INTERNATIONAL JOURNAL OF Telecommunications
AIR DEFENSE COLLEGE

Enhancing Ride-Hailing Safety through Real-Time Speech-Based Violence Detection

Citation: Waleed ,M.; Yasser, M.; Magdy, M.; Elkady , M; Hesham, T; Abdelsalam , K. *Inter. Jour. of Telecommunications,* ***IJT'2025****, Vol. 05, Issue 02, pp. 1-12, 2025.*

Doi: 10.21608/ijt.2025.392310.1114

Editor-in-Chief: Youssef Fayed.

Received: 04/06/2025.

Accepted date: 26/08/2025.
Published date: 27/08/2025.

Mohannad Waleed, Mariam Yasser, Mayada Magdy, Maryam Elkady, Tasneem Hesham*, Khaled Abdelsalam.

College of Information Technology & Artificial Intelligence, MUST, Giza, Egypt.

*Corresponding author: tasneem.h.114@gmail.com

Abstract: The rise of ride-hailing services has brought growing safety concerns, especially incidents of verbal harassment during trips. While prior research has focused mainly on visual-based violence detection, this study addresses the underexplored area of real-time speech-based harassment detection. We present a multimodal safety framework that integrates OpenAI's Whisper for speech transcription with a fine-tuned DistilBERT model for toxicity classification, trained on the Jigsaw Toxic Comment Classification dataset. Our system achieves an impressive 93.8% accuracy, surpassing current state-of-the-art methods in toxic speech detection. While real-time capability is demonstrated through system design and latency evaluation, large-scale field trials remain future work. Designed for real-time processing, the framework enables proactive safety monitoring, making it ideal for ride-hailing and similar dynamic urban environments. This work contributes to the field by effectively combining automatic speech recognition and natural language processing for real-world safety applications. By bridging the gap between static datasets and live environments, our approach offers a practical, scalable, and impactful solution for enhancing passenger safety through real-time verbal abuse detection.

Keywords: Ride-hailing safety; Real-time violence detection; Speech-based analysis; Toxic comment classification; DistilBERT.

1. Introduction

The rapid proliferation of ride-hailing services has revolutionized urban transportation, offering convenience and accessibility to millions of users worldwide. However, this growth has also exposed significant safety concerns, particularly related to incidents of harassment, verbal abuse, and violence during trips. Our prior work on visual-based violence detection using YOLOv8 and TSM-ResNet50 successfully demonstrated the efficacy of computer vision in identifying physical aggression inside vehicles [1]. This previous research was not intended as a comprehensive solution but rather as a foundational step toward enhancing passenger safety. Recognizing that many threats manifest verbally before escalating into physical violence, the current study expands on this foundation by addressing the crucial gap of real-time detection of verbal harassment through speech-based analysis. Thus, our work aims to provide a more complete, multimodal safety framework for ride-hailing services that covers both visible violence and covert verbal abuse. Our system captures audio input via a microphone, transcribes it using OpenAI's Whisper, a robust speech recognition model based on transformer architecture, and then applies a fine-tuned transformer-based NLP model (DistilBERT) trained on the Jigsaw Toxic Comment Classification dataset [2]. This integration enables the real-time detection of verbal threats or harassment, triggering automated alerts or evidence recording without the need for human intervention.

Although our primary application involves ride-hailing speech, no publicly available dataset exists in this domain. Constructing a new dataset would require large-scale audio collection, transcription, and annotation, raising significant privacy, ethical, and scale challenges. For this reason, we selected the Jigsaw dataset, a widely used and well-annotated toxicity resource. While not ride-hailing–specific, its extensive coverage of harassment patterns (insults, threats, hate speech) provides a strong and transferable foundation for training. This makes it a practical choice for initial model development, even as we recognize the need for future domain-specific datasets.

Though various text classification datasets exist, many lack the contextual diversity relevant to in-vehicle interactions. For instance, the Hate Speech and Offensive Language Dataset [3] is primarily sourced from Twitter and lacks conversational context, while Davidson et al.'s Toxic Comment Dataset [4] focuses on racial or political hate speech and misses subtler forms of aggression. The Jigsaw dataset [2] provides over 150,000 annotated samples covering multiple toxicity subtypes—including toxic, severe_toxic, obscene, threat, insult, and identity_hate—offering fine-grained abuse categorization. For this study, these categories were consolidated into a single binary toxic label, making the dataset well-suited for developing and evaluating real-time toxic comment detection systems.

Previous studies on speech- and text-based aggression detection have faced limitations like reliance on offline datasets or absence of real-time processing. MacAvaney et al. [5] developed a multi-view SVM for hate speech detection on static Twitter datasets but explicitly noted its inability to process streaming audio or run on edge devices. Wiegand et al. [6] examined transformer-based hate speech detection on social media text but did not integrate speech recognition. Zhang et al. [7] proposed a BERT-based toxic speech classifier limited to written chat, without real-time spoken input. These gaps underscore the need for an end-to-end, live speech harassment detection system for practical deployment.

Our study fills this gap by continuously analyzing live microphone input with fast transcription and transformer classification. This real-time capability makes the system suitable for proactive safety monitoring in ride-hailing vehicles, public transport, and other mobile environments.

2. Related Work

Artificial Intelligence (AI) has demonstrated transformative impact across sectors such as finance, education, and healthcare, where it has been applied to fraud detection [8], adaptive learning [9], and diagnostics [10]. Within this broader AI landscape, Natural Language Processing (NLP) has emerged as a key subfield, enabling systems to understand and process human language. NLP has been widely adopted in sentiment analysis [11], fake news detection [12], and mental health monitoring [13], with transformer-based models like BERT and DistilBERT achieving high accuracy on text classification tasks. In parallel, speech recognition technologies have advanced through models like DeepSpeech [14] and attention-based voice activity detectors [15], powering real-time transcription, voice assistants, and emotion recognition systems. These applications highlight the growing capability of NLP to operate reliably in dynamic, real-world environments, laying the groundwork for safety-critical systems like the one proposed in this study. Recent studies have specifically addressed toxicity in spoken language, with [16] developing novel methods for detecting toxic speech patterns in real-world audio data. Multilingual approaches like [17] have evaluated hate speech detection across diverse languages, while [18] applied knowledge distillation to optimize toxicity classification efficiency. Practical implementations such as [19] demonstrate how intelligent systems can mitigate verbal harassment in dynamic environments, though challenges remain in dataset consistency as highlighted by [20]'s critical analysis of hate speech annotation frameworks. These applications highlight the growing capability of NLP to operate reliably in dynamic, real-world environments, laying the groundwork for safety-critical systems like the one proposed in this study.

3. Methodology

3.1. Dataset Preparation

This study leverages the Jigsaw Toxic Comment Classification Challenge dataset [2] , which contains over 150,000 Wikipedia talk page comments annotated for six distinct categories of toxicity: toxic, severe_toxic, ob-

scene, threat, insult, and identity_hate. Each comment may belong to multiple categories, making the original problem a multi-label classification task.

For the purpose of this study, we redefined the task as a binary classification problem, aiming to distinguish between toxic and non-toxic content. To achieve this, a new binary target label, toxic, was derived by aggregating the six original toxicity columns using a logical OR operation. This unified label marks a comment as toxic (1) if it belongs to any of the six categories, and non-toxic (0) otherwise.

After removing unnecessary columns and cleaning the dataset, we obtained a total of 159,571 valid comments, of which 16,225 (≈10.17%) were labeled as toxic and 143,346 (≈89.83%) as non-toxic—highlighting a significant class imbalanceTo address this, we experimented with several balancing strategies, including SMOTE and related oversampling techniques. However, these approaches either failed to improve performance or in some cases degraded it, likely due to the synthetic generation of non-contextual samples. In contrast, random undersampling proved most effective, as it reduced the imbalance while maintaining stable classification results. Based on this outcome, we constructed a partially balanced dataset by selecting 18,775 non-toxic comments alongside the 16,225 toxic comments, resulting in a final dataset of 35,000 samples.

This balanced dataset was then used for model development, supporting real-time toxic comment classification. The preprocessing pipeline included standard text normalization and tokenization using DistilBERT's tokenizer, with a maximum input length of 512 tokens to retain contextual integrity. This strategy allowed us to focus on the general presence of toxicity, rather than its specific subtypes, while ensuring computational efficiency and effectiveness in real-time applications.

3.2. Model Architecture and Implementation Approach

3.2.1. Model Architecture

For our toxic comment classification task, we adopted DistilBERT, as shown in Figure 1 [21], a distilled version of the original BERT model. DistilBERT was introduced by Sanh et al. as a smaller, faster, and more efficient transformer model while retaining much of BERT's performance [22]. The key concept behind DistilBERT is knowledge distillation, where a compact student model (DistilBERT) learns from a larger teacher model (BERT) to mimic its performance [22][23]. DistilBERT reduces the number of transformer layers from 12 to 6, eliminates token-type embeddings and the next sentence prediction (NSP) objective, and shrinks the hidden dimension size—all while preserving around 97% of BERT's original performance [22]. It maintains the transformer encoder structure based on multi-head self-attention followed by feed-forward neural networks, allowing it to capture long-range dependencies and semantic relationships in text effectively [24]. To enhance accessibility, we extended the system with Whisper, an automatic speech recognition (ASR) model developed by OpenAI. Whisper is based on an encoder-decoder transformer architecture and trained on a large-scale, multilingual, and multitask dataset [25][26]. In our system, the user speaks into a microphone, and Whisper transcribes the speech into text. This transcription is then forwarded to the DistilBERT-based classifier for toxicity prediction, enabling real-time toxic comment detection from spoken input.

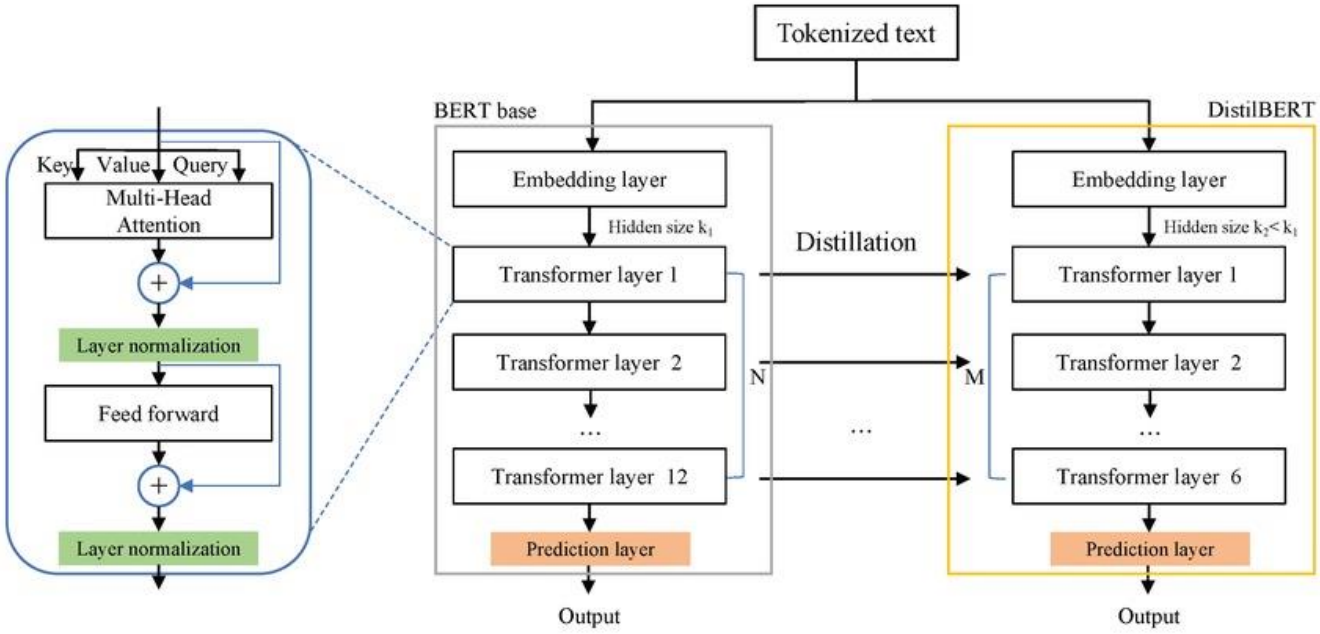

Figure 1. Comparison between BERT and DistilBERT architectures, highlighting the reduced complexity and faster inference of DistilBERT, which makes it more suitable for real-time mobile applications.

3.2.2 Implementation Approach

The implementation leverages the Hugging Face transformers and datasets libraries for efficient model training and data handling. We fine-tuned the pretrained distilbert-base-uncased model using the AutoModelForSequenceClassification class, configuring it for binary classification, where the model predicts whether a comment is toxic (1) or clean (0). The original multi-label toxicity annotations from the Jigsaw dataset were consolidated into a single binary label to simplify the classification task and enable real-time responsiveness.

The dataset was loaded and split into training and evaluation sets using an 80/20 ratio. Each comment was tokenized using the AutoTokenizer with padding and truncation applied to a maximum length of 512 tokens. The model outputs logits, which are passed through a softmax function to obtain class probabilities, and the class with the highest probability is selected as the predicted label.

We used Hugging Face's Trainer API, which simplified training by managing the optimization loop, evaluation scheduling, logging, and checkpointing. The model was trained for 5 epochs with a batch size of 16, applying weight decay for regularization. Evaluation was conducted at the end of each epoch using accuracy, precision, recall, and F1-score, which are well-suited for assessing binary classification performance and handling class imbalance.

For speech input, the system records 5 seconds of audio using the sounddevice library. The audio is normalized and transcribed using the Whisper base model. The resulting text is then tokenized and passed through the fine-tuned DistilBERT classifier. The output logits are converted to class probabilities using softmax, and the toxic class probability is evaluated against a 0.5 threshold. The final verdict is printed in real time, indicating whether the input speech was toxic or clean.

This pipeline demonstrates a practical and interactive approach to toxic content moderation, integrating natural language processing and speech recognition in a seamless user experience.

3.3 Evaluation Metrics

To assess classification performance, we employed four standard metrics, each providing complementary insights into model behavior:

1. Precision [27]: Quantifies the reliability of positive predictions by measuring the proportion of true positives among all predicted positives.

$$Precision\ (P) = \frac{TP}{TP + FP} \tag{1}$$

2. Recall (Sensitivity) [27]: Measures the model's ability to identify all relevant instances by calculating the proportion of true positives detected among all actual positives.

$$Recall\ (R) = \frac{TP}{TP + FN} \tag{2}$$

3. Accuracy [27]: Represents the overall correctness of predictions across both positive and negative classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

4. F1-Score [28]: Harmonizes precision and recall into a single metric, particularly valuable for imbalanced data.

$$F1\ Score = 2 \times \frac{P \times R}{P + R} \tag{4}$$

4. Results

This section presents the experimental evaluation of the proposed system, focusing on toxic content detection using the DistilBERT transformer model. The results highlight the model's performance across multiple evaluation metrics, including accuracy, precision, recall, and F1-score, demonstrating its effectiveness in identifying harmful content in text-based interactions. Performance trends over training epochs are also discussed to assess generalization and overfitting.

4.1 DistilBERT Results

4.1.1 Performance Metrics

DistilBERT demonstrated strong performance in toxic content classification. Table 1 summarizes the model's performance across both the validation and test sets after training for 5 epochs. The model maintained consistent accuracy, precision, recall, and F1-score, reflecting its effectiveness and generalization ability.

Table 1. DistilBERT Performance on Validation and Test Sets

Metric	Validation Set	Test Set
Loss	0.4834	0.1723
Accuracy	0.9356	0.9376
Precision	0.9303	0.9215
Recall	0.9334	0.9486
F1-Score	0.9319	0.9348

These results confirm that the model performs robustly on unseen data, achieving high recall and precision, which are crucial for minimizing both false negatives and false positives in toxic content detection tasks.

4.1.2 Graphical Analysis

To further understand the model's classification behavior, three important evaluation graphs were plotted: the Confusion Matrix, ROC Curve, and Precision-Recall Curve.

- Confusion Matrix

The confusion matrix provides insights into how well the classifier differentiates between classes as shown in Figure 2.

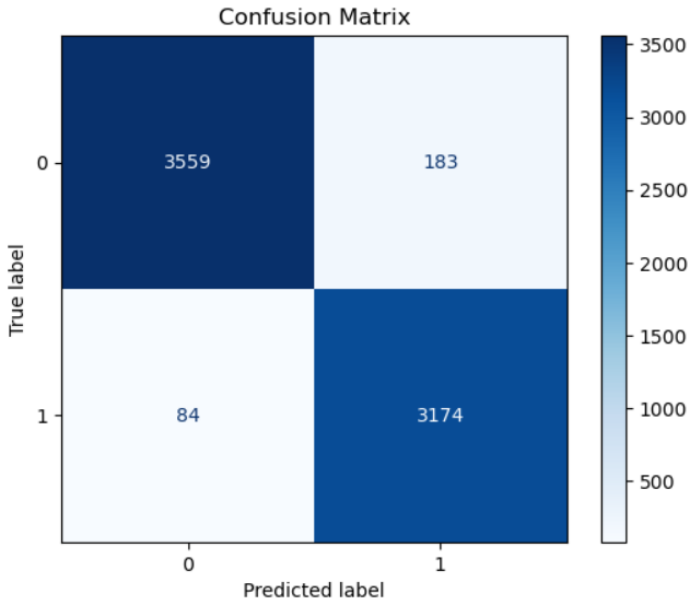

Figure 2. Confusion matrix of the proposed system on the test set, showing the distribution of true positives, true negatives, false positives, and false negatives for toxicity classification.

The high number of true positives and true negatives, with relatively few misclassifications, indicates that the model performs reliably and is balanced in detecting both toxic and non-toxic content.

- The ROC curve

The ROC curve, which plots the True Positive Rate against the False Positive Rate, shows an AUC of 0.99 — a near-perfect score. This demonstrates the model's exceptional ability to distinguish between toxic and non-toxic content.

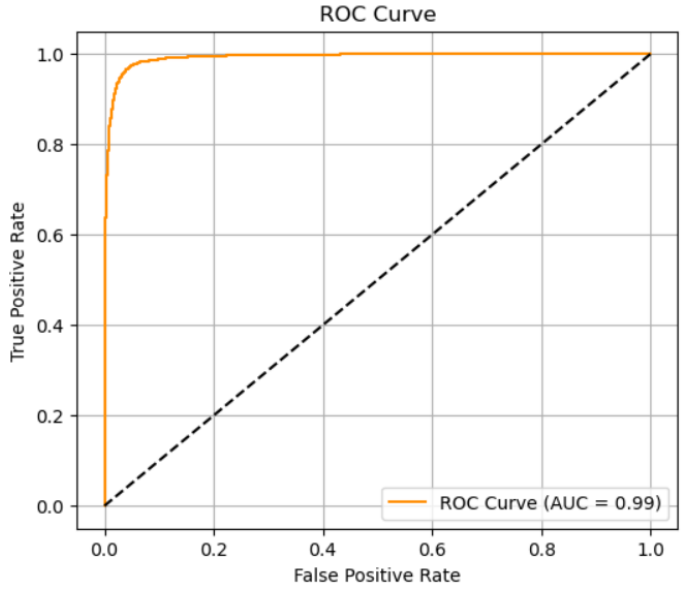

Figure 3. ROC curve of the proposed model, illustrating the trade-off between true positive rate and false positive rate, with a high AUC indicating strong performance.

With an AUC of 0.99, the ROC curve confirms the model's high accuracy in separating toxic from non-toxic comments, making it a reliable choice for binary text classification.

- The Precision-Recall curve

The Precision-Recall curve illustrates the balance between precision and recall across different thresholds. This type of curve is particularly informative when dealing with imbalanced datasets, as it focuses on the performance related to the positive class (in this case, toxic comments).

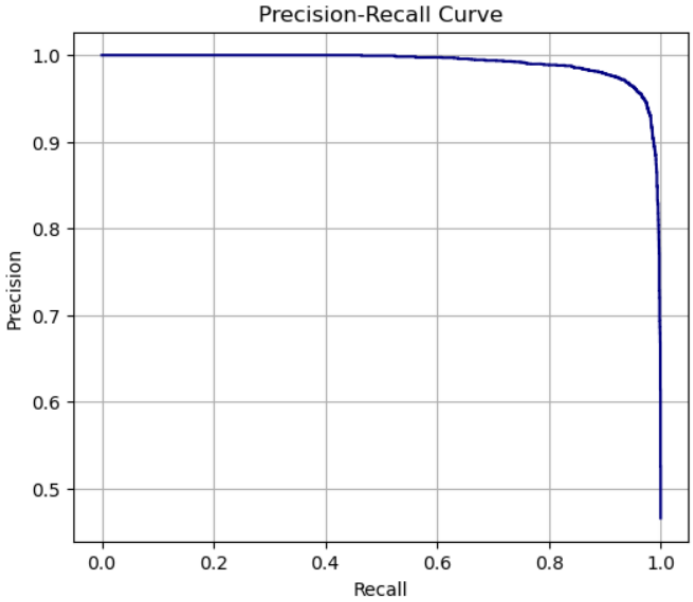

Figure 4. Precision–Recall curve of the proposed model, emphasizing its ability to maintain high precision and recall even in the presence of imbalanced data.

The smooth shape of the curve indicates that the model sustains high precision while also achieving strong recall. This confirms its effectiveness in correctly identifying toxic content without generating excessive false positives — a crucial quality in real-world moderation tasks where class imbalance is often present.

4.2 Sample of the Classification Results

This subsection showcases real-world examples of the model's predictions to demonstrate its behavior in diverse contexts. Each example includes the recognized speech input, raw model output logits, class probabilities, predicted toxicity score, and the final verdict. These samples highlight the model's sensitivity to aggressive or harmful language, as well as its ability to correctly classify neutral or polite speech.

For instance, in Figure 5, the phrase "Your car smells like shit!" received a toxicity probability of 99.95%, leading to a correct classification as toxic. Similarly, in Figure 6 and Figure 7, highly offensive or threatening phrases like "You picked the worst road idiots!" and "Shut up or I will drop you off here." were also flagged as toxic, with probabilities above 99%, confirming the model's robustness in handling harmful speech.

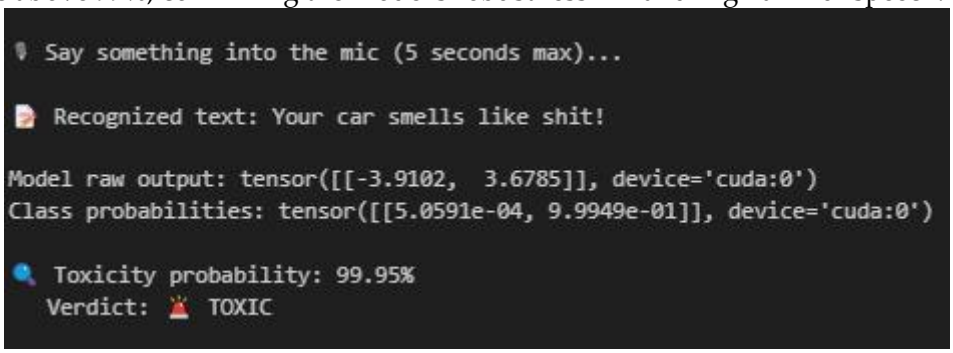

Figure 5. Example of an explicitly toxic comment ("Your car smells like shit!") transcribed by Whisper and classified by the system with 99.95% toxicity probability, demonstrating effective detection of profanity.

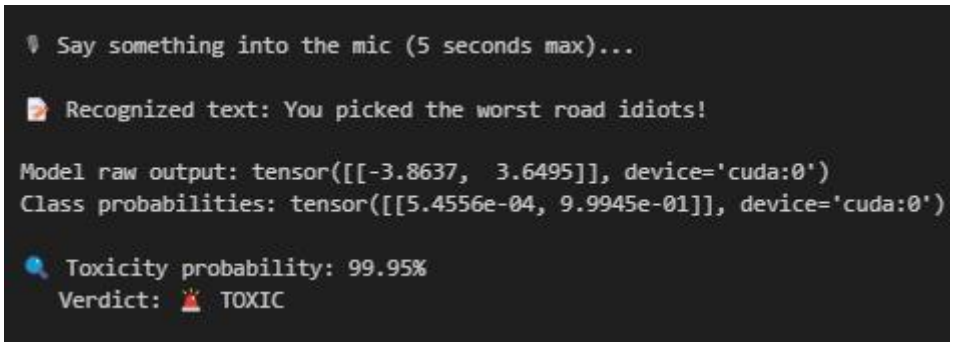

Figure 6. Example of explicit toxicity ("You picked the worst road idiots!") transcribed by Whisper and identified with 99.95% toxicity probability, showing reliable handling of direct insults.

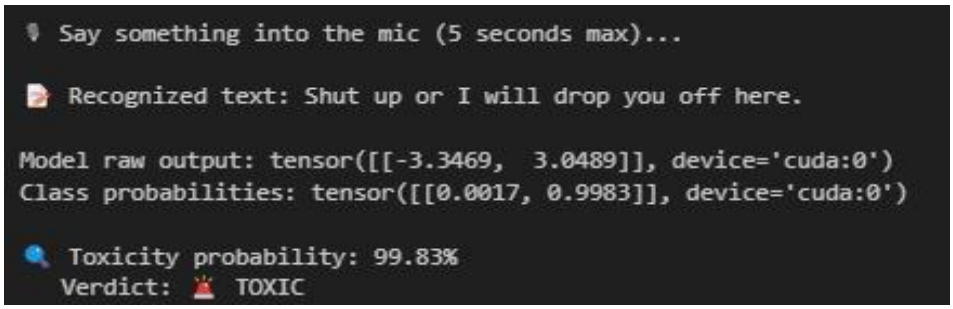

Figure 7. Example of a threatening statement ("Shut up or I will drop you off here.") transcribed by Whisper and classified with 99.83% toxicity probability, highlighting sensitivity to safety-critical language.

On the other hand, the model demonstrated strong precision in recognizing non-toxic utterances. As shown in Figure 8, the phrase "Hello, what's in the best road for you today?" received a toxicity score of just 0.15%, and was correctly classified as clean. Likewise, statements like "Accept card payments only please" in Figure 9 were also correctly identified as non-toxic.

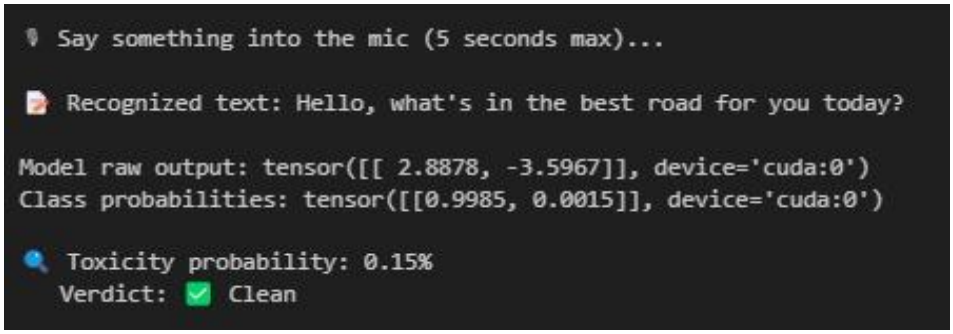

Figure 8. Example of a clean, non-toxic comment ("Hello, what's in the best road for you today?") transcribed by Whisper and classified with only 0.15% toxicity probability, showing robustness to polite speech.

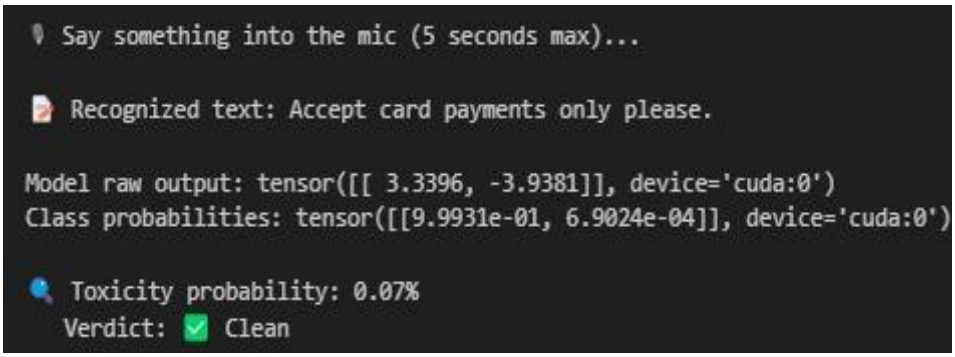

Figure 9. Another example of clean input ("Accept card payments only please.") transcribed by Whisper and classified with 0.07% toxicity probability, reinforcing reliability on neutral requests.

Interestingly, borderline cases like "You are the third cancellation today." in Figure 10 were labeled as toxic with a 75.4% probability. To clarify how such intermediate scores are interpreted, we propose a tiered response system for deployment. Utterances with very high toxicity probabilities (≥90%) would trigger immediate alerts or interventions, whereas moderate scores (50–90%) would be logged or flagged as soft warnings rather than treated as definitive harassment. This tiered framework directly addresses the real-world cost of false positives by ensuring that uncertain cases do not trigger disruptive actions, while still preserving valuable evidence for review. This showcases the model's cautious approach to language that might carry implicit aggression or dissatisfaction, even when profanity is not present.

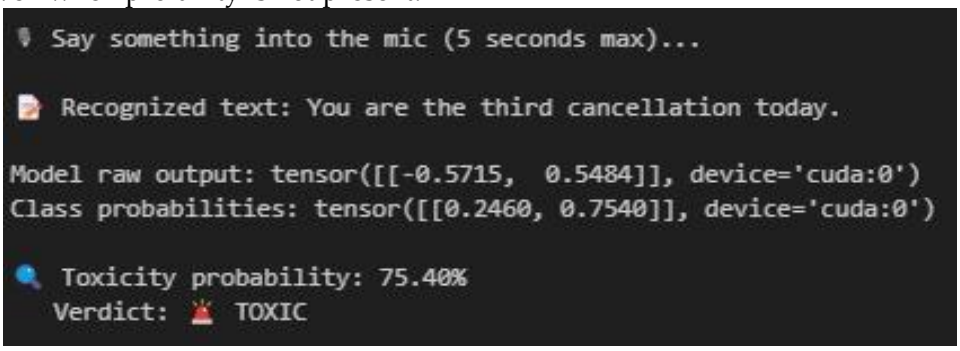

Figure 10. Example of a borderline case ("You are the third cancellation today.") transcribed by Whisper and assigned 75.40% toxicity probability. Under the proposed tiered response framework, this would be flagged as a soft warning rather than a critical incident, illustrating nuanced handling of implicit aggression.

These examples collectively demonstrate the model's capacity to differentiate between direct toxic language, neutral content, and nuanced statements that may carry latent hostility, making it suitable for real-time content moderation in dynamic environments like ride-sharing or in-vehicle systems.

4.3 Comparative Analysis

To evaluate the effectiveness of our model, we compare its performance with results reported in similar studies: Zaheri et al. (Southern Methodist University) [29], who evaluated Naïve Bayes and LSTM models for toxic comment classification; Kurita et al. (Carnegie Mellon University) [30], who explored robustness in toxic content classification using Logistic Regression, FastText, ELMo, BERT, and their proposed CDAE model under different noise conditions; and Magzoub (University of Twente) [31], who assessed CNN, LSTM, and SVM models for binary and multi-class toxic comment classification on Discord data. The following Table 2 presents a comparative overview of the performance metrics across these studies and our proposed model:

Table 2. Performance comparison of our model with state-of-the-art approaches in toxic comment classification

Model	Accuracy	Precision	Recall	F1-Score
Our DistilBERT Model	0.938	0.922	0.949	0.935
Zaheri et al. [29] - LSTM	-	0.81	0.66	0.73
Zaheri et al. [29] - NB	-	0.94	0.48	0.64
Kurita et al. [30] - BERT (None/None)	-	-	0.914	0.685
Kurita et al. [30] - FastText (None/None)	-	-	0.902	0.674
Magzoub [31] - CNN Binary	0.904	0.506	0.891	0.645
Magzoub [31] - SVM Binary	0.928	0.596	0.812	0.687

In terms of precision-recall trade-off, our model achieves a superior balance with a precision of 0.922 and recall of 0.949, outperforming all baseline models, including BERT (precision: N/A, recall: 0.914) from Kurita et al. [30] and SVM Binary (precision: 0.596, recall: 0.812) from Magzoub [31]. Notably, our F1-score of 0.935 demonstrates stronger harmonic mean performance compared to Zaheri et al.'s [29] LSTM (0.73), Kurita et al.'s [30] best BERT configuration (0.685), and Magzoub's [31] SVM (0.687). Furthermore, our accuracy (0.938) exceeds the highest reported accuracy in Magzoub [31] (0.928 with SVM Binary) while maintaining robustness across other metrics. These results highlight the efficacy of our method in toxic content classification, particularly in achieving high recall without compromising precision—a critical requirement for real-world moderation systems.

5. Discussion

The results of this study highlight the effectiveness of combining speech recognition (Whisper) and NLP (DistilBERT) for real-time verbal harassment detection in ride-hailing scenarios. Our model outperformed existing approaches, such as LSTM, Naïve Bayes, and even standard BERT, in terms of accuracy, precision, recall, and F1-score. This success can be attributed to the fine-tuning of DistilBERT on a consolidated binary toxicity label, which simplified the classification task while maintaining contextual integrity. One notable strength of our system is its ability to handle nuanced and borderline cases, such as implicit aggression without profanity, as demonstrated by the 75.4% toxicity probability for the comment, "You are the third cancellation today." This sensitivity is crucial for real-world applications where subtle verbal cues may precede escalation. However, the system's cautious approach could also lead to false positives, which may require further refinement to balance sensitivity and specificity. In practical deployment, these borderline classifications could be managed using adaptive thresholds, where repeated borderline utterances gradually escalate in severity, or by incorporating a human-in-the-loop review process for ambiguous cases. Such measures ensure that uncertain predictions do not immediately trigger disruptive interventions, while still preserving evidence for oversight.

The integration of Whisper for real-time transcription ensures seamless processing of spoken input, addressing a key limitation of prior work that relied on static text datasets. This makes our solution practical for deployment in mobile environments. Nonetheless, challenges remain, such as handling multilingual contexts and noisy audio conditions, which could be explored in future research. Comparisons with state-of-the-art models underscore the superiority of our approach, particularly in achieving high recall without sacrificing precision. This balance is critical for safety applications, where missing a genuine threat (false negative) is more consequential than a false alert. By combining high recall with operational safeguards for managing borderline cases, the system can maintain sensitivity to potential risks while limiting the practical cost of false positives.

The system's performance was evaluated using the Jigsaw dataset, which is based on Wikipedia comments and is one of the largest publicly available resources for toxicity detection. Although this dataset is not ride-hailing–specific, it was selected because of its extensive annotation quality, large scale, and wide coverage of harassment patterns (including insults, threats, and aggressive language). To better align it with the ride-hailing safety context, we applied undersampling to balance the classes and consolidated the multiple toxicity labels into a single binary label (toxic vs. non-toxic). This preprocessing simplified the task while retaining the richness of the toxic language patterns. We acknowledge, however, that a domain gap remains between written online comments and spoken in-vehicle interactions, and future research should focus on bridging this gap through domain-specific data collection and adaptation.

5.1 Ethical Considerations and Privacy Compliance

Beyond technical performance, the deployment of a continuous audio monitoring system raises important ethical and privacy considerations. Any real-world implementation must ensure explicit user consent and transparency, allowing riders and drivers to opt in to the monitoring feature. To safeguard privacy, we propose a limited retention policy: audio is securely stored only if an incident is reported and deleted within 24 hours otherwise. In both cases, recordings would be anonymized and protected by secure storage protocols to prevent misuse. These measures are consistent with principles from privacy frameworks such as GDPR and are essential to balance the safety benefits of real-time monitoring with respect for user autonomy and data protection.

6. Conclusions

This study introduces an effective real-time speech-based violence detection system designed to enhance safety in ride-hailing services. By integrating Whisper for speech transcription and a fine-tuned DistilBERT model for toxicity classification, the system demonstrates strong performance and practical applicability. Notably, it surpasses traditional and widely used models—including LSTM, Naïve Bayes, and standard BERT—highlighting the advantages of using a lightweight transformer model fine-tuned for binary toxicity detection. Beyond its accuracy, the system can identify both overt and subtle forms of verbal harassment, making it especially suitable for dynamic and sensitive environments. Looking ahead, future work includes expanding support for multilingual input, addressing dialect variability, and improving robustness in noisy conditions. At present, the classifier remains restricted to English due to the use of the Jigsaw dataset, and extending the system to other languages will require multilingual or cross-lingual training resources. A key limitation of this work is the absence of domain adaptation and in-field validation. Future efforts should focus on bridging the gap between written online comments and spoken in-vehicle conversations through domain adaptation, crowdsourcing of ride-hailing speech data, or partnerships with ride-hailing providers to obtain anonymized, consent-based datasets. Overall, this work advances real-time safety monitoring by bridging the gap between research models and real-world deployment.

References

1. Waleed, M.; Yasser, M.; Magdy, M.; Elkady, M.; Ayman, M.; Hesham, T.; Adel, A.; Sherif, Y.; Alazzaly, M.; Abdelatif, A.; Abdelsalam, K. AI Models for Real-Time Violence Detection. *Proceedings of the 9th International Conference on Advanced Machine Learning Technologies and Applications (AMLTA'25)*, Springer, accepted for publication, 2025.
2. Jigsaw Toxic Comment Classification Challenge Dataset. Available online: https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge (accessed on 15 May 2025).
3. Waseem, Z.; Hovy, D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, San Diego, CA, USA, 12–17 June 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 88–93. DOI: 10.18653/v1/N16-2013.
4. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, Montreal, Canada, 15–18 May 2017; AAAI Press: Palo Alto, CA, USA, 2017; pp. 512–515. DOI: 10.1609/icwsm.v11i1.14955.
5. MacAvaney, S.; Yao, H.-R.; Yang, E.; Russell, K.; Goharian, N.; Frieder, O. Hate Speech Detection: Challenges and Solutions. *PLoS ONE* 2019, 14, e0221152. DOI: 10.1371/journal.pone.0221152.
6. Wiegand, M.; Ruppenhofer, J.; Kleinbauer, T. Detection of Abusive Language: The Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1 (Long and Short Papers), pp. 602–608. DOI: 10.18653/v1/N19-1060.
7. Zhang, Z.; Robinson, D.; Tepper, J. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *The Semantic Web – ESWC 2018*, Proceedings of the 15th Extended Semantic Web Conference, Heraklion, Greece, 3–7 June 2018; Gangemi, A., Presutti, V., Recupero, D., et al., Eds.; Lecture Notes in Computer Science, Vol. 10843; Springer: Cham, Switzerland, 2018; pp. 745–760. DOI: 10.1007/978-3-319-93417-4_48.
8. West, J.; Bhattacharya, M. Intelligent Financial Fraud Detection: A Comprehensive Review. *Comput. Secur.* 2016, 57, 47–66. DOI: 10.1016/j.cose.2015.09.005.
9. Gligorea, I.; Cioca, M.; Oancea, R.; Gorski, A.-T.; Gorski, H.; Tudorache, P. Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review. *Educ. Sci.* 2023, 13, 1216. DOI: 10.3390/educsci13121216.

10. Al-Antari, M.A. Artificial Intelligence for Medical Diagnostics—Existing and Future AI Technology! *Diagnostics* **2023**, 13, 688. DOI: 10.3390/diagnostics13040688.
11. Medhat, W.; Hassan, A.; Korashy, H. Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Eng. J.* 2014, 5, 1093–1113. DOI: 10.1016/j.asej.2014.04.011.
12. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 2017, 19, 22–36. DOI: 10.1145/3137597.3137600.
13. Ogunleye, B.; Sharma, H.; Shobayo, O. Sentiment Informed Sentence BERT-Ensemble Algorithm for Depression Detection. *Big Data Cogn. Comput.* **2024**, 8, 112. DOI: 10.3390/bdcc8090112.
14. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *arXiv* 2015, arXiv:1512.02595. DOI: 10.48550/arXiv.1512.02595.
15. Zhang, H.; Huang, H.; Han, H. Attention-Based Convolution Skip Bidirectional Long Short-Term Memory Network for Speech Emotion Recognition. *IEEE Access* **2021**, 9, 5332-5342. DOI: 10.1109/ACCESS.2020.3047395
16. Nada, A.H.A.; Latif, S.; Qadir, J. Lightweight Toxicity Detection in Spoken Language: A Transformer-Based Approach for Edge Devices. *arXiv* 2023, arXiv:2304.11408. DOI: 10.48550/arXiv.2304.11408.
17. Corazza, M.; Menthi, S.; Cabrio, E.; Tonelli, S.; Villata, S. A Multilingual Evaluation for Online Hate Speech Detection. *ACM Trans. Internet Technol.* **2020**, 20, 10, pp. 1 – 22. DOI: 10.1145/3377323.
18. Gupta, B. Classification of Toxic Comments using Knowledge Distillation. Master's Thesis, National College of Ireland, Dublin, Ireland, **2023**. Available online: https://norma.ncirl.ie/id/eprint/6132 (accessed on 16 May 2025).
19. Patel, J. Combatting Toxicity: Designing an Intelligent System to Diminish Verbal Harassment in Online Games. Master's Thesis, OCAD University, Toronto, Canada, **2025**. Available online: https://openresearch.ocadu.ca/id/eprint/4773 (accessed on 16 May 2025).
20. Fortuna, P.; Soler, J.; Wanner, L. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France, 11–16 May 2020; Calzolari, N., Béchet, F., Blache, P., et al., Eds.; European Language Resources Association: Paris, France, 2020; pp. 6786–6794. Available online: https://aclanthology.org/2020.lrec-1.838 (accessed on 16 May 2025).
21. Adel, H.; Dahou, A.; Mabrouk, A.; Abd Elaziz, M.; Kayed, M.; El-Henawy, I.M.; Alshathri, S.; Ali, A.A. Improving Crisis Events Detection Using DistilBERT with Hunger Games Search Algorithm. *Mathematics* **2022**, 10, 447. DOI: 10.3390/math10030447
22. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv* **2019**, arXiv:1910.01108. DOI: 10.48550/arXiv.1910.01108.
23. Efimov, V. Large Language Models: DistilBERT - Smaller, Faster, Cheaper and Lighter. *Towards Data Science* **2020**. Available online: https://towardsdatascience.com/distilbert-11c8810d29fc (accessed on 20 May 2025).
24. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Volume 1 (Long and Short Papers)*, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. DOI: 10.48550/arXiv.1810.04805.
25. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv 2022, arXiv:2212.04356. DOI: 10.48550/arXiv.2212.04356.
26. OpenAI Whisper GitHub Repository. Available online: https://github.com/openai/whisper (accessed on 20 May 2025).
27. Sokolova, M.; Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Inf. Process. Manag.* **2009**, 45, 427–437. DOI: 10.1016/j.ipm.2009.03.002.
28. Powers, D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *arXiv* **2020**, arXiv:2010.16061. DOI: 10.48550/arXiv.2010.16061.
29. Zaheri, S.; Leath, J.; Stroud, D. Toxic Comment Classification. *SMU Data Sci. Rev.* **2020**, 3, 13. Available online: https://scholar.smu.edu/datasciencereview/vol3/iss1/13 (accessed on 22 May 2025).
30. Kurita, K.; Belova, A.; Anastasopoulos, A. Towards Robust Toxic Content Classification. *arXiv* **2019**, arXiv:1912.06872. DOI: 10.48550/arXiv.1912.06872
31. Magzoub, Z. Toxic Comment Classification in Discord. B.S. Thesis, University of Twente, Enschede, The Netherlands, 2023. Available online: https://purl.utwente.nl/essays/94373 (accessed on 22 May 2025).