Gendered Abuse Detection in Indic Languages

Harsh Nangia 2022199
Karan Yadav 2022234
Harshit Gautam 2022208

Abstract

Gendered abuse in online communication results in harassment alongside psychological distress and suppression of free speech especially toward marginalized gender and sexuality groups. Research on abusive language detection for English has advanced yet there is limited investigation in Indic languages especially Hindi and Tamil because of their complex linguistic structure and scarce quality annotated datasets. In this paper, we have used the ULI dataset (Arora et al., 2024) for gendered abuse detection. For Task2 [Transfer Learning], we have used dataset (Dementieva et al., 2024) for english and hindi, and for tamil we used MACD dataset (Gupta et al., 2022). We used three approaches, first we finetuned XLM Roberta (Conneau et al., 2019), In second approach we used CNN on top of XLM Roberta Embeddings for feature classification and applied softmax on top. In third approach we used BiLSTM on top of XLM Roberta Embeddings. We achieved better results than already published research studies on the task, also on kaggle competition (Vaidya and Tech, 2023).

1 Introduction

1.1 Task Description

The proposed task is to develop a gendered abuse detection model based on the labels and languages in the dataset. There are total three tasks.

1. **Build a classifier using the provided dataset only to detect gendered abuse (label 1):** This subtask involves training a classifier solely on the provided data to identify instances of gendered abuse.

2. **Use transfer learning from other open datasets for hate speech and toxic language detection in Indic languages to build a classifier to detect gendered abuse (label 1):** This subtask involves use of transfer learning by leveraging external datasets related to hate speech and toxic language in Indic languages to enhance gendered abuse detection.

3. **Build a multi-task classifier that jointly predicts both gendered abuse (label 1) and explicit language (label 3):** This subtask involves the design of a multi-task learning model capable of simultaneously predicting both gendered abuse and the presence of explicit language.

2 Related Works

2.1 Breaking the Silence Detecting and Mitigating Gendered Abuse in Hindi, Tamil, and Indian English Online Spaces (Vetagiri et al., 2024)

- A gendered abuse detection model was created by the researchers who used an combination of CNN and BiLSTM networks on English and Hindi and Tamil Twitter posts.

- They used Fasttext embeddings. The CNN model extracted abusive patterns from localized word embeddings through its convolution process while BiLSTM models tracked sequence dependencies over time.

2.2 Gender-Abusive Language Detection in Bengali Using Machine Learning Algorithms (Farjana et al., 2024)

- It explored the problem of gender-abusive language detection in Bengali by applying various classical machine learning algorithms.

- Their study highlighted the importance of feature engineering and data preprocessing in improving the classification performance of models such as Support Vector Machines and Random Forests.

2.3 ARGUABLY at ComMA@ICON: Detection of Multilingual Aggressive, Gender Biased, and Communally Charged Tweets Using Ensemble and Fine-Tuned IndicBERT (Kohli et al., 2021)

- It addressed the rising concern of offensive and aggressive language on social media by proposing multi-class classification methods to detect aggression, gender bias, and communal hate.

- They introduced two main approaches: (i) an ensemble model combining XGBoost, LightGBM, and Naive Bayes on vectorized English-transliterated data originally in Meitei, Bangla, Hindi, and English, and (ii) a fine-tuned IndicBERT-based architecture leveraging contextual embeddings for detecting misogyny and aggression.

2.4 Multi-label Categorization of Accounts of Sexism using a Neural Framework (Parikh et al., 2019)

- They presented the first study on multi-label classification of sexist content, addressing the co-occurrence of various forms of sexism such as objectification and stereotyping.

- They introduced the largest dataset for sexism categorization and proposed a hierarchical neural architecture that combines BERT-based sentence representations with distributional and linguistic embeddings.

- Their model incorporates recurrent and optional convolutional layers, and leverages unlabeled data to enhance domain-specific learning. The approach significantly outperformed traditional and deep learning baselines.

3 Methodology and Experimental Setup

We tried three approaches to get the best results possible for all the three tasks.

3.1 Taskwise

3.1.1 Task1

We used the provided training dataset and testing dataset [ULI dataset] (Arora et al., 2024) for label 1 for all the three proposed models.

3.1.2 Task2

We used the (Dementieva et al., 2024) dataset for english and hindi [5000 samples each], and (Gupta et al., 2022) [MACD] for tamil [5000 samples] for training and we used the provided testing dataset [ULI dataset](Arora et al., 2024) for label 1 for all the three proposed models.

3.1.3 Task3

We used the provided training dataset and testing dataset [ULI dataset] (Arora et al., 2024) for label 1 and label 3 and created a dataframe for each language that contains both labels. Here for tamil and english, all contained same sentences but in hindi, all sentences were not same hence we were forced to use only common sentences. We tried all the three proposed models for joint prediction.

3.2 Proposed Models :

3.2.1 Approach-1 (XLM-ROBERTa Finetuned)

We used pretrained XLM-Roberta model from hugging face and finetuned it on the training dataset.

3.2.2 Approach -2 (XLM-ROBERTa + BiLSTM)

We used pretrained XLM-Roberta embeddings and on top of that Bidirectional LSTM layer with pooling.

- This architecture uses XLM-Roberta-base as a pre-trained encoder for producing contextual embeddings from text inputs.

- These embeddings are passed out and then undergo a Bidirectional LSTM for recognizing sequential patterns in both directions.

- Global Average Pooling reduces the sequence to a fixed-sized vector that is normalized, fed through dense layers with ReLU and dropout, and finally projected into class logits using a final linear layer for binary classification (abuse or otherwise).

3.2.3 Approach -3 (XLM-ROBERTa + CNN)

We used pretrained XLM-Roberta embeddings and on top of that convolution layers with with pooling.

- XLM-RoBERTa feeds the input through to create contextualized embeddings.

- The embeddings go through two 1D convolutional layers in order to get local n-gram features.

- The features are summed using both adaptive max and average pooling and concatenated.
- The concatenated representations are layer-normalized, sent through two fully connected layers with dropout in between, and output logits for classification (e.g., abusive or not).
- The model makes use of both transformer-based contextual comprehension and CNN's local pattern recognition for abuse robust detection.

4 Dataset

The dataset consists of **23,266** posts across three languages:

- 7,638 posts in Indian English
- 7,714 posts in Hindi
- 7,914 posts in Tamil

Language	Train	Test
English	6531	1107
Hindi	6197	1516
Tamil	6779	1135

Table 1: Train-Test Split for Each Language

Each post has been annotated based on three questions:

- **Label 1:** Is this post gendered abuse when not directed at a person of marginalized gender and sexuality?
- **Label 2:** Is the post gendered abuse when directed at a person of marginalized gender and sexuality?
- **Label 3:** Is this post explicit/aggressive?

Results

The results of existing models in (Vaidya and Tech, 2023) are in *table 2*. Our Models in all approaches showed better results.

Team	**Subtask 1**	**Subtask 2**	**Subtask 3**	
	label 1	label 1	label 1	label 3
CNLP-NITS-PP	0.616	0.572	0.616	0.582
SCalAR	0.228	-	-	-

Table 2: Results of Teams in the Shared Task

Language	**XLM Roberta**	**XLM + CNN**	**XLM + BiLSTM**	**XLM + BiLSTM + CNN**
Hindi	0.7382	0.7299	0.7366	**0.7414**
English	0.7413	0.7199	**0.7438**	0.7285
Tamil	**0.8403**	0.8363	0.8334	0.8334

Table 3: Task1 Results Table [Macro Avg. F1 scores]

Language	**XLM Roberta**	**XLM + CNN**	**XLM + BiLSTM**	**XLM + BiLSTM + CNN**
Hindi	**0.5303**	0.5172	0.4758	0.5226
English	0.5414	0.5624	0.5485	**0.5791**
Tamil	0.6088	**0.6722**	0.6197	0.6357

Table 4: Task2 Results Table [Macro Avg. F1 scores]

Language	**XLM Roberta**		**XLM + CNN**		**XLM + BiLSTM**	
	Label1	**Label3**	**Label1**	**Label3**	**Label1**	**Label3**
Hindi	0.6818	0.8029	**0.7428**	0.7950	0.7219	**0.8094**
English	0.7418	0.7050	**0.7494**	**0.7074**	0.7317	0.6984
Tamil	0.8313	0.9083	0.8314	0.9191	**0.8436**	**0.9214**

Table 5: Task3 Results Table [Macro Avg. F1 scores]

- For First Task, our primary approach is XLM + BiLSTM. others are considered Baselines.
- For Second Task, our primary approach is XLM + CNN, others are considered as Baselines.
- For Third task, our primary approach is XLM + BiLSTM, others are considered as Baselines.

5 Observations and Analysis

- For **Task1**, XLMR + BiLSTM is showing the most optimal performance for all the three languages *[Table 3]* because XLM-Roberta is a multilingual transformer trained on massive cross-lingual data, capturing rich semantic, syntactic, and cross-lingual information and BiLSTM adds temporal understanding by processing sequences left-to-right and right-to-left.

- For **Task2**, XLMR + CNN is showing the most optimal performance for all the three languages *[Table 4]* because XLM-Roberta is a multilingual transformer trained on massive cross-lingual data, capturing rich semantic, syntactic, and cross-lingual information and CNN are known for detecting local patterns (like n-gram features).

- Task 2 involves shorter, noisier, and less-structured texts from open source datasets, where local context matters more than global structure.

- For **Task3**, XLMR + BiLSTM is showing the most optimal performance for all the three languages *[Table 5]* because XLM-Roberta is a multilingual transformer trained on massive cross-lingual data, capturing rich semantic, syntactic, and cross-lingual information and BiLSTM adds temporal understanding by processing sequences left-to-right and right-to-left.

6 Conclusion

This research aims to build an efficient system for detecting gendered abuse in Hindi, English, and Tamil.

7 Future Work

Future work includes developing and deploying an extension for classifying tweets or comments on any social media website for gendered abuse and hiding it or labelling it as potentially matured content.

8 Appendix

- **Google Drive Link 1:** Click here to access Models.

- **Google Drive Link 2:** Click here to access Models.

- **Kaggle Link:** Click here to access code and outputs.

References

Arnav Arora, Maha Jinadoss, Cheshta Arora, Denny George, Brindaalakshmi, Haseena Dawood Khan, Kirti Rawat, Div, Ritash, Seema Mathur, Shivani Yadav, Shehla Rashid Shora, Rie Raut, Sumit Pawar, Apurva Paithane, Sonia, Vivek, Dharini Priscilla, Khairunnisha, and 6 others. 2024. The uli dataset: An exercise in experience led annotation of ogbv. *Preprint*, arXiv:2311.09086.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Daryna Dementieva, Valeriia Khylenko, Nikolay Babakov, and Georg Groh. 2024. Toxicity classification in Ukrainian. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 244–255, Mexico City, Mexico. Association for Computational Linguistics.

Mayeesha Farjana, Barisha Chowdhury, Farhana Rahman, Zuairia Raisa Bintay Makin, Sumaiya Rahman, and Azmain Yakin Srizon. 2024. Gender-abusive language detection in bengali using machine learning algorithms. In *Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning*, pages 861–875, Singapore. Springer Nature Singapore.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, hastagiri prakash vanchinathan, and Animesh Mukherjee. 2022. Multilingual abusive comment detection at scale for indic languages. In *Advances in Neural Information Processing Systems*, volume 35, pages 26176–26191. Curran Associates, Inc.

Guneet Kohli, Prabsimran Kaur, and Jatin Bedi. 2021. ARGUABLY at ComMA@ICON: Detection of multilingual aggressive, gender biased, and communally charged tweets using ensemble and fine-tuned IndicBERT. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 46–52, NIT Silchar. NLP Association of India (NLPAI).

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. *Preprint*, arXiv:1910.04602.

Aatman Vaidya and Tattle Civic Tech. 2023. Gendered abuse detection in indic languages. https://kaggle.com/competitions/gendered-abuse-detection-shared-task. Kaggle.

Advaitha Vetagiri, Gyandeep Kalita, Eisha Halder, Chetna Taparia, Partha Pakray, and Riyanka Manna. 2024. Breaking the silence detecting and mitigating gendered abuse in hindi, tamil, and indian english online spaces. *Preprint*, arXiv:2404.02013.