

1 Baseline Replication

For the baseline replication, one core model and one core dataset from the original study were selected:

- **Model:** ALBERT-v2
- **Dataset:** EMGSD

The baseline implementation was reproduced using the publicly released code and configurations. Data preprocessing steps, model architecture, and training procedures were aligned as closely as possible with the original experimental setup. Minor deviations were required due to differences in data availability and computational constraints. Baseline performance metrics were successfully reproduced. Model performance is reported below.

Table 1: Performance of the replicated model

Class	Precision	Recall	F1-score
0 (Non-stereotype)	0.876	0.884	0.880
1 (Stereotype)	0.771	0.759	0.765
Macro Avg	0.824	0.821	0.823
Overall Accuracy: 0.841			

Table 2: F1-score compared with the original study

Model	Dataset	F1-score (Macro Avg)
ALBERT-v2 (Original)	EMGSD	81.5%
ALBERT-v2 (Replicated)	EMGSD	82.3%

2 GBV Dataset Construction

To adapt the HEARTS framework to a gender-based violence (GBV) detection context, a new dataset was constructed using the *Jigsaw Unintended Bias in Toxicity Classification* dataset.

A GBV-focused subset was constructed by filtering comments that explicitly reference women or girls, using a combination of identity-based annotations and keyword-driven rules. A binary label was subsequently derived, where a comment is classified as *hostile* if it exhibits both high overall toxicity and at least one strong harm-related subtype (e.g., insult, threat, or identity attack), with a threshold greater than 0.3.

Manual inspection shows that, although the filtering procedure substantially increases the proportion of comments about women, a minority of retained instances are not explicitly gender-based or GBV-related. This reflects noise in the underlying toxicity annotations and the limitations of rule-based filters.

This filtering process resulted in a dataset comprising 86,719 comments, including 10,339 hostile and 76,380 non-hostile instances. The resulting class distribution reflects a realistic yet imbalanced setting commonly observed in real-world online moderation tasks.

3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to examine the characteristics of the constructed GBV-focused dataset and to validate the effectiveness of the filtering strategy.

4.2 Data Splits

The GBV dataset contains comments with a binary label (*non-hostile* / *hostile*). The data were split as follows:

- Train–test split: 80% train, 20% test,
- Stratified by label to preserve class proportions,
- Fixed random seed for reproducibility: `RANDOM_STATE = 42`.

Within the training portion, a further 20% of the data were reserved as a validation set for monitoring training:

- Train–validation split: 80% / 20% of the training data,
- Stratified by label, using the same random seed.

4.3 Training Configuration

Fine-tuning was performed using the `Trainer` API with the following hyperparameters:

Table 3: Hyperparameters for ALBERT-v2 fine-tuning on the GBV dataset

Hyperparameter	Value
Base model	<code>albert-base-v2</code>
Max sequence length	128
Batch size (train / eval)	32 / 32
Number of epochs	3
Learning rate	2×10^{-5}
Weight decay	0.01
Random seed	42
Loss function	Cross-entropy (via <code>AutoModelForSequenceClassification</code>)
Optimizer / scheduler	HuggingFace defaults for <code>TrainingArguments</code> (AdamW-style optimiser)

4.4 Performance Evaluation

During training, a held-out validation set was used for intermediate evaluation. After training, the best model was saved and evaluated on the test set. The following metrics were computed:

Table 4: Performance of ALBERT-v2 on the GBV dataset

Class	Precision	Recall	F1-score
0 (Non-hostile)	0.938	0.966	0.952
1 (Hostile)	0.676	0.525	0.591
Macro Avg	0.807	0.745	0.771
Overall Accuracy: 0.913			

The adapted model achieved an accuracy of 91.3% and a macro-averaged F1 score of 0.771. Performance was notably stronger on the non-hostile class, with high precision and recall, while recall for the hostile class was lower. This discrepancy reflects the inherent class imbalance and the linguistic diversity of hostile expressions.

Compared to the baseline HEARTS results, the adapted model demonstrates comparable overall performance while operating in a more challenging, context-specific classification setting.

5 Model Interpretability

Table 5: LIME explanations for representative prediction cases

Text Instance	Predicted	Actual	Top Token Rankings (LIME)
RAPO in the White House. Terrorists and fanatics at your local Planned Parenthood. Get your Glocks women.	1	0	“Terrorists”: 0.692, “your”: 0.159, “RAPO”: 0.123, “fanatics”: 0.099, “White”: 0.087, “Glocks”: 0.078
You must be male. Your favorite politicians certainly want domination over my body and all other women’s. So I guess your idea of “slavery” and “tyranny” is sliiii-iggghhhhtttlllly...	0	1	“silly”: 0.391, “idea”: -0.106, “and”: -0.085, “liberty”: -0.069, “favorite”: -0.066, “death”: 0.063
The “broad” is the dead kid’s mother you insensitive idiot.	1	1	“idiot”: 0.889, “insensitive”: 0.038, “broad”: -0.036, “dead”: 0.018, “you”: 0.012, “mother”: -0.010

Model interpretability was examined using Local Interpretable Model-agnostic Explanations (LIME). Analysis of **true positive** examples shows that explicit insults and overtly hostile language strongly influence correct hostile predictions. **False positive** cases indicate that the model can overemphasise emotionally charged or negative terms that are not necessarily gender-targeted, leading to misclassification of non-hostile comments. In contrast, **false negative** cases often involve indirect, contextual, or descriptive language that lacks explicit hostility markers. Together, these findings highlight the model’s sensitivity to explicit cues while revealing limitations in detecting subtle or implicit forms of gender-based hostility.

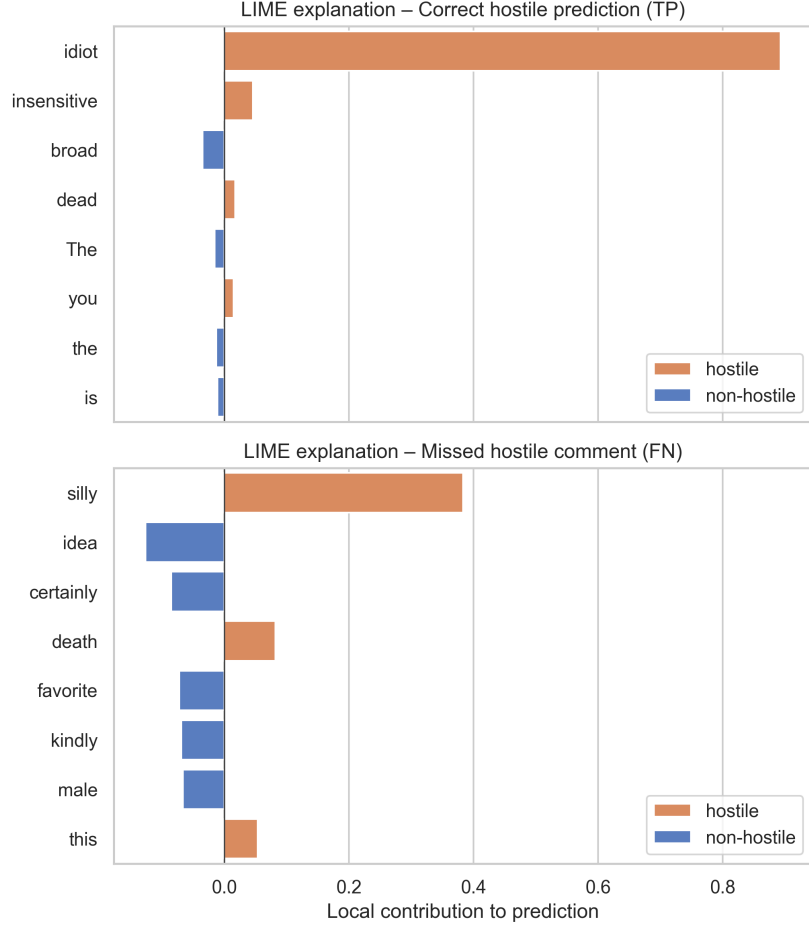


Figure 3: LIME explanations for a true positive and a false negative prediction

6 Ethical Considerations

Automated detection of hostile language raises important ethical concerns, including the risk of false positives suppressing legitimate speech and false negatives allowing harmful content to persist. Dataset bias, annotation subjectivity, and uneven representation of identity groups can further exacerbate these risks.

This project mitigates some concerns through transparency, using explainability tools to expose model decision logic. However, human oversight remains essential, and such systems should be deployed as decision-support tools rather than fully autonomous moderation solutions.

7 Scalability and Sustainability

The use of ALBERT significantly reduces memory requirements and computational cost compared to larger transformer models, supporting scalability in real-world moderation systems. Once trained, the model can efficiently process large volumes of text, making it suitable for platform-level deployment.

From a sustainability perspective, parameter-efficient architectures reduce energy consumption during training and inference. Nevertheless, regular retraining is required to adapt to evolving language use, which carries ongoing environmental and computational costs.