DataFrame Column	Description
id_new	Created
text	Unaltered original text
extracted_target	Extracted from original targets
label_hatespeech_binary_offensive_not_included	Offensive not considered hate speech (binary)
label_hatespeech_binary_offensive_included	Offensive included as hate speech (binary)
label_normal_offensive_hatespeech	Normal, offensive, or hate speech (multi-class)
platform	Source of text
original_dataset_title	Original dataset name
original_id	Original dataset ID
original_label	Original label from dataset
original_target	Majority-annotated targets

Table 1: Merged dataset column descriptions

Setting	Value / Description
Seed	42 (for 'random', 'numpy', 'torch', CUDA)
Models	Cardiff ('twitter-roberta') and Facebook ('roberta-hate-speech')
Prediction Method	HuggingFace 'transformers.pipeline'
Batch Size	512
LIME Settings	
Number of Features	4
Number of Samples	500
SHAP Settings	
Explainer Type	`shap.Explainer`
Max Evaluations	100
Masker	`shap.maskers.Text` with tokenizer

Table 2: Unified experimental settings for LIME and SHAP explanations

Table 3: Consolidated Experiment Settings and XAI Parameters

Component	Setting	Value / Description
General	Seeds	4 random seeds between 0 and 100000: [83811, 14593, 3279, 97197] (used for random, numpy, torch, CUDA)
	Model	CardiffNLP (`cardiffnlp/twitter-roberta-base-hate-latest`)
	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	Explainer Type	`LimeTextExplainer`
	random_state	Same as experimental seeds: [83811, 14593, 3279, 97197]
	Other Settings	Default values used
SHAP	Explainer Type	`shap.Explainer`
	Masker	`shap.maskers.Text` with corresponding model tokenizer
	seed	Same as experimental seeds: [83811, 14593, 3279, 97197]
	Other Settings	Default values used