



ViTOSA: Audio-Based Toxic Spans Detection on Vietnamese Speech Utterances

Huy Ba Do^{1,3}, Vy Le-Phuong Huynh^{2,3}, Luan Thanh Nguyen^{2,3}

¹Faculty of Computer Science, University of Information Technology, Ho Chi Minh City, Vietnam

²Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam

³Vietnam National University, Ho Chi Minh City, Vietnam

{21522137,20520951}@gm.uit.edu.vn, luannt@uit.edu.vn

Abstract

Toxic speech on online platforms is a growing concern, impacting user experience and online safety. While text-based toxicity detection is well-studied, audio-based approaches remain underexplored, especially for low-resource languages like Vietnamese. This paper introduces ViTOSA (Vietnamese Toxic Spans Audio), the first dataset for toxic spans detection in Vietnamese speech, comprising 11,000 audio samples (25 hours) with accurate human-annotated transcripts. We propose a pipeline that combines ASR and toxic spans detection for fine-grained identification of toxic content. Our experiments show that fine-tuning ASR models on ViTOSA significantly reduces WER when transcribing toxic speech, while the text-based toxic spans detection (TSD) models outperform existing baselines. These findings establish a novel benchmark for Vietnamese audio-based toxic spans detection, paving the way for future research in speech content moderation¹.

Disclaimer: This paper includes real examples from social media platforms that may be perceived as toxic or offensive.

Index Terms: audio-based toxic spans detection, automatic speech recognition, spans detection

1. Introduction and Related Work

In the context of robust digital content development, online platforms have become increasingly popular for community interaction and information sharing; however, the rise of toxic audio utterances has become a significant concern [1, 2, 3]. Furthermore, the widespread dissemination of sensitive and toxic phrases and audio clips is having a negative impact on users' mental well-being as well as on individual honor and dignity [4, 5, 6]. Such content can incite violence, promote hatred, and inflict deep psychological harm on listeners, especially children and teenagers who are particularly vulnerable. The uncontrolled spread of toxic speech in online audio environments not only degrades communication quality but also creates a negative atmosphere, making many users feel concerned and even withdraw from discussions, as mentioned in the work of Qayyum et al. [7]. This growing issue undermines public trust in digital platforms and threatens a safe, healthy communication environment.

Research on detecting toxic speech in audio has gained attention, but it remains relatively underdeveloped compared to text-based approaches. Efforts such as the DeToxy dataset [8] and MuTox [9] have introduced toxic speech datasets along

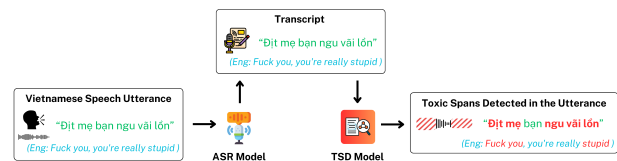


Figure 1: Framework of ViTOSA.

with classification models, including end-to-end and multilingual approaches. Besides, Nada et al. [10]'s studies have explored efficient models for real-time toxicity detection, while Liu et al. [11] have investigated the integration of speech and text modalities to improve detection accuracy. However, existing studies often focus on utterance-level classification and lack the ability to detect toxic segments within speech.

This issue affects many languages but is particularly severe in under-resourced contexts like Vietnamese, which lacks sufficient datasets and where research on toxic content detection has largely focused on text. Notable contributions include UIT-ViCTSD [12] and ViHSD [13], which aim to identify toxicity and hate speech in user comments; the ViHOS dataset [14], designed for detecting toxic textual phrases; and the ViHateT5 model [15], which utilizes a text-to-text transformer for various hate-speech-related tasks in Vietnamese. While these studies have advanced text-based toxicity detection, there is currently no dedicated dataset or research specifically addressing toxic speech in Vietnamese audio. This gap highlights the urgent need for developing resources and methodologies to detect and mitigate toxic speech in Vietnamese, ensuring a more comprehensive approach to online safety.

In this research, we aim to address the limitations of existing studies and meet practical needs by developing a comprehensive approach for detecting toxic speech in Vietnamese audio. To address these challenges, our contributions are as follows: (1) We introduce the novel ViTOSA dataset, the first specifically focused on Vietnamese toxic audio segments, comprising 25 hours of speech, along with a dedicated evaluation test set serving as a benchmark for both ASR and TSD tasks; (2) We propose an effective pipeline that integrates a domain-specific speech recognition model with a text-based language model to accurately detect toxic audio segments; (3) We highlight the impact of our approach, demonstrating its potential to advance toxic speech detection in low-resource languages and lay the groundwork for future research in this field.

¹<https://github.com/vitosa-research/ViToSA-Dataset>

2. ViTOSA Dataset

We begin by conducting preliminary experiments to evaluate the performance of existing ASR and TSD models in Vietnamese. These experiments are designed to assess the effectiveness of current models in transcribing toxic speech, identify specific challenges faced by ASR systems when handling toxic content, and evaluate the accuracy of TSD models in identifying toxic spans. Both tasks are evaluated using our ViTOSA test set, which was constructed prior to the training set to serve as a reliable benchmark. The test set includes triplets of audio, manually transcribed text, and annotated toxic spans, providing a comprehensive resource for analyzing model performance.

2.1. Preliminary Experiments

ASR Task. Recent state-of-the-art ASR models for Vietnamese, such as Whisper [16], Wav2Vec2 (W2V2)² [17], and PhoWhisper [18], demonstrate strong performance on standard benchmarks and widely used speech corpora. However, these models are typically trained on datasets with limited coverage of toxic vocabulary, leading to subpar performance in recognizing toxic speech, as shown in Table 1.

Table 1: *Outputs of four models on a Vietnamese toxic utterance.*

Ground Truth	thì đéo nói địt mẹ mày nó đến bốn năm lần thì đỡ thể lớn nào được (Eng: hadn't fucking said anything, but your fucking mother came four or five times, how the fuck could it be better?)
Whisper	đé theo nói bị mày á mày nó đến bốn năm lần đi đó thể là một nào được
wav2vec2-base-vi-vlsp2020	thì đơ nói địt mày mày nó đến bưng năm lần thì nói thể lớn nào được
wav2vec2-base-vietnamese-250h	vì đi nói đi mày á nói đến bốn năm lần thì nói thể lớn nào được
PhoWhisper	vì đéo nói mấy ảnh mày nó đến bốn năm lần vì đỡ thể lớn vào đưông.

Due to the frequent misrecognition of toxic words in ASR outputs, we propose constructing a Vietnamese ASR dataset specifically for the toxic speech domain to improve model robustness in handling such content.

TSD Task. We utilize the ViHOS dataset [14], which has been annotated to identify toxic text spans in Vietnamese. This dataset is used to fine-tune language models, optimizing their performance for the toxic spans detection task. After training, we evaluate these models on our test set to assess their effectiveness in detecting toxic content.

2.2. Data Creation

In this paper, we first release ViTOSA, a high-quality dataset for Vietnamese speech processing, with a focus on toxic speech research. We strictly follow the process shown in Figure 2 to collect and annotate audio data.

Data Collection. The Data Collection Process begins with the Video Collection phase. Short video clips containing toxic content are manually gathered from social media platforms such as Facebook, YouTube, and TikTok.

Audio Extraction. Once videos are collected, we extract audio files using the open-source library SoundFile³ to convert videos into audio files. We hired undergraduate students from various academic backgrounds as annotators⁴, training them to

²Vietnamese variants of the W2V2 architecture include wav2vec2-base-vi-vlsp2020 and wav2vec2-base-vietnamese-250h, available on HuggingFace.

³<https://python-soundfile.readthedocs.io/en/0.13.1/>

⁴Paid according to the local minimum wage.

identify toxic content and use the Audio Cutter⁵ tool for annotation. The selected audio segments range from 1 to 14 seconds in length. We discard segments shorter than 1 second due to a lack of meaningful context for toxicity identification, while those longer than 14 seconds are split to prevent cognitive overload for annotators.

Human-annotated Transcription Phase involves a challenge designed to help annotators understand the guidelines and improve transcript quality. We randomly select 50 audio files from the dataset for assessment. Annotators in group A are paired into sets $A_i = \{A_i \mid i \in N\}$, with each pair consisting of two members. Each annotator independently listens and transcribes the audio. The transcripts from each pair are then compared to calculate WER. The challenge consists of three rounds:

- If a pair's WER is less than 8%, they are considered to have met the standard and understood the guideline.
- If WER is greater than 8%, we analyze the errors, update the guideline for clarification, and proceed to the next round.

After three rounds, all pairs achieved WER below 8%, ensuring high-quality data. Once all annotators in group A met the standard, they proceeded to transcribe the remaining samples. To prevent fraud, members within the same pair were unaware of each other's identities, avoiding the risk of one member transcribing while the other merely copied.

Quality Control Phase is conducted by annotators in group B. 20% of the samples from each annotator in group A are randomly selected and assigned to group B as ground truth. We continue using the WER threshold of 8% to evaluate transcripts. If the WER between group A and group B is below 8%, the transcript from that group A annotator is deemed valid. Otherwise, annotators exceeding the threshold must re-transcribe the entire set.

The final dataset contains 24.75 hours of Vietnamese-speaking utterances across 11,802 audio-transcript pairs⁶, split into training, validation, and test sets, with 1,000 samples for testing and the rest divided 8:2 for training and validation.

3. Methodology

Having established the dataset, we now introduce our proposed detection framework, ViTOSA, for detecting toxic speech segments in Vietnamese utterances. As shown in Figure 1, it consists of two key components: Automatic Speech Recognition (ASR), which transcribes spoken utterances, and Toxic Spans Detection (TSD), which identifies toxic segments in the transcriptions.

3.1. Automatic Speech Recognition

We utilize state-of-the-art transformer-based pre-trained models specifically optimized for Vietnamese automatic speech recognition. Trained on large-scale multilingual and monolingual corpora, these models effectively transcribe spoken language into text while demonstrating robustness to variations in accent, background noise, and speech patterns.

⁵<https://mp3cut.net/>

⁶https://huggingface.co/datasets/ViTOSAResearch/ViTOSA_Dataset

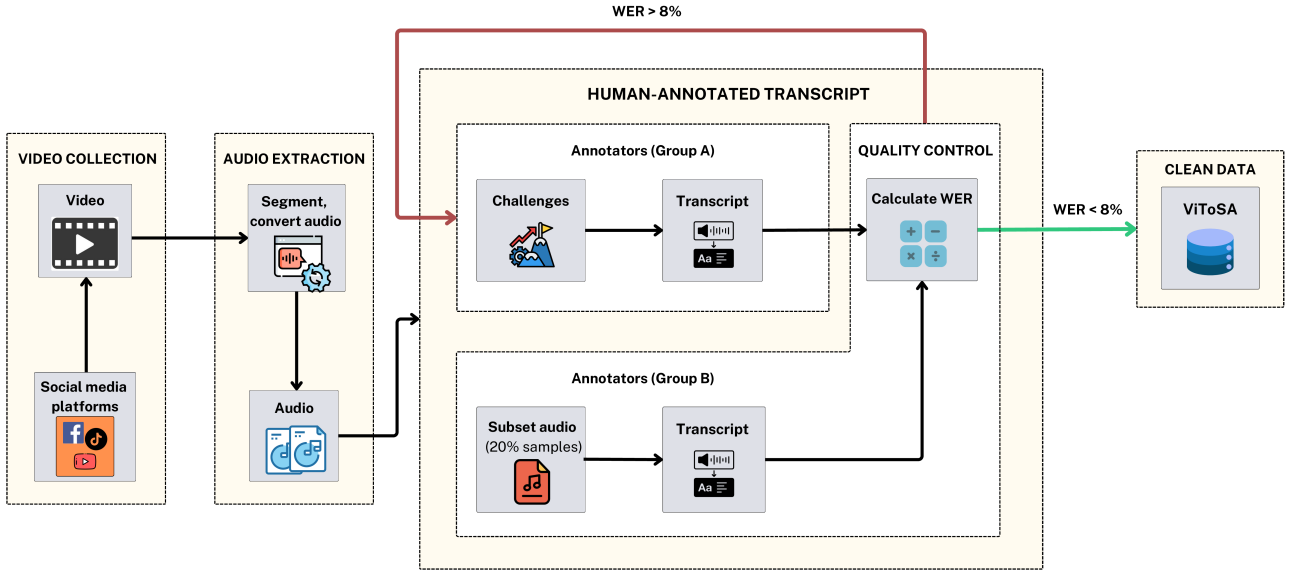


Figure 2: Pipeline for Collecting, Processing, and Quality Checking Transcribed Audio for the ViToSA dataset (train and validation).

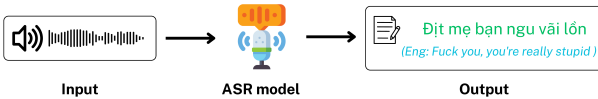


Figure 3: Input and output of the ASR component.

To further improve their performance in recognizing toxic speech, we fine-tune these models on our domain-specific ASR dataset, ViToSA, enabling more accurate transcription of Vietnamese utterances containing toxic content.

3.2. Toxic Spans Detection

After obtaining transcriptions from the ASR component, we apply BERT-based language models, either Vietnamese-specific or multilingual, to detect and precisely localize toxic words or phrases within the text.

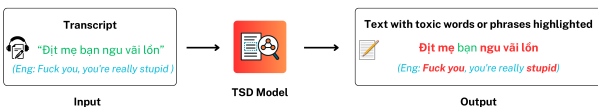


Figure 4: Illustration of the input and output of the TSD component. Predicted text-based spans are highlighted in bold red.

By leveraging deep contextual embeddings, these models effectively capture semantic subtleties and syntactic patterns, enabling accurate identification of both explicit and context-dependent toxic language.

4. Experiments

We perform experiments on the ViToSA dataset, focusing on two key tasks: ASR and TSD. The process is outlined in the following sections: data pre-processing, evaluation metrics, and speech recognition experimental results.

4.1. Data

We use our ViToSA dataset to perform ASR. All audio files are resampled to 16kHz and converted to mono channel to ensure consistency. For the ASR task, text pre-processing includes lowercasing, removing punctuation, and converting numbers into words, enhancing model readability and accuracy. In the TSD task, unnecessary whitespace is eliminated, line breaks are standardized, and toxic word positions are formatted into a structured labeling scheme, optimizing the data for precise identification.

For TSD, we use ViHOS training data for fine-tuning (as explained in Section 2.1). Those fine-tuned models are then evaluated on our ViToSA test set.

4.2. Models and Settings

We present the models and experimental settings used in our research on two tasks. Note that we use a single NVIDIA A100 GPU for all experiments in this study.

ASR Models. For the ASR task, we utilize several models, including wav2vec2 [17] variants fine-tuned for Vietnamese, namely wav2vec2-base-vi-vlsp2020⁷ and wav2vec2-base-vietnamese250h⁸. Additionally, we employ the multilingual Whisper (base) model [16] and its Vietnamese fine-tuned version, PhoWhisper (base) [18]. All ASR models are trained for 10 epochs with a batch size of 8, using the AdamW optimizer with a learning rate of 5e-5, a warmup ratio of 0.1, and a weight decay of 0.05.

TSD Models. We fine-tune current high-performance models in Vietnamese for TSD tasks, including multilingual pre-trained models such as XLM-R (base) [19], BERT (base, multilingual, cased) [20], DistilBERT (base, multilingual, cased) [21], and monolingual ones such as PhoBERT (base, v2) [22], ViSoBERT [23], CafeBERT [24], and ViHateT5 [15]. Note that ViHateT5 is already fine-tuned on the TSD task, we only use the

⁷<https://huggingface.co/nguyenvulebinh/wav2vec2-base-vi-vlsp2020>

⁸<https://huggingface.co/nguyenvulebinh/wav2vec2-base-vietnamese-250h>

Table 2: *Speech recognition experimental results on ViTOSA test set.*

Models	Toxic	Non-toxic	All
<i>w/o ViTOSA dataset</i>			
Whisper	1.660	0.593	1.149
wav2vec2-base-vi-vlsp2020	0.988	0.984	0.986
wav2vec2-base-vietnamese-250h	0.997	0.999	0.998
PhoWhisper	0.615	0.212	0.418
<i>with ViTOSA dataset</i>			
Whisper	0.325 ↓ 1.335	0.264 ↓ 0.329	0.289 ↓ 0.860
wav2vec2-base-vi-vlsp2020	0.319 ↓ 0.669	0.302 ↓ 0.682	0.310 ↓ 0.676
wav2vec2-base-vietnamese-250h	0.342 ↓ 0.655	0.280 ↓ 0.719	0.311 ↓ 0.687
PhoWhisper	0.302 ↓ 0.313	0.192 ↓ 0.020	0.257 ↓ 0.161

Table 3: *Toxic spans detection experimental results on ViTOSA test set.*

Models	Acc	WF1	MF1
ViHateT5	0.765	0.785	0.500
DistilBERT	0.937	0.934	0.732
BERT	0.940	0.940	0.768
XLM-R	0.940	0.943	0.790
CafeBERT	0.927	0.932	0.807
ViSoBERT	0.945	0.947	0.817
PhoBERT	0.951	0.955	0.837

model in the original paper without further fine-tuning. They are then fine-tuned for 4 epochs with a batch size of 8, using the AdamW optimizer. The training process is conducted with a learning rate of $2e-5$ and a warmup ratio of 0.1 to optimize model performance.

4.3. Evaluation Metrics

Word Error Rate (WER) is a widely used metric for assessing the accuracy of speech recognition models. Moreover, following the methodology mentioned in the work of Hoang et al. [14], we evaluate the toxic spans detection task using Accuracy (Acc), Macro F1 (MF1), and Weighted F1 (WF1) scores.

4.4. Results and Discussions

We fine-tune ASR models on our annotated training data ViTOSA and evaluate them on the test set, depicted in Table 2. For TSD, we fine-tune language models on ViHOS and assess their performance on our ViTOSA test set, listed in Table 3.

The need for a domain-specific toxic audio dataset. Table 2 underscores the importance of using a dedicated toxic audio dataset, such as ViTOSA, for training ASR models. Without ViTOSA, all models exhibit significantly higher word error rates (WER), particularly for toxic speech, where errors are more pronounced. After fine-tuning with ViTOSA, WER drops considerably across all models, with Whisper experiencing the most significant improvement (from 1.149 to 0.289 overall). The wav2vec2-based models also benefit from ViTOSA, showing WER reductions of approximately 0.676 and 0.687. Even PhoWhisper, which initially had a lower WER, further improves. These results confirm that general ASR models struggle

Table 4: *Outputs of four models on a Vietnamese toxic utterance after fine-tuning on ViTOSA dataset.*

Ground Truth	thì đéo nói địt mẹ mày nó đến bốn năm lần thì đờ thể lôn nào được (Eng: hadn't fucking said anything, but your fucking mother came four or five times, how the fuck could it be better?)
Whisper	thì đéo nói địt mẹ mày nói đến bốn năm lần thì đéo thể lôn nào được
wav2vec2-base-vi-vlsp2020	địt đéo nói địt mẹ mày nói đến bốn năm lần thì đờ thể lôn nào được
wav2vec2-base-vietnamese-250h	thì đéo nói địt mẹ mày nó đến bốn năm lần thì đờ thể lôn nao đường
PhoWhisper	thì đéo nói địt mẹ mày nó đến bốn năm lần thì đờ tể lôn nào được

with toxic speech due to data scarcity, and incorporating a domain-specific dataset significantly enhances performance, making ASR more reliable in toxic speech recognition tasks.

TSD on normalized text (from ASR models) achieves higher performance than direct evaluation on social-media texts of ViHOS. The results in Table 4 indicate that performing TSD on normalized text, generated by ASR models, yields a higher MF1 compared to direct evaluation on the ViHOS dataset⁹. Among all models, PhoBERT achieves the highest performance with 0.837 MF1, demonstrating its effectiveness in detecting toxic spans. ViSoBERT follows closely, particularly excelling in MF1 (0.817), indicating better generalization to minority toxic spans. Other transformer-based models, such as XLM-R and BERT, also perform well, with MF1 scores above 0.75. However, ViHateT5, which is mainly pre-trained and fine-tuned on social media texts, lags behind, particularly in MF1 (0.500), suggesting difficulties in handling the normalized toxic language. The superior performance on normalized text suggests that ASR-generated outputs, after normalization, may simplify TSD for general transformer-based models. However, this normalization process appears to challenge domain-specific pre-trained models like ViHateT5, which have been trained exclusively on social media data.

Result Analysis. To further assess model performance, we conduct inference again on a representative toxic utterance from Table 2.1 using the trained models. The predictions show that after training on ViTOSA, nearly all toxic words in the utterance are accurately detected, a significant improvement compared to the initial results before fine-tuning on ViTOSA. This demonstrates the necessity of constructing a dedicated dataset for toxic word recognition in Vietnamese utterances. Although the WER of the models remains relatively high, leading to some inaccuracies in full-sentence ASR outputs, our primary focus in this task is toxic word detection, for which the results are well-aligned with our objectives.

5. Conclusion

This paper introduces ViTOSA, the first benchmark for detecting toxic spans in Vietnamese speech, addressing the gap in audio-based toxicity detection for low-resource languages. Our findings highlight the limitations of current ASR models in accurately transcribing toxic speech and demonstrate that fine-tuning ASR on ViTOSA significantly reduces WER for toxic content. Furthermore, our Transformer-based toxic span detection models achieve strong results, showing the effectiveness of ASR-transcribed text for toxicity detection. We hope ViTOSA fosters further research in speech-based toxicity detection and supports the development of a safer online environment.

⁹According to the best performance in the original paper [14] that obtained 0.772 MF1 with PhoBERT (large).

6. Acknowledgement

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

7. References

- [1] A. S. Namin, R. Hewett, K. S. Jones, and R. L. Pogrund, "The sounds of cyber threats," *ArXiv*, vol. abs/1805.08272, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:46894197>
- [2] M. Yousefi and D. Emmanouilidou, "Audio-based toxic language classification using self-attentive convolutional neural network," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 11–15.
- [3] H. Kwak and J. Blackburn, "Linguistic analysis of toxic behavior in an online video game," in *Social Informatics*, L. M. Aiello and D. McFarland, Eds. Cham: Springer International Publishing, 2015, pp. 209–217.
- [4] K. Browne and C. Hamilton-Giachritsis, "The influence of violent media on children and adolescents: A public-health approach," *Lancet*, vol. 365, pp. 702–10, 02 2005.
- [5] A.-M. Bucur, M. Zampieri, and L. P. Dinu, "An exploratory analysis of the relation between offensive language and mental health," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 3600–3606. [Online]. Available: <https://aclanthology.org/2021.findings-acl.315/>
- [6] N. D. Volkow, J. A. Gordon, and G. F. Koob, "Choosing appropriate language to reduce the stigma around mental illness and substance use disorders," *Neuropsychopharmacology*, vol. 46, no. 13, pp. 2230–2232, Dec 2021. [Online]. Available: <https://doi.org/10.1038/s41386-021-01069-4>
- [7] H. Qayyum, M. Ikram, B. Z. H. Zhao, I. D. Wood, N. Kourtellis, and M. A. Kaafar, "Exploring the distinctive tweeting patterns of toxic twitter users," in *2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 3624–3633.
- [8] 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.
- [9] 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*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 5725–5734. [Online]. Available: <https://aclanthology.org/2024.findings-acl.340/>
- [10] A. H. A. Nada, S. Latif, and J. Qadir, "Lightweight toxicity detection in spoken language: A transformer-based approach for edge devices," *arXiv preprint arXiv:2304.11408*, 2023.
- [11] J. Liu, M. K. Nandwana, J. Pylkkönen, H. Heikinheimo, and M. McGuire, "Enhancing multilingual voice toxicity detection with speech-text alignment," in *Interspeech 2024*, 2024, pp. 4298–4302.
- [12] L. T. Nguyen, K. Van Nguyen, and N. L.-T. Nguyen, *Constructive and Toxic Speech Detection for Open-Domain Social Media Comments in Vietnamese*. Springer International Publishing, 2021, p. 572–583. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-79457-6_49
- [13] S. T. Luu, K. V. Nguyen, and N. L.-T. Nguyen, "A large-scale dataset for hate speech detection on vietnamese social media texts," in *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*. Springer, 2021, pp. 415–426.
- [14] P. G. Hoang, C. D. Luu, K. Q. Tran, K. V. Nguyen, and N. L.-T. Nguyen, "ViHOS: Hate speech spans detection for Vietnamese," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 652–669. [Online]. Available: <https://aclanthology.org/2023.eacl-main.47/>
- [15] L. Thanh Nguyen, "ViHateT5: Enhancing hate speech detection in Vietnamese with a unified text-to-text transformer model," in *Findings of the Association for Computational Linguistics ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 5948–5961. [Online]. Available: <https://aclanthology.org/2024.findings-acl.355>
- [16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.
- [17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [18] T.-T. Le, L. T. Nguyen, and D. Q. Nguyen, "PhoWhisper: Automatic Speech Recognition for Vietnamese," in *Proceedings of the ICLR 2024 Tiny Papers track*, 2024.
- [19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schuster, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: <https://aclanthology.org/2020.acl-main.747/>
- [20] 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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423/>
- [21] V. Sanh, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [22] D. Q. Nguyen and A. Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1037–1042. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.92/>
- [23] N. Nguyen, T. Phan, D.-V. Nguyen, and K. Nguyen, "ViSoBERT: A pre-trained language model for Vietnamese social media text processing," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5191–5207. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.315/>
- [24] P. N.-T. Do, S. Q. Tran, P. G. Hoang, K. V. Nguyen, and N. L.-T. Nguyen, "VLUE: A new benchmark and multi-task knowledge transfer learning for Vietnamese natural language understanding," in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 211–222. [Online]. Available: <https://aclanthology.org/2024.findings-naacl.15/>