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

1st NAVEEN B
Computer Science and Engineering
Dhanalakshmi Srinivasan College of Engineering and Technology
Chennai,India
contactnaveenb@gmail.com

2nd MOULIMONISH S
Computer Science and Engineering
Dhanalakshmi Srinivasan College of Engineering and Technology
Chennai,India
moulimonishs@gmail.com

3rd NIMALAN R
Computer Science and Engineering
Dhanalakshmi Srinivasan College of Engineering and Technology
Chennai,India
nimalan004@gmail.com

4th SRIVATHSAN CB
Computer Science and Engineering
Dhanalakshmi Srinivasan College of Engineering and Technology
Chennai,India
srivathsan539@gmail.com

5th prof. DR.K.JOHN PETER
Computer Science and Engineering
Dhanalakshmi Srinivasan College of Engineering and Technology
Chennai,India
kjohnpeter@gmail.com

***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.

***Index Terms*—Hate Speech Detection, Multilingual Analysis, Llama-2, Whisper, Multimedia Content, Real-Time Processing**

I. 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.

II. Related Work

The diverse forms of web content coupled with different language variations have increased the difficulty of identifying hate speech. The study demonstrates that low multimodal techniques and large language models (LLMs) as well as collaborative tools are effective for managing these detection difficulties. The following discussion analyzes key papers regarding their contribution to building complete systems that detect hate speech.

1. Fine Tuning Llama 2 Model for Detection of Multimodal Hate Speech.

Sasidaran and Geetha (2024) introduced a specialized Llama 2 model trained for the detection of multimodal hate speech within text speech and video contents. Research that appeared at the International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS) demonstrates the system improves accuracy by utilizing media of different types including audio, video, and text. The refinement of Llama 2 provides the model with effective capabilities to identify various forms of hate speech while demonstrating a comprehensive hate speech detection solution [1].

2. Advances in Hate Speech Moderation: Multimodality and Large Models

The authors Hee et al. (2024) explained their views on how multimodal approaches together with large models should be utilized for hate speech moderation. ArXiv researchers present the findings that show single-modality text-only methods are inadequate while large models which support text as well as video and audio content are essential to solve the challenge. The authors state that Llama 2 serves as an optimal solution for detecting intricate patterns of hate speech throughout all languages and media thus advancing online content moderation [2].

3. Large Language Models in Low-Resource Bangla Language for Hate Speech Detection

Faria together with Lihamir Hani Baniata and Kang (2024) conducted a performance evaluation of LLMs against transformer models in identifying hate speech content in limited-resource Bangla language. The investigators established that LLMs perform best among traditional models when there are restricted language resources during their research. The use of Llama 2 along with other LLMs helps expand hate speech detection strategies for minority languages [3].

4. The research examines public large language models for detecting hate speech.

The author Malode (2024) evaluated public large language models against specialized applications such as Hate Speech Detection to measure accuracy and scalability in their deployment. His research at Technische Hochschule Ingolstadt aims to investigate Llama 2 effectiveness in dealing with hateful content for his doctoral thesis. Malode presents evidence in his work which demonstrates the significance of model robustness and scalability when executing LLMs on actual hate speech detection tasks [4].

5. Cross-Language Framework for Harmful Content Detection with Sentiment Analysis

Dehghani (2024) applies a framework for multi-language detection to sentiment analysis and language modeling to increase the detection of harmful content. The research paper available on arXiv preprint presents evidence that sentiment detection layers improve detection accuracy specifically during cross-linguistic analysis. The research describes sentiment analysis as a process for understanding hate speech tones which leads to more complex multilingual hate speech detection systems [5].

6. MultiSentimentArcs for Coherence in Multimodal Sentiment Analysis

The researcher Chun (2024) details how to assess Multimodal Sentiment Analysis coherence in lengthy film narratives by using MultiSentimentArcs. The method Chun developed can be adapted to merge textual content with auditory and visual elements in hate speech analysis because the three elements are core elements of hate speech detection. Systems that utilize this method would have improved capacity to discover complex multimedia information containing semantic and intentional hate speech in non-straightforward content [6].

7. Brinjal: A Web-Plugin for Collaborative Hate Speech Detection

Brinjal serves as "a collaborative web plugin for improving hate speech detection with user interaction and real time feedback" according to Hee et al. (2024). Brinjal provides users with interactive capabilities to improve detection algorithms and makes an effective addition to systems that depend on continuous learning presented at the ACM Web Conference 2024. Real time applications benefit greatly from model tuning operated by users because accuracy requirements and adaptability take precedence in these scenarios [7].

8. Review of Multimodal Large Language Models Across Different Tasks

The performance and challenge aspects of multimodal large language models receive detailed evaluation from Wang et al. (2024) through their task-based review. The authors demonstrate through their arXiv paper that multimodal LLMs display potential as hate speech detection tools even though these models require significant computational resources. Real-time hate speech detection systems require efficient processing performance as well as model efficacy and the authors describe this trade-off [8]

.

9. Comprehensive Survey of Large Language Models:

Hadi et al. (2024) investigated the operations of large language models including their practical uses and limitations and potential ways to advance their functionality. The research paper uploaded to Authorea reveals the barriers to hate speech detection which occurs despite complex language variation while the models need precise accuracy in multilingual conditions. Authors suggest different approaches to scale down model efficiency so the methods stay effective across various media frameworks.

10. Argumentative Stance Prediction Using Multimodality and Few-Shot Learning

Sharma, Gupta and Bilalpur (2023) established argumentative stance predictions through the application of multimodal and few-shot learning approaches in their research paper. The authors demonstrate on arXiv that understanding user intent together with an argumentative stance directly impacts hate speech detection therefore highlighting their value in this field. Such few learning techniques help model detection of new hate speech patterns across different languages with minimal retraining requirements while maintaining their importance for dynamic hate detection systems [10].

III. Proposed System

A. System Architecture

The main parts of our system comprise input processing and hate speech classification alongside user interface functions.

1. **Multimedia Input Processing**
   The input process starts by adding audio then converts it into video that leads to text during the Whisper model of OpenAI's operation with a URL. The system achieves this capability because its preprocessing step converts all incoming media content into a standard textual format.

2. **Hate Speech Classification**
   The Llama-2 base model functions as the main hate speech classification tool that focuses on achieving peak accuracy in predictions. Additionally this model received training from diverse multilingual content involving hate speech to develop language-independent functionality.

3. **Web Interface**
   The created Streamlit web application presents an intuitive interface where users can interact with the system. Real sense analysis results submitted through the user interface will expand the user base that interacts with this system.

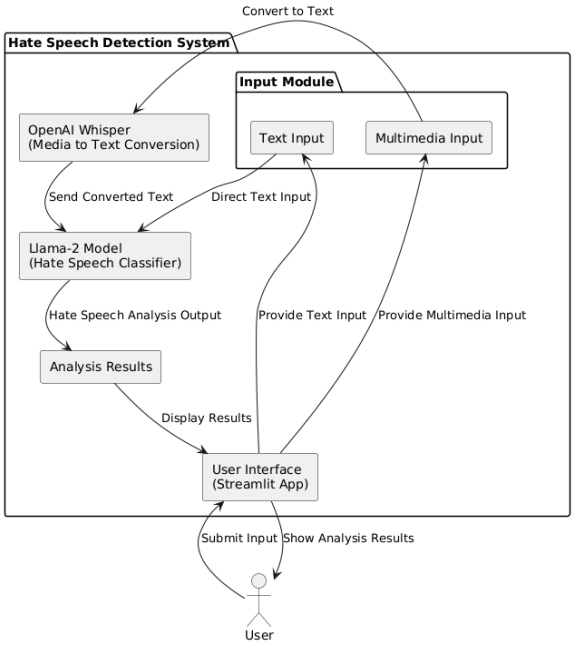

Fig 1 - System Architecture

B. Model Training and Fine-Tuning

The Llama-2 model underwent specific training for processing a comprehensive hate speech data collection which extended across multiple languages including English, Spanish, Arabic among others. During training the model achieved optimization to recognize different hate speech elements including slang expressions and coded expressions as well as implicit forms of hate speech.

IV. Experimental Setup

A. Dataset

The training parameters and evaluation criteria are supplied with annotations through various languages as well as manually transcribed audiovisual material for assessment purposes. Text representation remained standard because the multimedia data from Whisper served as the source input.

B. Evaluation Metrics

The system will use standard classification metrics which include accuracy, precision, recall together with F1 score for evaluation purposes. The combination of these metrics enables us to achieve a complete evaluation of model performance when determining hate speech from benign or neutral content.

C. Implementation Details

The complete system was implemented in Python with two of its following libraries:

- **streamlit** for the web interface,
- **torch** and **transformers** for model handling,
- **openai-whisper** for multimedia transcription,
- **youtube_transcript_api** and **pytube** for video processing,
- **SpeechRecognition** for additional audio processing.

V. Results

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.

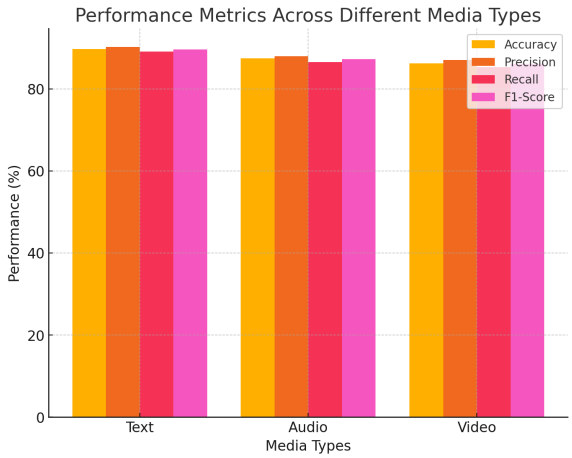

Fig 2 - Performance Metrics

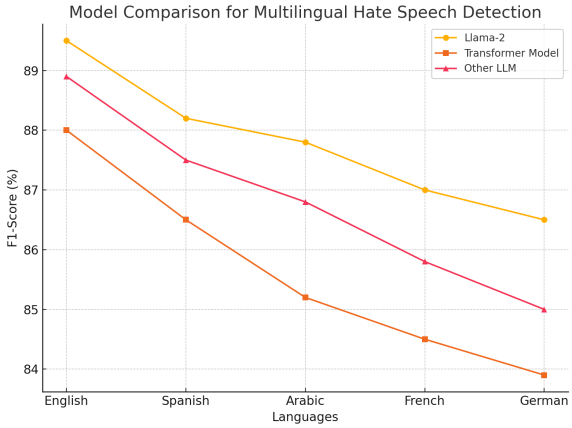

Fig 3 - Comparison for Multilingual hate speech

Table I: Performance Metrics for Multimedia Hate Speech Detection

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%

VI. Discussion

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.

VII. 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.

REFERENCES

1. Sasidaran, K., & Geetha, J. (2024, August). Multimodal Hate Speech Detection using Fine-Tuned Llama 2 Model. In *2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS)* (pp. 1-6). IEEE.
2. Hee, M. S., Sharma, S., Cao, R., Nandi, P., Chakraborty, T., & Lee, R. K. W. (2024). Recent advances in hate speech moderation: Multimodality and the role of large models. arXiv preprint arXiv:2401.16727.
3. Faria, F. T. J., Baniata, L. H., & Kang, S. (2024). Investigating the Predominance of Large Language Models in Low-Resource Bangla Language Over Transformer Models for Hate Speech Detection: A Comparative Analysis.
4. Malode, V. M. (2024). *Benchmarking public large language model* (Doctoral dissertation, Technische Hochschule Ingolstadt).
5. Dehghani, M. (2024). A comprehensive cross-language framework for harmful content detection with the aid of sentiment analysis. *arXiv preprint arXiv:2403.01270*.
6. Chun, J. (2024). MultiSentimentArcs: a novel method to measure coherence in multimodal sentiment analysis for long-form narratives in film. *Frontiers in Computer Science*, *6*, 1444549.
7. Hee, M. S., Singh, K., Si Min, C. N., Choo, K. T. W., & Lee, R. K. W. (2024, May). Brinjal: A Web-Plugin for Collaborative Hate Speech Detection. In *Companion Proceedings of the ACM on Web Conference 2024* (pp. 1063-1066).
8. Wang, J., Jiang, H., Liu, Y., Ma, C., Zhang, X., Pan, Y., ... & Zhang, S. (2024). A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319*.
9. Hadi, M. U., Al Tashi, Q., Shah, A., Qureshi, R., Muneer, A., Irfan, M., ... & Shah, M. (2024). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.
10. Sharma, A., Gupta, A., & Bilalpur, M. (2023). Argumentative Stance Prediction: An Exploratory Study on Multimodality and Few-Shot Learning. *arXiv preprint arXiv:2310.07093*.