Using OpenAI-Whisper and NLP to Prevent Toxicity in Online Communication

Rishabh Bedidha, Apurwa Khanal, Alex Kim, David Kim, Brenton Lian

Abstract

Online toxicity is a growing challenge across social media and gaming platforms. Most existing moderation tools rely on user reports and stored audio recordings, raising privacy concerns and limiting scalability. We present a lightweight, real-time toxic-speech detection system that uses OpenAI Whisper for transcription and a transformer-based Toxic-BERT classifier with a custom scoring algorithm. Our approach avoids storing raw audio and retains only flagged text snippets, improving both privacy and storage efficiency. Across two experiments, our method maintains perfect precision while improving recall by 142% after incorporating contextual transformers. A qualitative case study examines the strengths and weaknesses of our approach. Our findings demonstrate that privacy-preserving, real-time voice moderation is feasible using modern STT and NLP models.

1 Introduction

Online voice-chat toxicity is a persistent problem across social media and gaming platforms, harming users and reducing engagement [6]. Traditional moderation depends on user reports and manual review, which is slow and difficult to scale, and storing full audio for review also creates significant privacy concerns.

Recent advances in speech-to-text (STT) technology and natural language processing (NLP) provide opportunities to detect toxic speech automatically in real time. However, challenges remain: STT models may introduce transcription errors that reduce detection accuracy, and existing toxicity classifiers often fail to capture context, sarcasm, or multi-lingual speech. [3]. In this work, we propose a real-time system for detecting and classifying toxic speech using OpenAI Whisper for STT combined with a contextual NLP pipeline and custom weighting for severity categorization. Our system offers several advantages over existing approaches:

- **Real-time monitoring and detection**: allows immediate response to toxic speech
- **Lower storage demands for platforms**: Only flagged speech is retained instead of full recordings
- **Preservation of user privacy**: mitigates concerns about the storage of extensive recordings
- **Flexible severity thresholds**: enables platform-specific tuning of what constitutes toxic speech
- **Reduced reporting burden**: reduced for users and moderators.

We evaluate the system on real-world and generated speech and demonstrate that adding contextual transformer models dramatically improves recall while maintaining perfect precision.

2 Related Work

Prior text-based hate-speech work, like as Waseem & Hovy (2016)[10], Davidson et al. (2017)[1], and Google Jigsaw's Perspective API (Google Jigsaw, 2017)[5], shows effective toxicity scoring, but these methods rely on written content and cannot directly handle real-time voice communication.

Recent advances in speech-to-text (STT) systems, such as DeepSpeech[2] and Google Cloud STT[4], have greatly improved transcription accuracy across speakers and noisy settings. Whisper offers a robust multilingual model with strong performance[7], making it well-suited for real-time toxicity pipelines. Prior studies highlight challenges in contextual interpretation, multilingual detection, and disambiguating toxic vs. gaming-specific language.

Our work builds on these findings by integrating Whisper with transformer-based toxicity modeling and a custom scoring strategy tailored to harassment patterns in real-time voice communication.

3 Method

3.1 Overview

The pipeline has four stages:

1. **Transcription** — Whisper transcribes streaming audio into text.
2. **Keyword scoring** — A rule-based system applies severity multipliers for direct harassment, slurs, or harmful phrases.
3. **Context-aware scoring** — Toxic-BERT predicts toxicity probability; we apply a calibration transform to reduce overconfidence.
4. **Final decision** — A weighted score determines whether the speech exceeds the toxicity threshold (0.65).

3.2 Keyword Scoring

After normalization (lowercasing, punctuation removal), we apply multipliers: We apply length-normalization to reduce false positives for long sentences:

$$\text{keyword\_score} = \min\left(1, \left(\frac{2 \cdot S_{\text{initial}}}{X}\right)^{1.7}\right)$$

Where S_{initial} is the raw keyword score

Type	**Multiplier**
Keywords used in direct harassment (e.g., "you are X")	15x
Normal keyword use	2x
Phrase match (e.g., "kill yourself")	10x

Table 1: Table of Keyword multipliers

3.3 Contextual Toxicity Model

We use Toxic-BERT, a transformer classifier trained on Twitter and hate-speech datasets. Let ρ be the model's predicted toxicity probability.

$$\rho^{'} = (100\rho)^{0.7}$$

and

$$\text{context\_score} = 1 - \rho^{'}$$

This exponent reduces model overconfidence and better reflects uncertainty from STT errors.

3.4 Final Score

$$\text{final} = 0.5 \cdot \text{keyword\_score} + 0.5 \cdot \text{context\_score}$$

Speech is classified as toxic if final > 0.65

4 Experiments/Evaluation

4.1 Dataset

We evaluated 123 total clips:

- **Toxic Speech**: 45 YouTube clips + 48 spoken phrases
- **Non-toxic Speech**: 20 YouTube clips + 10 spoken phrases
- **Hardware**: Macbook Air (2023) microphone

4.2 Experiment 1 - Baseline: Keyword-Only Detection

Setup: 15 toxic YouTube clips + 15 team-spoken toxic phrases; 15 non-toxic YouTube clips + 15 team non-toxic phrases

Metric	Value
Precision	100%
Recall	26%
F1 Score	58.5%

Table 2: Experiment 1

Analysis: Keyword-only detection maintains perfect precision but misses a majority of toxic speech (low recall) due to a lack of context awareness

4.3 Experiment 2 - Transformer-Enhanced Detection

Setup: Same as Experiment 1, but with combined keyword + Toxic-BERT context scoring.

Metric	Value
Precision	100%
Recall	63%
F1 Score	87.2%

Table 3: Experiment 2

Analysis: Adding Toxic-BERT increases recall by 142% while maintaining perfect precision.

4.4 Case Study

We evaluated the system qualitatively on three example inputs:

Sample 1: "Fuck, I missed my shot"

- **Scores**: keyword=1, context=0.76, final=0.88
- **Flagged Words**: “fuck”
- **Observation**: Although non-toxic, keyword normalization based on length combined with context score results in moderately high toxicity

Sample 2: "We need to kill the enemy healer"

- **Scores**: keyword=0, context=0.28, final=0.14
- **Flagged Words**: none
- **Observation**: Correctly avoids false positives; “kill” is not flagged in a gaming context.

Sample 3: "We don’t need an elaborate build; something basic should be fine."

- **Scores**: keyword=0.9375, context=0, final=0.46875
- **Flagged Words**: “rat”, “basic”
- **Observation**: Context score reduces the impact of spurious keywords, preventing false positives.

5 Discussion

Our hybrid approach achieves strong real-time performance without audio retention. Whisper enables accurate transcription in typical gaming conditions, and transformer augmentation corrects most keyword-based over-flagging. High precision across experiments indicates minimal false positives, essential for real-world deployment.

However, several limitations remain. The keyword list requires manual curation and may miss emerging slang. Whisper still struggles with heavy noise or overlapping speakers. Toxic-BERT is trained primarily on Twitter text, limiting generalization to voice-chat domains. Short utterances remain difficult for contextual models.

Despite these limitations, results demonstrate that real-time, privacy-preserving voice moderation is technically viable today and benefits substantially from hybrid scoring.

6 Conclusion and Future Work

We introduce a real-time toxic speech detection pipeline that integrates Whisper transcription, keyword scoring, and contextual transformer classification. The hybrid method significantly improves recall while maintaining perfect precision and aligns with privacy requirements by avoiding audio storage.

Future directions include:

- Multilingual toxicity modeling
- Domain-specific transformer fine-tuning for gaming and social platforms
- Improved context modeling using conversation history
- Adaptive keyword lists learned from user feedback

References

[1] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, pages 512–515, 2017.

[2] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Ng. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[3] Duc Hoang et al. Audio-to-text alignment for multilingual automatic speech recognition. In *NAACL 2024*, 2024.

[4] Ioana Iancu et al. Evaluating google speech-to-text api's performance for romanian e-learning resources. *ResearchGate*, 2019.

[5] Google Jigsaw. Perspective api. `https://www.perspectiveapi.com`, 2017.

[6] Jacob Morrier et al. How do the effects of toxicity in competitive online video games vary by source and match outcome? *PLoS One*, 20(6):4, 2025.

[7] Alec Radford et al. Robust speech-to-text with whisper. OpenAI, 2023.

[8] Shubham Saraf et al. Multilingual translation for speech and text using whisper ai: A deep learning approach. *ResearchGate*, 2023.

[9] Siyuan Song, Andrew Kim, Jennifer Chien, Christopher Lewis, and Brian Wu. Detecting toxic language on social media using transformer models. *Information*, 12(5), 2021.

[10] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, 2016.