Beyond Content: A Comprehensive Speech Toxicity Dataset and Detection Framework Incorporating Paralinguistic Cues

Anonymous submission

Abstract

Toxic speech detection has become a crucial challenge in maintaining safe online communication environments. However, existing approaches to toxic speech detection often neglect the contribution of paralinguistic cues, such as emotion, intonation, and speech rate, which are key to detecting speech toxicity. Moreover, current toxic speech datasets are predominantly text-based, limiting the development of models that can capture paralinguistic cues. To address these challenges, we present ToxiAlert-Bench, a large-scale audio dataset comprising over 30,000 audio clips annotated with seven major toxic categories and twenty fine-grained toxic labels. Uniquely, our dataset annotates toxicity sources—distinguishing between textual content and paralinguistic origins—for comprehensive toxic speech analysis. Furthermore, we propose a dual-head neural network with a multi-stage training strategy tailored for toxic speech detection. This architecture features two task-specific classification headers: one for identifying the source of sensitivity (textual or paralinguistic), and the other for categorizing the specific toxic type. The training process involves independent head training followed by joint fine-tuning to reduce task interference. To mitigate data class imbalance, we incorporate class-balanced sampling and weighted loss functions. Our experimental results show that leveraging paralinguistic features significantly improves detection performance. Our method consistently outperforms existing baselines across multiple evaluation metrics, with a 21.1% relative improvement in Macro-F1 score and a 13.0% relative gain in accuracy over the strongest baseline, highlighting its enhanced effectiveness and practical applicability.

Introduction

Toxic speech, as part of toxic behaviors, can occur virtually and physically, resulting in a negative psychological impact (Nada, Latif, and Qadir 2023). "Toxic speech" often includes hostile intent that is threatening, abusive, discriminatory, etc. (Garg et al. 2023; Fortuna, Soler, and Wanner 2020). Such behavior can target individuals or groups, leading to severe outcomes such as cyberbullying, harassment, and the spread of discriminatory ideas.

Voice-based social platforms have surged in recent years, amplifying the risk of spreading toxicity through audio. Popular social platforms like Twitter and Facebook have mature text-based content moderation systems (Zeng et al. 2024; Inan et al. 2023). However, voice-involving social platforms for live streaming, multiplayer online gaming, or voice/video chatting, such as Twitch and Slack, require more than text-based moderation (Hamilton, Garretson, and Kerne 2014). Researchers have discovered that when it comes to detecting intents embedded within audio signals, purely text-based models are not sufficient (Lin and Emmanouilidou 2022). The combination of verbal and non-verbal cues can express some toxic intentions. For example, explicit adult content can be conveyed through non-verbal cues such as moaning or Autonomous Sensory Meridian Response (ASMR)-like sounds, which can bypass text-based detection. Therefore, *how to leverage paralinguistic cues for effective detection of toxic speech urges more study.*

Recent studies have realized the importance of acoustic features other than semantic content in toxic speech identification. Relevant studies can be divided into three categories: Generic Acoustic-Based (Yousefi and Emmanouilidou 2021; Ghosh et al. 2022), Feature Fusion-Based (Lin and Emmanouilidou 2022; Rana and Jha 2022; Mandal et al. 2024), and Textual Task-Assisted Multi-task Learning (Nada, Latif, and Qadir 2023; Liu et al. 2024b; Kumar Nandwana et al. 2024).

Despite these advances, several critical limitations hinder progress in paralinguistic-aware toxic speech detection.

Lack of Suitable Datasets. The scarcity of publicly available datasets poses a significant barrier to research development. Among existing works, only DeToxy (Ghosh et al. 2022) has released a public dataset (i.e., DeToxy-B), but its toxicity classification is solely based on textual content. It lacks samples where toxicity originates from paralinguistic cues (Scherer, London, and Wolf 1973) alone or from both textual and paralinguistic sources combined. Consequently, existing public datasets are insufficient to support the development of detection systems capable of identifying paralinguistic-based toxicity. While some studies focus on non-textual information of audio signals as sources of toxicity and develop corresponding detection systems, their datasets remain private and undisclosed, with unclear and non-transparent construction methodologies.

Technical Limitations of Existing Methods. Existing *MTL and acoustic feature-based approaches* are highly dependent on textual information. This text-dependency bias may inherently limit the applicability of these methods when semantics are benign but paralinguistic properties, such as intonation and emotion, convey harmful intent. Current *feature fusion methods* focus on combining specific acoustic dimensions, which may miss subtle paralinguistic toxic signals that exist beyond their explicitly extracted dimensions. Some *acoustic feature-based methods* rely on traditional handcrafted features for toxicity detection, which may fail to capture rich dimensions representing harmful intent. DeToxy (Ghosh et al. 2022) applies Self-Supervised Learning (SSL) (Liu et al. 2022; Gong et al. 2022) pre-trained foundation models, but underutilizes their representational capabilities through simple feature extraction without sophisticated architectural design. Moreover, DeToxy focuses solely on textual content analysis, failing to address toxicity that originates from paralinguistic sources.

Evaluation and Reproducibility Limitations. The benchmarking practices in this field are incomplete and inconsistent. Existing works primarily compare against their own baselines, lacking broader evaluations. The lack of open-source code further hinders reproducibility and collaboration in the field. These limitations collectively create substantial obstacles to advancing research in paralinguistic-aware toxic speech detection, highlighting the urgent need for comprehensive datasets, transparent methodologies, and reproducible evaluation frameworks.

In this work, we address these limitations from both data and methodological perspectives:

(1) We develop and open-source ToxiAlert-Bench, the first large-scale toxic speech dataset specifically designed for paralinguistic-aware detection, comprising over 60 hours of annotated audio clips. It features comprehensive toxicity source annotations, including four distinct categories: safe for both textual and paralinguistic sources, textually toxic but paralinguistically safe, textually safe but paralinguistically toxic, and toxic for both sources. The dataset encompasses seven toxic categories and a safe category, with twenty fine-grained toxic labels. We systematically document and open-source the complete dataset construction pipeline, enabling reproducible research and facilitating future dataset-building studies in this domain.

(2) We propose a novel dual-head neural network architecture built upon pre-trained SSL foundation models for robust toxic speech detection. Our model leverages large-scale pre-trained representations to capture both semantic and paralinguistic features effectively. The architecture incorporates two specialized classification heads with a multi-stage training strategy. To address data imbalance challenges, we integrate class-balanced sampling and weighted loss functions. Extensive experimental results, including benchmarking and ablation results, validate the effectiveness of our model architecture and training strategies.

Our contributions are summarized as follows:

1. We fill the gap in the research domain of toxic speech detection, with the documentation and open source ToxiAlert-Bench, a comprehensive paralinguistic-aware toxic speech dataset.
2. We design a dual-head speech detection framework, employing a multi-stage training strategy with class-balanced sampling, weighted loss functions, and sequential head-specific training followed by joint fine-tuning.
3. Through comprehensive benchmarking against established baselines, including DeToxy-B and state-of-the-art (SOTA) multimodal large language models (MLLM), our approach demonstrates significant improvements, achieving a 21.1% relative improvement in Macro-F1 and a 13.0% relative gain in accuracy over the strongest baseline.

Related Work

Textual Content Moderation

Conventional toxic speech detection often ignores the non-verbal properties of speech signals. Some early content moderation (CM) methods heavily depend on manual examination, which is costly and non-scalable. Platforms commonly employ automated CM to ensure that content aligns with behavioral standards by removing inappropriate posts and spam. Most moderators on social platforms utilize conventionally text-based frameworks (Lin and Emmanouilidou 2022; Nada, Latif, and Qadir 2023; Koratana and Hu 2018). They identify whether a post or comment contains toxic information by analyzing the textual features.

Audio-Based Toxic Speech Detection

Studies in this area can be divided into three categories. **General Acoustic Features:** Yosefi et al. (Yousefi and Emmanouilidou 2021) propose a self-attentive Convolutional Neural Networks framework to detect audio-based toxic language. DeToxy (Ghosh et al. 2022) proposes to use acoustic features (F-Bank and wav2vec2.0) for classification. Liu et al. (Liu et al. 2024a) propose a cross-modal learning to incorporate semantic information of text into audio feature representatives, facilitating speech toxicity classification only requiring audio. **Multi-Task Learning with text information:** Nada et al. (Nada, Latif, and Qadir 2023) applies an Automatic Speech Recognition (ASR) task to assist toxicity detection. Liu et al. (Liu et al. 2024b) predict the toxicity labels of a speech signal with the assistance of text information alignment. Nandwana et al.(Kumar Nandwana et al. 2024) utilize multi-task learning to predict the toxicity of speech with the assistance of an auxiliary Audio Keyword detection task. **Feature Fusion:** Lin et al. (Lin and Emmanouilidou 2022) explore the relationship between speech emotion and toxic speech and propose a framework combining speech emotion recognition (SER) and audio-based CM models. Rana et al. (Rana and Jha 2022) combines acoustic features representing emotion and text features for hate speech detection. Attentive fusion (Mandal et al. 2024) fuses audio and text representation for hate speech identification.

ToxiAlert-Bench

Overview

ToxiAlert-Bench is a comprehensive English toxic speech dataset comprising 32,561 audio samples totaling 60.82 hours. It uniquely combines both real-world and synthesized audio, including 19745 samples from established speech corpora and 12,816 samples specifically synthesized for

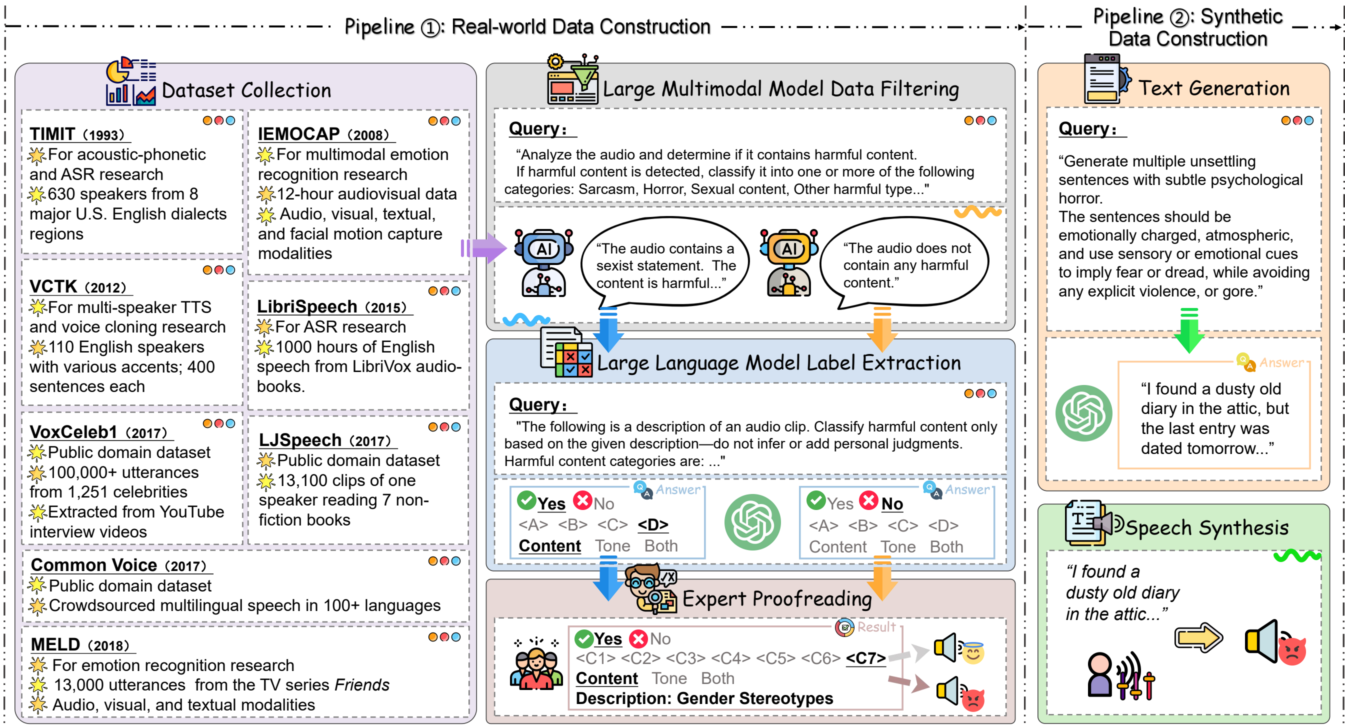

Figure 1: Overview of the ToxiAlert-Bench dataset construction framework. Pipeline1 (left) illustrates the collection and annotation process for bonafide real-world audio data. Pipeline2 (right) depicts the generation of synthetic toxic speech.

toxicity analysis. To facilitate rigorous experimentation, ToxiAlert-Bench is split into training, validation, and test sets using a 7:1:2 ratio. Each audio sample is annotated with three key attributes: (1) toxicity classification across 7 major categories—**C1: Sarcasm, C2: Horror, C3: Sexual, C4: Mental Health & Risk Behavior**(Mental & Risk), **C5: Political & Ideological Sensitivity**(Ideology), **C6: Violence & Harm, and C7: Discrimination** plus a **Safe** category; (2) toxicity source identification, distinguishing between textual-only toxic, paralinguistic-only toxic, both textual and paralinguistic toxic, or safe, and (3) fine-grained categorization, using 20 specific toxicity labels within the taxonomy. Notably, our dataset focuses on paralinguistic toxicity detection—6,728 samples exhibit toxicity solely through paralinguistic cues, addressing a critical gap in existing datasets that primarily focus on textual toxicity.

Compared to the existing DeToxy-B dataset (Ghosh et al. 2022), ToxiAlert-Bench demonstrates advantages in both scale and annotation comprehensiveness. While DeToxy-B contains 20,271 utterances (24.4 hours), ToxiAlert-Bench includes 60% more samples and 150% more total duration. Importantly, DeToxy-B defines toxicity purely based on textual content, with all 5,077 toxic samples labeled through text-only analysis. In contrast, ToxiAlert-Bench introduces a novel toxicity source annotation framework, explicitly labeling samples as textual-only toxic (6,953), paralinguistic-only toxic (6,728), both textual and paralinguistic toxic (2,551), and safe content (16,329). This granular source-based annotation enables researchers to develop models capable of identifying paralinguistic toxicity. Furthermore, while DeToxy-B does not provide detailed toxicity type labels, our dataset provides comprehensive coverage with 7 major toxic categories and 20 fine-grained labels, supporting more nuanced toxic speech research. Please refer to Appendix A for more details on ToxiAlert-Bench.

Dataset Construction Framework

Bonafide Data Sources. We collect bonafide speech samples from eight datasets widely accepted in the domain of speech-related studies. They are: (1) TIMIT (Garofolo et al. 1993), (2) IEMOCAP (Busso et al. 2008), (3) VCTK (Veaux et al. 2017), (4) LibriSpeech (Panayotov et al. 2015), (5) VoxCeleb1 (Nagrani, Chung, and Zisserman 2017), (6) LJSpeech-1.1 (Ito and Johnson 2017), (7) CommonVoice (Ardila et al. 2019), and (8) MELD (Poria et al. 2018). For each one, we collect both non-toxic and toxic samples, and these samples are further categorized with a multi-stage annotation pipeline.

Annotation Pipeline for Bonafide Data. Our annotation pipeline for bonafide data involves a systematic multi-stage approach, as shown in Figure 1. First, we employ two large multimodal models, Gemini-1.5-Flash (Team 2024) and R1-AQA (Li et al. 2025)[1], for initial data filtering and preliminary toxicity assessment. Each audio sample is pro-

[1]R1-AQA is based on Qwen2-Audio-7B-Instruct (Chu et al. 2024), optimized through reinforcement learning (RL), achieving SOTA performance on the MMAU benchmark (Sakshi et al. 2024) with only 38k post-training samples.

cessed using a structured query that determines whether contains harmful content(see Appendix B for the complete prompt and question-answer pairs). If detected, classifies it into one of four predefined categories: Sarcasm, Horror, Sexual Content, and other harmful categories. During this stage, both multimodal models analyze each audio sample and its description across three critical dimensions: toxicity (toxic/non-toxic), major toxic category, and source. When both models reach consensus on all three aspects, the corresponding labels are automatically assigned to the sample. However, if disagreement occurs between the models on any dimension, the sample is forwarded to the Expert Proofreading stage for human validation. This dual-model consensus mechanism ensures automated screening efficiency while maintaining annotation reliability.

In the next phase, we leverage GPT-4o (Achiam et al. 2023) to extract label suggestions based on audio descriptions, followed by expert annotation. Human annotators assess each sample for undergoes detailed evaluation: (1) whether the content contains harmful elements, (2) the source of toxicity (textual content, paralinguistic cues, or both), and (3) the specific toxic category. This multi-layered approach guarantees that our toxicity source distinctions and category labels accurately reflect both semantic content toxicity and paralinguistic cues such as sarcastic tone, threatening intonation, or manipulative speech patterns.

Synthesized Data Construction. To enhance diversity and ensure comprehensive coverage of paralinguistic toxicity patterns, we implement a synthetic data generation pipeline, as illustrated in the right portion of Figure 1. We utilize the text-to-speech (TTS) method for synthesizing data (Eskimez et al. 2024; Chen et al. 2024; Anastassiou et al. 2024).

Our synthesis begins with GPT-4o generating emotionally charged sentences across toxic categories like psychological horror and subtle sexual tension. Carefully designed prompts (see Appendix B) emphasize subtle and context-dependent expressions of toxicity while avoiding explicit language. GPT-4o's strong safety mechanisms ensure all generated sentences are non-toxic on the textual level, allowing paralinguistic cues to be the sole carriers of toxicity.

Following text generation, we use DubbingX (DubbingX 2025) to synthesize audio. The TTS engine features extensive character personality configurations, enabling the speech generation with distinct vocal styles by simply providing the input text and selecting a specific character role. We strategically select personas likely to produce speech with different toxic paralinguistic characteristics. The synthesis process enables us to produce naturalistic speech with varying paralinguistic features, ensuring that the toxicity of the resulting audio manifests itself through non-textual cues, such as intonation, rhythm, and emotional expression.

Fine-grained Toxicity Categorization. Following the initial annotation pipeline for bonafide data, we obtain four broad categories from the multimodal model consensus or expert proofreading stages. However, category D ("other harmful") lacks specificity for detailed toxicity analysis. To refine this, we further establish fine-grained labels.

Each Class-D audio sample is first annotated by humans with detailed descriptions of its toxic characteristics. We then employ unsupervised clustering algorithms(Likas, Vlassis, and Verbeek 2003) to group similar descriptions, followed by manual review to correct errors (see Appendix A for clustering details). This hybrid clustering approach results in 20 distinct toxic labels plus one safe category. These labels are organized as 7 major toxic classes with hierarchical grouping accomplished by a human annotator.

The above procedures are applied to both bonafide and synthetic speech samples, ensuring the consistency of toxic type classification across both bonafide and synthetic portions of ToxiAlert-Bench.

ToxiAlert

To address the limitations of existing approaches, we propose ToxiAlert, a unified detection model designed to identify toxic speech where toxicity may arise from textual content, paralinguistic cues, or their combination.

Design Principles

We desire to utilize the multi-dimensional information of speech signals to perform toxicity detection. Inspired by recent advancements in the domain of deepfake detection (Tak et al. 2022), SOTA methods explore the use of self-supervised learning to obtain better representations trained on diverse speech data and other tasks with only bonafide samples for the purpose of generalization improvement. The pre-trained SSL model, combined with a classifier, is then finetuned with the downstream task dataset, achieving leading performance. Specifically, we adopt Wav2Vec 2.0 (Baevski et al. 2020) as the speech encoder $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^d$, where input audio waveform $x \in \mathcal{X}$ is mapped to a latent representation:

$$\mathbf{h} = f_\theta(x) \tag{1}$$

This representation $\mathbf{h} \in \mathbb{R}^{T \times d}$ is passed through two classification heads:

- **Source Head** ($g_\phi^{(s)}$): This is a multi-label classification head designed for toxicity source identification. It predicts whether the toxicity in the audio arises from *textual content*, *paralinguistic cues*, or *both*.

$$\hat{\mathbf{y}}^{(s)} = \sigma(g_\phi^{(s)}(\mathbf{h})) \tag{2}$$

  where $\hat{\mathbf{y}}^{(s)} \in [0, 1]^2$ represents the independent probabilities assigned to the two binary toxicity sources—textual and paralinguistic. $\sigma(\cdot)$ represents the element-wise *sigmoid* activation function.

- **Category Head** ($g_\phi^{(c)}$): This is a multi-class classification head designed for toxic category classification. It determines the specific type of toxicity in the input audio.

$$\hat{\mathbf{y}}^{(c)} = \text{softmax}(g_\phi^{(c)}(\mathbf{h})) \tag{3}$$

  where $\hat{\mathbf{y}}^{(c)} \in [0, 1]^K$ the softmax-normalized likelihoods across $K = 8$ mutually exclusive classes, including seven toxic categories and one safe category.

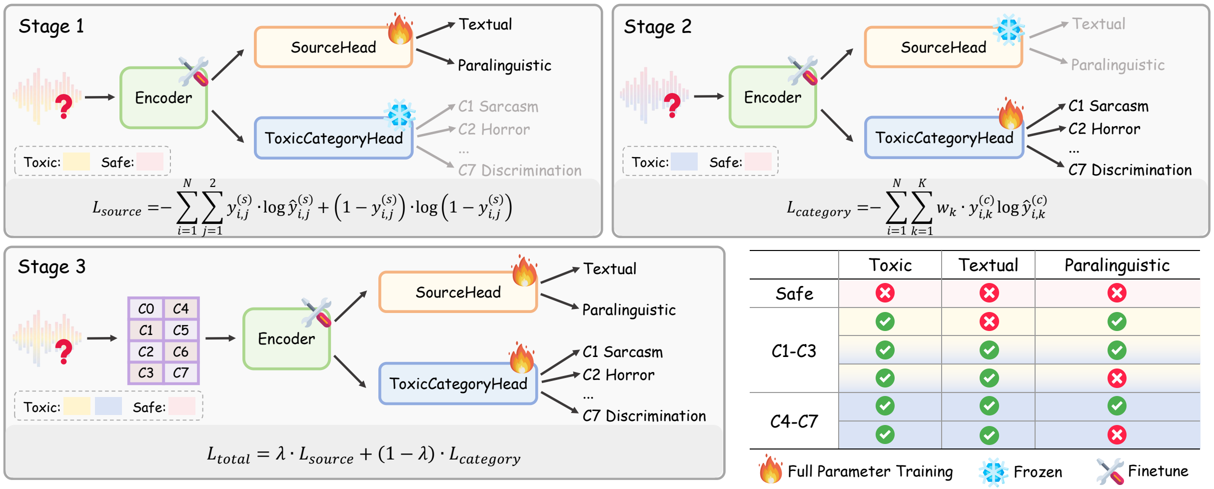

Figure 2: Overview of the ToxiAlert training framework. Multi-Stage Training Strategy: Stage 1 trains the source head to detect toxicity sources; Stage 2 trains the category head for toxicity classification; Stage 3 jointly fine-tunes both heads.

Multi-Stage Training Strategy

Our training method employs a multi-stage approach designed to optimize both task-specific performance and inter-task coordination, as illustrated in Figure 2. Let $\mathcal{D}^{(s)}$, $\mathcal{D}^{(c)}$, and $\mathcal{D}^{(full)}$ represent the datasets used in each stage.

Stage 1: Source Head Training. We first train the source head $g_\phi^{(s)}$, freezing the category head $g_\phi^{(c)}$. This stage focuses on learning to distinguish between textual and paralinguistic toxicity sources. $\mathcal{D}^{(s)}$ has all data from categories "Sarcasm," "Horror," and "Sexual," which exhibit diverse source characteristics. To achieve class balance, we supplement the toxic samples with safe samples equivalent to approximately 1/3 of the total toxic sample count from these three categories. The objective is a binary cross-entropy loss:

$$\mathcal{L}_{\text{source}} = -\sum_{i=1}^{N}\sum_{j=1}^{2}\left[y_{i,j}^{(s)} \log \hat{y}_{i,j}^{(s)} + (1 - y_{i,j}^{(s)}) \log(1 - \hat{y}_{i,j}^{(s)})\right] \quad (4)$$

where $y_{i,j}^{(s)} \in \{0, 1\}$ are the ground-truth binary labels.

Stage 2: Category Head Training. In the second stage, we freeze the source head $g_\phi^{(s)}$ and train the category head $g_\phi^{(c)}$ on $\mathcal{D}^{(c)}$, which includes textually toxic but paralinguistically safe samples, allowing the model to focus on textual discrimination. We add safe speech samples of approximately 1/7 of the toxic subset to maintain balance. The training minimizes a weighted cross-entropy loss:

$$\mathcal{L}_{\text{category}} = -\sum_{i=1}^{N}\sum_{k=1}^{K} w_k \cdot y_{i,k}^{(c)} \log \hat{y}_{i,k}^{(c)} \quad (5)$$

where w_k is the inverse frequency of class k for balancing, and $y_{i,j}^{(c)} \in \{0, 1\}$ are one-hot toxicity category labels.

Stage 3: Joint Fine-tuning. The final stage performs end-to-end joint training of both heads using the complete dataset $\mathcal{D}^{(full)}$, and a composite objective function is optimized:

$$\mathcal{L}_{\text{total}} = \lambda \cdot \mathcal{L}_{\text{source}} + (1 - \lambda) \cdot \mathcal{L}_{\text{category}} \quad (6)$$

where $\lambda = 0.2$ to reflect the relatively auxiliary nature of the source task.

To ensure training stability and mitigate label imbalance, we employ a class-balanced sampler that selects m samples per category for every batch of size $B = m \cdot K$. In our experiments, we use $m = 3$, resulting in $B = 24$.

Experiments

Settings

Baselines. We compare ToxiAlert with several SOTA open-source and commercial systems. DeToxy and NetEase Yidun Audio Moderation API (YIDUN) (NetEase 2025) are specifically built for toxic speech detection. DeToxy is an open-source toxicity classifier, while YIDUN is a commercial platform supporting real-time moderation in multiple languages. In contrast, Qwen2-Audio, GPT-4o Audio, and Gemini-2.5-Flash (Comanici et al. 2025) are general-purpose MLLMs not explicitly trained for toxicity detection, these models have demonstrated strong capabilities in speech comprehension and multimodal reasoning, due to their large parameter scale and training on vast datasets.

Evaluation Setup. We train ToxiAlert and DeToxy on the ToxiAlert-Bench training set and directly evaluate other baselines. Toxicity classification performance is assessed at both the category level (7 toxic categories) and the label level (20 fine-grained labels). In addition to overall performance, we emphasize a challenging subset of the benchmark where the toxicity is conveyed solely through paralinguistic cues. This setting remains underexplored in prior work, yet it is highly relevant for real-world applications. For generalization evaluation, all models are tested on the DeToxy-B test set[2]. Evaluation prompts are detailed in Appendix B.

[2]After excluding CMU-MOSEI, CMU-MOSI, MSP-Improv, MSP-Podcast, Social-IQ, and SwitchBoard due to their discontin-

Model	Sarcasm	Horror	Sexual	Mental & Risk	Ideology	Violence & Harm	Discrim.	ACC	Macro-F1	Binary ACC
DeToxy	-	-	-	-	-	-	-	-	-	85.70
YIDUN	-	-	0.50	-	0.50	0.65	-	-	-	50.49
Qwen2-Audio	4.42	0.00	12.21	0.00	2.51	26.83	9.73	55.15	19.24	60.41
Gemini-2.5-Flash	53.00	58.89	34.32	47.15	21.61	61.64	36.19	70.84	57.55	75.38
GPT-4o Audio	27.08	12.22	20.17	29.27	18.09	34.88	21.01	61.89	39.91	64.52
ToxiAlert	**81.10**	**90.94**	**81.85**	**48.78**	**52.76**	**65.95**	**39.30**	**80.04**	**69.69**	**86.33**

Table 1: Comparison of ToxiAlert with baselines on ToxiAlert-Bench. We report per-category accuracy across seven toxicity categories. Note that models without category-level predictions leave corresponding entries blank (-).

		Label-Level				Sample-Level
Model	**Type**	ACC	F1	Macro F1	Micro F1	Subset ACC
Qwen2	Para.	71.84	3.79	19.28	20.72	55.35
	Tex.	77.00	34.77			
Gemini	Para.	69.48	19.57	31.11	31.31	52.90
	Tex.	77.48	42.66			
GPT-4o	Para.	71.50	0.32	13.81	15.04	53.20
	Tex.	75.06	27.30			
ToxiAlert	**Para.**	**91.18**	**83.30**	**79.48**	**79.34**	**80.21**
	Tex.	**86.21**	**75.66**			

Table 2: Comparison of model performance on the source identification task. Both label-level and sample-level results are reported.

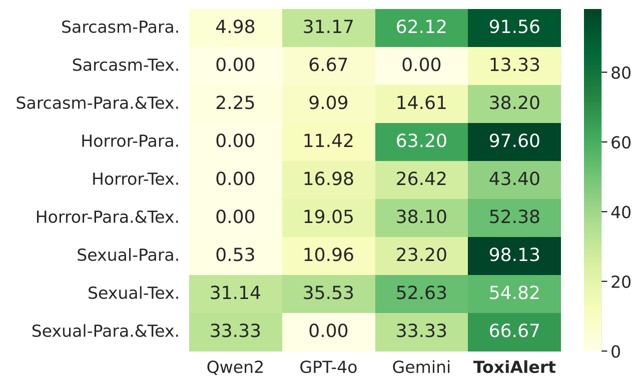

Figure 3: Performance comparison on source-specific toxicity detection across three toxicity types and three source settings. ACC denotes per-class accuracy.

Metrics. To comprehensively evaluate model performance, we adopt metrics from two tasks: (1) Toxicity category classification, reporting overall accuracy, per-category accuracy, and Macro-F1 to capture both global and class-specific performance. For binary classifiers like DeToxy, we compute binary accuracy by merging all toxic classes for fair comparison. (2) Toxicity source identification is formulated as a multi-label task, evaluated with label-level metrics—accuracy, F1 score, Macro-F1, and Micro-F1. We also report subset accuracy at the sample-level, which measures the percentage of samples with all labels predicted correctly.
Implementation Details. We adopt wav2vec2-large-960h as the audio encoder, followed by three fully connected layers for toxicity classification and source identification. All audio samples are resampled to 16kHz, converted to mono, and truncated to a maximum length of 25 seconds. All experiments are conducted on NVIDIA A100 GPUs using the PyTorch framework.

Toxic Speech Classification

Category Level: As shown in Table 1, ToxiAlert consistently achieves the best overall performance across all 7 toxicity categories. Compared with the strongest baseline, Gemini-2.5-Flash, it improves Macro-F1 by 21.1% and overall accuracy by 13.0%. While DeToxy reports high binary accuracy, but being a binary classifier, it lacks the capacity to distinguish between toxicity types, making it less suitable for fine-grained moderation. In contrast, ToxiAlert delivers both higher binary accuracy and comprehensive multi-class prediction.
Label Level: Given that Gemini-2.5-Flash achieves the best performance at the category level, we adopt it as the baseline for assessing fine-grained classification capabilities. As shown in Figure 4, ToxiAlert outperforms Gemini-2.5-Flash on the majority of labels. These improvements suggest that ToxiAlert is better equipped to distinguish subtle differences among overlapping or co-occurring toxic behaviors.

Source-Specific Toxicity Detection

We further assess model performance under varied source conditions of toxic expression. Specifically, we focus on three challenging categories—Sarcasm, Horror, and Sexual, and evaluate classification accuracy when the toxic signal is conveyed through paralinguistic cues (Para.), textual content (Tex.), or both (Para.&Tex.). Results are presented in Figure 3. ToxiAlert consistently outperforms all baselines across all categories and source types. In cases where toxic intent is expressed exclusively through Para., ToxiAlert achieves 91.56% on Sarcasm, 97.60% on Horror, and 98.13% on Sexual. In contrast, baselines show notable performance degradation, as they typically overlook non-verbal signals during training or inference.

ued open access, the test set contains 2,035 samples.

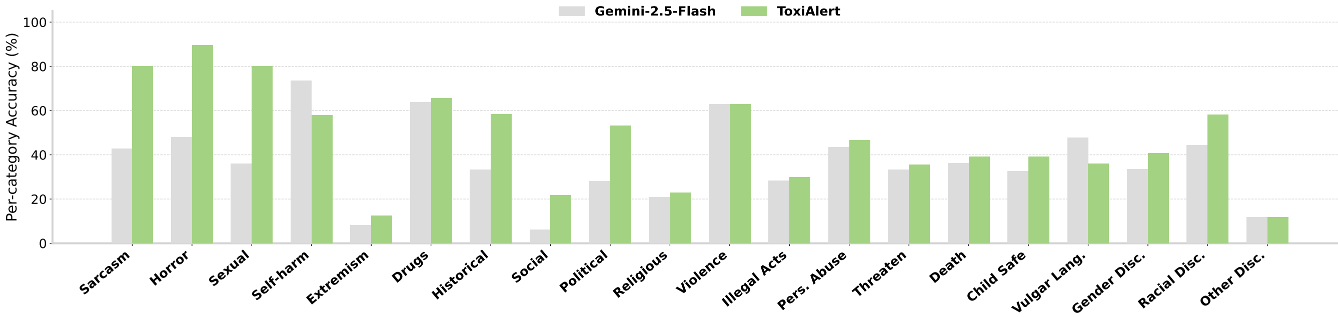

Figure 4: Fine-grained comparison of ToxiAlert and Gemini-2.5-Flash on ToxiAlert-Bench. We report per-category accuracy across twenty fine-grained toxicity labels, spanning all seven major toxicity categories.

Model	Balanced ACC	F1-Binary	Toxic ACC
DeToxy	66.95	50.33	67.78
YIDUN	49.97	0.40	0.20
Qwen2-Audio	52.91	12.50	6.88
Gemini-2.5-Flash	59.89	37.17	29.47
GPT-4o Audio	69.20	54.32	48.51
ToxiAlert	**72.29**	**55.83**	**80.94**

Table 3: Comparison of model generalization performance. Balanced ACC mitigates the effect of class imbalance; Toxic ACC is the accuracy on toxic samples.

Model	**Toxic Cls.**			**Source ID**	
	ACC	Macro-F1	Binary ACC	Macro-F1	Subset ACC
w/o SourceHead	75.04	66.01	81.67	–	–
w/o Multi-stage	78.25	68.79	84.72	78.35	77.80
w/o Sampler	78.34	68.00	85.47	79.05	79.51
ToxiAlert	**80.04**	**69.69**	**86.33**	**79.48**	**80.21**

Table 4: Ablation study on the effectiveness of ToxiAlert components. Performance is shown for toxicity category classification (Toxic Cls.) and toxicity source identification (Source ID).

Toxicity Source Identification

We assess the model's ability to identify toxicity sources—textual or paralinguistic. As shown in Table 2, ToxiAlert consistently outperforms all baselines. For paralinguistic cues, which are inherently subtle and challenging to detect, ToxiAlert achieves an accuracy of 91.18% and an F1 score of 83.30%, significantly surpassing all competing models. For textual content sources, it also delivers strong results with 86.21% accuracy and 75.66% F1 score. Moreover, ToxiAlert achieves the highest overall performance on Subset Accuracy, improving over the strongest baseline by 44.9%, underscoring its robustness in capturing both explicit and implicit forms of toxic expression.

Generalization Evaluation

To assess the generalization ability of ToxiAlert, we evaluate it on DeToxy-B. For fair comparison, we train a binary version of ToxiAlert using the training set of ToxiAlert-Bench and evaluate it directly on the DeToxy-B test set without any additional fine-tuning. As shown in Table 3, ToxiAlert surpasses the best-performing baseline, GPT-4o Audio, by 4.5% in balanced accuracy, 2.8% in F1 score, and 32.9% in accuracy on toxic samples. These results demonstrate that ToxiAlert generalizes effectively to out-of-distribution data.

Ablation Study

To investigate the impact of each core component in ToxiAlert, we conduct an ablation study focusing on three core modules: the dual-head architecture, the multi-stage training strategy, and the class-balanced sampler. As shown in Table 4, removing any of these components results in a noticeable performance decline.

Removing the source head leads to a significant drop in classification accuracy and Macro-F1, which drops to 75.04% and 66.01%, respectively. Moreover, the model is no longer capable of performing source identification, highlighting the necessity of this joint modeling approach. Removing the multi-stage training strategy results in consistent degradation across metrics, with overall accuracy reduced to 78.25% and subset accuracy to 77.80%. These results underscore the importance of progressive training in improving convergence and generalization. Finally, without the class-balanced sampler, performance in both tasks degrades. This result highlights the importance of structured sampling in improving model stability in multi-class classification tasks.

Conclusion

In this work, we propose ToxiAlert, a paralinguistic-toxic-aware speech toxicity dataset, the first in the area, featuring various combinations of toxicity sources and capable of facilitating the development of a more comprehensive toxic speech detection system. In addition, we propose an SSL-based model capable of predicting the toxic/safe binary labels, the specific major toxicity category, and the toxicity source. The model features our dual-head design and multi-stage training strategy, outperforming existing academic methods and commercial MLLM-based solutions.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anastassiou, P.; Chen, J.; Chen, J.; Chen, Y.; Chen, Z.; Chen, Z.; Cong, J.; Deng, L.; Ding, C.; Gao, L.; et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Baevski, A.; Zhou, H.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359.

Chen, Y.; Niu, Z.; Ma, Z.; Deng, K.; Wang, C.; Zhao, J.; Yu, K.; and Chen, X. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; Zhou, C.; and Zhou, J. 2024. Qwen2-Audio Technical Report. *arXiv preprint arXiv:2407.10759*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DubbingX. 2025. DubbingX TTS.

Eskimez, S. E.; Wang, X.; Thakker, M.; Li, C.; Tsai, C.-H.; Xiao, Z.; Yang, H.; Zhu, Z.; Tang, M.; Tan, X.; et al. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 682–689. IEEE.

Fortuna, P.; Soler, J.; and Wanner, L. 2020. 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*, 6786–6794.

Garg, T.; Masud, S.; Suresh, T.; and Chakraborty, T. 2023. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s): 1–32.

Garofolo, J. S.; Lamel, L. F.; Fisher, W. M.; Fiscus, J. G.; and Pallett, D. S. 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93: 27403.

Ghosh, S.; Lepcha, S.; Sakshi, S.; Shah, R. R.; and Umesh, S. 2022. DeToxy: A Large-Scale Multimodal Dataset for Toxicity Classification in Spoken Utterances. In *Interspeech 2022*, 5185–5189.

Gong, Y.; Lai, C.-I.; Chung, Y.-A.; and Glass, J. 2022. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10699–10709.

Hamilton, W. A.; Garretson, O.; and Kerne, A. 2014. Streaming on twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1315–1324.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Ito, K.; and Johnson, L. 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/.

Koratana, A.; and Hu, K. 2018. Toxic speech detection. *URL: https://web. stanford. edu/class/archive/cs/cs224n/cs224n*, 1194.

Kumar Nandwana, M.; He, Y.; Liu, J.; Yu, X.; Shang, C.; Du Bois, E.; McGuire, M.; and Bhat, K. 2024. Voice Toxicity Detection Using Multi-Task Learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 331–335.

Li, G.; Liu, J.; Dinkel, H.; Niu, Y.; Zhang, J.; and Luan, J. 2025. Reinforcement Learning Outperforms Supervised Fine-Tuning: A Case Study on Audio Question Answering. *arXiv preprint arXiv:2503.11197*.

Likas, A.; Vlassis, N.; and Verbeek, J. J. 2003. The global k-means clustering algorithm. *Pattern recognition*, 36(2): 451–461.

Lin, W.-C.; and Emmanouilidou, D. 2022. Toxic Speech and Speech Emotions: Investigations of Audio-based Modeling and Intercorrelations. In *2022 30th European Signal Processing Conference (EUSIPCO)*, 115–119.

Liu, J.; Li, G.; Zhang, J.; Dinkel, H.; Wang, Y.; Yan, Z.; Wang, Y.; and Wang, B. 2024a. Enhancing Automated Audio Captioning via Large Language Models with Optimized Audio Encoding. In *Interspeech 2024*, 1135–1139.

Liu, J.; Nandwana, M. K.; PylkkÃ¤nen, J.; Heikinheimo, H.; and McGuire, M. 2024b. Enhancing multilingual voice toxicity detection with speech-text alignment. *arXiv preprint arXiv:2406.10325*.

Liu, S.; Mallol-Ragolta, A.; Parada-Cabaleiro, E.; Qian, K.; Jing, X.; Kathan, A.; Hu, B.; and Schuller, B. W. 2022. Audio self-supervised learning: A survey. *Patterns*, 3(12).

Mandal, A.; Roy, G.; Barman, A.; Dutta, I.; and Naskar, S. K. 2024. Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection. arXiv:2401.10653.

Nada, A. H. A.; Latif, S.; and Qadir, J. 2023. Lightweight Toxicity Detection in Spoken Language: A Transformer-based Approach for Edge Devices. arXiv:2304.11408.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

NetEase. 2025. NetEase Yidun: AI-Powered Business Security Platform. https://dun.163.com/. Accessed: 2025-08-01.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210. IEEE.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Rana, A.; and Jha, S. 2022. Emotion Based Hate Speech Detection using Multimodal Learning. arXiv:2202.06218.

Sakshi, S.; Tyagi, U.; Kumar, S.; Seth, A.; Selvakumar, R.; Nieto, O.; Duraiswami, R.; Ghosh, S.; and Manocha, D. 2024. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. arXiv:2410.19168.

Scherer, K. R.; London, H.; and Wolf, J. J. 1973. The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality*, 7(1): 31–44.

Tak, H.; Todisco, M.; Wang, X.; weon Jung, J.; Yamagishi, J.; and Evans, N. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. arXiv:2202.12233.

Team, G. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.

Veaux, C.; Yamagishi, J.; MacDonald, K.; et al. 2017. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 6: 15.

Yousefi, M.; and Emmanouilidou, D. 2021. Audio-based Toxic Language Classification using Self-attentive Convolutional Neural Network. In *2021 29th European Signal Processing Conference (EUSIPCO)*, 11–15.

Zeng, W.; Liu, Y.; Mullins, R.; Peran, L.; Fernandez, J.; Harkous, H.; Narasimhan, K.; Proud, D.; Kumar, P.; Radharapu, B.; et al. 2024. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*.

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

---

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) yes

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) yes

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) yes

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) partial

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) partial

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) no

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) no

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) partial