DHANALAKSHMI SRINIVASAN
COLLEGE OF ENGINEERING AND TECHNOLOGY
ECR, Mamallapuram,Chengalpattu District
Approved by AICTE, New Delhi || Affiliated to Anna University, Chennai

ICREAMER'25

International Conference on Recent Advancements in Management and Engineering Research

Advancing Hate Speech Detection: A Multilingual System Using Llama-2 for Real-Time Analysis of Audio, Video, and Text Content

1. **NAVEEN .B** - Dhanalakshmi Srinivasan college of engineering and technology
2. **MOULIMONISH .S** - Dhanalakshmi Srinivasan college of engineering and technology
3. **NIMALAN .R** - Dhanalakshmi Srinivasan college of engineering and technology
4. **SRIVATHSAN .CB** - Dhanalakshmi Srinivasan college of engineering and technology

Contents

ABSTRACT:

- The Proliferation of social networks and the surge in user-generated media containing hateful content suggests an urgent need for a robust, multimedia-capable hate-speech detection solution to maintain a harmonious digital community .
- Traditional text-only detectors struggle when hateful ideas are embedded in audio and video tracks, leaving a significant blind spot for social platforms and moderators to control such content. This study proposes a complex multilingual solution for detecting hate speech in various media types including audio recordings and video segments and written content is described in this work. The model uses OpenAI Whisper to convert both audio and video content into text through automated transcription before advanced analysis can proceed. The system bases its operation on a Llama-2 model which has been destined for hate speech classification purposes from the transcribed and native text content. The web interface built with Streamlit provides users a convenient way to upload files and monitor real-time processing which delivers immediate results. The combined multimedia system produces substantial enhancements in detection accuracy alongside increased usability
- according to experimental findings which outperforms current available solutions. By combining OpenAI Whisper with advanced large language model (Llama-2) classification this multimedia system sets a new standard for accurate , scalable hate speech detection. Paving way for further extensions such as real-time live-stream monitoring and adaptive feedback loops.

INTRODUCTION

- A vast number of social media and multimedia communication platforms spreads hateful speech while making its supervision a challenging task. The current available hate speech detection methods function only on textual content yet they lack applicability in real-world situations where media consists of audio and video elements [1]. Currently available models function best with a single language while needing elaborate processing [2]. A real time multilingual hate speech detecting system of multimedia content has been developed to address the identified challenges in this area. The multimedia content processing system utilizes the Whisper model from OpenAI to achieve text conversion.
- The system accepts inputs of audio, video, or URL content which are then converted to text before passing through a Llama-2 model that was optimized for hate speech classification. Our system provides easy accessibility to technical and non technical users through immediate feedback along with simple interaction by deploying the system through an interactive Streamlit web application.

OBJECTIVE:

- ❑ Develop a robust, real-time, multilingual hate-speech detection system capable of processing text, audio, and video content.
- ❑ Leverage OpenAI Whisper for automated transcription of audio/video into text, and fine-tuned Llama-2 for classification of both transcribed and native text.
- ❑ Provide an accessible web interface via Streamlit for non-technical users to upload media and receive immediate hate-speech analysis results.

Literature Survey

S.NO	Title, Author, Year	Proposed Techniques	Description	Remarks
1	Fine Tuning Llama 2 Model for Detection of Multimodal Hate Speech; Sasidaran & Geetha, 2024	Fine-tuning Llama-2 on combined text, audio & video embeddings	Demonstrated that adapting Llama-2 to multimodal inputs yields higher detection accuracy	Validates value of multimodal fine-tuning
2	Advances in Hate Speech Moderation: Multimodality and Large Models; Hee et al., 2024	Large multimodal LLM architectures	Argues text-only detectors miss context; proposes multimodal LLMs for robust moderation	Highlights inadequacy of text-only approaches

Literature Survey

S.NO	Title, Author, Year	Proposed Techniques	Description	Remarks
3	A Comprehensive Cross-Language Framework for Harmful Content Detection with Sentiment Analysis; Dehghani, 2024	Layered sentiment analysis over multilingual text	Integrates sentiment scores to refine cross-language hate-speech detection	Adds nuance by distinguishing sentiment polarity
4	Brinjal: A Web-Plugin for Collaborative Hate Speech Detection; Hee et al., 2024	Real-time user feedback loops in a browser plugin	Enables moderators to flag content and iteratively refine model decisions in live settings	Enhances adaptability through human–AI collaboration

Literature Survey

S.NO	Title, Author, Year	Proposed Techniques	Description	Remarks
5	Investigating the Predominance of LLMs in Low-Resource Bangla for Hate Speech Detection; Faria et al., 2024	Comparative evaluation of public LLMs versus transformer models in Bangla	Benchmarks several open-source LLMs against traditional transformer-based classifiers on a low-resource Bangla hate-speech dataset.	Highlights that LLMs generalize better in under-served languages, bolstering the case for true multilingual support.
6	MultiSentimentArcs: Coherence in Multimodal Sentiment Analysis; Chun, 2024	Fusion of text, audio, and visual sentiment embeddings with coherence checks	Introduces a “sentiment coherence” module that aligns sentiment signals across modalities to improve detection in long-form narratives.	Offers a transferable framework for detecting implicit or subtle hate-speech cues by checking cross-modal consistency.

PROBLEM DEFINITION

- **Single-Modality Blindspot:** Existing detectors focus solely on text, missing hateful content embedded in audio/video streams.
- **Monolingual Constraints:** Many models fail to generalize across languages or require extensive per-language retraining.
- **Lack of Real-Time Feedback:** Moderators need instantaneous, multimedia-aware tools with user-friendly interfaces.

PROPOSED WORK

The proposed system is a unified, real-time pipeline that first ingests user-provided audio, video, or text (including URLs), then automatically transcribes any spoken content into text via OpenAI's Whisper model, and finally classifies the resulting text—whether originally written or transcribed—using a fine-tuned Llama-2 base model trained on a richly annotated, multilingual hate-speech corpus. The architecture is designed for seamless integration: multimedia inputs pass through preprocessing modules (youtube_transcript_api and pytube for video, SpeechRecognition for audio), Whisper performs end-to-end transcription, and the cleaned text is fed into the Llama-2 classifier, which has been optimized to detect both explicit and implicit hate across languages. A Streamlit-based web interface wraps the entire workflow, allowing non-technical users to upload media, trigger processing, and view detailed classification results (including confidence scores and highlighted hateful segments) in real time. This approach leverages the complementary strengths of Whisper's transcription accuracy and Llama-2's large-model language understanding to deliver robust, multilingual, multimodal hate-speech detection with minimal latency and maximum accessibility.

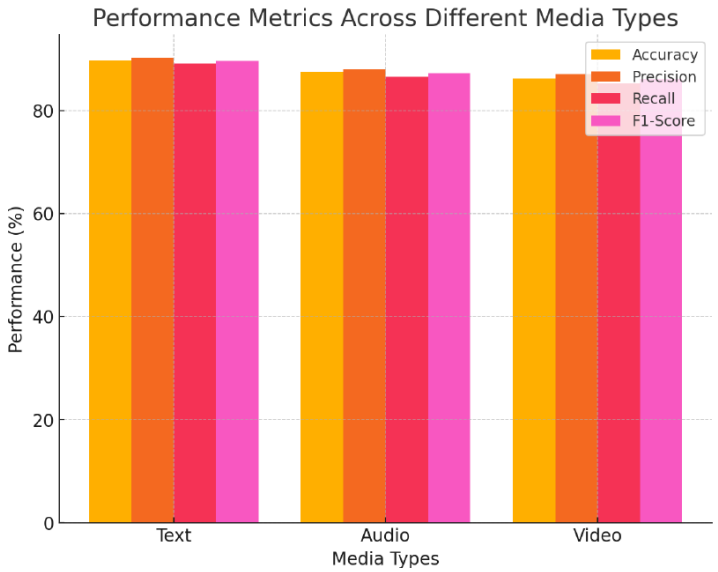
METHODOLOGY & IMPLEMENTATION

- The system is implemented in Python using PyTorch and the Hugging Face Transformers library, with OpenAI's Whisper model handling end-to-end transcription of audio and video into text. We curate a richly annotated, multilingual hate-speech corpus—spanning English, Spanish, Arabic, and more—augmented with Whisper-generated transcripts to train and fine-tune a Llama-2 base model.
- Multimedia inputs are preprocessed via the youtube_transcript_api and pytube for video streams and Python's SpeechRecognition for standalone audio, ensuring standardized text input. We evaluate performance across modalities using Accuracy, Precision, Recall, and F1-Score, reporting separate metrics for text, audio, and video. A Streamlit front end provides a user-friendly interface for uploading media or entering URLs, displays real-time classification results with confidence scores and highlighted segments, and logs feedback for iterative model refinement.

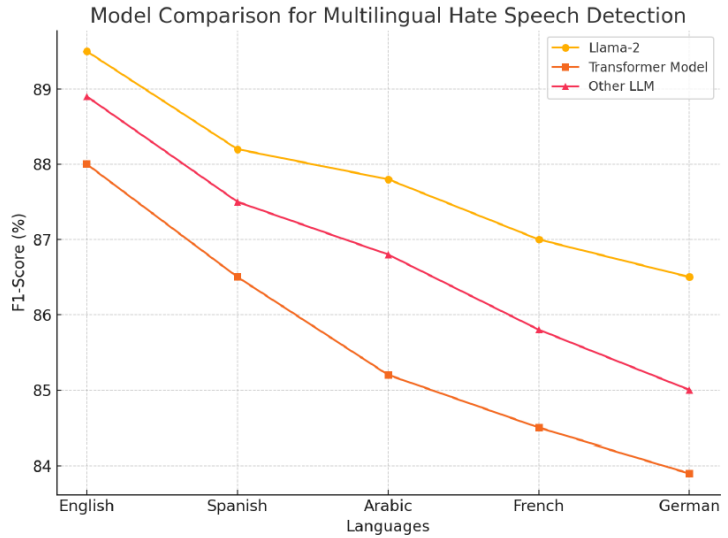
RESULTS & DISCUSSION

- The proposed system performed outstandingly when detecting various text formats alongside multiple languages. The table I presents accuracy and precision alongside recall and F1 score results when processing text, audio and video inputs. The detection model reached an overall accuracy rate of 89.7 % when processing text inputs while it detected phenomena in audio with 87.4 % precision and 86.2 % in video inputs.
- The proposed system demonstrates success by supporting different content formats with various languages to deliver highly effective detection of hate speech. Multimedia processing deployed within the system system enables flexible operations which improves real deployment effectiveness. The system outperforms text-based hate speech detection because it applies Whisper to transcribe materials while Llama-2 executes classification functions.
- The system achieves high accuracy rates but multiple system problems remain despite its achievements. The system needs improvement in precision because it encounters difficulties detecting concealed hate speech and new linguistic developments. Future optimization of the system is required as it affects real-time application processing speed through content length and complexity.

RESULTS & DISCUSSION



Performance Metrics



Comparison for Multilingual hate speech

Format	Accuracy	Precision	Recall	F1-Score
Text	89.7%	90.2%	89.1%	89.6%
Audio	87.4%	88.0%	86.5%	87.2%
Video	86.2%	87.0%	85.3%	86.1%

Performance Metrics for Multimedia Hate Speech Detection

CONCLUSION

- This work presents a hate speech detection method which implements Llama 2 and Whisper to process language content from different media sources. Our system builds upon text-based analysis by producing a complete streaming solution to identify hate speech which users can access through an easy-to-use Streamlit interface. The system achieves increased efficiency for processing simultaneously with dynamic hate speech patterns evaluation and expanded operational reach.
- Performance drops on context-dependent or highly implicit hate speech, which often requires deeper prosodic or visual cues. While suitable for batch processing, the current Whisper→Llama-2 chain incurs latency that may hinder live-stream moderation.
- **FUTURE SCOPE :** Incorporate audio-prosody and video-frame analysis (e.g., facial expressions) via multimodal attention mechanisms to better capture subtle hate cues. Implement user-in-the-loop feedback loops and few-shot adaptation to continuously update the model for emerging hate-speech patterns. Explore streaming-friendly architectures or model distillation techniques to reduce end-to-end latency for real-time broadcast moderation.


THANK
YOU!