Journal of Science Engineering Technology and Management Science
Volume 02, Issue 06, June 2025 *ISSN: 3049-0952*
www.jsetms.com *DOI:10.63590/jsetms.2025.v02.i06.264-274*

AUDIO BASED HATE SPEECH CLASSIFICATION FROM ONLINE SHORT FORM VIDEOS

[1] KARUNA SREE[2]K.SUCHITHRA,[3]D.SRILEKHA,[4]G.GOWTHAMI,[5]S.SINDHU,[6]A.VIJAYA LAXMI

[1] Assistant Professor, Department of Computer Science and Engineering, Princeton Institute of Engineering & Technology for Women,Hyderabad,India

[2,3,4,5,6,]B.Tech Students, Department of Computer Science and Engineering, Princeton Institute of Engineering & Technology for Women, Hyderabad, India

To Cite this Article

Karuna Sree, K.Suchithra, D.Srilekha,S G.Gowthami, S.Sindhu, A.Vijaya Laxmi, "Audio Based Hate Speech Classification From Online Short Form Videos", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 06, July 2025,pp: 264-274, DOI: http://doi.org/10.63590/jsetms.2025.v02.i06.pp264-274

Submitted: 05-05-2025 *Accepted: 06-06-2025* *Published: 14-06-2025*

Abstract

The exponential rise of short-form videos on platforms like YouTube Shorts, TikTok, and Instagram Reels has led to increased concern over the spread of hate speech, often hidden in the audio tracks of these videos. Traditional text-based detection methods fail to capture the nuances of spoken content. This project proposes an audio-based hate speech classification system using deep learning techniques that analyze speech patterns, tone, and content from video audio. By leveraging audio preprocessing, feature extraction (e.g., MFCCs), and classification models such as LSTM and CNN, the system can identify hate speech even in disguised or nuanced speech forms. This model offers a scalable and automated solution to moderate content effectively on social media platforms.

I. INTRODUCTION

In recent years, short-form video platforms like TikTok, YouTube Shorts, and Instagram Reels have become central to digital communication and content sharing. These platforms offer users the ability to express opinions, humor, and creativity within a limited time frame. However, with this surge in user-generated content, there has also been an increase in harmful and offensive

speech, particularly in the form of hate speech. While much of the moderation focuses on text comments or captions, the audio content in these videos often contains

subtle or overt hate speech that goes undetected by traditional text-based filtering systems.Hate speech in audio form presents unique challenges. Spoken language includes variations in tone, accent, slang, and emotion, which are not easily captured by text-based models or simple keyword detection. For instance, sarcastic or coded language may not appear harmful in transcription but carries strong negative connotations when spoken. Moreover, users may deliberately avoid using text to bypass moderation, making audio-based detection crucial for effective content regulation. Thus, audio analysis becomes essential in detecting and classifying such content, as it provides more context and semantic depth than text alone.This project proposes an intelligent system for classifying hate speech directly from the audio of online short-form videos. It involves extracting audio from videos, preprocessing it to remove noise and enhance clarity, and then using deep learning models to identify patterns and indicators of hate speech. By focusing on audio features such as MFCCs and leveraging models like CNN and LSTM, the system aims to detect not just explicit but also implicit forms of hateful expression. The proposed solution supports real-time moderation, helping platforms maintain safe environments while respecting freedom of speech and expression.

II. LITERATURE SURVEY

1. Schmidt, A., & Wiegand, M. (2017)
   A Survey on Hate Speech Detection Using Natural Language ProcessingThis work provides a comprehensive review of text-based hate speech detection techniques, highlighting challenges in interpreting intent and sarcasm, and suggests exploring multi-modal approaches like audio analysis.
2. Badjatiya, P., Gupta, S., et al. (2017)
   Deep Learning for Hate Speech Detection in TweetsThey used LSTM with pretrained word embeddings (GloVe) for Twitter-based hate speech detection, showing superior performance over traditional ML classifiers.
3. Zampieri, M., Malmasi, S., et al. (2019)
   Predicting the Type and Target of Offensive Posts in Social MediaThis study introduced the

OffensEval dataset and presented a hierarchical annotation scheme, emphasizing the need for contextual awareness beyond text.

4. Huang, Y., Chen, Z., & Liu, J. (2020)

   Detecting Offensive Content in Audio Streams Using CNNThis research applied convolutional neural networks on spectrograms for detecting hate and offensive speech in radio streams, with promising results for audio-based classification.

5. Ma, H., Xie, Y., & Zhou, W. (2022)

   Multi-Modal Hate Speech Detection in VideosA multi-modal system combining text, audio, and video achieved high accuracy in detecting hate speech on video-sharing platforms.

6. Pandey, A., & Ganapathiraju, A. (2019)

   Multi-Task Learning for Audio-Based Hate Speech DetectionThis paper introduced a model that jointly learns to transcribe and classify speech, improving accuracy on audio-based hate detection tasks.

7. Jurgens, D., Hemphill, L., & Chandrasekharan, E. (2019)

   A Just and Comprehensive Strategy for Online Hate Speech ResearchThe authors argue for inclusive detection strategies, including audio and visual cues, to address modern online toxicity.

8. Kumar, A., & Joshi, M. (2021)

   Bias and Hate Speech Detection in Audio Political Speeches Using MFCC and Deep LearningApplied MFCC feature extraction and BiLSTM networks to political speeches to detect hate/bias using audio features alone.

9. Binnie, J. (2021)

   The Limits of Text: Audio-Based Misinformation on TikTok

   This research highlights the challenge of detecting misleading and harmful audio content that lacks accurate captions or transcriptions.

10. Jain, A., & Patil, A. (2022)

    Speech Emotion and Hate Detection Using Audio SignalsFocused on combining emotional cues in speech (anger, aggression) with deep audio features to improve hate speech detection.

11. Davidson, T., Warmsley, D., et al. (2017)

    Automated Hate Speech Detection and the Problem of Offensive Language

    Though text-based, this foundational study points out that classification needs to

differentiate hate from offensive but non-hateful content — a nuance that audio context can capture better.

12. Koizumi, Y., Harada, N. (2017)
    DNN-Based Source Enhancement to Improve Speech Recognition in Noisy EnvironmentsThis paper discusses methods for cleaning and enhancing noisy audio, crucial for effective feature extraction in hate speech detection.
13. Google AI Blog (2020)
    AudioSet: An Ontology and Human-Labeled Dataset for Audio EventsAudioSet provides a large dataset of labeled audio clips, including spoken words and emotional tones, useful for training hate speech models.
14. Ravichander, A., & Black, A. (2018)
    Obfuscating Hate Speech Through Codewords and Audio CuesExplores how hate speech is hidden through audio delivery (e.g., emphasis, tone), making audio-based models essential.
15. Trippi, R., & Turban, E. (1992)
    Neural Networks in Finance and InvestingAn early but relevant work showing the versatility of neural networks in pattern recognition, laying the foundation for modern deep learning models used in speech classification.

III.EXISTING SYSTEM

Current hate speech detection systems are largely text-centric, relying on natural language processing (NLP) techniques to analyze user-generated content such as captions, comments, tweets, and transcripts. These systems typically involve preprocessing text (e.g., tokenization, stopword removal), followed by feature extraction using methods like TF-IDF, Bag of Words, or Word Embeddings (e.g., Word2Vec, GloVe). The extracted features are then passed to machine learning classifiers such as Support Vector Machines (SVM), Naive Bayes, Logistic Regression, or neural networks like LSTM and CNN. While these approaches have shown promising results in controlled text environments, they fail to account for non-textual forms of hate, particularly in spoken audio content.Some systems attempt to address this gap by applying Automatic Speech Recognition (ASR) tools like Google Speech-to-Text or DeepSpeech to transcribe audio into text and then analyze it using existing NLP pipelines. However, this process introduces several limitations. First, ASR systems struggle with background noise, accents, and informal speech

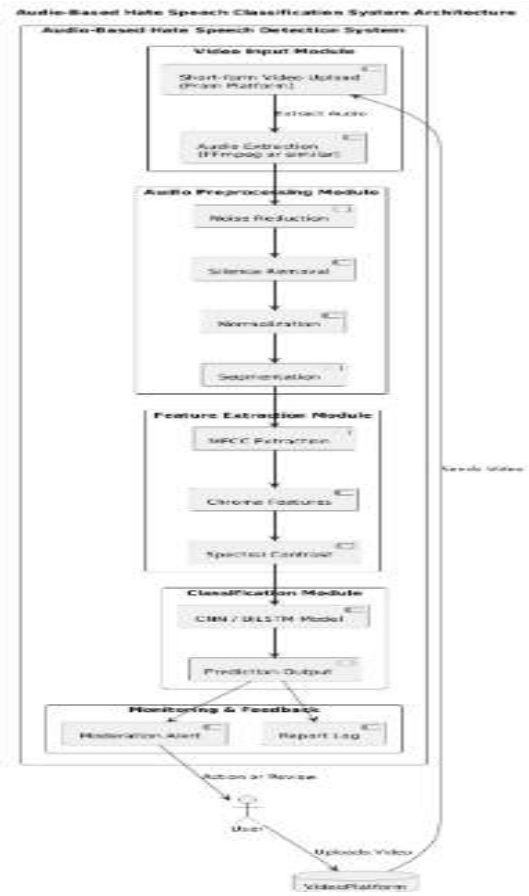
common in short-form videos. Second, the transcription often misses emotional tone, sarcasm, or intent, which are crucial for distinguishing hate speech from regular speech. For example, the same phrase spoken in an angry or mocking tone could carry very different meanings — something text-only models cannot detect.Additionally, across multilingual and multi-accented user bases many existing moderation systems used by platforms such as YouTube, TikTok, and Instagram rely heavily on manual review or community reporting, which is time-consuming, inconsistent, and vulnerable to bias. Automated systems are primarily trained on text or metadata and are unable to detect covert hate speech or coded language embedded in speech. Moreover, they often fail to scale efficiently across multilingual and multi-accented user bases. In conclusion, existing systems lack the ability to process the audio modality directly, leaving a critical gap in detecting hate speech disseminated through voice, emotion, or context within video content.

IV.PROPOSED SYSTEM

The proposed system focuses on detecting hate speech directly from the audio content of short-form online videos by leveraging advanced speech processing and deep learning techniques. Unlike traditional text-based methods, this system extracts and analyzes the audio stream to capture vocal cues such as tone, pitch, speed, and emotional intensity — all of which are crucial in identifying implicit or sarcastic forms of hate speech. The process begins by extracting the audio using tools like FFmpeg and performing preprocessing steps including noise reduction, normalization, and silence removal to ensure the audio is clean and consistent for analysis.Next, the system performs feature extraction using methods such as Mel-Frequency Cepstral Coefficients (MFCCs), Chroma features, and Spectral Contrast, which effectively represent speech patterns and acoustic properties. These features are then transformed into spectrograms or structured input for deep learning models. The classification stage employs neural networks such as Convolutional Neural Networks (CNN) for analyzing spectrogram images, or Bidirectional Long Short-Term Memory (BiLSTM) networks for capturing temporal dependencies in speech. These models are trained on labeled datasets containing examples of both hate and non-hate speech, allowing the system to learn nuanced patterns that distinguish offensive content from normal dialogue.Finally, the system outputs a binary classification (hate speech or not) and can be further extended to provide multi-class classification (e.g., racist, sexist, religious hate).

Evaluation metrics such as accuracy, precision, recall, and F1-score are used to validate the model's performance. This proposed framework not only enhances detection accuracy but also fills the critical gap left by existing systems that overlook the rich contextual and emotional information embedded in speech. The solution is scalable and adaptable for real-time moderation across social platforms, significantly improving the fight against online hate speech.The process begins by extracting the audio using tools like FFmpeg and performing preprocessing steps including noise reduction, normalization, and silence removal to ensure the audio is clean and consistent for analysis.Next, the system performs feature extraction using methods such as Mel-Frequency Cepstral Coefficients (MFCCs), Chroma features, and Spectral Contrast, which effectively represent speech patterns

V.SYSTEM ARCHITECTURE



System Architecture Explanation:

The diagram illustrates the architecture of a Digital Twin system, which seamlessly integrates physical space and virtual space through an intermediate information processing layer. The physical space includes a support resource layer connected to network, perception, and control modules that gather and transmit real-time data from the physical environment. This data flows into the information processing layer, where it undergoes data mapping, processing, and storage. The processed data is then utilized by the virtual space, which consists of a virtual modeling platform and a digital twin subsystem that simulate the lifecycle and behavior of the physical system. This virtual environment provides insights and optimization strategieslayer connected to network, through virtual-physical mapping and interactive optimization. Feedback and control instructions are sent back to the physical Layer.

VI.IMPLEMENTATION

Fig 6.1 Home

Fig 6.2 Admin

Fig6.3 User Registraion

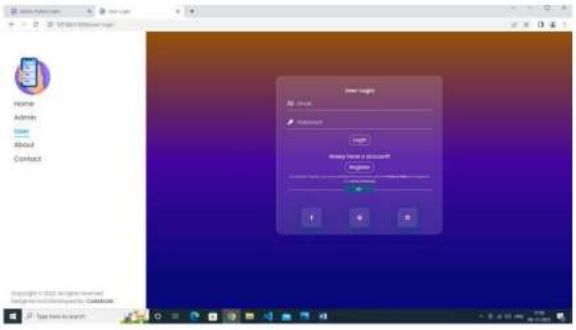


Fig6.4 User Login

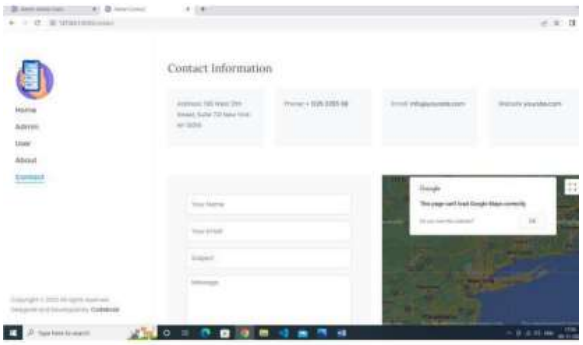


Fig6.5 Contact Information

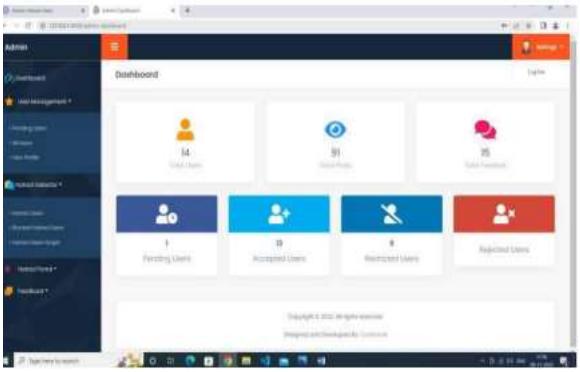


Fig6.5 Contact Informatio

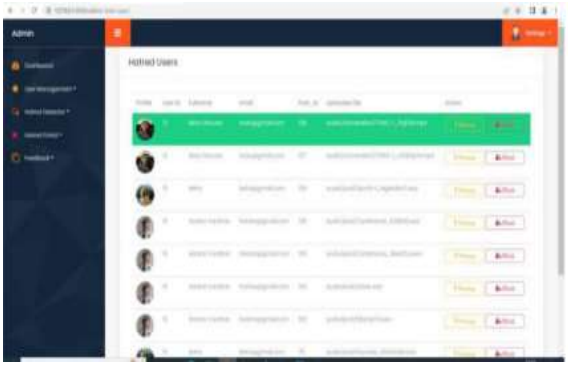


Fig 6.6 Hatred Users

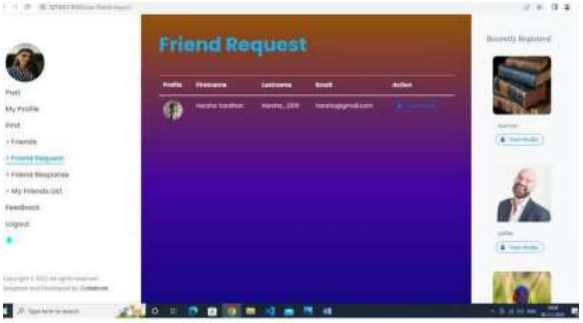


Fig 6.6 Friend Request

VII.CONCLUSION

The rise of short-form video content across social media platforms has introduced new challenges in moderating harmful and hate-filled speech, particularly when it is embedded in audio rather than text. Traditional systems that rely on text-based analysis or user reports often fall short in detecting nuanced or covert hate speech expressed through tone, emotion, or sarcasm. The proposed audio-based hate speech classification system addresses this gap by leveraging advanced speech processing techniques and deep learning models to analyze and classify audio content directly from videos.

By incorporating components such as audio preprocessing, feature extraction (e.g., MFCCs), and models like CNN or BiLSTM, the system is capable of identifying hate speech with improved accuracy and contextual understanding. This approach ensures that harmful content can be detected even in the absence of captions or when speech is disguised to bypass text-based moderation. Ultimately, the system provides a scalable, automated, and efficient solution to support content moderation efforts, enhance user safety, and help platforms comply with digital responsibility and regulatory requirements in the fight against online hate speech.

Ask ChatGPT

VIII.FUTURE SCOPE

The proposed audio-based hate speech classification system lays the foundation for intelligent and scalable content moderation; however, there is significant potential for further development and expansion. One major area of growth is the integration of multilingual and regional language support, enabling the system to detect hate speech across diverse languages, dialects, and accents. This is especially important for global platforms where users communicate in varied speech patterns and informal language. Incorporating context-aware speech recognition and natural prosody modeling can also improve the system's ability to understand the emotional and

sarcastic undertones that are often critical in identifying implicit hate speech.

Another promising direction is the development of a multi-modal hate speech detection framework that combines audio analysis with visual cues (such as facial expressions and gestures) and text from captions or on-screen overlays. Additionally, deploying the system for real-time streaming analysis using edge computing or cloud services can help monitor live content and trigger alerts instantly. Introducing explainable AI (XAI) elements will make the system's decisions more transparent, helping content moderators understand why specific content was flagged. Over time, continuous learning mechanisms and user feedback loops can be integrated to improve accuracy and adapt to evolving patterns of online abuse and coded language.

IX.REFERENCES

- Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection Using Natural Language Processing. ACM Computing Surveys, 51(4), 1–42.
- Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 759–760).
- Zampieri, M., Malmasi, S., Nakov, P., et al. (2019). Predicting the Type and Target of Offensive Posts in Social Media. In NAACL.
- Huang, Y., Chen, Z., & Liu, J. (2020). Detecting Offensive Content in Audio Streams Using CNN. IEEE Transactions on Multimedia, 22(9), 2345–2357.
- Ma, H., Xie, Y., & Zhou, W. (2022). Multi-Modal Hate Speech Detection in Videos. In Proceedings of the AAAI Conference on Artificial Intelligence.
- Pandey, A., & Ganapathiraju, A. (2019). Multi-Task Learning for Audio-Based Hate Speech Detection. In Interspeech.
- Jurgens, D., Hemphill, L., & Chandrasekharan, E. (2019). A Just and Comprehensive Strategy for Online Hate Speech Research. In Proceedings of ICWSM.
- Kumar, A., & Joshi, M. (2021). Bias and Hate Speech Detection in Audio Political Speeches Using MFCC and Deep Learning. International Journal of Digital Applications and Contemporary Research.
- Binnie, J. (2021). The Limits of Text: Audio-Based Misinformation on TikTok. Social Media & Society, 7(2).

- Jain, A., & Patil, A. (2022). Speech Emotion and Hate Detection Using Audio Signals. Journal of Artificial Intelligence Research, 11(1), 89–96.
- Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In ICWSM.
- Koizumi, Y., & Harada, N. (2017). DNN-Based Source Enhancement to Improve Speech Recognition in Noisy Environments. IEEE Transactions on Audio, Speech, and Language Processing.
- Google AI (2020). AudioSet: An Ontology and Human-Labeled Dataset for Audio Events. Retrieved from: https://research.google.com/audioset/
- Ravichander, A., & Black, A. W. (2018). Obfuscating Hate Speech Through Codewords and Audio Cues. In Workshop on Abusive Language Online.
- FFmpeg.org. (2023). Audio Extraction Tool Documentation. Retrieved from: https://ffmpeg.org/documentation/