Table 1: Overview of Hate Speech Detection Studies Using XAI

Reference	Dataset Source	Classification Task	Models Used	XAI Method	XAI Use
Ansari, Kaur and Saxena (2023)	Youtube, Facebook, Twitter	Binary	LSTM, CNN	LIME, Integrated Gradient	4 visual examples: Local explanations using and using quantitative metrics of AOPC, log-odds, and coherence
Hashmi et al. (2024)	Public datasets and multilingual corpora	Binary	DT, RF, LR, SVM, mBERT, LSTM	LIME	23 visual examples: Two for each language, except one with 20 examples (1.00 prob.)
Wich et al. (2021)	Twitter	Binary	DistilBERT, BoW, GraphSage	SHAP	6 visual examples: Explanation of each submodel, 1 example with 1.00 prob.
Imbwaga, Chittaragi and Koolagudi (2024)	Youtube	Binary	SVM, RF, XGBoost, Bi-LSTM, BERT, GPT-J-6B	LIME	2 visual examples: Hate-labeled examples in English and Kiswahili
Nandini and Schmid (2023)	Twitter	Multiclass	BERT	LIME	4 visual examples: One hate-labeled + aggregated features from 50 examples/class
Siddiqui et al. (2024)	Twitter and cyberbullying datasets	Fine-grained	mBERT, XLM-RoBERTa, Distil-RoBERTa	LIME	9 visual examples: TP, TN, FP, FN + 5 fine-grained samples, 4 with 1.0 prob.
Tiwari (2024)	Twitter	Binary	LSTM	LIME	4 visual examples: One for each LSTM word embedding method
Yadav, Kaushik and McDaid (2023)	Twitter	Binary	LR, SVM, NB-G, NB-B, NB-M, RF, KNN, DT	LIME	2 visual examples: One TP and one FP, with one at 1.0 prob.
Babaeianjelodar et al. (2022)	Twitter	Multiclass	XGBoost, LSTM	SHAP	1 visual example: Explained hate-labeled tweet
Hashmi and Sule Yildirim Yayilgan (2024)	Facebook, Twitter, Resset	Multiclass	LSTM, GRU, FAST-RNN	LIME	12 visual examples: Two per class + two misclassified
Mehta and Passi (2022)	Twitter, Gab, Wikipedia	Binary	DT, RF, NB, LR, LSTM, BERT+ANN, BERT+MLP	LIME	4 visual examples: DT, RF, LR, NB models and quantit metrics: IOU F1, Token F1, AUPRC, Comprehensiveness, Subgroup AUC, etc.
Mazhar Qureshi, Qureshi and Rashwan (2023)	Twitter	Multiclass	EBM, SVC	LIME, SHAP, Counterfactual	21 visual examples: SHAP/LIME examples across classes + counterfactual text edits
Hareem Kibriya et al. (2024)	Twitter	Fine-grained	LSTM	LIME, SHAP	7 visual examples: 4 Lime (weighted features), 2 SHAP-based instances

Table 2: Distribution of Detrimental Content Types

Detrimental Content	**Total Count**
hate	13
fake_news	9
rumours	3
toxic	3
misinformation	2
troll	2
sexism	2
cyberbullying	2
fake_reviews	2
controversy	1
misogyny	1
suspension	1