IVD Indic Vulgar Detection: A vulgar detection dataset for Hindi and Telugu languages

Anonymous ARR submission

Abstract

In this paper, we introduce a novel dataset specifically curated for detecting vulgar content in audio, focusing on two low-resource Indic languages, Hindi and Telugu. Unlike previous work, we propose a new class, *Playful*, which distinguishes vulgar expressions that lack intent to incite hate from more extreme forms. The dataset is sourced from diverse platforms and contains audio recordings featuring potentially offensive or inappropriate language. To evaluate the dataset, we employed state-of-the-art models as baselines, achieving F1 scores of 0.66 for Hindi and 0.58 for Telugu, highlighting the unique challenges and opportunities this dataset presents for further research in low-resource language processing. Disclaimer: This manuscript includes sensitive and extreme examples.

1 Introduction and Background

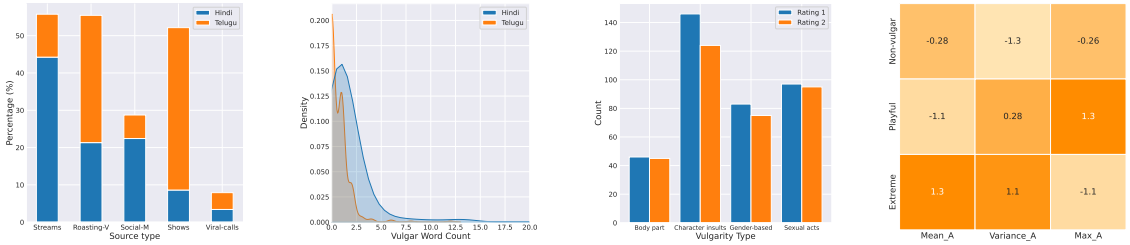
Social media has transformed global communication, offering unprecedented access to platforms for sharing information, connecting with others, and engaging in public discourse, encompassing audio, video, and text content used for education, entertainment, and social interaction. However, with the democratization of content creation, challenges in moderating harmful language have arisen, particularly concerning vulgar audio content, such as hate speech, offensive language, and slurs. The normalization of offensive language on platforms like Twitter, Reddit, and TikTok, where children and young users are increasingly exposed to vulgar content, further exacerbates this issue. This exposure not only affects online interactions but can also influence offline behavior. Manual moderation, while effective to some extent, is insufficient due to the sheer volume of content being uploaded. As such, automated systems capable of detecting and filtering vulgar language in real-time are imperative to ensure a safer online environment, particularly for younger users.

In response to this pressing need, we introduce a novel dataset specifically designed to detect vulgar audio content in two low-resource Indic languages: Hindi and Telugu. Unlike existing datasets, our dataset goes beyond simple toxicity detection by introducing a new class, *Playful*, which helps distinguish between vulgar expressions used in a non-serious or humorous context and those intended to incite hate or harm. This distinction is crucial for developing more nuanced content moderation systems that can flag genuinely harmful language while allowing playful, non-offensive expressions to remain.

Several previous works have contributed to the detection of abusive or toxic audio content. A novel dataset for toxic audio detection was introduced in (Costa-jussà et al., 2024), which focuses on classifying audio content as either toxic or non-toxic. However, this dataset lacks finer-grained labels for categorizing different levels or types of toxicity, limiting its applicability to more specific use cases. Additionally, research in (Spiesberger et al., 2023) demonstrated that acoustic features, rather than textual features, can be effectively used to detect abusive content in audio, highlighting the potential of non-textual cues in audio moderation tasks.

The ADIMA dataset, introduced in (Gupta et al., 2022), aimed at abusive audio detection, has some notable limitations. It lacks samples in Telugu, and its annotations only distinguish between abusive and non-abusive content without differentiating between varying levels of severity. Furthermore, the dataset primarily focuses on data from the ShareChat platform, neglecting other popular social media outlets. In contrast, our dataset encompasses a wider range of sources, including social media, streaming platforms, and roasting videos, and introduces the Playful category to further differentiate the nature of vulgar language.

The contributions of our work are as follows:



((a)) Source distribution of the IVD Indic Vulgar Detection dataset is illustrated in this sub-figure.

((b)) Frequency distribution of vulgar words in the dataset for Hindi and Telugu.

((c)) Categorization of vulgar instances words in the IVD dataset with classes.

((d)) Heatmap of spectrogram metrics (mean amplitude, variance, and max amplitude).

Figure 1: Comprehensive overview of the IVD dataset, including source distribution, vulgar word frequency, vulgar word categorization, and spectrogram analysis. The figure provides a detailed visualization of the dataset's characteristics and the observed patterns in speech data.

1. We introduce a novel dataset, IVD Indic Vulgar Detection, featuring Telugu and Hindi language data, annotated by native language experts.

2. We introduce a new category within the vulgar class, called Playful, being the first to introduce this distinction, which we believe holds significant value for the linguistic community.

3. We conduct extensive experiments with 12 benchmark models, providing a detailed analysis of the proposed dataset.

Reproducibility. We commit to releasing the code and dataset upon acceptance. A sample dataset is available here[1].

2 Dataset - IVD

The IVD Indic Vulgar Detection dataset focuses on two low-resource indic languages, Hindi and Telugu. Audio data were extracted from various sources such as streams, social media, roasting videos..etc. The source distribution of the dataset is shown in Figure 1(a).

Three language experts (ages 18 to 27) annotated the audio files for each language. The dataset achieved a Fleiss Kappa score (Krippendorff, 2011) of 0.8514 indicates almost perfect agreement among annotators. The dataset is categorized into three classes: *Non-Vulgar*, *Playful*, and *Extreme Vulgar*.

1. **Non-vulgar**: Non-Vulgar: The audio does not contain any vulgar language. Although it may include harsh or hateful speech, there is a clear absence of vulgar or obscene content.

2. **Playful-Vulgar**: In this category, vulgar words are used in a casual or lighthearted manner, with no intent to harm or offend. The tone is often friendly or joking, and any hateful remarks are unintentional. An example would be casual banter between friends.

3. **Extreme-Vulgar**: This involves the use of vulgar language with the clear intention to insult or provoke, often during heated or aggressive conversations. The speech is both offensive and hateful, and the vulgarity is deliberate and targeted.

The detailed distribution of the IVD dataset for both Hindi and Telugu is shown in Table 1. Figure 1(b) shows the frequency distribution of vulgar words in the dataset. Furthermore, the classification of vulgarity as demonstrated in figure 1(c) highlights that character insults and sexual acts dominate across different rating levels, indicating that these categories hold a more pervasive role in defining vulgarity across languages. Notably, the playful class, often overlooked, shows a significant overlap in vulgar word usage with extreme cases, challenging the binary notion of vulgarity and underscoring the complexity of speech patterns. Spectrogram analysis revealed mean amplitude, variance, and maximum amplitude for each audio. These metrics were grouped by category (Non-Vulgar, Playful, Extreme Vulgar) to identify differences in communication patterns for initial findings. The heatmap in Figure 1(d) highlights key tendencies: **Playful Vulgar (1)**: Dynamic tone with positive correlations to variance

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

Table 1: Dataset Frequency Distribution for Hindi and Telugu languages in IVD dataset.

Language	Split	Not-Vulgar	Playful	Extreme
Hindi	Train	320	265	142
	Test	80	67	35
Telugu	Train	201	70	108
	Test	50	18	27

and max values, reflecting expressive communication. **Extreme Vulgar (2)**: High intensity and variability in speech, indicating aggressive or offensive language, though not always at maximum loudness.

3 Experiments

In this section, we present a detailed description of the experimental setup used to evaluate the performance of various multilingual models for the task of vulgar speech detection in low-resource languages, Hindi and Telugu.

3.1 Model selection

For this task, we chose multilingual models including *Gemini-1.5 Flash* (Team et al., 2024), *mHuBERT-147* (Marcely Zanon Boito, 2024), and Facebook's *wav2vec-xlsr-300m* model[2]. Additionally, we used models fine-tuned specifically for Telugu and Hindi: *wav2vec2-large-xlsr-53-telugu*[3] and *Wav2Vec2-large-xlsr-hindi*[4], respectively. During experimentation, *GPT-4o* did not support direct audio input. Instead, the API used Whisper for transcription and then fed the text into *GPT-4o* for multimodal generation. However, this approach results in the loss of prosodic features, which are crucial for the task at hand. So we didn't consider *GPT-4* for the baselines.

Table 2: Performance of models on the Hindi dataset, showing precision for *Not-Vulgar* and *Vulgar* classes and the weighted F1 score (W-F1).

Model Name	Not-Vulgar	Vulgar	W-F1
Gemini-1.5	0.51	0.61	0.61
mHubert-147	0.74	0.79	**0.76**
Facebook-XLSR	0.53	0.68	0.62
Theainerd-XLSR	0.62	0.74	0.69

The responses demonstrated in Appendix A.1 confirm that *Gemini-1.5 Flash* is capable of effectively understanding and interpreting both Telugu and Hindi languages.

3.2 Training and Metrics

In the initial step, we merged the ***playful*** class into the ***Extreme-Vulgar*** class to create a unified category. This approach was used to determine whether the models could effectively classify vulgarity in the audio data. It also facilitated hyperparameter tuning for the dataset.

Subsequently, we conducted a wide range of experiments with the given models and evaluated their performance. We report F1-scores for each class and use the weighted F1-score to represent the overall performance of the models. Given that the dataset is in a low-resource language, there may be slight class imbalances. Therefore, the weighted F1-score is the ideal metric for accurately reflecting the model's performance across these imbalanced classes. All the hyperparameters used in the experiments are detailed in Section A.2 of the Appendix.

4 Results and Analysis

In this section, we provide a detailed analysis and benchmarking of the models evaluated in this proposed work. This includes a broad classification of vulgar content, as well as a more specific classification by dividing it into ***Playful Vulgar*** and ***Extreme Vulgar*** categories.

4.1 Vulgar detection

The playful class was combined with the extreme vulgar class to simplify vulgar detection into a binary classification task for the initial baselines. The performance of various models in classifying vulgar content was evaluated. As shown in Table 2, *mHubert-147* outperformed other models, including *Gemini-1.5 Flash*, by a significant margin. As expected fine-tuned version of XLSR performed well when compared to the base version. The models other than *Gemini-1.5* are fine-tuned on the dataset and then tested, resulting in better performance than the *zero-shot Gemini-1.5* model.

4.2 Analysing Playful vulgar detection

Distinguishing the ***Playful*** category from others was challenging, even with state-of-the-art models. Table 3 shows that models had lower F1 scores for the ***Playful*** category compared to ***Non-vulgar*** and ***Extreme*** categories in both Hindi and Telugu

[2]https://huggingface.co/facebook/wav2vec2-xls-r-300m

[3]https://huggingface.co/anuragshas/wav2vec2-large-xlsr-53-telugu

[4]https://huggingface.co/theainerd/Wav2Vec2-large-xlsr-hindi

Model Name	Hindi				Telugu			
	Not-Vulgar	Playful	Extreme	Weigh-F1	Not-Vulgar	Playful	Extreme	Weigh-F1
Gemini-1.5	0.53	0.29	0.55	0.44	0.72	0.18	0.42	**0.58**
mHubert	0.66	0.35	0.51	0.52	0.60	0.46	0.59	0.57
Facebook-XLSR	0.68	0.40	0.80	**0.66**	0.67	0.01	0.62	0.53
Anuragshas-XLSR	–	–	–	–	0.68	0.17	0.66	**0.58**
Theainerd-XLSR	0.67	0.36	0.78	0.64	–	–	–	–

Table 3: Performance comparison of various models on Hindi and Telugu datasets, including precision for each category and weighted F1 score.

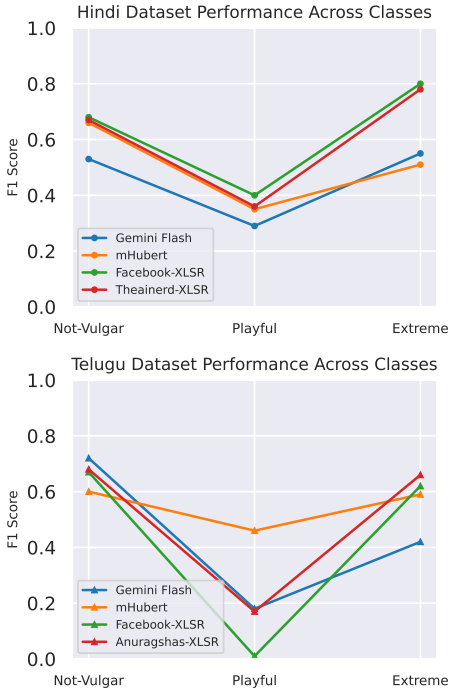


Figure 2: This figure illustrates illustrates a V-shaped dip in performance for the ***Playful*** category, where the weighted F1 score is presented across different models.

datasets. Figure 2 illustrates this challenge, with a notable dip in performance for the ***Playful*** category. Fine-tuned XLSR models generally perform better in the language they were trained on, which is evident in the Telugu data where the base model, not trained on Telugu, struggled with the ***Playful*** class. However, in Hindi, it's surprising that the base model outperformed the fine-tuned model in every class prediction. For binary vulgar detection, *mHubert-147* performed the best among all models. However, for more specific classifications, the XLSR models excelled, especially in distinguishing patterns between ***Playful*** and ***Extreme Vulgar*** classes, including tone variations. Detecting the ***Playful*** class is challenging, and further research is needed to improve accuracy in this area. More sophisticated models should be developed using the IVD dataset, which is highly relevant for research on vulgar and offensive content detection. This dataset provides valuable insights that can help advance the field and create more effective solutions for distinguishing subtle differences in tone and context.

5 Conclusion and Future Work

In conclusion, this paper introduces a novel, multi-class dataset curated specifically for detecting vulgar audio content, sourced from various social media platforms and websites. The dataset provides annotations distinguishing both the presence of vulgar language and the tone of the conversation, whether friendly or serious. The results highlight its potential as a valuable resource for advancing research on nuanced content, such as the playful class, and for moderating inappropriate audio across digital platforms.

For future work, we aim to extend the dataset to include a broader range of Indic languages, enhancing its diversity and applicability in low-resource language contexts. Additionally, we plan to involve the development and evaluation of advanced AI models tailored for multilingual and context-aware vulgar language detection, leveraging both supervised and unsupervised learning techniques to improve the robustness and scalability of content moderation systems. This expansion will further support the creation of comprehensive solutions for detecting and moderating inappropriate audio in diverse linguistic and cultural settings.

6 Limitations

A key limitation of this work is the dataset's focus on only two Indic languages, Hindi and Telugu, which restricts its applicability to other linguistic contexts. Expanding the dataset to cover more Indic languages is crucial for improving its generalizability and effectiveness in diverse real-world scenarios.

7 Ethical Considerations

We strongly condemn any misuse of the proposed dataset and emphasize that it is intended solely for ethical research and development purposes in the detection of vulgar audio content.

References

Marta R. Costa-jussà, Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alex Mourachko, Christophe Ropers, and Carleigh Wood. 2024. Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector.

Vikram Gupta, Rini Sharon, Ramit Sawhney, and Debdoot 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.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Nikolaos Lagos Laurent Besacier Ioan Calapodescu Marcely Zanon Boito, Vivek Iyer. 2024. mHuBERT-147: A Compact Multilingual HuBERT Model. In *Interspeech 2024*.

Anika A. Spiesberger, Andreas Triantafyllopoulos, Iosif Tsangko, and Björn W. Schuller. 2023. Abusive speech detection in indic languages using acoustic features. In *INTERSPEECH 2023*, pages 2683–2687.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, and Ryan Doherty. 2024. Gemini: A family of highly capable multimodal models.

A Experiments

A.1 Can Gemini-1.5 Flash understand indic languages ?

To evaluate whether Gemini-1.5 Flash (zero-shot) is a good fit for the task, we used a prompt template to determine if the model can understand and interpret the language correctly. The evaluation involved the following steps:

Template: Listen to the audio file and provide the following three outputs:

1. Translate the transcribed text into English.
2. Give a brief explanation of the audio content in English.

***Telugu output*:**

- *English Translation:* Who knows, maybe she's talking. She said that a girl shouldn't have male friends. I didn't say anything.
- *Explanation:* The audio clip is a short conversation between two people. The first speaker is expressing uncertainty about what the other person is saying. The second speaker then states that a girl should not have male friends, but the first speaker remains silent.

***Hindi output*:**

- *English Translation:* Puneet bhai, played any game Such a dirty man. I am going to lose completely brother, I am telling the truth
- *Explanation:* The audio is a short conversation between two people, possibly during a game. The speaker is expressing frustration and anger towards another player ("Puneet bhai") who is being described as "dirty." The speaker implies he is going to lose due to the actions of the other player and emphasizes that he is speaking the truth.

The above responses confirm that Gemini-1.5 Flash can understand and interpret both Telugu and Hindi languages effectively.

A.2 Experimenta setup

For the baseline training, we used pre-trained models from Hugging Face and trained them for 10 epochs. On average, it took approximately 30 minutes per model using 3 NVIDIA Tesla V100

GPUs for all training and inference operations. For the models, we unfreezed the last two layers and allowed gradients for backpropagation.

We employed the AdamW optimizer with a learning rate of 5×10^{-5}, β_1 of 0.9, and β_2 of 0.999. To prevent overfitting, we applied a dropout rate of 0.2 and a weight decay of 1×10^{-2}. For single-label classification tasks, we used Cross Entropy Loss. The model's performance was evaluated using Weighted F1 Scores to provide a comprehensive assessment across different class distributions and task types.

The learning rate scheduler was set to constant, and we used a batch size of 16. The RMS Norm Epsilon was set to 1×10^{-5}, and the Adam Epsilon to 1×10^{-8}. The maximum sequence length was capped at 512 tokens. Gradient clipping was applied with a threshold of 1.0 to stabilize training.

These hyperparameters were carefully tuned to optimize model performance while balancing computational efficiency. The detailed hyperparameters are provided in Table 4, which includes additional parameters such as warmup steps and specific optimizer settings.

Hyperparameter	**Value**
Learning Rate (lr)	5×10^{-5}
Adam Beta1	0.9
Adam Beta2	0.999
Adam Epsilon	1×10^{-8}
RMS Norm Epsilon	1×10^{-5}
Dropout	0.2
Batch Size	16
Learning Rate Scheduler	Constant

Table 4: Hyperparameters used in the experiment.