Gendered Abuse Detection in Indic Languages

Natural language processing

CSE 556

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Group No. 8

Dushyant Singh(2022181)
Mridul Goel(2022303)
Sujal Soni(2022513)

Problem Statement and Motivation

The Problem:

- **Gendered Abuse**: Derogatory language targeting gender, often veiled in cultural or contextual expressions.
  - Example: Hindi slurs or Tamil phrases that traditional filters miss.
- **Challenges**:
  - Subtlety: Implicit insults evade keyword-based detection.
  - Multilingualism: Indic languages have diverse scripts, syntax, and slang.
  - Imbalance: Abusive content is a minority class, complicating model training.

Motivation:

- **Social Impact**: Protect marginalized groups from online harassment.
- **Technological Gap**: Existing tools (e.g., English-focused BERT) underperform in Indic contexts.
- **Research Opportunity**: Leverage XLM-R and transfer learning for scalable, robust detection.

Objective:

- Build a classifier using the provided dataset only to detect gendered abuse (label 1)
- Use transfer learning from other open datasets for hatespeech and toxic language detection in Indic languages to build a classifier to detect gendered abuse (label 1)
- Build a multi-task classifier that jointly predicts both gendered abuse (label 1) and explicit language (label 3)

Related Work and Foundations

- **Early Approaches**:
  - Keyword-based filters and bag-of-words models (e.g., TF-IDF + Naive Bayes).
  - Limitations: Missed context (e.g., "hot" as slang vs. temperature) and struggled with multilingualism.
  - Statistical models (e.g., TF-IDF + SVM): Limited by shallow feature extraction.
- **Deep Learning Era**:
  - CNNs: Captured local patterns but struggled with long dependencies.
  - LSTMs: Improved sequence modeling but computationally heavy.
- **Transformers**:
  - BERT: Strong for monolingual tasks, less effective for multilingual Indic data.
  - XLM-R: Pretrained on 100+ languages, ideal for cross-lingual transfer.

Key Studies:

- **ULI Dataset** (arXiv:2311.09086): Labeled Indic language posts with annotator guidelines.
- **Hate Speech Benchmarks**:
  - Davidson et al. (English hate speech dataset).
  - Kaggle Hindi and Dreamspace Tamil datasets for pretraining.
- **Transfer Learning**: Prior work shows pretraining on related tasks boosts recall in low-resource settings.

Our Positioning:

- Extend XLM-R with custom GRU and attention layers.
- Focus on gendered abuse, explicit language, and their intersection.

Pros of XLM:

- Better handling of multilingual text.
- Richer semantic understanding.

Cons of XLM:

- High memory/computational cost.
- Limited by pretraining domain (e.g., news, Wikipedia).

Dataset and Preprocessing

Dataset Details:

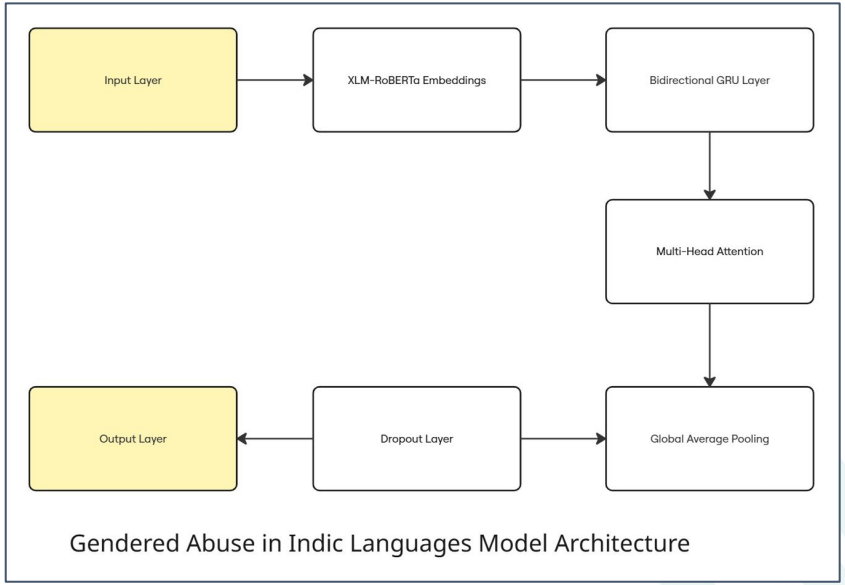
- **Source**: ULI Dataset (train_en_l1.csv, train_hi_l1.csv, train_ta_l1.csv).
- **Posts**:
  - English: 7,638
  - Hindi: 7,714
  - Tamil: 7,914
- **Splits**: Train and test sets per language
- **Labels**: Binary (0 = non-abusive, 1 = abusive).
  - **Label 1**: Gendered abuse (not targeting marginalized genders/sexualities)
  - **Label 2**: Gendered abuse (targeting marginalized genders/sexualities)
  - **Label 3**: Explicit or aggressive content
- **Aggregation**:
  - Majority voting (e.g., [0, 0, 1, 1, 1, 0] → 1).
  - Handling NaNs: Dropped rows with <3 valid annotations to ensure reliability.
- **Challenges**:
  - Missing Annotations.
  - Imbalance: requiring mitigation strategies like class weights.

Preprocessing Pipeline:

1. **Text Cleaning**:
   - Removed URLs (http://...), HTML tags (<p>), digits, and special characters (#@!).
   - Example: "Check this out http://xyz.com 😊" → "check this out [smiley]".
2. **Emoji Handling**:
   - Converted emojis to text (e.g., 😂 → "laughing face") using Python's emoji library.
   - Preserves sentiment critical for abuse detection.
3. **Normalization**:
   - Lowercased text (e.g., "WOMEN" → "women").
   - Replaced contractions (e.g., "can't" → "cannot").
4. **Tokenization**:
   - XLM-R tokenizer splits text into subwords (e.g., "playing" → ["play", "##ing"]).
   - Max lengths: English=128, Hindi=150, Tamil=160 (due to script complexity).
5. **Language-Specific Steps**:
   - **Hindi**: Normalized Devanagari Unicode (e.g., "हिन्दी" → consistent glyphs).
   - **Tamil**: Handled agglutination (e.g., "வேலைக்கு" split into morphemes).

Model Architecture

1. **Input**:
   - Token IDs and attention masks from XLM-R tokenizer.
   - Padding to max sequence length per language.
2. **XLM-RoBERTa**:
   - These are pretrained on 100+ languages, outputs 768-dimensional embeddings per token.
   - Why Chosen: Superior cross-lingual transfer and handling of code-switching (e.g., "Bro, ये too much है").
3. **Bidirectional GRU (512 units, 2 layers)**:
   - It processes embeddings bidirectionally, capturing past/future context.
   - Example: "She is useless" → GRU links "useless" to "she" for gendered intent.
4. **Multi-Head Attention (4 heads, key_dim=64)**:
   - It focuses on critical tokens (e.g., slurs or pronouns) across the sequence.
   - Attention weights reveal model focus (e.g., "she" and "useless" weighted higher).
5. **Global Average Pooling**:
   - It aggregates sequence into a fixed 512-D vector, emphasizing overall context.
   - Why Not Max Pooling: Avoids over-focusing on single extreme tokens.
6. **Dropout (0.2)**:
   - It prevents overfitting, tested with 0.1-0.3 (0.2 optimal).
7. **Dense Output**:
   - **Binary**: Sigmoid (0-1 probability).
   - **Multi-Class**: Softmax (4 classes).



Gendered Abuse in Indic Languages Model Architecture

Model Architecture

Why This architecture is used?

- **Purpose-Built for Indic Languages:** It is especially built for detecting gendered abuse in Hindi, Tamil, and Indian English to progress in this task.
- **Multilingual Understanding:** The architecture uses pretrained **XLM-RoBERTa**, a model that understands multiple languages effectively which helps to maintain consistency across dataset and It can also understand mixed languages.
- **Context-Aware Processing:** It includes a **Bidirectional GRU** to understand sentence meaning in both directions.
- **Focus on Important Words:** It uses **Multi-Head Attention** to highlight key parts of the text that indicate abuse and get more contextually aware meaning.
- **Simplified and Stable Outputs:** It uses **Global Average Pooling** and **Dropout** to reduce complexity and prevent overfitting.
- **Flexible Output Layer:** It can adapts to different tasks which is required in our project
  - Binary classification (e.g., is this abusive?)
  - Multi-class classification (e.g., Predicting gendered abuse and hat speech?)

Task Wise Analysis

Task 1: Binary Classification: Gendered Abuse Detection

Detect gendered abuse in posts using the ULI dataset (English, Hindi, Tamil).

Model

- **Architecture**: XLM-R + BiGRU + Multi-Head Attention (Section 3.3)
- **Output**: Sigmoid for binary predictions
- **Training**: Preprocessed multilingual data

Threshold Optimization

- **Challenge**: Class imbalance (fewer abusive instances) solved using class weights.
- **Process**:
  - Evaluated thresholds beyond default 0.5
  - Computed precision, recall at each threshold
  - Selected threshold with highest F1-score
- **Outcome**: Balanced detection of abuse vs. false positives

Task 2: Transfer Learning for Hate Speech Detection

- Detect gendered abuse (0 = non-abusive, 1 = abusive) using only ULI data.

Datasets: were preprocessed

- **English**: Davidson et al. (GitHub)
- **Hindi**: Kaggle Multi-labeled Hate Speech
- **Tamil**: Dreamspace Academy Dataset

Approach:

- The Model with architecture as mentioned earlier was trained on these indic datasets.
- Then they were finetuned on train_*_l1.csv and finally tested on test_*_l1.csv

Benefits

- Improves generalization across languages
- Enhances recall for underrepresented abusive cases
- Adapts to specific gendered abuse patterns

Task 3: Multi-Class Classification: Gendered + Explicit Language

- **0**: Neutral (no abuse) **1**: Gendered abuse only
- **2**: Explicit language **3**: Both gendered and explicit

Label Generation: label_1: Aggregated from 6 annotators (max frequency) for gendered abuse, label_3: Aggregated from 6 annotators (max frequency) for explicit language

Logic: Label 0: Neither label_1 nor label_3 positive, Label 1: Only label_1 positive, Label 2: Only label_3 positive, Label 3: Both label_1 and label_3 positive

Layers:

- XLM-R for multilingual embeddings
- Bi-GRU to capture context
- Multi-Head Attention for key token focus
- Global Average Pooling

Output: Softmax for 4-class prediction

Training: 5 epochs, categorical cross-entropy, class weights

Evaluation

Metrics: **F1-Score**: Balances precision/recall, ideal for imbalanced data.

- **Baseline 1 (Table 1)**:
  - Tamil outperforms others (T1: 0.79, T2: 0.78, T3: 0.21) across all tasks.
  - English and Hindi show moderate performance (T1: 0.68-0.60, T2: 0.53-0.69, T3: 0.20-0.21).
  - Task 3 (explicit/aggressive) is consistently weak across languages (0.20-0.21).
- **Baseline 2 (Table 2)**:
  - Tamil again leads (T1: 0.85, T2: 0.75, T3: 0.39).
  - English improves significantly (T1: 0.85, T2: 0.65, T3: 0.32).
  - Hindi shows steady performance (T1: 0.78, T2: 0.65, T3: 0.27).
  - Task 3 remains the weakest, with Tamil at 0.39.
- **XLM-BiGRU (Table 3)**:
  - Tamil excels (T1: 0.78, T2: 0.53, T3: 0.42).
  - English performs well (T1: 0.66, T2: 0.72, T3: 0.43).
  - Hindi lags (T1: 0.59, T2: 0.45, T3: 0.40).
  - Task 3 shows slight improvement across all languages.

Language	T1	T2	T3
English	0.68	0.53	0.20
Hindi	0.60	**0.69**	0.20
Tamil	0.79	**0.78**	0.21

Table 1: Macro F1 score of Baseline 1 across English, Hindi, and Tamil on Tasks 1, 2, and 3

Language	T1	T2	T3
English	**0.85**	0.65	0.32
Hindi	**0.78**	0.65	0.27
Tamil	**0.85**	0.75	0.39

Table 2: Macro F1 score of Baseline 2 across English, Hindi, and Tamil on Tasks 1, 2, and 3

Language	T1	T2	T3
English	0.66	**0.72**	**0.43**
Hindi	0.59	0.45	**0.40**
Tamil	0.78	0.53	**0.42**

Table 3: Macro F1 score of XLM-BiGRU across English, Hindi, and Tamil on Tasks 1, 2, and 3

Insights

- **Language Performance**: Tamil consistently outperforms English and Hindi, possibly due to dataset quality or annotation consistency.
- **Task Difficulty**: Task 3 (explicit/aggressive content) is the hardest to detect, with low F1 scores across all models and languages, suggesting a need for better feature engineering or data.
- **Model Improvement**: Baseline 2 and XLM-BiGRU show better results than Baseline 1, indicating that advanced architectures (e.g., BiGRU) enhance detection, especially for Task 2.

Conclusion and Future Works

Conclusion:

- Built a high-performing system for gendered abuse detection, leveraging XLM-R, BiGRU, and attention.
- Transfer learning and threshold optimization were key to success.
- Contributes to NLP for social good in underrepresented languages.

Future Directions:

- **More Languages**: Add Bengali, Telugu, Marathi with new datasets.
- **Real-World Deployment**: API integration for platforms like Twitter.
- **Model Enhancements**: Test GPT-style models or graph networks.
- **User Feedback**: Active learning with moderator input.
- **Ethical Considerations**: Mitigate bias (e.g., over-flagging neutral female mentions).