Training-Time Explainability for Multilingual Hate Speech Detection: Aligning Model Reasoning with Human Rationales

M.D.M Qureshi†*
Technological University Dublin
D22124696@mytudublin.ie

Sannaan Khan†
National University of Sciences and Technology
mkhan.msse24sines@student.nust.edu.pk

M. Atif Qureshi
Technological University Dublin
atif.qureshi@tudublin.ie

Wael Rashwann
Maynooth University
wael.rashwan@mu.ie

Abstract

Online hate against Muslim communities often appears in culturally coded, multilingual forms that evade conventional AI moderation. Such systems, though accurate, remain opaque and risk bias, over-censorship, or under-moderation, particularly when detached from sociocultural context. We propose a *training-time* explainability framework that aligns model reasoning with human-annotated rationales, improving both classification performance and interpretability. Our approach is evaluated on HateXplain (English) and BullySent (Hinglish), reflecting the prevalence of anti-Muslim hate across both languages. Using LIME, Integrated Gradients, Grad×Input, and attention, we assess accuracy, explanation quality, and cross-method agreement. Results show that gradient- and attention-based regularization improve F-scores, enhance plausibility and faithfulness, and capture culturally specific cues for detecting implicit anti-Muslim hate, offering a path toward multilingual, culturally aware content moderation.

1 Introduction

Online hate against Muslim communities often appears in subtle, implicit, or culturally coded forms, frequently expressed in multilingual or code-switched language such as Hinglish. These complexities challenge moderation systems, risking both *under-moderation*, which allows harmful narratives to persist, and *over-moderation*, which suppresses legitimate religious or cultural expression [1, 2]. Given its prevalence in both English and South Asian online spaces, tackling such content is essential for Muslim and Global digital safety and inclusivity.

While transformer-based systems achieve high accuracy, they remain opaque and risk reinforcing biases [3, 4], especially when divorced from sociocultural context. For Muslim communities, like other underrepresented groups, these shortcomings can amplify harm and erode trust in AI moderation.

Explainable AI (XAI) offers transparency [5], but post-hoc methods applied after training rarely influence how models learn, limiting their ability to produce faithful, culturally aware explanations [6, 7, 8, 9].

We address this gap with a multilingual, training-time XAI regularization framework that aligns model reasoning with human rationales. Using RoBERTa and XLM-R classifiers, and four explanation strategies - LIME [10], Integrated Gradients [11], Grad×Input [12], and attention [13]-we guide models toward linguistically and culturally meaningful features. Our evaluation on HateXplain (English) and BullySent (Hinglish) demonstrates cross-linguistic generalization for detecting implicit anti-Muslim hate.

*Corresponding Author. †Student Authors

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MusIML.

Contributions: (1) a training-time XAI regularization framework aligning reasoning with human rationales; (2) comparative evaluation of four integrated XAI methods; (3) extension to a multilingual, code-switched setting; (4) analysis of plausibility, faithfulness, and cross-method agreement, highlighting improved detection of implicit, culturally coded anti-Muslim hate; and (5) evidence that gradient- and attention-based regularization offer the best balance of accuracy, explanation quality, and efficiency.[2]

2 Methodology

2.1 Datasets and Pre-processing

We evaluate on two culturally and linguistically distinct datasets to test monolingual and multilingual performance.

HateXplain [14] contains ∼20k English social media posts (Twitter, Gab) labeled as Hate, Offensive, or Normal, with multi-annotator token-level rationales. Following prior work, Hate and Offensive are merged into a *Detrimental Content* class, including *15%* posts explicitly targeting Muslims.

BullySent is a Hinglish (Hindi–English code-switched) cyberbullying dataset from Indian social media, comprising ∼*6.5k* posts. While it does not explicitly mark anti-Muslim hate, it labels religion-based hate, serving as a relevant proxy for South Asian contexts where Hinglish is the dominant medium of informal online discourse [15]. For comparability, all labels are unified into overarching categories: *Detrimental* (e.g., Hate, Offense, Bullying) vs. *Non-Detrimental* content.

2.2 Model Architectures

We employ two transformer encoders to match the linguistic characteristics of each dataset: RoBERTa-base for HateXplain (English) and XLM-RoBERTa-base for BullySent (Hinglish). Each encoder is followed by a Dense ReLU layer, a dropout of 0.3, and a sigmoid output. Models are optimized using Adam with a learning rate of 2×10^{-5}.

2.3 Training-Time Explanation-Based Regularization

Our training augments binary cross-entropy loss with an explanation alignment term, applied only when the gold label is *Detrimental Content*.

The classification loss is:

$$Loss_{BCE} = -\frac{1}{N}\sum_{i=0}^{N} y_t \log(y_p) - (1 - y_t)\log(1 - y_p) \tag{1}$$

where y_t is the true label and y_p the predicted probability.

Let $rat(i)$ and $exp(i)$ denote normalized rationale and explanation scores for each term i in the text input. We aggregate explanation score at word level to ensure compatibility with human-annotated rationales. The explanation loss is:

$$Loss_{EXP}(rat \parallel exp) = \sum_i rat(i) \log\left(\frac{rat(i)}{exp(i)}\right) \tag{2}$$

The total loss is:

$$Loss_{COMP} = Loss_{BCE} + \lambda Loss_{EXP}(rat \parallel exp) \tag{3}$$

with $\lambda \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$ tuned by validation F1-score and early stopping. Algorithm 1 summarizes the process. We use the non-regularized model as our baseline for comparison.

2.4 XAI Methods

We compare four complementary explanation strategies, covering both model-agnostic and model-specific paradigms:

- **LIME** [10]: Model-agnostic and perturbation-based; fits a sparse linear surrogate by masking tokens. We limit $n\_samples$ to 128 for efficiency.

[2]Code and Data at https://github.com/DeedahwarMazhar/NeurIPS-MulsimsInML-TrainingTimeReg

Algorithm 1: Training with Explanation-Based Regularization

Input: Training set $D = \{(x_i, y_i, rat_i)\}_{i=1}^{N}$, XAI method E, regularization weight λ
Output: Trained model f_θ

for *each epoch* **do**
  for *each batch* (x, y, rat) *in* D **do**
    $y_p \leftarrow f_\theta(x)$; // Forward pass
    $Loss_{BCE} \leftarrow \text{BinaryCrossEntropy}(y, y_p)$
    if $y == 1$ **then**
      $exp \leftarrow E(f_\theta, x)$; // Generate explanation
      if *exp is token-level* **then**
        $exp \leftarrow \text{ConvertTokenToWord}(exp)$; // Aggregate word scores
      $Loss_{EXP} \leftarrow \text{KL}(rat \parallel exp)$; // Explanation loss
    $Loss_{COMP} \leftarrow Loss_{BCE} + \lambda \cdot Loss_{EXP}$
    $\theta \leftarrow \theta - \eta \cdot \nabla_\theta Loss_{COMP}$; // Backpropagation
  if *early stopping criterion met on validation set* **then**
    break

- **Integrated Gradients (IG)** [11]: Integrates gradients from a zero embedding baseline to the input over 15 steps, capturing non-linear dependencies.
- **Grad×Input (GXI)** [12]: Multiplies token embeddings with their gradients to estimate local feature sensitivity in one backward pass.
- **Attention** [13]: Uses averaged final-layer attention weights across heads as token importance; not always faithful but internally consistent for regularization.

2.5 Post-Hoc Explanation Evaluation

We evaluate XAI methods by comparing post-hoc attributions to human-annotated rationales. Scores are normalized to probability distributions over tokens, and subword attributions (IG, GXI, Attention) are merged to word-level before alignment. Word scores are binarized into rationale masks using:

1. **Top-K:** highest $K \in \{3, 5, 7, 10, 15, 20, 25\}$ tokens (LIME: $K \leq 10$).
2. **Thresholding:** tokens with scores $> \theta \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$.

Human rationales are aggregated via the *Union* strategy.
We report:

- **Plausibility:** *IoU* (overlap proportion) and *Token-F1* (token-level precision–recall) vs. human rationales.
- **Faithfulness:** *Prediction Drop*—confidence decrease when identified tokens are removed.

This token-level alignment captures fine-grained linguistic/contextual cues, and comparing Top-K vs. thresholding highlights trade-offs between fixed- and variable-length rationales.

3 Results and Discussion

Table 1 reports classification (Acc, F1), plausibility (IoU, Token-F1), and faithfulness (Prediction Drop) for HateXplain (HX) and BullySent (BS).

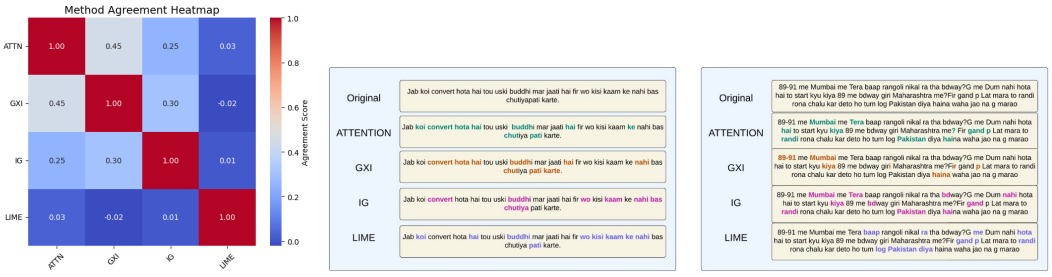
Classification. Attention achieves the best overall F1 (HX: **0.84**, BS: **0.85**) and highest BS accuracy (**0.86**). LIME and IG perform competitively on HX, while G×I is efficient but slightly less accurate.

Plausibility. On HX, Attention leads in IoU/Token-F1 (**0.38/0.60**), with IG second. On BS, LIME attains the best Top-K plausibility (IoU **0.27**, Token-F1 **0.45**), suggesting perturbations capture Hinglish cues better than attention or gradients.

Faithfulness. IG yields the largest drops on both datasets (HX: **0.72/0.74**; BS: **0.82**), indicating strong causal linkage. G×I performs well on BS, while LIME's drops remain low. Gradient-based methods transfer faithfulness more robustly across languages, while plausibility varies by dataset and binarization strategy.

Method	Data	Perf.		IoU		Token-F1		Drop	
		Acc	F1	Top-K	Thresh.	Top-K	Thresh.	Top-K	Thresh.
No Regularizer	HX	0.78	0.82	—	—	—	—	—	—
	BS	0.84	0.83	—	—	—	—	—	—
LIME	HX	0.80	0.84	0.09	0.11	0.19	0.22	0.25	0.26
	BS	0.85	0.84	**0.27**	0.11	**0.45**	**0.34**	0.41	0.45
Integrated Gradients	HX	0.79	0.83	0.21	0.28	0.39	0.44	**0.72**	**0.74**
	BS	0.84	0.84	0.14	0.12	0.23	0.22	0.41	**0.82**
Grad×Input	HX	0.78	0.82	0.21	0.19	0.38	0.35	0.63	0.35
	BS	0.85	0.85	0.07	0.06	0.16	**0.12**	**0.68**	0.54
Attention	HX	**0.80**	**0.84**	**0.33**	**0.38**	**0.58**	**0.60**	0.64	0.60
	BS	**0.86**	**0.85**	0.11	0.12	0.21	0.17	0.30	0.30

Table 1: Combined classification and XAI performance. HX = HateXplain, BS = BullySent. Perf. = Accuracy, F1. IoU and Token-F1 measure plausibility; Drop measures faithfulness. Best scores per metric (HX and BS) are in bold.



(a) Method Agreement Heatmap for XAI methods with Kendall's Agreement Correlation (τ).

(b) Qualitative comparison of token-level attributions from four XAI methods across two code-switched hate speech examples. While gradient-based methods (Attention, G×I, IG) frequently assign salience to subword fragments, LIME operates at the word level and often misses parts of compound slurs.

Figure 1: Kendall's τ agreement between XAI methods (left), showing higher correlation among gradient-based approaches and divergence from LIME. Qualitative token-level attributions for two representative code-switched hate speech examples (right) illustrate method-specific highlighting patterns.

Inter-Method Agreement: We measure similarity via Kendall's τ, suitable for comparing ordered token attributions. Figure 1a shows highest alignment between Attention and G×I (≈ 0.45), with moderate scores for Attention–IG and G×I–IG, and near-zero or negative correlations for LIME due to its word-level attributions.

Qualitative Comparison. Figure 1b visualizes token-level highlights for a representative post with different XAI methods. Gradient-based approaches often mark subword units within compounds, enabling finer-grained detection of culturally specific or offensive expressions. In contrast, perturbation-based methods tend to operate at the whole-word level and can overlook embedded cues.

4 Conclusion

We present a training-time explainability framework that embeds human-aligned rationales into hate speech classifiers, improving both plausibility and faithfulness across English and Hinglish datasets. Results show that gradient- and attention-based regularization generalize well across languages, capturing culturally specific cues often missed by perturbation-based methods, and offering a path toward more accountable multilingual moderation.

Acknowledgemets

This publication has emanated from research conducted with the financial support of Research Ireland (RI) Center for Research Training in Machine Learning (ML-Labs), under Grant number 18/CRT/6183. For Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

[1] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[2] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

[3] Sahana Udupa, Antonis Maronikolakis, and Axel Wisiorek. Ethical scaling for content moderation: Extreme speech and the (in) significance of artificial intelligence. *Big Data & Society*, 10(1):20539517231172424, 2023.

[4] Eugenia Siapera. Ai content moderation, racism and (de) coloniality. *International journal of bullying prevention*, 4(1):55–65, 2022.

[5] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18, 2018.

[6] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28:689–707, 2018.

[7] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[8] Muhammad Deedahwar Mazhar Qureshi, M Atif Qureshi, and Wael Rashwan. Toward inclusive online environments: Counterfactual-inspired xai for detecting and interpreting hateful and offensive tweets. In *World Conference on Explainable Artificial Intelligence*, pages 97–119. Springer, 2023.

[9] Muhammad Deedahwar Mazhar Qureshi, M Atif Qureshi, and Wael Rashwan. Explainable ai for hate speech moderation: A stakeholder-centered and socially grounded review. *Authorea Preprints*, 2025.

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[11] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.

[12] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[13] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[14] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875, 2021.

[15] Meghana Bhange and Nirant Kasliwal. Hinglishnlp: Fine-tuned language models for hinglish sentiment detection. *arXiv preprint arXiv:2008.09820*, 2020.

[16] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, 2019.

[18] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

[19] Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300, 2020.

[20] Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. Disembodied machine learning: On the illusion of objectivity in nlp. *arXiv preprint arXiv:2101.11974*, 2021.

[21] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*, 2021.

[22] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[23] Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. *arXiv preprint arXiv:2205.04238*, 2022.

[24] Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. Counterfactual inference for text classification debiasing. 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)*, pages 5434–5445, 2021.

[25] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics.

[26] Sofia Serrano and Noah A. Smith. Is attention interpretable? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics.

[27] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670, 2017.

[28] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR, 2020.

[29] Samuel Carton, Surya Kanoria, and Chenhao Tan. What to learn, and how: Toward effective learning from rationales. *arXiv preprint arXiv:2112.00071*, 2021.

[30] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics.

Related Studies	Multiple XAI Methods	XAI-based Regularization	Explanation Optimization	XAI Evaluations
Mathew et al. [14]		✓		✓
Ross et al. [27]	✓	✓		
Hancock et al. [32]		✓	✓	
Balkir et al. [33]	✓			✓
Our Study	✓	✓	✓	✓

Table 2: Comparison of selected related studies with our work.

[31] Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. Measuring association between labels and free-text rationales. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[32] Braden Hancock, Martin Bringmann, Paroma Varma, Percy Liang, Stephanie Wang, and Christopher Ré. Training classifiers with natural language explanations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 1884, 2018.

[33] Esma Balkir, Isar Nejadgholi, Kathleen C Fraser, and Svetlana Kiritchenko. Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection. *arXiv preprint arXiv:2205.03302*, 2022.

A Related Work

AI-based content moderation has evolved from keyword matching and simple classifiers [16] to transformer-based models such as BERT [17] and multilingual encoders [18]. While these improve accuracy, they still struggle with sarcasm, implicit hate [19], and culturally coded language, often producing opaque or biased decisions across linguistic and cultural boundaries [20]. Human-in-the-loop approaches [21] highlight that moderation requires cultural and political context [3, 4], especially for communities facing targeted hostility.

Explainable AI (XAI) offers interpretability via methods such as LIME [10], SHAP [22], attention [13], and counterfactuals [23, 24]. However, post-hoc explanations can be unfaithful [25, 26] and often lack integration with domain knowledge, which is critical for detecting subtle, culturally specific hate speech against marginalized groups.

Training-time XAI regularization, enabled by datasets like HateXplain [14], aligns model explanations with human rationales [27, 28, 29], but most work uses a single XAI method and monolingual data. Few evaluate both plausibility and faithfulness [30, 31] or assess consistency across explanation methods. This gap is pronounced in multilingual and code-switched contexts like Hinglish, where implicit anti-Muslim hate is prevalent. Our work addresses this by integrating multiple XAI methods into training, applying them to both English and Hinglish datasets, and jointly evaluating cultural relevance, explanation quality, and cross-method agreement (Table 2).

B Extended Methodology

B.1 Datasets

We evaluate on two datasets with token-level human rationales: **HateXplain** [14] contains $\sim$19k social media posts from Twitter and Gab, annotated for *Hate*, *Offensive*, or *Normal* content, along with rationales from multiple annotators. Following prior work, we merge *Hate* and *Offensive* into a single **Detrimental Content** label.

BullySent is a Hinglish (Hindi–English code-switched) dataset of $\sim$6.4k posts annotated for abusive content, also with token-level rationales. Compared to HateXplain, BullySent contains greater lexical

Dataset	Detrimental	Non-Detrimental	Total
HateXplain	11,415	7,814	19,299
BullySent	3,451	2,985	6,436

Table 3: Dataset distribution across both corpora.

variation, frequent spelling inconsistencies, and transliteration noise, making rationale alignment more challenging.

For evaluation, we aggregate annotator rationales using the **Union** strategy to ensure inclusive coverage of all highlighted tokens.

B.2 Training Pipeline

Figure 2 outlines our training procedure. For HateXplain, we fine-tune a RoBERTa encoder, while for BullySent (Hinglish), we use XLM-R to better handle multilingual and code-switched text. Both models employ a binary classification head and are optimized with a composite loss combining standard binary cross-entropy with an explanation-alignment term. For correctly predicted detrimental-content examples, we generate token-level importance scores using a given XAI method, normalize them, and align them with human rationales via KL divergence.

The regularization weight λ controls the trade-off between accuracy and explanation alignment, tuned separately for each dataset. The optimal value was $\lambda = 5$ for HateXplain and $\lambda = 0.05$ for BullySent, suggesting that in the multilingual, noisier BullySent setting, heavy regularization can hurt performance, whereas HateXplain benefits from stronger alignment pressure.

This approach is applied consistently across all four XAI methods studied: Attention, Gradient$\times$Input, Integrated Gradients, and LIME.

C Extended Explainability Analysis Across Datasets

To assess the generality of our findings, we evaluate explanation plausibility and faithfulness on both HateXplain and BullySent using the same four XAI methods (Attention, Gradient$\times$Input, Integrated Gradients, and LIME) and two rationale selection strategies (Top-K and Thresholding). For each dataset, we plot metric trends over varying K and τ values, with all curves using identical scales to enable direct visual comparison. Figures 3 present all twelve plots (six per dataset), while the following discussion synthesizes the results across datasets.

C.1 Intersection over Union

Across both datasets, Attention consistently achieves the highest IoU values under both selection strategies, indicating strong alignment with human rationales. Integrated Gradients (IG) follows closely, especially under thresholding, where it benefits from capturing non-linear token dependencies. Gradient$\times$Input lags in plausibility on both datasets, while LIME exhibits the lowest IoU scores, particularly for higher K values. Notably, BullySent scores are generally lower than HateXplain across methods, suggesting that shorter, noisier, and more code-switched inputs make exact token overlap harder to achieve.

C.2 Token-F1

Token-level F1 patterns are remarkably consistent across datasets: attention explanations best match individual human-annotated tokens, followed by IG. Thresholding continues to outperform Top-K for most methods, reflecting that selecting tokens above a score threshold better preserves rationale boundaries—especially beneficial for HateXplain's longer texts. On BullySent, all methods show reduced F1 scores relative to HateXplain, likely due to higher annotation sparsity and linguistic variability, but the relative ranking of methods remains stable.

C.3 Prediction Drop

Prediction Drop trends reaffirm IG as the most faithful method in both datasets: masking its top-ranked tokens produces the largest confidence decreases, indicating strong causal alignment with the model's decision process. Attention also yields substantial drops, though with more variability across

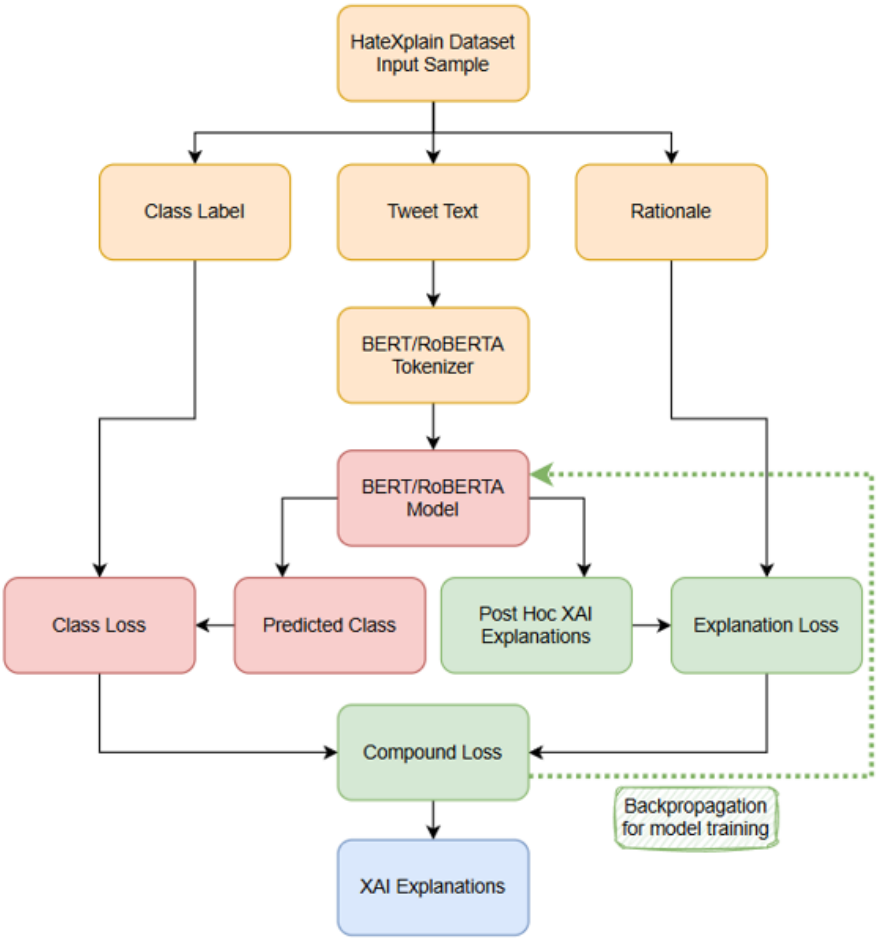


Figure 2: Training pipeline with integrated explanation regularization.

K and τ. LIME maintains low prediction drop values across datasets, reinforcing critiques of its faithfulness in high-dimensional, context-dependent NLP settings. Interestingly, BullySent exhibits smaller overall drop values compared to HateXplain, which may be due to its shorter, more direct statements where a few key tokens suffice for classification.

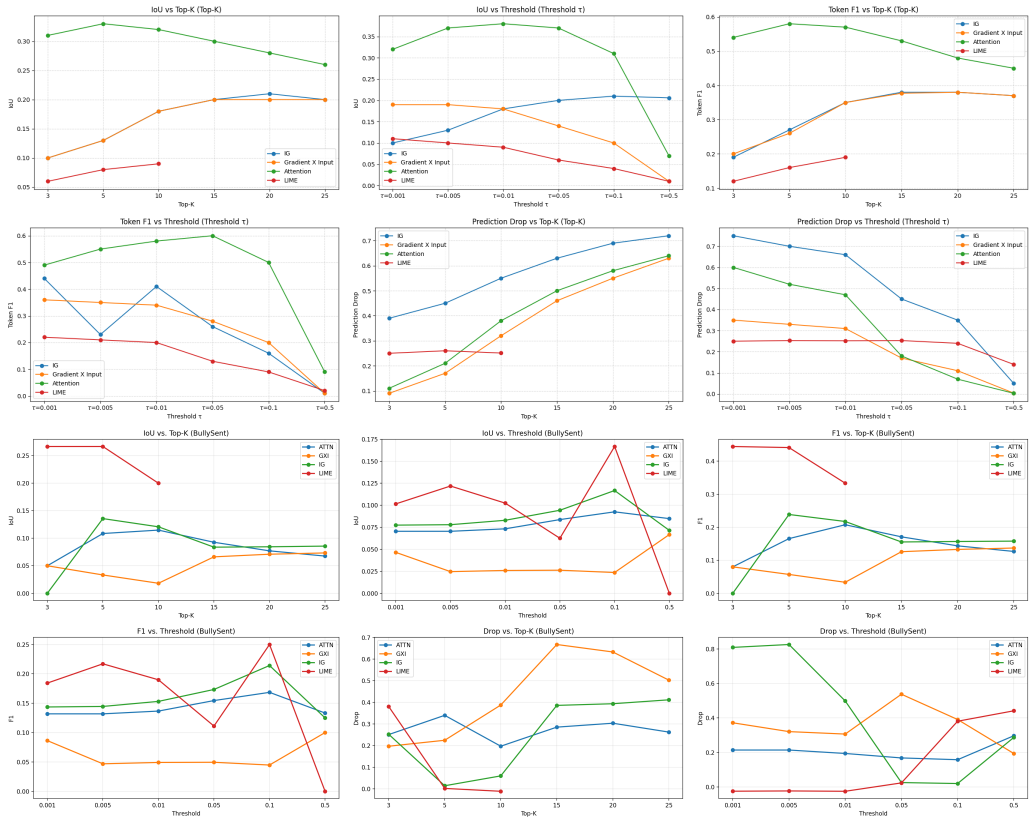


Figure 3: Variation of IoU, Token-F1, and Prediction Drop across Top-K and Threshold values for HateXplain (top two rows) and BullySent (bottom two rows). LIME's Top-K curves terminate at $K{=}10$ due to sampling constraints.