# Towards Cross-Lingual Audio Abuse Detection in Low-Resource Settings with Few-Shot Learning

**Aditya Narayan Sankaran, Reza Farahbakhsh, Noel Crespi**
SAMOVAR, Télécom SudParis
Institut Polytechnique de Paris
91120 Palaiseau, France
aditya.sankaran@ip-paris.fr

## Abstract

Online abusive content detection, particularly in low-resource settings and within the audio modality, remains underexplored. We investigate the potential of pre-trained audio representations for detecting abusive language in low-resource languages, in this case, in Indian languages using Few Shot Learning (FSL). Leveraging powerful representations from models such as Wav2Vec and Whisper, we explore cross-lingual abuse detection using the ADIMA dataset with FSL. Our approach integrates these representations within the Model-Agnostic Meta-Learning (MAML) framework to classify abusive language in 10 languages. We experiment with various shot sizes (50-200) evaluating the impact of limited data on performance. Additionally, a feature visualization study was conducted to better understand model behaviour. This study highlights the generalization ability of pre-trained models in low-resource scenarios and offers valuable insights into detecting abusive language in multilingual contexts.

## 1 Introduction

The widespread adoption of social media for everyday communication requires safeguards and moderation to create a safe space for the user and the social community. With audio-based social media platforms like Twitter (now X) Spaces, Clubhouse, Discord, ShareChat etc, moderating offensive language and hate speech has become essential in maintaining a safe space for people online. These platforms host users from diverse linguistic backgrounds, especially in multilingual countries like India. India is home to several languages with more than 30 Million speakers of languages and has experienced a phenomenal increase in the use of online social media services, including Facebook, Twitter (now X), Instagram, LinkedIn, and YouTube, with over 250 million users (and growing) helping them in Social interactions and conversations (Ganguly and Kumaraguru, 2019; Palakodety and KhudaBukhsh, 2020). 76% of Indians spend an estimated 1 hour and 29 minutes daily using social media through smartphones, with adolescents between the ages of 13 and 19 making up 31% of the overall number of people who use social media (Dar and Nagrath, 2022; Srivastava et al., 2019).

Given that a large share of users of social media are teenagers and young adults, there is an absolute need to create a safe online space for them to express their views freely without being exposed to hate-filled and offensive content. In an era where social media and entertainment companies are being scrutinized more carefully than ever before, laxity related to user privacy or community rule violations could prove harmful and costly for social audio platforms. There have been reports of incidents where Clubhouse rooms engaged in racist and anti-Semitist talks[1]. This issue raises concerns about how it can be more difficult to safeguard user privacy and security on Audio Social Media platforms as traditional content moderation techniques utilized for text-based media do not necessarily work well with audio-based platforms.

Abusive language detection and Hate Speech detection in Low Resource Language settings has been a popularly researched topic, especially in the text modality. Transfer Learning-based approaches have shown incredible performance in abuse detection using existing popular models like BERT, RoBERTa, XLM-RoBERTa etc (Mozafari et al.,

[1] https://tinyurl.com/fslanti

*Proceedings of the 31st International Conference on Computational Linguistics*, pages 5558–5569
January 19–24, 2025. ©2025 Association for Computational Linguistics

2019; Ranasinghe and Zampieri, 2021). Works by Mozafari et al. (2022) and Awal et al. (2024) also show the ability of meta-learning-based models to outperform transfer learning-based models in a cross-lingual abuse detection task in Low resource settings. Even abusive content detection on images and videos has been accelerated with the contribution of multimedia datasets and progress in Deep Learning techniques (Gao et al., 2020; Alcântara et al., 2020). Abuse detection in audio is relatively unexplored given the limited availability of such datasets to work with and the complex task of collecting and annotating a large set of audio samples, that too in a variety of languages and accents, especially in highly multilingual countries like India where there are 1369 rationalised languages and more than 500 dialects, with 22 scheduled languages having over 1.17 billion speakers (Khanuja et al., 2021; Sengupta, 2018).

A simple approach for abuse detection is to employ an Automatic Speech Recognition (ASR) model to transcribe the audio to text and then work with them using existing NLP techniques with models like Whisper and Wav2Vec boasting low Word-Error Rates for the transcription task. This is what Ghosh et al. (2021) did by crawling through popular voice datasets for abusive data called DeToxy and performed a Two-Step approach by transcribing audio to text and using BERT for downstream classification. A major problem with these approaches is that they tend to miss out on abusive words since they are usually not spoken clearly and completely, thereby limiting the ability to spot abusive keywords (Gupta et al., 2022).

Recent advancements in Pre-Trained Audio Representations, such as those demonstrated by Shor et al. (2022) and Saeed et al. (2021), show promising results in various audio abuse detection-related tasks, particularly embeddings pre-trained on large datasets like Audioset have been successfully transferred to multiple classification tasks, including various audio pattern recognition tasks (Kong et al., 2020). Building on the success of Meta-Learning techniques, such as Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017), Prototypical Networks (Snell et al., 2017), and Contrastive Learning (Yang et al., 2022), our work applies MAML in a low-resource setting, leveraging its demonstrated efficacy in tasks such as document classification (van der Heijden et al., 2021), speech recognition (Singh et al., 2022), and abstract summarization (Huh and Ko, 2022).

Thus, we present the contributions[2] of our work enumerated below:

1. We propose a MAML-based few-shot cross-lingual audio abuse classification methodology, leveraging pre-trained audio representations from Whisper (Radford et al., 2022) and Wav2Vec (Baevski et al., 2020). We also assess the effectiveness of these pre-trained features under two feature normalization strategies: L2 normalization and Temporal Mean.

2. Our method is evaluated on the ADIMA dataset (Gupta et al., 2022), with Whisper achieving top accuracy scores ranging from 78.98% to 85.22% in the 100-shot setting, using L2-Norm feature normalization.

3. We provide a visual analysis of pre-trained audio features from the best-performing normalization setting, examining how language similarity can enhance cross-lingual abuse detection, especially in Low-Resource Languages. This study offers valuable insights into optimal strategies for audio abuse detection and identifies potential directions for future research.

## 2 Related Works

### 2.1 Audio Abuse Detection

The introduction of DeToxy (Ghosh et al., 2021), a large-scale multi-modal dataset was a significant advancement in the field of audio toxicity detection, particularly in the context of spoken utterances. DeToxy improved text-based methods' performance and lessened keyword bias, paving the way for more robust audio abuse detection. Building on this work, Facebook AI extended on DeToxy by creating MuTox (Costa-jussà et al., 2024), one of the first large-scale multilingual audio datasets for toxicity detection. MuTox includes 20,000 audio utterances in English and Spanish and 4,000 utterances across 19 other languages. However, it is worth noting that this dataset only includes three Indian languages: Bengali, Hindi, and Urdu.

The first Bengali Audio Abuse dataset was introduced by (Rahut et al., 2020), where Transfer Learning was applied to extract features from 960 voice recordings of native Bengali speakers. Following this, ADIMA (Gupta et al., 2022) intro-

[2] Codebase: `https://github.com/callmesanfornow/fsl-audio-abuse.git`

duced the first Indian audio abuse dataset, comprising abusive audio clips in 10 Indian languages. This work sought to democratize audio-based content moderation in Indic languages through quantitative experiments conducted in both monolingual and cross-lingual zero-shot settings. A follow-up study by (Sharon et al., 2022) explored a multi-modal approach to improve abuse detection in multilingual settings. Additionally, (Spiesberger et al., 2023) demonstrated that abuse detection could be effectively performed using only acoustic and prosodic features on the ADIMA dataset, thereby avoiding the need for transcriptions.

Sharon and Mukherjee (2024) leverage well-established Natural Language Processing techniques for abuse detection and introduce a cascaded model that combines an ASR system with textual keyword spotting and compares it with an end-to-end model utilizing audio-level feature embeddings and neural classifiers. A two-step process for abuse detection by first transcribing spoken audio into text using ASR systems followed by natural language processing-based methods was explored by (Sharon et al., 2022). While this approach captured semantic information, it missed important audio cues such as pitch, volume, tone, and emotions, which are crucial in detecting abusive behaviour, as abusive speech often involves anger, agitation, or loudness (Rana and Jha, 2022; Plaza-Del-Arco et al., 2021).

### 2.2 Few-Shot Learning and Meta-Learning

Few-shot learning (FSL) is particularly significant in low-resource settings where data scarcity is a major challenge. Model-agnostic Meta-learning (MAML) has emerged as a powerful method within this domain. Singh et al. (2022) proposed a MAML-based Low Resource ASR methodology using a multi-step loss (MSL) approach, which significantly improved the stability and accuracy of low-resource speech recognition systems compared to the traditional MAML approach. Gu et al. (2018) demonstrated the effectiveness of MAML in low-resource scenarios for Neural Machine Translation, significantly outperforming multilingual transfer learning methods on the Romanian-English WMT'16 dataset with only limited translated words. Xia et al. (2021) proposed MetaXL, a method that effectively transforms representations from auxiliary languages to target languages, enhancing cross-lingual learning in tasks such as sentiment analysis and named entity recognition.

### 2.3 Automatic Speech Recognition

Meta-learning approaches have also shown promise in automatic speech recognition. Hsu et al. (2019) proposed MetaASR, which significantly outperformed the state-of-the-art multitask pretraining approach across various target languages with different combinations of pretraining languages. Furthermore, (Conneau et al., 2020) introduced XLSR, which learns cross-lingual speech representations by pretraining a single model from raw speech waveforms in multiple languages, enabling a single multilingual speech recognition model that is competitive with strong individual models.

Hou et al. (2021a) explored the combination of adapter modules with meta-learning algorithms to achieve high ASR performance in low-resource settings while improving parameter efficiency. In another study, (Hou et al., 2021b) proposed SimAdapter, a novel algorithm for learning knowledge from adapters for cross-lingual speech adaptation, showing that these approaches can be integrated to achieve significant performance improvements, including a relative Word Error Rate (WER) reduction of up to 3.55%.

While these works have highlighted the progress of various tasks like Abuse Detection, Machine Translation, Automatic Speech Recognition etc using Meta-Learning techniques in low-resource settings, there is an avenue for using Meta-Learning for Audio Abuse detection. This work contributes to the research gap and serves as a foundation for Audio Abuse Detection in Low Resource Languages, especially in Indian Languages, employing Few-Shot Learning and Pre-Trained Audio Representations.

## 3 Methodologies

Representations that are effective across general audio tasks, capture multiple robust features of the input sound thereby using these learned embeddings for classification tasks like Music Information Retrieval, Industrial Sound Analysis, etc (Niizumi et al., 2022; Grollmisch et al., 2021), we propose our method. Building on these studies, we employed a Model-Agnostic Meta-Learning (MAML) approach (Finn et al., 2017) to develop a few-shot classifier for cross-lingual audio abuse detection using pre-trained audio features. This approach leverages the adaptability of meta-learning to handle the complexities of low-resource and multilingual settings, ensuring improved performance in audio-based abusive content moderation. Additionally, we also perform a feature study of abusive language in the 10 languages with the best-performing normalised feature set that has the best classification accuracy.

### 3.1 Pre-Trained Audio Feature Extractions

Features from Pre-Trained Audio Models were used for few-shot classification using MAML. We employed the CLSRIL-23 variant of Wav2Vec (Gupta et al., 2021), which is a self-supervised learning-based audio pre-trained model that learns cross-lingual speech representations from raw audio across 23 Indic languages. It is built on top of Wav2Vec 2.0 (Baevski et al., 2020) and solved by training a contrastive task over masked latent speech representations and jointly learning the quantization of latents shared across all languages. We also used Whisper (Radford et al., 2022), a pre-trained model for automatic speech recognition (ASR) and speech translation. Trained on 680k hours of labelled data, Whisper models demonstrated a strong ability to generalise to many datasets and domains without the need for fine-tuning.

These extracted features were then normalised using the two methods:

**Temporal Mean:** This process involves computing the mean of the vectors along the temporal dimension for each tensor.

$$V_i[j] = \frac{1}{x_i} \sum_{t=1}^{x_i} T_i[1, t, j] \tag{1}$$

where $x_i$ is the temporal length for the $i-th$ tensor and $j$ ranges from 1 to the size of the feature dimension.

**L2-Norm:** This process involves computing the Euclidean Norm of the features for each tensor along the temporal dimension and then normalizing the features using the norm. After normalization, the mean vector is computed similarly to the previous function.

$$f_{i,t}^{\text{norm}} = \frac{T_i[1, t, :]}{\sqrt{\sum_{j=1}^{768} (T_i[1, t, j])^2}} \tag{2}$$

The $j-th$ element of the mean vector $V_i$ is:

$$V_i[j] = \frac{1}{x_i} \sum_{t=1}^{x_i} f_{i,t}^{\text{norm}}[j] \tag{3}$$

where $x_i$ is the temporal length for the $i-th$ tensor and $j$ ranges from 1 to the size of the feature dimension.

### 3.2 Model Agnostic Meta-Learning (MAML)

Meta Agnostic Meta Learning is a technique introduced by (Finn et al., 2017). The goal of few-shot meta-learning is to train a model that can quickly adapt to a new task using only a few data points and training iterations. For this, the model or learner is trained during a meta-learning phase on a set of tasks, such that the trained model can quickly adapt to new tasks using only a small number of examples or trials. In effect, the meta-learning problem treats entire tasks as training examples.

For our few-shot learning setup, we perform stratified sampling of $k$ samples per class for each language. Formally, let $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ denote the dataset of audio samples for language $l$, where $x_i$ represents the feature vector and $y_i$ the class label (abusive or non-abusive). For a given $k$-shot scenario, we construct a support set $\mathcal{S}_l \subseteq \mathcal{D}_l$ such that:

$$\mathcal{S}_l = \{(x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2}), \ldots, (x_{i_k}, y_{i_k})\} \tag{4}$$

where $y_{i_j} \in \{\text{abusive}, \text{non-abusive}\}$ and each class is represented equally within the support set. Specifically, for a $k$-shot learning task with $k = 2$, $\mathcal{S}_l$ contains one abusive and one non-abusive sample per language.

Assuming there are $L$ languages, the total number of samples for the $k$-shot scenario is:

$$|\mathcal{S}| = k \times L \tag{5}$$

where $\mathcal{S}$ represents the combined support sets across all languages. For instance, in a 50-shot scenario across 10 languages, this results in $50 \times 10 = 500$ samples.

### 3.3 Cross-Lingual Training and Testing

The few-shot model is trained using a cross-lingual approach, which is key to ensuring the model's ability to generalise across different languages. During training, the model is exposed to data from all $L$ languages, so that learning from the pre-trained representations captures the nuances of abusive and non-abusive speech across different contexts and languages. This cross-lingual training strategy enables the model to recognize similarities in abusive language across Indian languages, which is essential for achieving strong performance in low-resource scenarios where data for individual languages is limited. By leveraging a cross-lingual setting, the model can better generalize and identify abusive patterns even when specific language data

| Language | Abusive | | Non-Abusive | | Total |
|---|---|---|---|---|---|
| | Train | Test | Train | Test | |
| Bengali | 394 | 148 | 428 | 222 | 1192 |
| Bhojpuri | 253 | 122 | 506 | 214 | 1095 |
| Gujarati | 516 | 255 | 301 | 107 | 1179 |
| Haryanvi | 419 | 193 | 399 | 173 | 1184 |
| Hindi | 449 | 186 | 373 | 183 | 1191 |
| Kannada | 530 | 243 | 289 | 126 | 1188 |
| Malayalam | 582 | 257 | 237 | 115 | 1191 |
| Odia | 491 | 209 | 323 | 156 | 1179 |
| Punjabi | 405 | 176 | 413 | 191 | 1185 |
| Tamil | 572 | 267 | 248 | 104 | 1191 |
| **Total** | 4611 | 2056 | 3517 | 1591 | 11775 |

Table 1: ADIMA Dataset distribution across languages and classes. Train and Test being the ones provided by authors.

is sparse. The testing phase involves evaluating the model's performance on individual languages and assessing its ability to adapt and accurately classify audio samples in a language-specific context after having been trained on a diverse set of languages. This setup simulates real-world conditions where a model, trained on data from multiple languages, must be capable of quickly adapting to and performing well on new languages with limited labelled examples. The effectiveness of this approach is demonstrated by the model's performance across various shot settings, which we analyze in detail in subsequent sections.

### 3.4 Feature Study

Observing the best-performing normalised feature set, this study aims to understand the specific characteristics of the features that contributed to improved classification accuracy in our few-shot learning setup. For the study, L2-Norm feature normalisation with Whisper was selected based on its superior performance in its accuracy scores as presented in Figure 2b. We performed a feature study by plotting the tSNE projection of the features to 2 dimensions and performing a visual analysis and an outlier analysis.

## 4 Experiments

### 4.1 Dataset

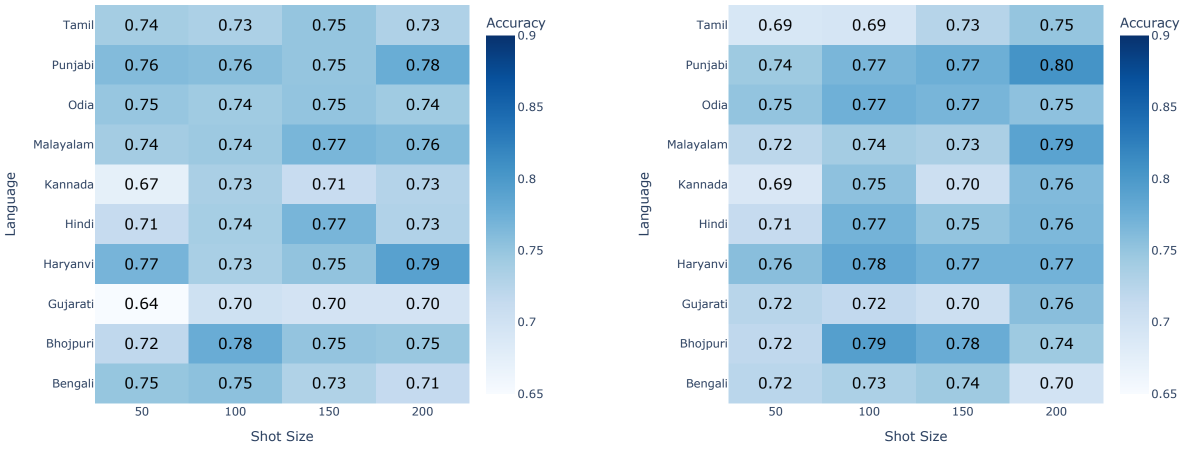
To carry out our study, we employed the ADIMA dataset (Gupta et al., 2022) by ShareChat. It contains 11,775 audio clips, sourced from real-life conversations, in 10 Indian Languages annotated for a binary classification task. It is an evenly distributed dataset comprised of 5,108 abusive and 6,667 non-abusive samples from 6,446 unique users. Table 1 presents the distribution of samples across the classification category and across 10 languages where the average and median number of samples are 1177.5 and 1186.5 respectively. We can observe that Bhojpuri has the least number of samples, but the maximum number of abusive samples, whereas Bengali has the highest number of audio samples. Another observation is the lesser amount of abusive samples Malayalam and Tamil have compared to non-abusive samples.

### 4.2 Feature Extraction

The code for extracting features for all the audio clips was with PyTorch (Paszke et al., 2019) and HuggingFace libraries. In this task, we compared two pre-trained audio models: Whisper (Radford et al., 2022) and Wav2Vec (Baevski et al., 2020), specifically, the `whisper-large` variant for Whisper and the `CLSRIL-23` variant of Wav2Vec (Gupta et al., 2021) were used for extracting pre-trained features. Firstly, embeddings were extracted by passing raw audio files through these models. The embeddings were then feature-normalised to generate decision-level features (Wang et al., 2023) for our task in two ways: Temporal Mean (equation 1) and L2-Norm (equation 2) to generate two different feature sets.

### 4.3 Few-Shot Experimental Setup

We employed the Model-Agnostic Meta-Learning (MAML) algorithm (Finn et al., 2017) to address the challenge of few-shot cross-lingual audio abuse detection. The few-shot learning methodology is particularly suited to scenarios where only a limited number of labelled examples are available for each language, ensuring efficient and effective learning from small datasets. We utilized stratified sampling (Mitchell, 1996) to sample audio clips for the few-shot task ensuring that the proportion of abusive and non-abusive samples remained balanced across languages. As discussed in Section 3.2, for a shot size "$k$", $k$-samples are chosen in the 10 languages, thereby making the number of samples for cross-lingual training 10 x $k$-samples (refer equation 5). To evaluate the model's performance across different few-shot settings, we conducted experiments with varying shot sizes: 50, 100, 150, and 200. These shot sizes were selected to investigate the impact of sample size on model performance, particularly in scenarios where the number of available samples is less than half of the average number of audio clips per language, which is approximately



(a) Temporal Mean Wav2Vec

(b) Temporal Mean Whisper

Figure 1: **Temporal Mean**: Few Shot Accuracies in 50, 100, 150 and 200 shot cases

1177.5 (refer to Table 1). By exploring these shot sizes, we aimed to understand how well the model could generalize to unseen examples with limited training data, a key concern in real-world applications where extensive labelled data is often unavailable. The data was split into training and testing sets based on the splits provided by (Gupta et al., 2022).

### 4.4 Model Architecture and Training

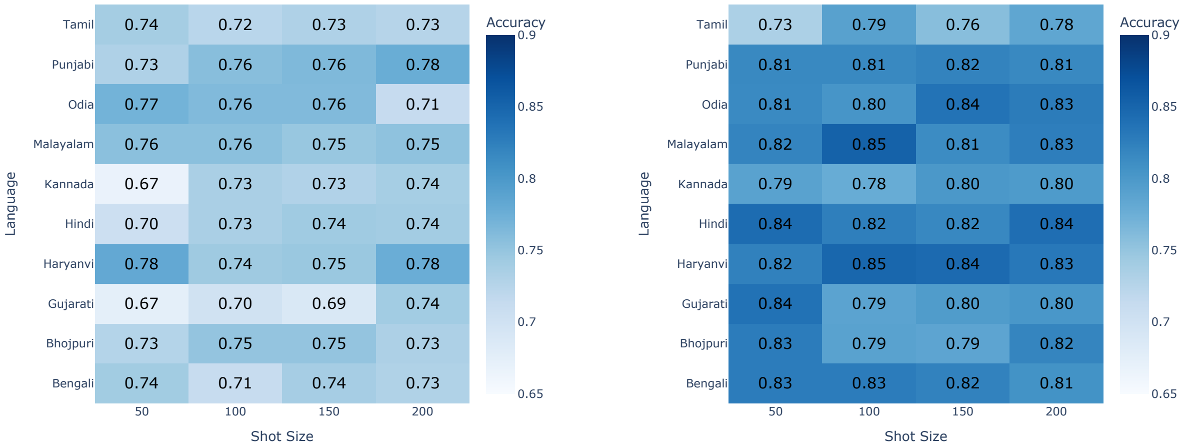
The learner model utilized in our experiments is an Artificial Neural Network (ANN) consisting of three fully connected layers. The network architecture was designed as follows: an input layer with a size corresponding to the dimension of the extracted feature vectors (1024 for Whisper and 768 for Wav2Vec), followed by hidden layers with sizes 256 and 128 respectively, and a final output layer with size 2, with leaky ReLu for non-linearity, corresponding to the binary classification task. The output layer utilized a softmax activation function to convert the raw logits into probabilities for each class, determining whether an audio clip was abusive or non-abusive. Training of the learner model was done using the Adam optimizer (Kingma and Ba, 2017). We employed a task-specific learning rate and a meta-learning rate of 0.001, both of which were managed by a linear learning rate scheduler with default parameters as provided by the PyTorch optimizer library. The model was trained with a batch size of 128 and for a total of 150 epochs, based on repeated testing, which provided a good balance between training time and model performance.

## 5 Results

### 5.1 Classification Results

We present accuracy scores from both feature settings in 4 shot settings [50, 100, 150, 200] as heatmaps in Figures 1 and 2. More Detailed results with macro-F1 scores are provided in the Appendix A. An aggregate macro-f1 score table with the baseline from the Dataset paper (Gupta et al., 2022) is also presented in the Appendix in Table 4.

It is evident that Whisper with the L2-Norm feature normalisation has consistently better scores across languages, with no top accuracy scores in the 200-shot scenario and most of the best-performing accuracy scores are in the 50 and 100-shot settings. Comparing normalisation settings, we can observe that L2-Norm has much better classification performance compared to the temporal mean normalisation setting. For most languages, L2-Norm offers higher accuracy and macro-F1 scores compared to Temporal Mean. For Bhojpuri with Whisper, L2-Norm gives significantly better accuracy (82.75% at 50 shots) compared to Temporal Mean (79.17% at 100 shots). Some other results that are evident are the not-so-consistent performance of Tamil and Kannada, especially Tamil, with its highest Accuracy scores being 74.93% in the Temporal Mean Normalisation setting of Wav2Vec and 78.98% in the L2-Norm Mean Normalisation of Whisper. Across both models and normalization methods, F1 scores generally tracked closely with



(a) L2-Norm Wav2Vec

(b) L2-Norm Whisper

Figure 2: **L2-Norm**: Few Shot Accuracies in 50, 100, 150 and 200 shot cases

accuracy. Languages like Haryanvi, Punjabi, and Odia generally perform better than other languages across both models and normalization strategies. For Whisper with L2-Norm, Haryanvi has strong accuracy (84.7% at 100 shots), and similarly for Punjabi and Odia. Gujarati, Kannada, and Tamil show lower accuracy and F1 scores overall compared to others, regardless of the model or normalization method. An interesting observation is the abuse detection accuracy score of Malayalam with 85.22% accuracy in the L2-Norm feature normalisation setting with Whisper features. Given that Tamil, Kannada and Malayalam are of the Dravidian Language family (Srivatsa and Kulkarni, 2017), it is interesting to observe Malayalam's superior accuracy scores compared to its other Language family counterparts in the cross-lingual setting.

### 5.2 Pre-Trained Feature Study of Abusive Language

Pre-Trained audio representations offer a powerful perspective for studying abusive language, especially in Low-Resource contexts where labelled data can be scarce or imbalanced. Models like Whisper, employing vast amounts of training data to learn robust audio feature representations enable transfers to downstream tasks like abuse detection. This approach allows us to bypass the need for large annotated datasets and explore the underlying audio characteristics that distinguish abusive language across diverse linguistic groups. To understand the improved performance of L2-norm feature normalization in conjunction with Whisper's ability to perform cross-lingual abuse detection, as demonstrated by its superior performance in Section 5.1, we visualized the extracted features by plotting them in 2D with t-SNE and observed the language clustering by performing a visual study of the features.

As shown in Figure 3, three distinct clusters emerge for the Dravidian languages (Kannada, Malayalam, and Tamil) present in the ADIMA dataset. These clusters are separated from others, with Tamil and Malayalam appearing closer together, which is expected given their linguistic similarities. Malayalam shares significant grammatical and literary ties with Tamil, particularly from the historical period spanning late Old Tamil to early Middle Tamil (G., 2022; Menon, 1990). In contrast, the Indo-Aryan language family (Jain and Cardona, 2007) forms a denser and more overlapping cluster, encompassing languages such as Bengali, Bhojpuri, Gujarati, Haryanvi, Hindi, Odia, and Punjabi. Within this cluster, languages like Bengali and Odia appear grouped, while Hindi and Punjabi show a similar pattern clustered close together. The distinct clusters observed in the t-SNE plot suggest that Whisper's audio features capture language-specific patterns effectively in the L2-Norm feature normalization. This finding is promising for cross-lingual classification tasks, as it indicates the preservation of key phonetic characteristics across languages. For instance, Tamil and Malayalam form tighter clusters, reflecting their clearer phonetic or acoustic features, which can aid classification. However, languages such as Bho-

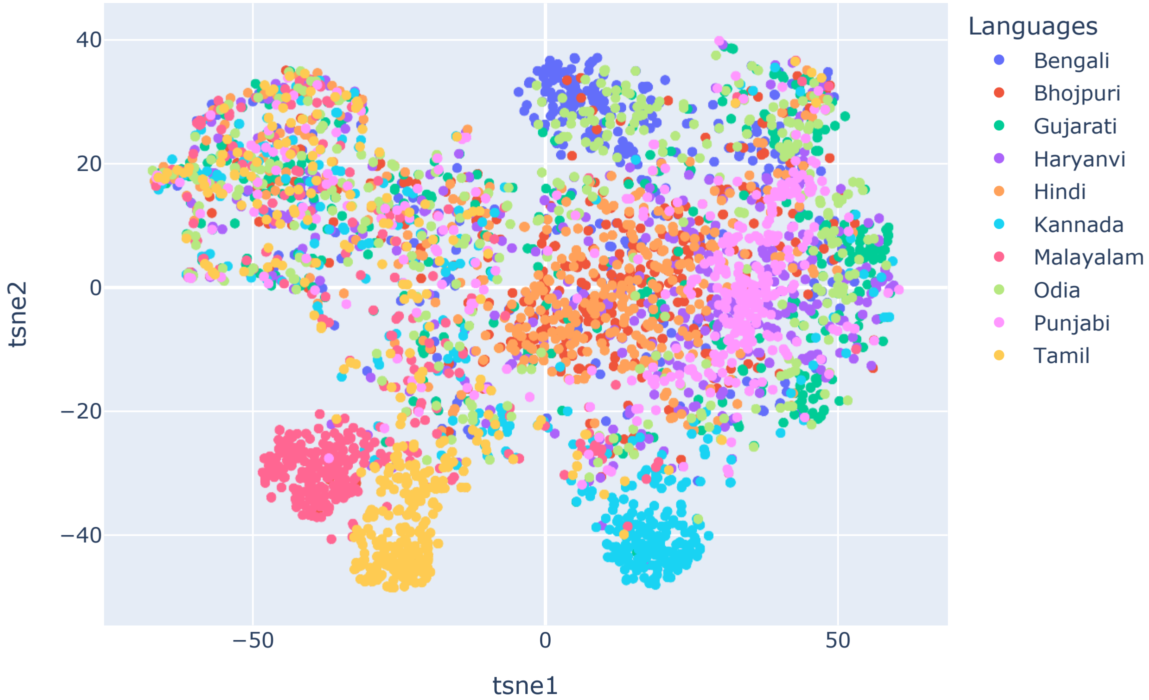


Figure 3: tSNE plot of L2-Norm Feature normalisation of Whisper Features extracted from the ADIMA dataset

jpuri or Haryanvi exhibit less defined boundaries, likely due to limited data and findings of them being dialects of Hindi (Sinha et al., 2014), which results in noisier feature representations. These overlaps help support the study that Hindi-like languages are similar acoustically as well and provide insight that maybe language does play a part in cross-lingual audio abuse detection.

## 6 Conclusion

With the widespread adoption of social media for everyday communication, the need for effective content moderation has become critical to keeping malicious actors in check. While there has been extensive research on text-based moderation, safeguarding against abusive content in audio remains underexplored, particularly for low-resource languages. In this work, we propose a few-shot cross-lingual audio abuse detection method in low-resource languages, specifically focusing on Indian languages by employing the Model-Agnostic Meta-Learning (MAML) framework (Finn et al., 2017). Our method addresses the challenge of detecting abusive audio content with limited training samples per language, a key issue in low-resource settings by benchmarking our approach on the ADIMA dataset (Gupta et al., 2022), which provides binary-labeled audio clips for abuse detection in 10 Indian Languages, leveraging pre-trained audio representations from Wav2Vec and Whisper.

To provide a deeper understanding to enhance the performance of the few-shot classification, we investigated the impact of various feature normalization techniques, such as Temporal Mean Normalization and L2-Norm, applied to the extracted features from the pre-trained audio models and we also conducted a comparative study across these normalization methods and pre-trained models. To better understand the effect of feature normalization and to identify the best-performing model, we conduct a visual analysis of the learned audio representations by plotting the features using a 2D t-SNE plot. We were able to observe language similarities and have discussed why L2-Norm with Whisper Features performed well compared to other feature normalisation techniques and Pre-Trained Audio Representations (Section 5.2).

Our research contributes to the ongoing efforts to combat abusive content on social media platforms, particularly in the audio domain, by providing insights into meta-learning, feature normalization, and performance in low-resource language settings.

## Limitations

While this work explores cross-lingual few-shot audio abuse detection in the 10 languages ADIMA provides, we believe this methodology can also be expanded to other low-resource languages. Further research will be needed to assess efficiency. Exploring other Meta Learning Algorithms like ProtoMAML (Triantafillou et al., 2020) and Contrastive Learning (Saeed et al., 2021) will also contribute to addressing these issues. While this is a cross-lingual abuse detection task, there are avenues for Mono-Lingual Experiments too for more specific languages and also with other languages for cross-lingual tasks. Whisper and Wav2Vec have been the only models that have been used but future works will involve exploring other pre-trained audio models like SeamlessM4T (Communication, 2023) and different feature normalisation such as other L-N Normalisation, Weighted Averaging (Phukan et al., 2024) and more.

Since this work deals with Low Resource Languages in the Indian context, an important limitation is the absence of training data in other Indian Languages since Languages like Telugu and Marathi have been missed out, which are other major spoken languages with about 71 million people who speak Marathi as their native tongue. (Garje et al., 2016) and Telugu with 82.7 million native speakers (Jaswanth et al., 2022). Training data in these languages will add diversity and more languages to detect, improving variety and diversity, thereby being inclusive of all languages. Given the scarcity, we would also like to see the creation of curated datasets for offensive speech detection in other Low-Resource Languages, including ones from the Global South, in the audio modality.

## Ethics Statement

This study does not involve any personal or public data pointing to an individual or a group of individuals and thus does not break any ethical guidelines.

## References

Cleber Alcântara, Viviane Moreira, and Diego Feijo. 2020. Offensive video detection: Dataset and baseline results. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4309–4319, Marseille, France. European Language Resources Association.

Md Rabiul Awal, Roy Ka-Wei Lee, Eshaan Tanwar, Tanmay Garg, and Tanmoy Chakraborty. 2024. Model-agnostic meta-learning for multilingual hate speech detection.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Seamless Communication. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdel rahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition.

Marta Costa-jussà, Mariano Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alexandre Mourachko, Christophe Ropers, and Carleigh Wood. 2024. MuTox: Universal MUltilingual audio-based TOXicity dataset and zero-shot detector. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5725–5734, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

S. Dar and Dolly Nagrath. 2022. The impact that social media has had on today's generation of indian youth: An analytical study.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *Preprint*, arXiv:1703.03400.

S. G. 2022. Family relations in the moral values expressed by dravidian literature.

Niloy Ganguly and P. Kumaraguru. 2019. The positive and negative effects of social media in india.

Zhiwei Gao, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. Offensive language detection on video live streaming chat. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1936–1940, Barcelona, Spain (Online). International Committee on Computational Linguistics.

G. V. Garje, A. Bansode, Suyog Gandhi, and Adita Kulkarni. 2016. Marathi to english sentence translator for simple assertive and interrogative sentences.

Sreyan Ghosh, Samden Lepcha, Sahni Sakshi, Rajiv Ratn Shah, and Srinivasan Umesh. 2021. Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances.

Sascha Grollmisch, Estefanía Cano, Christian Kehling, and Michael Taenzer. 2021. Analyzing the potential of pre-trained embeddings for audio classification tasks. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 790–794. IEEE.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and V. Li. 2018. Meta-learning for low-resource neural machine translation.

Anirudh Gupta, Harveen Singh Chadha, Priyanshi Shah, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. 2021. Clsril-23: Cross lingual speech representations for indic languages.

V. Gupta, R. Sharon, R. Sawhney, and D. Mukherjee. 2022. Adima: Abuse detection in multilingual audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6172–6176. IEEE.

Wenxin Hou, Yidong Wang, Shengzhou Gao, and Takahiro Shinozaki. 2021a. Meta-adapter: Efficient cross-lingual adaptation with meta-learning.

Wenxin Hou, Hanlin Zhu, Yidong Wang, Jindong Wang, Tao Qin, Renjun Xu, and Takahiro Shinozaki. 2021b. Exploiting adapters for cross-lingual low-resource speech recognition.

Jui-Yang Hsu, Yuan-Jui Chen, and Hung yi Lee. 2019. Meta learning for end-to-end low-resource speech recognition.

Taehun Huh and Youngjoong Ko. 2022. Lightweight meta-learning for low-resource abstractive summarization.

Danesh Jain and George Cardona. 2007. The indo-aryan languages.

M. Jaswanth, N. V. L. Narayana, Sreedharreddy Rahul, Susmitha Vekkot, and Sreedharreddy Rahul. 2022. A comparative study of feature modelling methods for telugu language identification.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, D. Margam, Pooja Aggarwal, R. Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vishnu Subramanian, and P. Talukdar. 2021. Muril: Multilingual representations for indian languages.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *Preprint*, arXiv:1912.10211.

A. Menon. 1990. Some observations on the sub-group tamil-malayalam: differential realizations of the cluster *nt.

Don P Mitchell. 1996. Consequences of stratified sampling in graphics. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 277–280.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *"Springer"*.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2022. Cross-lingual few-shot hate speech and offensive language detection using meta learning.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. 2022. Byol for audio: Exploring pre-trained general-purpose audio representations.

Shriphani Palakodety and Ashiqur KhudaBukhsh. 2020. Annotation efficient language identification from weak labels. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 181–192, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Orchid Chetia Phukan, Yashasvi Chaurasia, Arun Balaji Buduru, and Rajesh Sharma. 2024. Collab: A collaborative approach for multilingual abuse detection.

Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, L Alfonso Ureña-López, and María Teresa Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Shantanu Kumar Rahut, Riffat Sharmin, and Ridma Tabassum. 2020. Bengali abusive speech classification: A transfer learning approach using vgg-16. In *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, pages 1–6.

Aneri Rana and Sonali Jha. 2022. Emotion based hate speech detection using multimodal learning. *Preprint*, arXiv:2202.06218.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual offensive language identification for low-resource languages.

Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879. IEEE.

S. Sengupta. 2018. Vision assessment in regional indian languages.

R. Sharon, H. Shah, D. Mukherjee, and V. Gupta. 2022. Multilingual and multimodal abuse detection.

Rini Sharon and Debdoot Mukherjee. 2024. Study of abuse detection in continuous speech for indian languages. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11756–11760.

Joel Shor, Aren Jansen, Wei Han, Daniel Park, and Yu Zhang. 2022. Universal paralinguistic speech representations using self-supervised conformers. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Satwinder Singh, Ruili Wang, and Feng Hou. 2022. Improved meta learning for low resource speech recognition.

S. Sinha, Aruna Jain, and S. Agrawal. 2014. Speech processing for hindi dialect recognition.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. *Preprint*, arXiv:1703.05175.

Anika A. Spiesberger, Andreas Triantafyllopoulos, Iosif Tsangko, and Björn W. Schuller. 2023. Abusive Speech Detection in Indic Languages Using Acoustic Features. In *Proc. INTERSPEECH 2023*, pages 2683–2687.

K. Srivastava, S. Chaudhury, J. Prakash, and Sana Dhamija. 2019. Social media and mental health challenges.

B. Srivatsa and Annarao Kulkarni. 2017. The curious case of kannada 'maadu'.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. 2020. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*.

Niels van der Heijden, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2021. Multilingual and cross-lingual document classification: A meta-learning approach. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1966–1976, Online. Association for Computational Linguistics.

Yuanyuan Wang, Yu Gu, Yifei Yin, Yingping Han, He Zhang, Shuang Wang, Chenyu Li, and Dou Quan. 2023. Multimodal transformer augmented fusion for speech emotion recognition.

Mengzhou Xia, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Neubig, and Ahmed Hassan Awadallah. 2021. MetaXL: Meta representation transformation for low-resource cross-lingual learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 499–511, Online. Association for Computational Linguistics.

Zhanyuan Yang, Jinghua Wang, and Yingying Zhu. 2022. Few-shot classification with contrastive learning. *Preprint*, arXiv:2209.08224.

# A Appendix

## A.1 Accuracy Tables

To improve readability and facilitate analysis of the extensive results, we have organized the accuracy and F1 scores for various shot sizes and normalization settings into dedicated tables here in the appendix section.

| | **Temporal Mean** | | | | | | | | **L2-Norm** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Shot Size | 50 | | 100 | | 150 | | 200 | | 50 | | 100 | | 150 | | 200 | |
| Language | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| Bengali | 74.86 | 74.3 | **75.41** | **74.57** | 72.97 | 72.38 | 71.35 | 70.73 | **74.05** | **73.57** | 71.08 | 70.61 | 73.78 | 72.9 | 72.97 | 72.24 |
| Bhojpuri | 71.73 | 69.59 | **77.68** | **75.91** | 75 | 73.25 | 74.7 | 73.06 | 72.62 | 70.97 | **75** | **73.58** | 75 | 72.97 | 73.21 | 71.04 |
| Gujarati | 64.09 | 62.61 | **70.17** | **68.29** | 69.61 | 66.76 | 69.61 | 67.61 | 67.4 | 65.64 | 69.89 | 67.35 | 68.78 | 66.48 | **74.03** | **71.28** |
| Haryanvi | 76.78 | 76.68 | 73.5 | 73.47 | 75.14 | 75.02 | **78.96** | **78.83** | **78.14** | **78.1** | 74.32 | 74.27 | 74.59 | 74.49 | 77.6 | 77.57 |
| Hindi | 71.27 | 71.25 | 73.71 | 73.7 | **76.69** | **76.69** | 72.9 | 72.88 | 70.46 | 70.44 | 73.44 | 73.43 | **74.25** | **74.23** | 73.98 | 73.98 |
| Kannada | 67.21 | 66.44 | **73.44** | **72.36** | 71 | 69.42 | 72.63 | 71.06 | 66.67 | 65.7 | 73.44 | 71.88 | 72.9 | 71.3 | **73.71** | **72.05** |
| Malayalam | 73.92 | 72.27 | 74.46 | 72.52 | **76.61** | **74.67** | 76.08 | 74 | **75.54** | **73.83** | 75.54 | 73.33 | 74.73 | 72.68 | 75.27 | 73.58 |
| Odia | 74.79 | 74.16 | 74.25 | 73.85 | **75.07** | **74.55** | 74.25 | 73.77 | **76.99** | **76.28** | 76.16 | 75.59 | 76.16 | 75.59 | 71.23 | 70.92 |
| Punjabi | 76.02 | 75.62 | 75.75 | 75.21 | 74.93 | 74.62 | **77.66** | **77.43** | 73.02 | 72.51 | 75.75 | 75.36 | 76.29 | 76.01 | **77.66** | **77.4** |
| Tamil | 73.85 | 71.23 | 73.32 | 70.64 | **74.93** | **72.11** | 73.05 | 70.4 | **73.85** | **71.13** | 72.24 | 68.87 | 73.05 | 70.4 | 72.51 | 69.8 |

Table 2: Few-shot Classification Results for Wav2Vec
Acc: Accuracy, F1: Macro F1-Score

| | **Temporal Mean** | | | | | | | | **L2-Norm** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Shot Size | 50 | | 100 | | 150 | | 200 | | 50 | | 100 | | 150 | | 200 | |
| Language | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| Bengali | 72.16 | 71.74 | 73.24 | 73 | **74.32** | **73.65** | 69.73 | 69.66 | **82.7** | **82.45** | 82.7 | 82.33 | 82.16 | 81.91 | 80.54 | 79.9 |
| Bhojpuri | 71.73 | 70.63 | **79.17** | **77.85** | 77.68 | 75.56 | 74.4 | 73.62 | **82.74** | **81.34** | 78.87 | 77.12 | 78.57 | 76.58 | 81.85 | 80.19 |
| Gujarati | 72.38 | 69.09 | 71.55 | 68.23 | 70.44 | 68.26 | **75.69** | **72.09** | **83.7** | **81.73** | 78.73 | 76.43 | 79.56 | 77.56 | 80.11 | 77.75 |
| Haryanvi | 75.68 | 75.67 | **78.14** | **78.14** | 77.05 | 76.99 | 77.05 | 76.93 | 82.24 | 82.21 | **84.7** | **84.69** | 84.43 | 84.4 | 83.06 | 83.04 |
| Hindi | 71.27 | 71.26 | **76.96** | **76.94** | 75.07 | 75.03 | 76.42 | 76.2 | **84.28** | **84.26** | 82.38 | 82.37 | 81.57 | 81.53 | 84.01 | 84 |
| Kannada | 68.83 | 68.2 | 75.34 | 73.21 | 70.46 | 70.26 | **76.15** | **74.14** | 78.86 | 77.95 | 77.78 | 76.77 | **79.95** | **78.77** | 79.67 | 78.67 |
| Malayalam | 72.04 | 68.92 | 73.92 | 70.48 | 73.39 | 71.36 | **78.76** | **75.08** | 81.99 | 80.23 | **85.22** | **83.33** | 81.18 | 78.83 | 82.53 | 80.3 |
| Odia | 75.07 | 74.38 | **77.26** | **76.58** | 76.71 | 76.48 | 75.34 | 74.07 | 81.37 | 80.8 | 80.27 | 79.78 | **83.56** | **83.18** | 82.74 | 82.35 |
| Punjabi | 74.39 | 74.2 | 76.57 | 76.38 | 77.38 | 77.11 | **80.38** | **80.36** | 81.47 | 81.3 | 81.47 | 81.21 | **82.02** | **81.87** | 81.47 | 81.32 |
| Tamil | 69 | 65.24 | 69.27 | 65.34 | 72.51 | 70.31 | **74.93** | **71.42** | 73.32 | 70.74 | **78.98** | **75.88** | 75.74 | 72.85 | 78.44 | 76.14 |

Table 3: Few-shot Classification Results for Whisper
Acc: Accuracy, F1: Macro F1-Score

## A.2 Aggregate Scores Table

This section and the presented table is with regard to a reviewer's concern about the need to compare with a baseline. The original Dataset paper for ADIMA (Gupta et al., 2022) presents Macro F1 Scores by training a Zero-shot model on the source language and evaluating the performance on the target language using CLSRIL-23 (Gupta et al., 2021) and Max-Pooling.

| **Language** | ADIMA | Ours |
|---|---|---|
| Bengali | 79.1 | **82.45** |
| Bhojpuri | - | **81.34** |
| Gujarati | - | **81.73** |
| Haryanvi | - | **84.69** |
| Hindi | 80.7 | **84.26** |
| Kannada | 78.4 | **78.77** |
| Malayalam | - | **83.33** |
| Odia | - | **83.18** |
| Punjabi | **83.4** | 81.87 |
| Tamil | 75.2 | **75.88** |

Table 4: Baseline ADIMA vs Aggregate Macro F1 Scores for comparison

In table 4, we present a comparison of our best scores versus their approach for better clarity.