# GENDERED ABUSE DETECTION IN INDIC LANGUAGES

Kushal Mitra MT24050
Sankaranarayanan Senguntheer MT24081
Satyam Singh MT24082

# INTRODUCTION

- Online gender-based harassment restricts freedom of expression in digital spaces
- Detection system for abusive and sexist content across English, Hindi, and Tamil
- Dataset: 18,000+ labeled posts from ICON 2023 shared task (Tattle Civic Tech)
- Our architecture combines GRU with restricted self-attention and Transformer with Gated CNNs
- Outperforms baselines including CNN + BiLSTM, and custom transformer models with restricted self attention.

# DATASET DESCRIPTION

- Primary Dataset: ICON 2023 Shared Task (Tattle Civic Tech)
- 6,532 English, 6,198 Hindi, and 6,780 Tamil posts
- Three classification dimensions:
- **Label1**: Gendered abuse when not directed at marginalized gender/sexuality
- **Label2**: Gendered abuse when directed at marginalized gender/sexuality
- **Label3**: Explicit or aggressive content

- Additional Datasets (for transfer learning):
- MACD Dataset: Multilingual abusive comments in Hindi, Tamil, etc.
- Hate Speech Dataset: English tweets classified as hate speech, offensive language, or neither

# EXPERIMENTAL TASKS

- **Task 1:** Detect gendered abuse in combined multilingual dataset (Label 1)
- Transformer + Gated CNN architecture
- mBERT embeddings

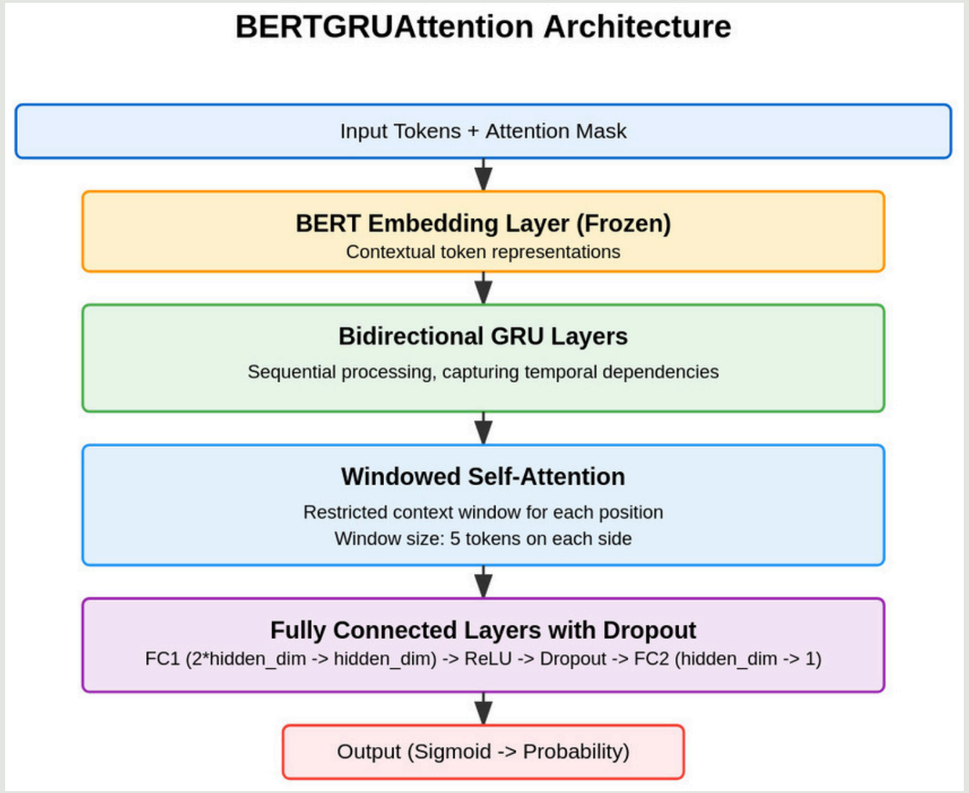
- **Task 2:** Transfer learning approach
- Pre-train on hate speech datasets (MACD, English hate speech)
- Fine-tune on ICON dataset
- BiGRU with restricted self-attention

- **Task 3:** Multi-task classification
- Simultaneous prediction of Label 1 and Label 3
- Transformer + Gated CNN architecture

# ARCHITECTURE 1: GRU WITH RESTRICTED SELF-ATTENTION

Advantages:

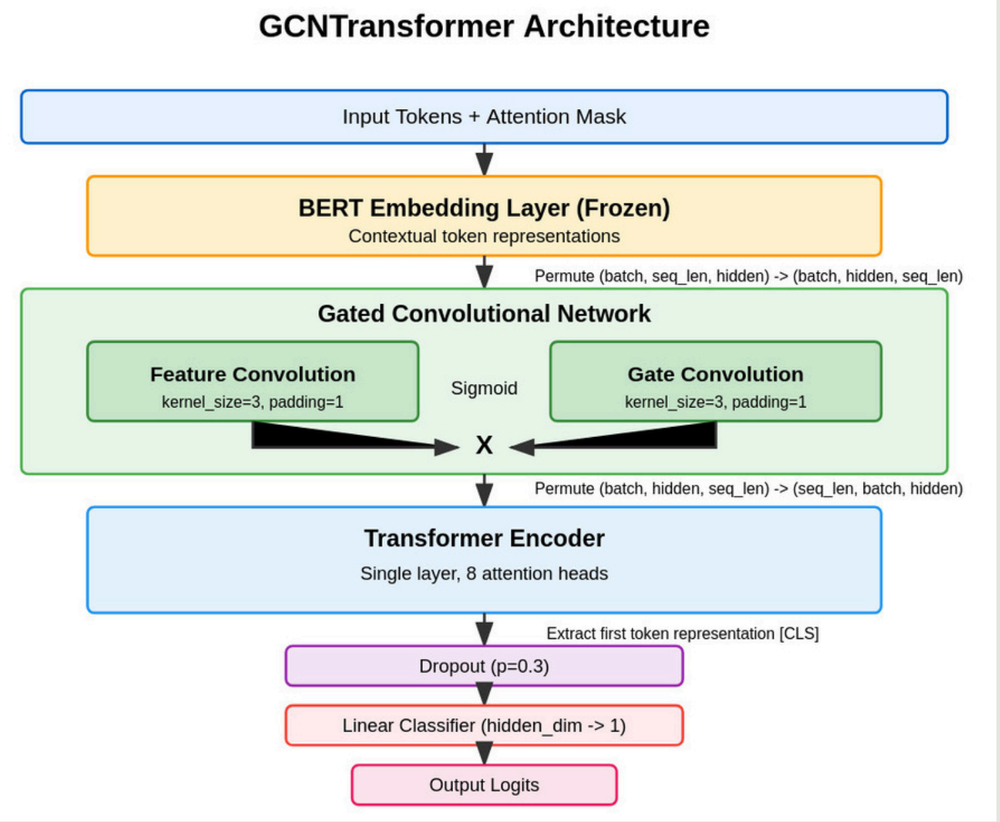
- Efficiently captures local linguistic patterns
- Better handling of code-switched text
- Focused attention on abusive language markers



# ARCHITECTURE 2: TRANSFORMER WITH GATED CNN

Advantages:

- Captures hierarchical n-gram features
- Models global dependencies in text
- Handles noisy social media content effectively



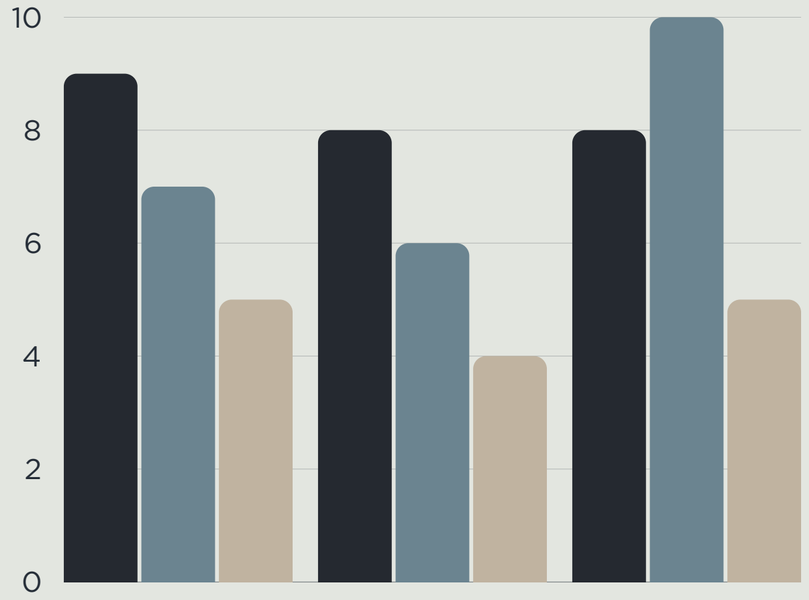
# EXPERIMENTAL SETUP & HYPERPARAMETERS

| Parameter | Task 1 (Transformer+GCN) | Task 2 (GRU+RSA) | Task 3 (Transformer+GCN) |
|---|---|---|---|
| Epochs | 10 | 10 | 20 |
| Learning Rate | 1e-5 | 3e-5 | 2e-6 |
| Weight Decay | 1e-9 | - | 0.01 |
| Batch Size | 16 | 16 | 16 |
| Optimizer | Adam | Adam | AdamW |

- Preprocessing:
- Lowercase conversion, URL and HTML tag removal
- Language-specific processing (non-alphabetic character removal)
- Extra whitespace elimination

# EVALUATION METRICS & RESULTS

| Task | Model | Test Macro F1 | False Negatives |
|---|---|---|---|
| Task 1 | Transformer+GCN | 0.752 | 417 |
| Task 1 | Custom Transformer (baseline) | 0.699 | 431 |
| Task 2 | GRU+RSA | 0.707 | 413 |
| Task 2 | Custom Transformer (baseline) | 0.609 | 467 |
| Task 3 (Label 1) | Transformer+GCN | 0.659 | 454 |
| Task 3 (Label 3) | Transformer+GCN | 0.780 | - |



# CONCLUSION

- Successfully developed models for detecting gendered abuse in Indic languages
- Architecture combining Transformers with GCNs performs best for direct classification (Task 1, Task 3)
- GRU with restricted self-attention superior for transfer learning scenarios (Task 2)
- All proposed architectures significantly outperform baselines

- **Key contributions:**
- Effective handling of multilingual and code-switched content
- Focused attention on abusive language patterns
- Balanced performance across languages
- **Limitations:** Still challenges with heavily code-switched content
- Provides robust detection system to promote safer, more inclusive digital spaces

# Thank You

For your “attention”