INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Investigating Acoustic Cues for Multilingual Abuse Detection

Yash Thakran[1], *Vinayak Abrol*[2]

ECE Department[1], Infosys Centre for AI & CSE Department[2], IIIT Delhi, India
yash20269,abrol@{iiitd.ac.in}

Abstract

This work focuses on audio abuse detection from an acoustic cue perspective in a multilingual social media setting. While textual abuse detection has been widely researched, comparatively, abuse detection from audio remains unexplored. Our key hypothesis is based on the fact that abusive behavior leads to distinct acoustic cues. Such cues can help detect abuse directly from audio signals without the need to transcribe them. We first demonstrate that employing a generic large pre-trained acoustic/language model is suboptimal. This proves that incorporating the right acoustic cues might be the way forward to improve performance and achieve generalization in a large-scale setting. Our proposed method explicitly focuses on two modalities: the underlying emotions expressed and the language features of audio. On the recently proposed ADIMA benchmark for this task, our approach achieves the state-of-the-art performance of 96% on the test set and outperforms existing best models by a large margin.

Index Terms: multilingual abuse detection, abusive speech detection, speech processing, transfer learning

1. Introduction

Due to the extensive use of social media platforms, abusive content detection in online material has drawn a lot of attention. Profanity, cyberbullying, racial slurs, hate speech, and other prevalent abusive behaviors call for strong content moderation algorithms to guarantee safe and healthy communication [1] as many users suffer harassment in the form of targeted personal or communal attacks. These attcks create a negative user experience and may have long-lasting psychological impacts that demand timely detection and prevention of abusive behavior [2, 3]. The majority of recent research in this domain has been on finding abusive conduct in textual data extracted from social media posts/comments [4, 5, 6] or uploaded multimedia content like images, memes, or videos [7, 8, 9, 10]. As a workaround to handling audio data, existing works employ automatic speech recognition (ASR) systems for audio transcription followed by text processing over the transcriptions to detect profanity in audio [11]. However, this calls for precise ASR systems, which are expensive to train, especially in a multilingual setting. Further, since there is a dearth of profane words in the training corpora, ASR accuracy on these words may be subpar. Also, the effectiveness of an ASR is further diminished by a lack of clarity and incompleteness with which abusive words are typically spoken. Alternatively, few works have formulated and studied this problem as a keyword detection task using a dictionary of audio examples of abusive words and a template-matching approach [12, 13]. Updating the dictionary with novel words and creating templates that capture the variances in style, accent, dialect, and environmental factors is challenging and computationally expensive, whereas a neural model just needs fine-tuning without increasing the model size.

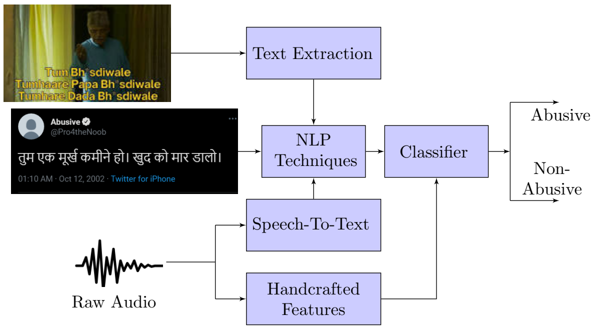


Figure 1: *Abuse detection pipeline for various data modalities.*

In contrast to existing works on abuse detection in the visual [14, 8] and textual [15, 16, 17] domains, the detection of abusive content in raw audio has received little attention. In this work, we try to establish the importance of acoustic cues in detecting abusive content in multilingual spoken utterances. We demonstrate that ASR transcriptions (even from large vocabulary SOTA models) are suboptimal since they only capture semantic information. We also demonstrate that employing large pre-trained acoustic or language models performs poorly compared to small raw-waveform models. Abusive behavior leads to distinct and important acoustic cues such as pitch, tone, and emotions. This is intuitive as people tend to be irate, irritated, or loud when acting abusively [18], and there are inherent dynamics due to the spoken language & accent, which is often multilingual in practical scenarios.

Considering these issues, we propose an approach that uses two important facets of information present in audio utterances: language features of the audio and emotions expressed in them for robust detection of abuse in spoken audio. We use modality-specific models for extracting the features for each modality followed by multimodal fusion and observe substantial improvements by using these modalities in an end-to-end setup for multiple languages. To this aim, we used the recently proposed *Abuse Detection In Multilingual Audio* (ADIMA) [19] benchmark for this task, where our approach achieves the state-of-the-art performance of 96% on the test set and outperforms existing best models by a large margin. We observe a strong correlation

This work is supported by SERB-Startup Research Grant (SRG/2021/002348), Govt. of India and Infosys Foundation via Infosys Centre for AI, IIIT Delhi.

10.21437/Interspeech.2023-1311

between the display of abusive behavior, language features, and emotions expressed in the corresponding audio utterances.

2. Related work

Significant efforts in the domain of abuse detection have been devoted to abusive text [20, 15], images [9], videos [14, 8], and comments [16, 17]. A pictorial summary of existing abuse detection pipelines with various data modalities is presented in Figure1. Existing approaches that are employed to perform abuse detection vary from using lexical features together with metadata [21] to machine/deep learning models and their derivatives [22]. Furthermore, various natural language processing (NLP) approaches involving transfer learning, distillation [23], and keyword-based protocols have also gained popularity. Most of the widespread abuse detection approaches in NLP deal with offensive content published via comments on online platforms [15, 16]. While audio-based abuse detection is challenging and relatively uncommon, there have been previous attempts to perform isolated limited-vocabulary swear word detection [24], read-out abuse classification [25], and small-vocabulary keyword spotting [26]. However, these approaches do not capture the overall context as they do not deal with real-world conversational examples and require set vocabulary or templates of abusive words. Moreover, abusive words are ever-evolving and usually not spoken completely entirely, making understanding the overall context very important for this task.

Authors in [19] recently proposed ADIMA, a novel, and linguistically diverse multilingual profanity detection audio dataset. They presented baselines for monolingual and zero-shot cross-lingual settings using medium (VGG) to large-scale (Wav2Vec2) pre-trained deep learning models as backbones feature extractors. Incorporating emotional attributes (which are strongly linked to offensive behavior in real-life) boosts the performance of abuse detection as opposed to using only video or text-based features, as shown in [18, 27]. As established in the literature, emotion can be detected with greater accuracy from raw audio compared to its corresponding text [28, 29]. Exploiting this, in [30], authors proposed *Multimodal Abuse Detection in Audio* (MADA) on the ADIMA benchmark. They investigate the significance of self-supervised modeling of the audio (context encoder), its underlying emotions (emotion encoder), and the semantic information present in its transcribed text (text encoder) to show gains over the original ADIMA benchmark. Their detailed ablation experiments evaluate the contribution of every modality and performance gain by using information from all the modalities together. They also presented results with a two-stage process (TSP) originally proposed in [11] for a different dataset. Here, the first stage involves transcribing the audio into text, and the second stage comprises training a text-based classifier over the transcriptions.

3. Abuse Detection from Raw Audio Signal

In order to motivate our approach, we first highlight the current pitfalls of the existing works. In particular, we demonstrate that without understanding the requirements of the task at hand, the use of large pre-trained models for this task in [19, 30] is not the right way forward. Similarly, using multiple modalities and various handcrafted acoustic features only lead to minuscule performance gains. To elaborate on this and support our argument, we present an experiment where we trained a shallow 1D-convolution based ResNet and InceptionNet architecture with 6 million and 1.6 million parameters respectively, by completely relying on audio signals without any text transcriptions or handcrafted acoustic features. Results of this experiment are reported in Table1 (see columns 5 & 6), while the model & experimental details are presented in Sections 4 & 5.

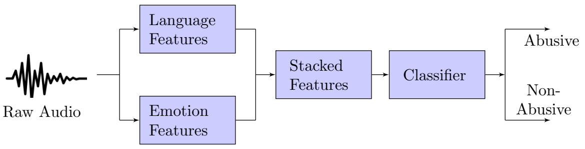


Figure 2: *Block diagram of the proposed ACMAD framework*

It can be observed that a shallow model trained on raw audio signals achieves comparable results (on a majority of individual languages) to the previous SOTA approaches employing pre-trained models and/or multiple modalities. This shows that generic unsupervised/semi-supervised pre-trained acoustic models (irrespective of how large they are) do not capture the important acoustic cues for abuse detection. They either need to be fine-tuned or augmented with the right feature modality. While with combined modalities, MADA achieved improved overall performance, the performance gain over our low-footprint raw-waveform models is small considering the overall computational complexity in terms of size and inference time of context, emotion, and text encoders along with the classifier training for abuse detection.

3.1. Acoustic Cues for Multilingual Abuse Detection (ACMAD)

Motivated by the above study, we present the proposed ACMAD framework that exploits important acoustic cues to achieve SOTA performance on the ADIMA benchmark. Figure2 shows the ACMAD detection pipeline comprising of the following blocks: 1) pre-trained acoustic model for language features; 2) pre-trained acoustic model for emotion features; 3) dimensionality reduction and feature stacking module; 4) neural network classifier. The novelty of ACMAD lies in carefully selecting the models/algorithms in each of its modules.

3.1.1. Language features

Language, accent, and dialect profoundly impact how abusive words are spoken or delivered. For instance, a given keyword might be easier to detect in one language than another. Hence incorporating language information during feature extraction is very important. Since the ADIMA benchmark deals with Indic languages, ACMAD employs IndicWav2Vec-Base multilingual model [31] that is pre-trained on 40 Indian languages in a self-supervised manner[1]. We further performed a supervised fine-tuning for the downstream language identification task to boost the performance. Here, embedding are obtained from the encoder module (feature extractor) of the fairseq Wav2Vec model.

3.1.2. Emotion features

Various studies in the literature have established the correlation between abusive behavior and human emotions. This was, in fact, the motivation of MADA [30] to incorporate various handcrafted acoustic features for abuse detection. However, such features (like MFCC, mel-spectrogram, and chroma) are very generic in that they not only capture emotions but are useful for various other speech/audio tasks. Hence, they can become

[1] https://tinyurl.com/48pcumyz

Table 1: *Abuse detection test accuracy (%) of various approaches on different languages. Results for 1D-ResNet, 1D-IncNet, and ACMAD have been averaged over 10 trials.*

Language	**ADIMA**	**TSP**	**MADA**	**1D-ResNet (ours)**	**1D-IncNet (ours)**	**ACMAD (ours)**	
						Acc	**F1 Score**
Hindi	79.67	79.0	84.82	77.51	77.78	96.75	0.957
Bengali	81.08	76.0	82.43	72.43	74.86	96.76	0.958
Bhojpuri	76.48	73.0	78.27	70.83	70.83	97.02	0.959
Gujarati	80.94	69.0	82.04	80.11	81.22	96.69	0.9584
Haryanvi	81.15	68.0	83.06	78.69	79.78	96.72	0.9581
Kannada	82.92	58.0	83.47	79.95	79.95	96.75	0.9580
Malayalam	86.29	77.0	85.48	82.26	83.33	96.77	0.9580
Odia	83.29	82.0	82.46	80.55	79.45	96.71	0.9578
Punjabi	82.01	78.0	85.01	82.02	80.93	96.73	0.957
Tamil	80.59	82.0	82.21	80.78	79.78	96.77	0.9577
Average	81.44	74.2	82.92	78.51	78.79	96.76	0.9579

a bottleneck for the downstream classifier that has to exploit information from multiple modalities for abuse detection. In contrast, ACMAD employs XLS-R 300M model [32] pre-trained on 128 languages and further finetuned on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [33] for predicting 8 different emotion classes[2]. Here, features are extracted from the projector block of the model, i.e., the 2nd last linear layer.

3.1.3. Feature stacking

ACMAD uses adaptive average pooling on the extracted embedding from the pre-trained language and emotion models to obtain features of size 256. Since both the pre-trained models are built on the Wav2Vec2 backbone, we found the obtained features to be in a similar range/scale. Hence, no normalization or PCA was performed to remove redundant or non-discriminative dimensions. Finally, the obtained features are stacked before feeding them to a CNN classifier.

3.1.4. Classifier

The extracted and stacked features from audio are used to train 1D-convolutional network that learns to classify them into abusive and non-abusive classes. Since the pre-trained models have their own footprint, we ensured the classifier has a very low footprint. As discussed in Section 5, we also experimented with various other ML classifiers that resulted in suboptimal performance compared to the CNN classifier.

4. Experimental Setup

4.1. Dataset

In this work, we used the recently proposed ADIMA dataset that contains audio recordings of ShareChat chat rooms totaling 65 hours for 10 Indian languages - Hindi (Hi), Bengali (Be), Punjabi (Pu), Haryanvi (Ha), Kannada (Ka), Odia (Od), Bhojpuri (Bh), Gujarati (Gu), Tamil (Ta) and Malayalam (Ma). The dataset is balanced across languages and contains records from 6446 users, making it a diverse, multilingual, and multi-user dataset. Each audio recording is labeled as abusive or non-abusive. Extracted from real conversations, these recordings inherently have corruption due to environmental noise and distortions due to poor recording devices or improper placement of the recording device. We use the standard train/test splits (70:30 for each language) provided by [19] across all the experiments for fair comparison and report accuracy on the test set.

4.2. Model architecture

In this work, we experimented with 1D-convolution based Residual network (ResNet) and Inception network (IncNet) adapted from [34]. Both the models share the same first layer, a single 1D convolution layer that acts as a trainable filter bank [35] and operates on the input using a short kernel of size 3 and stride 1. The input layer is followed by multiple ResNet-style or Inception-style blocks. Global max-pooling combines the channel outputs of the last block into a fixed-size feature vector fed to linear layers. The classifier comprises two linear layers for downsampling and another linear layer for classification into abusive or non-abusive categories.

For experiments with raw waveforms, these models operate on raw samples, while in the case of the proposed ACMAD framework, they operate on the stacked features obtained from pre-trained models. Each model is trained for 100 epochs with cross-entropy loss, SGD optimizer with momentum, learning rate (LR) of 1e-4, exponential-decay ($\gamma = .95$) LR scheduler, batch size set to 32, and with the same experimental setup using a single Nvidia RTX3090 GPU for a fair comparison or benchmarking.

4.3. Reproducible research

All the experiments are reproducible & implementation is available on GitHub[3]

5. Results and Discussion

5.1. Experiment with raw waveform models

As described in Section3, our experiments with shallow low-footprint raw-waveform models were done to establish the importance of relevant acoustic cues for the task of abuse detection over text transcriptions. In particular, we trained a 1D-ResNet and 1D-IncNet model on raw audio, and the results of this experiment are reported in Table1 (columns 5 and 6). It can be observed that in all languages, the average test accuracy of these

[2] https://tinyurl.com/2vm9stjn

[3] `https://github.com/Cross-Caps/ACMAD`

Table 2: *Abuse detection test accuracy (%) over 10 trials of various classifiers using language and emotion features.*

ML classifier	Acc
CNN	**96.76**
GPC	82.42
SVM (polynomial kernel)	80.61
SVM (linear kernel)	78.94
MLP	75.78
RF	74.84
MADA	82.92

models differs by approximately 2.6% and 4.1% as compared to ADIMA and MADA, respectively, with much lower computational complexity. Although not the aim of this work, we argue that one can easily boost the accuracy by using a raw-waveform based model with more parameters. These results suggest that generic pre-trained models either need to be fine-tuned or augmented with the right feature modality. Further, note the performance gain in the case of MADA pertains to the inclusion of emotional features that have been shown to be well correlated with abusive behavior.

In comparison to TSP, raw-waveform models achieve better results for five languages with an average gain of 4.5% of all languages. The performance of TSP degrades substantially for some of the languages, which can be attributed to the inferior quality of text transcriptions. This is intuitive as the ASR models used Wav2Vec2 are not trained on data containing instances of abusive words in their vocabulary resulting in substitutions or deletions, justifying the dip in the test accuracy. Furthermore, the ADIMA dataset contains code-mix languages, which further makes recognition challenging.

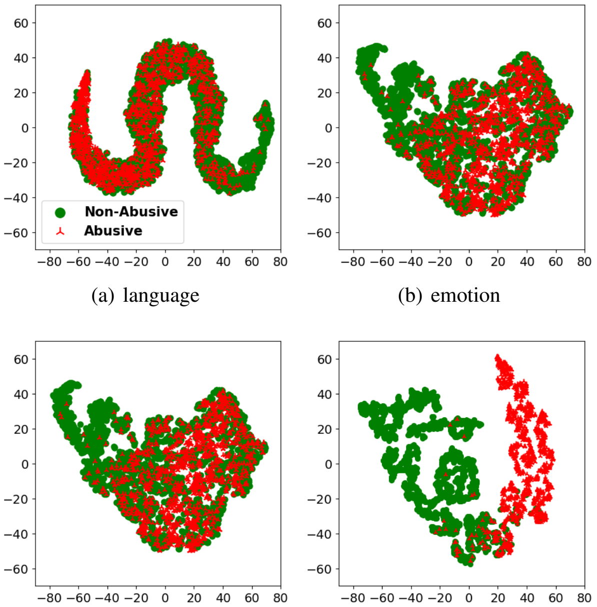
5.2. Experiments with stacked features

We now present our main experiments with language & emotional features extracted from pre-trained models as described in Section3.1. Irrespective of the inference time, since pre-trained models have their own footprint; our aim is to achieve the best classification results with a lightweight model having few parameters. To this aim, we first experimented with four popular ML classifiers, namely support vector machines (SVM) [linear & degree two polynomial kernel, regularization parameter C=0.025], random forest (RV) [50 trees with maxdepth 5 & gini metric], gaussian process classifier (GPC) [RBF kernel with length scale=1], and multilayer perception (MLP) [2 layers with 100 & 2 neurons]. These models were trained on concatenated language & emotional features.

Results of these experiments in terms of average test accuracy are reported in Table2. It can be observed that a combination of proposed features and a powerful classifier like GPC can achieve performance at par with MADA that use multiple modalities/features. Other classifiers do reasonably well in comparison to ADIMA and TSP. Our best results are achieved with a CNN classifier where we trained a 1D-IncNet with only 14000 parameters. Here inputs are stacked features as they are easier to work with. CNN, due to its local modeling capability, is able to better capture the variability in features as compared to classical classifiers. Our results establish a new SOTA on the ADIMA dataset outperforming all existing best models by a large margin.

5.2.1. *Visualization of embedding*

In order to highlight the effectiveness of our proposed approach, in Figure3 we show t-SNE visualizations of the embedding/features. It can be observed that while the language & emotional embedding are not inherently discriminative, the processed embedding extracted from the trained CNN show a high discriminative property. This further provides evidence of the effectiveness of a CNN classifier in modeling rich and complex features extracted from pre-trained models.



(a) language

(b) emotion

(c) language + emotion

(d) CNN

Figure 3: *t-SNE visualization of various embedding*

6. Conclusions

Prevention of abusive content is crucial for facilitating safe and healthy interactions. In this work, we explore audio abuse detection from the perspective of using acoustic cues in a multilingual setting over 10 Indic languages. We investigate the significance of language information and underlying emotions present in the audio signals. We also highlighted the current pitfalls in existing works and proposed a framework that addresses them. And in doing so, we achieved the SOTA performance on ADIMA dataset that outperforms existing best results by a 13.84% average accuracy difference.

7. References

[1] I. Sarridis, C. Koutlis, O. Papadopoulou, and S. Papadopoulos, “Leveraging large-scale multimedia datasets to refine content moderation models,” in *IEEE Eighth International Conference on Multimedia Big Data (BigMM)*, December 2022, pp. 125–132.

[2] V. Chavan and S. S S, “Machine learning approach for detection of cyber-aggressive comments by peers on social media network,” in *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, August 2015, pp. 2354–2358.

[3] M. O’Reilly, N. Dogra, N. Whiteman, J. Hughes, S. Eruyar, and P. Reilly, “Is social media bad for mental health and wellbeing? exploring the perspectives of adolescents,” *Clinical Child Psy-*

chology and Psychiatry, vol. 23, no. 4, pp. 601–613, October 2018.

[4] 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 Conference on Artificial Intelligence*, vol. 35, 2021, pp. 14 867–14 875.

[5] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *International Workshop on Semantic Evaluation*, June 2019, pp. 54–63.

[6] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "ETHOS: a multi-label hate speech detection dataset," *Complex & Intelligent Systems*, pp. 1–16, 2020.

[7] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," in *International Conference on Neural Information Processing Systems*, December 2020.

[8] C. Alcântara, V. Moreira, and D. Feijo, "Offensive video detection: Dataset and baseline results," in *Language Resources and Evaluation Conference*, May 2020, pp. 4309–4319.

[9] A. Gangwar, E. Fidalgo, E. Alegre, and V. González-Castro, "Pornography and child sexual abuse detection in image and video: A comparative evaluation," in *International Conference on Imaging for Crime Detection and Prevention (ICDP)*, December 2017, pp. 37–42.

[10] C. S. Wu and U. Bhandary, "Detection of hate speech in videos using machine learning," in *International Conference on Computational Science and Computational Intelligence (CSCI)*, December 2020, pp. 585–590.

[11] 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*, September 2022, pp. 5185–5189.

[12] A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, "Offensive language detection using multi-level classification," in *Advances in Artificial Intelligence*, December 2010, pp. 16–27.

[13] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *International Conference on Privacy, Security, Risk and Trust and International Confernece on Social Computing*, September 2012, pp. 71–80.

[14] Z. Gao, S. Yada, S. Wakamiya, and E. Aramaki, "Offensive language detection on video live streaming chat," in *International Conference on Computational Linguistics*, December 2020, pp. 1936–1940.

[15] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media," in *Workshop on Semantic Evaluation*, December 2020, pp. 2054–2059.

[16] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *International World Wide Web Conferences Steering Committee*, April 2016, p. 145–153.

[17] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on Twitter," in *Workshop on Abusive Language Online*, August 2017, pp. 41–45.

[18] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "A multi-task learning approach to hate speech detection leveraging sentiment analysis," *IEEE Access*, vol. 9, pp. 112 478–112 489, August 2021.

[19] V. Gupta, R. Sharon, R. Sawhney, and D. Mukherjee, "ADIMA: abuse detection in multilingual audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6172–6176.

[20] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *International AAAI Conference on Web and Social Media*, vol. 11, pp. 512–515, March 2017.

[21] A. Koufakou, E. W. Pamungkas, V. Basile, and V. Patti, "HurtBERT: Incorporating lexical features with BERT for the detection of abusive language," in *Workshop on Online Abuse and Harms*, Novenmer 2020, pp. 34–43.

[22] S. Kaur, S. Singh, and S. Kaushal, "Abusive content detection in online user-generated data: A survey," *Procedia Computer Science*, vol. 189, pp. 274–281, 2021.

[23] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," in *International Workshop on Complex Networks & Their Applications*, December 2019.

[24] S. N. Endah, D. M. K. Nugraheni, S. Adhy, and Sutikno, "The automation system censor speech for the indonesian rude swear words based on support vector machine and pitch analysis," *IOP Conference Series: Materials Science and Engineering*, vol. 190, p. 012039, April 2017.

[25] S. K. Rahut, R. Sharmin, and R. Tabassum, "Bengali abusive speech classification: A transfer learning approach using vgg-16," in *Emerging Technology in Computing, Communication and Electronics (ETCCE)*, December 2020, pp. 1–6.

[26] F. I. Ablaza, T. O. D. Danganan, B. P. L. Javier, K. S. Manalang, D. E. V. Montalvo, and L. U. Ambata, "A small vocabulary automatic filipino speech profanity suppression system using hybrid hidden markov model/artificial neural network (hmm/ann) keyword spotting framework," in *International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, November 2014, pp. 1–5.

[27] S. Rajamanickam, P. Mishra, H. Yannakoudakis, and E. Shutova, "Joint modelling of emotion and abusive language detection," *Association for Computational Linguistics*, pp. 4270–4279, July 2020.

[28] P. Sharma, V. Abrol, A. Sachdev, and A. D. Dileep, "Speech emotion recognition using kernel sparse representation based classifier," in *European Signal Processing Conference (EUSIPCO)*, August 2016, pp. 374–377.

[29] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech emotion recognition using speech feature and word embedding," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, November 2019, pp. 519–523.

[30] R. Sharon, H. Shah, D. Mukherjee, and V. Gupta, "Multilingual and multimodal abuse detection," in *Interspeech*, September 2022, pp. 4631–4635.

[31] T. Javed, S. Doddapaneni, A. Raman, K. S. Bhogale, G. Ramesh, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Towards building asr systems for the next billion users," *AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10 813–10 821, June 2022.

[32] 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," *Interspeech*, pp. 2278–2282, September 2022.

[33] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," April 2018.

[34] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards end-to-end synthetic speech detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1265–1269, 2021.

[35] V. Singh, V. Abrol, and K. Nathwani, "On spectral and temporal feature encoding behaviour in stacked architectures," in *NeurIPS workshop on Efficient Natural Language and Speech Processing (ENLSP-II)*, December 2022, pp. 1–5.