



# ToxicTone: A Mandarin Audio Dataset Annotated for Toxicity and Toxic Utterance Tonality

Yu-Xiang Luo\*, Yi-Cheng Lin\*, Ming-To Chuang\*, Jia-Hung Chen, I-Ning Tsai, Pei Xing Kiew, Yueh-Hsuan Huang, Chien-Feng Liu, Yu-Chen Chen, Bo-Han Feng, Wenze Ren, Hung-yi Lee

National Taiwan University, Taiwan

{b10902037, f12942075, r13942091, hungyilee}@ntu.edu.tw

## Abstract

Despite extensive research on toxic speech detection in text, a critical gap remains in handling spoken Mandarin audio. The lack of annotated datasets that capture the unique prosodic cues and culturally specific expressions in Mandarin leaves spoken toxicity underexplored. To address this, we introduce ToxicTone—the largest public dataset of its kind—featuring detailed annotations that distinguish both forms of toxicity (e.g., profanity, bullying) and sources of toxicity (e.g., anger, sarcasm, dismissiveness). Our data, sourced from diverse real-world audio and organized into 13 topical categories, mirrors authentic communication scenarios. We also propose a multimodal detection framework that integrates acoustic, linguistic, and emotional features using state-of-the-art speech and emotion encoders. Extensive experiments show our approach outperforms text-only and baseline models, underscoring the essential role of speech-specific cues in revealing hidden toxic expressions.

**Index Terms:** Toxicity detection; Mandarin Chinese; Annotation; Ensemble

**Warning:** This paper may contain uncomfortable content.

## 1. Introduction

Toxic speech in online media is a serious global problem that creates hostile environments, discourages participation, and silences vulnerable voices. For individuals, exposure to toxic speech can cause psychological problems like stress, anxiety, and depression [1]. This concern is more serious on social platforms, where users—especially those from marginalized groups—are often targeted by harmful language, resulting in exclusion and emotional harm [2]. Over time, repeated exposure can cause long-term mental health issues like trauma and reduced self-esteem.

Although previous research has focused on toxicity detection in English and other well-supported languages, progress for Mandarin Chinese remains slow, mainly because large labeled datasets are scarce [3]. In addition, most work focuses on text input and does not address the unique challenges of spoken communication (for example, voice chats and audio streams). Furthermore, Mandarin includes unique colloquialisms and code-switching influenced by local languages [4], meaning that detection models must account for culture-specific toxic expressions [5]. Existing Chinese datasets also often lack detailed labels (for instance, separating toxic speech from general profanity or marking hidden insults [6]), limiting the creation of culturally aware and strong toxicity detection solutions.

In this paper, we tackle these problems by building a large Mandarin Chinese toxic speech dataset, ToxicTone, which we

believe is the biggest public resource of its kind. Unlike past datasets, ours includes detailed labels on the source of toxicity, capturing not only explicit or hateful words but also prosodic features (intonation, emphasis, rhythm) that show harmful intent yet cannot be seen from text alone. By showing how aggression emerges through vocal cues, ToxicTone uncovers toxic content that may be hidden behind seemingly polite words.

We also show that combining multiple model architectures trained on our dataset leads to better performance in toxic speech detection than text-only approaches. In particular, combining speech, emotion, and semantic encoders achieves the best results. This finding highlights the need to include prosodic and emotional information in speech, and supports the idea that we need special, speech-focused methods.

Our main contributions are as follows:

- We release the largest public Mandarin Chinese toxic speech detection dataset, with detailed labels on both the form and the source of toxicity<sup>1</sup>.
- We demonstrate that combining acoustic, emotional, and semantic features significantly boosts toxicity detection accuracy, showing that speech data is essential for dealing with toxic content.

By filling these data and modeling gaps, we aim to support safer online spaces and advance spoken-language research on toxicity detection, both for Mandarin Chinese and for broader uses in speech analytics.

## 2. Related work

Previous works on Chinese toxic speech detection focus mainly on text. COLA [7] represents the first Chinese offensive language classification dataset, comprising 18.7k comments sourced from YouTube and Weibo. The dataset categorizes texts into four classes: neutral, insulting, antisocial, and illegal. TOCP [8], which focuses on Chinese profanity, contains 16k toxic comments collected from the PTT Bulletin Board and Twitch livestream chatrooms. Similarly, COLD [3] serves as the first Chinese offensive language detection benchmark, consisting of 37k offensive language samples and anti-bias content related to race, gender, and region. Lastly, ToxiCN [6] provides 12k hierarchical annotation for texts from Zhihu or Tieba, including toxic type, targeted group, and expression.

These datasets are limited to text-based content, which does not capture the complexity of spoken language. Unlike spoken utterances, written text lacks prosodic features such as intonation, pitch, and stress, which can convey subtle expressions of toxicity or sarcasm. Furthermore, text datasets from social platforms cannot reflect real-world interactions, especially sponta-

\*Equal contribution

<sup>1</sup><https://github.com/YuXiangLo/ToxicTone>

neous language, emotion, and interruption.

The currently available datasets for detecting toxic speech include Detoxy [9], Mutox [10], and ADIMA [11]. Detoxy labels subsets of pre-existing speech datasets, such as CMU-MOSI [12], Common Voice [13], and Switchboard [14], by determining whether the samples are toxic or non-toxic. Notably, all the samples in Detoxy are in English, which limits the dataset’s ability to generalize to other languages. On the other hand, Mutox annotates multilingual segments derived from SeamlessAlign [15] and Common Voice. However, the dataset includes only 2,000 samples per language, and the audio primarily originates from podcast recordings. This focus limits its representation of real-world conversational scenarios, such as phone calls, live-streamed gameplay, or drama. The ADIMA dataset, focused on abuse detection in 10 Indic languages, provides a diverse multilingual approach but is limited to detecting profanity.

### 3. Dataset collection

#### 3.1. Definition of toxicity

We define toxicity via two aspects: the **form** of toxicity and the **source** of toxicity.

**Forms of toxicity:** These describe the specific manifestations of harmful or offensive language. They include:

- Profanities (Prof.): Offensive words that demonstrate disrespect or negativity. For example, *fuck, bastard, sissy, idiot*.
- Hate speech (Hate): Language that explicitly or implicitly expresses hostility, discrimination, or hatred toward groups based on their inherent or perceived characteristics. For example, *bitch, XX party dog*.
- Pornographic language (Porn. Lang.): Language that refers to sexual acts or body parts in a vulgar way, primarily intended to arouse sexual interest or evoke shock. For example, *cum, tits, boobs, cock, pussy*.
- Bullying speech (Bully): Threatening, offensive, or aggressive speech that does not target a specific group. For example, *shut up, you’ll die, you suck*.
- Sarcasm (Sarc.): Utterances that convey a meaning opposite or significantly different from the literal words used, often degrading or mocking the target. For example, *You are so smart, you are a genius*.
- Other toxic speech (Oth. Tox.): Language that does not fit the above categories, but still makes the listener feel disgusted or uncomfortable.

**Sources of toxicity:** These refer to the origin of the harmful intent or tone in communication. They include:

- Specific Words (Spec.): Use of explicitly aggressive or offensive words that carry direct insults or defamatory meanings.
- Angry or Violent Tone (Ang./Viol.): Speech that directly expresses anger with emotional or provocative content, potentially implying violent actions.
- Dismissive or Impatient Tone (Dism./Imp.): A tone marked by derogatory adjectives or dismissiveness, often appearing indifferent or unfriendly.
- Sarcastic or Satirical Tone (Sarc./Satir.): A condescending tone used to mock or ridicule the target, often through double entendre or implied meanings.
- Explicit/Implicit Threatening Tone (Threat.): Speech that directly or indirectly intimidates the target, causing mental or emotional fear.

#### 3.2. Preprocessing

We used web-crawled speech data as the basis of our research. After downloading the audio recordings, we employed a speaker diarization model<sup>2</sup> to differentiate and segment speech from multiple speakers, following [16]. Subsequently, the segmented audio was transcribed into text using the K2D model [17] and splitting the results into 2–10 second clips.

Given the enormous number of generated segments, a preliminary filtering step was required. To this end, we applied a text-based toxicity classifier from Alibaba-pai<sup>3</sup>—based on Chinese BERT-base [18]—to the transcriptions. The classifier assigns a toxicity score between 0 (no toxicity) and 1 (very toxic), and only segments scoring above 0.75 were retained for subsequent human annotation and analysis, because this value effectively balances the need to filter out non-toxic content while capturing segments that are likely to contain significant toxic elements. This filtering reduced the total number of clips from 770k to 52k.

In addition, because our preliminary filtering seldom detected speech with explicit sexual content, we employed a rule-based word list<sup>4</sup> to extract an additional 600 samples that potentially exhibit pornographic toxicity.

#### 3.3. Human Annotation

The dataset was annotated by a team of 11 native Chinese annotators. All annotators are informed that they might encounter uncomfortable audio content, and they can quit the task at any time. For each sample, annotators could select one or more forms of toxicity and their corresponding sources, or indicate that the sample was non-toxic. Annotators would also filter out the audio not in Mandarin. Initially, each sample was annotated by four annotators. Samples receiving an equal number of “toxic” and “non-toxic” annotations were subsequently reviewed and annotated by an additional annotator to resolve discrepancies.

The final dataset was separated into train, development, and test sets. The statistics of our dataset are depicted in Table 1 and Figure 1. Compared to other datasets in Table 2, our dataset is the largest publicly available toxic speech detection dataset.

The labels are imbalanced in our dataset both in the forms and sources of toxicity. For example, there are nearly 5,800 bullying speech samples, while pornographic language has fewer than 400 samples. Similarly, samples labeled with specific words appear almost 8,000 times, while those with a threatening tone occur only about 560 times. This imbalance reflects the real-world distribution of toxic speech. Research shows that some forms of toxic language—such as casual profanity and bullying—are more common in everyday online interactions because they are often tolerated or even normalized in many communities [19, 20]. The samples of the ToxicTone dataset are in the supplementary material.

#### 3.4. Category distribution

Motivated by the observation that the topical focus of a speech segment can influence both the prevalence and expression of toxic language, we divide our dataset into 13 topical categories. These categories—Society & News, Technology & Science, Education, Gaming, Entertainment, Culture & Arts, Psychology & Lifestyle, Movie & Book Reviews, Food, Health

<sup>2</sup> <https://huggingface.co/pyannote/speaker-diarization-3.1>

<sup>3</sup> <https://huggingface.co/alibaba-pai/pai-bert-base-zh-llm-risk-detection>

<sup>4</sup> <https://github.com/facebookresearch/flores/blob/main/toxicity>

Table 1: *Statistic of ToxicTone. TL refers to the total length of the dataset, in format hh:mm:ss.*

Split	# Utt.	# Toxic	# Non-Toxic	TL
Train	41,649	13,401	28,248	74:29:52
Dev	5,206	1,654	3,552	9:16:49
Test	5,207	1,672	3,535	9:22:09
<b>Total</b>	<b>52,062</b>	<b>16,727</b>	<b>35,335</b>	<b>93:08:50</b>

Table 2: *Comparison with MuTox (English+Spanish), DeToxy-B, and ADIMA. TL refers to the total dataset length.*

Dataset	# Utt.	# Toxic	# Non-Toxic	TL
DeToxy-B	20,217	5,307	14,910	24:39:59
MuTox	40,000	7,143	31,919	43:12:00
ADIMA	11,775	5,108	6,667	65:00:00
<b>ToxicTone</b>	<b>52,062</b>	<b>16,727</b>	<b>35,335</b>	<b>93:08:50</b>

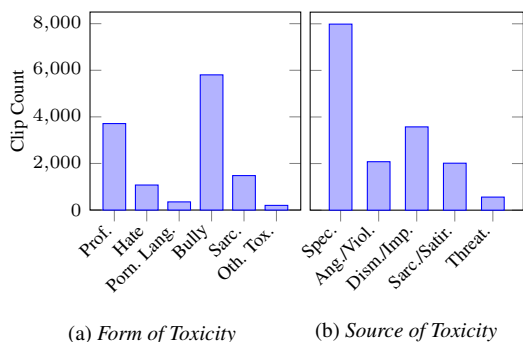


Figure 1: *Comparison of clip counts by Form and Source of Toxicity.*

& Fitness, Parenting & Family, Beauty & Fashion, and Business—are designed in line with the categorization systems used by Apple Podcasts<sup>5</sup> and Spotify Podcasts<sup>6</sup>. To assign each audio sample to one of these categories, we use GPT-4o mini [21] to classify the samples in batches of 20, based on transcripts from the first 30 minutes of the spoken content before segmenting to 2-10 seconds.

Figure 2 shows the number of clips in each category, highlighting the variety in our dataset. The largest groups are Society & News and Entertainment, with 14,247 and 13,684 clips respectively, and they also contain a high number of toxic clips. Gaming also has strong representation with 8,326 total clips, 4,133 of which are toxic. On the other hand, smaller categories such as Beauty & Fashion (589 clips with 135 toxic clips) and Movie & Book Reviews (249 clips with 56 toxic clips) are well represented too. This varied distribution, covering both high-volume mainstream topics and more niche areas, highlights the diversity of our dataset and its value for studying toxic speech.

## 4. Experiments

### 4.1. Experiment Type

We evaluate two classification tasks. The first, toxicity detection, determines whether a given speech segment contains toxic content. This task is a binary classification problem, where the model outputs a score between 0 and 1, with a threshold-based

<sup>5</sup> <https://podcasters.apple.com/support/1691-apple-podcasts-categories>

<sup>6</sup> <https://open.spotify.com/genre/0JQ5DArNBzkmxXHCqFLx2U>

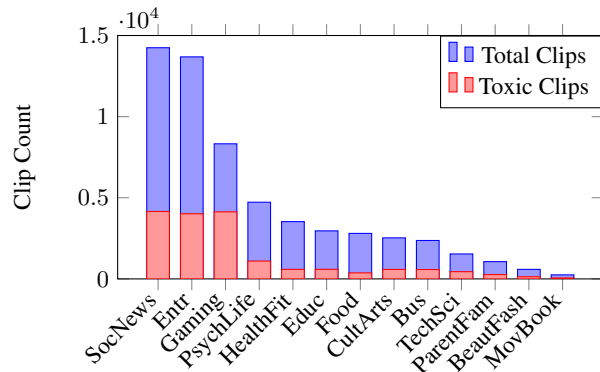


Figure 2: *Total and Toxic Clip Counts by Category.*

decision to classify an utterance as toxic or non-toxic.

The second task, toxicity source classification, identifies the origin of toxicity in speech. We classify toxic utterances into source categories, including aggressive wording, sarcastic tone, or threatening intent. Since toxicity in speech is often implicit, this task helps distinguish between overtly toxic expressions and subtler toxic cues embedded in speech patterns.

### 4.2. Experiment Setup

To evaluate toxicity detection in spoken Mandarin, we compare our approach against three baseline systems: MuTox, ETOX<sup>7</sup> [22], and COLDETECTOR<sup>8</sup> [3]. MuTox is a multilingual speech-based toxicity classifier that employs a pre-trained speech encoder (SONAR) [23] with a three-layer feedforward network for binary classification. In contrast, ETOX is a lexicon-based toxicity detection system operating on text. It relies on predefined wordlists covering multiple languages to flag toxic content, making it highly interpretable. However, since it operates on explicit lexical cues, it struggles with context-dependent or implicit toxicity, which is common in spoken communication. Additionally, COLDETECTOR is built upon bert-base-chinese and fine-tuned on a large-scale Chinese offensive language dataset, enabling it to capture both explicit and subtle offensive cues.

We investigate three pre-trained models to encode audio into features. SONAR (S) [23] is a multilingual sentence embedding model supporting both text and audio inputs; we experiment with both its text encoder (applied to ASR-transcribed speech,  $S_T$ )<sup>9</sup> and speech encoder (applied to raw audio,  $S_A$ )<sup>10</sup>. Additionally, we employ XLS-R 1B (X)<sup>11</sup> [24], a multilingual self-supervised speech model that captures rich prosodic and acoustic features; for XLS-R, we compute a layerwise weighted sum of its features to obtain a robust embedding [25]. To incorporate emotional cues, we use Emotion2Vec+ Large (E)<sup>12</sup> [26], a speech emotion representation model pre-trained on large-scale emotion datasets, and we extract its last-layer embedding. These models provide complementary information, capturing semantic, acoustic, and emotional aspects of toxicity in spoken Mandarin. In addition, we explore ensemble models by concatenating the individual features along the feature dimension, allowing us to leverage the strengths of each model.

<sup>7</sup> [https://github.com/facebookresearch/seamless\\_communication/tree/main](https://github.com/facebookresearch/seamless_communication/tree/main)

<sup>8</sup> <https://huggingface.co/thu-coai/roberta-base-cold>

<sup>9</sup> [https://dl.fbaipublicfiles.com/SONAR/sonar\\_text\\_encoder.pt](https://dl.fbaipublicfiles.com/SONAR/sonar_text_encoder.pt)

<sup>10</sup> <https://dl.fbaipublicfiles.com/SONAR/spenc.v5ap.cmn.pt>

<sup>11</sup> <https://dl.fbaipublicfiles.com/fairseq/wav2vec/xlsr2.960m.1000k.pt>

<sup>12</sup> <https://huggingface.co/emotion2vec/emotion2vec-plus-large>

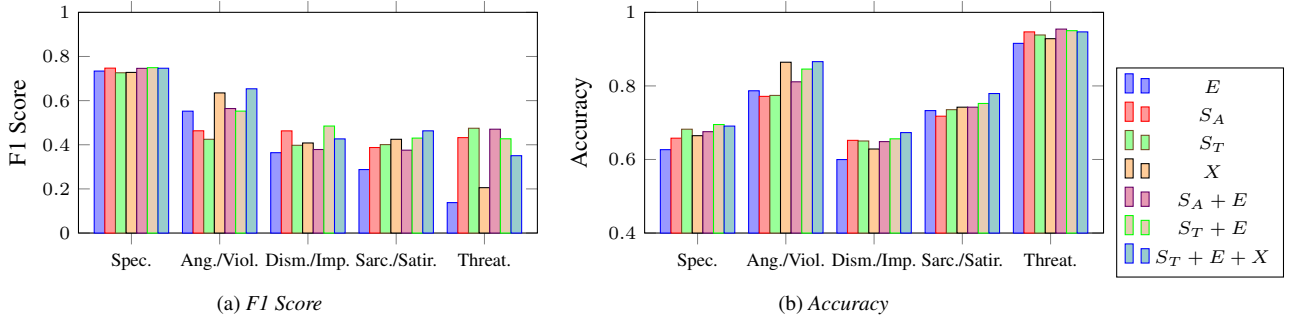


Figure 3: Performance of toxicity source classification models across different toxic sources. Plot (a) shows F1 scores and plot (b) shows Accuracy. The right panel shows the common legend of different embeddings used.

For downstream classification, all tasks utilize a three-layer linear classifier trained with binary cross-entropy loss. The toxicity detection model predicts a binary toxicity score. For toxicity source classification, we train separate classifiers per category using a one-vs-all approach, treating each class independently. This setup allows us to analyze the different manifestations of toxicity in a structured and interpretable manner.

### 4.3. Toxicity detection

Table 3: Toxicity detection performance with existing models and different embedding configurations. All results are in %. In each metric, the best performance is marked in **bold** and the second best is underlined.

Model	F1	Acc.	Prec.	Rec.
Mutox	29.41	67.19	45.79	21.66
Etox	50.54	72.40	58.85	44.28
COLD <sub>DETECTOR</sub>	42.47	65.32	45.01	40.20
$E$	47.96	72.18	57.49	41.14
$X$	60.89	74.98	59.35	62.52
$S_A$	54.04	73.36	58.43	50.26
$S_T$	57.49	72.76	55.95	59.11
$S_A + E$	56.68	74.72	60.81	53.08
$S_T + E$	<u>62.47</u>	73.76	56.35	<b>70.09</b>
$X + S_A + E$	61.93	<u>76.52</u>	<u>62.58</u>	61.30
$X + S_T + E$	<b>64.16</b>	<b>77.90</b>	<b>64.85</b>	<u>63.48</u>

Table 3 summarizes the classification performance across different models and feature combinations. Our model trained on our dataset outperforms all baselines. Baseline performance shows that Mutox scores low, while ETOX attains higher performance, with COLD<sub>DETECTOR</sub> slightly lagging behind. Individually, the emotion-based embedding ( $E$ ) scores an F1 of 47.96%, suggesting that emotional signals alone are insufficient, whereas XLS-R ( $X$ ) reaches an F1 of 60.89% with balanced metrics, which is the best among all single encoder settings. The SONAR encoders are competitive, with  $S_T$  achieving an F1 of 57.49% and a recall of 59.11%. Notably, combining embeddings improves performance:  $S_T + E$  yields an F1 of 62.47% and a recall of 70.09%, and the multimodal  $X + S_T + E$  configuration attains the best results, underscoring the value of fusing acoustic, linguistic, and emotional features for effective toxic speech detection.

### 4.4. Toxicity source detection

The performance of toxicity source detection is depicted in Figure 3. For the “Specific Words” category, all models ex-

hibit consistently high F1 scores (approximately 0.73–0.75) and comparable accuracy, indicating that explicit lexical cues are reliably detected even with single model input. For categories like “Angry/Violent” and “Sarcastic/Satirical,” the ensemble configuration ( $S_T + E + X$ ) excels with the highest F1 scores and improved accuracy. In contrast, threat detection, despite uniformly high accuracy, yields a lower F1 score with the best performance achieved by the semantic encoder ( $S_T$ ). This reduced performance can be attributed to the imbalanced distribution of threat-labeled samples, which are considerably fewer.

## 5. Limitation and future work

While our dataset and models establish a strong foundation for Chinese spoken toxic speech detection, several areas offer opportunities for further refinement. The dataset reflects real-world toxicity distributions, including natural class imbalances, such as the higher prevalence of toxic speech in gaming content. While this alignment with authentic usage patterns enhances model relevance, future work can expand coverage of underrepresented categories to improve generalization.

Like most toxic speech detection datasets, our experiments reflect majority opinion, which may not fully capture individual perceptions influenced by personal experience, culture, and context [27, 28, 29, 30]. To enhance transparency and enable further research on annotation biases, we will release the annotator-level annotations. Future work can leverage these detailed annotations to explore alternative aggregation strategies and develop adaptive models that better account for cultural and individual differences in toxicity perception.

## 6. Conclusion

This work introduces the first large-scale Mandarin Chinese toxic speech dataset, addressing a critical gap in spoken toxic speech detection. Unlike prior text-based datasets, our dataset incorporates prosodic cues and detailed toxicity labels, enabling a more nuanced understanding of harmful speech. Our experiments demonstrate that multimodal approaches—leveraging acoustic, emotional, and linguistic features—significantly outperform text-only models, underscoring the importance of speech-specific cues in detecting toxicity. Additionally, our analysis reveals domain-specific trends in toxicity, with gaming content exhibiting the highest prevalence. These findings highlight the necessity of dedicated speech-based detection models to capture the complexities of spoken toxicity. By establishing a strong benchmark for Chinese spoken toxic speech detection, our work lays the foundation for future advancements in multimodal toxicity detection, dataset expansion across Chinese dialects, and personalized toxicity perception models to further enhance content moderation in online speech interactions.

## 7. References

- [1] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth." *Psychological bulletin*, 2014.
- [2] A.-M. Bliuc, N. Faulkner, A. Jakubowicz, and C. McGarty, "Online networks of racial hate: A systematic review of 10 years of research on cyber-racism," *Computers in Human Behavior*, 2018.
- [3] J. Deng, J. Zhou, H. Sun, C. Zheng, F. Mi, H. Meng, and M. Huang, "COLD: A benchmark for Chinese offensive language detection," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 11 580–11 599.
- [4] M.-c. Chiu, "Code-switching and identity constructions in taiwan tv commercials," *Monumenta Taiwanica*, 2012.
- [5] A. de Wynter, I. Watts, T. Wongsangaroonsri, M. Zhang, N. Farra, N. E. Altuntoprak, L. Baur, S. Claudet, P. Gajdusek, C. Gören, Q. Gu, A. Kaminska, T. Kaminski, R. Kuo, A. Kyuba, J. Lee, K. Mathur, P. Merok, I. Milovanović, N. Paananen, V.-M. Paananen, A. Pavlenko, B. P. Vidal, L. Strika, Y. Tsao, D. Turcato, O. Vakhno, J. Velcsov, A. Vickers, S. Visser, H. Widarmanto, A. Zaikin, and S.-Q. Chen, "RTP-LX: Can LLMs evaluate toxicity in multilingual scenarios?" vol. AAAI AISI, 2025.
- [6] J. Lu, B. Xu, X. Zhang, C. Min, L. Yang, and H. Lin, "Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 16 235–16 250.
- [7] X. Tang and X. Shen, "Categorizing offensive language in social networks: A Chinese corpus, systems and an explainable tool," in *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, 2020, pp. 1045–1056.
- [8] H. Yang and C.-J. Lin, "TOCP: A dataset for Chinese profanity processing," in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 2020, pp. 6–12.
- [9] S. Ghosh, S. Lepcha, S. Sakshi, R. R. Shah, and S. Umesh, "Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances," in *Interspeech 2022*, 2022, pp. 5185–5189.
- [10] M. Costa-jussà, M. Meglioli, P. Andrews, D. Dale, P. Hansanti, E. Kalbassi, A. Mourachko, C. Ropers, and C. Wood, "MuTox: Universal Multilingual audio-based TOXicity dataset and zero-shot detector," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 5725–5734.
- [11] V. Gupta, R. Sharon, R. Sawhney, and D. Mukherjee, "Adima: Abuse detection in multilingual audio," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- [12] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.
- [13] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020.
- [14] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, speech, and signal processing, ieee international conference on*, 1992.
- [15] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenhaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haasheim *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.
- [16] C.-K. Yang, Y.-K. Fu, C.-A. Li, Y.-C. Lin, Y.-X. Lin, W.-C. Chen, H. L. Chung, C.-Y. Kuan, W.-P. Huang, K.-H. Lu, T.-Q. Lin, H.-H. Wang, E.-P. Hu, C.-J. Hsu, L.-H. Tseng, I.-H. Chiu, U. Sanga, X. Chen, P. chun Hsu, S. wen Yang, and H. yi Lee, "Building a taiwanese mandarin spoken language model: A first attempt," 2024.
- [17] L.-H. Tseng, Z.-C. Chen, W.-S. Chang, C.-K. Lee, T.-R. Huang, and H.-y. Lee, "Leave no knowledge behind during knowledge distillation: Towards practical and effective knowledge distillation for code-switching asr using realistic data," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024.
- [18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, Volume 1 (Long and Short Papers)*, 2019.
- [19] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," in *Proceedings of the international aaii conference on web and social media*, 2015.
- [20] A. Lenhart, M. Ybarra, K. Zickuhr, and M. Price-Feeney, "Online harassment, digital abuses, and cyberstalking in america," 2016.
- [21] OpenAI *et al.*, "Gpt-4o system card," 2024. [Online]. Available: <https://arxiv.org/abs/2410.21276>
- [22] M. Costa-jussà, E. Smith, C. Ropers, D. Licht, J. Maillard, J. Ferrando, and C. Escolano, "Toxicity in multilingual machine translation at scale," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.
- [23] P.-A. Duquenne, H. Schwenk, and B. Sagot, "Sonar: sentence-level multimodal and language-agnostic representations," *arXiv e-prints*, pp. arXiv–2308, 2023.
- [24] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022.
- [25] S. wen Yang *et al.*, "Superb: Speech processing universal performance benchmark," in *Interspeech 2021*, 2021.
- [26] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," *Proc. ACL 2024 Findings*, 2024.
- [27] Y.-C. Lin, H. Wu, H.-C. Chou, C.-C. Lee, and H. yi Lee, "Emobias: A large scale evaluation of social bias on speech emotion recognition," in *Interspeech 2024*, 2024.
- [28] Y.-C. Lin, T.-Q. Lin, H.-C. Lin, A. T. Liu, and H. yi Lee, "On the social bias of speech self-supervised models," in *Interspeech 2024*, 2024.
- [29] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, "Handling bias in toxic speech detection: A survey," *ACM Computing Surveys*, 2023.
- [30] Y.-C. Lin, W.-C. Chen, and H.-Y. Lee, "Spoken stereoset: on evaluating social bias toward speaker in speech large language models," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024.