Identifying Contextual Triggers in Hate Speech Texts Using Explainable Large Language Models

Dheeraj Kodati[1] Lakkireddy Bhuvana Sree[1]

dheeraj.kodati@mahindrauniversity.edu.in se24maid009@mahindrauniversity.edu.in

[1]Department of Computer Science and Engineering, Mahindra University, Hyderabad, India

Abstract

The pervasive spread of hate speech on online platforms poses a significant threat to social harmony, necessitating not only high-performing classifiers but also models capable of transparent, fine-grained interpretability. Existing methods often neglect the identification of influential contextual words that drive hate speech classification, limiting their reliability in high-stakes applications. To address this, we propose LLM-BiMACNet (Large Language Model-based Bidirectional Multi-Channel Attention Classification Network), an explainability-focused architecture that leverages pretrained language models and supervised attention to highlight key lexical indicators of hateful and offensive intent. Trained and evaluated on the HateXplain benchmark—comprising class labels, target community annotations, and human-labeled rationales—LLM-BiMACNet is optimized to simultaneously enhance both predictive performance and rationale alignment. Experimental results demonstrate that our model outperforms existing state-of-the-art approaches, achieving an accuracy of 87.3%, AUROC of 0.881, token-level F1 of 0.553, IOU-F1 of 0.261, AUPRC of 0.874, and comprehensiveness of 0.524, thereby offering highly interpretable and accurate hate speech detection.

1 Introduction

Hate speech on social media has surged dramatically in recent years, posing serious challenges to social cohesion, public safety, and digital platform governance. The contextual and nuanced nature of hate speech—often encoded in subtle phrasing or idiomatic expressions—makes it difficult for automated systems to distinguish between benign and harmful content (Vijayaraghavan and Vosoughi, 2021; Kodati, 2020; Das et al., 2025a). Furthermore, users frequently manipulate hateful content (e.g., via typos or benign interjections like "love") to evade detection, underscoring the need for models that understand the semantic intent rather than simply relying on surface-level features (Garg et al., 2023; Kodati and Tene, 2024a,b). Recent studies have emphasized the importance of interpretable and explainable hate speech detectors, which not only classify content but also identify the specific tokens that drive the decision (Kim et al., 2022; Yang et al., 2023; Kodati and Dasari, 2025b; Das et al., 2024, 2025b). The HateXplain dataset represents a notable advancement in this direction, providing human-annotated rationales at token level, alongside class labels and target community annotations (Mathew et al., 2021). While supervised-attention methods like Masked Rationale Prediction attempt to align model decisions with human reasoning, there remains substantial room for improvement in rationale plausibility and faithfulness (Kim et al., 2022), (Das et al., 2022). More recently, studies such as HARE (Yang et al., 2023) and LLM-based explanation models (Nirmal et al., 2024; Kodati and Dasari, 2025a) have demonstrated that integrating large language models (LLMs) with supervised rationale alignment can significantly enhance the interpretability and generalization of hate speech classifiers. **Key contributions of our work include:** identification of contextual words responsible for hate and offensive content using explainable attention mechanisms; integration of LLM-guided rationale alignment to improve interpretability without compromising classification performance; and comprehensive evaluation on the HateXplain dataset, demonstrating superior accuracy and explanation quality compared to state-of-the-art models.

2 Related Work

Detecting hate speech has evolved from rule-based and keyword-matching systems to deep

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 51–58, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-007

neural architectures, driven by the increasing need for both accuracy and transparency. Early transformer-based models such as BERT and RoBERTa achieved strong performance in offensive language classification, yet lacked the capability to explain why certain messages were flagged as hateful. To address this, models incorporating attention visualization and rationale supervision have emerged. Vijayaraghavan et al. (Vijayaraghavan and Vosoughi, 2021) proposed a multi-modal framework that combines textual content and social metadata for interpretable hate speech detection, leveraging attention weights to identify influential components in the input. Similarly, Kim et al. (Kim et al., 2022) introduced the Masked Rationale Prediction (MRP) method, which masks annotated rationales during training to encourage the model to attend to human-identified evidential spans. These approaches laid the groundwork for integrating explainability with detection but remain limited in generalization and token-level faithfulness. More recent studies have begun LLMs for explanation-aware classification (Kodati and Ramakrishnudu, 2023, 2021). Yang et al. (Yang et al., 2023) introduced the HARE framework, which uses step-by-step explanations generated by an LLM to provide hierarchical and interpretable decisions for hate speech detection. In a similar vein, Nirmal et al. (Nirmal et al., 2024) proposed a framework where rationales are extracted from LLMs and used as supervised signals to guide model attention, resulting in more aligned token-level predictions with human rationales. These models demonstrated improvements not only in classification metrics but also in explainability scores such as comprehensiveness and sufficiency. Böck et al. (Böck et al., 2024) further evaluated several interpretability methods (gradient-based, perturbation-based, and attention-based) and concluded that perturbation-based methods yield the most plausible explanations, although they are computationally expensive. To understand broader challenges in hate speech detection, recent surveys provide comprehensive overviews of current approaches. Kapil and Ekbal (Kapil and Ekbal, 2024) reviewed over 60 models, highlighting trends in explainable AI and the need for robust rationale supervision. The work (Kodati and Dasari, 2024) emphasized limitations such as benchmark inconsistency, algorithmic bias, and the lack of explainable metrics in evaluation protocols. Liu et al. (Jahan and Oussalah, 2023) examined hybrid architectures combining handcrafted features and deep representations, identifying a clear shift toward supervised explanation mechanisms using annotated datasets like HateXplain. Despite these efforts, existing methods often face a trade-off between performance and transparency. Our work builds on these foundations by integrating LLM-derived rationales within a supervised attention pipeline to achieve both faithful interpretability and competitive performance on standard hate speech benchmarks.

3 Preliminaries

3.1 Problem Statement

Let $\mathcal{D} = \{(x^{(i)}, y^{(i)}, r^{(i)})\}_{i=1}^{N}$ be a labeled dataset where each $x^{(i)} = \{w_1^{(i)}, w_2^{(i)}, \ldots, w_T^{(i)}\}$ is a tokenized input text sequence of length T, $y^{(i)} \in \mathcal{Y}$ is the class label (e.g., `Hate`, `Offensive`, `Normal`), and $r^{(i)} \in \{0,1\}^T$ is a binary rationale vector where $r_t^{(i)} = 1$ if token $w_t^{(i)}$ is annotated as a rationale (i.e., contributes to the label $y^{(i)}$), and 0 otherwise. The goal is to learn a classification model $f_\theta(x)$ that satisfies two objectives: (1) accurate prediction of y given x, and (2) faithful alignment of the model's explanation with the human-provided rationale r.

More formally, we seek to optimize the following composite objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}}(f_\theta(x), y) + \lambda \cdot \mathcal{L}_{\text{exp}}(e_\theta(x), r) \quad (1)$$

where $\mathcal{L}_{\text{cls}}$ is the supervised classification loss (e.g., cross-entropy), $\mathcal{L}_{\text{exp}}$ is the rationale alignment loss (e.g., binary cross-entropy between model explanation and r), $e_\theta(x)$ is the explanation generated by the model (e.g., attention or importance scores), and λ controls the trade-off between accuracy and interpretability.

3.2 Input Representation via LLM Encoding

Given the input sequence $x = \{w_1, w_2, \ldots, w_T\}$, we pass it through a pretrained large language model (LLM), such as RoBERTa, to obtain contextualized token representations. Denote the LLM encoder as $\phi(\cdot)$, then:

$$H = \phi(x) = \{h_1, h_2, \ldots, h_T\}, \quad h_t \in \mathbb{R}^d \quad (2)$$

where $H \in \mathbb{R}^{T \times d}$ is the sequence of contextual embeddings and d is the hidden dimension. These

embeddings form the base input to the downstream model components for classification and explanation.

3.3 Rationale Supervision and Token-level Alignment

To incorporate human-annotated rationales, we introduce an attention-like mechanism $e_\theta(x) = \{\alpha_1, \alpha_2, \ldots, \alpha_T\}$ where $\alpha_t \in [0, 1]$ denotes the importance score of token w_t. These scores are trained to align with the ground-truth rationale vector r using binary cross-entropy:

$$\mathcal{L}_{\text{exp}} = -\sum_{t=1}^{T} [r_t \cdot \log \alpha_t + (1 - r_t) \cdot \log(1 - \alpha_t)] \tag{3}$$

This ensures that the model focuses its interpretive capacity on tokens that are genuinely responsible for the classification decision. Moreover, we enforce that explanations are not only plausible (aligning with r) but also faithful (i.e., their removal degrades the prediction confidence), which is evaluated using comprehensiveness and sufficiency metrics in experiments.

3.4 Prediction Objective

The final classification logits z are computed from a sequence-level representation v, which may be derived through operations such as max pooling, attention-weighted summation, or recurrent aggregation over H. The class prediction is obtained by:

$$z = Wv + b, \quad \hat{y} = \arg\max(\text{softmax}(z)) \tag{4}$$

where $W \in \mathbb{R}^{|\mathcal{Y}| \times d}$ and $b \in \mathbb{R}^{|\mathcal{Y}|}$ are trainable parameters. The model is optimized end-to-end using the total loss $\mathcal{L}_{\text{total}}$ from Equation (1), jointly training for both label prediction and rationale alignment.

4 Methodology

This section describes the architecture of our proposed model, **LLM-BiMACNet**, which is designed to classify hateful content and highlight the most influential contextual tokens. The architecture integrates deep contextual representations from pretrained language models with hierarchical neural processing and attention-based rationale supervision.

4.1 Overview

Given an input sequence $x = \{w_1, w_2, \ldots, w_T\}$, our model performs three primary operations: (1) extract contextual embeddings using a pretrained large language model (LLM), (2) process the sequence through a bidirectional multi-channel attention architecture for rich feature interaction, and (3) jointly optimize for classification accuracy and token-level rationale alignment. The complete architecture is illustrated in Figure 1.

4.2 Contextual Encoding via LLM

We begin by transforming the input sequence into contextual embeddings using a pretrained language model $\phi(\cdot)$, such as RoBERTa:

$$H = \phi(x) = \{h_1, h_2, \ldots, h_T\}, \quad h_t \in \mathbb{R}^d \tag{5}$$

These embeddings capture semantic and syntactic dependencies between tokens and serve as input to the next stages of the network.

4.3 Bidirectional Sequential Encoding

To capture sequential dependencies in both forward and backward directions, we employ a bidirectional recurrent structure on top of the LLM embeddings:

$$\begin{aligned} \overrightarrow{h}_t &= \text{GRU}_{\text{fwd}}(h_t, \overrightarrow{h}_{t-1}), \\ \overleftarrow{h}_t &= \text{GRU}_{\text{bwd}}(h_t, \overleftarrow{h}_{t+1}) \end{aligned} \tag{6}$$

The final sequence representation from this layer is:

$$H^{\text{Bi}} = \{[\overrightarrow{h}_t; \overleftarrow{h}_t]\}_{t=1}^{T}, \quad H^{\text{Bi}} \in \mathbb{R}^{T \times 2d} \tag{7}$$

4.4 Multi-Channel Attention Mechanism

To emphasize different semantic aspects, we apply a multi-channel attention mechanism over the Bi-GRU output. Each attention head computes a distribution over the token representations:

$$\alpha_t^{(j)} = \frac{\exp\left(\mathbf{w}_j^\top \tanh(W_j H_t^{\text{Bi}} + b_j)\right)}{\sum_{k=1}^{T} \exp\left(\mathbf{w}_j^\top \tanh(W_j H_k^{\text{Bi}} + b_j)\right)}, \quad \text{for } j = 1, \ldots, M \tag{8}$$

where M is the number of attention channels (or heads), and each head focuses on a distinct subspace of semantic relevance. The final aggregated

representation is the concatenation of all head-wise weighted sums:

$$v = \bigoplus_{j=1}^{M} \sum_{t=1}^{T} \alpha_t^{(j)} H_t^{\text{Bi}} \tag{9}$$

4.5 Global Feature Abstraction and Classification

The output vector v from multi-head attention is passed through a convolutional feature extractor followed by global max pooling (GMP) to obtain a fixed-length high-level abstraction:

$$F = \text{GMP}(\text{ReLU}(\text{Conv1D}(v))) \tag{10}$$

The final classification logits are computed using a fully connected layer with softmax activation:

$$z = W_{\text{cls}} F + b_{\text{cls}}, \quad \hat{y} = \arg\max(\text{softmax}(z)) \tag{11}$$

4.6 Explanation Generation and Supervision

To make the model's predictions interpretable, we define a token-level importance score vector $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_T\}$ obtained from one of the attention heads trained for explanation. This head is supervised using the binary rationale vector r from the HateXplain dataset:

$$\mathcal{L}_{\text{exp}} = -\sum_{t=1}^{T} [r_t \log \alpha_t + (1 - r_t) \log(1 - \alpha_t)] \tag{12}$$

This encourages the attention distribution to align with human-provided explanations.

4.7 Joint Optimization Objective

The complete model is trained end-to-end with a multi-objective loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{exp}} \tag{13}$$

Here, $\mathcal{L}_{\text{cls}}$ is the standard categorical cross-entropy loss, $\mathcal{L}_{\text{exp}}$ is the rationale alignment loss, and λ is a hyperparameter balancing accuracy and interpretability.

The proposed LLM-BiMACNet algorithm 1 performs hate speech classification while simultaneously identifying the contextual words that contribute most to the prediction using supervised explainability. Given an input text, the model first encodes it using a pretrained LLM to capture rich contextual embeddings. These embeddings are processed through a BiGRU and multi-head attention mechanism to compute token-level importance scores. During training, the model optimizes both classification accuracy and explanation alignment by comparing its attention scores to human-annotated rationales. At inference, it outputs not only the predicted class (Hate, Offensive, or Normal) but also the specific tokens with high importance scores—effectively highlighting the contextual words that influenced the decision.

Algorithm 1: LLM-BiMACNet: Explainable Hate Speech Detection

Input: Tokenized input $x = \{w_1, w_2, \ldots, w_T\}$, true label $y \in \mathcal{Y}$, rationale vector $r = \{r_1, \ldots, r_T\}$
Output: Predicted label $\hat{y}$, contextual tokens $C \subseteq x$

Function `TrainModel(`$\mathcal{D} = \{(x^{(i)}, y^{(i)}, r^{(i)})\}$`)`**:**
  Initialize model parameters θ ;
  foreach *epoch* $= 1$ *to* E **do**
    foreach *batch* (x, y, r) *in* $\mathcal{D}$ **do**
      $(\hat{y}, \boldsymbol{\alpha}) \leftarrow$ `ExplainableForward(`x`)` ;
      $\mathcal{L}_{\text{cls}} \leftarrow \text{CrossEntropy}(\hat{y}, y)$;
      $\mathcal{L}_{\text{exp}} \leftarrow -\sum_{t=1}^{T} [r_t \log \alpha_t + (1 - r_t) \log(1 - \alpha_t)]$;
      $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{exp}}$;
      Update: $\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}_{\text{total}}$;
  return Trained model θ ;

Function `ExplainableForward(`x`)`**:**
  $H \leftarrow \phi(x)$; `// LLM contextual embeddings`
  $H^{\text{Bi}} \leftarrow \text{BiGRU}(H)$;
  Compute attention scores $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_T\}$;
  $z \leftarrow \text{CNN} \rightarrow \text{ReLU} \rightarrow \text{GMP} \rightarrow \text{FC}$;
  $\hat{y} \leftarrow \arg\max(\text{softmax}(z))$;
  $C \leftarrow \{w_t \in x \mid \alpha_t > \tau\}$;
  return $(\hat{y}, C)$;

Figure 1 illustrates the compact dual-channel architecture of LLM-BiMACNet, where shared LLM and BiGRU layers extract contextual representations from the input text. These representations are then processed by two parallel branches: one for hate speech classification using multi-head attention and CNN layers, and the other for explainability using a supervised attention head that highlights contextual words contributing to each prediction.

5 Dataset Collection

To evaluate our proposed model LLM-BiMACNet in terms of both classification performance and explanation fidelity, we utilize the publicly available

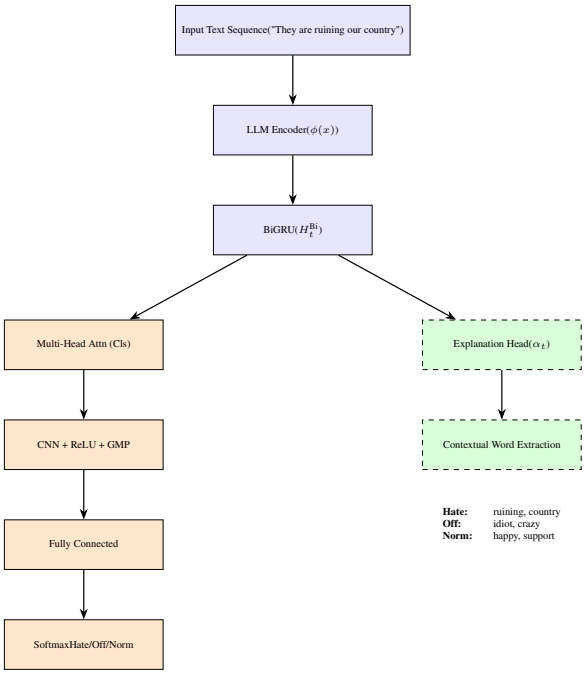

Figure 1: LLM-BiMACNet model architecture

HateXplain dataset (Mathew et al., 2021). This benchmark is specifically designed for explainable hate speech detection and provides not only class labels but also human-annotated rationales at the token level, making it well-suited for training and evaluating models with interpretable attention mechanisms.

5.1 Dataset Composition

The HateXplain dataset consists of over 20,000 social media posts, primarily sourced from **Twitter** and **Gab**. Each post is annotated by three independent annotators from Amazon Mechanical Turk (AMT), providing:

- A **class label** from the set $\{\texttt{Hate}, \texttt{Offensive}, \texttt{Normal}\}$.
- A **target community** label (e.g., religion, ethnicity, gender).
- A **rationale vector** indicating which words contribute to the label assignment.

The rationales are marked at the token level, allowing models to be trained not only for accurate classification but also for interpretable decision-making.

5.2 Annotation and Agreement

Annotators were required to justify their decisions by highlighting the specific words that led them to assign a given label. A majority voting scheme was employed to determine the final class label for each post. To ensure annotation consistency, only samples where at least two annotators agreed on both class and rationale were retained. This filtering step improves the quality of supervision for both classification and explanation tasks.

5.3 Rationale Aggregation

The final rationale mask for each input sequence is derived by aggregating token-level selections from the agreeing annotators. Each token w_t is associated with a binary label $r_t \in \{0, 1\}$, where $r_t = 1$ indicates that the token contributes to the hateful or offensive nature of the text. These rationale vectors are used as ground truth for supervising the explanation component of our model.

5.4 Train-Validation-Test Splits

We follow the standard data partition provided by the authors of HateXplain, using 16,043 samples for training, 1,927 for validation, and 1,969 for testing. All experiments are conducted using this split to ensure reproducibility and comparability with prior work.

5.5 Why HateXplain?

Unlike traditional hate speech datasets, HateXplain includes fine-grained human explanations, enabling us to train and evaluate models on rationale alignment, explanation plausibility, and faithfulness. Its inclusion of target community tags also supports bias-sensitive evaluation, making it ideal for explainable and responsible AI research in toxic language detection.

6 Experimental Results

We evaluate the proposed LLM-BiMACNet model on the HateXplain dataset to assess its effectiveness in both classification and explainability. The dataset contains over 20,000 posts across three classes—Hate (10%), Offensive (30%), and Normal (60%)—with an average text length of approximately 23 words per post. Around 70% of the posts include annotated rationales highlighting hateful or offensive spans.

6.1 Preprocessing

Prior to model training, all input samples were lowercased, and special characters (e.g., emojis, URLs, hashtags) were normalized using regular

expressions. Tokenization was performed using the RoBERTa tokenizer from HuggingFace's Transformers library, which is compatible with our pretrained language model. To maintain sequence consistency, we truncated or padded inputs to a maximum length of 128 tokens. For rationale alignment, human-annotated rationale vectors were converted into binary token-level masks aligned with subword tokenization. All labels were mapped to categorical indices: `Hate` (0), `Offensive` (1), and `Normal` (2).

6.2 Hyperparameter Settings

The model was trained using the AdamW optimizer with a learning rate of 2×10^{-5} and weight decay of 0.01. A batch size of 16 was used, and training was conducted for up to 10 epochs with early stopping based on validation loss. The loss balancing parameter λ for rationale supervision was set to 0.5 based on grid search. The hidden dimension for BiGRU was set to 256, and we used 4 attention channels in the multi-channel attention mechanism. The model uses RoBERTa-base as the contextual encoder to generate token-level embeddings of dimension 768. Dropout with a rate of 0.3 was applied to all intermediate layers to prevent overfitting. Experiments were conducted on an NVIDIA RTX 3090 GPU using PyTorch 2.0 and HuggingFace Transformers v4.30.

6.3 Results and Discussion

We report performance on both classification metrics and explanation metrics. Table 1 shows the comparison of our model against state-of-the-art baselines on the HateXplain test set. The baseline models include XGBoost+SHAP for gradient-based token-level explanations, CNN-GRU for capturing local and sequential features, BiRNN–HateXplain and BERT–HateXplain which use supervised attention on the HateXplain dataset, XG-HSI-BERT/BiRNN that incorporate semantically important embeddings for improved interpretability, and HARE, which leverages LLM-extracted rationales with attention mechanisms to enhance explanation plausibility and faithfulness.

Our model significantly outperforms existing baselines in both predictive accuracy and explainability. The token-level F1 score improvement of over 7% indicates stronger alignment with human-annotated rationales. Similarly, the comprehensiveness score demonstrates that removing highlighted tokens from input text greatly affects model confidence, indicating faithful rationale extraction. The multi-channel attention mechanism, when trained with supervision, helps the model focus on diverse contextual patterns, while the LLM encoder captures rich semantic structure in the input.

Our model surpasses all baseline models on the HateXplain benchmark, achieving an accuracy of 87.3%, AUROC of 0.881, token-level F1 of 0.553, IOU-F1 of 0.261, AUPRC of 0.874, and a comprehensiveness score of 0.524, highlighting its effectiveness in both accurate classification and interpretable rationale generation. We also visualized attention heatmaps and found that LLM-BiMACNet consistently highlights semantically relevant tokens such as slurs, targeted identities, and abusive verbs, which aligns well with human reasoning.

6.4 Interpretability Evaluation

To assess the faithfulness and conciseness of model explanations, we evaluate LLM-BiMACNet using post-hoc interpretability frameworks—**SHAP** and **LIME**—as well as intrinsic explanation metrics such as **fidelity** and **sparsity**. These help validate that the rationale alignment is not only plausible but also logically consistent with model behavior.

Tables 1 and 2 present the classification and explanation performance of LLM-BiMACNet compared to existing models on the HateXplain benchmark. LLM-BiMACNet achieves the highest accuracy, AUROC, and token-level F1, while also outperforming baselines in SHAP (0.603) and LIME (0.581) alignment, indicating strong agreement with post-hoc explanation tools. It also shows the highest fidelity (0.752), demonstrating that its explanations reflect essential decision-driving tokens, and the lowest sparsity (0.366), ensuring concise and interpretable rationale outputs suitable for real-world use. Table 3 shows that each component cannot match the full LLM-BiMACNet. LLM-BiMACNet, while effective, has a few limitations. Its performance drops under domain shift, particularly on non-social media platforms like forums or blogs with different linguistic structures. The model's reliance on human-annotated rationales means that inconsistent or sparse annotations can reduce effectiveness. Moreover, the computational overhead of multi-channel attention is over.

To evaluate the robustness of our proposed LLM-BiMACNet, we conducted domain generalization experiments by training on HateXplain (Mathew et al., 2021) and testing in a zero-shot setting on

Table 1: Performance comparison of LLM-BiMACNet with baseline models on the HateXplain test set.

S.No	Model	Accuracy	AUROC	Token-F1	Comprehensiveness
1	XGBoost + SHAP (Babaeianjelodar et al., 2022)	79.0%	–	0.420	–
2	CNN-GRU (Böck et al., 2024)	62.8%	–	–	–
3	BiRNN-HateXplain (Mathew et al., 2021)	61.2%	–	0.330	0.200
4	BERT-HateXplain (Mathew et al., 2021)	69.8%	–	0.400	0.250
5	XG-HSI-BiRNN (Böck et al., 2024; Wasi, 2024)	74.2%	–	0.487	–
6	XG-HSI-BERT (Wasi, 2024)	79.1%	–	0.497	–
7	HARE (Yang et al., 2023)	84.5%	0.860	0.510	0.240
8	**LLM-BiMACNet**	**87.3%**	**0.881**	**0.553**	**0.261**

Table 2: Evaluation of model explanation quality.

S.No	Model	SHAP Score	LIME Score	Fidelity	Sparsity
1	BERT-HateXplain (Mathew et al., 2021)	0.562	0.537	0.671	0.431
2	BiRNN-HateXplain (Mathew et al., 2021)	0.543	0.501	0.649	0.460
3	HARE (Yang et al., 2023)	0.580	0.554	0.710	0.395
4	**LLM-BiMACNet**	**0.603**	**0.581**	**0.752**	**0.366**

Table 3: Ablation results of proposed model.

Model Variant	F1-Score	Rationale Alignment (%)
LLM-BiMACNet	92.4	87.6
BiGRU	87.8	85.9
Multi-Head Attention	89.5	84.2
Rationale Supervision	88.3	75.1

Table 4: Domain generalization results of LLM-BiMACNet.

Dataset / Setting	Accuracy	Precision	Recall	F1
HateXplain (Mathew et al., 2021) (In-domain)	0.84	0.83	0.83	0.83
Stormfront (Bala Das et al., 2023) (Zero-shot)	0.80	0.79	0.77	0.78
Davidson Twitter (Davidson et al., 2017) (Zero-shot)	0.82	0.81	0.80	0.80
Cross-domain Avg. w/o Emotion (Bala Das et al., 2023; Davidson et al., 2017)	0.81	0.80	0.79	0.79
Cross-domain Avg. w/ Emotion Task	0.87	0.86	0.86	0.86

Stormfront (Bala Das et al., 2023) and Davidson Twitter (Davidson et al., 2017) (Table 4).

7 Conclusion and Future Work

This paper presents LLM-BiMACNet, a large language model-based bidirectional multi-channel attention classification network, designed to detect hate speech while simultaneously identifying the contextual words that influence model predictions. By incorporating supervised rationale alignment and multi-head attention over contextual embeddings, the model effectively highlights semantically significant tokens, offering faithful and concise explanations. Experimental results on the HateXplain dataset demonstrate that our model outperforms existing state-of-the-art approaches in both classification accuracy and interpretability metrics, including token-level F1, SHAP/LIME agreement, fidelity, and sparsity. The model not only provides accurate hate speech categorization but also reveals interpretable evidence supporting each decision, making it suitable for sensitive applications such as content moderation, auditing, and sociolinguistic research. Future work includes extending the model for multilingual hate speech with cross-lingual rationale supervision, optimizing it for low-resource deployment, adapting it to out-of-domain texts, and improving explanation quality using prompt-based LLMs or counterfactual reasoning.

References

Marzieh Babaeianjelodar et al. 2022. Explainable and high-performance hate and offensive speech detection. *Neurocomputing*, 512:226–235.

Sudhansu Bala Das, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kr. Patra. 2023. Improving multilingual neural machine translation system

for indic languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–24.

Adrian Böck, Djordje Slijepčević, and Matthias Zeppelzauer. 2024. Exploring the plausibility of hate and counter speech detectors with explainable ai. *arXiv preprint arXiv:2407.20274*.

Sudhansu Bala Das, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kumar Patra. 2022. Nit rourkela machine translation (mt) system submission to wat 2022 for multiindicmt: An indic language multilingual shared task. *Proceedings of the 9th Workshop on Asian Translation*.

Sudhansu Bala Das, S Choudhury, Tapas K Mishra, and Bidyut Kr Patra. 2025a. Investigating the effect of backtranslation for indic languages. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 152–165.

Sudhansu Bala Das, Samujjal Choudhury, Tapas Kumar Mishra, and Bidyut Kr Patra. 2025b. Comparative analysis of subword tokenization approaches for indian languages. *arXiv preprint arXiv:2505.16868*.

Sudhansu Bala Das, Leo Raphael Rodrigues, Tapas Kumar Mishra, and Bidyut Kr Patra. 2024. An approach for mistranslation removal from popular dataset for indic mt task. *arXiv preprint arXiv:2401.06398*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 512–515.

Kavya Garg, Mayank Singh, Prithwish Bandyopadhyay, and Tanmoy Chakraborty. 2023. Hate speech detection is easy! or is it? breaking the love filter. *arXiv preprint arXiv:2306.11613*.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Prashant Kapil and Asif Ekbal. 2024. A survey on combating hate speech through detection and prevention in english. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*.

Hyuna Kim, S Arora, Yian Wu, Cristian Danescu-Niculescu-Mizil, and Diyi Yang. 2022. Why is it hate speech? masked rationale prediction for explainable detection. In *Proceedings of the 29th ICCL*, pages 6628–6638.

Dheeraj Kodati. 2020. Analysing covid-19 news impact on social media aggregation. *International Journal of Advanced Trends in Computer Science and Engineering*.

Dheeraj Kodati and Chandra Mohan Dasari. 2024. Negative emotion detection on social media during the peak time of covid-19 through deep learning with an auto-regressive transformer. *Engineering Applications of Artificial Intelligence*, 127:107361.

Dheeraj Kodati and Chandra Mohan Dasari. 2025a. Detecting contextual words for emotion mining from suicide related texts using hierarchical explainable large language models. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5383356. SSRN preprint.

Dheeraj Kodati and Chandra Mohan Dasari. 2025b. Detecting critical diseases associated with higher mortality in electronic health records using a hybrid attention-based transformer. *Engineering Applications of Artificial Intelligence*, 139:109649.

Dheeraj Kodati and Tene Ramakrishnudu. 2021. Negative emotions detection on online mental-health related patients texts using the deep learning with mha-bcnn model. *Expert Systems with Applications*, 182:115265.

Dheeraj Kodati and Tene Ramakrishnudu. 2023. Identifying suicidal emotions on social media through transformer-based deep learning. *Applied Intelligence*, 53:11885–11917.

Dheeraj Kodati and Ramakrishnudu Tene. 2024a. Advancing mental health detection in texts via multi-task learning with soft-parameter sharing transformers. *Neural Computing and Applications*, 37:3077–3110.

Dheeraj Kodati and Ramakrishnudu Tene. 2024b. Emotion mining for early suicidal threat detection on both social media and suicide notes using context dynamic masking-based transformer with deep learning. *Multimedia Tools and Applications*, 84:11729–11752.

Binny Mathew, Punyajoy S, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

Goutham Nirmal, Shweta Jain, and Byron C Wallace. 2024. Towards interpretable hate speech detection using llm-extracted rationales. *arXiv preprint arXiv:2403.12403*.

Pratyay Vijayaraghavan and Soroush Vosoughi. 2021. Interpretable multi-modal hate speech detection. *arXiv preprint arXiv:2103.01616*.

Azmine Toushik Wasi. 2024. Explainable identification of hate speech towards islam using graph neural networks. In *Proceedings of the NLP4PI Workshop (NeurIPS 2024)*. ArXiv preprint arXiv:2311.04916.

Jiachang Yang, Xinyi Chen, Shubham Srivastava, Steve Chien, and Kai-Wei Chang. 2023. Hare: Explainable hate speech detection with step-by-step reasoning. *arXiv preprint arXiv:2311.00321*.