Multimodal Zero-Shot Framework for Deepfake Hate Speech Detection in Low-Resource Languages

Rishabh Ranjan[*1], *Likhith Ayinala*[*2], *Mayank Vatsa*[1], *Richa Singh*[1]

[1]Indian Institute of Technology Jodhpur, India
[2]Columbia University, USA

{ranjan.4@iitj.ac.in, mvatsa, richa}@iitj.ac.in, la3073@columbia.edu

Abstract

This paper introduces a novel multimodal framework for hate speech detection in deepfake audio, excelling even in zero-shot scenarios. Unlike previous approaches, our method uses contrastive learning to jointly align audio and text representations across languages. We present the first benchmark dataset with 127,290 paired text and synthesized speech samples in six languages: English and five low-resource Indian languages (Hindi, Bengali, Marathi, Tamil, Telugu). Our model learns a shared semantic embedding space, enabling robust cross-lingual and cross-modal classification. Experiments on two multilingual test sets show our approach outperforms baselines, achieving accuracies of 0.819 and 0.701, and generalizes well to unseen languages. This demonstrates the advantage of combining modalities for hate speech detection in synthetic media, especially in low-resource settings where unimodal models falter. The Dataset is available at https://www.iab-rubric.org/resources.

Index Terms: Hate Speech, audio deepfakes, audio classification

1. Introduction

Social media has transformed global communication, connecting nearly 6.3 billion users and exhibiting a remarkable compound annual growth rate-driven primarily by emerging markets in Asia, such as India, China, and Indonesia. However, this digital expansion has also amplified the reach of hate speech. Online hate speech demonstrably harms society, while current moderation techniques struggle to keep pace with the fast-evolving online landscape. This issue is further exacerbated as platforms increasingly adopt multi-modal communications across diverse languages [1].

Existing research on hate speech detection predominantly focuses on textual analysis, often overlooking the rich information contained in other modalities. Prior studies and datasets [2, 3, 4, 5] have primarily centered on conversational text, leaving audio-based modalities relatively underexplored due to technical challenges and limited datasets. In the audio domain, approaches are typically categorized into two types: cascaded systems, which extend traditional text toxicity detection by incorporating speech recognition pipelines [6], and end-to-end systems, which classify toxicity directly from audio signals [7]. While the latter approach shows promise, it has primarily been validated on English datasets, demonstrating significant advantages over text-based models in handling out-of-domain content, as shown in studies using proprietary datasets [8]. Recent advancements, such as MuTox [9], have introduced scalable
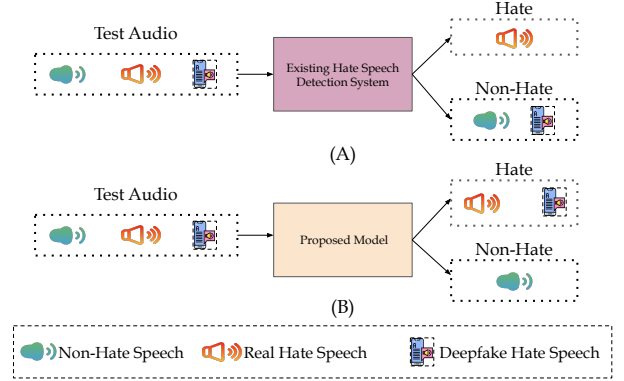

Figure 1: *(A) Current hate speech detectors often misclassify deepfake-generated hate speech audio as non-hateful, exposing moderation systems to manipulation. (B) The proposed multimodal framework effectively distinguishes deepfake hate speech from genuine non-hateful content, enhancing detection accuracy and robustness.*

multilingual audio-based toxicity detectors capable of zero-shot detection across multiple languages, representing a significant leap forward in this field.

Despite these advancements, a critical research gap remains: the detection of hate speech in deepfake audios. The rise of synthetic hate content online has raised concerns about its potential to manipulate public discourse; however, current literature and datasets have not addressed this challenge. Traditional hate speech detection datasets primarily focus on textual content [2, 10, 11], and while some efforts have expanded into audio [9, 7, 12], none have tackled the unique challenges posed by deepfake audio manipulations, such as their synthetic nature and potential for deception. Figure 1 shows the limitations of the current deepfake detection system.

To bridge this gap, we introduce a novel dataset comprising paired audio and text samples from English and five low-resource languages-Hindi, Bengali, Marathi, and Tamil, and curated explicitly for hate speech detection in deepfake audios. These languages are considered low-resource due to limited datasets and computational tools. Alongside this dataset, we propose a multi-modal architecture that leverages state-of-the-art text and audio encoders to project heterogeneous data into a unified embedding space using contrastive loss that enhances similarity learning. This joint embedding approach not only enables robust cross-lingual and cross-modal detection but also equips our system with zero-shot capabilities, allowing it to generalize to unseen languages and modalities.

Our contributions are threefold. First, we provide a multi-

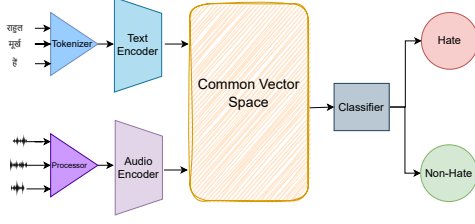[*]These authors contributed equally to this work

Figure 2: *Proposed multimodal hate speech detection pipeline: Audio and text inputs are encoded separately and mapped into a common semantic embedding space, enabling effective cross-modal and cross-lingual classification through contrastive learning.*

modal dataset addressing the gap in deepfake audio hate speech detection. Second, we introduce a novel, contrastive learning-based architecture that effectively fuses textual and acoustic features. Finally, we establish strong multi-modal baselines and demonstrate the superiority of our approach over traditional text-based detectors, particularly in zero-shot scenarios, where "zero-shot" refers to detecting hate speech in new languages. This work lays a foundation for future research in multimodal hate speech detection, paving the way for safer digital environments through effective content moderation strategies.

2. Proposed Hate Audio Detector Model

The proposed framework introduces a novel two-stage contrastive learning approach for multimodal hate speech detection, distinctly combining audio and text modalities. Unlike existing methods, we employ a unique alignment strategy in the pre-training stage, leveraging state-of-the-art encoders, SONAR for text and SeamlessM4T for audio, to project embeddings into a unified semantic space. This multimodal alignment significantly enhances zero-shot generalization capabilities, particularly in low-resource languages. Formally, our objective is to learn a classifier f_θ.

$$f_\theta : (a_i, t_i) \mapsto s_u, \quad \text{where} \quad a_i \in \mathbb{R}^{d_a}, t_i \in \mathbb{R}^{d_t}, s_u \in [0, 1] \tag{1}$$

By minimizing intra-class distances and maximizing inter-class distances in the embedding space, the proposed approach effectively addresses cross-lingual and cross-modal classification challenges, significantly advancing the state-of-the-art in hate speech detection for synthetic audio.

2.1. Pre-training Phase

In developing the hate speech classification network, we implement a sophisticated pre-training approach that leverages transfer learning and contrastive learning techniques. This method enables us to exploit rich representations from existing models while fine-tuning for our target multimodal hate speech detection task. Let a_i represent the audio input and T_i represent the text input. Our network architecture comprises two primary components: a text encoder f_T and an audio encoder f_A. The weights of the text encoder f_T are initialized with weights θ_T from the SONAR encoder (i.e., $\theta_T = \theta_{SONAR}$), and the audio encoder f_A is initialized with weights θ_A from SeamlessM4T (i.e., $\theta_A = \theta_{SeamlessM4T}$). This choice is motivated by the zero-shot capabilities of SONAR: once trained on a set of languages, the classifier head can seamlessly integrate with any compatible SONAR encoder for different languages.

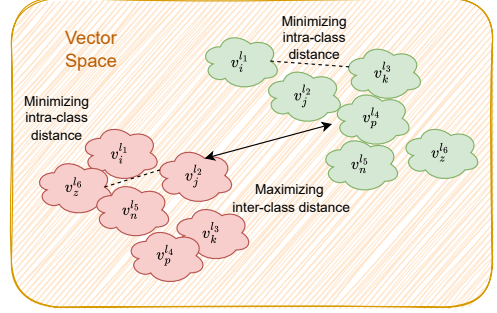The core of our pre-training process lies in the contrastive

Figure 3: *Illustration of contrastive training to maximize inter-class distance while reducing intra-class distance between embeddings of different languages.*

learning framework. We aim to create a common vector space $\mathcal{V} \subset \mathbb{R}^{m \times d}$ where textual and audio representations can be meaningfully compared, independent of language. Here, $m \times d$ denotes the dimensions of the output embedding space. For a batch of N samples, let $\{(a_i, T_i, y_i)\}_{i=1}^N$ represent the audio inputs, text inputs, and their corresponding labels, respectively. The encoders produce embeddings as

$$e_{A_i} = f_A(a_i), \quad e_{T_i} = f_T(T_i).$$

These embeddings are normalized to ensure they reside on a unit hypersphere, as shown in Equation 2:

$$\hat{e}_{A_i} = \frac{e_{A_i}}{\|e_{A_i}\|_2}, \quad \hat{e}_{T_i} = \frac{e_{T_i}}{\|e_{T_i}\|_2}. \tag{2}$$

We compute a similarity matrix $S \in \mathbb{R}^{N \times N}$ by taking the dot product between all pairs of audio and text embeddings, defined as $S_{ij} = \hat{e}_{A_i}^T \hat{e}_{T_j}$. Positive and negative masks are then created based on the labels: the positive mask is defined as $M_{pos_{ij}} = \mathbb{1}[y_i = y_j]$ and the negative mask as $M_{neg_{ij}} = \mathbb{1}[y_i \neq y_j]$.

During pre-training, we optimize the contrastive loss $\mathcal{L}$ via $\min_{\theta_T, \theta_A} \mathcal{L}$. This process encourages the model to maximize the similarity of positive pairs while minimizing that of negative pairs in the shared embedding space $\mathcal{V}$. The use of a common vector space facilitates cross-modal learning, leveraging the relationship

$$\text{sim}(f_A(a_i), f_T(T_j)) \approx \mathbb{1}[y_i = y_j],$$

where, sim is a similarity function (e.g., cosine similarity).

2.2. Downstream Phase

Following the pre-training phase, we proceed to the downstream phase, where we fine-tune our model for the specific task of hate speech classification. In this phase, we train a classifier on top of our pre-trained encoders while continuing to refine their weights. We employ a novel approach that combines triplet loss and binary cross-entropy loss to enhance the model's discriminative power across different languages. Figure 3 illustrates the triplet-based contrastive learning phase.

Our model architecture in this phase consists of the pre-trained audio encoder $f_A : \mathbb{R}^d \to \mathbb{R}^m$ and text encoder $f_T : \mathbb{R}^{n \times d} \to \mathbb{R}^m$, along with a newly introduced classifier $g : \mathbb{R}^{2m} \to [0, 1]$ that takes the concatenated embeddings from both encoders and outputs a probability indicating hate speech. For an input pair (a_i, T_i), the classification process is defined as

$$e_i = [f_A(a_i); f_T(T_i)] \quad \text{and} \quad \hat{y}_i = g(e_i),$$

Table 1: *Distribution of hate and non-hate samples in the proposed dataset.*

Language	Hate		Non-Hate		Total
	Train	Test	Train	Test	
English	6251	9132	1142	782	17307
Hindi	2149	2433	529	617	5728
Marathi	15000	15000	1875	1875	33750
Tamil	1926	1831	815	791	5363
Telugu	259	24340	111	10432	35142
Bengali	7000	14000	3000	600	30000
Total : 127290					

where, $[;]$ denotes concatenation and $\hat{y}_i$ is the predicted probability of hate speech. The loss functions used in this phase are crucial to our approach. We utilize a combination of triplet loss and binary cross-entropy loss. The triplet loss is designed to maximize inter-class distance while minimizing intra-class distance in the embedding space, thereby creating a language-independent vector space for hate and non-hate representations. For each language l, we form triplets (e_a, e_p, e_n) where e_a (anchor) is the embedding of a hate speech sample in language l, e_p (positive) is the embedding of another hate speech sample in language l, and e_n (negative) is the embedding of a non-hate speech sample in a different language $l' \neq l$.

The total loss for the downstream phase is a weighted sum of the triplet loss, $\mathcal{L}_{triplet}$, and the binary cross-entropy loss, $\mathcal{L}_{BCE}$. The complete loss formulation is given by Equation 3:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{triplet} + (1 - \alpha)\mathcal{L}_{BCE}, \quad (3)$$

where, $\alpha \in [0, 1]$ is a hyperparameter controlling the balance between the two losses. During the downstream phase, we optimize

$$\min_{\theta_A, \theta_T, \theta_g} \mathcal{L}_{total},$$

where, θ_A, θ_T, and θ_g are the parameters of the audio encoder, text encoder, and classifier, respectively. This downstream training phase refines our model's hate speech detection capabilities while maintaining robustness across different languages. The triplet loss drives the model to learn language-invariant features of hate speech, while the binary cross-entropy loss ensures accurate classification. The result is a model that can effectively identify hate speech in a multilingual context by leveraging both audio and textual information.

3. Proposed Dataset and Experimental Details

We introduce a novel multimodal, multilingual dataset designed to address the scarcity of resources for hate speech detection, with a focus on low-resource languages. The dataset encompasses both textual and audio modalities across six languages: English, Hindi, Telugu, Tamil, Marathi, and Bengali, chosen for their linguistic diversity and representation in hate speech.
Data Collection and Processing: The dataset was constructed by converting existing text corpora into audio format using state-of-the-art text-to-speech (TTS) synthesis technology, specifically Meta's Massive Multi-Lingual model [13], chosen for its robust multilingual capabilities. Audio samples were standardized to a 16 kHz sampling rate, consistent with common speech processing standards, and limited to 10 seconds in duration to balance information retention and efficiency.

Dataset Composition: The textual sources were selected from peer-reviewed collections to ensure data quality and reliability. The English subset, derived from the HateXplain dataset [2], contains 17,307 social media samples annotated via Amazon Mechanical Turk, ensuring high-quality labels for hate speech, offensive language, and neutral content. The Hindi subset comprises 5,728 samples from the Hostility Dataset [10]. The Marathi subset, with 33,750 samples from [14], is particularly significant due to its large size, addressing the scarcity of hate speech data in this language. The Telugu subset, focused on news website comments, contributes 35,142 samples from [11]. The Tamil subset includes 5,363 samples from the Tamil Hate Speech Project, a collaborative academic initiative annotated by linguistic experts. Lastly, the Bengali subset offers 30,000 samples from [15], sourced from YouTube comments on controversial topics, which were carefully filtered to remove irrelevant content. Table 1 provides the detailed distribution of hate and non-hate samples across languages and train-test splits, highlighting notable variations in balance and size stemming from the differing scales of the underlying text corpora.

To the best of our knowledge, this is the first dataset to leverage synthetically generated speech data for multilingual, multimodal hate speech detection, addressing the lack of paired text-audio resources in low-resource languages. By combining text and audio modalities across multiple low-resource languages, this dataset enables research in cross-lingual and cross-modal hate speech detection, as well as the development of robust multimodal models for real-world applications. The dataset will be publicly available to support further research and development.

3.1. Experimental Setup

We focus on two primary research objectives: *RO1 - Multilingual Deepfake Hate Speech Detection:* To evaluate the accuracy of the proposed dataset and model in detecting hate speech in deepfake audio across multiple languages. *RO2 - Zero-Shot Hate Speech Detection:* To assess the zero-shot learning capabilities of the dataset and model for hate speech detection in languages unseen during training.
Dataset Protocols: To thoroughly evaluate zero-shot learning capabilities, we employ a cross-lingual train-test split. First, we partition each language's data into training and testing subsets using a 70:30 ratio. Next, we define two distinct language sets: (i) **Set-A**: Marathi, Bengali, and Tamil, and (ii) **Set-B**: English, Hindi, and Telugu. These languages were selected to represent diverse linguistic families and ensure a rigorous evaluation of multilingual generalization. We conduct two experiments: (1) training on Set-A and testing on Set-B, and (2) training on Set-B and testing on Set-A. This approach allows us to assess generalization across entirely unseen languages, providing a stringent test of zero-shot learning and multilingual robustness.

3.2. Baselines and Implementation Details

To benchmark the performance of the proposed model, we compare it against six baseline classifiers, selected for their state-of-the-art performance in hate speech detection, multilingual tasks, and multimodal learning:
Text-Based Models: (i) *SentenceBERT* [16]: a fine-tuned BERT model with attention-based pooling, (ii) *HASOC22* [17]: a linear-layer-based architecture optimized on multilingual hate speech data, and (iii) *CNNGRU* [18]: a hybrid model combining convolutional and recurrent (GRU) layers.
Multimodal Models: (i) *ASTBERT* [19]: which combines an Audio Spectrogram Transformer with BERT-based text em-

Table 2: *Results of the Proposed and baseline models when training and testing are on the same set.*

Model	Input	Set-A		Set-B	
		ACC	EER	ACC	EER
ASTBERT	Text + Audio	0.777	0.223	0.669	0.331
HUBERT	Text + Audio	0.762	0.237	0.635	0.369
WAVELMBERT	Text + Audio	0.790	0.210	0.669	0.331
SENTENCEBERT	Text	0.773	0.228	0.653	0.346
HASOC'22	Text	0.755	0.244	0.637	0.359
CNN-GRU	Text	0.738	0.262	0.620	0.380
Proposed	Text + Audio	0.819	0.181	0.701	0.301

Table 3: *Results for the cross-subset evaluation. The results highlight the Zero-Shot capabilities of our Proposed model.*

Model	Input	Train Set-A, Eval Set-B		Train Set-B, Eval Set-A	
		ACC	EER	ACC	EER
ASTBERT	Text + Audio	0.602	0.398	0.750	0.250
HUBERT	Text + Audio	0.572	0.428	0.723	0.277
WAVELMBERT	Text + Audio	0.596	0.403	0.752	0.248
SENTENCEBERT	Text	0.595	0.406	0.746	0.254
HASOC'22	Text	0.581	0.421	0.731	0.270
CNN-GRU	Text	0.560	0.442	0.707	0.293
PROPOSED	Text + Audio	0.625	0.373	0.786	0.214

beddings, (ii) *HUBERTWavBERT* [20]: which integrates a HuBERT-based audio encoder with a BERT text encoder, and (iii) *WAVELMBERT* [21]: which merges WaveLM-based audio embeddings with a BERT text encoder.

All models use the multilingual BERT-uncased tokenizer and are trained for five epochs using cross-entropy loss and the Adam optimizer (learning rate = 0.0001, batch size = 32, dropout rate = 0.2). Performance is evaluated using accuracy, F1-score, and AUC-ROC metrics to ensure a comprehensive assessment of the proposed model's capabilities.

4. Results and Analysis

RO1: Multilingual Deepfake Hate Speech Detection. We evaluate the performance of our proposed model on two language sets: Set-A (Tamil, Marathi, Bengali) and Set-B (English, Hindi, Telugu). These experiments address RO1 by demonstrating the effectiveness of our approach for multilingual deepfake hate speech detection across diverse language groups. For Set-A (see Table 2), our proposed model achieves an accuracy of 0.819 and an Equal Error Rate (EER) of 0.181, highlighting its strong performance. Among the baseline methods, WAVELMBERT performs best with an accuracy of 0.790, followed closely by ASTBERT and SENTENCEBERT. Notably, text-only models (SENTENCEBERT, HASOC'22, CNN-GRU) generally underperform compared to multimodal models, suggesting that incorporating audio features significantly enhances hate speech detection.

The results for Set-B (also shown in Table 2) indicate that our model maintains robust multilingual capabilities, achieving an accuracy of 0.701 and an EER of 0.301. ASTBERT and WAVELMBERT achieve comparable performance, with an accuracy of 0.669. Interestingly, the performance gap between multimodal and text-only models is less pronounced in Set-B than in Set-A. This may suggest that audio features are less informative for these languages or that the models face challenges in generalizing audio features across linguistically distant languages. Overall, our proposed model demonstrates the most consistent performance across all languages and both sets, indicating superior cross-lingual generalization capabilities. Additionally, WAVELMBERT and ASTBERT exhibit strong and consistent performance, making them viable alternatives for multilingual hate speech detection tasks.

RO2: Zero-Shot Hate Speech Detection. We conducted cross-subset experiments to evaluate the models' ability to generalize across different language groups, thereby assessing the zero-shot capabilities of the proposed model. Table 3 presents the results for two scenarios: (1) training on Set-A (Tamil, Marathi, Bengali) and evaluating on Set-B (English, Hindi, Telugu), and (2) training on Set-B and evaluating on Set-A. When training on Set-A and evaluating on Set-B, our proposed model

achieves the highest accuracy of 0.625 and the lowest Equal Error Rate (EER) of 0.373, outperforming all baselines. ASTBERT follows with an accuracy of 0.602, while other models show performance ranging from 0.560 to 0.596. These results highlight the challenge of generalizing from the languages in Set-A to those in Set-B, potentially due to linguistic differences and data characteristics.

Interestingly, when training on Set-B and evaluating on Set-A, all models demonstrate a significant improvement. Our proposed model achieves the highest accuracy of 0.786 and an EER of 0.214, with WAVELMBERT and ASTBERT following closely at 0.752 and 0.750, respectively. This substantial improvement suggests that the models can better generalize from the languages in Set-B to those in Set-A. The presence of English in Set-B - a widely used language with abundant training data - may have contributed to more robust feature learning, facilitating cross-lingual transfer. The consistent outperformance of multimodal models in these scenarios further indicates that incorporating audio features provides valuable information that generalizes well across language sets. This emphasizes the importance of multimodal approaches in multilingual hate speech detection tasks. Overall, the superior performance of our proposed model in both cross-subset experiments demonstrates its robust zero-shot capabilities and highlights the need for models that can effectively generalize across diverse language groups, particularly when transferring from less commonly spoken to more widely spoken languages.

5. Conclusion

This paper presents a novel zero-shot hate speech detection approach for audio, addressing challenges posed by multimodal content in low-resource languages. Our contributions include: (1) a framework that integrates audio and text modalities using contrastive learning; (2) the first comprehensive multimodal dataset for deepfake hate speech detection, covering English and five low-resource Indian languages; and (3) a two-stage training process combining contrastive pre-training and supervised contrastive learning. Our proposed model consistently outperforms existing baselines, with multimodal approaches demonstrating superiority over text-only models in both same-language and cross-language scenarios. This work highlights the value of incorporating audio features alongside text for robust hate speech detection in low-resource contexts. Future work will expand the dataset with diverse deepfake audio samples featuring varied voice characteristics, accents, and emotional tones to further enhance detection capabilities.

6. Acknowledgement

7. References

[1] K. Müller and C. Schwarz, "Fanning the flames of hate: Social media and hate crime," June 5 2020, available at SSRN: https://ssrn.com/abstract=3082972 or http://dx.doi.org/10.2139/ssrn.3082972.

[2] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in *AAAI*, 2021, pp. 14 867–14 875.

[3] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis," in *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Nov. 2019. [Online]. Available: https://aclanthology.org/D19-1474

[4] Ò. G. i Orts, "Multilingual detection of hate speech against immigrants and women in twitter at semeval-2019 task 5: Frequency analysis interpolation for hate in speech detection," in *SemEval@NAACL-HLT*. Association for Computational Linguistics, 2019, pp. 460–463.

[5] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "Ethos: a multi-label hate speech detection dataset," *Complex & Intelligent Systems*, vol. 8, no. 6, pp. 4663–4678, 2022.

[6] S. Communication, L. Barrault, Y. Chung, M. C. Meglioli, D. Dale, N. Dong, P. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, C. Klaiber, P. Li, D. Licht, J. Maillard, A. Rakotoarison, K. R. Sadagopan, G. Wenzek, E. Ye, B. Akula, P. Chen, N. E. Hachem, B. Ellis, G. M. Gonzalez, J. Haaheim, P. Hansanti, R. Howes, B. Huang, M. Hwang, H. Inaguma, S. Jain, E. Kalbassi, A. Kallet, I. Kulikov, J. Lam, D. Li, X. Ma, R. Mavlyutov, B. N. Peloquin, M. Ramadan, A. Ramakrishnan, A. Y. Sun, K. Tran, T. Tran, I. Tufanov, V. Vogeti, C. Wood, Y. Yang, B. Yu, P. Andrews, C. Balioglu, M. R. Costa-jussà, O. Celebi, M. Elbayad, C. Gao, F. Guzmán, J. Kao, A. Lee, A. Mourachko, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, P. Tomasello, C. Wang, J. Wang, and S. Wang, "Seamlessm4t-massively multilingual & multimodal machine translation," *CoRR*, vol. abs/2308.11596, 2023.

[7] S. Ghosh, S. Lepcha, S. Singh, R. R. Shah, and S. Umesh, "Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances," in *INTERSPEECH*. ISCA, 2022, pp. 5185–5189.

[8] M. Yousefi and D. Emmanouilidou, "Audio-based toxic language classification using self-attentive convolutional neural network," in *29th European Signal Processing Conference, EUSIPCO 2021, Dublin, Ireland, August 23-27, 2021*. IEEE, 2021, pp. 11–15. [Online]. Available: https://doi.org/10.23919/EUSIPCO54536.2021.9616001

[9] M. R. Costa-jussà, M. C. 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, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 5725–5734. [Online]. Available: https://aclanthology.org/2024.findings-acl.340

[10] M. Bhardwaj, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Hostility detection dataset in hindi," *ArXiv*, vol. abs/2011.03588, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:226281866

[11] M. Marreddy, S. R. Oota, L. S. Vakada, V. C. Chinni, and R. Mamidi, "Am I a resource-poor language? data sets, embeddings, models and analysis for four different NLP tasks in telugu language," *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, vol. 22, no. 1, pp. 18:1–18:34, 2023.

[12] 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, pp. 6172–6176.

[13] H. Abelson, G. J. Sussman, and J. Sussman, *Structure and Interpretation of Computer Programs*. Cambridge, Massachusetts: MIT Press, 1985.

[14] H. Patil, A. Velankar, and R. Joshi, "L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models," in *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, 2022, pp. 1–9.

[15] N. Romim, M. F. Ahmed, H. Talukder, and M. S. Islam, "Hate speech detection in the bengali language: A dataset and its baseline evaluation," *ArXiv*, vol. abs/2012.09686, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:229298046

[16] P. Burnap and M. L. Williams, "Hate speech, machine classification and statistical modelling of information flows on twitter: interpretation and communication for policy decision making," 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:142840744

[17] S. Satapara, P. Majumder, T. Mandl, S. Modha, H. Madhu, T. Ranasinghe, M. Zampieri, K. North, and D. Premasiri, "Overview of the hasoc subtrack at fire 2022: Hate speech and offensive content identification in english and indo-aryan languages," in *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, ser. FIRE '22. New York, NY, USA: Association for Computing Machinery, 2023, p. 4–7. [Online]. Available: https://doi.org/10.1145/3574318.3574326

[18] Z. Zhang, D. Robinson, and J. A. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *ESWC*, ser. Lecture Notes in Computer Science, vol. 10843. Springer, 2018, pp. 745–760.

[19] Y. Gong, Y. Chung, and J. R. Glass, "AST: audio spectrogram transformer," *CoRR*, vol. abs/2104.01778, 2021. [Online]. Available: https://arxiv.org/abs/2104.01778

[20] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *CoRR*, vol. abs/2106.07447, 2021. [Online]. Available: https://arxiv.org/abs/2106.07447

[21] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *CoRR*, vol. abs/2110.13900, 2021. [Online]. Available: https://arxiv.org/abs/2110.13900