GENDER BASED ABUSE DETECTION

Presented by Akshat W, Akshat T and Shashank

PROJECT INTRODUCTION

- The rise of online gender-based abuse in India's multilingual digital spaces poses a significant challenge, as content in languages like Hindi and Tamil often goes undetected due to inadequate automated systems. This issue marginalizes certain groups and reveals a gap in NLP research.
- Current tools struggle with code-mixing, regional slurs, and culturally specific sexism. To combat this, we aim to develop an AI system that accurately detects gendered abuse while addressing class imbalance. The macro-F1 score will be used as the primary evaluation metric to balance the consequences of false negatives and positives in this critical classification task.

GOALS OF GENDER ABUSE DETECTION

- Identify if a post is gendered abuse when it is directed at a person of marginalized gender and sexuality and when it is not directed at a person of marginalized gender and sexuality.

- Identify if the post by a user is explicit or aggressive.

GENDER ABUSE DETECTION (GAD) TASKS

Classifier on ULI Dataset

Transfer Learning using other dataset

Multi-task classifier that predicts both gendered abuse and explicit language

DEFINING DATASET

- The Dataset consists of over 5K lines in English, Hindi and Tamil which have been collected from twitter/X. The data was curated by focusing on often used slurs and observing the accounts which received them the most often and those which most frequently used the highlighted slurs

- Researchers converged on the following two labels:
  - Is the post gendered abuse
  - Does the post contain explicit or aggressive language.

TASK 1: HATE SPEECH DETECTION

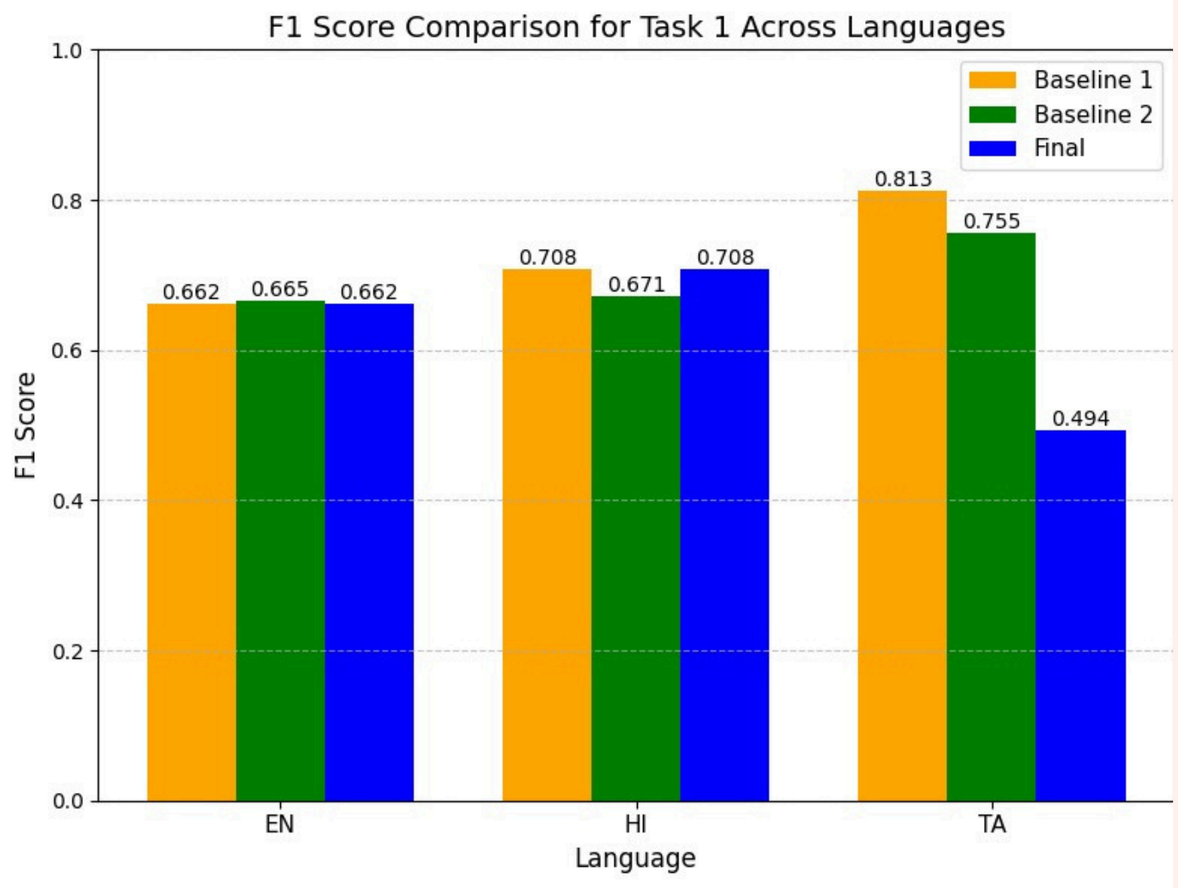
- Model Highlights:
- Preprocessing: Removes noise (URLs, mentions, hashtags), transliterates Hindi/Tamil to native scripts.
- Data Augmentation: Uses back-translation (EN → HI → EN) to improve generalization.
- Loss Function: Uses Focal Loss with class weights to handle class imbalance.
- Architecture: Fine-tunes distilroberta-base (English) and indic-bert (Hindi, Tamil).
- Outcome: Robust performance on detecting gendered abuse and explicit content.

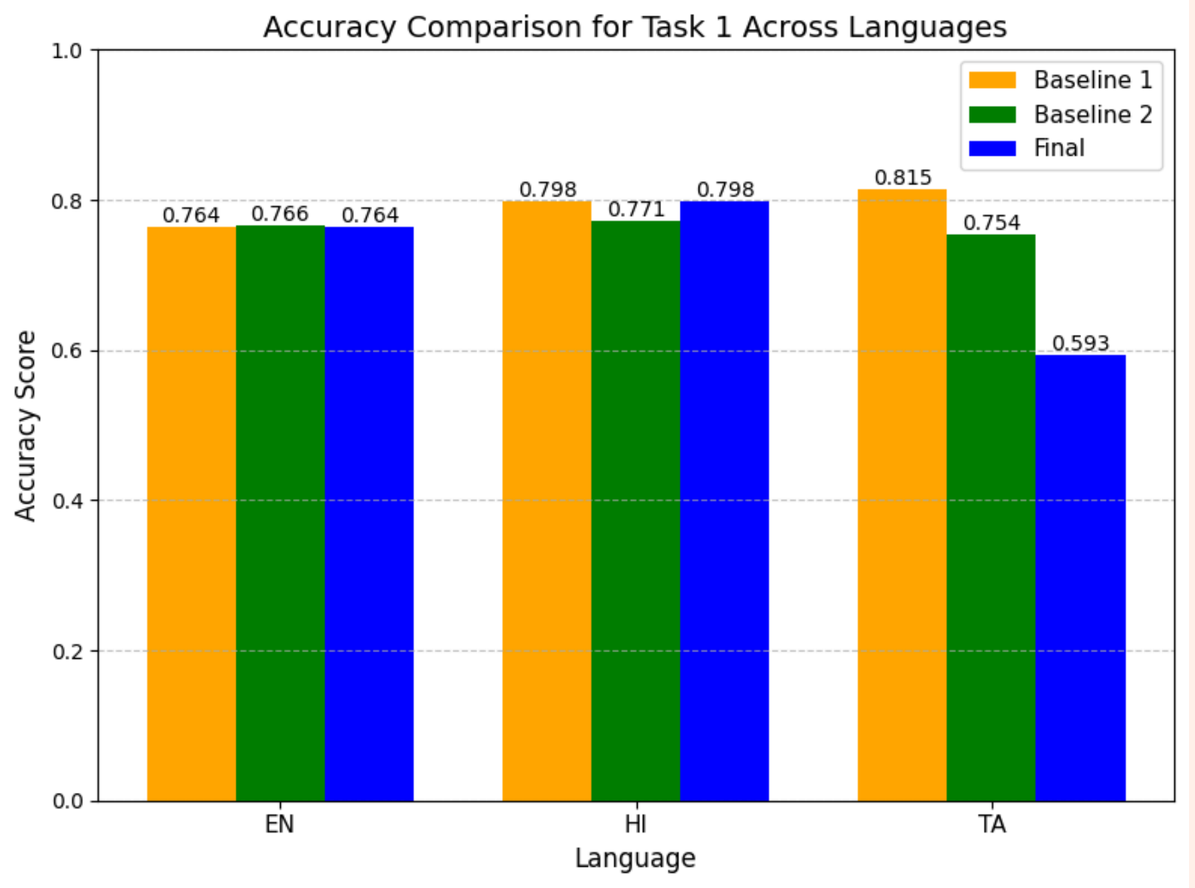
TASK 2: MULTILINGUAL TRANSFER LEARNING

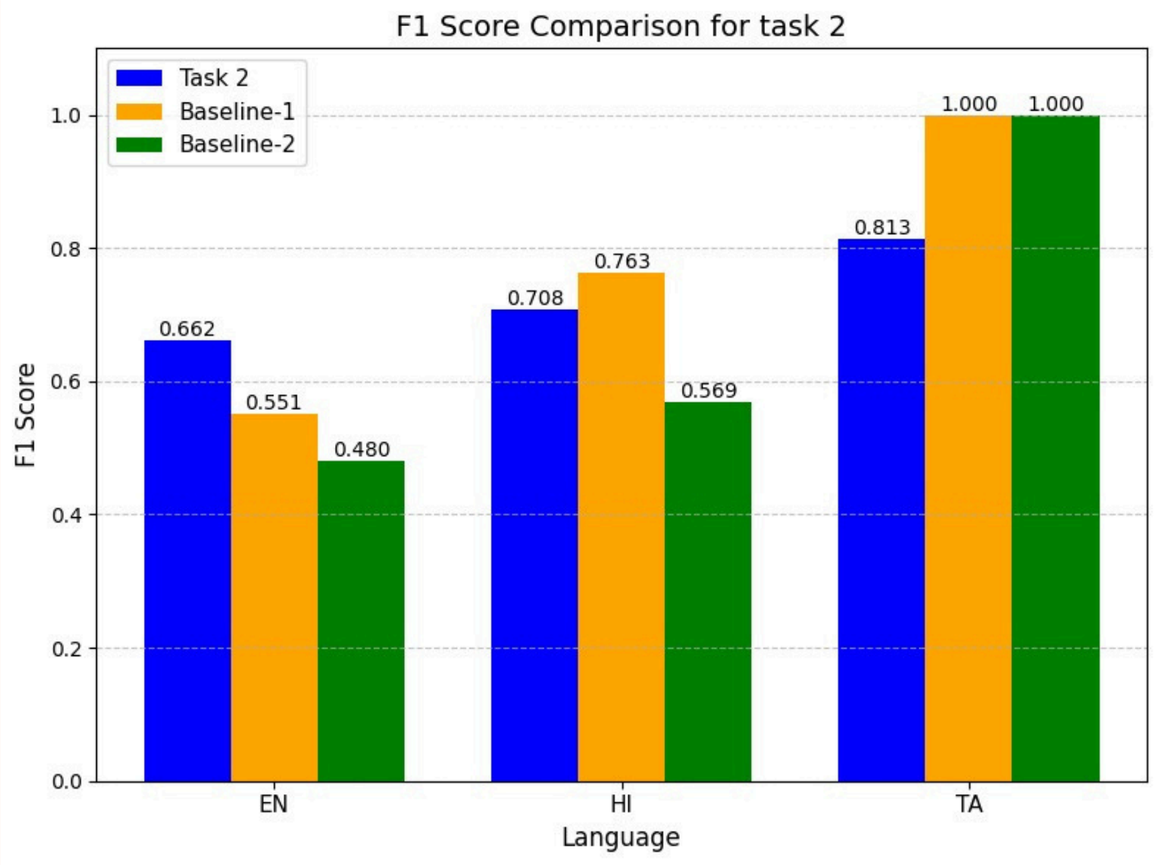
- Architecture: Custom DeBERTa-v3 model with language-specific adapters for EN, HI, TA.
- Forward Pass: Applies language adapter to last hidden state before classification.
- Loss Function: Cross-entropy loss computed per batch with language-specific inputs.
- Data Augmentation: Includes synonym replacement, word swaps (EN), character swaps (HI, TA).
- Training: 3 epochs, learning rate 2e-5, early stopping, and gradient accumulation.
- Custom Trainer: Computes cross-entropy loss, tracks F1 & accuracy.
- Outcome: Multilingual model learns language-specific nuances and generalizes across domains.

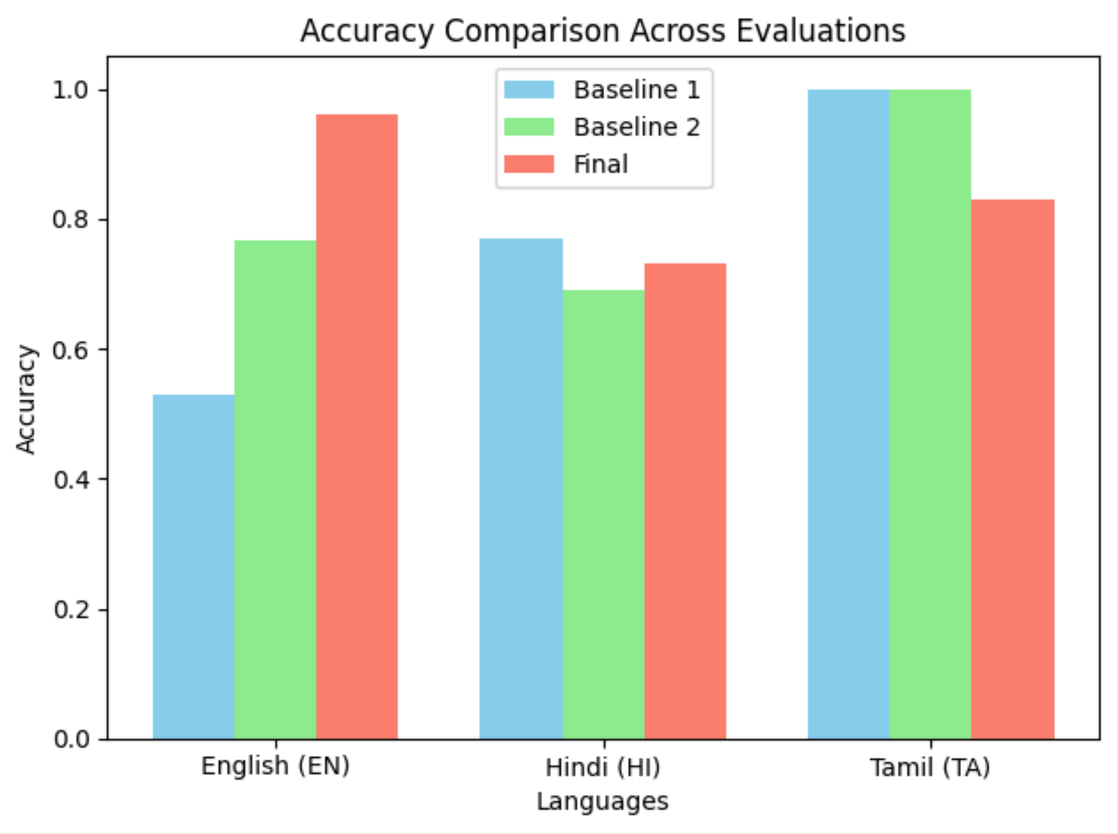
TASK 3 MULTITASK LEARNING FOR GENDERED & EXPLICIT ABUSE

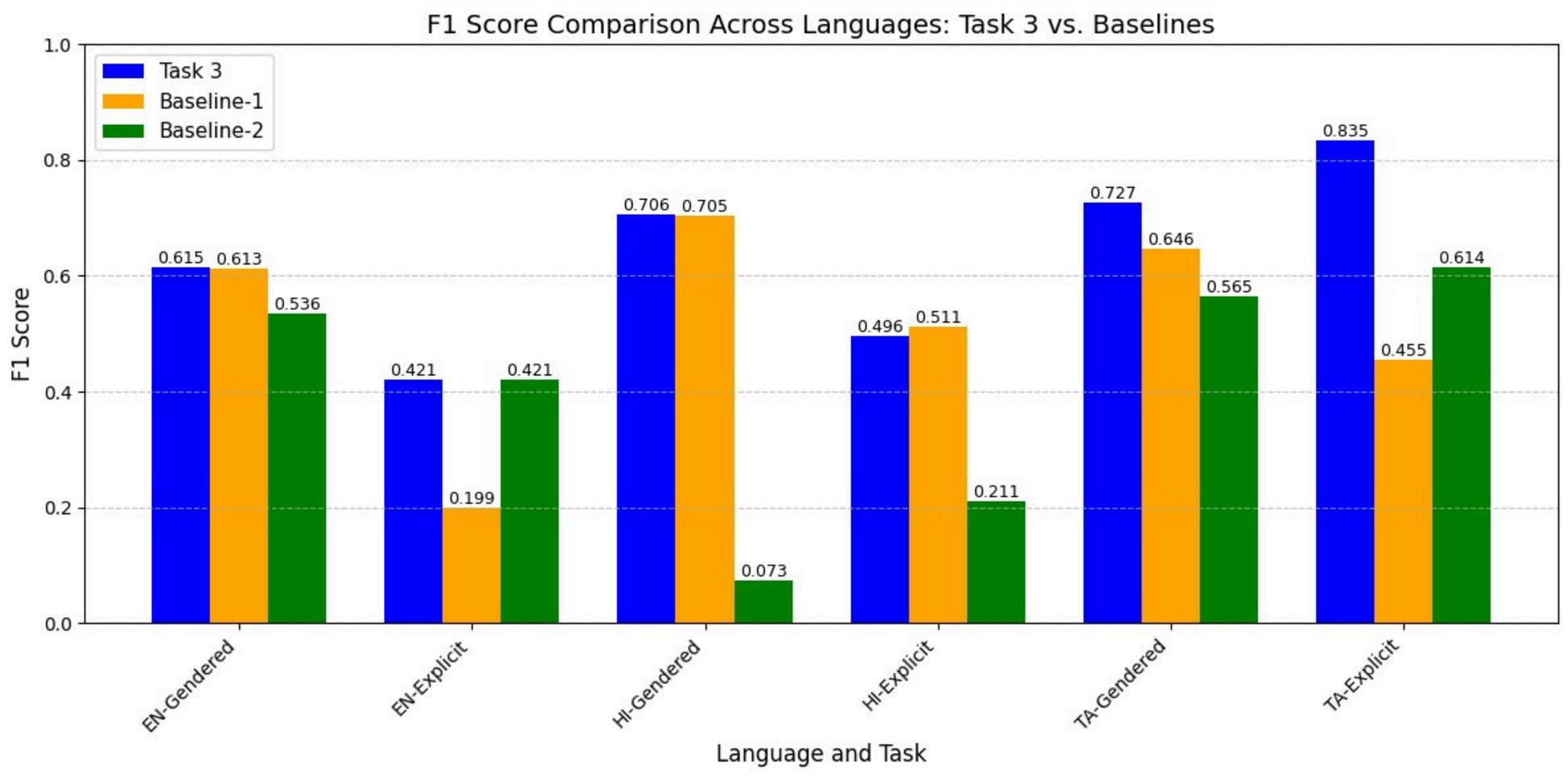
- Architecture: BiLSTM-based multitask model for joint classification of:
- Gendered abuse
- Explicit language
- Preprocessing: Standard text cleaning and language-specific tokenization.
- Shared Learning: Multitask design captures common patterns across both labels.
- Training: Done using HuggingFace's Trainer API with accuracy and F1 evaluation.
- Outcome: Efficient model that handles both subtasks simultaneously across 3 languages.


F1 Score Comparison for Task 1 Across Languages
Baseline 1
Baseline 2
Final
F1 Score
Language
EN
HI
TA
0.662
0.665
0.662
0.708
0.671
0.708
0.813
0.755
0.494


Accuracy Comparison for Task 1 Across Languages
Baseline 1
Baseline 2
Final
Accuracy Score
Language
EN
HI
TA
0.764
0.766
0.764
0.798
0.771
0.798
0.815
0.754
0.593


F1 Score Comparison for task 2
Task 2
Baseline-1
Baseline-2
F1 Score
Language
EN
HI
TA
0.662
0.551
0.480
0.708
0.763
0.569
0.813
1.000
1.000
0.0
0.2
0.4
0.6
0.8
1.0


Accuracy Comparison Across Evaluations
Baseline 1
Baseline 2
Final
Accuracy
Languages
English (EN)
Hindi (HI)
Tamil (TA)
0.0
0.2
0.4
0.6
0.8
1.0


F1 Score Comparison Across Languages: Task 3 vs. Baselines
Task 3
Baseline-1
Baseline-2
F1 Score
1.0
0.8
0.6
0.4
0.2
0.0
0.615
0.613
0.536
0.421
0.199
0.421
0.706
0.705
0.073
0.496
0.511
0.211
0.727
0.646
0.565
0.835
0.455
0.614
EN-Gendered
EN-Explicit
HI-Gendered
HI-Explicit
TA-Gendered
TA-Explicit
Language and Task