Table 1: Experimental settings

Setting	Value / Description
Experiment Type	Single factor experiment
Platform	Google Colab T4 GPUs
Seeds	0, 42, 123, 2025 (for `random`, `numpy`, `torch`, CUDA)
Model	Cardiff (`cardiffnlp/twitter-roberta-base-hate-latest`)
Prediction Method	HuggingFace `transformers.pipeline` (`text-classification`)
Batch Size	512
Dataset Focus	1% of the dataset with `cardiff_score` ≥ 0.95 , divided into quarters of True Positives, True Negatives, False Positives, and False Negatives. Sample size per category: 163.
LIME Settings	
Explainer Type	`LimeTextExplainer`
Other Settings	Default values used
SHAP Settings	
Explainer Type	`shap.Explainer`
Masker	`shap.maskers.Text` with corresponding model tokenizer
Other Settings	Default values used