ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746718

ADIMA: ABUSE DETECTION IN MULTILINGUAL AUDIO

Vikram Gupta, Rini Sharon, Ramit Sawhney, Debdoot Mukherjee

ShareChat, India
{vikramgupta, rinisharon, ramitsawhney, debdoot}@sharechat.co

ABSTRACT

Abusive content detection in spoken text can be addressed by performing Automatic Speech Recognition (ASR) and leveraging advancements in natural language processing. However, ASR models introduce latency and often perform suboptimally for abusive words as they are underrepresented in training corpora and not spoken clearly or completely. Exploration of this problem entirely in the audio domain has largely been limited by the lack of audio datasets. Building on these challenges, we propose **ADIMA**, a novel, linguistically diverse, ethically sourced, expert annotated and well-balanced multilingual abuse detection audio dataset comprising of 11,775 audio samples in 10 Indic languages spanning 65 hours and spoken by 6,446 unique users. Through quantitative experiments across monolingual and cross-lingual zero-shot settings, we take the first step in democratizing audio based content moderation in Indic languages and set forth our dataset to pave future work. Dataset and code are available at: https://github.com/ShareChatAI/Adima

Index Terms— Abusive Content Detection, Multilingual Audio Analysis, Indic Dataset, Crosslingual Audio Analysis

1. INTRODUCTION

Detecting abusive content in online content has gained a lot of attention due to the widespread adoption of social media platforms. Use of profane language, cyber-bullying, racial slur, hate speech etc. are common examples of abusive behaviour demanding robust content moderation algorithms to ensure healthy and safe communication. Majority of the existing work has focused on detecting abusive behaviour in textual data [1–6]. Abusive content detection on images and videos has been accelerated with the contribution of multimedia datasets [7–11]. However, abusive content detection in audio has been underexplored primarily due to the absence of audio datasets. Abuse detection in audio can also be addressed by transcribing audio into text using automatic speech recognition (ASR) followed by textual search over the transcriptions. However, this requires accurate ASR systems which require large amount of expensive training data, especially in multilingual setups. Moreover, accuracy of ASR on abusive words can be low as they are under-represented in the training corpora. Another paradigm is to formulate this as *keyword spotting task* by using a dictionary of audio exemplars of abusive words and then use template matching approach. However, this does not exploit underlying cues and overall context which can be helpful for identifying profanity. Template matching approaches also fail for novel words and continuously updating the dictionary is time-consuming. Moreover, these approaches have high time complexity and require collection of significant number of reference templates that capture the variations in style/accent/dialect and environmental conditions. Abusive words are usually not spoken clearly and completely, further limiting the effectiveness of *keyword spotting*.

To tackle these challenges, we contribute a novel and highly diverse multilingual abuse detection audio dataset - **A**buse **D**etection **I**n **M**ultilingual **A**udio (ADIMA). ADIMA contains 11,775 audio recordings from ShareChat chatrooms with a total duration of 65 hours for 10 Indic languages - Hindi (Hi), Bengali (Be), Punjabi (Pu), Haryanvi (Ha), Kannada (Ka), Odia (Od), Bhojpuri (Bh), Gujarati (Gu), Tamil (Ta) and Malyalam (Ma). The dataset is balanced across the languages and has recordings spoken by 6446 different users making it a highly diverse multilingual and multi-user dataset. The recordings have been extracted from real-life conversations and capture natural and in-the-wild conversations. We also formulate the abuse detection task, where the objective is to classify the audio as *Abusive* or *Non-Abusive*. Since the classifier analyzes the complete audio, it is able to effectively leverage the context and underlying audio properties like pitch, intensity, tone and emotion for robust profanity detection. We setup baselines for monolingual and zero-shot cross-lingual setting for encouraging further research in this direction. ADIMA presents promise in supervised settings such as automatic moderation of live/recorded audio/video content, social media chatrooms etc. to enable safer interactions. In unsupervised settings also, ADIMA can be used for large-scale pretraining of models for Indic languages. Competitive crosslingual performance also showcases the strength of ADIMA to address more languages. Our contributions can be summarized as:

- We release ADIMA, a highly diverse, multilingual, expert annotated audio dataset for abuse detection in 10

Table 1. ADIMA statistics across sample distribution, linguistic diversity and audio duration.

Data Description	Value
# Languages	10
# Total Samples	11,775
# Abusive Samples	5,108
# Non-Abusive Samples	6,667
# Unique Users	6,446
Total Duration	65 hours
Average Duration	20 (±3) seconds
Min/Max Duration	5/58 seconds

Indic languages spanning a total of 65 hours comprising 11,775 total samples.

- We introduce *abuse detection task* and report baseline monolingual results capturing nuances of different languages and architectures.

- Competitive crosslingual and joint training results exhibit the potential of ADIMA for abuse detection in other languages and possibility of having a single unified model for all the languages.

2. RELATED WORK

Abuse detection in textual data under multilingual and monolingual settings has received lot of attention from the community [1–3, 5, 6, 12]. Video datasets for identifying offensive videos and a weakly annotated dataset of videos from YouTube labelled for profanity detection are contributed by [8] and [9] respectively. Datasets to identify specific cases such as pornography and child abuse in images and videos [10] and multimodal hateful memes identification using visual and textual features [7] are also present. A YouTube video dataset to identify racist, sexist and normal videos is also contributed by [11]. While the above mentioned datasets exist for abusive content detection in text, image and video modality, audio modality has been rather under explored. Recently, [13] explored self-attentive networks for toxic language classification. Our dataset ADIMA is an attempt to reduce this gap. Audio classification is a well studied area and has been accelerated with the presence of large scale datasets [8, 14, 15]. Audio classification has been performed using Gaussian Mixture Models, Support Vector Machines over Mel Frequency Cepstrum Coefficients, Convolutional Neural Network (CNN) [14] and Recurrent Neural Networks (RNN) [16]. Finetuning and extracting representations from transformer based models [17] which are trained using unlabelled raw audios has also gained significant interest and we leverage these methods for our task.

3. ADIMA DATASET

3.1. Data Collection

Recordings are collected from *public* audio chatrooms of ShareChat[1]. ShareChat is a leading, Indian social media application supporting over 10 Indic languages with penetration across India. ShareChat's *public* chatrooms are open for anyone to join and informed consent of the users is requested for recording and broadcasting the discussions. The data was collected for a period of 6 months (January-June, 2021) from audio chatrooms pertaining to 10 Indic languages. These chatrooms provide an interactive, audio-only platform for users to speak in regional dialect. To build ADIMA, we sampled audio from conversations which were reported as abusive by the users in the chatroom. Focusing on creating a well-balanced and diverse dataset, we select audio samples across 10 languages in similar proportions from a total of 6,446 users of the ShareChat platform.

3.2. Data Annotation

Specific to each language, an independent set of three annotators per language were employed on a contract basis and fairly compensated to annotate each data sample as abusive or non-abusive. We considered the presence of swear, cuss and abusive words/phrases for annotating an audio as abusive. The abusive words/phrases were catalogued and reviewed to ensure consensus among the three reviewers. On average, the inter-annotator agreement measured by Cohen's Kappa $\kappa = 0.88$ was observed to indicate a high degree of agreement amongst the annotators for each language. The Cohen's Kappa varied from $\kappa = 0.77$ to $\kappa = 1.0$ across different languages, indicating the variation in annotation complexity and annotator diversity across languages for the same task. In the case of disagreements, the final label was selected based on review by a fourth, expert annotator. Further, we removed low quality recordings after which the final dataset comprises 11,775 audio recordings spanning over 65 hours of audio across 10 languages.

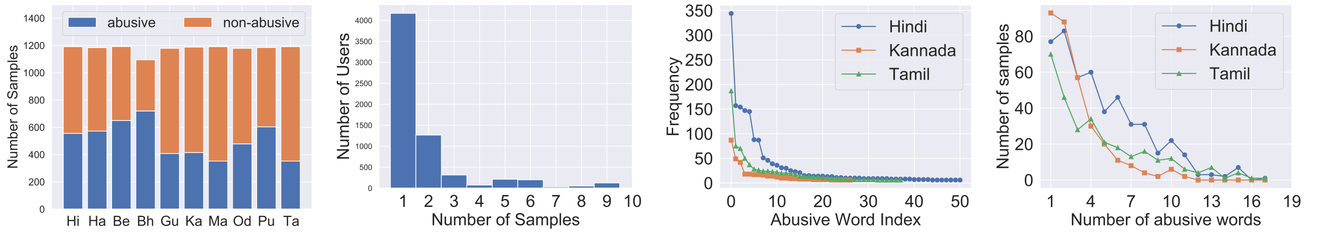
3.3. Dataset Analysis

We briefly summarize the key statistics pertaining to ADIMA in Table 1. The dataset is well balanced (43.38%) with 6,667 non-abusive and 5,108 abusive recordings. We now analyze ADIMA across four different dimensions.

Language Distribution: Figure 1(a) shows the statistics for each language. We note that on average, each language is almost equally represented in ADIMA. Further, while ADIMA is well balanced on average, there exist class imbalance-related variations across languages.

User Distribution: Figure 1(b) depicts the frequency distribution of the number of samples in the dataset spoken by each

[1] https://sharechat.com/



(a) Dataset Distribution (b) User Distribution (c) Unique abusive words (d) Abusive words per recording

Fig. 1. Analyzing ADIMA across multiple dimensions: (a) Number of samples for *Abusive* and *Non-Abusive* categories for all the languages (b) Unique users for all the languages (c) Frequency of abusive words in the dataset (words with more than 5 instances) (d) Frequency distribution of the number of abusive words in each sample. (Best viewed in color)

individual user. On average, there are around 1,500 (23.7%) users that have spoken more than one audio recording in the dataset, indicating a diverse set of users while also presenting a promising opportunity of potential user profiling and leveraging similarity across the samples spoken by the same user for improved performance.

Vocabulary Distribution: In Figure 1(c), we plot the frequency of abusive words which occur more than 5 times in the dataset. We note that Hindi has around 50 dominant abusive words, followed by Tamil and Kannada with 35 and 25 words. Overall, there are 1059, 820 and 690 unique abusive words in Hindi, Tamil and Kannada respectively. The distribution of unique words showcases the diversity of the vocabulary.

Profanity Distribution: In Figure 1 (d), we plot the distribution of instances of abusive words present in the dataset for three languages. Lot of recordings have lesser than 5 abusive words making it challenging to spot the profanity while some recordings are majorly abusive. Each abusive recording contains atleast one profane word.

ADIMA is highly diverse across languages, users, vocabulary and density. The recordings are sampled at 16kHz, mono-channel and range from 5-60 seconds with an average duration of 20 seconds. We randomly split the dataset in 70:30 ratio for each language to form the train and test set.

4. FORMULATION AND METHODOLOGY

4.1. Problem Formulation

We consider the task of classifying audio recording x into $c \in \{abusive, non\text{-}abusive\}$ categories. We extract features using VGG [14] (pretrained over audio dataset) and Wav2Vec2 [17] (pretrained over speech datasets) as backbones. The features are then aggregated across temporal dimension and are passed through a fully connected classifier for classification.

4.2. Feature Representations

VGG: Log-mel spectrograms are extracted from raw audios with the window and hop length as 25ms and 10 ms for short-time fourier transform following [14]. We use 64 mel-spaced frequency bins and transform the magnitude using log to arrive at the log-mel spectogram features for the recordings. Following [14], we train VGG network over AudioSet dataset [18] for extracting features from the spectograms.

Wav2Vec2: Wav2Vec2 models [17] are transformer based models and are trained in a semi-supervised way on unlabelled raw audios. The models can be finetuned for downstream tasks with task-specific labelled data. We explore XLSR-53 model (trained over 53 languages with little overlap with Indic languages), CLSRIL-23 [19] (trained on Indic languages) and Him-4200[2] which was finetuned using 4200 hours of labelled Hindi data over CLSRIL-23. For extracting Wav2Vec2 features, we pass the raw recordings as input.

4.3. Model Architecture

We experiment with Mean-Pool, Max-Pool, and recurrent networks (GRU, LSTM). We represent audio recordings by taking average and maximum of the features across temporal dimension in Mean-Pool and Max-Pool, respectively. The accumulated features are passed into a fully connected classifier $(512 \rightarrow 256 \rightarrow 128 \rightarrow 2)$ with ReLU activation and 0.1 dropout. For RNNs, audio features are processed through single-layer bidirectional Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM). The output of the final time step is used as input to the classifier.

4.4. Training Setup and Evaluation

We train the networks using cross entropy loss with Adam optimizer with learning rate of 0.001 and batch size of 16 for 50 epochs. The recordings are normalized by zero-padding shorter audios. We augment data by applying temporal jittering of 0.1 and mask the temporal and frequency/feature dimension randomly between 0 and 10%. We report Macro

[2] https://huggingface.co/Harveenchadha/vakyansh-wav2vec2-hindi-him-4200

Table 2. Accuracy (Acc), Macro F1 (MaF1), Area under ROC curve (AUC) and Area under Precision-Recall curve (AUCpr) for different architectures and backbones for Hindi.

Backbone	Model	Acc	MaF1	AUC	AUCpr
VGG	Max-Pool	78.05	78.01	0.84	0.85
	Mean-Pool	78.59	78.57	0.85	0.86
	LSTM	77.77	77.71	0.84	0.86
	GRU	78.57	78.56	0.85	0.86
XLSR-53	Max-Pool	76.96	76.90	0.84	0.84
	Mean-Pool	77.51	77.34	0.83	0.85
Him-4200	Max-Pool	79.13	79.03	0.85	0.86
	Mean-Pool	78.86	78.69	0.85	0.85
CLSRIL-23	Max-Pool	**79.67**	**79.48**	0.86	0.86
	Mean-Pool	78.59	78.59	0.86	0.84
	LSTM	70.19	69.53	0.78	0.79
	GRU	75.34	75.23	0.82	0.83

F1 (MaF1), Accuracy (Acc), area under the ROC (AUC) and precision-recall curve (AUCpr) on the test set.

5. RESULTS

5.1. Monolingual Experiments

From Table 2, we note that CLSRIL-23 outperforms other backbones which can be attributed to the pretraining of CLSRIL-23 in Indic languages. Surprisingly, VGG which has been trained for identifying different sounds instead of spoken text demonstrates superior performance than XLSR-53 model for Hindi. However, Him-4200 which has been finetuned for Hindi outperforms XLSR-53 showing the advantage of language specific finetuning. The RNN baselines (GRU and LSTM) do not improve results. We evaluate the baselines on other languages in Table 3 and note that Wav2Vec2 models work better for majority of the languages. This can be attributed to the pretraining of these models on speech data.

Table 3. Accuracy (Acc) and F1 (MaF1) for languages for VGG and XLSR-53 (Mean-Pool) and CLSRIL-23 features with Max-Pool aggregation.

	VGG		XLSR-53		CLSRIL-23	
Lang	Acc	MaF1	Acc	MaF1	Acc	MaF1
Hi	78.59	78.57	77.51	77.34	**79.67**	**79.48**
Be	78.65	77.63	**81.08**	**79.46**	79.73	77.95
Pu	82.01	81.97	82.01	81.99	**82.01**	**82.01**
Ha	**81.15**	**81.12**	80.05	79.91	79.23	79.10
Ka	82.91	80.06	**82.92**	**80.15**	79.67	75.39
Od	81.64	81.46	**83.29**	**82.21**	81.64	80.21
Bh	76.19	71.85	**76.48**	71.10	75.89	**72.30**
Gu	79.56	74.11	79.28	69.21	**80.94**	**76.38**
Ta	79.78	70.77	**80.59**	**75.04**	80.59	73.39
Ma	81.72	77.31	81.70	75.45	**86.29**	**83.41**

5.2. Cross-lingual Experiments

In Table 4, we train zero-shot models on the source language and evaluate the performance on the target language using CLSRIL-23 and Max-Pool. The cross-lingual performance is competitive and even better for some languages showing strong cross learning among languages for this task. We hypothesize that models are able to leverage audio properties like pitch, emotions, intensity etc. for this task instead of relying on the actual words, which is highly encouraging. We also combine the data (All) for all these languages together for training and evaluate on each language separately. We note that combination of all the languages shows improvement for majority of the languages paving path for truly multilingual models for profanity detection.

Table 4. Macro F1 score across languages using CLSRIL-23 model with Max-Pool aggregation. **Bold** represent the best combinations.

source/target	Hi	Be	Pu	Ka	Ta
Hi	**79.5**	77.0	81.5	**79.1**	**74.7**
Be	78.3	**77.9**	82.0	75.2	74.1
Pu	78.5	77.1	82.0	77.6	71.0
Ka	78.3	77.4	82.4	75.4	74.1
Ta	77.1	76.0	**83.4**	77.1	73.3
All	80.7	79.1	83.4	78.4	75.2

6. CONCLUSION AND FUTURE WORK

Detection of abusive content in spoken text is an important problem. Performing ASR followed by a NLP layer for processing the transcription introduces complexity and cost of developing ASR models. In this paper, we contribute a novel and diverse multilingual audio dataset - ADIMA for tackling this problem entirely in the audio domain. The dataset covers 10 Indic languages with 11,775 samples (65 hours) spoken by 6446 unique users and annotated by expert team of reviewers. We also perform comprehensive experiments and report baselines for encouraging further exploration in this direction.

Ethical Considerations: Keeping in mind the sensistive nature of the task, we ensure to mandate certain ethical considerations throughout the course of this research and public release of data. Specifically, ShareChat's *public* chatrooms are open for anyone to join and users' informed consent is sought for recording and broadcasting the discussions. Further, we remove any Personally Identifiable Information (PII) from the dataset and anonymize it. The raw data is kept on secure servers with strong access restrictions to prevent any malicious usage.

7. REFERENCES

[1] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee, “Hatexplain: A benchmark dataset for explainable hate speech detection,” in *AAAI*, 2021.

[2] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan 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.

[3] Òscar Garibo 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 *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019.

[4] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros, “Hate speech dataset from a white supremacy forum,” in *Proc. of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium, Oct. 2018, Association for Computational Linguistics.

[5] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis, “Large scale crowdsourcing and characterization of twitter abusive behavior,” in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[6] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas, “ETHOS: an online hate speech detection dataset,” *CoRR*, 2020.

[7] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine, “The hateful memes challenge: Detecting hate speech in multimodal memes,” *Advances in Neural Information Processing Systems*, 2020.

[8] Cleber Alcântara, Viviane Moreira, and Diego Feijo, “Offensive video detection: dataset and baseline results,” in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020.

[9] Vishal Anand, Ravi Shukla, Ashwani Gupta, and Abhishek Kumar, “Customized video filtering on youtube,” *arXiv preprint:1911.04013*, 2019.

[10] Abhishek Gangwar, Eduardo Fidalgo, Enrique Alegre, and Víctor González-Castro, “Pornography and child sexual abuse detection in image and video: A comparative evaluation,” 2017.

[11] Ching Seh Wu and Unnathi Bhandary, “Detection of hate speech in videos using machine learning,” in *International Conference on Computational Science and Computational Intelligence (CSCI)*, 2020.

[12] Cristina Bosco, Dell’Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio, “Overview of the evalita 2018 hate speech detection task,” in *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*. CEUR, 2018.

[13] Midia Yousefi and Dimitra Emmanouilidou, “Audio-based toxic language classification using self-attentive convolutional neural network,” in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., “Cnn architectures for large-scale audio classification,” in *International conference on acoustics, speech and signal processing*. IEEE, 2017.

[15] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, “Vggsound: A large-scale audio-visual dataset,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[16] Yong Xu, Qiuqiang Kong, Wenwu Wang, and Mark D Plumbley, “Large-scale weakly supervised audio classification using gated convolutional neural network,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018.

[17] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, 2020.

[18] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[19] Anirudh Gupta, Harveen Singh Chadha, Priyanshi Shah, Neeraj Chimmwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan, “Clsril-23: Cross lingual speech representations for indic languages,” *arXiv preprint:2107.07402*, 2021.