RESEARCH & INNOVATION DIVISION
IJIST

Beyond Accuracy: Explainability of Transformer Models for Sentiment Analysis

Shahriyar Shahid, Abdul Aziz, Muhammad Bilal, Muhammad Shoaib Lashari, Muhammad Aaqib
Aror University of Art, Architecture, Design & Heritage Sukkur, Pakistan
*Correspondence: f23ari63@aror.edu.pk

Citation | Shahid. S, Aziz. A, Bilal. M, Lashari. M. S, Aaqib. M, "Beyond Accuracy: Explainability of Transformer Models for Sentiment Analysis", IJIST, Vol. 07 Issue. 10 pp 300-310, December 2025

Received| November 26, 2025 **Revised|** December 18, 2025 **Accepted|** December 22, 2025 **Published|** December 26, 2025.

Transformer-based models achieve strong performance on Urdu sentiment analysis; however, their predictions are often difficult to interpret, which can undermine trust in practical applications. In this paper, we present a Hybrid LIMESHAP Attribution (HLSAE) Module. We fine-tune Urdu BERT and introduce a token-level explanation pipeline that combines LIME and SHAP. The method is evaluated on a dataset of 50,000 Urdu movie reviews. Compared with LIME or SHAP individually, the hybrid approach produces more stable attributions across multiple runs and perturbations. In a human evaluation involving native Urdu speakers, the explanations achieve a fidelity score of 0.72, while the classifier maintains an F1-score of 79.8%. The resulting visualizations highlight sentiment-bearing linguistic cues and reveal instances where the model relies on spurious correlations. These results demonstrate that explainability can be incorporated without sacrificing classification accuracy in low-resource sentiment analysis, and the proposed workflow can be extended to similar languages.

Keywords: Explainable AI, Hybrid Explainability, Urdu NLP, Transformer Models, Sentiment Analysis

Introduction:

Urdu is spoken by more than 253 million people worldwide, and it plays a central role in online communication and cultural expression. With the rapid growth of Urdu content on social media, platforms also face familiar risks such as misinformation, hate speech, and highly polarizing discourse, which can undermine social cohesion [1][2]. Sentiment analysis provides a practical tool for monitoring such content at scale. However, developing reliable Urdu sentiment classifiers remains challenging due to rich morphology, frequent codeswitching with English, dialectal variation, and the limited availability of high-quality annotated datasets [3].

Recent transformer-based models, particularly Urdu BERT, have reported strong results by transferring knowledge from large unlabeled corpora through fine-tuning [4][5]. Their self-attention mechanism [6] captures long-range dependencies and subtle contextual cues, which is especially useful in low-supervision scenarios [7][8]. Despite these gains, transformers are often difficult to interpret. They produce a label without clearly indicating which words or linguistic cues contributed to the decision [9]. Transformers encode rich linguistic structure in their hidden representations [10][11]; identifying which internal components are responsible for specific predictions remains challenging [12]. This lack of transparency has practical consequences. Without explanations, stakeholders cannot easily verify or audit predictions [13], making deployment in sensitive settings riskier. The model may also rely on spurious correlations, such as dialectal markers or domain-specific terms rather than genuine sentiment evidence [14]. When failures occur, it is difficult to diagnose whether the issue stems from biased training data, annotation noise, or the model's misunderstanding of linguistic structure [15]. Furthermore, policy trends and emerging regulations increasingly emphasize the need for transparent and interpretable automated decisions [16].

Our Solution and Contributions:

To address the interpretability gap in Urdu sentiment analysis, we present a Hybrid LIME-SHAP Attribution Explainability (HLSAE) Module, which combines a fine-tuned Urdu BERT classifier with post-hoc attribution techniques. In addition to reporting classification accuracy, we evaluate the reliability and usefulness of the explanations, treating explanation quality as a first-class objective alongside predictive performance.

Key Contributions of This Work Include:

An end-to-end pipeline that produces token-level attributions for every model prediction.

A hybrid LIME–SHAP fusion strategy that leverages perturbation-based local explanations while incorporating the consistency benefits of game-theoretic attributions.

A large-scale study on 50,000 translated Urdu movie reviews, highlighting linguistic cues that are most influential for sentiment decisions.

Quantitative evaluation of explanation quality, including stability under perturbations, fidelity with human judgments, and the accuracy–interpretability trade-off.

A qualitative user study with native Urdu speakers showing that the proposed hybrid explanations are easier to interpret than single-method baselines.

Overall, this work bridges strong transformer-based performance with more accountable decision-making, supporting trustworthy NLP systems for Urdu and other low-resource languages.

Paper Organization:

Section II reviews related work. Section III describes the dataset, model architecture, and explanation methods. Section IV reports classification and explanation results. Section V discusses insights and limitations. Section VI concludes the paper and outlines future directions.

Related Work:

Urdu Natural Language Processing:

Urdu NLP faces significant challenges including limited annotated resources [17], frequent code-switching with English [18], and rich morphological structures. Early sentiment analysis systems relied on lexicon-based methods and conventional machine learning models, which struggled to capture context-dependent polarity. Recent transformer-based approaches, particularly multilingual BERT and Urdu-specific variants, have substantially improved performance [19]; however, most studies focus primarily on classification accuracy without examining model interpretability.

Explainable AI for NLP:

Post-hoc explainability methods have become essential as NLP models grow increasingly complex. LIME [20] explains predictions by perturbing input instances and fitting local surrogate models, whereas SHAP [21] employs a game theoretic framework to assign feature-level contributions. Both approaches have inherent trade-offs: LIME can produce unstable explanations due to sampling sensitivity, while SHAP is computationally intensive for large models. Our proposed approach combines the strengths of both methods to balance explanation, stability and computational cost.

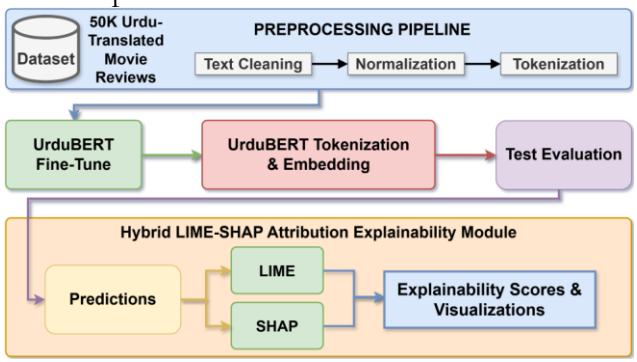

Figure 1. Overall system workflow of the proposed UrduBERT-based sentiment analysis model.

Low-Resource Language Technologies:

Multilingual pre-trained transformers such as BERT have become standard for low-resource NLP tasks, enabling knowledge transfer from large corpora to tasks with limited supervision. UrduBERT and related language-adapted models leverage this strategy; however, interpretability has received comparatively less attention in low-resource contexts [22]. The field has largely prioritized predictive accuracy over explainability, with limited discussion on model validation or auditability. This paper demonstrates that interpretability can be integrated into system design without compromising classification performance.

Methodology:

We propose an end-to-end framework for explainable Urdu sentiment analysis that couple's transformer-based classification with post-hoc attribution techniques. The overall workflow consists of four stages:

Dataset preprocessing
Fine-tuning UrduBERT
Generating explanations with LIME and SHAP
Evaluating both predictive performance and explanation quality
Figure 1 summarizes the complete pipeline.

Model Architecture:

Transformer Backbone:

We employ UrduBERT, a BERT-base model that has been further pre-trained on large Urdu corpora. It adheres to the standard BERT-base configuration, comprising L = 12

transformer encoder layers, H = 12 self-attention heads, and a hidden size of dh = 768. Given an input token sequence x = [x1, x2, xn] of length n, UrduBERT produces contextualized token representations:

$$H = \text{UrduBERT}(x) \in Rn \times dh \quad (1)$$

where H = [h[CLS], h1, hn] contains the embedding for the [CLS] token and the embeddings for all input tokens.

Classification Head. For binary sentiment prediction, we attach a lightweight feed-forward classification layer on top of the [CLS] embedding. The classifier maps h[CLS] to the two sentiment classes (positive, negative) using:

$$\hat{y} = \text{softmax}(Wh[CLS] + b) \quad (2)$$

where W ∈ RC×dh, b ∈ RC, and C = 2 for binary classification. Figure 2 illustrates the full flow, from tokenization through the classifier output.

We train the model using the standard cross-entropy objective:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^{(c)} \log \hat{y}_i^{(c)} \quad (3)$$

where N is the batch size, is the ground-truth indicator for class c, and is the predicted probability assigned to class c.

Data Preprocessing:

For our experiments, we use the Kaggle IMDb dataset compromising 50,000 Urdu-translated movie reviews. [23]. The dataset is balanced, containing 25,000 positive and 25,000 negative reviews. These reviews were originally authored in English and then subsequently professionally translated into Urdu, providing a large-scale benchmark despite the general scarcity of native Urdu sentiment resources.

We preprocess the raw text in four steps. (1) Punctuation, special characters, URLs, and other noise are removed using regex-based cleaning. (2) Unicode normalization is performed by standardizing Urdu diacritics and unifying character variants. (3) Tokenization is conducted using UrduBERT's WordPiece tokenizer, with sequences capped at a maximum length of 128 tokens. (4) The dataset is partitioned into training and test sets using an 80/20 split, resulting in 40,000 training samples and 10,000 test samples.

Fine-Tuning Strategy:

The pre-trained UrduBERT parameters are fine-tuned using the AdamW optimizer [24] with a learning rate of η = 2 × 10−5 and a weight decay of λ = 0.01. A linear warmup is applied over 10% of the total training steps:

$$\eta t = \eta \cdot \min\left(1, \frac{t}{t_{\text{warmup}}}\right) \quad (4)$$

Table 1 presents the complete set of training hyperparameters. To stabilize optimization, gradient clipping is applied with a maximum norm of 1.0. Training is performed for 3 epochs, with validation conducted after each epoch to monitor convergence and mitigate the risk of overfitting.

Table 1. Hyperparameters used for fine-tuning

Hyperparameter	Value
Learning Rate	2×10^{-5}
Batch Size	8
Number of Epochs	3
Optimizer	Adam W
Weight Decay	0.01
Gradient Clipping	1.0
Warmup Steps	10% of total
Evaluation Strategy	Per Epoch

Explainability Framework:

To reduce the black-box nature of transformer predictions, we implement a hybrid explanation pipeline combining LIME and SHAP. The objective is to generate token-level attributions that are both informative and robust, leveraging the complementary strengths of the two methods.

LIME: Local Perturbation-Based Explanations:

LIME (Local Interpretable Model-agnostic Explanations) [20] provides an explanation for a single prediction by generating perturbed versions of the input and fitting an interpretable surrogate model locally around that instance. For given x, we create K perturbations by randomly masking tokens, and subsequently then learn a sparse linear explanation model by solving:

$$\xi(x) = \text{argmin} L_{LIME}(f,g,\pi x) + \Omega(g) \quad (5)$$

$g \in G$ where f is the black-box classifier (UrduBERT), $g \in G$ is an interpretable linear model, πx weights samples by proximity to x, and $\Omega(g)$ controls explanation complexity. The LIME loss is:

$$\mathcal{L}_{\text{LIME}}(f, g, \pi_{\mathbf{x}}) = \sum_{k=1}^{K} \pi_{\mathbf{x}}(\mathbf{x}'_k)[f(\mathbf{x}'_k) - g(\mathbf{x}'_k)]^2 \quad (6)$$

using the kernel $\pi x(x') = \exp(-D(x,x')2/\sigma 2)$, where D measures feature distance. We set K = 1000 perturbations and use Lasso regression to obtain sparse token weights $w_{LIME} \in R^n$.

SHAP: Game-Theoretic Attributions:

SHAP (SHapley Additive exPlanations) [21] assigns each token a Shapley value that quantifies its average marginal contribution across all possible subsets of tokens. For a given token i, the Shapley value is computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (7)$$

where N is the full token set and S denotes a coalition (subset of tokens).

Exact Shapley computation requires O(2n) model evaluations, which is infeasible for long sequences. We therefore use Kernel SHAP, which approximates the solution by fitting a weighted linear model:

$$\mathbf{w}_{\text{SHAP}} = \arg\min_{\mathbf{w}} \sum_{k=1}^{K} \pi_K(z'_k)[f(z'_k) - \mathbf{w}^T z'_k]^2 \quad (8)$$

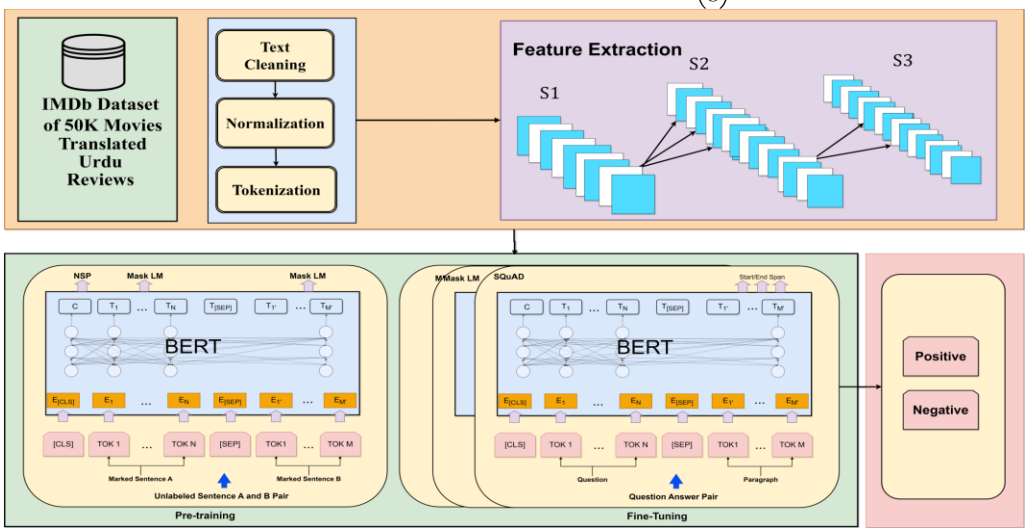

Figure 2. Hybrid LIME–SHAP Attribution Explainability (HLSAE) module architecture with sentiment classification head

with kernel, which preserves the SHAP Consistency Property.

Hybrid Consensus Attribution: LIME is effective for analyzing local model decisions, analysis, whereas SHAP providess theoretically grounded and consistent attributions. To leverage the strengths of both methods, we construct a consensus explanation by linearly combining the two attribution vectors:

$$w_{hybrid} = \alpha \cdot w_{LIME} + (1 - \alpha) \cdot w_{SHAP} \quad (9)$$

We set $\alpha = 0.5$ to give equal weight to both methods. Because negative scores can be harder to interpret for sentiment evidence, we retain only positive contributions via:

$$s_i = \max(0, w_{hybrid,i}) \quad (10)$$

As shown in our results (Section IV), this hybrid strategy produces more stable explanations than using LIME or SHAP alone.

Experimental Setup and Results:

Dataset:

All experiments were conducted on the Kaggle IMDb dataset of 50,000 Urdu-translated movie reviews [23]. The dataset contains an equal number of positive and negative reviews, as illustrated in Figure 3. The reviews were originally composed in English and subsequently professionally translated into Urdu, providing a large-scale benchmark despite the limited availability of native Urdu sentiment datasets.

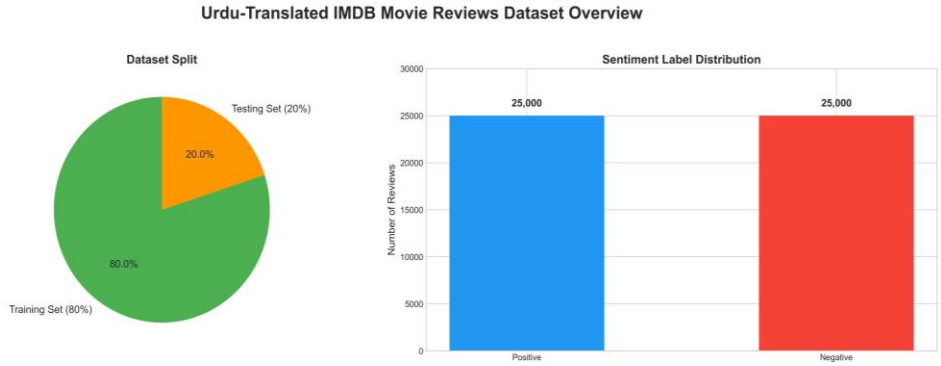

Figure 3. Distribution of sentiment labels in the dataset: 25,000 positive and 25,000 negative reviews.

Following the preprocessing pipeline described in Section III-B, the dataset was split into 40,000 training and 10,000 test samples.

Training Configuration:

Urdu BERT was fine-tuned for 3 epochs using the hyperparameters listed in Table 1, with validation performed at the end of each epoch. Training behavior was monitored using Weights & Biases, and the resulting training and validation curves are presented in Figure 4.

Classification Performance:

The fine-tuned Urdu BERT achieves 79.8% accuracy, with an F1 score of 79.2% precision and 80.5% recall, demonstrating competitive performance, as shown in Table 2. While these results are favorable compared to prior work (summarized in Table 3), it is important to note that [5][25] evaluated their models on different datasets, limiting direct comparison. Importantly, our study differs in scope: To the best of our knowledge, we are the first to investigate the interpretability accuracy trade-off in Urdu sentiment analysis, demonstrating that explainability and predictive performance can coexist as an aspect not addressed in prior approaches.

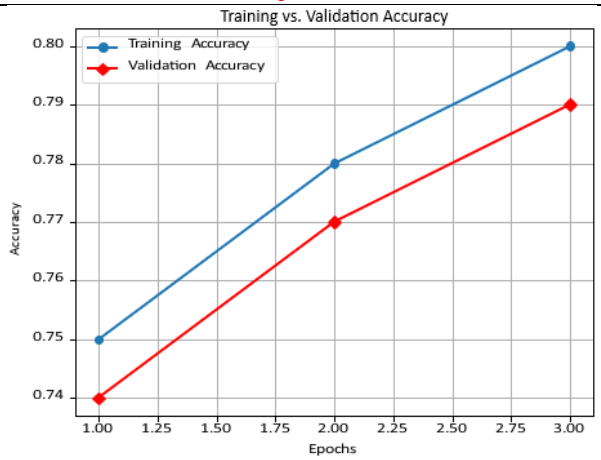

Figure 4. Training and validation accuracy over 3 epochs, reaching 79.8% on the validation set.

Table 2. Sentiment classification performance

Metric	Score (%)
Accuracy	79.8
Precision	79.2
Recall	80.5
F1-Score	79.8

Explainability Analysis:

Attribution Visualization: To illustrate the behavior of the explanation methods on real inputs, Table 4 presents token-level attribution scores produced by LIME, SHAP, and the proposed hybrid method for representative positive and negative reviews. The hybrid approach combines LIME's robust local explanations with SHAP's theoretically grounded attributions, providing more stable and interpretable token-level insights. In our examples, it produces attributions that are less erratic across tokens. Highly weighted tokens correspond to sentiment-bearing cues such as intensifiers (bahut, ziada), negation (nahi), and opinionated words (shandar, kharab), indicating that the model relies on linguistically meaningful evidence rather than arbitrary artifacts.

Quantitative Evaluation: We evaluate explanation quality across three dimensions. For stability, we generate synonym-substituted variants of test samples using Urdu WordNet and measure the cosine similarity between attribution vectors. The hybrid method achieves a stability score of 0.84.

Table 3. Classification performance compared to previous urdu sentiment analysis methods

Method	Accuracy	F1-Score
UrduBERT + Hybrid LIME-SHAP (this work)	79.8%	79.8%[a]
[5]	78.1%	77.8%
[25]	72.5%	71.2%

Evaluated on 50K Kaggle IMDb Urdu-translated dataset with 80/20 train-test split.
Results reported on Urdu social media corpus; different dataset and preprocessing pipeline.
Results reported on custom Roman-Urdu dataset; different annotation scheme and preprocessing.

Note: Urdu tokens shown in romanized transliteration with English glosses. Cells with attribution > 0.6 are highlighted. Original reviews: Positive – "This movie was very good and the acting was excellent"; Negative – "This movie was very bad and the story was boring".

(std. 0.09), outperforming LIME (0.76) and SHAP (0.81), demonstrating greater robustness to input perturbations.

For fidelity, three native Urdu speakers annotated 200 test reviews to identify sentiment-critical tokens (Cohen's $\varkappa = 0.81$). At k = 5 tokens, the hybrid method achieves 0.72 fidelity compared to 0.65 for LIME and 0.68 for SHAP, indicating superior alignment with human judgment.

Table 4. Comparison of token-level attribution scores for lime, shap, and hybrid methods

Review	Token (Romanized)	LIME	SHAP	Hybrid
Positive	*yeh* (this)	0.02	0.03	0.025
	film (movie)	0.05	0.04	0.045
	bahut (very)	0.85	0.88	0.865
	ziada (very)	0.82	0.84	0.83
	achhi (good)	0.92	0.78	0.85
	thi (was)	0.01	0.02	0.015
	aur (and)	0.03	0.04	0.035
	adakari (acting)	0.78	0.82	0.80
	shandar (excellent)	0.95	0.88	0.915
	thi (was)	0.02	0.03	0.025
Negative	yeh (this)	0.03	0.04	0.035
	film (movie)	0.06	0.05	0.055
	bahut (very)	0.83	0.86	0.845
	nahi (no)	0.81	0.79	0.80
	Kharab (bad)	0.94	0.89	0.915
	thi (was)	0.02	0.03	0.025
	aur (and)	0.04	0.05	0.045
	bilkul nahi (absolutely not)	0.87	0.85	0.86
	kahani (story)	0.75	0.78	0.765
	boring (boring)	0.91	0.87	0.89
	thi (was)	0.03	0.04	0.035

For sufficiency, we measure whether predictions remain correct when using only the top-k attributed tokens. With k = 10, the hybrid approach retains 89.3% accuracy, confirming that high-attribution tokens effectively capture the sentiment signal.

Table 5. Ablation Study: Effect of Fusion Weight α

a	Stability	Fidelity@5	Sufficiency@10
0.0 (SHAP only)	0.81	0.68	0.87
0.3	0.83	0.70	0.88
0.5 (Hybrid)	0.84	0.72	0.89
0.7	0.82	0.69	0.88
1.0 (LIME only)	0.76	0.65	0.86

Ablation Study and Computational Analysis:

We examine the impact of the fusion weight α in Equation 9. Table 5 reports stability, fidelity@5, and sufficiency@10 for different settings. An equal balance ($\alpha = 0.5$) yields the best stability and fidelity, motivating our default choice to weight LIME and SHAP equally.

Beyond ablation experiments, we also evaluate the computational overhead introduced by the hybrid explanation pipeline. Inference without explanations takes 0.02 s per sample. With explanations, LIME requires 1.2 s per sample, SHAP requires 0.8 s, and the hybrid method takes 1.3 s. While this adds overhead, it remains practical for real-time use. Importantly, the per-sample cost is independent of training dataset size. Larger datasets may increase fine-tuning time, but the inference and explanation pipeline scales linearly, making it suitable for deployment.

Discussion:

Key Findings:

Three key observations stand out. First, Urdu BERT achieves F1 score of 79.8%, demonstrating that transfer learning is effective for Urdu sentiment classification. Second, hybrid LIME– SHAP explanations are both more stable (0.84) and better aligned with human judgment (0.72 fidelity) than either method individually. Third, generating explanations introduces additional computational overhead (1.3 s vs. 0.02 s inference), but remains practical when interpretability is required.

Linguistic Insights:

Attributions highlight expected Urdu sentiment cues: intensifiers (bahut, ziada), negations (nahi, bilkul nahi), and affective words (shandar, kharab) frequently dominate. The explanations also maintain meaningfulness under common Urdu-English code-switching, reflecting the model's robustness to mixed-language input.

Limitations and Impact:

As experiments rely on translated movie reviews with binary sentiment labels, results may not fully generalize to other domains (e.g., social media) or multi-class sentiment tasks without additional data. The human evaluation (200 samples, 3 annotators) provides useful insights but is limited in scale. Nevertheless, token-level explanations enable auditing of model behavior and contribute to more transparent analysis of Urdu content, supporting accountability in practical applications.

Conclusion and Future Work:

We introduced an explainable Urdu sentiment analysis framework, the Hybrid LIME–SHAP Attribution Explainability (HLSAE) Module, which fine-tunes Urdu BERT and generates token-level explanations using a hybrid LIME-SHAP approach. The model achieves F1 score of 79.8% with explanations that are more stable (0.84) and better aligned (0.72) than those produced by either method individually. These results demonstrate that interpretability and predictive performance are not mutually exclusive in low-resource NLP settings.

Future work includes extending the framework to multiclass sentiment classification and evaluating it across diverse domains such as social media and news. Additionally, we plan to apply the approach to other low-resource languages including Sindhi, explore inherently interpretable architectures, and deploy the system in real-world moderation pipelines to validate its practical utility.

Overall, this advances Urdu sentiment analysis toward models that are not only accurate but also transparent and trustworthy.

References:

[1] M. A. H. Muhammad Irzam Liaqat, "Sentiment analysis techniques, challenges, and opportunities: Urdu language-based analytical study," *PeerJ Comput. Sci.*, vol. 8, 2022, [Online]. Available: https://www.researchgate.net/publication/363184417_Sentiment_analysis_techniques_challenges_and_opportunities_Urdu_language-based_analytical_study

[2] "Ethnologue: Languages of the World, 24th Edition | Request PDF." Accessed: Mar. 29, 2026. [Online]. Available: https://www.researchgate.net/publication/352064261_Ethnologue_Languages_of_the_World_24th_Edition

[3] U. Khan, M. Bin Ahmad, F. Shafiq, and M. Sarim, "Urdu Natural Language Processing Issues and Challenges: A Review Study," *Commun. Comput. Inf. Sci.*, vol. 1198, pp. 461–470, 2020, doi: 10.1007/978-981-15-5232-8_39.

[4] B. Tahir and M. A. Mehmood, "UBERT22: Unsupervised Pre-training of BERT for Low Resource Urdu Language," *2022 16th Int. Conf. Open Source Syst. Technol. ICOSST 2022 - Proc.*, 2022, doi: 10.1109/ICOSST57195.2022.10016821.

[5] S. Tariq, T. A. Rana, and F. Shahzadi, "A comparative study of sentiment analysis in urdu and roman urdu: the neglected realms," *CSI Trans. ICT*, vol. 13, no. 2–3, pp. 193–211, Sep. 2025, doi: 10.1007/s40012-025-00418-8.

[6] Z. Ali, A. Aziz, A. Ali, A. Ullah, M. Aurangzeb, and M. A. Nazir, "Fine-Tuned training method for semantic text similarity measurement using SBERT, Bi-LSTM and Attention Network," *Proc. - 2023 Int. Conf. Mach. Vision, Image Process. Imaging Technol. MVIPIT 2023*, pp. 134–140, 2023, doi: 10.1109/MVIPIT60427.2023.00028.

[7] Abdul Aziz Ansari, M. Abdul Rehman, "Spatial Data Analysis: Recommendations for Educational Infrastructure in Sindh," *Sukkur IBA J. Comput. Math. Sci.*, vol. 1, no. 1, 2017, [Online]. Available: https://www.researchgate.net/publication/318987116_Spatial_Data_Analysis_Recommendations_for_Educational_Infrastructure_in_Sindh

[8] A. Aziz *et al.*, "Leverage Diagnosis Intensity in Medication Recommendations," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 14880 LNAI, pp. 38–50, 2024, doi: 10.1007/978-981-97-5678-0_4.

[9] Sarthak Jain, Byron C. Wallace, "Attention is not Explanation," *arXiv:1902.10186*, 2019, [Online]. Available: https://arxiv.org/abs/1902.10186

[10] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, Marco Baroni, "What you can cram into a single vector: Probing sentence embeddings for linguistic properties," *arXiv:1805.01070*, 2018, [Online]. Available: https://arxiv.org/abs/1805.01070

[11] A. E. M. Anna Mai, "Linguistic structure as a guiding principle for human neuroscience," *Neurosci. Biobehav. Rev.*, vol. 177, p. 106322, 2025, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0149763425003239

[12] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, Wojciech Samek, "Explaining Recurrent Neural Network Predictions in Sentiment Analysis," *arXiv:1706.07206*, 2017, [Online]. Available: https://arxiv.org/abs/1706.07206

[13] M. Rehan Ashraf, M. Hussain, M. Arfan Jaffar, W. Yousuf Ramay and M. Faheem, "Revolutionizing Urdu Sentiment Analysis: Harnessing the Power of XLM-R and GPT-2," *IEEE Access*, vol. 12, pp. 99779–99793, 2024, doi: 10.1109/ACCESS.2024.3429496.

[14] I. A. Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, "Survey of Cultural Awareness in Language Models: Text and Beyond," *arXiv:2411.00860*, 2024, [Online]. Available: https://arxiv.org/abs/2411.00860

[15] S. S. Javier Troya, "Model Transformation Testing and Debugging: A Survey," *ACM Comput. Surv.*, vol. 55, no. 4, 2022, [Online]. Available: https://dl.acm.org/doi/full/10.1145/3523056

[16] "Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future." Accessed: Feb. 08, 2026. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

[17] Guang Xiang, Bin Fan, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," *ACM Int. Conf. Proceeding Ser.*, 2012, [Online]. Available: https://dl.acm.org/doi/10.1145/2396761.2398556

[18] Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, Alan W Black, "A Survey of Code-switched Speech and Language Processing," *arXiv:1904.00784*, 2019, [Online]. Available: https://arxiv.org/abs/1904.00784

[19] K. T. Jacob Devlin, Ming-Wei Chang, Kenton Lee, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805*, 2018,

[Online]. Available: https://arxiv.org/abs/1810.04805

[20] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *arXiv:1602.04938*, 2016, [Online]. Available: https://arxiv.org/abs/1602.04938

[21] Hammad Rizwan, Muhammad Haroon Shakeel, "Hate-Speech and Offensive Language Detection in Roman Urdu," *EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, 2020, [Online]. Available: https://www.researchgate.net/publication/347236442_Hate-Speech_and_Offensive_Language_Detection_in_Roman_Urdu

[22] Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, Dietrich Klakow, "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios," *Assoc. Comput. Linguist.*, 2021, [Online]. Available: https://aclanthology.org/2021.naacl-main.201/

[23] "IMDB Dataset of 50K Movie translated Urdu Reviews." Accessed: Feb. 08, 2026. [Online]. Available: https://www.kaggle.com/datasets/akkefa/imdb-dataset-of-50k-movie-translated-urdu-reviews

[24] Ilya Loshchilov, Frank Hutter, "Decoupled Weight Decay Regularization," *arXiv:1711.05101*, 2017, [Online]. Available: https://arxiv.org/abs/1711.05101

[25] Muhammad Bilal, Huma Israr, Muhammad Shahid,Amin Khan, "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 28, no. 3, pp. 330–344, 2016, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157815001330