Towards Explainable Hate Speech Detection in Roman Urdu

Ubaid Azam[1]*, Hammad Rizwan[2]*, Faizad Ullah[3], Ali Faheem[3], Faisal Kamiran[4], Asim Karim[3]
[1]University of Southampton, UK [2]Dalhousie University, Canada
[3]Lahore University of Management Sciences, Pakistan [4]Information Technology University, Pakistan
u.azam@soton.ac.uk, hammad.rizwan@dal.ca, faisal.kamiran@itu.edu.pk
{20030057, ali.faheem, akarim}@lums.edu.pk

***Abstract*—Automatic hate speech detection in low-resource languages remains limited to basic classification, while fine-grained tasks such as span detection and intensity classification are largely confined to English. Moreover, temporal generalization and model interpretability, both critical for real-world deployment, remain underexplored. This work presents a methodology for fine-grained, robust, and interpretable hate speech detection in Roman Urdu. We introduce RUHSOLD++, an enriched dataset featuring annotations for hate intensity and hateful span detection, enabling interpretability evaluation through plausibility and faithfulness metrics. Using multilingual transformer models in a multitask framework, we assess temporal generalizability via a time-skip dataset and robustness via adversarial word variations. Our results demonstrate that while interpretability metrics enhance prediction reliability, model performance degrades significantly under temporal distribution shifts and linguistic variations, underscoring the need for periodic retraining or improved training paradigms.**

***Index Terms*—Roman Urdu; Hate Speech; Fine-grained Classification; Explainable AI**

I. Introduction

Social media platforms, while beneficial, are often used to spread hateful content targeting users based on politics, race, and religion. To combat this, platforms rely on machine learning models trained on annotated datasets in various languages [1]. These datasets have evolved from basic classification to fine-grained tasks such as hate span detection [2], with transformer models achieving state-of-the-art results [3], [4].

However, deep learning models remain black boxes [5]. In hate speech detection, justifying decisions is essential for trust and reliability. Recent work addresses this through interpretable models that predict hateful spans using attention extraction or token classification compared to hand-labeled annotations. Some research incorporates real-world context using user history and social network links [6], achieving interpretability through surrogate models [7] or marginal contribution analysis [8]. Additionally, temporal generalizability is rarely evaluated, most datasets cover only 1–3 months, ignoring long-term changes from real-world events. Many works use stratified rather than chronological sampling, masking temporal effects.

While fine-grained, robust, and interpretable hate speech detection exists for some languages, no work addresses Roman Urdu (RU). RU uses Latin script for Urdu and is primarily used on social media by Pakistani users. Like Roman Hindi and Roman Bengali, RU presents NLP challenges due to colloquial language, improper grammar, spelling variations, and code-switching with English. Pakistan has experienced a surge in online hate speech driven by political and religious extremism and gender prejudice, prompting the government to form a committee to address it[1]. Despite legislation criminalizing such speech, investigations and convictions are lengthy and unsustainable. An automatic, explainable detection system is essential for prompt action by social media platforms and government agencies.

This work advances interpretable and robust hate speech detection in Roman Urdu through the following contributions. First, we provide enhanced data resources: the dataset from [9] annotated for intensity and hateful spans; a supplementary set of 2,005 tweets collected one year after the original dataset to evaluate temporal generalizability; and a word variation dataset to assess robustness against adversarial prompts. Second, we evaluate multilingual transformer models for Roman Urdu in a multitask setup predicting hate type, intensity, and spans, finding larger models perform best with smaller models only slightly behind. Third, we analyze interpretability by examining how hateful spans affect hate type predictions via class probability changes. Finally, we test robustness against adversarial word augmentations, showing that linguistic modifications notably reduce performance across all tasks. The dataset is publicly available at https://github.com/hammadrizwan/RUHSOLD-.git.

II. Related Work

Related work is divided into two subsections: hate speech detection models and datasets, and NLP model interpretability.

A. Datasets and Models

Early work classified obscene content with labels for racism, sexism [10], [11], religious hate [12], cyberbullying [13], or differentiating hate and offensive content [14]. Research progressed to target identification (group or individual) [15], abuse type (direct or indirect) [16], and multilingual tasks [17], [18].

*Equal contribution.

Funded by the University of Southampton (grant no. 522886110).

[1]https://thestandard.com.pk/paigham-e-aman-committee-to-counter-extremism-hate-speech/

Early modelling used POS tags, TF-IDF, emotion lexicon, and n-grams with Logistic Regression, Naive Bayes, SVM, Decision Tree, and Random Forests [14]. Deep learning shifted focus to LSTM and transformers. Syntactic and n-gram features have been fused with deep learning [11]. Recent work uses data augmentation [19], multitask modelling [20], and prompt-based techniques [21].

1) Existing Urdu Datasets: The most prominent existing Urdu datasets for hate speech and toxicity detection include RUHSOLD [9], a multiclass Roman Urdu dataset with 10,012 instances; RUT [22], comprising 72,000 samples for binary toxicity detection; a binary-labeled offensive YouTube comments dataset [23] with 147,000 instances; and an Urdu Twitter dataset [24] with 10,526 instances. More recently, CRU [25] was released with 7,342 instances annotated for cybercrime-related content, and the Urdu Toxicity Corpus (UTC) [26] was introduced as a multilabel dataset of 18,000 instances.

B. Interpretability

As models improved, opacity increased. GDPR's "right of explanation" shifted focus toward interpretable NLP [27]. Interpretability approaches are categorized as local versus global, and as post hoc versus self-explaining [28]. Interpretability is defined through "plausibility" (human convincingness) and "faithfulness" (accurate reasoning reflection) [29].

For faithfulness with text-only inputs, erasure methods remove text portions to assess class probability changes via attention weights [30], [31]. Recent work integrates prompt engineering, few-shot learning, and fine-tuning for deeper explanations [32].

For plausibility, multitask modelling has been employed for toxicity and span detection, comparing transformers with logistic regression using softer metrics for partial credit [33]. Span extraction methods have been applied to long texts [34]. Multi-label detection with target groups and explanatory spans has been explored [35].

Prior work using span-based and multitask approaches is largely limited to high-resource languages. We propose a multitask framework that jointly models span detection, intensity prediction, and hate classification, improving the explainability of hate speech detection in Roman Urdu and supporting legal and regulatory use cases. To the best of our knowledge, this is the first work to address this problem setting for this under-resourced language.

III. Methodology

Our methodology for robust and interpretable hate speech detection in Roman Urdu includes a new dataset, additional prediction tasks, multilingual deep learning models, and comprehensive evaluation.

We develop RUHSOLD++ by enhancing RUHSOLD [9] with annotations for hate intensity classification and hateful span detection. These tasks enable detailed analysis by identifying words contributing to hate classification (rationales), facilitating interpretation and plausibility of predictions. The dataset also includes over 2,000 annotated tweets collected one year after RUHSOLD's release and a word variation dataset simulating adversarial prompts to test robustness. RUHSOLD was selected as foundation because it addresses multiple hate dimensions and types rather than binary classification, essential for interpretability analysis.

TABLE I
Distribution of Tweets and their labels in RUHSOLD++

Label	RUHSOLD	New Tweets	RUHSOLD++
Abusive/Offensive	2402	502	2,904
Sexism	839	187	1,026
Religious hate	782	166	948
Profane	640	180	820
Normal	5349	970	6,319
Total	10,012	2,005	12,017

We fine-tune multilingual models of varying size and pre-training in a multitask setup to detect implicit and explicit abusive language in Roman Urdu, evaluating on random stratified and temporal splits. Figure 1 illustrates our methodology.

A. RUHSOLD++ Dataset

RUHSOLD++ extends RUHSOLD [9] into a multi-task benchmark for Roman Urdu hate speech analysis, supporting robustness, interpretability, and fine-grained evaluation. It increases annotation depth and data coverage, enabling hate category prediction, hate intensity classification, and hateful span detection.

RUHSOLD contains 10,012 Roman Urdu tweets labeled into five classes: Abusive/Offensive, Sexism, Religious Hate, Profane, and Normal. While it captures multiple hate dimensions beyond binary classification, it lacks supervision for interpretability and robustness.

RUHSOLD++ addresses these limitations by (i) adding hate intensity and hateful span annotations to all tweets, enabling word-level rationales; (ii) introducing 2,005 newly collected and manually annotated tweets gathered one year after RUHSOLD's release for temporal robustness analysis; and (iii) including a word-variation dataset simulating adversarial spelling and lexical variations common in Roman Urdu hate speech.

New tweets were collected using the same lexicon-based protocol as RUHSOLD and filtered to remove retweets, non–Roman Urdu content, and very short texts. To reduce user-level bias [36], contributions were capped at 50 tweets per user. After resolving annotation disagreements, 2,005 tweets were selected via stratified sampling from 2,500 candidates to match the original class distribution.

The final RUHSOLD++ dataset contains 12,017 Roman Urdu tweets with balanced class proportions. Table I summarizes the class-wise distribution.

1) Annotation Procedure: Manual annotation was conducted for four tasks: (i) hateful span identification (zero or more contiguous word sequences expressing hate), (ii) hate intensity, (iii) hate class labeling for newly collected tweets, and (iv) generation of word-level spelling variations of hateful span vocabulary absent from the RUHSOLD lexicon.

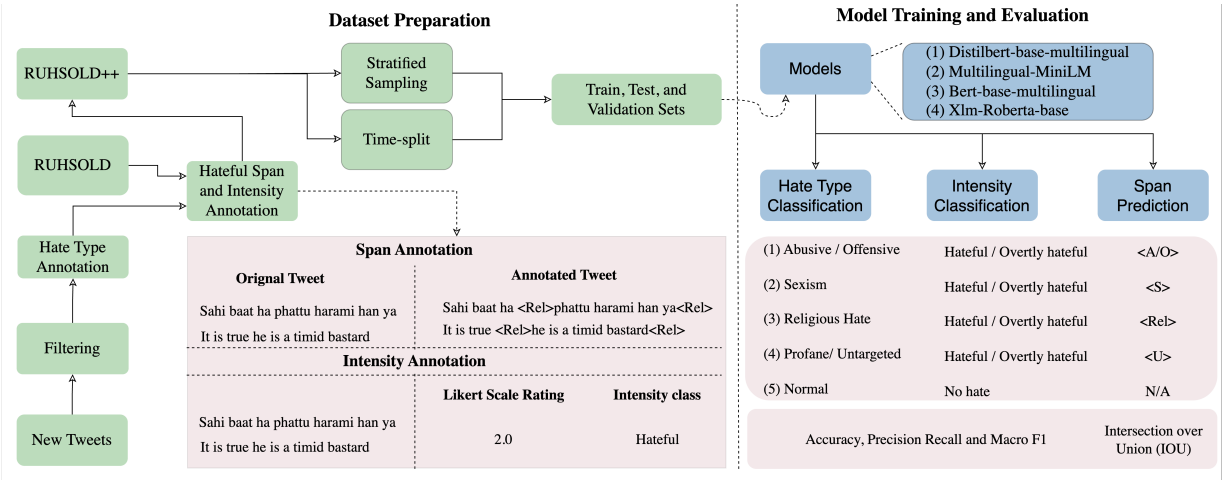

Fig. 1. Pipeline for robust and interpretable Roman Urdu hate speech detection

TABLE II
SAMPLES OF TWEETS FOR EACH LABEL FROM RUHSOLD++ DATASET

Tweet	Translation	Type	Intensity
Ye < A/O >bc<A/O> peeke kuch b likhta hai <A/O>charsi hai ek number<A/O>	this<A/O>sisterfucker<A/O> write anything after drinking <A/O>he is the number one druggy<A/O>	Abusive /Offensive	Overtly Hateful
<S>Vo toh hai hei sasti randi<S> aur koi bata	<S>that is a cheap whore<S> tell someone else	Sexist	Hateful
Sahi baat ha <Rel>phattu harami han ya<Rel>	It is true <Rel>he is a timid bastard<Rel>	Religious	Hateful
<U>Bc<U> ab kia keh dia me	<U>Sisterfucker<U> what have I said now	Untargeted	Hateful
Us sy acha to ny ga lana tha	you could have sung better than her	Neutral	No Hate

Annotations were performed by three native Urdu speakers who are active Roman Urdu users and familiar with hate speech nuances. Annotators were provided with written guidelines, illustrative examples, and training to reduce bias. A tweet label was retained only when at least two annotators agreed on the hate class. Sample annotated tweets with English translations, hate categories, and intensity labels are shown in Table II.

For hateful span annotation, annotators marked words directly contributing to hateful content. Hate intensity was initially rated on a 0–5 Likert scale (0: no hate, 5: highest intensity), and hate type was assigned using the five RUHSOLD classes.

Inter-annotator agreement was measured using Fleiss's Kappa:

$$k = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_o denotes observed agreement and p_e chance agreement. Substantial agreement was achieved for hateful spans and hate type ($k = 0.67$). Intensity annotation yielded low agreement ($k = 0.22$), primarily due to confusion between adjacent levels (1–3 and 4–5). To address this, intensity labels were merged into three classes: *neutral* (0), *hateful* (1–3), and *overtly hateful* (4–5), increasing agreement to $k = 0.82$ (almost perfect).

For robustness evaluation, annotators additionally created a spelling variation dataset, producing an average of 3–4 variants per hateful vocabulary term. Only variants absent from both the RUHSOLD++ lexicon and dataset were retained. Example variations are provided in Table III.

TABLE III
SAMPLE OF VARIATION OF WORDS

Word	Variations
samjaya	[samjhaya, samjaeya, smjaya, samjaia, smjaya, smjaia]
kuttya	[kutia, kutiaa, kutia, kuttiya]
tareqah	[tareqa, tareeka, tareka, treka, tariqa, tareeqa, tarikah]
chudwai	[chudwae, chodwai, chodwaiyee, cudweed, chudwaye]
jahnum	[jahannum, jahannom, jahanum]

B. Tasks and Evaluation Metrics

1) Hate Type and Hate Intensity Classification: Hate type classification is formulated as a 5-class task, while hate intensity is modeled as a 3-class classification problem. Performance is evaluated using precision, recall, accuracy, and macro F1 score. Macro F1 is reported for primary comparison due to its robustness to class imbalance.

2) Hate Span Detection and Interpretability: Hate span detection is framed as a sequence labeling task, where each token is assigned a B (beginning), I (inside), or O (outside) label. The detected spans serve as rationales for interpreting both hate type and hate intensity predictions. Span detection performance is measured using precision, recall, and F1 score.

Interpretability assessed via plausibility and faithfulness. Multitask learning evaluates plausibility for span detection,

intensity classification, and hate type classification. For faithfulness, we use Eraser Benchmark [31] similar to [30], but use predicted span words for erasure instead of attention. Minimal prediction change after removing predicted words suggests independent task learning or reliance on spurious correlations.

For plausibility, we report the Intersection-over-Union (IoU) F1 score for hate span detection. Since exact rationale matching is challenging, IoU F1 provides a softer evaluation by awarding partial credit for overlapping spans at thresholds of 0.50, 0.60, 0.70, 0.80, 0.90, and 1.00.

For faithfulness, we report *comprehensiveness* and *sufficiency*. Comprehensiveness measures the extent to which the identified rationales influence the model's prediction and is defined as

$$\text{Comprehensiveness} = m(x_i)_j - m(x_i \setminus r_i)_j, \qquad (2)$$

where x_i denotes the text of tweet i, r_i represents the predicted rationales for class j, and $x_i \setminus r_i$ is the input with the rationales removed. Higher comprehensiveness values indicate that the rationales play a substantial role in the prediction and that the model does not rely on spurious features.

Sufficiency evaluates whether the rationales alone are adequate for making the prediction and is defined as

$$\text{Sufficiency} = m(x_i)_j - m(r_i)_j. \qquad (3)$$

Lower sufficiency values indicate that the rationales are essential for the model's decision.

IV. Experimental Design

This section discusses the evaluation strategy and models employed in our experiments.

A. Evaluation Strategy

We perform three experiment sets. First, 'Random' uses stratified sampled train, test, and validation sets (8,649, 2,403, and 961 tweets) maintaining class distribution. Second, 'Time-skip' uses temporal splits with RUHSOLD tweets for training (9,007) and validation (1,001), and newly collected tweets for testing (2,005). Third, we evaluate robustness via black-box adversarial perturbation using word variations to replace predicted hateful span words. All experiments report task performance and explainability metrics.

B. Models

We employ four multilingual pre-trained transformer models in our experiments: DistilBERT-base-Multilingual-cased, BERT-base-Multilingual-uncased, XLM-RoBERTa-base, and Multilingual-MiniLM [37]–[40]. All models are fine-tuned in a multitask setting on the relevant splits of the RUHSOLD++ dataset, where each model is trained jointly on all three tasks. Fine-tuning is performed for up to 100 epochs using the Adam optimizer with default epsilon and early stopping based on validation loss. All models support over 100 languages and have been pre-trained on large-scale multilingual corpora, including Wikipedia and Common Crawl, making them well-suited for multilingual and code-mixed text classification.

TABLE IV
HATE TYPE AND INTENSITY CLASSIFICATION RESULTS FOR RANDOM AND TIME-SKIP EXPERIMENTS.

Model Names	Time-skip				Random			
	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
	Type Classification				Type Classification			
DistilBERT-M	0.70	0.69	0.69	0.77	0.70	0.69	0.70	0.79
BERT-M	0.71	0.70	0.70	**0.79**	0.73	0.73	0.73	0.81
XLM-R	**0.74**	**0.71**	**0.72**	**0.79**	**0.75**	**0.76**	**0.75**	**0.82**
MiniLM-M	0.70	0.67	0.68	0.77	0.72	0.74	0.73	0.80
	Intensity Classification				Intensity Classification			
DistilBERT-M	0.73	0.73	0.73	0.76	0.74	0.74	0.74	0.79
BERT-M	**0.77**	**0.78**	**0.78**	**0.80**	0.77	0.78	0.78	0.82
XLM-R	0.73	0.75	0.74	0.77	**0.78**	**0.79**	**0.79**	**0.83**
MiniLM-M	0.75	0.76	0.76	0.79	0.77	0.78	0.77	0.81

TABLE V
INTERPRETABILITY METRICS FOR RANDOM AND TIME-SKIP EXPERIMENTS.

Model Names	Plausibility						Faithfulness	
	IOU F1 on thresholds						Comprehens.	Sufficiency
	0.50	0.60	0.70	0.80	0.90	1.00	-	-
	Random							
DistilBERT-M	0.71	0.67	0.63	0.59	0.55	0.51	**0.855**	**0.047**
BERT-M	0.80	0.75	0.71	0.66	0.61	0.58	0.852	0.350
XLM-R	**0.82**	**0.77**	**0.73**	**0.68**	**0.64**	**0.61**	0.848	0.153
MiniLM-M	0.80	0.76	0.72	**0.68**	**0.64**	**0.61**	0.835	0.136
	Time-skip							
DistilBERT-M	0.72	0.67	0.63	0.60	0.56	0.54	0.826	**0.034**
BERT-M	**0.81**	**0.77**	**0.73**	0.67	0.62	0.60	**0.902**	0.045
XLM-R	0.80	0.76	0.71	0.67	**0.64**	**0.62**	0.838	0.093
MiniLM-M	0.80	0.76	0.71	**0.68**	**0.64**	**0.62**	0.803	0.080

V. Results and Discussion

This section presents our experimental findings across three dimensions: (1) classification and span detection performance, (2) interpretability analysis, and (3) adversarial robustness. We analyze not only model performance but also the practical implications for deploying hate speech detection systems in real-world settings.

A. Classification and Span Detection

Tables IV and V present classification and interpretability results for both evaluation settings.

1) Random Split: XLM-RoBERTa achieves the best performance with F1 scores of 0.75 (hate type) and 0.79 (intensity), while BERT-Multilingual and MiniLM trail by only 0.01–0.02 points despite fewer parameters, suggesting smaller models offer viable accuracy-efficiency trade-offs for deployment. All models exhibit slight recall bias, producing more false positives than false negatives. This characteristic benefits human-review settings where missing hateful content carries greater risk than over-flagging.

For interpretability, models achieve high comprehensiveness (0.835–0.855), indicating predictions rely on identified spans rather than spurious features. However, IoU F1 scores decline substantially (minimum 0.20 drop) as thresholds increase from 0.5 to 1.0. Error analysis reveals models over-predict

span boundaries by including adjacent non-hateful words. Combined with tweet brevity, where span removal eliminates large input portions, this inflates comprehensiveness scores, suggesting the need for boundary-aware span detection or binary token classification instead of BIO tagging.

2) Time-Skip Split: Under temporal distribution shift, all models show degraded performance. XLM-RoBERTa's hate type F1 drops from 0.75 to 0.72; intensity drops more sharply from 0.79 to 0.74. The shift from recall-dominant to precision-dominant predictions is concerning: models become conservative, missing emerging hate speech patterns while maintaining accuracy on familiar ones. Given hateful content's low base rate online, this recall degradation poses significant deployment challenges.

Interpretability metrics show divergent trends. BERT-Multilingual's comprehensiveness improves to 0.902, the only model gaining, suggesting more generalizable span-classification associations. All models achieve sufficiency below 0.1, indicating that despite degraded overall performance, identified spans become more predictive. Models appear to extract partial but highly indicative rationales from evolved hate speech.

B. Adversarial Experiments

To evaluate robustness, we examine word variation substitution effects. Data augmentation improves hate speech detection for RU [19], but linguistic variation space is expansive and dynamic, making comprehensive annotation challenging. This experiment highlights real-world deployment challenges. We extract rationales from test sets, randomly replace words within predicted spans with word variations dataset alternatives, then reprocess for type, intensity, and span prediction. Tables VI and VII report the results of these adversarial robustness experiments.

1) Random Split with Variations: Performance drops across all models, with intensity classification more affected than hate type. For hate type, only BERT-Multilingual maintains its F1 (0.73); XLM-RoBERTa shows the largest drop (0.75→0.69). This reversal suggests XLM-RoBERTa's cross-lingual pre-training creates brittleness to character-level perturbations absent from pre-training data. The shift to higher precision with lower recall indicates models fail to recognize perturbed hateful words, producing false negatives, mirroring real-world dynamics where misspelled slurs evade detection.

2) Time-Skip Split with Variations: Combining temporal shift with adversarial perturbation reveals compounding vulnerabilities. XLM-RoBERTa's hate type F1 collapses to 0.48, while BERT-Multilingual proves most robust at 0.61. Intensity classification shows relative resilience (XLM-RoBERTa: 0.64 F1), suggesting intensity signals partially encode non-lexical features stable across perturbations. Interpretability metrics decline less severely than classification, with BERT-Multilingual maintaining highest IoU F1 across thresholds.

These findings have direct deployment implications: (1) static models require periodic retraining as language evolves, (2) adversarial robustness needs explicit attention beyond standard fine-tuning, and (3) the varying degradation patterns across models suggest ensemble approaches may provide more stable performance.

TABLE VI
HATE TYPE AND INTENSITY CLASSIFICATION RESULTS FOR MODELS TRAINED ON RANDOM AND TIME-SKIP SPLITS WITH ADDED VARIATIONS.

Model Names	Time-skip				Random			
	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
	Type Classification				Type Classification			
DistilBERT-M	0.59	0.54	0.56	0.68	0.69	0.67	0.68	0.72
BERT-M	**0.66**	**0.58**	**0.61**	**0.73**	**0.74**	**0.72**	**0.73**	**0.76**
XLM-R	0.60	0.47	0.48	0.66	0.71	0.68	0.69	0.73
MiniLM-M	0.61	0.49	0.51	0.68	0.73	0.70	0.71	0.75
	Intensity Classification				Intensity Classification			
DistilBERT-M	0.62	0.57	0.59	0.72	0.70	0.68	0.69	0.75
BERT-M	0.68	0.60	0.63	**0.76**	**0.74**	**0.71**	**0.72**	**0.78**
XLM-R	**0.69**	**0.61**	**0.64**	**0.76**	**0.74**	**0.71**	**0.72**	0.77
MiniLM-M	0.66	0.58	0.61	0.73	0.73	0.70	0.71	0.76

TABLE VII
INTERPRETABILITY METRICS FOR MODELS TRAINED ON RANDOM AND TIME-SKIP SPLITS WITH WORD VARIATIONS.

Model Names	Plausibility						Faithfulness	
	IOU F1 on thresholds						Comprehens.	Sufficiency
	0.50	0.60	0.70	0.80	0.90	1.00	-	-
	Random							
DistilBERT-M	0.66	0.61	0.58	0.54	0.50	0.48	0.812	**0.039**
BERT-M	0.72	0.68	0.64	0.60	0.55	0.53	0.818	0.393
XLM-R	**0.72**	**0.68**	**0.64**	**0.61**	**0.57**	**0.55**	0.815	0.155
MiniLM-M	0.70	0.66	0.62	0.58	0.56	0.53	**0.820**	0.147
	Time-skip							
DistilBERT-M	0.66	0.61	0.58	0.55	0.52	0.50	0.809	**0.040**
BERT-M	**0.74**	**0.70**	**0.65**	**0.62**	**0.58**	**0.57**	**0.885**	0.050
XLM-R	0.70	0.66	0.62	0.58	0.55	0.54	0.807	0.118
MiniLM-M	0.73	0.70	0.65	0.61	0.58	0.56	0.801	0.090

VI. CONCLUSION AND FUTURE WORK

This work addresses hate and offensive speech detection for Roman Urdu (RU), a low-resource language lacking prior dedicated resources. We introduce RUHSOLD++, an enhanced dataset that extends RUHSOLD with additional tweets and tasks, including hate span and intensity annotations for interpretability, temporally distinct data for evaluating generalization, and spelling variations for robustness analysis. Experiments using multilingual transformer models show that performance degrades as hate-related topics evolve over time and that adversarial spelling variations significantly impact detection accuracy. Although evaluated on Roman Urdu, the framework is language-agnostic and applicable to other low-resource, code-mixed languages. Future work will explore prompt-tuning methods to reduce reliance on large annotated datasets, addressing a key challenge in hate speech detection for low-resource languages.

REFERENCES

[1] N. S. Mullah and W. M. N. W. Zainon, "Advances in machine learning algorithms for hate speech detection in social media: a review," *IEEE access*, vol. 9, pp. 88 364–88 376, 2021.

[2] N. Jafari, J. Allan, and S. M. Sarwar, "Target span detection for implicit harmful content," in *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, 2024, pp. 117–122.

[3] J. S. Malik, H. Qiao, G. Pang, and A. van den Hengel, "Deep learning for hate speech detection: a comparative study," *International Journal of Data Science and Analytics*, pp. 1–16, 2024.

[4] M. Usman, M. Ahmad, G. Sidorov, I. Gelbukh, and R. Q. Tellez, "A large language model-based approach for multilingual hate speech detection on social media," *Computers*, vol. 14, no. 7, p. 279, 2025.

[5] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. A. Hale, and H. Margetts, "Challenges and frontiers in abusive content detection," in *Proceedings of the third workshop on abusive language online*, 2019, pp. 80–93.

[6] S. Nagar, F. A. Barbhuiya, and K. Dey, "Towards more robust hate speech detection: using social context and user data," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 47, 2023.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "" 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*, 2016, pp. 1135–1144.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[9] H. Rizwan, M. H. Shakeel, and A. Karim, "Hate-speech and offensive language detection in roman urdu," in *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2020, pp. 2512–2522.

[10] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL student research workshop*, 2016, pp. 88–93.

[11] M. Sajjad, F. Zulifqar, M. U. G. Khan, and M. Azeem, "Hate speech detection using fusion approach," in *2019 International Conference on Applied and Engineering Mathematics (ICAEM)*. IEEE, 2019, pp. 251–255.

[12] Q. Mehmood, A. Kaleem, and I. Siddiqi, "Islamophobic hate speech detection from electronic media using deep learning," in *Mediterranean conference on pattern recognition and artificial intelligence*. Springer, 2021, pp. 187–200.

[13] S. Chandrasekaran, A. K. Singh Pundir, T. B. Lingaiah *et al.*, "Deep learning approaches for cyberbullying detection and classification on social media," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[14] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[15] Z. Yu, I. Sen, D. Assenmacher, M. Samory, L. Fröhling, C. Dahn, D. Nozza, and C. Wagner, "The unseen targets of hate: A systematic review of hateful communication datasets," *Social Science Computer Review*, vol. 43, no. 5, pp. 1114–1144, 2025.

[16] M. Haim and E. Hoven, "Hate speech's double damage: A semi-automated approach toward direct and indirect targets," *Journal of Quantitative Description: Digital Media*, vol. 2, 2022.

[17] E. Hashmi, S. Y. Yayilgan, I. A. Hameed, M. M. Yamin, M. Ullah, and M. Abomhara, "Enhancing multilingual hate speech detection: From language-specific insights to cross-linguistic integration," *IEEE Access*, 2024.

[18] A. Pal, P. Hansdak, G. Raj, A. S. Verma, and A. P. Agrawal, "Filtering toxicity in multilingual social media conversations: A step towards safer online interactions," in *Artificial Intelligence and Sustainable Innovation*. CRC Press, 2026, pp. 170–176.

[19] U. Azam, H. Rizwan, and A. Karim, "Exploring data augmentation strategies for hate speech detection in roman urdu," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 4523–4531.

[20] M. M. Abdelsamie, S. S. Azab, and H. A. Hefny, "The dialects gap: A multi-task learning approach for enhancing hate speech detection in arabic dialects," *Expert Systems with Applications*, vol. 295, p. 128584, 2026.

[21] F. Ghorbanpour, D. Dementieva, and A. Fraser, "Can prompting llms unlock hate speech detection across languages? a zero-shot and few-shot study," in *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, 2025, pp. 413–425.

[22] H. H. Saeed, M. H. Ashraf, F. Kamiran, A. Karim, and T. Calders, "Roman urdu toxic comment classification," *Language Resources and Evaluation*, vol. 55, no. 4, pp. 971–996, 2021.

[23] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, and M. T. Sadiq, "Automatic detection of offensive language for urdu and roman urdu," *IEEE Access*, vol. 8, pp. 91 213–91 226, 2020.

[24] R. Ali, U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg, "Hate speech detection on twitter using transfer learning," *Computer Speech & Language*, vol. 74, p. 101365, 2022.

[25] F. Ullah, A. Faheem, U. Azam, M. S. Ayub, F. Kamiran, and A. Karim, "Detecting cybercrimes in accordance with pakistani law: Dataset and evaluation using plms," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 4717–4728.

[26] A. Rashid, S. Mahmood, U. Inayat, and M. F. Zia, "Urdu toxicity detection: A multi-stage and multi-label classification approach," *AI*, vol. 6, no. 8, p. 194, 2025.

[27] B. A. Juliussen, "The right to an explanation under the gdpr and the ai act," in *International Conference on Multimedia Modeling*. Springer, 2025, pp. 184–197.

[28] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A survey of the state of explainable ai for natural language processing," *arXiv preprint arXiv:2010.00711*, 2020.

[29] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?" *arXiv preprint arXiv:2004.03685*, 2020.

[30] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 14 867–14 875.

[31] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "Eraser: A benchmark to evaluate rationalized nlp models," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 4443–4458.

[32] G. Damo, N. B. Ocampo, E. Cabrio, and S. Villata, "Unveiling the hate: Generating faithful and plausible explanations for implicit and subtle hate speech detection," in *International Conference on Applications of Natural Language to Information Systems*. Springer, 2024, pp. 211–225.

[33] T. Xiang, S. MacAvaney, E. Yang, and N. Goharian, "Toxccin: Toxic content classification with interpretability," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2021, pp. 1–12.

[34] L. Zhou, A. Caines, I. Pete, and A. Hutchings, "Automated hate speech detection and span extraction in underground hacking and extremist forums," *Natural Language Engineering*, vol. 29, no. 5, pp. 1247–1274, 2023.

[35] Z. Delbari, N. S. Moosavi, and M. T. Pilehvar, "Spanning the spectrum of hatred detection: a persian multi-label hate speech dataset with annotator rationales," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17 889–17 897.

[36] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation (extended version)," *Information Systems*, vol. 105, p. 101584, 2022.

[37] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[39] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 8440–8451.

[40] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in neural information processing systems*, vol. 33, pp. 5776–5788, 2020.