Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reliability of Urdu–Arabic Abusive Language Detection with Cross- Lingual Transformers and Explainable AI

Malik Daler Ali Awan
daler.ali@iub.edu.pk
Islamia University of Bahawalpur
Farhan Ansari
Islamia University of Bahawalpur
Sunnia Ikram
Islamia University of Bahawalpur
Nadia Khan
Islamia University of Bahawalpur
Amna Ikram
Government Sadiq College Women University
Najah Alsubaie
Princess Nourah bint Abdulrahman University
Hadia Saeed
Islamia University of Bahawalpur

Article

Keywords: Abusive language detection, Urdu and Arabic, Explainable AI (XAI), cross-lingual transfer, Integrated Gradients, faithfulness versus plausibility

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11337779/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Reliability of Urdu–Arabic Abusive Language Detection with Cross-Lingual Transformers and Explainable AI

Malik Daler Ali Awan[1*], Farhan Ansari[1], Sunnia Ikram[1], Nadia Khan[1], Amna Ikram[2], Najah Alsubaie[3], Hadia Saeed[1]

[1]Department of Software Engineering, Faculty of Computing, The Islamia University of Bahawalpur, Pakistan
[2]Department of Computer Science and Information Technology, The Government Sadiq College Women University Bahawalpur Pakistan
[3]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University (PNU), P.O. Box 84428, Riyadh 11671, Saudi Arabia.

*Corresponding author:
Malik Daler Ali Awan (Email: daler.ali@iub.edu.pk)

Abstract

Automated detection of abusive language remains challenging for Urdu and Arabic, and high classification scores do not establish that a model transfers reliably or explains its decisions faithfully. We evaluate a bilingual XLM-RoBERTa detector trained on 65,344 records, combining Integrated Gradients (IG), KernelSHAP, and an open-weight Aya-23-8B explainer with tests of cross-lingual transfer, robustness, calibration, and context sensitivity. The classifier achieved 0.821 accuracy (Urdu, 0.816; Arabic, 0.826), with mean accuracy of 0.819 ± 0.001 across three seeds. Zero-shot transfer between Urdu and Arabic remained near chance (0.50–0.53), indicating that the small in-language performance gap depended on joint bilingual training. IG and KernelSHAP showed moderate agreement (pooled Spearman ρ = 0.369). Although Aya-23 explanations echoed 69% of attributed words, deletion tests found that quoted words were not more decisive than random words, whereas IG-selected words were. Synthetic Roman-Urdu transliteration reduced Urdu accuracy by 0.218. A contrast set sharing sensitive keywords exposed context-insensitive predictions (accuracy, 0.571); context-aware fine-tuning raised contrast accuracy to 0.857 while leaving full-test accuracy nearly unchanged. These findings support evaluating multilingual abusive-language detectors through targeted reliability tests rather than aggregate accuracy alone.

Keywords: Abusive language detection; Urdu and Arabic; Explainable AI (XAI); cross-lingual transfer; Integrated Gradients; faithfulness versus plausibility

Introduction

The scale of abusive and toxic content on social platforms has outpaced human moderation, motivating automated detection. Most toxicity-detection research, however, targets English, leaving low-resource, right-to-left languages such as Urdu and Arabic under-served despite their hundreds of millions of speakers [1,2]. The harm is well documented in the region: in a survey of 426 Saudi university students, 17.6% reported cyberbullying others [3], and among young Saudi online gamers, half of those who had been cyberbullied showed depressive symptoms [4]. Following common usage [1], we use *abusive language* as an umbrella term for hate speech, offensive and toxic content, and classify single posts as abusive or not abusive; cyberbullying proper, which requires conversation histories, is outside our scope. A second gap is that most detectors are black boxes: they assign a label without justifying it, which undermines moderator trust, due process for users and any meaningful audit of bias [5,6].

Urdu abusive-language research covers native Nastaliq script, through shared tasks, hatred benchmarks, transfer learning and recurrent and transformer models [7–13], and Roman (Latin-script) Urdu, which dominates informal social media [14–18]. Arabic research is anchored by dialectal corpora and shared tasks [2,19,20] and by monolingual transformers such as AraBERT, MARBERT and SaudiBERT [21–23], with recent work on dialectal breadth [24–26] and class imbalance [27–29]. These systems are strong but monolingual, single-corpus and rarely interpretable, so they cannot show whether one model treats the two languages equally.

Multilingual encoders such as mBERT and XLM-RoBERTa learn a representation shared across languages [30–32], but cross-lingual transfer of hate-speech models is fragile and language-pair dependent [33,34]. Proposed remedies include label bootstrapping and auxiliary-task training [35,36], and for Arabic a translate-then-classify approach outperformed a jointly trained multilingual model [37]. On the explanation side, token attributions are dominated by Integrated Gradients (IG) [38] and SHapley Additive exPlanations (SHAP) [39], with LIME [40] as a perturbation

baseline, all available in Captum [41]. The two families often disagree [42,43], and benchmarks such as HateXplain and ERASER [5,44], together with recent surveys, show that faithfulness, whether an explanation reflects the model's reasoning, differs from plausibility [6,45,46]. Multilingual rationale benchmarks [47,48] and the Urdu HateInsights study [49] compare attribution methods but add no generative stage, while LLM-based explainers [50–52] produce fluent text whose faithfulness is only conditional [53,54].

No prior system combines a single bilingual Urdu–Arabic classifier, two independent attribution methods with a quantified agreement, and an open-weight generative explainer evaluated for faithfulness under one reliability protocol. We address this gap with the three-stage framework in Fig. 1 and four research questions: whether one cross-lingual model detects abusive language reliably in both languages (RQ1); whether IG and KernelSHAP agree on the words that drive a decision (RQ2); whether an open-weight generative model can turn attributions into faithful explanations (RQ3); and how robust the detector is to character noise, masking, diacritic removal and Roman-Urdu transliteration (RQ4). Our contributions are (i) a balanced, dual-script, gold-labelled Urdu–Arabic corpus of 65,344 records; (ii) a jointly trained XLM-RoBERTa detector that reaches near-parity across the two languages, with evidence that this parity comes from joint training rather than from multilingual pre-training; (iii) a quantified IG–KernelSHAP agreement for this language pair; (iv) a locally run Aya-23-8B explainer measured for faithfulness, plausibility and deletion impact; and (v) a context-sensitivity test and a reliability card that report, for each decision, a calibrated confidence, the evidence word, the attacked target and the reason. An extended review of related work is given in the Supplementary Information.

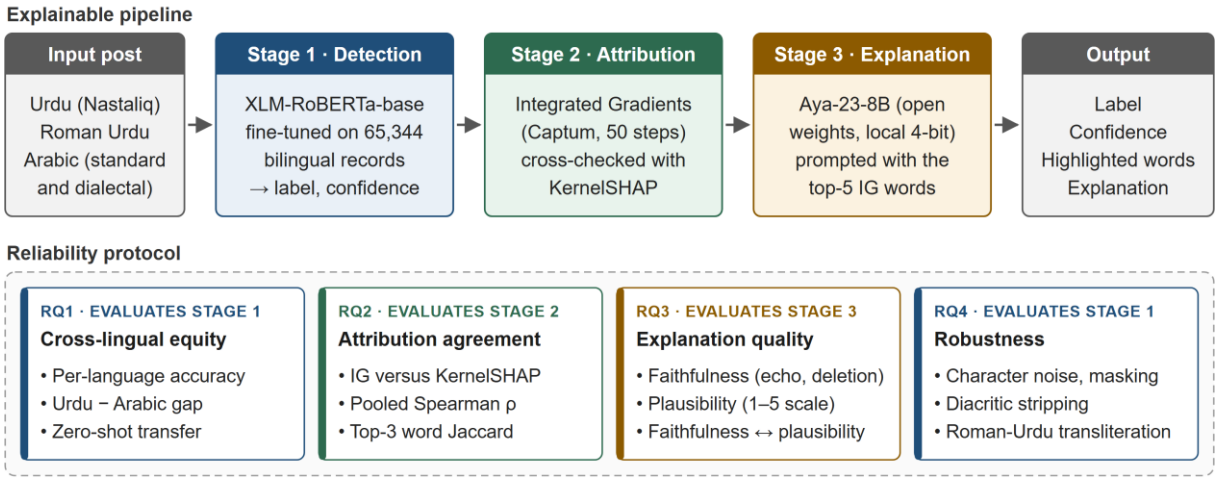


Figure 1. Overview of the proposed framework. The upper panel shows the three-stage explainable pipeline; the lower panel maps each research question to the pipeline stage it evaluates.

Results

One bilingual detector reaches near-parity across languages

On the 9,802-post held-out test set the detector attains 0.821 accuracy and 0.820 macro-F1, with an abusive-class recall of 0.866 (Supplementary Table S1). Accuracy is 0.816 on Urdu and 0.826 on Arabic, a gap of −0.010 (95% bootstrap interval −0.025 to 0.006). Retraining with three random seeds (Table 1) changes accuracy little (0.819 ± 0.001, mean ± s.d.; Matthews correlation coefficient (MCC) 0.639 ± 0.003), whereas abusive recall varies more (0.800–0.866) because the runs settle at different trade-offs between recall and specificity. The Urdu − Arabic gap is small in every run (−0.010 to −0.018), but its interval excludes zero for seeds 43 and 44, so the detector shows near-parity with a small Arabic advantage of one to two accuracy points. Missed abusive posts are the minority error (680 of 5,057 abusive test posts; Supplementary Fig. S1). Validation macro-F1 peaked at epoch 5 (0.813) while validation loss rose after epoch 2 (Supplementary Fig. S2), a sign of growing over-confidence that we address by temperature scaling below.

Table 1. XLM-RoBERTa trained with three random seeds on identical splits. Test set, n = 9,802; Urdu–Arabic accuracy gap with 95% bootstrap interval.

Seed	Accuracy	Macro-F1	Recall	Specificity	MCC	Urdu / Arabic acc.	Gap (95% CI)
42 (main model)	0.821	0.820	0.866	0.773	0.642	0.816 / 0.826	−0.010 (−0.025 to 0.006)
43	0.818	0.818	0.825	0.810	0.635	0.809 / 0.827	−0.018 (−0.032 to −0.003)
44	0.819	0.819	0.800	0.839	0.639	0.812 / 0.827	−0.015 (−0.031 to −0.000)
Mean ± SD	0.819 ± 0.001	0.819 ± 0.001	0.830 ± 0.033	0.807 ± 0.034	0.639 ± 0.003	0.812 / 0.826	−0.014

Parity depends on joint bilingual training

In-language accuracy is high (Urdu→Urdu 0.814, Arabic→Arabic 0.829), but zero-shot transfer collapses to near chance: Urdu→Arabic reaches 0.532 and Arabic→Urdu 0.504, with Arabic→Urdu abusive recall falling to 0.167 (Supplementary Table S2). The shared multilingual embedding does not by itself carry abusive-language cues across the two scripts; the near-parity reported above therefore comes from joint training on both languages, a finding that averaged accuracy would mask.

Attribution methods agree only moderately

IG and KernelSHAP agree positively but moderately, with a pooled Spearman ρ of 0.369 (mean per-sample ρ: Arabic 0.360, Urdu 0.301) and a mean top-3 Jaccard overlap of about 0.25 (Supplementary Table S3). The positive correlation supports the validity of the highlighted words, but no single attribution method should be treated as ground truth.

Generated explanations echo attributions but are only weakly faithful

The Aya-23 explainer quoted 69.0% of the IG top-5 words (faithfulness 0.690; Urdu 0.71, Arabic 0.67) across 40 explanations, all produced by the live model (Supplementary Table S4). An automated plausibility score (mean 2.2 of 5, with Aya-23 judging its own outputs) correlated only weakly with faithfulness (Spearman $\rho \approx 0.20$), so the two properties must be reported separately. The deletion test (Supplementary Table S5) is stricter. Removing the IG top-k words lowered the probability of the predicted class by 0.289, against 0.106 for the same number of random words (Wilcoxon $p = 0.023$), and keeping only the IG words preserved almost the whole prediction (sufficiency 0.065 versus 0.297; $p < 0.001$). Removing the words quoted by Aya-23 lowered it by 0.218, which did not differ significantly from random words ($p = 0.49$). Token echo therefore overstates how closely the generated text follows the detector: the explanations mix decisive words with words that matter little. All attribution and explanation estimates rest on a 40-post probe and carry correspondingly wide uncertainty.

Robustness to noise and transliteration

From a clean accuracy of 0.809, diacritic stripping cost essentially nothing (−0.001), character noise −0.029 and character masking −0.103 (Supplementary Table S6). The largest weakness is synthetic Roman-Urdu transliteration, which lowered Urdu accuracy by 0.218 (to 0.595) while leaving Arabic unaffected. Because naturally occurring Roman Urdu is already in the training data, this drop reflects the synthetic transliteration shift rather than a gap in script coverage.

Context sensitivity and calibration

A reliable detector must read a sensitive word together with its neighbours. The Urdu sentence S1, لڑکیاں تنگ لباس پہنتی ہیں اور بے حیا ہیں ("girls wear tight clothes and are shameless"), insults a group, whereas S2, ہمیں بیہودگی نہیں کرنی چاہیے ("we should not behave indecently"), is moral advice. On 49 held-out contrast sentences that share sensitive words between insults and harmless uses (Methods), the base detector reached only 0.571 accuracy: it labelled three of every four harmless sentences abusive, with a mean confidence of 0.957 on its errors (Supplementary Table S7). The training data explain why. The tokenizer splits بیہودگی into the same opening pieces as the insult بیہودہ, which is abusive in 75% of its training occurrences, and posts containing "نہیں … چاہیے" are abusive in 53% of the 107 Urdu training posts that contain them, mostly political attacks. Context-aware fine-tuning raised contrast accuracy to 0.857 without lowering test accuracy (0.823 versus 0.821; McNemar $p = 0.26$). Figure 2 shows the mechanism: in the base detector,

context pushed harmless sentences further toward abusive (mean p(abusive) 0.50 for the keyword alone, 0.77 for the full sentence), whereas in the context-aware model the words to the right of the keyword lowered it from 0.48 to 0.14, while insults stayed at or above 0.75. The word-level evidence for S1 and S2 is shown in Supplementary Fig. S3. Both models were over-confident, and temperature scaling reduced the expected calibration error about fourfold without changing any prediction.

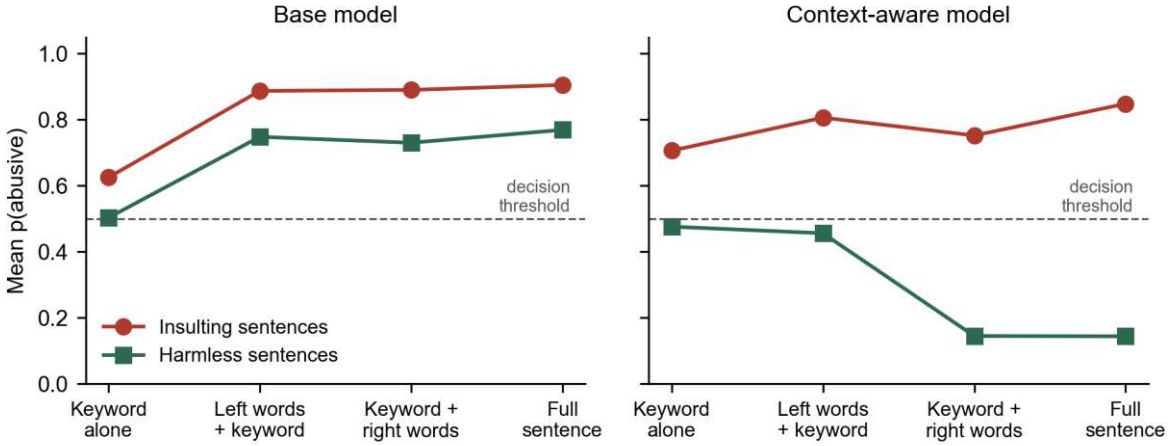


Figure 2. Effect of context on the abusive-class probability. Mean abusive-class probability as context is added around a sensitive keyword in insulting and harmless contrast sentences, for the base and context-aware detectors.

Comparison with baselines and language specialists

XLM-RoBERTa outperformed every classical baseline on accuracy, macro-F1, abusive recall, area under the ROC curve (AUC) and MCC (Supplementary Table S8). The strongest, TF-IDF with logistic regression, reached 0.796 accuracy and an MCC of 0.592, against 0.821 and 0.642, with non-overlapping confidence intervals and McNemar $p < 0.001$ for all four classical models. Multilingual BERT came closest (0.804 accuracy, MCC 0.608) but was significantly weaker for each of the three seeds ($p < 0.001$); a run capped at two epochs scored 0.811 and was also weaker than the main model ($p = 0.015$). The two XLM-RoBERTa models have the highest ROC curves (AUC 0.900 and 0.899; Fig. 3). Against single-language specialists (Table 2), MARBERTv2 was more accurate on Arabic (0.851 versus 0.826; $p < 0.001$), as expected of an encoder pretrained on about one billion Arabic tweets [22], whereas MuRIL did not outperform the bilingual model on Urdu (0.806 versus 0.816; $p = 0.08$). Specialised single-language systems report higher scores on their own benchmarks (for example, MARBERT F1 ≈ 0.92), but none covers both languages in one model or measures the reliability of its explanations (Supplementary Table S9).

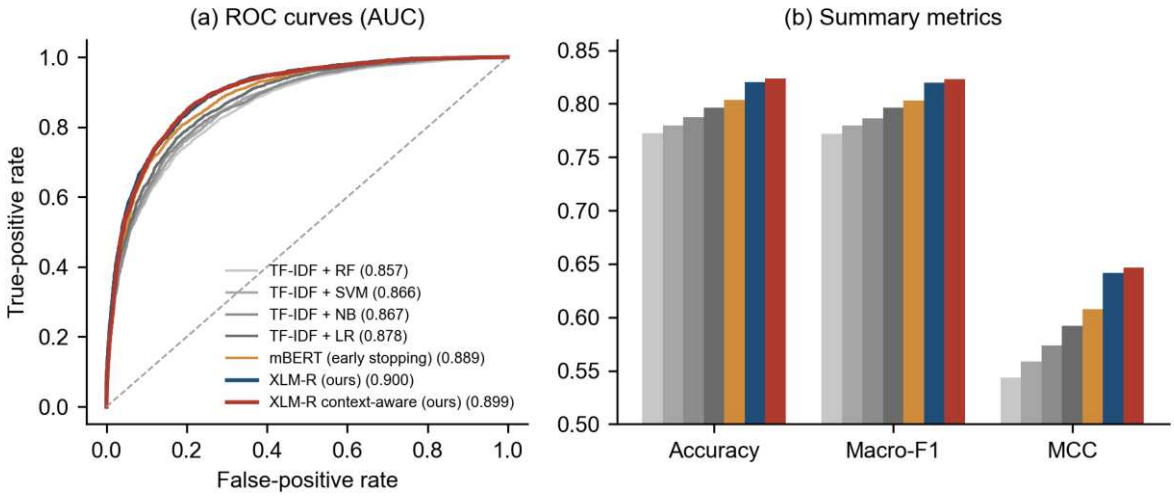


Figure 3. Comparison of all models on the identical test set. (**a**) Receiver operating characteristic curves with the area under each curve. (**b**) Accuracy, macro-F1 and Matthews correlation coefficient.

Table 2. Bilingual XLM-RoBERTa against single-language specialists on each language's test posts. Accuracy with 95% bootstrap interval; exact McNemar test against the bilingual model.

Language (test posts)	Model	Accuracy (95% CI)	Macro-F1	Recall	Specificity	AUC	MCC	McNemar p
Arabic (4,895)	MARBERTv2, Arabic only	0.851 (0.840–0.860)	0.850	0.898	0.803	0.924	0.705	< 0.001
Arabic (4,895)	XLM-RoBERTa, bilingual (ours)	0.826 (0.815–0.836)	0.825	0.859	0.793	0.904	0.653	—
Urdu (4,907)	MuRIL, Urdu only	0.806 (0.795–0.816)	0.804	0.860	0.744	0.879	0.611	0.08
Urdu (4,907)	XLM-RoBERTa, bilingual (ours)	0.816 (0.804–0.826)	0.813	0.872	0.751	0.896	0.630	—

Reliability card: target and reason

Each prediction carries a reliability card that states who is attacked and why (Fig. 4; Supplementary Table S10). On 24 held-out insults the target tagger found the annotated target in every sentence, against 0.71 for a gazetteer rule, and assigned the target type with accuracy 0.96, gender with 1.00 and topic with 0.96. For the 20 of these insults flagged as abusive, the reason category matched the category of the sentence's word family in every case; because both use the same four categories defined by us, this is a consistency check rather than an independent validation. For S1 the card reads: abusive, 92.6% calibrated confidence, target لڑکیاں (a female group), reason "modesty and morality, in the context of dress", evidence بے حیا; for S2 it reads: not abusive, no target, with the negation and advice cues that explain the decision.



Figure 4. Reliability cards for the example sentences S1 and S2. Each card shows the decision, calibrated confidence, evidence word, target, reason and context.

Discussion

A single explainable model can moderate abusive language in Urdu and Arabic at near-parity, provided it is trained jointly on both languages. The zero-shot collapse shows that this parity is earned by balanced joint training rather than granted by the multilingual backbone, which agrees with evidence that cross-lingual transfer of hate-speech models is fragile [33,34] and supports joint or auxiliary-task training as a remedy [35,36]. The bilingual model matches an Urdu specialist and trails the strongest Arabic specialist by 2.5 accuracy points, while serving both languages with one model and one explanation pipeline.

Explanations are useful but imperfect. IG–KernelSHAP agreement is moderate, and although Aya-23 echoes 69% of the attributed words, only the IG words, not the words the explanations quote, are more decisive than random words. This mirrors evidence that LLM-generated explanations are only conditionally faithful [53,54], and it supports grounding the generator in attribution evidence and presenting its text as decision support rather than as the model's reasoning. The contrast set exposes a failure that aggregate accuracy hides: the base detector judged sensitive words without their context, and a small context-aware fine-tuning step corrected most of these errors at no cost on the test set. The residual transliteration weakness echoes the dialect- and script-robustness concerns of recent Arabic and Roman-Urdu work [17,25,26].

Several limitations apply. Results come from one stratified split, with run-to-run variation estimated from three seeds; the Arabic portion has limited dialectal breadth; the attribution and explanation analyses rest on a 40-post probe; Aya-23 does not officially cover Urdu, and plausibility was scored by Aya-23 itself; the contrast set and reliability card

rest on a small, AI-assisted and self-annotated Urdu set that does not yet cover Arabic or Roman Urdu; the robustness perturbations are deterministic approximations; and the classical baselines were not tuned, while AraBERT and a dedicated Urdu encoder were not evaluated (see Supplementary Discussion). Automated moderation is also dual-use: public corpora can encode regional and demographic bias, and recall-first detection raises false positives, so the system is designed to support a human moderator with an explanation and an appeal path rather than to act autonomously.

Future work should add natural and back-translated Roman Urdu to training, broaden Arabic dialect coverage with k-fold evaluation, compare generative explainers that officially cover Urdu, extend the contrast set to Arabic and Roman Urdu with native-speaker review, and test the explanation layer with calibrated confidence in a live moderation workflow. More broadly, our results argue for evaluating multilingual abusive-language detectors through targeted reliability tests rather than aggregate accuracy alone.

Methods

Bilingual corpus

We assembled a balanced bilingual corpus of 65,344 records with a binary target (non-abusive = 0, abusive = 1) under a unified text, label, language and script schema (Supplementary Table S11). The Urdu portion (32,711) combines 24,719 Nastaliq records consolidated from public Urdu toxicity and hate-speech corpora with 7,992 Roman-Urdu records from the RUHSOLD corpus [15]. The Arabic portion (32,633) merges L-HSAB [19], the Arabic subset of MLMA [55], an annotated Arabic offensive-comment set and a sample of the Arabic hate-speech superset released with the dataset survey of Tonneau et al. [56]. Native labels (hate, offensive, abusive or toxic) were mapped to the binary target, and duplicate texts were removed within and across sources before splitting. The corpus was stratified 70/15/15 on language × label (train 45,740, validation 9,802, test 9,802; seed 42), and the test set was never used for training or model selection. Only publicly available, de-identified social-media text from published research corpora was used; no human participants were recruited and no personal data were collected.

Classification

We fine-tuned *xlm-roberta-base* [32] for binary classification with the Hugging Face Transformers library [57], using AdamW (learning rate 2×10^{-5}, warmup ratio 0.1, weight decay 0.01), a batch size of 16, 16-bit mixed precision, a maximum length of 128 tokens and up to 10 epochs with early stopping on validation macro-F1 (patience 1; training stopped after epoch 6 and the epoch-5 checkpoint was restored), on one NVIDIA GeForce RTX 3060 GPU (12 GB). Abusive-class recall was prioritised because false negatives let harmful content through moderation. The settings and computational costs of all fine-tuned models are listed in Supplementary Table S12.

Attribution and generative explanation

For each prediction we computed IG attributions [38] for the abusive-class logit over the input embeddings with Captum [41] (50 steps from a pad-token baseline), merged sub-word scores to word level and cross-checked them with KernelSHAP [39] on the same samples. Agreement was measured by the pooled Spearman correlation over shared word–attribution pairs and by top-k Jaccard overlap. The five words pushing most strongly toward and against the predicted label, together with the label and its confidence, were passed to Aya-23-8B [58,59], an open-weight model that covers 23 languages including Arabic but not, officially, Urdu. Aya-23 ran locally with 4-bit NF4 quantisation and greedy decoding (at most 170 new tokens, repetition penalty 1.2), so no paid application programming interface was used. The prompt asked for two to three plain-English sentences that explain the label using only the supplied evidence, quote every driving word verbatim and note any Urdu or Arabic sociolinguistic nuance.

Reliability protocol

Cross-lingual equity (RQ1) was assessed by per-language metrics, the Urdu − Arabic gap and zero-shot transfer, for which separate XLM-RoBERTa models were fine-tuned on one language (3 epochs) and evaluated on both. Attribution agreement (RQ2) and explanation quality (RQ3) were measured on a probe of 40 test posts stratified by language and label (10 per cell). Faithfulness is the fraction of the IG top-5 words quoted in the explanation, and plausibility is an automated score on a 1–5 scale. Following ERASER [44], the probability of the predicted class was recomputed after removing (comprehensiveness) or keeping only (sufficiency) the k quoted words, the IG top-k words or k random words (20 draws), with k equal to the number of quoted words (at most 5). Robustness (RQ4) was measured on 1,200 random test posts (seed 42; posts of four characters or fewer excluded) under character masking (each non-space

character replaced with probability 0.20), adjacent-character swaps (probability 0.15 in words longer than three characters), removal of diacritics and tatweel, and a phonetic Roman-Urdu transliteration of the Urdu rows.

Context sensitivity, calibration and reliability card

The contrast set contains 133 Urdu sentences, drafted with the assistance of a large language model and labelled by the authors, in which each of 22 sensitive-word families appears both in targeted insults and in non-abusive uses (advice, reporting, negation, general or self-referential statements). Fourteen families (84 sentences) were used for training and eight unseen families (49 sentences, including S1 and S2) for testing. The context-aware detector was obtained by training the base detector for one further epoch (learning rate 1×10^{-5}) on the full training set plus the 84 training sentences repeated 30 times (about 5% of the mixture). Calibration was measured by the expected calibration error (15 bins) before and after temperature scaling fitted on the validation set. For the reliability card, the target is found by a token-level logistic-regression tagger over the word, its neighbours, a pronoun list and a gazetteer of Urdu person and group nouns, trained on the 42 annotated insults of the contrast-set training split and tested on the 24 held-out insults; the reason is the attack category (modesty and morality, intelligence and competence, honesty and loyalty, or degradation and contempt) of the strongest IG evidence word, combined with any topic cue in the sentence.

Baselines and statistical testing

We trained four classical baselines on the identical splits, using TF-IDF word 1–2-grams and character 2–5-grams (300,000 features) with a linear support vector machine, logistic regression, a random forest (300 trees) and multinomial naive Bayes, and fine-tuned multilingual BERT [30] with the XLM-RoBERTa protocol, plus a run capped at two epochs. XLM-RoBERTa was retrained with seeds 43 and 44 to estimate run-to-run variation. As single-language specialists we fine-tuned MARBERTv2 [22] on the Arabic portion and MuRIL [60] on the Urdu portion. All models were evaluated on the same 9,802 test posts with accuracy, macro-F1, precision, recall, specificity, AUC and MCC; 95% confidence intervals come from 1,000 bootstrap resamples, and each comparison with XLM-RoBERTa uses an exact McNemar test on the posts where the two models disagree. The complete procedure is summarised in Supplementary Table S13, and further implementation details are given in the Supplementary Methods.

Use of generative AI tools

Aya-23-8B is part of the method. In addition, during the preparation of this manuscript the authors used Claude (Anthropic) to assist with drafting, language editing, verification of references against their primary sources, writing and checking analysis code, and drafting the Urdu contrast sentences, whose labels were reviewed by the authors. The authors reviewed and edited all content and take full responsibility for the publication.

Data availability

The datasets analysed during the current study are publicly available from the sources cited in the Methods. The consolidated bilingual splits, the Urdu contrast set with its target annotations, the fine-tuned models and the code for training, attribution, explanation and evaluation are available from the corresponding author on reasonable request for non-commercial research use.

References

1. Schmidt, A. & Wiegand, M. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* 1–10 (Association for Computational Linguistics, 2017). https://doi.org/10.18653/v1/W17-1101
2. Mubarak, H., Darwish, K. & Magdy, W. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online* 52–56 (Association for Computational Linguistics, 2017). https://doi.org/10.18653/v1/W17-3008
3. Al Qudah, M. F. et al. Psychological security, psychological loneliness, and age as the predictors of cyber-bullying among university students. *Community Ment. Health J.* **56**, 393–403 (2020). https://doi.org/10.1007/s10597-019-00455-z
4. AlJaffer, M. et al. Cyberbullying among young Saudi online gamers and its relation to depression. *J. Nature Sci. Med.* **4**, 142–147 (2021). https://doi.org/10.4103/jnsm.jnsm_78_20
5. Mathew, B. et al. HateXplain: A benchmark dataset for explainable hate speech detection. *Proc. AAAI Conf. Artif. Intell.* **35**, 14867–14875 (2021). https://doi.org/10.1609/aaai.v35i17.17745

6. Jacovi, A. & Goldberg, Y. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 4198–4205 (Association for Computational Linguistics, 2020). https://doi.org/10.18653/v1/2020.acl-main.386
7. Amjad, M. et al. Overview of abusive and threatening language detection in Urdu at FIRE 2021. In *Working Notes of FIRE 2021 – Forum for Information Retrieval Evaluation* (CEUR Workshop Proc., vol. 3159) 744–762 (CEUR-WS.org, 2021). https://ceur-ws.org/Vol-3159/T4-1.pdf
8. Das, M., Banerjee, S. & Saha, P. Abusive and threatening language detection in Urdu using boosting based and BERT based models: A comparative approach. In *Working Notes of FIRE 2021 – Forum for Information Retrieval Evaluation* (CEUR Workshop Proc., vol. 3159) 791–798 (CEUR-WS.org, 2021). https://ceur-ws.org/Vol-3159/T4-5.pdf
9. Akram, M. H., Shahzad, K. & Bashir, M. ISE-Hate: A benchmark corpus for inter-faith, sectarian, and ethnic hatred detection on social media in Urdu. *Inf. Process. Manag.* **60**, 103270 (2023). https://doi.org/10.1016/j.ipm.2023.103270
10. Malik, M. S. I., Nawaz, A. & Jamjoom, M. M. Hate speech and target community detection in Nastaliq Urdu using transfer learning techniques. *IEEE Access* **12**, 116875–116890 (2024). https://doi.org/10.1109/ACCESS.2024.3444188
11. Shoukat, E., Irfan, R., Basharat, I., Tahir, M. A. & Shaukat, S. Attention based bidirectional GRU hybrid model for inappropriate content detection in Urdu language. Preprint at https://doi.org/10.48550/arXiv.2501.09722 (2025).
12. Rashid, A., Mahmood, S., Inayat, U. & Zia, M. F. Urdu toxicity detection: A multi-stage and multi-label classification approach. *AI* **6**, 194 (2025). https://doi.org/10.3390/ai6080194
13. Malik, M. S. I., Cheema, U. & Ignatov, D. I. Contextual embeddings based on fine-tuned Urdu-BERT for Urdu threatening content and target identification. *J. King Saud Univ. Comput. Inf. Sci.* **35**, 101606 (2023). https://doi.org/10.1016/j.jksuci.2023.101606
14. Akhter, M. P., Zheng, J., Naqvi, I. R., Abdelmajeed, M. & Sadiq, M. T. Automatic detection of offensive language for Urdu and Roman Urdu. *IEEE Access* **8**, 91213–91226 (2020). https://doi.org/10.1109/ACCESS.2020.2994950
15. Rizwan, H., Shakeel, M. H. & Karim, A. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 2512–2522 (Association for Computational Linguistics, 2020). https://doi.org/10.18653/v1/2020.emnlp-main.197
16. Bilal, M., Khan, A., Jan, S., Musa, S. & Ali, S. Roman Urdu hate speech detection using transformer-based model for cyber security applications. *Sensors* **23**, 3909 (2023). https://doi.org/10.3390/s23083909
17. Ashiq, W. et al. Roman Urdu hate speech detection using hybrid machine learning models and hyperparameter optimization. *Sci. Rep.* **14**, 28590 (2024). https://doi.org/10.1038/s41598-024-79106-7
18. Hussain, N. et al. Fine-tuning large language models with QLoRA for offensive language detection in Roman Urdu–English code-mixed text. Preprint at https://doi.org/10.48550/arXiv.2510.03683 (2025).
19. Mulki, H., Haddad, H., Bechikh Ali, C. & Alshabani, H. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online* 111–118 (Association for Computational Linguistics, 2019). https://doi.org/10.18653/v1/W19-3512
20. Husain, F. OSACT4 shared task on offensive language detection: Intensive preprocessing-based approach. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools* 53–60 (European Language Resources Association, 2020). https://aclanthology.org/2020.osact-1.8/
21. Antoun, W., Baly, F. & Hajj, H. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools* 9–15 (European Language Resources Association, 2020). https://aclanthology.org/2020.osact-1.2/
22. Abdul-Mageed, M., Elmadany, A. & Nagoudi, E. M. B. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* 7088–7105 (Association for Computational Linguistics, 2021). https://doi.org/10.18653/v1/2021.acl-long.551
23. Qarah, F. SaudiBERT: A large language model pretrained on Saudi dialect corpora. Preprint at https://doi.org/10.48550/arXiv.2405.06239 (2024).
24. Khezzar, R., Moursi, A. & Al Aghbari, Z. arHateDetector: Detection of hate speech from standard and dialectal Arabic tweets. *Discov. Internet Things* **3**, 1 (2023). https://doi.org/10.1007/s43926-023-00030-9
25. Charfi, A., Besghaier, M., Akasheh, R., Atalla, A. & Zaghouani, W. Hate speech detection with ADHAR: A multi-dialectal hate speech corpus in Arabic. *Front. Artif. Intell.* **7**, 1391472 (2024). https://doi.org/10.3389/frai.2024.1391472
26. Abdelsamie, M. M., Azab, S. S. & Hefny, H. A. The dialects gap: A multi-task learning approach for enhancing hate speech detection in Arabic dialects. *Expert Syst. Appl.* **295**, 128584 (2026). https://doi.org/10.1016/j.eswa.2025.128584
27. Alzaqebah, M. et al. Cyberbullying detection framework for short and imbalanced Arabic datasets. *J. King Saud Univ. Comput. Inf. Sci.* **35**, 101652 (2023). https://doi.org/10.1016/j.jksuci.2023.101652

28. Aljohani, A., Alharbe, N., Al Mamlook, R. E. & Khayyat, M. M. A hybrid combination of CNN attention with optimized random forest with grey wolf optimizer to discriminate between Arabic hateful, abusive tweets. *J. King Saud Univ. Comput. Inf. Sci.* **36**, 101961 (2024). https://doi.org/10.1016/j.jksuci.2024.101961
29. Itriq, M. & Mohd Noor, M. H. TE-SFMTL: Enhanced multi-task learning framework with triple-encoder for extremely imbalanced Arabic toxicity detection. *J. King Saud Univ. Comput. Inf. Sci.* **38**, 484 (2026). https://doi.org/10.1007/s44443-026-00756-w
30. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 4171–4186 (Association for Computational Linguistics, 2019). https://doi.org/10.18653/v1/N19-1423
31. Pires, T., Schlinger, E. & Garrette, D. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 4996–5001 (Association for Computational Linguistics, 2019). https://doi.org/10.18653/v1/P19-1493
32. Conneau, A. et al. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 8440–8451 (Association for Computational Linguistics, 2020). https://doi.org/10.18653/v1/2020.acl-main.747
33. Jiang, A. & Zubiaga, A. Cross-lingual offensive language detection: A systematic review of dataset, approach and challenge. *ACM Comput. Surv.* **58**, 1–37 (2026). https://doi.org/10.1145/3801736
34. Eronen, J., Ptaszynski, M., Masui, F., Arata, M., Leliwa, G. & Wroczynski, M. Transfer language selection for zero-shot cross-lingual abusive language detection. *Inf. Process. Manag.* **59**, 102981 (2022). https://doi.org/10.1016/j.ipm.2022.102981
35. Bigoulaeva, I., Hangya, V., Gurevych, I. & Fraser, A. Label modification and bootstrapping for zero-shot cross-lingual hate speech detection. *Lang. Resour. Eval.* **57**, 1515–1546 (2023). https://doi.org/10.1007/s10579-023-09637-4
36. Montariol, S., Riabi, A. & Seddah, D. Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022* 347–363 (Association for Computational Linguistics, 2022). https://doi.org/10.18653/v1/2022.findings-aacl.33
37. El-Alami, F., Ouatik El Alaoui, S. & En Nahnahi, N. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *J. King Saud Univ. Comput. Inf. Sci.* **34**, 6048–6056 (2022). https://doi.org/10.1016/j.jksuci.2021.07.013
38. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 3319–3328 (PMLR, 2017). https://proceedings.mlr.press/v70/sundararajan17a.html
39. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30** (2017). https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
40. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (Association for Computing Machinery, 2016). https://doi.org/10.1145/2939672.2939778
41. Kokhlikyan, N. et al. Captum: A unified and generic model interpretability library for PyTorch. Preprint at https://doi.org/10.48550/arXiv.2009.07896 (2020).
42. Feng, T., Zhou, Z., Joshi, T. & Nair, V. N. Comparing baseline Shapley and integrated gradients for local explanation: Some additional insights. Preprint at https://doi.org/10.48550/arXiv.2208.06096 (2022).
43. Dhara, T. & Sheth, S. Beyond accuracy: An explainability-driven analysis of harmful content detection. Preprint at https://doi.org/10.48550/arXiv.2603.18015 (2026).
44. DeYoung, J. et al. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 4443–4458 (Association for Computational Linguistics, 2020). https://doi.org/10.18653/v1/2020.acl-main.408
45. Lyu, Q., Apidianaki, M. & Callison-Burch, C. Towards faithful model explanation in NLP: A survey. *Comput. Linguist.* **50**, 657–723 (2024). https://doi.org/10.1162/coli_a_00511
46. Muscato, B., Chen, B., Gezici, G., Plank, B. & Giannotti, F. Disagreeing rationales: Rethinking classification and explainability evaluation in hate speech detection. Preprint at https://doi.org/10.48550/arXiv.2605.31563 (2026).
47. Rehman, M. Z. U. et al. X-MuTeST: A multilingual benchmark for explainable hate speech detection and a novel LLM-consulted explanation framework. Preprint at https://doi.org/10.48550/arXiv.2601.03194 (2026).
48. Nirmal, A., Bhattacharjee, A., Sheth, P. & Liu, H. Towards interpretable hate speech detection using large language model-extracted rationales. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)* 223–233 (Association for Computational Linguistics, 2024). https://doi.org/10.18653/v1/2024.woah-1.17

49. Arshad, M. U. & Shahzad, W. Understanding hate speech: The HateInsights dataset and model interpretability. *PeerJ Comput. Sci.* **10**, e2372 (2024). https://doi.org/10.7717/peerj-cs.2372
50. Damo, G., Ocampo, N. B., Cabrio, E. & Villata, S. Unveiling the hate: Generating faithful and plausible explanations for implicit and subtle hate speech detection. In *Natural Language Processing and Information Systems* (Lecture Notes in Computer Science) 211–225 (Springer, 2024). https://doi.org/10.1007/978-3-031-70239-6_15
51. Nghiem, H. & Daumé III, H. HateCOT: An explanation-enhanced dataset for generalizable offensive speech detection via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024* 5938–5956 (Association for Computational Linguistics, 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.343
52. Inan, H. et al. Llama Guard: LLM-based input-output safeguard for human–AI conversations. Preprint at https://doi.org/10.48550/arXiv.2312.06674 (2023).
53. Madsen, A., Chandar, S. & Reddy, S. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024* 295–337 (Association for Computational Linguistics, 2024). https://doi.org/10.18653/v1/2024.findings-acl.19
54. Agarwal, C., Tanneru, S. H. & Lakkaraju, H. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. Preprint at https://doi.org/10.48550/arXiv.2402.04614 (2024).
55. Ousidhoum, N., Lin, Z., Zhang, H., Song, Y. & Yeung, D.-Y. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 4675–4684 (Association for Computational Linguistics, 2019). https://doi.org/10.18653/v1/D19-1474
56. Tonneau, M. et al. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)* 283–311 (Association for Computational Linguistics, 2024). https://doi.org/10.18653/v1/2024.woah-1.23
57. Wolf, T. et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 (Association for Computational Linguistics, 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6
58. Aryabumi, V. et al. Aya 23: Open weight releases to further multilingual progress. Preprint at https://doi.org/10.48550/arXiv.2405.15032 (2024).
59. Üstün, A. et al. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 15894–15939 (Association for Computational Linguistics, 2024). https://doi.org/10.18653/v1/2024.acl-long.845
60. Khanuja, S. et al. MuRIL: Multilingual representations for Indian languages. Preprint at https://doi.org/10.48550/arXiv.2103.10730 (2021).

Funding

This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R321), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author contributions

M.D.A.A.: conceptualization, methodology, supervision, project administration, writing (review and editing). F.A.: data curation, software, investigation, formal analysis, writing (original draft). S.I.: methodology, investigation, validation. N.K.: resources, visualization, data curation. A.I.: investigation, validation, writing (review and editing). N.A.: funding acquisition, resources, validation, writing (review and editing). H.S.: software, visualization, data curation. All authors reviewed the manuscript.

Additional information

Competing interests: The authors declare no competing interests.

Correspondence and requests for materials should be addressed to M.D.A.A.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ScientificReportsSupplementaryInformation.docx