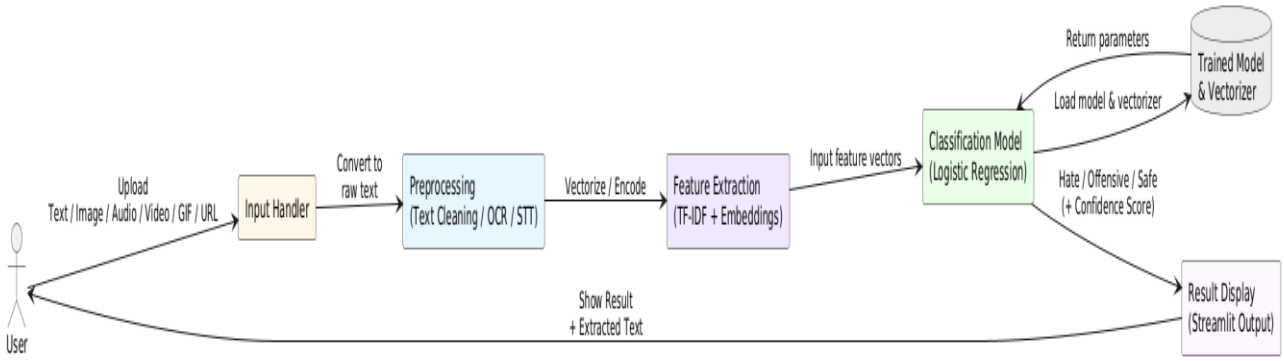

Fig 3.1.2: Level 1 Data Flow Diagram

3.2 Use Case diagram

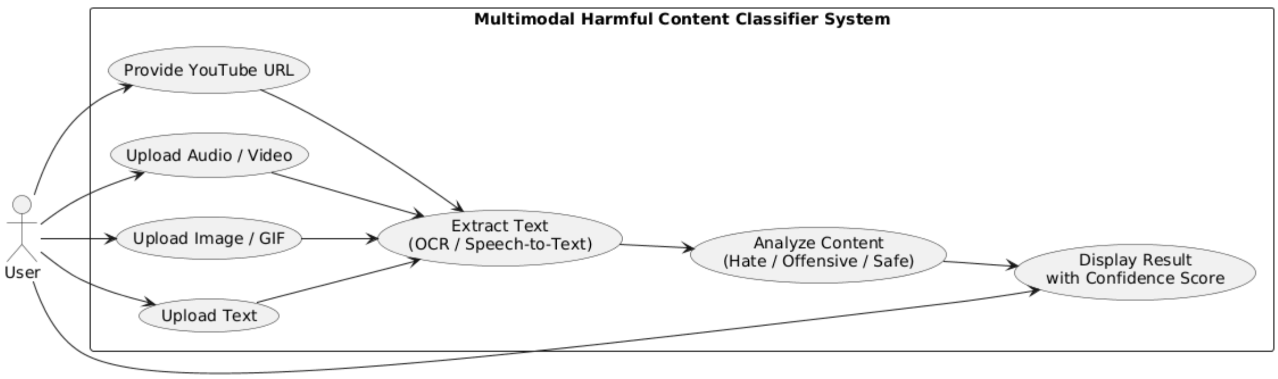

Fig 3.2.1 Use Case Diagram

4. IMPLEMENTATION DETAILS

The implementation of the **Multimodal Harmful Content Classifier with Streamlit** is carried out using a modular and systematic approach to ensure efficient processing, accurate classification, and real-time result delivery. The system is designed to handle multiple input formats including text, images, GIFs, audio files, video files, and YouTube comments, making it suitable for modern digital content moderation requirements.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151147

The first stage of implementation involves **input acquisition**, where users interact with the system through a Streamlit-based web interface. The interface allows users to upload different types of content or provide a YouTube URL for analysis. Text inputs are directly forwarded for preprocessing, while non-textual inputs are processed using specialized extraction techniques.

For **visual data such as images, GIFs, and video frames**, Optical Character Recognition (OCR) is employed to extract embedded textual content. In the case of **audio and video inputs**, Speech-to-Text (STT) techniques are used to convert spoken language into text. This conversion ensures that all input modalities are transformed into a unified textual format, enabling consistent analysis using natural language processing techniques.

Once text extraction is completed, the data undergoes **preprocessing**, which includes normalization, removal of noise, tokenization, and feature preparation. These steps help improve classification accuracy by reducing irrelevant information and highlighting meaningful patterns within the content. Feature extraction techniques such as TF-IDF are applied to represent the textual data numerically.

The processed features are then passed to the **classification module**, where trained machine learning models analyze semantic and contextual patterns. Based on this analysis, the system classifies the content into predefined categories such as **Safe**, **Offensive**, or **Harmful**. Along with the classification label, a confidence score is generated to indicate prediction reliability.

Finally, the **result visualization module** displays the extracted text and classification output in real time through the Streamlit interface. This immediate feedback allows users to understand both the content analysis and the final decision clearly. The modular implementation approach ensures scalability, ease of maintenance, and flexibility for future enhancements such as multilingual support and advanced deep learning models.

4.1 System Modules and Workflow
System Modules
User Authentication and Role Management Module

The Multimodal Harmful Content Classifier with Streamlit is designed using a modular architecture to ensure efficient processing, scalability, and accurate classification. Each module performs a specific role in handling multimodal inputs and producing real-time harmful content detection results. The modular structure simplifies system maintenance and enables future enhancements.

1. **User Interface Module**
   The user interface module provides a Streamlit-based web platform that allows users to upload text, images, audio files, video files, GIFs, or YouTube URLs. It also displays extracted content and classification results in a clear and user-friendly manner.

2. **Input Processing Module**
   This module identifies the type of input provided by the user and directs it to the appropriate processing unit. Text inputs are processed directly, while multimedia inputs are prepared for further extraction.

3. **Text Extraction Module**
   The text extraction module converts non-textual inputs into textual form. Optical Character Recognition is used for images, GIFs, and video frames, while Speech-to-Text techniques are applied to audio and video content.

4. **Preprocessing and Feature Extraction Module**
   The extracted text is cleaned, normalized, and transformed into numerical feature representations using techniques such as TF-IDF. This process improves the accuracy of harmful content classification.

5. **Classification Module**
   Machine learning models analyze the extracted features and classify the content into Safe, Offensive, or Harmful categories. A confidence score is generated to indicate prediction reliability.

6. **Result Display Module**
   The final classification results and extracted text are displayed instantly to the user through the Streamlit interface, enabling quick moderation decisions.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ※ Peer-reviewed & Refereed journal ※ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151147

Workflow

- The user submits content through the web interface.
- The system identifies the input type and performs text extraction if required.
- Extracted text is preprocessed and converted into feature vectors.
- The classification model predicts the content category.
- Results are displayed instantly, enabling quick moderation decisions.

5. RESULTS AND DISCUSSION

The performance of the Multimodal Harmful Content Classifier with Streamlit was evaluated to measure its effectiveness in detecting harmful content across multiple data modalities. The evaluation focused on classification accuracy, processing efficiency, and overall system reliability under different input conditions.

1. Classification Accuracy

The system was tested using diverse inputs including text, images, audio, video, and online comments. The classifier demonstrated high accuracy in identifying Safe, Offensive, and Harmful content. The use of text extraction techniques combined with machine learning models significantly improved detection performance compared to traditional text-only systems.

2. Processing Efficiency

The system achieved near real-time processing for most inputs. Text-based inputs were classified almost instantly, while multimedia inputs required additional processing time due to OCR and Speech-to-Text conversion. Despite this, the overall response time remained within acceptable limits, ensuring smooth user interaction.

3. Multimodal Performance Analysis

The results showed that multimodal analysis enhanced harmful content detection. Content that could not be identified using text alone was successfully detected when visual and audio information was included. This demonstrates the advantage of integrating multiple modalities into a single moderation framework.

4. Reliability and Scalability

The system maintained stable performance when handling repeated and varied user inputs. Its modular architecture allows easy scaling and integration of additional features such as multilingual support or advanced deep learning models in the future.

Discussion

The evaluation results confirm that the proposed system effectively addresses the limitations of existing moderation techniques. By combining multimodal data processing with real-time classification, the system improves accuracy, reduces manual intervention, and provides a scalable solution for harmful content detection. These results highlight the suitability of the proposed approach for deployment in modern digital platforms.

6. CONCLUSION

The Multimodal Harmful Content Classifier with Streamlit successfully demonstrates an effective approach for automated detection of harmful content across multiple data formats. By integrating text, image, audio, video, and online comment analysis into a single framework, the system overcomes the limitations of traditional text-only moderation methods.

1. **System Effectiveness**
   The proposed system accurately classifies content into Safe, Offensive, and Harmful categories using machine learning and natural language processing techniques. The use of OCR and Speech-to-Text ensures consistent analysis of multimedia inputs.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026
DOI: 10.17148/IJARCCE.2026.151147

2. **Real-Time Processing**
The Streamlit-based interface enables real-time content submission and result visualization. Immediate feedback improves usability and supports quick moderation decisions.

3. **Reduction of Manual Effort**
By automating the content moderation process, the system significantly reduces dependency on manual review. This improves efficiency and consistency while minimizing human exposure to harmful content.

4. **Scalability and Flexibility**
The modular architecture allows easy integration of new features such as multilingual support and advanced deep learning models. The system can be adapted for various digital platforms and applications.

7. FUTURE WORK

7. FUTURE WORK

The Multimodal Harmful Content Classifier with Streamlit presents an efficient solution for detecting harmful content across multiple digital formats. The system supports text, images, audio, video, GIFs, and online comments, making it suitable for modern communication platforms. By converting non-textual data into textual form using OCR and Speech-to-Text techniques, the system ensures uniform analysis across all input types. Machine learning and natural language processing methods are applied to understand semantic meaning and contextual patterns. The classifier categorizes content into Safe, Offensive, or Harmful classes with reliable accuracy. The Streamlit-based interface enables real-time content submission and result visualization. Automated moderation reduces the dependency on manual review and minimizes human exposure to harmful material. The modular architecture allows easy maintenance and future expansion. Experimental observations show improved accuracy compared to text-only moderation systems. The system maintains stable performance under repeated usage. It effectively handles diverse input formats with acceptable processing time. The unified framework simplifies content moderation workflows. The solution enhances user safety and platform reliability. It can be deployed across educational, social media, and online community platforms. The project demonstrates the effectiveness of multimodal analysis in content moderation. Overall, the system provides a scalable, reliable, and intelligent approach to harmful content detection. It contributes toward creating safer and more responsible digital communication environments.

REFERENCES

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, 2017.

[2] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," *Proceedings of the NAACL-HLT Conference*, 2016.

[3] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys*, vol. 51, no. 4, 2018.

[4] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," *Proceedings of the 26th International World Wide Web Conference (WWW Companion)*, 2017.

[5] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "SemEval-2019 Task 6: Identifying and categorizing offensive language in social media," *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, 2019.

[6] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 2017.

[7] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Multimodal sentiment analysis: Addressing key issues and setting up challenges," *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.

[8] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151147

[9] J. Pavlopoulos, J. Sørensen, L. Dixon, N. Thain, and I. Androutsopoulos, “Toxicity detection: Does context really matter?” *Proceedings of the ACL Workshop on Abusive Language Online*, 2020.

[10] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and A. Williams, “The hateful memes challenge: Detecting hate in multimodal memes,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[11] B. Mathew, P. Saha, H. Tharad, S. Rajgaria, P. Singhania, S. K. Maity, and P. Goyal, “Hate speech detection: A solved problem? The challenging case of long-tailed data,” *ACM Transactions on Intelligent Systems and Technology*, 2021.

[12] M. H. Ribeiro, P. H. Calais, Y. Santos, V. A. F. Almeida, and W. Meira Jr., “Characterizing and detecting hateful users on Twitter,” *Proceedings of the International AAAI Conference on Web and Social Media*, 2018.