IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 5, Ser. I (Sep. – Oct. 2026), PP 72-77
www.iosrjournals.org

Explainable and Robust Multitask Transformer for Hate Speech Detection

Jayashree Kalawa[1], Dr. Arpana Chourasia[2], Dr. Mohan Ashwala[3]
[1]*Research Scholar, Department of CSE, Madhyanchal Professional University, Bhopal, MP (India)*
[2]*Supervisor, Department of CSE, MPU Bhopal.MP (India)*
[3]*Co-Supervisor, Department of ECE-, School of Engineering, Malla Reddy University, Hyderabad, Telangana State*
Corresponding author mail id: jayshreekalawa11@gmail.com

Abstract
Hate speech detection systems must do more than assign a harmful-content label. Moderators and affected communities need to understand why a prediction was produced, while attackers continually modify spelling, punctuation, and wording to evade detection. This paper proposes an explainable and robust multitask transformer framework that jointly predicts hate speech, offensive language, target group, and token-level rationales. The model combines a shared contextual encoder with classification heads and a rationale extraction head. A consistency objective encourages the rationale to support the classification decision, while robustness training evaluates character perturbations, obfuscation, paraphrasing, and code-mixed text. The proposed study uses rationale-annotated and multilingual datasets where available. It defines evaluation through classification metrics, rationale overlap, faithfulness, perturbation robustness, and subgroup error analysis. Results are reserved for actual experiments. The paper offers a second, distinct research direction that extends ordinary multitask classification toward interpretable and resilient moderation assistance.

Keywords
Explainable AI; robust NLP; multitask transformer; rationale extraction; hate speech; adversarial text; multilingual learning.

I. Introduction

Automated hate speech detection is increasingly used to support content moderation. However, a prediction without an explanation can be difficult to audit. A model may rely on demographic terms, identity mentions, or superficial profanity rather than hateful meaning. Such behavior creates false positives and may disproportionately affect discussions about protected groups.

A second challenge is evasion. Users may insert punctuation, replace characters, use deliberate misspellings, or switch languages. Robust detection requires evaluation beyond clean test sets. This paper therefore develops a multitask model that predicts both the content category and the textual evidence supporting the prediction.

The main contributions are: a rationale-aware multitask architecture; a consistency loss between rationales and classification; a robustness evaluation protocol; and a multilingual extension suitable for English, Hindi, Telugu, or code-mixed text when adequate data are available.

II. Background and Related Work

Explainable hate speech research has shown that human rationales can be useful for studying model plausibility and faithfulness. HateXplain provides labels, target communities, and rationale spans. Recent multilingual benchmark work has explored rationale annotations for Hindi, Telugu, and English, demonstrating the importance of language coverage.

Robust NLP research commonly evaluates character-level noise, word substitutions, paraphrases, and distribution shifts. In hate speech detection, robustness is complicated because harmful content often contains intentional spelling variation. A robust model should preserve its prediction when meaning is unchanged, but it should also change its prediction when a perturbation changes the meaning.

Multitask learning provides a natural framework: classification tasks teach semantic distinctions, target prediction teaches contextual information, and rationale extraction teaches evidence localization. However, rationale supervision can introduce negative transfer if annotations are inconsistent or too sparse.

III. Proposed Deep Neural Network Architecture

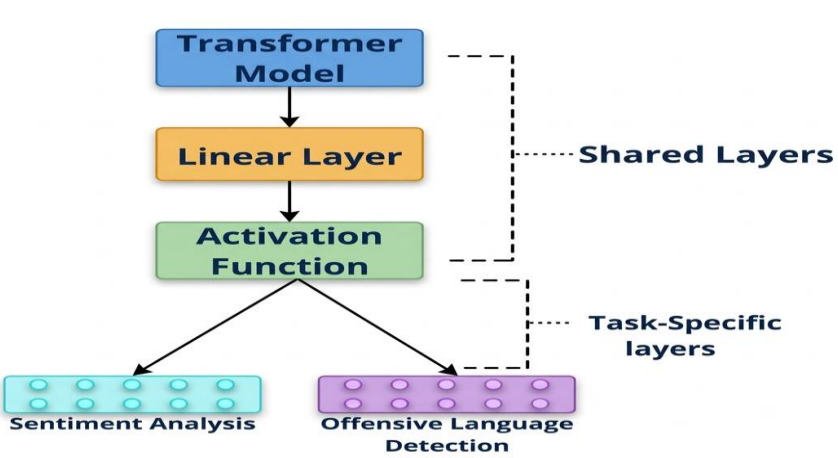

Figure 1

IV. Proposed Framework

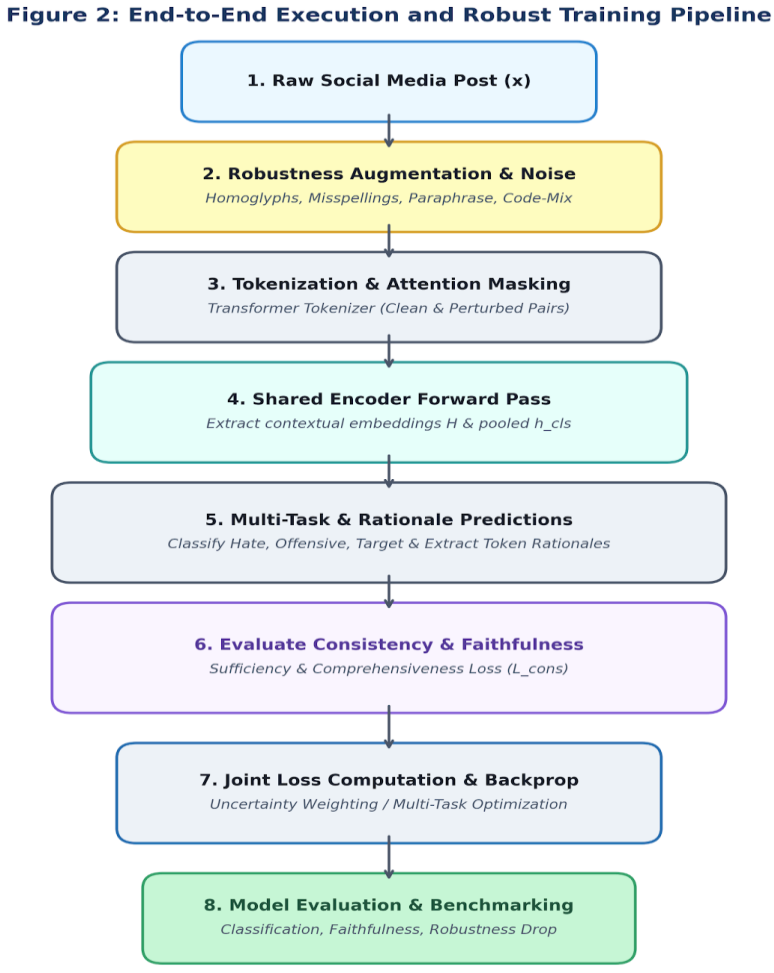

Figure 2

4.1 Input representation

Each post is tokenized using a pretrained transformer tokenizer. Special tokens identify the sequence boundary. The encoder produces contextual token representations H={h_1,...,h_n} and a pooled representation h_cls. Attention masks prevent padding tokens from contributing to the representation.

4.2 Prediction heads

The classification head predicts hate/non-hate. Additional heads predict offensive language and target group. The rationale head assigns a score r_j to each token. A sigmoid or softmax can convert token scores into rationale probabilities. The final explanation is obtained by selecting tokens above a threshold or selecting the top-k tokens.

4.3 Consistency objective

Let m_j be the rationale mask and p_h the hate probability. The rationale loss is token-level binary cross-entropy or a span-level loss. A sufficiency-style objective measures whether the selected rationale preserves the prediction, while a comprehensiveness-style objective measures whether removing the rationale changes the prediction. These can be approximated during training or used during evaluation.
The total loss is L=λ_hL_h+λ_oL_o+λ_tL_t+λ_rL_r+λ_cL_cons. The weights may be fixed initially and then adapted using uncertainty weighting or gradient-balancing methods. The consistency term must be tuned carefully because excessive pressure can produce short but unfaithful explanations.

4.4 Robustness training

Training can include meaning-preserving perturbations such as character swaps, repeated characters, punctuation insertion, homoglyph substitution, masked-word replacement, and controlled paraphrasing. Each augmented example retains the original label only when semantic equivalence is verified. A consistency loss encourages similar predictions for clean and meaning-preserving variants.

V. Dataset and Experimental Protocol

Figure 1: Explainable and Robust Multitask Transformer Architecture

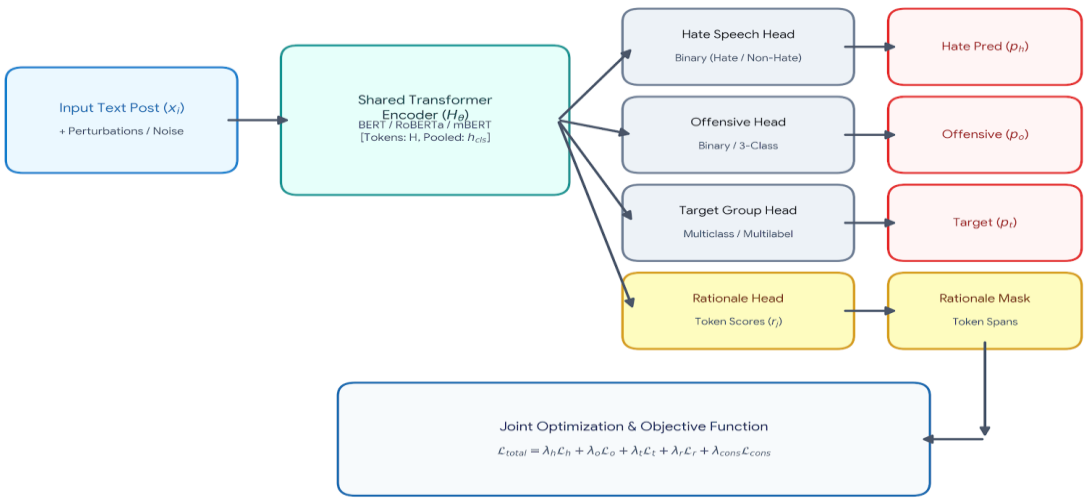

Figure 3

HateXplain is suitable for rationale and target experiments. HatEval and OLID can support hate/offensive tasks. Multilingual work may use datasets containing Hindi, Telugu, and English examples, provided that licensing and annotation quality are verified. Dataset-specific label mappings must be documented.

- Use stratified splits and avoid leakage from duplicated or near-duplicated posts.
- Evaluate clean and perturbed test sets separately.
- Report macro-F1, precision, recall, accuracy, and calibration measures.
- For rationales, report Token-F1, IOU-F1, sufficiency, and comprehensiveness where annotations permit.
- Report robustness drop: clean score minus perturbed score.
- Perform language-wise and target-wise error analysis.
- Compare single-task transformer, multitask transformer, rationale multitask model, and rationale-plus-robustness model.

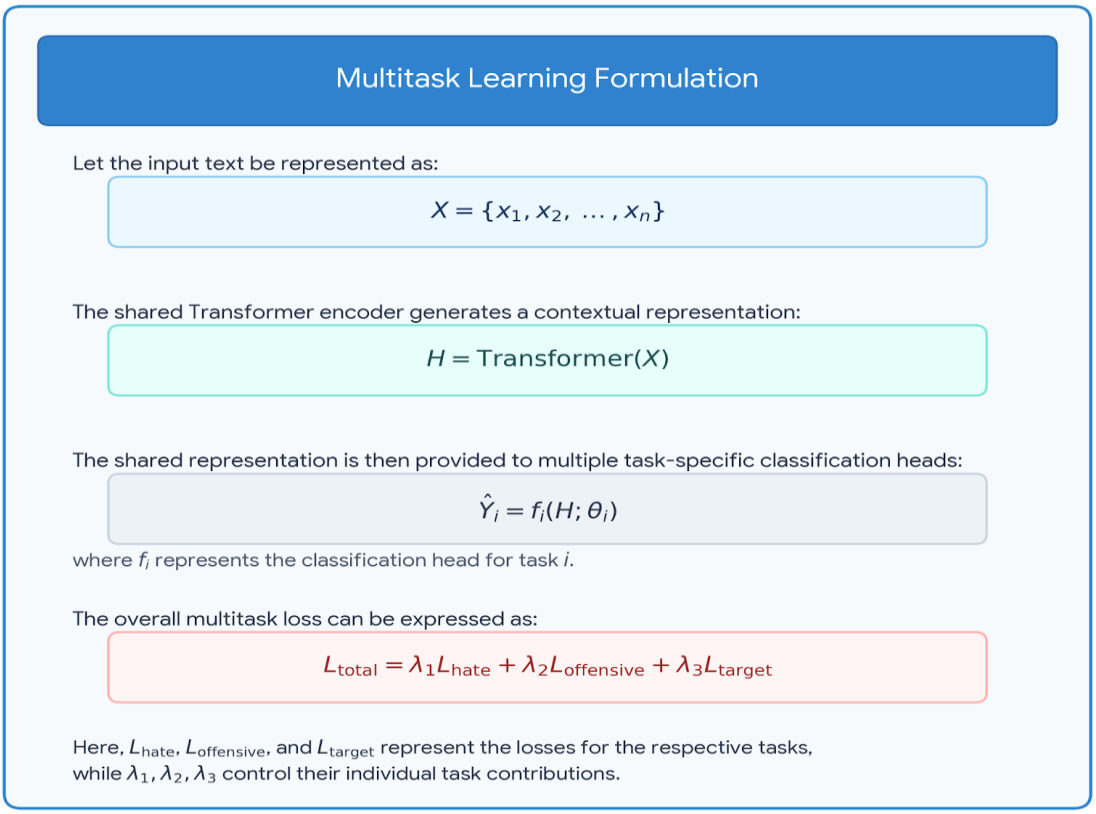

Figure 4

VI. Evaluation Metrics

6.1 Classification

Macro-F1 is important under class imbalance because it gives equal weight to classes. Recall should be reported for the hate class, but precision must also be monitored to avoid overmoderation. Calibration can be measured using expected calibration error or reliability diagrams.

6.2 Explainability

Plausibility compares predicted rationales with human annotations. Faithfulness asks whether the rationale actually supports the prediction. A model can achieve high plausibility by highlighting socially salient words while still being unfaithful. Therefore both categories of metrics are required.

6.3 Robustness

For each perturbation family, report absolute performance and relative degradation. Include examples where the perturbation changes meaning, because a robust system should not be forced to preserve an incorrect label under semantic change.

VII. Results and Discussion Framework

The results section should contain four tables: clean classification, rationale quality, robustness performance, and ablation results. All values are TBD until the implementation is executed. The discussion should identify whether rationale supervision improves the primary task, whether robustness training trades precision for recall, and whether multilingual performance is consistent.

Ablation analysis should remove the rationale head, target head, consistency loss, and robustness augmentation separately. If a component improves explanation quality but reduces classification, the trade-off should be reported rather than hidden. Qualitative examples should include true positives, false positives, false negatives, and obfuscated examples.

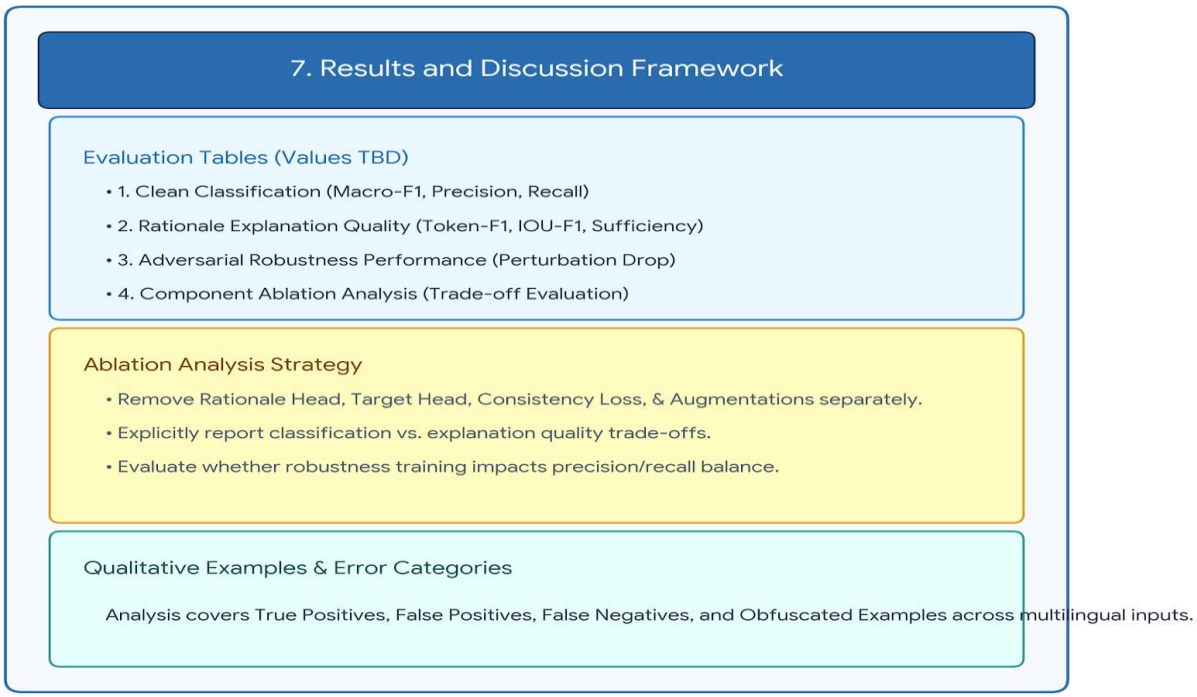

Figure 5

VIII. Ethical, Legal, and Practical Considerations

Explanations must not be treated as proof of intent. Token highlights are model evidence, not a legal or moral judgment. Data may contain slurs and traumatic content; researchers should minimize unnecessary exposure and follow institutional review requirements. Multilingual deployment requires consultation with speakers and domain experts. A model trained on one platform may not generalize to another.

IX. Conclusion and Future Work

This paper proposed an explainable and robust multitask transformer for hate speech detection. The model combines hate classification, offensive-language detection, target identification, rationale extraction, and perturbation consistency. The framework is designed to produce auditable predictions and to measure degradation under realistic evasion strategies. Future work can incorporate conversation context, multimodal memes, retrieval of policy definitions, and human-in-the-loop moderation.

References

[1]. Mathew et al. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. AAAI.
[2]. Rehman et al. (2026). X-MuTeST: A Multilingual Benchmark for Explainable Hate Speech Detection and a Novel LLM-Consulted Explanation Framework. AAAI.
[3]. Devlin et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
[4]. Silva et al. (2026). A Multitask Transformer for Offensive Language Detection and Target Identification in HateBR. PROPOR 2026.
[5]. Kapil and Ekbal (2025). A Transformer Based Multi Task Learning Approach to Multimodal Hate Speech Detection. Natural Language Processing Journal.
[6]. Hegde and Shashirekha (2024). Transformer-driven Multi-task Learning for Fake and Hateful Content Detection. ICON.
[7]. Zampieri et al. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media.
[8]. Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 14867-14875.
[9]. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 512-515.
[10]. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). Proceedings of the 13th International Workshop on Semantic Evaluation, 75-86.
[11]. Vargas, F., Carvalho, I., de Gooijer, F., Zhang, L., & Pardo, T. A. (2022). HateBR: A Large Annotated Dataset for Offensive Language Detection in Brazilian Portuguese. Proceedings of LREC.
[12]. Roy, S., et al. (2026). X-MuTeST: A Multilingual Benchmark for Explainable Hate Speech Detection. arXiv preprint.
[13]. Caruana, R. (1997). Multitask Learning. Machine Learning, 28(1), 41-75.
[14]. Ruder, S. (2017). An Overview of Multi-Task Learning in Deep Neural Networks. arXiv preprint arXiv:1706.05098.
[15]. Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-Task Deep Neural Networks for Natural Language Understanding. Proceedings of ACL, 4487-4496.

[16]. Silva, G., Silva, P., Peixoto, M., Moreira, G., & Luz, E. (2026). A Multitask Transformer for Offensive Language Detection and Target Identification in HateBR. Proceedings of PROPOR, 1049-1054.
[17]. Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2020). Evaluating Multitask Learning for Multi-label Toxicity Classification. Proceedings of the Workshop on Online Abuse and Harassment (WOAH).
[18]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS 30), 5998-6008.
[19]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 4171-4186.
[20]. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.
[21]. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale (XLM-RoBERTa). Proceedings of ACL, 8440-8451.
[22]. Model Explainability & Interpretability
[23]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of ACM SIGKDD, 1135-1144.
[24]. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks (Integrated Gradients). International Conference on Machine Learning (ICML), 3319-3328.
[25]. DeYoung, J., Jain, S., Tan, F. A., Nottingham, A., Suri, C., & Wallace, B. C. (2020). ERASER: A Benchmark for Evaluating Explainable NLP Models. Proceedings of ACL, 4443-4458.
[26]. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. Proceedings of ICLR.
[27]. Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. Proceedings of AAAI, 8018-8025.
[28]. Garg, S., & Ramakrishnan, G. (2020). BAE: BERT-based Adversarial Examples for Text Classification. Proceedings of EMNLP, 6174-6181.