Project Report on Gendered Abuse Detection in Indic Languages

Prashant Shrotriya (MT24065) , Saarim Khan (MT24077) , Vatsaly Rai (MT24144)

Abstract

Online gendered abuse in context of Indic languages introduces various challenges the reasons for this can considered like Low labeled datasets , code-mixing/switching and huge linguistic diversity.This report explores a multi-aspect approach for detecting gendered abuse mainly in Hindi , Tamil and Indian English.The main problem is divided into three different task Subtask 1: Building a classifier to detect label1 4, Subtask2 : leveraging the transfer learning approach to build a classifier to detect label1 4 and Subtask3: in which goal is to build a multitask classifier that predicts both gendered abuse label1 4 and explicit language label3 4. we propose a solution for each subtask , finedtuned multilingual BERT model , on external database utilizes the domain-adaptive pretraining and a a model MultiTaskBert a custom model built on top of bert-base-multilingual-cased. our best model outperforms baseline BiLSTM architectures achieving macro-F1 scores

1 Introduction

Online gender-based discrimination has emerged as a serious concern and threat to digital equality , especially in diverse regions countries like India where digital platforms like X , YouTube , Facebook , Instagram etc reported many gendered abusive cases especially directed towards womens and marginalized communities. According to a survey, 76 % of the Indian women surveyed register cases of online harassment severity from bullying to threat of sexual violence. In different countries like India, the challenge is that abusive content is often coded in mixed form, e.g. "Tum bilkul bevajah argue kar rahi ho" and it may also incorporate region-particluar slurs and while labelling the data perspective difference can present in between different annotators.

The ICON 2023 shared Task(8) on Gendered Abuse Detection was an initiative which result in meetings of International COnference on Natural Language Processing. The objective is to develop automated tools for detection and prevention for online gender-based violence in context of Indic languages. This report explores how multilingual transformer models like BERT can be used with techniques like transfer learning to encounter the challenges stated above. By incorporating linguistic preprocessing means normalizing hinglish language to purely hindi language and phased wise training.

2 Related Work

The paper (3) which specially focuses on detecting gendered abuse in Hindi, Tamil and Indian English. Dataset contains around 7638 posts in English, 7714 in Hindi and 7914 in Tamil. Six annotators have annotated each language data, classifying each post in to one of the three labels, label 1: if this post is not directed towards a person, will it be abused, label 2: if this post is directed towards a person, will it be abused and label 3: indicates whether the post is aggressive/explicit.

Main Task that is gendered abuse detection is divided into three subtasks , Subtask 1: the goal is to develop a classifier that classifies the post according to label 1 only , Subtask 2 is concerned with the transfer learning ,external data is given to the model from open dataset that includes hate speech and gendered abuse comments and In Subtask 3 the aim is to built a multiclass classifier that classifies label 1 and label 3. The paper incorporates a ensemble model consisting of CNN (Convolutional Neural Network) and BiLSTM(Bidirectional Long Short Term Memory).The use of CNN is to target local features while BiLSTM captures long-term sequential dependencies. For converting text into embedding vectors pretrained embedding GloVe and FastText is used with embedding dimensions of 300 and maximum sequence

length is of 100 words. Convolutional layer is size of 1dimension with 64 filters and kernel size of 2. The output of convolutional layer is passed as an input to BiLSTM layer which have hidden size 128 with dropout rate of 0.1. After passing through all the layers prediction is made with the help of SoftMax function.

Paper(1) employs XLM-RoBERTa (XLM-R) model that is trained in 100 different languages. The benefit of using this model instead of some other XLM multilingual model is that XLM-R does not need language tensors to identify language, does by input id. This model is fine-tuned by adding layers to the core model. They named the architecture of the model AniMOJity which is a combination of three distinct models in the first model the input ids are given as input by the pre-trained XLM-R model ,for accounting the loss the binary cross entropy is employed and for Adam optimizer is used learning rate 5e-6.The Second model as the input to the XLM-R the combination of Input-Id ,token-type-id and attention mask is given here also binary cross entropy and Adam optimizer was employed with learning rate of 1e-5.and lastly for the third model , input is same as second model but output from last hidden state is passed to GRU cell and here also sigmoid activation , binary cross entropy loss and Adam optimizer is used with learning rate same as model 2. Weighted average of prediction of these models are taken and new stacked model is created. Paper also incorporates pseudo labelling that involves predicting unlabeled data using labeled data. Comment is hateful or not is detected by fine tuning model and passing it to a sigmoid function and then minimizing the loss. The baseline model of this approach is m-BERT which is a good standard for multilingual text classification, Training is performed with the help of Graphical Processing Unit available in Kaggle.

Coming to paper (2) The dataset used in this paper consists of 12000 samples for training and 3000 samples for testing. A Total of four languages Meitei, Bangla, Hindi and English was present in the dataset and task to classify each row of the dataset into one of the labels presents in set of aggressive, communally charges and gender biased. The overall task is divided into three sub tasks A, B and C. The main goal of sub-Task A is build a classifier that classifies the sample into one the three labels OAG (Overly aggressive) , CAG(Covertly Aggressive) and NAG (Non-aggressive). In sub-Task B a binary classifier is built that separates GEN (gendered) and NGEN (non gendered) samples from the dataset. Similarly for the third Sub Task C classifier is employed to separate COM (communal) and non-communal (NCOM). Data preparation in this is done by transliterating the languages followed by the labelling of the language. Since machine learning models understand data in numerical form, textual data is converted into vectors with the help of Count vectorization technique that is frequency count of each word is noted in the full text. This will create a matrix that is eventually passed as an input to the Ensemble model comprising of XGBoost , LightGBM and traditional Naïve Baye. The label with the highest vote is chosen as the final label. Second Model which is considered is IndicBERT, every sentence in the dataset is finalized to the maximum length of 150 in this case, tokenization of sentence is done and embeddings is generated which is given as an input tot the IndicBERT transformer which is a multilingual ALBERT model specifically trained on mainly 12 Indian languages.

The paper (4) is the overview of 2023 ICON shared task on gender abuse detection in Hindi, English and Tamil languages. The task was published as a Kaggle competition and open to the public, the participants were given a maximum of 3 weeks to build their model. After completion of the timeline test set was released with 4 days windows to test, Participants along with their model have to submit a page summary explaining their methodology. Dataset was collected from Twitter as tweets, out of the 9 teams registered for the task only 2 team submitted their models with paper explaining their model. One of team work has been discussed in third paper where they have used an ensemble model comprising of Convolutional Neural Networks and Bidirectional LSTM. The other team named SCalAR used the only the BiLSTM model , for converting input text into embeddings they used fastText word embedding , Adam optimizer and categorical cross-entropy as loss function.

3 Methodology

After carefully reviewing previous work we proposed our first Baseline "Bert Multilingual with simple preprocessing" where in the preprocessing step we aggregates the annotation values into one single value by majority vote and fine tuned the "bert-base-multilingual-cased" multilingual model for the three languages Hindi , Tamil and En-

glish.Further on carefull inspection ,we divided main task into three subtasks and proposed a second Baseline that leverages BiLSTM with advance preprocessing techniques.For the task 1 vanilla BiLSTM is trained on Uli gendered abuse dataset(7) to predict the label1 and for the Task 2 we utilizes the external dataset incorporating the concept of transfer learning for Hindi and Tamil we used MACD dataset(5) and for English Tweet hate speech detection(6) used .and for Task 3 we built our custom muitltask classifier that utilizes jointly predicts label1 and label3.In this Task specific classification heads are present.

For the Task 1 ,The Model to detect the gendered abuse(Label 1) in Indic Languages mainly Hindi , Tamil and Indian English consists of the following components: Converting Hinglish (Romanized Hindi) to Devanagari script using Indic Transliteration package. To finalize the label for a post we used majority voting on annotator values with threshold of 0.5. The model used is Bert-base-multilingual-cased , it has 110 million parameters with 12 attention heads ,768 hidden units per layer and 12 layers. All the BERT layers are fine-tuned , for classification a fully connected layer is placed after BERT embedding.Refer Table 1

Hyperparameter	**Task 1**
Base Model	bert-base-multilingual
Learning Rate	2e-5
Batch Size	16 (train), 32 (eval)
Epochs	5
Sequence Length	128 tokens
Optimizer	AdamW
Loss Function	Cross-Entropy
Data Augmentation	None
Key Additions	Script Normalization

Table 1: Hyperparameters for Task 1

Hyperparameter	**Task 2**
Base Model	bert-base-multilingual
Learning Rate	Phase 1: 2e-5, Phase 2: 1e-5
Batch Size	32 (DAPT), 16 (fine-tune)
Epochs	DAPT: 1, Phase 1: 5, Phase 2: 5
Sequence Length	128 tokens
Optimizer	AdamW
Loss Function	Focal Loss ($\gamma = 2$, $\alpha = 0.25$)
Data Augmentation	None
Key Additions	DAPT + Phased Training

Table 2: Hyperparameters for Task 2

For the Sub Task 2: The main goal is to improve the performance acheived in task1 on label 1 by leveraging external high quality and quantity dataset , we utilizes two external databases.For Hindi and Tamil MACD Dataset(5) and for the English Davidson's hate speech corpus(6) is utilized .Domain- Adaptive pretraining strategy is used where abusive patterns are exposed to model by training on Davidson's Hate Speech Corpus and MACD dataset and then with the help of Masked Language Modelling model learn abusive languague contextually , tokens in hate speech are masked and BERT predicts them and thus learn pattern for abusive languages. Fine Tuning is performed in two phases,In the first phase layers of model are frozen means weight remain unchanged , A fully connected linear layer called classification head is added on top of model to learn gender-abusive pattrerns. This phase levarages model's pretrained knowledge simulataneously adapting to new dataset. In the second phase once fully connected layer (classifier head) start producing good results , the top 4 layers of BERT are unfrozen and now the model is fined with a smaller learning rate .The reason for unfreezing top 4 layers only is that only top few layers hold more task-relevant attributes while down layers captures more general lingustic properties. Refer Table 2

for Sub task 3,here the goal is to jointly predict gendered-abuse(label 1) and explicit language(label 3) by incorporating shared encoder and label-specific classifier heads.The encoder layers of BERT multilingual model represents contextual embeddings and since there are two different labels to predict ,two separate classification heads is used.To prevent overfitting ,better generalization/model does not depends only on specific words ,random word deletion is applied where every word has 10 % probability of removal.Here Training happens in two phases where in first phase model layers are fronzen and in second phase classifier heads continue training with higher learning rate to adapt on labels specific features. Data Augmentation strategy , differential learning rates and phased wise training are major compnents of this task.Refer Table 3

4 Dataset and Experimental Setup

The dataset both task specific and external general domain specific are very important component of the system to encounter the challenges in gendered abuse detection in multilingual and code mixed context. The Task specific dataset is the Uli Dataset(7)

Hyperparameter	Task 3
Base Model	bert-base-multilingual
Learning Rate	Phase 1: 2e-5, Phase 2: 1e-5 / 2e-5
Batch Size	16 (train), 32 (eval)
Epochs	Phase 1: 4, Phase 2: 5
Sequence Length	128 tokens
Optimizer	AdamW
Loss Function	Combined Cross-Entropy
Data Augmentation	Random Word Deletion (10%)
Key Additions	Multi-Task Architecture

Table 3: Hyperparameters for Task 3

that was developed as part of Uli project which focuses to counter online gender based violence. It includes Indic languages where Hindi has 7713 total samples (Train 6197 and Test 1516) , English has 7638 total samples (Train 6531 and Test 1107) and Tamil has 7914 total samples (Train 6779 and Test 1135). Training data has total of 9 files 3 label specific file for each languages , There are three labels , Label 1 indicates , Is the post a genedered abuse when directed at a person of marginalized gender? Label 2 corresponds , Is this post a gendered abuse when it is not directed at a person of marginalized gender and last label 3 represent, Does this post contain explicit/aggressive language ? Dataset Column explanation

- **id**: A unique identifier for each row.
- **text**: The content of the post.
- **language**: The language in which the post is present.
- **key**: Indicates which label corresponds to the post.
- **en_a1** ... **en_a6**: Annotator values assigned by English annotators.
- **hi_a1** ... **hi_a5**: Annotator values assigned by Hindi annotators.
- **ta_a1** ... **ta_a5**: Annotator values assigned by Tamil annotators.

A total of 18 activists and researchers have annotated the data.

For the transfer learning we we used the Davidson's Hate Speech Dataset for the English language and Multilingual Abusive Comment Detection (MACD) Dataset fot the Hindi and Tamil language. Davidson's Hate Speech Dataset contains around 24k tweets and Thre are three types of labels "Hate Speech" , "Offensive Language" and "Neither" while MACD contains around 55k samples combined for Hindi and Tamil and there are two types of labels "Abusive" and "Non-Abusive" .

The project is setup on kaggle platform using T4 x2 GPU (16GB VRAM) to balance computational efficiency and cost.

5 Results

The evaluation metric used is the macro-F1 score.

ICON SHARED TASK Result(4)

Language	Task 1	Task 2
CNLP-NITS-PP	0.616	0.572
Scalar	0.228	-

Table 4: F1-Score Comparison for Task 1 and Task 2

Language	Task 3 (Label-1)	Task 3 (Label-3)
CNLP-NITS-PP	0.616	0.582
Scalar	-	-

Table 5: Macro-F1 Score for Task 3 Labels

Starting from the Baseline1 The macro-F1 score is reported in table 6.The highest F1 score acheived is for English language.

Language	F1-Score
English	0.6739
Hindi	0.3633
Tamil	0.5663

Table 6: F1-Score Comparison Across Languages

for Baseline 2 macro-F1 score is reported for all the task and for each language table 7 and 8 , Tamil language acheives the highest performance across all tasks with respect to all languages

Language	Task 1	Task 2	Task 3 (Label 1)
English	0.619	0.590	0.583
Hindi	0.636	0.592	0.627
Tamil	0.763	0.762	0.775

Table 7: F1-Score Comparison Across Different Tasks

The macro-F1 score for the main model after considering these two Baselines. Refer 9 and 10

These are macroF1 Score for Task1 and Task2 for all languages

Language	(Task 3 Label 3)
English	0.559
Hindi	0.702
Tamil	0.853

Table 8: F1-Score for Task 3 (Label 3) Across Languages

Language	Task 1	Task 2
English	0.736	0.738
Hindi	0.712	0.703
Tamil	0.825	0.815

Table 9: F1-Score Comparison for Task 1 and Task 2

These are macroF1 Score for Task3-label1 and Task3-label3 for all languages.

Language	Task 3 (Label-1)	Task 3 (Label-3)
English	0.713	0.728
Hindi	0.678	0.773
Tamil	0.809	0.876

Table 10: F1-Score Comparison for Task 3 Labels

Training and Validation Plots

Training and Validation plots for Task1 Refer 1 and 2

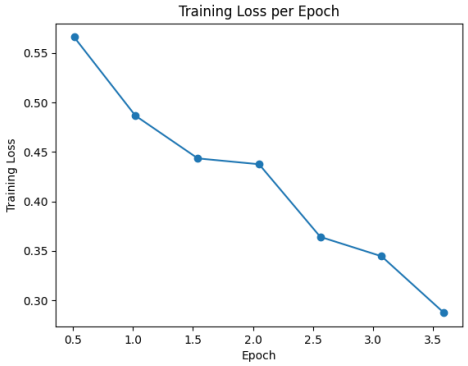

Figure 1: Example Plot Visualization

Training and Validation plots for Task2 Refer 3 ,4,5 and 6

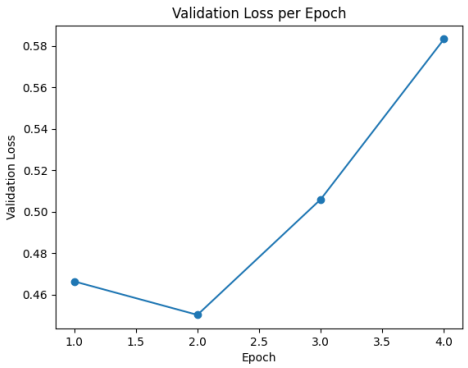

Figure 2: Example Plot Visualization

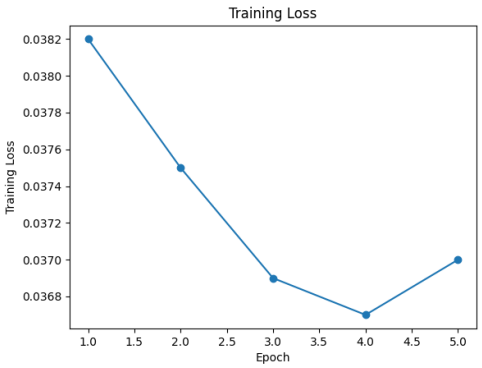

Figure 3: Training Loss Over Epochs for phase 1

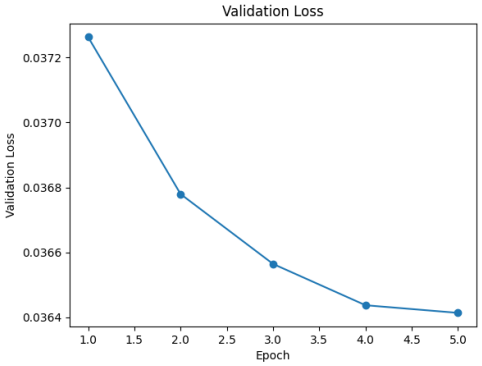

Figure 4: Evaluation Loss Over Epochs for phase1

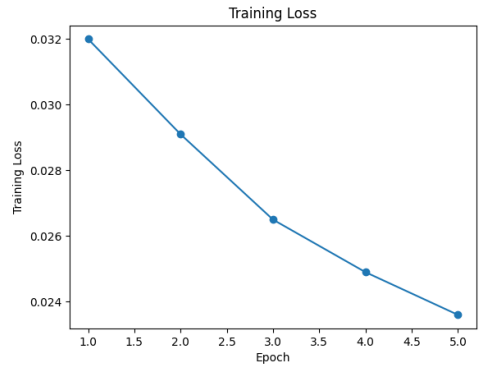

Figure 5: Training Loss Over Epochs for phase 2

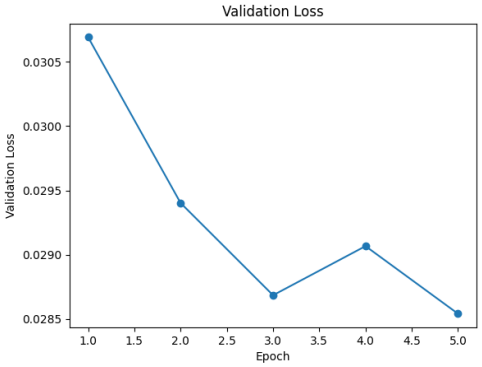

Figure 6: Evaluation Loss Over Epochs for phase 2

Training and Validation plots for Task3 Refer 7 and 8

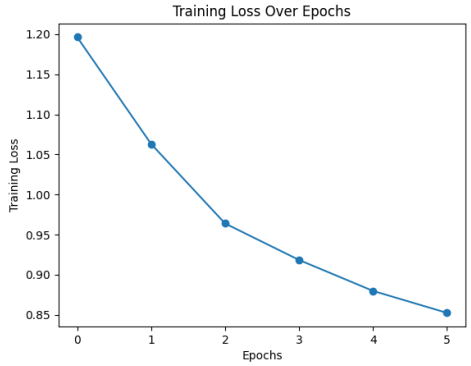

Figure 7: Training Loss Over Epochs

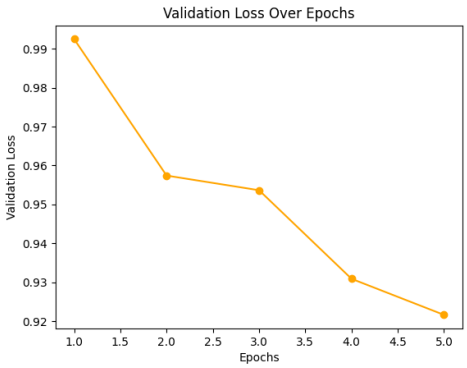

Figure 8: Evaluation Loss Over Epochs

6 Conclusion

This work presents a novel solution for this problem ,detecting gendered abuse in multilingual and code-mixed set up in social media plattforms like twitter , Youtube , Instagram for Hindi , English and Tamil Language. By utilizing the multilingual BERT , performing domain-adaptive pretraining part of transfer learning and building custom multitask classifier ,we achieved significant improvements over our Baseline models and previous study.

References

[1] Rahul Khurana, Chaitanya Pandey, Priyanshi Gupta, Preeti Nagrath. *AniMOJity: Detecting Hate Comments in Indic Languages and Analyzing Bias against Content Creators*. Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India.

[2] Guneet Singh Kohli, Prabsimran Kaur, Dr. Jatin Bedi. *ARGUABLYatComMA@ICON: Detection of Multilingual Aggressive, Gender Biased, and Communally Charged Tweets using Ensemble and Fine-Tuned IndicBERT*. Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab, India.

[3] Advaitha Vetagiri, Gyandeep Kalita, Eisha Halder, Chetna Taparia, Partha Pakray, Riyanka Manna. *Breaking the Silence: Detecting and Mitigating Gendered Abuse in Hindi, Tamil, and Indian English Online Spaces*. National Institute of Technology Silchar, Silchar, Assam, India.

[4] Aatman Vaidya, Arnav Arora, Aditya Joshi, Tarunima Prabhakar. *Overview of the 2023 ICON Shared Task on Gendered Abuse Detection in Indic Languages*. arXiv preprint, arXiv:2401.03677v1, 8 Jan 2024.

[5] Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Hastagiri Vanchinathan, Animesh

Mukherjee. *MACD: Multilingual Abusive Comment Detection at Scale for Indic Languages*. Available at: https://proceedings.neurips.cc/paper_files/paper/2022/file/a7c4163b33286261b24c72fd3d1707c9-Paper-Datasets_and_Benchmarks.pdf

[6] Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber. *Automated Hate Speech Detection and the Problem of Offensive Language*. Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM), 2017. Available at: https://github.com/t-davidson/hate-speech-and-offensive-language

[7] 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, Grace Banu, Ambika Tandon. *The Uli Dataset: An Exercise in Experience Led Annotation of oGBV*. University of Copenhagen, Denmark, Tattle Civic Tech, Bebaak Collective, National Council of Women Leaders, Center for Internet and Society, Independent, Chambal Media/Khabar Lahariya. Available at: https://arxiv.org/html/2311.09086v2

[8] ICON 2023 Tattle Shared Task. *Overview of the ICON 2023 Shared Task on Gendered Abuse Detection*. Available at: https://sites.google.com/view/icon2023-tattle-sharedtask/home