



Multilingual and Multimodal Abuse Detection

Rini Sharon, Heet Shah, Debdoot Mukherjee, Vikram Gupta

ShareChat, India

{rinisharon, heetshah, debdoot, vikramgupta}@sharechat.co

Abstract

The presence of abusive content on social media platforms is undesirable as it severely impedes healthy and safe social media interactions. While automatic abuse detection has been widely explored in textual domain, audio abuse detection still remains unexplored. In this paper, we attempt abuse detection in conversational audio from a multimodal perspective in a multilingual social media setting. Our key hypothesis is that along with the modelling of audio, incorporating discriminative information from other modalities can be beneficial for this task. Our proposed method, MADA, explicitly focuses on two modalities other than the audio itself, namely, the underlying emotions expressed in the abusive audio and the semantic information encapsulated in the corresponding text. Observations prove that MADA demonstrates gains over audio-only approaches on the ADIMA dataset. We test the proposed approach on 10 different languages and observe consistent gains in the range 0.6%-5.2% by leveraging multiple modalities. We also perform extensive ablation experiments for studying the contributions of every modality and observe the best results while leveraging all the modalities together. Additionally, we perform experiments to empirically confirm that there is a strong correlation between underlying emotions and abusive behaviour. Code is available at <https://github.com/ShareChatAI/MADA>

Index Terms: abusive speech detection, multimodal abuse detection, multilingual abuse detection

1. Introduction

Timely detection and prevention of abusive behaviour on social media is an important problem as multiple users face harassment in the form of targeted personal/communal attacks, which leads to an unpleasant user experience and can cause lasting psychological effects [1, 2]. With an exponential increase in the use of social media by people from different regional and cultural backgrounds, it is of paramount importance to develop accurate content moderation models for recognizing and filtering offensive interactions in multilingual social media environment. Abuse detection in visual [3, 4] and textual [5, 6, 7, 8] domains has been well explored. However, abuse detection in spoken audio is relatively under explored. In this work, we attempt to identify the presence of abusive words in spoken utterances in a multilingual setting. A two-stage process of transcribing the spoken audio into text using automatic speech recognition (ASR) systems followed by using natural language processing based abuse detection methods designed for text is a plausible approach. While the transcribed text captures the semantic information, it is not able to represent the audio cues like pitch, volume, tone, emotions and so on, which can often play an important role in abuse detection as humans are generally angry, agitated or loud while displaying abusive behaviour [9, 10].

Considering these issues, we propose a multimodal approach which leverages three important facets of information

present in the utterances: a) content present in the audio b) emotions expressed in them and c) textual semantics of the waveform, 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. We also observe that only modelling the emotion embeddings showcases competitive performance. These results indicate a strong correlation between the display of abusive behaviour and emotions expressed in the corresponding audio utterances. To the best of our knowledge, our work is the first one to explore the combination of audio, emotions and transcribed text for abuse detection in conversational speech. Our major contributions can be summarized as:

- We explore a multimodal approach for abuse detection in multilingual spoken text.
- We show that explicitly integrating emotions and semantic information present in the transcribed text results in promising state-of-the-art performance.
- We also demonstrate that there exists a strong correlation between emotions and abusive behaviour.
- We perform extensive experiments to highlight the contribution of each modality and compare them with a two-stage process (transcription followed by text classification) for abuse detection.

2. Related Work

Significant efforts in the domain of abuse detection have been devoted to abusive text [5, 6], image [11], video [3, 4] and comments [7, 8]. Existing approaches that are employed to perform abuse detection vary from using lexical features together with meta data [12], to machine/deep learning models, and their derivatives [13, 14]. Furthermore, various natural language processing (NLP) approaches involving transfer learning, distillation [15], and keyword-based protocols have also gained popularity. While audio-based abuse detection is challenging and fairly uncommon, there have been previous attempts to perform isolated limited-vocabulary swear word detection [16], read-out abuse classification [17] and small-vocabulary keyword spotting [18]. However, these approaches do not capture the overall context as they do not deal with conversational real-world examples and require set vocabulary or templates of abusive words. Moreover, abusive words are ever-evolving and are usually not spoken completely and clearly, making the understanding of the overall context very important for this task.

Incorporating emotion attributes (which are strongly linked to offensive behaviour in real-life) boosts the performance of abuse detection as opposed to using only the video or text based features as shown in [9, 10, 19]. Furthermore, research suggests that emotion can be detected with greater accuracy in speech than from its corresponding text [20, 21], making emotion one of our main choices among modalities.

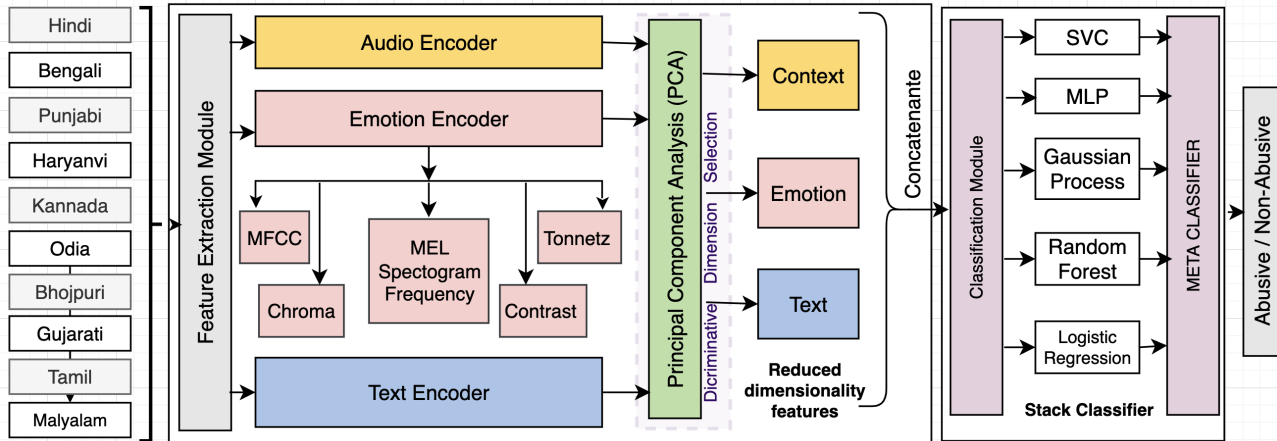


Figure 1: Overall architecture of MADA

Majority of the popular abuse detection approaches in NLP deal with offensive content published via comments in online platforms [14, 7, 6]. This motivates us to study the text modality as well. To the best of our knowledge, interplay of emotion and textual transcriptions for hate-speech detection in spoken audio has not been explored till date. A closely related work on audio abuse detection is studied in [22] using annotated English speech from various open-sourced databases. However, [22] do not explore the problem from a multimodal and multilingual perspective.

To fill this gap, in this paper, we explore abuse detection in audio by leveraging audio features alongside modelling emotions and semantics encapsulated in the transcribed text in a multimodal and multilingual setup. We also perform experiments to evaluate the effectiveness of emotions for this task and observe competitive performance by using the emotion representations alone as inputs.

3. Task

We consider the task of classifying audio utterances into abusive and non-abusive categories using different modalities individually and in combination. In this work, audio, emotion and textual features are extracted using modality-specific encoders and fused by concatenation before passing them through the classifier. Formally, let $X = (x_1, \dots, x_n)$ be the input features fused from multiple modalities of n data instances. Then our output $Y = (y_1, \dots, y_n)$ are corresponding target labels where $y_i \in \{0, 1\}$, with 0 indicating non-abusive utterance and 1 indicating the presence of abuse.

4. Multimodal Abuse Detection in Audio: MADA

In this section, we discuss the overall architecture and core components of MADA in detail. The following subsections explain in detail the three encoders proposed in our approach for multimodal feature extraction. The features extracted from the encoders are passed through a dimensionality reduction step before performing early fusion across modalities. The fused features are then passed through a binary classifier and trained using binary cross entropy loss. Figure 1 shows the overall architectural flow of our approach.

4.1. Audio Encoder

For extracting features from the audio, we use wav2vec2 [23, 24], which is a transformer-based model trained in a self-supervised way in a multilingual framework. We consider two pretrained wav2vec2 models, namely, XLSR-53 model (trained using 53 languages which include a few Indic languages) [25] and CLSRIL-23 (trained using Indic languages) [26]. We first pass the raw audio files to extract the context embeddings from the fully connected layer and then mean-pool across the temporal dimension to generate the final embeddings.

4.2. Emotion Encoder

Similar to previous work [27, 28], we extract five emotion-based features from the audio utterances¹ [29].

1. MFCC: Mel-frequency cepstral coefficients model human perception of frequency and loudness. We extract 40 coefficients for this study.
2. Chroma: Chromagram from a waveform represents the spectral energy with pitch profiling.
3. MEL: Mel-scaled power spectrogram is the audio spectrogram with frequencies converted to mel scale. We set the number of mel bands to 128.
4. Spectral Contrast: Spectral contrast is the decibel difference between peaks and valleys in the spectrum.
5. Tonnetz: Tonal centroid features arranges sounds according to pitch relationships into interdependent spatial and temporal structures.

These five features are then concatenated together and used as the emotion representation for the respective utterance.

4.3. Text Encoder

Firstly, we transcribe the spoken audio to text using pretrained wav2vec models which are finetuned to perform ASR on specific languages of interest (sourced from huggingface [30]). We then obtain text embeddings for the transcribed sentence using Sentence-BERT [31]. We apply the temporal mean-pooling operation on top of the generated word embeddings thereby mapping each input to a 768 dimensional vector space.

¹We use Librosa library for emotion feature extraction

4.4. Feature Fusion

The feature representations obtained in the previous step are concatenated together and used as the representation of the audio utterance. However, this results in high dimensional feature representations that impedes the training process and results in over-fitting. To address this, we use Principal Component Analysis (PCA) [32] to reduce the dimensions of the feature representations by removing redundant or non-discriminative dimensions. Since PCA is very sensitive to the ranges of the feature values, we apply z-score normalization prior to dimensionality reduction and select the number of PCA components such that the amount of variance to be explained is 95%. Once these projected features are obtained, they are concatenated to form the final feature vector.

4.5. Classification Module

After extracting the features from audio utterance, we train a binary classifier which uses these representations as input and learns to classify them into *abusive* and *non-abusive* categories. We experiment with the classifier architecture used by [33] and also propose a stack classifier(SC) architecture [34].

ADIMA classifier (AC): AC consists of two fully connected layers(512 \rightarrow 256 \rightarrow 128 \rightarrow 2) with ReLU activation. We train using Adam optimizer with learning rate of 0.001, a dropout of 0.1 and a batch size of 16 for 50 epochs. The output is mapped to the probability of the two output classes (*abusive* and *non-abusive*) using softmax layer.

Stack classifier (SC): SC is an ensemble learning method and achieves higher predictive performance by combining multiple classification models through a meta-classifier. We use five base classifiers, namely, Gaussian process classifier (kernel “1.0 * RBF(1.0)”), multi-layer perceptron classifier (1 layer, 100 neurons), support vector classifier (linear kernel, regularization parameter C=0.025), random forest classifier (10 trees with max-depth 5) and logistic regression classifier. We train a meta-classifier(logistic regression) over these classifiers for making the final predictions using the stacked predictions of base classifiers as inputs.

4.6. Two-Stage Process (TSP)

Besides reporting abuse-detection performances of ADIMA languages using our proposed method, we also compare their performances with the 2-step approach employed in [22]. In the first step, we use a fine-tuned ASR model for transcribing the audio utterances into text. For each language, we select the ASR model which has been trained for that particular language as quality of the transcription is important for the downstream classification task. In the second step, we use a text-based sequence classifier (IndicBERT [35]) for feature extraction from transcriptions followed by classification. TSP is trained with similar hyper-parameters as the ADIMA classifier (AC).

5. Abusive Behaviour and Emotions

Abusive behaviour in natural conversations is often accompanied with elevated emotions as humans are generally angry and agitated while displaying abusive behaviour. To validate this hypothesis, we train our models using only emotions as the input feature and compare them with the audio embeddings extracted using the audio encoder. From Table 2, we observe that both the features perform competitively for this task which shows that there is correlation between emotions and abusive behavior. In

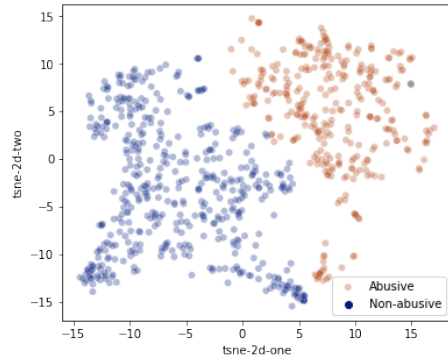


Figure 2: t-SNE visualization of emotion embeddings

Table 1: Comparison of abuse classification results between the ADIMA baselines[33], the proposed MADA and TSP. For ADIMA and MADA, we report the best accuracies. (*) involve XLSR feats and the others CLSRIL feats

Lang	ADIMA		TSP		MADA (Ours)	
	Acc	F1	Acc	F1	Acc	F1
Hi	79.67	79.48	79.0	78.49	84.82*	83.82
Be	81.08*	79.46	76.0	75.06	82.43*	85.58
Pu	82.01	82.01	78.0	77.87	85.01	85.09
Ha	81.15	81.12	68.0	68.26	83.06	81.87
Ka	82.92*	80.15	58.0	56.45	83.47*	79.94
Od	83.29*	82.21	82.0	82.33	82.46	77.78
Bh	76.48*	72.30	73.0	73.03	78.27	84.24
Gu	80.94	76.38	69.0	68.08	82.04	77.62
Ta	80.59*	75.04	82.0	82.94	82.21	81.13
Ma	86.29	83.41	77.0	76.77	85.48	80.27

Figure 2, we show the t-SNE plot of the emotion embeddings for the utterances and observe that the emotion embeddings of abusive and non-abusive words are fairly separable. This also explains the strong performance reported by ADIMA [33] while using features extracted using a CNN based model trained to perform sound classification instead of human speech understanding. ADIMA dataset contains audios collected from over 6k users across 10 languages which shows that this correlation holds for multiple users and languages. However, we acknowledge that ADIMA dataset might have a sampling bias and thus concretely making this claim requires a comprehensive study.

6. Dataset and Evaluation

We experiment with Abuse Detection in Multilingual Audio (ADIMA) [33] dataset. ADIMA is a well balanced multilingual dataset comprising of 10 Indic languages, Hindi (Hi), Bengali (Be), Punjabi (Pu), Haryanvi (Ha), Kannada (Ka), Odia (Od), Bhojpuri (Bh), Gujarati (Gu), Tamil (Ta) and Malayalam (Ma), with 1200 utterances for each language. Each utterance has been annotated by human annotators as *abusive* or *non-abusive* and is 5-60 seconds long in duration. The utterances containing cuss-words, swear-words, and other forms of explicit abuse are labelled as *abusive*. ADIMA dataset represents real-life scenarios because the utterances have been sourced from real-life conversations and often contain *environmental noise*. Moreover, poor recording devices or improper placement of the recording device at the user’s end further contribute to the loss of clarity. Additionally, the multilingual property of the dataset allows for exploration across multiple *languages, accents and dialects*, specifically for resource-impooverished languages.

Table 2: Ablation study on effect of different modalities (audio, emotion and text) and their combination (all) using concatenation on abuse detection performance (Accuracy (%))

	audio	emo	text	all-AC	all-SC
Hi	79.13	78.86	78.86	83.46	84.82
Ha	79.51	78.96	69.39	82.78	83.06
Ta	80.59	81.40	77.35	82.21	82.21
Be	80.81	80.54	79.19	78.91	82.43
Pu	81.74	80.38	77.80	83.65	85.01

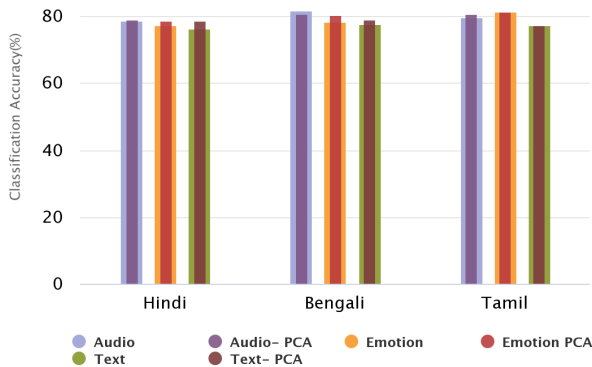


Figure 3: Classification accuracy of MADA and all modalities post dimensionality reduction. We note that for every modality, dimensionality reduction shows performance gains.

7. Results and Discussion

We train our unimodal and multimodal networks using binary cross entropy loss. We use the train/test splits (70:30 for each language) provided by [33] across all the experiments for fair comparison and report accuracy and F1 score on the test set.

7.1. Classification Results

We compare the performance of our proposed method (MADA) with the baselines reported in [33] and the binary classification results obtained using the two stage process (TSP) on the ADIMA dataset. ADIMA experimented with different backbones and feature processing techniques. In this study, we compare our method with the best results reported by ADIMA for each language irrespective of the set-up they used. From Table 1, we observe that our method outperforms the best of ADIMA results for a majority of the languages (except Odiya and Malayalam) with substantial margin demonstrating the importance of leveraging all the modalities. Similarly, we also note that our proposed method improves upon the TSP method significantly for a majority of the languages. Moreover, the performance of TSP degrades substantially for some of the languages which could be a result of inferior quality of the transcriptions. Moreover, since the ASR models used in TSP are pretrained models borrowed from different domains, they may not contain instances of abuse words in their vocabulary. This may result in domain gap resulting in the drop in evaluation accuracy. However, a more detailed study is warranted to establish these justifications conclusively.

Overall, our proposed method (MADA) demonstrates best performance for majority of the languages with an absolute accuracy gain ranging from **0.6-5.2%** (over ADIMA baselines) and **0.2-25.5%** (over TSP). These results highlight the significance of fusing different modalities for this task.

7.2. Ablation Studies

- Modality:** In this section, we study the contributions of different modalities and hyper-parameters for our task using different languages (Hi, Ha, Ta, Be and Pu). For analyzing the significance of modalities, we perform our experiments individually with each modality (audio, emotion and text) and by adding all the different modalities together. From Table 2, we can observe that adding the modalities is beneficial and improves the performance significantly for all the languages. For example, the accuracy of Hindi (Hi) improves from **79.13%** to **84.42%**. Similar trends are also observed for all the other languages.
- Correlation between Emotions and Abuse:** As discussed in previous section, it is interesting to note that *audio* and *emotion* modality have comparable abuse detection performance as shown in Table 2. This observation alludes towards strong correlation between emotions expressed in the audio and abusive behaviour displayed by the speakers. While dealing with emotion based abuse detection, the few false positives that occur may result from cases where the speaker expresses frustration or uses profanities for emphasis or humor, while the false negatives may be caused by missing out on subtler forms of abuse.
- Classifier Architecture:** We also compare the performance of ADIMA classifier (AC) and Stacked classifier (SC) in Table 2 to isolate the contribution of multimodal fusion and classifier architecture. We note that both classifiers achieve competitive results showcasing that the gains can be attributed to the fusion of the modalities and are classifier-agnostic.
- Dimensionality Reduction:** In Figure 3, we analyze the performance of MADA on performing dimensionality reduction before training the classifiers for three languages. We note that for all the modalities, dimensionality reduction helps in improving the performance of MADA. For Hindi, the performance improves by an absolute percentage of **1.14**, **1.35** and **2.44** for audio, emotion and text respectively. Similar patterns are also observed for the other languages as can be seen in Figure 3.

8. Conclusion

Prevention of abusive content is crucial for facilitating safe and healthy interactions. In this work, we explore audio abuse detection from a multimodal perspective in a multilingual setting over 10 Indic languages. We investigate the significance of modelling the audio, its underlying emotions and the semantic information present in its transcribed text to show gains over approaches which only work with the audio modality. Our detailed ablation experiments evaluate the contribution of every modality and show that we achieve substantial success in classifying abusive occurrences when we leverage information from all the modalities together. This therefore highlights the presence of complementary information in the different modalities. We also compare our methods with two-stage process where the first stage involves transcribing the audio into text and second stage comprises of training a text based classifier over the transcriptions. Our unimodal experiments show that using only emotion features is also competitive for this task which highlights that there is correlation between abusive behaviour and emotions.

9. References

- [1] V. S. 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 & Informatics (ICACCI)*, 2015, pp. 2354–2358.
- [2] 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 psychology and psychiatry*, vol. 23, no. 4, pp. 601–613, 2018.
- [3] Z. Gao and S. Yada, "Offensive language detection on video live streaming chat," in *Proceedings of the 28th International Conference on Computational Linguistics*. Spain: ICCL, Dec. 2020, pp. 1936–1940. [Online]. Available: <https://aclanthology.org/2020.coling-main.175>
- [4] C. Alcântara, V. Moreira, and D. Feijo, "Offensive video detection: Dataset and baseline results," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4309–4319.
- [5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, 2017.
- [6] A. Safaya, M. Abdullatif, and D. Yuret, "Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020, pp. 2054–2059.
- [7] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web*, 2016, pp. 145–153.
- [8] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on twitter," *arXiv preprint arXiv:1706.01206*, 2017.
- [9] A. Rana and S. Jha, "Emotion based hate speech detection using multimodal learning," *arXiv preprint arXiv:2202.06218*, 2022.
- [10] 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, 2021.
- [11] A. Gangwar, E. Fidalgo, E. Alegre, and V. González-Castro, "Pornography and child sexual abuse detection in image and video: A comparative evaluation," 2017.
- [12] A. Koufakou, E. W. Pamungkas, V. Basile, V. Patti *et al.*, "Hurtbert: incorporating lexical features with bert for the detection of abusive language," in *Workshop on Online Abuse and Harms*. Association for Computational Linguistics, 2020, pp. 34–43.
- [13] 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.
- [14] M. Jhaveri, D. Ramaiya, and H. S. Chadha, "Toxicity detection for indic multilingual social media content," <https://github.com/Sreyan88/Toxicity-Detection-in-Spoken-Utterances>, 2022.
- [15] B. Wei, J. Li, A. Gupta, H. Umair, A. Vovor, and N. Durzynski, "Offensive language and hate speech detection with deep learning and transfer learning," *arXiv preprint arXiv:2108.03305*, 2021.
- [16] S. Endah, D. Nugraheni, S. Adhy *et al.*, "The automation system censor speech for the indonesian rude swear words based on support vector machine and pitch analysis," in *IOP Conference Series: Materials Science and Engineering*, vol. 190, no. 1. IOP Publishing, 2017, p. 012039.
- [17] S. K. Rahut, R. Sharmin, and R. Tabassum, "Bengali abusive speech classification: A transfer learning approach using vgg-16," in *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*. IEEE, 2020, pp. 1–6.
- [18] F. I. Ablaza and Danganan, "A small vocabulary automatic filipino speech profanity suppression system using hybrid hidden markov model/artificial neural network (hmm/ann) keyword spotting framework," in *2014 (HNICEM)*. IEEE, 2014, pp. 1–5.
- [19] S. Rajamanickam, P. Mishra, H. Yannakoudakis, and E. Shutova, "Joint modelling of emotion and abusive language detection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 4270–4279. [Online]. Available: <https://aclanthology.org/2020.acl-main.394>
- [20] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech emotion recognition using speech feature and word embedding," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2019, pp. 519–523.
- [21] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhaji, "Emotion detection from text and speech: a survey," *Social Network Analysis and Mining*, vol. 8, no. 1, pp. 1–26, 2018.
- [22] S. Ghosh, S. Lepcha, S. Sakshi, and R. R. Shah, "Speech toxicity analysis: A new spoken language processing task," *arXiv preprint arXiv:2110.07592*, 2021.
- [23] 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.
- [24] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," *arXiv preprint arXiv:2105.11084*, 2021.
- [25] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.
- [26] A. Gupta, H. S. Chadha, and Shah, "Clsril-23: Cross lingual speech representations for indic languages," *arXiv preprint arXiv:2107.07402*, 2021.
- [27] R. Arya, D. Pandey, A. Kalia, B. J. Zachariah, I. Sandhu, and D. Abrol, "Speech based emotion recognition using machine learning," in *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*. IEEE, 2021, pp. 613–617.
- [28] "Deep Emotion Recognition," <https://github.com/x4nth055/emotion-recognition-using-speech>.
- [29] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.
- [30] "Huggingface models by Harveen Chadha," <https://huggingface.co/Harveenchadha>, <https://huggingface.co/gvs/wav2vec2-large-xlsr-malayalam>.
- [31] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of Conference on EMNLP*. Association for Computational Linguistics, 2019.
- [32] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.
- [33] V. Gupta, R. Sharon, R. Sawhney, and D. Mukherjee, "Adima: Abuse detection in multilingual audio," *arXiv preprint arXiv:2202.07991*, 2022.
- [34] S. Raschka, "Mlxtend," https://github.com/frank-ceballos/StackingClassifiers_Medium, 2016.
- [35] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhat-tacharyya, M. M. Khapra, and P. Kumar, "IndicNLPsuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages," in *Findings of EMNLP*, 2020.