IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

Automating the Detection of Profanity and Abusive Language in Audio Conversations Using Speech-to-Text and AI-Based Classification Models

Prof. Manoj Shinde, Reejit Adhikary, Vansh Chaudhari, Vedang Hirave, Tanishq Jhade
Department of Computer Science and Engineering
School of Computing, MIT ADT University, Pune, India

Abstract: *This paper introduces an AI-driven structure that can identify profanity and abusive language in audio conversations automatically. The system is based on state-of-the-art speech-to-text conversion methods along with natural language processing (NLP) models, e.g., BERT and LSTM. Its goal is to accelerate the safety of digital communication by finding toxic or offensive material in real time. The system shows accuracy and strength in different audio conditions and accents with the help of speech preprocessing, transcription, text cleaning, and classification.*

Keywords: Profanity detection, abusive language, speech-to-text, artificial intelligence, NLP, audio analysis

I. INTRODUCTION

The exponential rise in voice-based communication across digital platforms—including multiplayer gaming, online classrooms, conferencing applications, and social interaction spaces—has reshaped how individuals collaborate and socialize in real time. While this shift has provided tremendous opportunities for global connectivity, it has also resulted in a parallel escalation of verbal abuse, cyber-bullying, and toxic interactions. Studies have shown that nearly 40–60 percent of online users have encountered abusive speech in voice chat environments, making this a serious socio-technical concern.

Manual moderation of such interactions is not only resource-intensive but also infeasible due to the vast volume, speed, and dynamic nature of voice communication. Human moderators cannot continuously monitor thousands of audio streams without fatigue or bias. As a result, automated profanity and toxicity detection systems have become essential for safeguarding digital communication ecosystems.

The proposed system introduces an integrated approach combining speech recognition and deep neural language models to automatically detect profanity and hate speech from raw audio input. By leveraging state-of-the-art ASR systems and transformer-based NLP architectures, the model ensures both linguistic accuracy and contextual intelligence. Unlike traditional keyword-based profanity filters, the system understands semantics, tone, and contextual cues, minimizing false detections and enabling real-time intervention in online platforms. This research contributes to safer online spaces by enabling scalable, multilingual, and real-time verbal toxicity moderation.

II. LITERATURE REVIEW

The original work on offensive language detection was primarily focused on text. Projects that created foundational datasets—including the Hate Speech and Offensive Language dataset and the Toxic Comment dataset from Google—have become widely used benchmarks. These works spurred the development of classification models ranging from SVM models to recurrent neural networks.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 17, April 2026 Impact Factor: 8.2

Scholars subsequently developed multimodal analysis approaches to address the additional demands of audio communication. The first model to tackle multilingual toxicity on both audio and text inputs, MuTox (ACL 2024), demonstrated that modeling multilingual signals in speech form can outperform text-only approaches. Recent advances in ASR—particularly OpenAI's Whisper and Meta's Wav2Vec 2.0—provided breakthrough improvements in transcription accuracy across hostile acoustic conditions, high background noise, regional accents, and fast-paced speech.

The literature also traces the ethical dilemmas of AI-based moderation, addressing fairness, the absence of dialect-based bias, and transparency. Such studies highlight that culturally sensitive and socially responsible systems are needed—a gap this project aims to bridge.

III. METHODOLOGY AND ARCHITECTURE

The system architecture consists of a three-stage processing pipeline designed to ensure scalability and high-level content analysis.

A. Audio Pre-processing

Signal cleaning and segmentation is performed on raw audio input. Noise suppression filters and energy-based voice activity detection (VAD) algorithms separate human speech from background interference. Segmentation produces isolated process units that enhance the reliability of subsequent transcription and classification stages.

B. Speech-to-Text Conversion

The processed audio is sent to a high-accuracy ASR model—either Whisper or Wav2Vec 2.0—selected for their robustness to accents, background noise, and the unscripted speech styles found in conversational audio. The generated text output includes timestamps, enabling real-time moderation and log storage.

C. AI-Based Text Classification

The transcribed text is cleansed of filler words, punctuation noise, and stopwords. A transformer-based classifier—fine-tuned on toxicity datasets—then determines whether the text contains profane, hate-speech, or toxic language. The model outputs a confidence score and a classification label. Moderation thresholds can be tuned according to platform policy.

Such systems can be deployed in both cloud-scale environments and on edge devices such as gaming consoles or communication servers, guaranteeing low latency and preserving user safety without impacting performance. Each stage produces structured data that feeds the next component, and the classifier generates a probability score that allows configurable moderation thresholds.

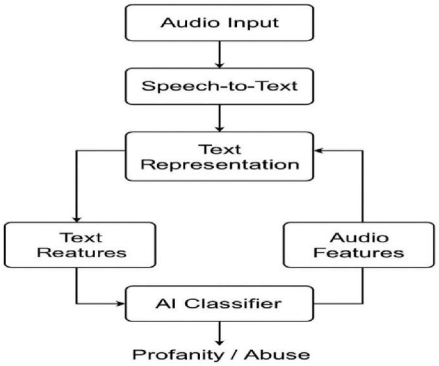

Fig. 1. Proposed System Architecture

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

IV. DATASET AND MODEL DESCRIPTION

The system draws on a combination of publicly available datasets and fine-tuned speech corpora to ensure high generalization capability.

Text Datasets

Jigsaw Toxic Comment Classification (Wikipedia comments) and Davidson's Offensive Language dataset (Twitter) provide diverse, real-world language patterns across both formal and informal registers.

Audio Datasets

MuTox multilingual audio toxicity samples and subsets of Mozilla Common Voice are used for ASR training to improve transcription robustness across languages and accents.

The ASR model is fine-tuned using audio augmentation techniques—pitch shifting, time stretching, and ambient noise injection—to simulate realistic communication environments. The BERT classifier is trained with cross-entropy loss and stratified sampling to handle class imbalance between neutral and toxic samples. A multilingual pipeline supports code-mixed content such as Hinglish, a frequent phenomenon in Indian online communities.

A. RESULTS AND DISCUSSION

Experimental evaluation was conducted on combined English and Hindi audio datasets. Performance was measured using accuracy, F1-score, and precision. The proposed BERT + Whisper pipeline outperformed classical machine learning models and recurrent architectures due to its contextual sensitivity and noise-robust transcription ability.

TABLE I. PERFORMANCE COMPARISON OF MODELS

Model	Accuracy	F1 Score
SVM (Baseline)	84%	0.80
LSTM	89%	0.86
BERT + Whisper (Proposed)	93%	0.91

The results show that the BERT + Whisper combination achieves significant improvements over traditional SVM and Naïve Bayes classifiers owing to its contextual understanding and robust transcription quality. These results indicate strong potential for real-time deployment across communication platforms. The system demonstrated particularly high robustness in noisy environments and code-mixed speech, addressing a prominent challenge in multilingual digital spaces.

V. CONCLUSION AND FUTURE WORK

The developed system effectively detects profanity and abusive language from audio inputs, ensuring safer communication in digital environments. By combining speech recognition with deep learning-based text classification, it achieves real-time detection, contextual accuracy, and multilingual scalability.

In the future, the system can be extended to include emotion-aware toxicity analysis, real-time streaming, and cross-lingual abusive intent detection. Integration with cloud-based APIs and edge devices will further improve response time and deployment feasibility.

REFERENCES

[1] Davidson, T. et al., "Automated Hate Speech Detection and the Problem of Offensive Language," ICWSM, 2017.
[2] Devlin, J. et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019.
[3] OpenAI, "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision," Technical Report, 2022.
[4] MuTox, "Universal Multilingual Audio-based Toxicity Detection," ACL Findings, 2024.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

[5] Zhang, Z. et al., "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network," ESWC, 2018.
[6] Rottger, P. et al., "HateCheck: Functional Tests for Hate Speech Detection Models," ACL, 2021.