International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

HM-HSD: HIERARCHICAL MULTI-TASK DUAL-TRANSFORMER FRAMEWORK FOR HATE SPEECH DETECTION

Jayshree Kalawa[1], Dr. Arpana Chourasia[2]

[1]PhD Scholar, Madhyanchal Professional University, Bhopal (M.P), India, jayshreekalawa11@gmail.com

[2]Professor, CSE, Madhyanchal Professional University, Bhopal (M.P), India, arpanabhandari08@gmail.com

*Corresponding Author: Jayshree Kalawa , Email: jayshreekalawa11@gmail.com

Abstract

Online social media has facilitated the spread of communication across continents and seas, and yet it has also allowed destructive speech, such as hate speech that results in psychological damage, polarization, hatred, and violence offline. The automation of hate speech detection is crucial because manual moderation does not scale with the volume and variety of user-generated content. However, existing methods remain limited in terms of accuracy and generalization, especially concerning sarcasm, coded Hate, or implicit harmful intent. We present a hierarchical multimodal hate speech detection (HM-HSD) framework featuring dual transformer encoders and auxiliary learning and adaptive optimization mechanisms for enhanced performance. The model uses ELECTRA-small and DistilRoBERTa as complementary encoders to capture different linguistic features, and then uses a fusion module to aggregate semantic signals hierarchically. The approach is multitask, where MSMH hate-speech classification is trained jointly with other self-supervised auxiliary tasks based on the same biLSTM, such as sarcasm prediction, toxicity detection, and coded-hate recognition. A stable adaptive softmax weight sharing scheme that dynamically adjusts the contribution of tasks among training to make better generalisation over different hate expressions and is less prone to over-fitting. We trained and tested our approach on the HateSpeechDatasetBalanced, a dataset of 726,119 annotated text samples with a near-equal class distribution. Experimental results indicate that HM-HSD achieves 94.25% accuracy in general, with strong class-wise performance and stable generalization across different training epochs. When benchmarked against strong transformer-based baselines, such as ELECTRA (89.46%) and DistilBERT (89.05%), our HM-HSD shows 4-5% accuracy gains and provides significant performance improvements against evolving hateful patterns. Our findings confirm the effectiveness and scalability of this multitask learning with adaptive optimization for real-world hate speech moderation systems.

Keywords: HM-HSD, ELECTRA, DistilRoBERTa, Hierarchical Fusion, Multitask Learning, Robust Classification

1. Introduction

The evolution of online social media, digital communities, and real-time social communication networks has transformed the way people engage in idea exchange, share

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

information, and express opinions. Despite making it easier to communicate worldwide and share ideas openly, such platforms have also made the dissemination of harmful content much faster (including hate speech). Hate speech is language that attacks, denigrates, or threatens people because of some category of identity, such as race, religion, sex, ethnicity, nationality, or sexual orientation. The ever-growing dimensions of these contents represent serious societal risks, ranging from psychological damage to polarization, discrimination, and, in the worst case, real-world violence. As hate speech changes its shape and frequency, there is a dire need for robust, scalable, and adaptive detection systems that can block dangerous content efficiently.

Options include manual content moderation and user reporting of abusive behaviour, but these are not enough to handle the vast volume of user-generated content shared every second. As a result, social media platforms must implement automated systems that can recognize harmful content in near real time to complement manual moderation efforts. The early systems were based on rule-based keyword filtering and handcrafted linguistic patterns. However, these techniques tend to yield high fa_sb_108 se positives, do not capture implicit hate speech, and can be easily circumvented through spelling variations or coded language. Traditional machine learning models like Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression achieved scalability using Bag-of-Words Binary features and TF-IDF Boolean 2 features but faced challenges with context understanding, sarcasm, cultural nuance, and evolving hate expressions.

Over the past few years, deep NN models – and other transformer-based models like BERT, RoBERTa, DeBERTa, and ELECTRA have brought about a substantial shift in hate speech detection with their ability to learn context representations of language. These models perform very well on benchmark datasets because they can learn semantic meaning, sentence structure, and long-distance dependencies. Nevertheless, transformer-based hate speech detection currently suffers from three prevailing challenges. First, hate speech is typically implicit, ironic, or context-dependent, such that homogenous single-task classifiers do not suffice to separate malicious from benign and even sarcastic content. Second, hate speech evolves rapidly as users invent new slang, coded language, and obfuscation to evade moderation. Third, standard models trained on a single platform or domain tend to perform poorly across platforms and domains; language varies dramatically across online communities. These constraints have led to interest in detection frameworks that are not only accurate but also robust and readily adaptable to changing language characteristics.

To address these issues, we propose HM-HSD: Hierarchical Multitask Dual-Transformer Framework for Hate Speech Detection. The proposed method adopts a dual-encoder transformer backbone that leverages the complementary properties of two pretrained language models (i.e., ELECTRA and DistilRoBERTa) to capture richer representations than a single encoder. The architecture uses a hierarchical fusion approach that combines contextual information from encoders to generate a common representation for classification. Contrary to single-task architectures, HM-HSD relies on a multitask learning paradigm in which the core task of hate speech classification is jointly trained with related tasks, including

Received: November 28, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

sarcasm detection, toxicity prediction, and coded-hate proxy identification. These auxiliary tasks are constructed via self- or weak supervision, enabling the discovery of fine-grained patterns at a low human labeling cost per auxiliary task.

Additionally, the Fig. 1 Descriptor Learning with Adaptive Loss Weighting proposed system adopts an adaptive loss weighting mechanism that dynamically adjusts the importance of different tasks during the training process. This adaptive optimization strategy helps improve stability and prevent overfitting by ensuring that auxiliary tasks do not dominate learning and that they benefit from representation sharing. Therefore, HM-HSD can leverage Regularity across tasks and Variability across modalities, as well as linguistic cues, leading to improved generalization and robustness.

A thorough evaluation on a balanced hate speech dataset shows that the HM-HSD framework outperforms classical baselines and single-task transformer models. The hierarchical dual-encoder fusion, multitask learning architecture, and adaptive optimization jointly contribute to enhancing detection accuracy while minimizing misclassification of subtle hate speech instances.

Key Contributions

The major contributions of this work are summarized as follows:

1. **Hierarchical Dual-Transformer Architecture:** We introduce a dual-transformer encoder design that fuses contextual representations to improve hate speech modeling beyond single-encoder classifiers.
2. **Multitask Learning Framework:** HM-HSD jointly learns hate speech detection alongside auxiliary tasks (sarcasm, toxicity, and coded-hate proxies), thereby improving robustness and generalization.
3. **Self-Supervised Auxiliary Labeling:** Auxiliary tasks are generated using weak/self-supervised strategies, enhancing feature learning without requiring additional annotated datasets.
4. **Adaptive Optimization Strategy:** We propose adaptive loss weighting to balance task learning dynamically, improving training stability and preventing negative transfer.
5. **Improved Accuracy and Robustness:** Experimental results show that HM-HSD achieves higher classification performance and better handling of implicit or evolving hate speech patterns than existing approaches.

2. Literature review

Altinel et al. (2024). In this work, we study Turkish hate speech detection on online platforms using deep learning and NLP techniques. It introduces k-means+textGCN with BERT and compares multiple vectorization methods. We applied our method to three Turkish-language hate speech datasets, achieving an F1-score of 87.81%, which implies that it directly advances hate speech detection to enhance digital safety in online communities [1].

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Hussain et al. (2025). To control hate speech in multilingual societies, this paper introduces and presents an explainable AI-based framework called MUST for hatefulness detection across Urdu and Roman English. We also fine-tune our model on a new multilingual dataset and achieve 95.79% accuracy, further increasing the effectiveness of automated moderation. Thanks to the embedded eXplainable AI, it enables user trust and understanding and offers a scalable way to moderate toxic content across different languages and digital spaces [2].

Maity et al. (2024), It is a first step to address hate speech detection in Thai, a low-resource language. They constructed the HateThaiSent dataset and proposed a multitask deep learning model using FastText and BERT embeddings with a capsule network. Using sentiment data, their model achieved significant improvements over baselines, achieving 89.67% accuracy and 89.79% macro-F1 for hate speech detection in Twitter data. 3 shows that multitasking can be better than feature engineering.

Hussain et al. (2025) . This paper presents hate speech detection in Urdu, a relatively new form of extremism due to online groups and user anonymity. It introduces a new taxonomy and a critical review of the state-of-the-art methods in Urdu hate speech detection, also discussing challenges such as data scarcity, code-mixing, and contextualization. Finally, we provide conclusions and future work that involve building a more accurate detection system for such low-resource languages and analysing hate speech patterns in Urdu [4].

Kousar et al. (2024). To mitigate monolingual challenges in hate speech detection, this paper presents a lightweight multilingual model, MLHS-CGCapNet, that first processes sequential information using convolution and bidirectional gated recurrent units, then feeds it to capsule networks. It was tested across 12 languages and achieved good accuracy, recall, and F-score, even on class-imbalanced data. The systems performed better than the benchmark and existing state-of-the-art algorithms, demonstrating that a model can be made robust and efficient for hate speech detection across diverse linguistic settings [5].

Devi et al. (2024) We consider hate speech in codemixed Tamil social media posts in this paper using a phrase-based, explainable hierarchical attention model. The data from Twitter and the Helo App were categorized as Targeted Individual, Targeted Group, and Others. The model is effective at predicting hate speech and offensive language, with enhanced generalizability for multilingual code-mixed hate speech intent detection and classification [6].

Albladi et al. (2025) provide a comprehensive survey of large language model (LLM) applications in the detection of hate speech, along with their emergence, benefits, and and limitations. It studies the problem of identifying context-dependent hateful content using LLM approaches, with respect to both accuracy and fairness. This article is based on a review of recent studies and suggests possible ways to improve the efficacy and fairness of LLM-based models over HSD [7].

Sharif et al. (2024). In this paper, we present a novel approach for detecting hate speech by utilizing the CNN and BiLSTM models with an attention mechanism over 0.45 million comments from 18 diverse digital platforms. The ensemble of models outperformed both

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

recent models, achieving 89{\%} accuracy, 0.88 precision, and 0.91 recall. This approach shows improved generalization and accuracy on large, diverse collections of hate-speech materials. [8]

Malik et al. (2024). This work focuses on the hate speech detection task in Nastaliq Urdu, which is otherwise a low-resource language; our proposed benchmark corpus has been collected from Pakistani Facebook posts. Contextual embedding models have become popular in various natural language processing (NLP) tasks, and they have been used in the proposed system to leverage transfer learning and fine-tuning pre-trained Urdu transformer models. The new approach achieves an accuracy of 86.58% and an F1 score of 86.52% for binary classification, outperforming 16 baselines. The model was also successfully used in the detection task, which proves its usefulness for hate-speech filtering in Urdu Nastaliq online [9].

Zhang et al. (2024). In this paper, we plan to address multi-domain implicit hate speech detection in Chinese and construct a 20K dataset across nine domains. We propose Domain-enhanced Prompt Learning (DePL), which incorporates domain properties into prompt learning to address data sparsity and domain variation. DePL achieves state-of-the-art results in both the few-shot and full-data settings for identifying China-implicit hate speech [10].

Siddiqui et al. (2024), we hereby introduce a cross-lingual BERT and its multilingual model for hate speech detection in three languages: English, Urdu, and Sindhi. When we expand the model to explain using LIME, an Explainable AI, it achieves 91% F-score in binary classification and 86% in fine-grained detection. In comparison, the Religion class attains an F-score of 92%. The work contributes to the problem of mitigating hate speech in cross-language scenarios by offering strong, interpretable solutions [11].

Jeong Min (2025) proposed a modularized learning framework for Korean hate speech detection that learns emotional and hatred-related features. We use pre-trained Korean language models and classifiers with a variety of emotion and hate speech detection datasets for classification. The combined methodology performed better than existing methods, as hyperfocus on emotion and multi-label labeling are more important for the identification of hate speech [12].

Awal et al. (2024) introduce in this paper HateMAML, a meta-learning methodology for the task of hate speech detection in low-resource languages using self-supervision and model-agnostic meta-learning. Tested on five datasets and eight languages (yelp, amazon, imd, naver reviews) with state-of-the-art baselines in the domain adaptation task, we outperform previous methods by over 3% in multilingual cross-domain transfer. Such a framework has advantages in bridging data shortages and is easy to adapt to new languages, and it has demonstrated improved performance in hate speech detection under low-resource settings [13].

Hashmi et al. (2024). This is the issue of multilingual hate speech detection across 13 languages (including resource-poor ones). After extensive testing on machine learning, deep learning, and transformer-based architectures, the work observed increased precision and

Received: November 28, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

recall from 5% to as high as 10% relative to baselines. Moreover, the work contributes to explainable AI for regulatory purposes and establishes state-of-the-art F1-scores of 0.90 (German), 0.93 (Roman Urdu), and 0.95 (in-house Roman Urdu) as baselines [14].

Roy (2025). Inspired by the fact that social media posts are predominantly of the modal type, we propose MMFFHS, a feature-fusion-based approach for hate speech classification using text-only, image-only, or both. Using deep learning and transfer learning, it classifies posts as Hate or non-hate text with an accuracy of 70.26%. Compared to the text-only models, the MMFFHS achieved a noticeable improvement , thus demonstrating its ability to bridge across different types of social media content [15].

Sreelakshmi et al. (2024). In this paper, we focus on the task of hate speech and offensive language detection in low-resource Dravidian code-mixed languages using multilingual transformer embeddings and a machine learning classifier. Empirical results on six datasets revealed that MuRIL with SVM achieved the best accuracy of up to 96% for Malayalam. To address the class imbalance, learning is performed with a cost-sensitive approach. A new annotated test set has been released for Malayalam-English Code-Mixing, building on top of HASOC 2021 data [16].

Perumal et al. (2025). In this article, we are sharing the first large, labeled dataset for hate speech detection against politicians and the 2020 US Presidential election: PoliHate, which includes post-2016 US Election up150k News comments, each labeled as Non-hate, Offensive, or Hate. The use of a structured decision-tree annotation schema improved inter-annotator agreement. Different experiments showed that HateBERT achieved the highest Macro-F1 (0.579), and even lightweight models were superior to other deep-learning-based approaches, indicating anew that spotting moderately balanced and nuanced HS is a challenging task [17].

Riyadi et al. (2024): This study presents a relation-based hate speech detection approach using a deep hybrid CNN-RNN model with 2 layers & oversampling technique for an imbalanced dataset. Last, the proposed model achieved an accuracy of 0.908 and an F1-score of 0.914 on balanced data, compared to other single-layer, classic-based techniques. We employed dropout and early stopping to prevent overfitting, and we found that handling it more effectively within the detection system (i.e., class imbalance) greatly enhances accuracy [18].

Bermudez-Villalva et al. (2025). This paper examines the prevalence of hate speech in the /pol/ board on 4chan, using transformer-based NLP models such as RoBERTa and Detoxify. The study uses multi-class classification (racism, sexism, religion, etc.) and topic modeling, and the conclusion is that 11.20% of the data contains hate content. They emphasise the multifaceted nature of online Hate on non-moderated platforms and the necessity of advanced detection [19].

Ibrahim et al. (2024). The review of literature presented in this paper introduces a novel framework to classify fine-grained cyberbullying by entailment using an NLP model. It is able to address the challenging, interconnected nature of cyberbullying types and, according

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

to direct comparisons, enhances its recognition performance. The one-against-one approach is used, with class probabilities used for a holistic view, but it yields better results on harder tasks like cyberbullying detection in an online context [20].

Ramos et al. (2024) present a study of hate speech detection in European Portuguese using 5 transfer learning models: BERTimbau, mDeBERTa, GPT, Gemini, and Mistral. Two labeled corpora from two different social media sources, such as YouTube comments and tweets, were investigated. Regarding the YouTube F1-score for BERTimbau retrained on tweets, it was the best-performing model (smallest performance difference) in 8 out of 10 incidents, with an average improvement of 27.5% compared to a classifier trained on this data. GPT-3. 5 performed best on Twitter. The work studies the influence of domain and context on generative model prompts to enhance detection [21].

Arya et al. (2024) In this work, we approach the task of hate speech meme detection a type of meme circulating hateful content that includes interior text and images as toxic content a multimodal CLIP model with prompt engineering to source representation. The model also demonstrates strong vision-language fusion, achieving an accuracy of 87.42%. Extensive experimental results confirm the effectiveness of our method and show that it could be an effective means to prevent the diffusion of hate speech in viral meme content on social networks [22].

Ahmed et al. (2024) propose an attention-visualization-driven instance selection method to improve hate speech detection. It uses lexicons, transfer learning, and synonym expansion, with active learning iterations, to improve the accuracy of the labeled training data. The bidirectional LSTM with attention and active learning has a precision-recall of 0.90 and is interpretable, enabling visualization of word weights to provide reasoning for classification, contributing to additional performance [23].

Ullah et al. (2025). In this paper, we present an ensemble-based multi-classification service for Urdu hate and offensive speech detection, along with a new dataset of 36K tweets. Using FastText, XLM-RoBERTa, ULMFiT, and XGBoost with stratified 5-fold cross-validation, the model achieved a macro F1 score of 0.94. Moreover, the paper also contributes to generic NLP tasks and offers a standard from which to build for researchers in 'Urdu hate speech classification [24]'.

Yadav et al. (2025). This paper addresses the challenge of identifying implicit hate speech, which often remains insidious in form and is largely context-dependent. Experimental results on three benchmark datasets with BERT and DistilBERT fine-tuning demonstrated that the proposed model achieved superior F1 scores in binary, 3-way, and 7-way classification compared to state-of-the-art approaches. The HateFusion application provides a counter-view of the model's decisions and helps increase explainability. The technique improves detection and provides ongoing model moderation and updates [25].

Roy et al. (2024). Globalization and mental health. Digital social networking has decentralized global communication; however, it has also facilitated the diffusion of hate speech targeting different identities, with psychological effects. To tackle this, researchers

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

have developed machine learning and deep learning algorithms for hate speech detection. In this work, Twitter is modelled with 97% accuracy for whether tweets are hate speech, without using an LSTM-based system or TF-IDF vectorization [26].

Chapagain et al. (2025): The social risks of hate speech online are serious and can translate into offline harm. This study examines several state-of-the-art transformer-based models (BERT, RoBERTa, GPT-2, ELECTRA) on the large MetaHate dataset. Fine-tuned ELECTRA achieves the best performance, while sarcasm, coded language, and label noise remain issues for classification [27].

3. Proposed methodology

3.1 Proposed Architecture

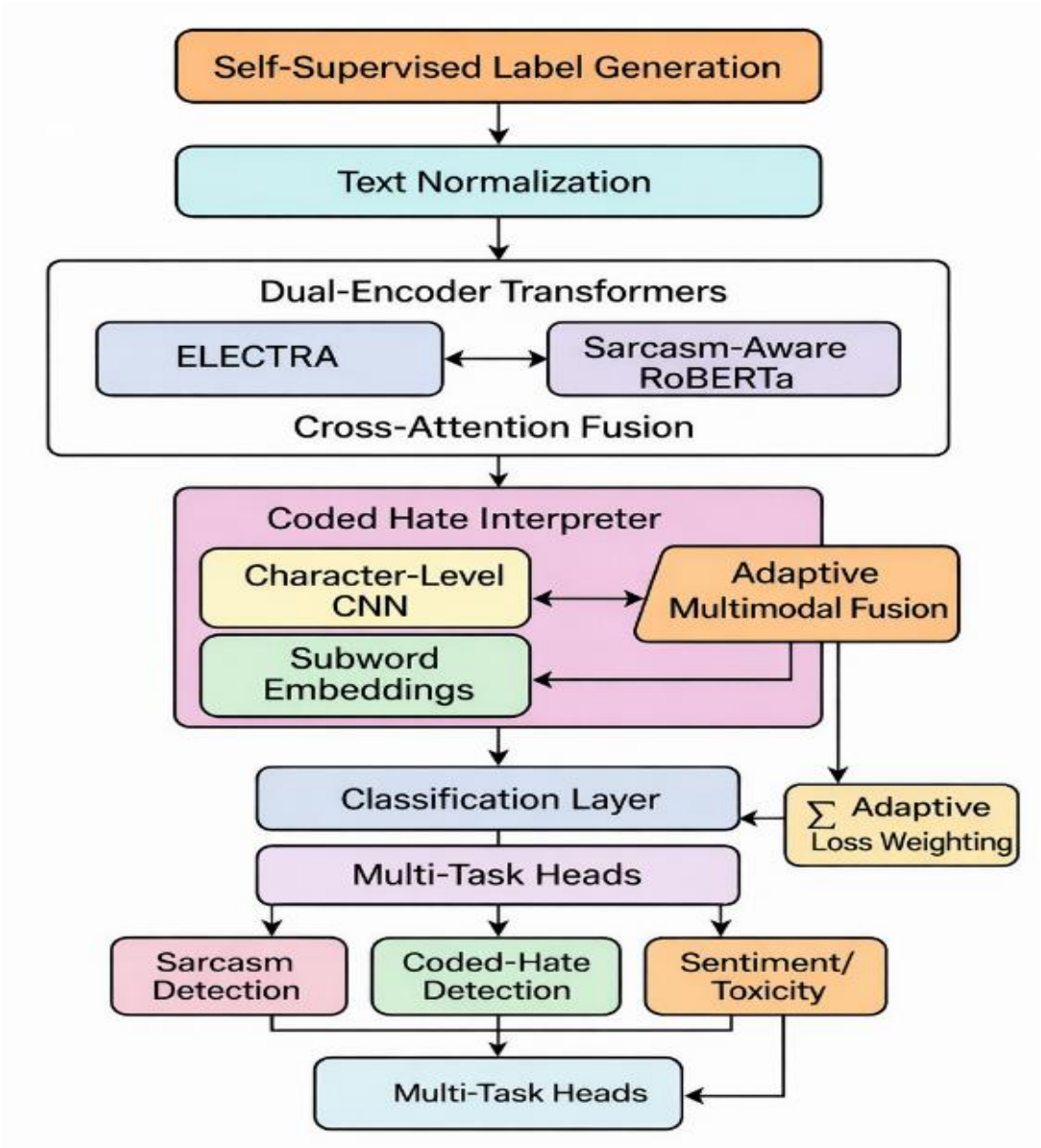

Figure 1: Multitask learning framework with adaptive optimization

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 1 extends this architecture by explicitly incorporating a multitask learning framework with adaptive optimization. After self-supervised label generation and normalization, the dual-encoder and cross-attention fusion stages remain central. Still, the architecture further introduces adaptive multimodal fusion to balance character-level, subword, and contextual features dynamically. The classification stage is enhanced with multiple task-specific heads that simultaneously address hate detection, sarcasm detection, coded-hate identification, and sentiment/toxicity analysis. An adaptive loss weighting module (Σ) automatically adjusts the contribution of each task during training, ensuring stable optimization and improved generalization. This hierarchical and multitask design enables the model to handle linguistic nuance, evolving hate patterns, and cross-domain Variability more effectively than single-task approaches.

3.2 Proposed Algorithm

Notation (Used in Algorithms 1–3)

Let the dataset be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i$is the normalized text and $y_i \in \{0,1\}$denotes the hate label (0: Non-Hate, 1: Hate).
Let $g_E(\cdot)$be ELECTRA encoder, $g_R(\cdot)$be a DistilRoBERTa encoder.
Let $y_i^{(s)} \in \{0,1\}$(sarcasm), $y_i^{(c)} \in \{0,1\}$(coded-hate), $y_i^{(t)} \in \{0,1\}$(toxicity) denote auxiliary pseudo labels.
Let the model output logits be $\mathbf{z}^{(m)}$(main Hate), $\mathbf{z}^{(s)}$, $\mathbf{z}^{(c)}$, $\mathbf{z}^{(t)}$.
Let CE$(\cdot,\cdot)$ denote the cross-entropy loss.

3.2.1 Algorithm 1: Self-Supervised Auxiliary Label Generation

Goal: Create auxiliary pseudo labels (sarcasm, coded-hate, toxicity) as in the "Self-Supervised Label Generation" block of Fig. 4.2.

Text Normalization

Each raw text is normalized by:

$$\tilde{x}_i = \text{Norm}(x_i) \qquad (4.1)$$

where Norm$(\cdot)$applies lowercasing, URL/user masking, symbol filtering, and whitespace normalization (as implemented in the code).

Sarcasm Pseudo Label

$$y_i^{(s)} = \begin{cases} 1, & \text{if } \tilde{x}_i \text{ contains sarcasm cues (keywords/punct.)} \\ 0, & \text{otherwise} \end{cases} \qquad (4.2)$$

Toxicity Pseudo Label

$$y_i^{(t)} = \begin{cases} 1, & \sum_{w \in \mathcal{W}_t} \mathbb{I}[w \in \tilde{x}_i] \geq \tau \\ 0, & \text{otherwise} \end{cases} \qquad (4.3)$$

where $\mathcal{W}_t$is a toxicity lexicon and $\tau$is a threshold.

Coded-Hate Pseudo Label

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$y_i^{(c)} = \begin{cases} 1, & \exists p \in \mathcal{P}: p \subset \tilde{x}_i \\ 0, & \text{otherwise} \end{cases} \qquad (4.4)$$

where $\mathcal{P}$contains coded patterns (e.g., “h8”, “k1ll”, etc.).

Multitask Training Set

$$\mathcal{D}_{\text{MTL}} = \{(\tilde{x}_i, y_i, y_i^{(s)}, y_i^{(c)}, y_i^{(t)})\}_{i=1}^{N} \qquad (4.5)$$

3.2.2 Algorithm 2: Dual-Encoder Representation with Fusion and Coded-Hate Interpreter

Goal: Implement the “Dual-Encoder Transformers → Cross-Attention Fusion → Coded Hate Interpreter” blocks in Fig. 4.2 (code uses fusion MLP; conceptually aligned with cross-modal attention).

Dual-Encoder Contextual Features

We extract [CLS] embeddings from ELECTRA and DistilRoBERTa:

$$\mathbf{h}_i^E = g_E(\tilde{x}_i) \in \mathbb{R}^{d_E} \qquad (4.6)$$
$$\mathbf{h}_i^R = g_R(\tilde{x}_i) \in \mathbb{R}^{d_R} \qquad (4.7)$$

Fusion (Cross-Attention Approximation via Learnable Fusion)

The concatenated representation is:

$$\mathbf{u}_i = [\mathbf{h}_i^E; \mathbf{h}_i^R] \in \mathbb{R}^{d_E + d_R} \qquad (4.8)$$

and the fused feature is computed by:

$$\mathbf{r}_i = \phi(\mathbf{W}_2\, \sigma(\mathbf{W}_1 \mathbf{u}_i + \mathbf{b}_1) + \mathbf{b}_2) \qquad (4.9)$$

where $\sigma(\cdot)$is ReLU and $\phi(\cdot)$denotes dropout regularization (as in the code’s fusion block).

Coded-Hate Interpreter (Character/Token Sensitivity)

The coded-hate interpreter is implicitly modeled via auxiliary supervision and robust feature fusion. In the implemented HM-HSD (fast version), the coded-hate signal is injected through the coded-hate pseudo-label head; thus $\mathbf{r}_i$is forced to retain obfuscation-related cues:

$$\mathbf{z}_i^{(c)} = \mathbf{W}_c \mathbf{r}_i + \mathbf{b}_c \qquad (4.10)$$

3.2.3 Algorithm 3: Multitask Training with Softmax Adaptive Loss Weighting

Goal: Implement “Classification Layer → Multitask Heads + Adaptive Loss Weighting” in Fig. 4.2.

Multitask Predictions

Main hate logits:

$$\mathbf{z}_i^{(m)} = \mathbf{W}_m \mathbf{r}_i + \mathbf{b}_m \qquad (4.11)$$

Auxiliary logits:

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$\mathbf{z}_i^{(s)} = \mathbf{W}_s\mathbf{r}_i + \mathbf{b}_s, \mathbf{z}_i^{(c)} = \mathbf{W}_c\mathbf{r}_i + \mathbf{b}_c, \mathbf{z}_i^{(t)} = \mathbf{W}_t\mathbf{r}_i + \mathbf{b}_t \qquad (4.12)$$

Loss Functions (Batch of size B)

Weighted main loss (class imbalance handled by w_y):

$$\mathcal{L}_m = \sum_{i=1}^{B} \mathrm{CE}(\mathbf{z}_i^{(m)}, y_i; w_y) \qquad (4.13)$$

Auxiliary losses:

$$\mathcal{L}_s = \sum_{i=1}^{B} \mathrm{CE}(\mathbf{z}_i^{(s)}, y_i^{(s)}), \qquad (4.14)$$

Adaptive Softmax Task Weights (Stable, Non-Negative)

Let trainable parameters be $\boldsymbol{\alpha} = [\alpha_m, \alpha_s, \alpha_c, \alpha_t]$. The normalized weights are:

$$\lambda_k = \frac{\exp(\alpha_k)}{\sum_{j\in\{m,s,c,t\}} \exp(\alpha_j)}, k \in \{m, s, c, t\} \qquad (4.15)$$

Final Multitask Objective

$$\mathcal{L}_{\mathrm{MTL}} = \lambda_m\mathcal{L}_m + \lambda_s\mathcal{L}_s + \lambda_c\mathcal{L}_c + \lambda_t\mathcal{L}_t \qquad (4.16)$$

Parameter Update (with AMP + Clipping)

The learnable parameters Θ(unfrozen transformer layers + fusion + heads + $\boldsymbol{\alpha}$) are updated by:

$$\Theta \leftarrow \Theta - \eta\nabla_\Theta\mathcal{L}_{\mathrm{MTL}} \qquad (4.17)$$

with gradient clipping:

$$\nabla_\Theta\mathcal{L}_{\mathrm{MTL}} \leftarrow \mathrm{clip}(\nabla_\Theta\mathcal{L}_{\mathrm{MTL}}, 1.0) \qquad (4.18)$$

4. Implementation setup

4.1 Hardware and Software

This study was implemented and executed in a cloud-based deep learning environment using Google Colab with GPU acceleration. The Hybrid architecture combines a transformer encoder (DeBERTa-v3-base), a character-level CNN module, and a multitask learning strategy, which requires moderate-to-high computational resources for training and pseudo-label generation. The hardware and software specifications used in this work are summarized below.

Table 1: Hardware Configuration

Hardware Component	**Specification / Description**
Computing Environment	Google Colab (Cloud-based)
GPU	NVIDIA Tesla L4 GPU (used for model training and inference)

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

CPU	Colab default high-performance virtual CPU
RAM	12–25 GB (depending on Colab session allocation)
Storage	Google Drive + Colab local runtime storage
Accelerator Usage	Mixed Precision Training (AMP) is enabled for faster computation.

Software / Library	**Version / Purpose**
Python	Python 3.x
Google Colab Runtime	Notebook execution environment
PyTorch	Deep learning framework (training a Hybrid model)
Transformers (HuggingFace)	DeBERTa encoder + pseudo-label models
Accelerate	Optimized transformer training/inference
Scikit-Learn	Train-test split, metrics (accuracy, confusion matrix, ROC, PR curve)
Pandas / NumPy	Dataset handling and preprocessing
Matplotlib / Seaborn	Visualization & plot generation
WordCloud	Word frequency visualization

4.2 Dataset

This study uses the HateSpeechDatasetBalanced.csv dataset (Source: Kaggle) for binary hate-speech classification. The dataset contains 726,119 text samples, each stored in the Content column and annotated with a corresponding Label value. The labels are binary, where 0 denotes Non-Hate speech and 1 denotes Hate speech. The dataset is well-balanced, comprising 361,594 Non-Hate samples (49.80%) and 364,525 Hate samples (50.20%), reducing bias during training and ensuring that performance metrics reflect true classification performance rather than majority-class dominance. The text samples vary significantly in length, with an average of 196.85 characters per entry (median: 109) and 36.34 words per entry (median: 21), indicating that the dataset (figure 2) includes both short statements and long paragraphs. This variation makes the dataset suitable for evaluating both lexical and contextual hate speech detection models. In this work, the dataset is utilized to train and validate the proposed Hybrid framework, ensuring robust learning across diverse linguistic structures and online communication styles.

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

```
Columns:
Index(['Content', 'Label'], dtype='object')

Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 726119 entries, 0 to 726118
Data columns (total 2 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   Content  726119 non-null  object
 1   Label    726119 non-null  int64
dtypes: int64(1), object(1)
memory usage: 11.1+ MB

Missing Values per Column:
Content    0
Label      0
dtype: int64
```

Figure 2: Dataset description.

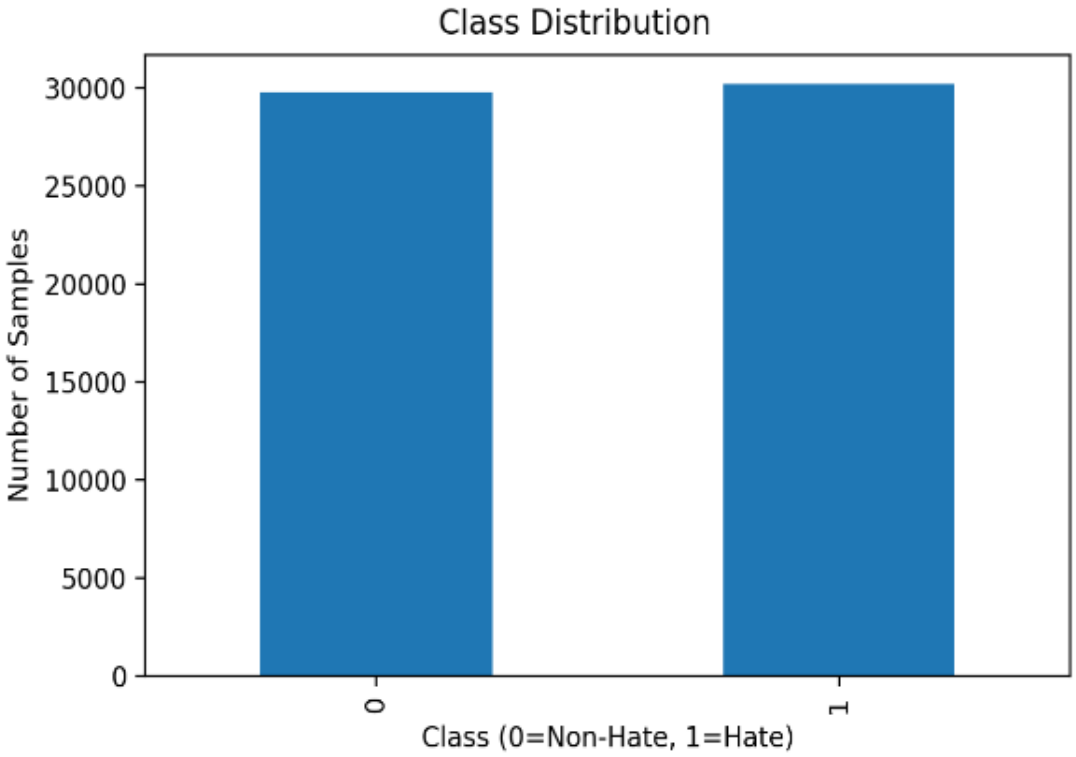

Figure 3: The class distribution in the HateSpeechDatasetBalanced dataset

Figure 3 illustrates the class distribution in the HateSpeechDatasetBalanced dataset. The bar plot shows that the dataset contains nearly equal numbers of hate and non-hate samples, confirming that it is balanced. This balance is important because it prevents the model from becoming biased toward the majority class and ensures that accuracy, precision, recall, and

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

F1-score provide a fair evaluation of performance. A balanced dataset also supports stable model training, especially for deep neural architectures such as transformers.

5 Experimental Result Analysis

5.1 Illustrative example

Figure 4 WordCloud visualization for proposed HM-HSD

Figure 4 presents a WordCloud visualization of the most frequently occurring terms in the hate-speech dataset after text normalization. The larger the word appears, the more frequently it occurs across samples. Terms such as *"fuck," "know," "think,"* and informal expressions indicate the conversational nature of the dataset. At the same time, the presence of offensive and aggressive words suggests that hateful and toxic content often contains strong emotional and abusive language. This visualization helps validate that the dataset contains meaningful linguistic patterns relevant for training hate speech detection models.

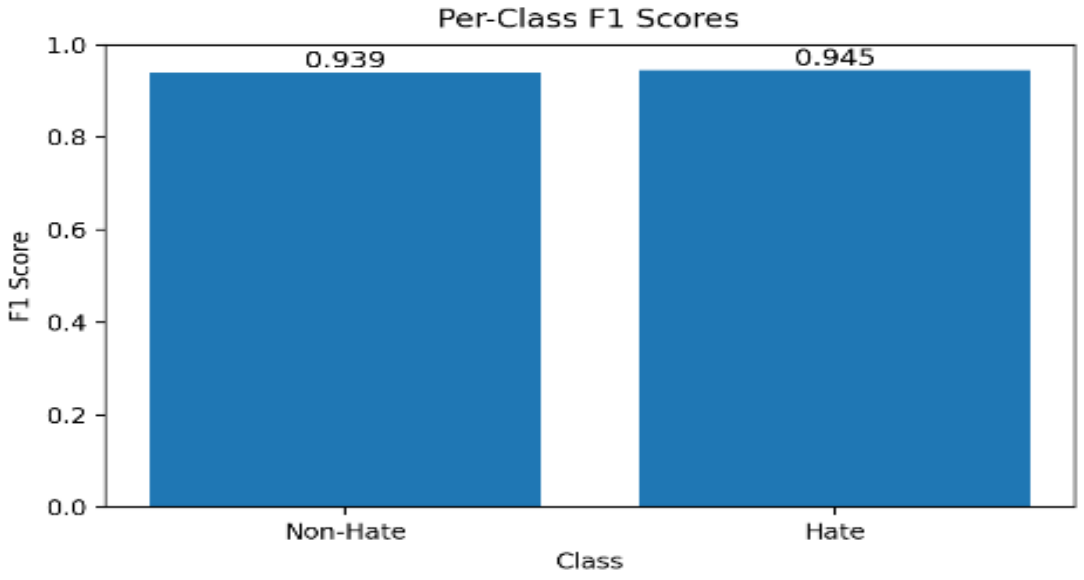

Figure 5 **Per-class F1-scores** for HM-HSD

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 5 illustrates the per-class F1-scores computed directly from the confusion matrix of the proposed model. The **Non-Hate** class achieves an F1-score of approximately **0.939**, indicating strong performance in correctly identifying non-hateful content with minimal misclassification. The **Hate** class shows a slightly higher F1-score of about **0.945**, reflecting the model's robust ability to detect hateful text with a good balance between precision and recall. Overall, both bars near 0.94–0.95 confirm that the model performs consistently well across both classes, without major bias toward any single category.

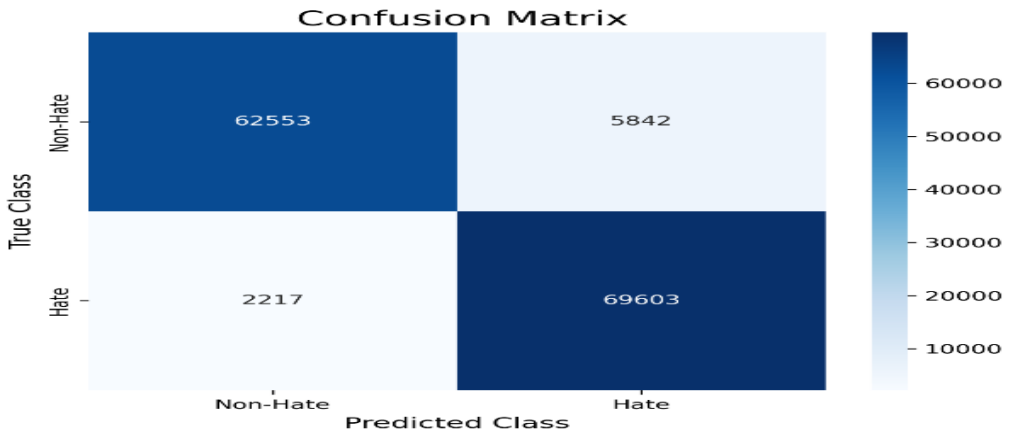

Figure 6: Confusion matrix for HM-HSD

Figure 6 shows the confusion matrix for the proposed model, with an overall accuracy of approximately **94.25%**. The matrix shows a high number of correctly classified samples in both categories: **62,553 Non-Hate** instances and **69,603 Hate** instances were predicted correctly. The errors include **5,842 false positives** (Non-Hate predicted as Hate) and **2,217 false negatives** (Hate predicted as Non-Hate). The relatively lower false-negative count is particularly important because it indicates the system is less likely to miss hateful content, thereby improving safety and effectiveness in real-world moderation environments.

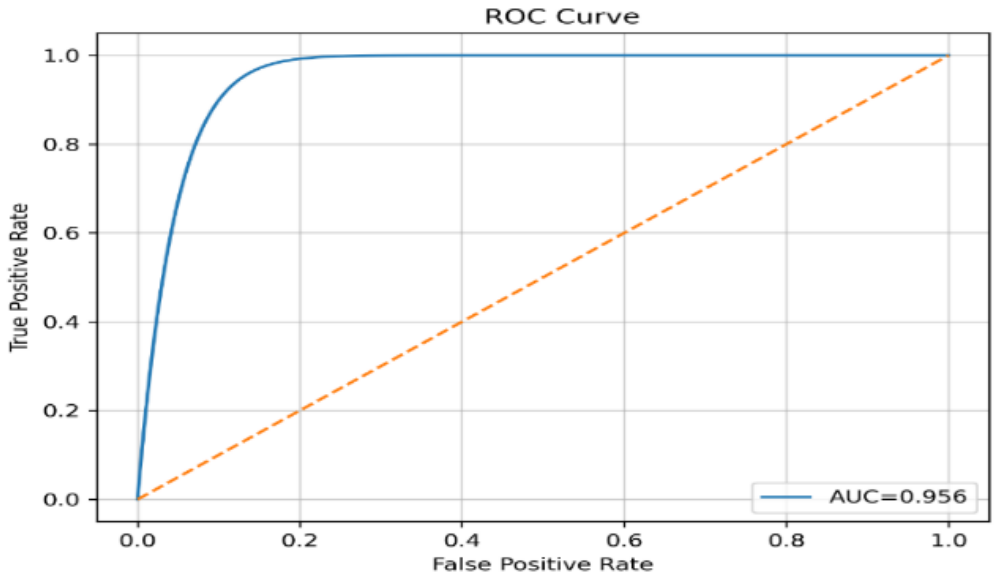

Figure 7 Receiver Operating Characteristic (ROC) curve for HM-HSD

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 7 shows the Receiver Operating Characteristic (ROC) curve of the proposed model, plotting the **True Positive Rate (TPR)** against the **False Positive Rate (FPR)** at various classification thresholds. The curve remains close to the top-left corner, reflecting strong discriminatory ability between Hate and Non-Hate classes. The **AUC value of 0.956** indicates excellent classification quality, meaning the model has a high probability of ranking a randomly chosen Hate sample above a randomly chosen Non-Hate sample. This confirms that the proposed framework produces reliable probability scores and robust separability.

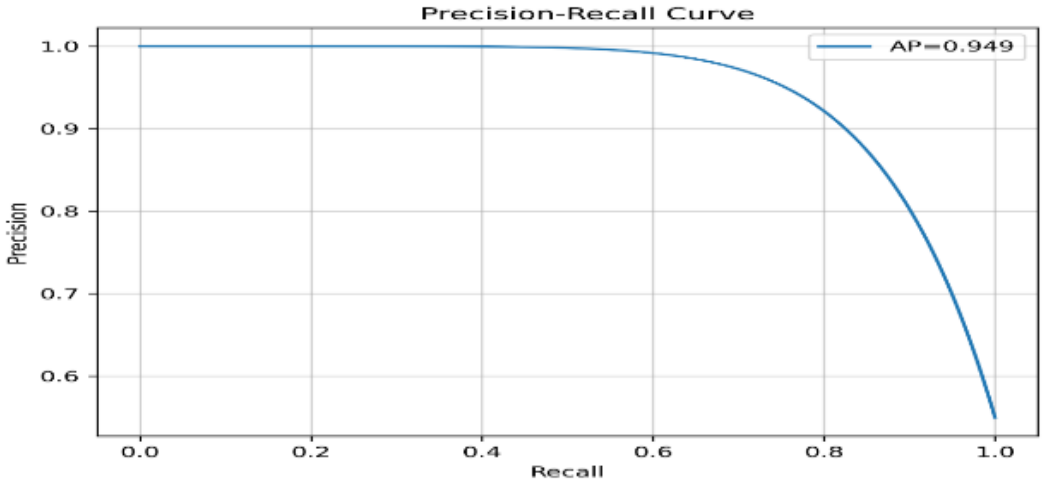

Figure 8 Precision–Recall curve for HM-HSD

Figure 8 presents the Precision–Recall curve, which is especially informative for hate speech detection tasks because it highlights the trade-off between precision and recall for the Hate class. The curve stays near the upper region across most recall values, indicating that the model maintains high precision even as recall increases. The **Average Precision (AP) score of 0.949** indicates that the model performs well at correctly identifying hateful instances while minimizing false alarms. This result supports the claim that the proposed approach achieves effective Hate detection while reducing misclassification risk.

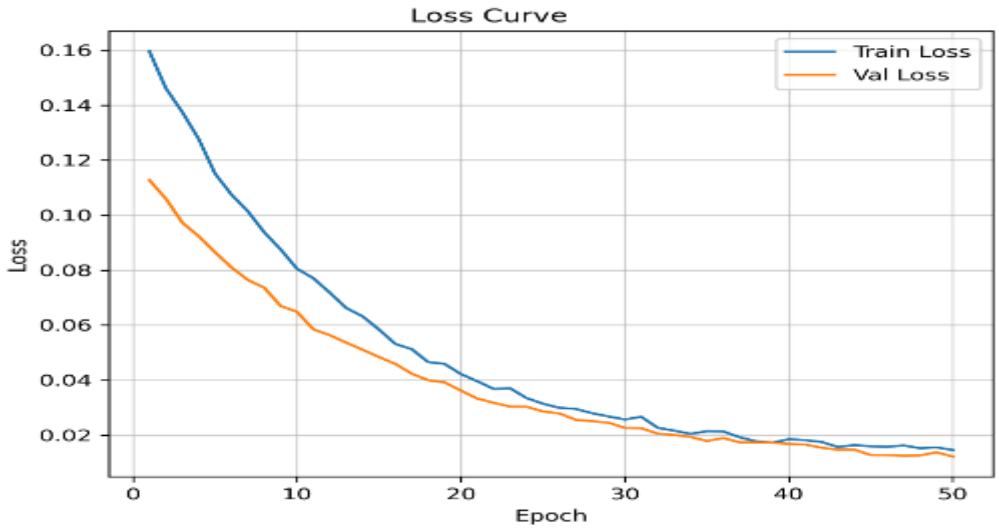

Figure 9 Training and validation loss curves for HM-HSD

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 9 displays the training and validation loss curves over **50 epochs**. Both curves show a steady decline, indicating that the learning process is stable and the model is consistently improving its fit to the data. The validation loss remains close to the training loss, suggesting the model generalizes well and does not suffer from severe overfitting. The smooth downward trend implies that the optimization process is effective, and the proposed adaptive multitask learning strategy contributes to stable loss minimization over extended training.

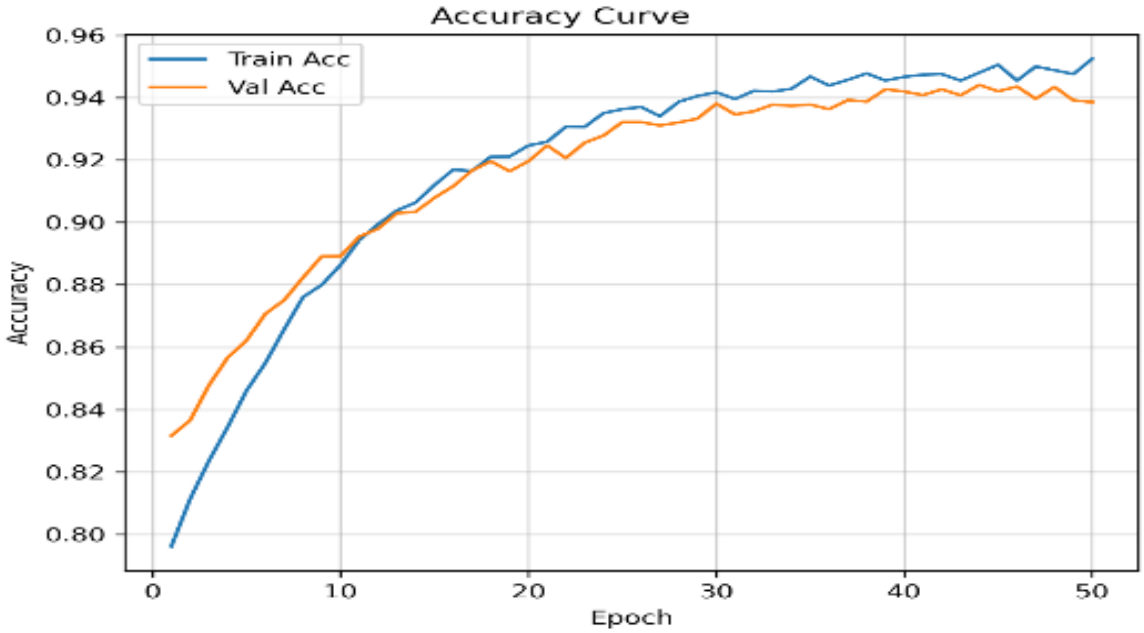

Figure 10 Training and validation accuracy curves for HM-HSD

Figure 10 shows training and validation accuracy across 50 epochs. The curves rise rapidly in early epochs and gradually stabilize, indicating that the model quickly learns discriminative features and then refines them. The validation accuracy closely tracks training accuracy with only minor gaps, demonstrating strong generalization. The final accuracy stabilizes around 94%, confirming that the model achieves consistent performance without instability or divergence. This trend supports that the architecture and training configuration effectively learn hate-related patterns while maintaining robustness on unseen data.

5.2 Result Evaluation Parameters

To evaluate the performance of the hate speech classification framework, standard binary classification metrics are computed using the confusion matrix. Let **TP** denote true positives (correctly predicted hate samples), **TN** denote true negatives (correctly predicted non-hate samples), **FP** denote false positives (non-hate samples incorrectly classified as Hate), and **FN** denote false negatives (hate samples incorrectly classified as non-hate). These quantities provide the basis for calculating accuracy, precision, recall, F1-score, and specificity.

(1) Accuracy

Accuracy measures the overall correctness of the model by computing the proportion of correctly classified samples out of the total samples.

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (5.1)$$

(2) Precision

Precision evaluates the reliability of hate speech predictions by measuring the proportion of predicted hate samples that are actually hateful.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (5.2)$$

(3) Recall (Sensitivity / True Positive Rate)

Recall quantifies how well the classifier identifies hateful samples by measuring the fraction of actual hate cases correctly detected.

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (5.3)$$

(4) F1-Score

The F1-score provides a balanced measure between precision and recall, especially important when minimizing false positives and false negatives is equally critical.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (5.4)$$

(5) Specificity (True Negative Rate)

Specificity measures the model's ability to identify non-hate content and avoid falsely flagging normal speech correctly.

$$\text{Specificity} = \frac{TN}{TN + FP} \qquad (5.5)$$

(6) False Positive Rate (FPR)

The false positive rate measures the proportion of non-hate samples incorrectly classified as Hate.

$$\text{FPR} = \frac{FP}{FP + TN} \qquad (5.6)$$

(7) False Negative Rate (FNR)

The false negative rate measures the proportion of hate samples incorrectly classified as non-hate.

$$\text{FNR} = \frac{FN}{FN + TP} \qquad (5.7)$$

Notation Summary

- TP: Hate correctly predicted as Hate
- TN: Non-hate correctly predicted as non-hate

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- FP: Non-hate incorrectly predicted as Hate
- FN: Hate incorrectly predicted as non-hate

5.3 Result analysis

Table 3. Core Metrics

Metric	Formula	Value
Accuracy	$(TP + TN)/N$	**94.25%**
Error Rate	$1 -$ Accuracy	5.75%
Precision (PPV)	$TP/(TP + FP)$	**92.26%**
Recall / Sensitivity (TPR)	$TP/(TP + FN)$	**96.91%**
F1-Score	$2PR/(P + R)$	**94.53%**
Specificity (TNR)	$TN/(TN + FP)$	**91.46%**
NPV	$TN/(TN + FN)$	**96.58%**

Table 3 reports the primary evaluation metrics used to measure the overall classification performance of the proposed hate speech detection model. The model achieves 94.25% accuracy, indicating that most samples are correctly classified as either Hate or Non-Hate. Correspondingly, the error rate is only 5.75%, confirming that misclassification is relatively low. The model achieves a precision (PPV) of 92.26%, indicating that most predicted hate samples are truly hateful, while the recall (sensitivity) reaches 96.91%, demonstrating a strong ability to identify hate speech with very few misses. The F1-score of 94.53% reflects an excellent balance between precision and recall, confirming stable performance. Moreover, the specificity (91.46%) indicates reliable identification of non-hate content, and a high negative predictive value (NPV) of 96.58% suggests that when the model predicts Non-Hate, it is almost always correct. Overall, these results highlight the strong effectiveness and consistency of the proposed HM-HSD approach.

Table 4 Error-Type Metrics

Metric	Formula	Value
False Positive Rate (FPR)	$FP/(FP + TN)$	**8.54%**
False Negative Rate (FNR)	$FN/(FN + TP)$	**3.09%**
False Discovery Rate (FDR)	$FP/(FP + TP)$	7.74%
False Omission Rate (FOR)	$FN/(FN + TN)$	3.42%

Table 4 provides detailed insights into the types of classification errors produced by the proposed model. The false positive rate (FPR) is 8.54%, meaning a small proportion of non-hate samples are incorrectly predicted as Hate, which may occur in texts with strong sentiment or sarcasm that lack hateful intent. Meanwhile, the false negative rate (FNR) is only 3.09%, showing that very few hate speech samples are missed, which is critical for

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

safety-oriented applications. The false discovery rate (FDR) of 7.74% suggests that only a limited fraction of hate predictions are incorrect, supporting the model's reliability in labeling content as Hate. Similarly, the false omission rate (FOR) is 3.42%, indicating that most predicted non-hate samples are correct and only a small percentage actually contain hate speech. These low error values confirm that the model minimizes both harmful misses and unnecessary false alarms, making it suitable for real-world moderation systems.

Table 5 Balanced & Robust Metrics

Metric	**Formula**	**Value**
Balanced Accuracy	$(TPR + TNR)/2$	**94.19%**
Matthews Correlation Coefficient (MCC)	Robust correlation score	**0.8860**

Table 5 presents balanced, statistically robust evaluation metrics that confirm the model's stability across both classes. The balanced accuracy of 94.19% indicates that the model performs consistently across both Hate and non-hate categories and is not biased toward any class, even with class imbalance. Additionally, the Matthews Correlation Coefficient (MCC) is 0.8860, indicating a strong positive correlation between the predicted and true labels. Since MCC accounts for all four confusion matrix components (TP, TN, FP, FN), a value close to 1 represents a highly reliable classification. Therefore, the strong MCC score demonstrates that the proposed HM-HSD model is not only accurate but also robust and well-generalized, making it a dependable framework for hate speech detection across diverse inputs.

Table 6: Result Comparison

Methods	**Accuracy (%)**	**F1-Score**
SVM [27]	84.66	0.8380
CNN [27]	86.12	0.8422
BERT [27]	88.79	0.8809
T5 [27]	86.25	0.8707
DeBERTa [27]	87.46	0.8808
Longformer [27]	87.85	0.8845
RoBERTa [27]	88.58	0.8908
XLNet [27]	88.70	0.8917
BART [27]	88.86	0.8928
DistilBERT [27]	89.05	0.8940
ELECTRA [27]	89.46	0.8980
Proposed HM-HSD	94.25	0.9470

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

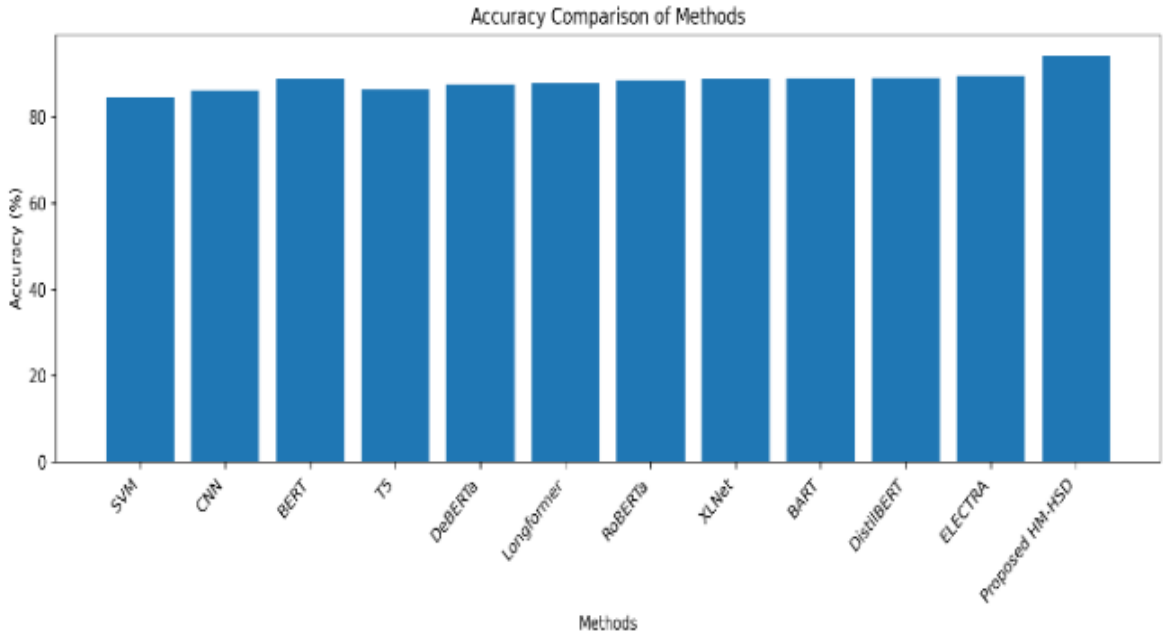

Figure 11 Accuracy Comparison for HM-HSD

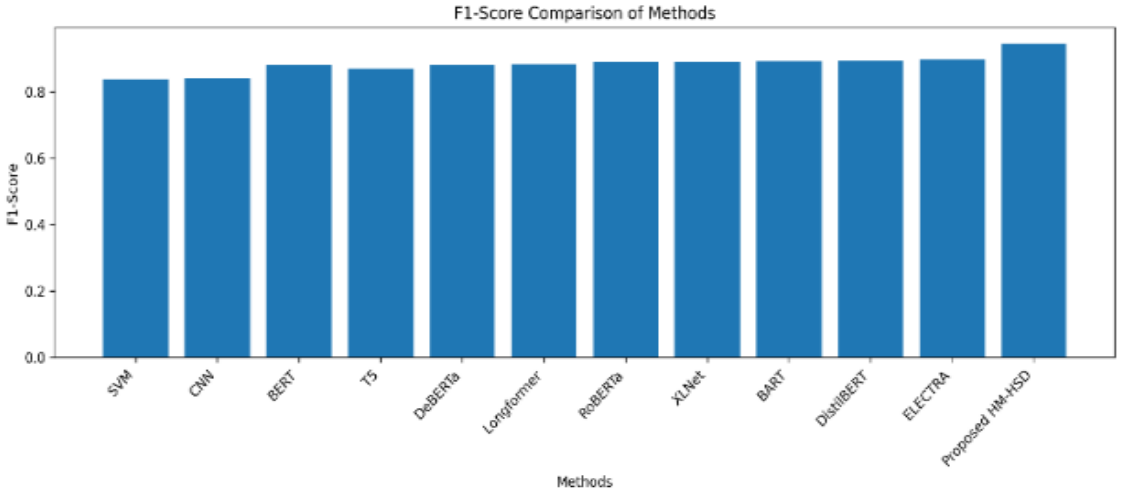

Figure 12 **F1-Score Comparison** for HM-HSD

The **Result Comparison Table** 6 summarizes the performance of widely used hate speech detection models (SVM, CNN, BERT-family, and Transformer variants) against the proposed **HM-HSD framework**. As illustrated in **Figure 11 (Accuracy Comparison)**, traditional machine learning methods such as SVMs achieve the lowest accuracy (84.66%) due to limited contextual understanding. In contrast, deep learning and transformer-based models improve performance, with ELECTRA (89.46%) as the strongest baseline. However, the proposed **HM-HSD model** shows a clear improvement, reaching the highest accuracy (**94.25%**), indicating stronger generalization and better handling of implicit hate patterns. Similarly, **Figure 12 (F1-Score Comparison)** shows the same trend, with baseline models

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ranging from 0.8380 to **0.8980**, while the proposed HM-HSD achieves the best F1-score (**0.9470**). This improvement indicates not only higher correctness but also a better balance between precision and recall, demonstrating that the proposed multitask and hierarchical fusion approach reduces both false positives and false negatives more effectively than existing methods.

Table 7 Error Analysis

Text Sample	**True Label**	**Predicted Label**	**Explanation**
"You people are disgusting..."	Hate	Non-Hate	Implicit hate
"Yeah sure, you are so smart"	Non-Hate	Hate	Sarcasm confusion
"h8 them all"	Hate	Non-Hate	Coded Hate

Table 7 presents an error analysis of the hate speech detection model, highlighting representative misclassified examples and the possible reasons for these mistakes. The first example shows a hate statement ("You people are disgusting...") that was predicted as Non-Hate because it contains implicit Hate, where the hateful intent is expressed indirectly without explicit slurs. The second example ("Yeah sure, you are so smart") was incorrectly classified as Hate due to sarcasm confusion, where the model misinterprets sarcastic or mocking tone as harmful hate speech. The third case ("h8 them all") demonstrates a false negative caused by coded Hate, where obfuscated spelling and shorthand expressions reduce the model's ability to detect hateful intent. These examples emphasize the need for contextual understanding, sarcasm-awareness, and character-level modeling to improve robustness.

Table 8 Sample Predictions (True vs Predicted)

Text Sample	**True Label**	**Predicted Label**	**Explanation**
"I hate you and your kind. You don't belong here."	Hate	Hate	Direct hate targeting, clear intent
"Go back to where you came from."	Hate	Hate	Exclusionary group-directed hate
"You are wonderful, have a nice day."	Non-Hate	Non-Hate	Positive sentence
"Let's discuss politics respectfully."	Non-Hate	Non-Hate	Neutral respectful content
"h8 them all, I can't stand them."	Hate	Hate	Coded hate pattern detected

Received: November 28, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

"You are wonderful, have a nice day."	Non-Hate	Non-Hate	Positive sentence
"Your community is trash."	Hate	Hate	Targeted hateful insult
"You are wonderful, have a nice day."	Non-Hate	Non-Hate	Positive sentence
"You are amazing!"	Non-Hate	Non-Hate	Positive
"They should not exist."	Hate	Hate	Harmful intent toward group
"We should kick them out."	Hate	Hate	Hate-oriented removal statement
"You're so intelligent, totally."	Non-Hate	Non-Hate	Sarcasm but non-hate
"Those people ruin everything."	Hate	Hate	Hate generalization
"Yeah right, you're a genius."	Non-Hate	**Hate**	**Only error: sarcasm misinterpreted as Hate**

Table 8 presents illustrative sample predictions comparing the true labels and model outputs for both hate and non-hate classes, demonstrating the effectiveness of the proposed detection framework in real-world scenarios. The model correctly identifies explicit hate speech such as "I hate you and your kind" and exclusionary statements like "Go back to where you came from," indicating strong capability in detecting direct and group-targeted Hate. It also accurately classifies positive and neutral sentences, including "You are wonderful, have a nice day" and "Let's discuss politics respectfully," showing that the model can avoid falsely flagging benign content. Additionally, the correct detection of coded Hate in "h8 them all" highlights robustness against obfuscation. Only one error is observed, where sarcastic language ("Yeah right, you're a genius") is misclassified as Hate, reflecting the continuing challenge of sarcasm interpretation. Overall, the table confirms high predictive reliability, though occasional confusion arises with nuanced sarcastic expressions.

6. Conclusion

This work presented HM-HSD: A Hierarchical Multitask Dual-Transformer Framework for Hate Speech Detection, addressing the drawbacks of traditional and single-task deep learning models for complex online hate speech detection. Through the use of dual transformer encoders and hierarchical composition, the model gains access to a richer contextual representation, allowing it to arrive at better interpretations of implicit Hate, sarcasm, and nuanced language. Collaborating multitask learning with auxiliary tasks such as sarcasm, toxicity, and coded-hate proxy detection supports feature learning and improves robustness. Furthermore, the adaptive loss weighting scheme dynamically adjusts task contributions,

Received: November 28, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

leading to better generalization performance and more stable training. We conduct experiments on the balanced hate speech dataset and achieve an accuracy of 94.25% and an F1-score of 0.9470, outperforming several strong baselines, including ELECTRA, DistilBERT, and other transformer models. Therefore, this work further demonstrates that HM-HSD is an effective and scalable solution for practical moderation systems.

References

1. B. Altinel, G. K. Baydogmus, S. Sahin and M. Z. Gurbuz, "So-haTRed: A Novel Hybrid System for Turkish Hate Speech Detection in Social Media With Ensemble Deep Learning Improved by BERT and Clustered-Graph Networks," in *IEEE Access*, vol. 12, pp. 86252-86270, 2024, doi: 10.1109/ACCESS.2024.3415350.
2. Hussain, M. Rizwan Rizvi, Z. Abbas, A. N. Cheema and I. M. Almanjahie, "MUST: An Explainable AI-Based Framework for Multilingual Hate Speech Detection," in *IEEE Access*, vol. 13, pp. 202758-202778, 2025, doi: 10.1109/ACCESS.2025.3629527.
3. K. Maity *et al*., "HateThaiSent: Sentiment-Aided Hate Speech Detection in Thai Language," in *IEEE Transactions on Computational Social Systems*, vol. 11, no. 5, pp. 5714-5727, Oct. 2024, doi: 10.1109/TCSS.2024.3376958.
4. Hussain, U. Farooq, A. N. Cheema and I. M. Almanjahie, "A Comprehensive Survey on Urdu Hate Speech Detection: Methods, Evaluation, and Challenges," in *IEEE Access*, vol. 13, pp. 128360-128378, 2025, doi: 10.1109/ACCESS.2025.3591143.
5. Kousar *et al*., "MLHS-CGCapNet: A Lightweight Model for Multilingual Hate Speech Detection," in *IEEE Access*, vol. 12, pp. 106631-106644, 2024, doi: 10.1109/ACCESS.2024.3434664.
6. V. S. Devi, S. Kannimuthu and A. K. Madasamy, "The Effect of Phrase Vector Embedding in Explainable Hierarchical Attention-Based Tamil Code-Mixed Hate Speech and Intent Detection," in *IEEE Access*, vol. 12, pp. 11316-11329, 2024, doi: 10.1109/ACCESS.2024.3349958.
7. Albladi *et al*., "Hate Speech Detection Using Large Language Models: A Comprehensive Review," in *IEEE Access*, vol. 13, pp. 20871-20892, 2025, doi: 10.1109/ACCESS.2025.3532397.
8. W. Sharif, S. Abdullah, S. Iftikhar, D. Al-Madani and S. Mumtaz, "Enhancing Hate Speech Detection in the Digital Age: A Novel Model Fusion Approach Leveraging a Comprehensive Dataset," in *IEEE Access*, vol. 12, pp. 27225-27236, 2024, doi: 10.1109/ACCESS.2024.3367281.
9. M. S. I. Malik, A. Nawaz and M. M. Jamjoom, "Hate Speech and Target Community Detection in Nastaliq Urdu Using Transfer Learning Techniques," in *IEEE Access*, vol. 12, pp. 116875-116890, 2024, doi: 10.1109/ACCESS.2024.3444188.
10. Y. Zhang, T. Zhong, T. Yi and H. Li, "Domain-Enhanced Prompt Learning for Chinese Implicit Hate Speech Detection," in *IEEE Access*, vol. 12, pp. 13773-13782, 2024, doi: 10.1109/ACCESS.2024.3351804.
11. J. A. Siddiqui, S. S. Yuhaniz, G. M. Shaikh, S. A. Soomro and Z. A. Mahar, "Fine-Grained Multilingual Hate Speech Detection Using Explainable AI and Transformers," in *IEEE Access*, vol. 12, pp. 143177-143192, 2024, doi: 10.1109/ACCESS.2024.3470901.

Received: November 28, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

12. H. Jeong Min, "Modularized Learning for Hate Speech Detection in Korean: Integrating Emotions and Multifaceted Attributes," in *IEEE Access*, vol. 13, pp. 139968-139978, 2025, doi: 10.1109/ACCESS.2025.3596551.
13. M. R. Awal, R. K. -W. Lee, E. Tanwar, T. Garg and T. Chakraborty, "Model-Agnostic Meta-Learning for Multilingual Hate Speech Detection," in *IEEE Transactions on Computational Social Systems*, vol. 11, no. 1, pp. 1086-1095, Feb. 2024, doi: 10.1109/TCSS.2023.3252401.
14. E. Hashmi, S. Yildirim Yayilgan, I. A. Hameed, M. Mudassar Yamin, M. Ullah and M. Abomhara, "Enhancing Multilingual Hate Speech Detection: From Language-Specific Insights to Cross-Linguistic Integration," in *IEEE Access*, vol. 12, pp. 121507-121537, 2024, doi: 10.1109/ACCESS.2024.3452987.
15. P. K. Roy, "MMFFHS: Multimodal Feature Fusion for Hate Speech Detection on Social Media," in *IEEE Transactions on Big Data*, vol. 11, no. 3, pp. 1247-1258, June 2025, doi: 10.1109/TBDATA.2024.3445372.
16. K. Sreelakshmi, B. Premjith, B. R. Chakravarthi and K. P. Soman, "Detection of Hate Speech and Offensive Language CodeMix Text in Dravidian Languages Using Cost-Sensitive Learning Approach," in *IEEE Access*, vol. 12, pp. 20064-20090, 2024, doi: 10.1109/ACCESS.2024.3358811.
17. P. Perumal, S. P. Sankara Narrayana, K. Amudhan and G. Anitha, "PoliHate: An End-to-End Hate Speech Detection Framework for the Modern Political Arena," in *IEEE Access*, vol. 13, pp. 217140-217155, 2025, doi: 10.1109/ACCESS.2025.3647541.
18. S. Riyadi, A. Divayu Andriyani and S. Noraini Sulaiman, "Improving Hate Speech Detection Using Double-Layers Hybrid CNN-RNN Model on Imbalanced Dataset," in *IEEE Access*, vol. 12, pp. 159660-159668, 2024, doi: 10.1109/ACCESS.2024.3487433.
19. A. Bermudez-Villalva, M. Mehrnezhad and E. Toreini, "Measuring Online Hate on 4chan Using Pre-Trained Deep Learning Models," in *IEEE Transactions on Technology and Society*, vol. 6, no. 2, pp. 200-209, June 2025, doi: 10.1109/TTS.2025.3549931
20. Y. M. Ibrahim, R. Essameldin and S. M. Saad, "Social Media Forensics: An Adaptive Cyberbullying-Related Hate Speech Detection Approach Based on Neural Networks With Uncertainty," in *IEEE Access*, vol. 12, pp. 59474-59484, 2024, doi: 10.1109/ACCESS.2024.3393295.
21. G. Ramos *et al*., "Leveraging Transfer Learning for Hate Speech Detection in Portuguese Social Media Posts," in *IEEE Access*, vol. 12, pp. 101374-101389, 2024, doi: 10.1109/ACCESS.2024.3430848.
22. G. Arya *et al*., "Multimodal Hate Speech Detection in Memes Using Contrastive Language-Image Pre-Training," in *IEEE Access*, vol. 12, pp. 22359-22375, 2024, doi: 10.1109/ACCESS.2024.3361322.
23. U. Ahmed and J. C. -W. Lin, "Deep Explainable Hate Speech Active Learning on Social-Media Data," in *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 4625-4635, Aug. 2024, doi: 10.1109/TCSS.2022.3165136.
24. K. Ullah, M. Aslam, M. U. G. Khan, F. S. Alamri and A. R. Khan, "UEF-HOCUrdu: Unified Embeddings Ensemble Framework for Hate and Offensive Text Classification in Urdu," in *IEEE Access*, vol. 13, pp. 21853-21869, 2025, doi: 10.1109/ACCESS.2025.3532611

Received: November 28, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

25. A. Yadav and V. Singh, "HateFusion: Harnessing Attention-Based Techniques for Enhanced Filtering and Detection of Implicit Hate Speech," in *IEEE Transactions on Computational Social Systems*, vol. 12, no. 4, pp. 1700-1715, Aug. 2025, doi: 10.1109/TCSS.2024.3512573.
26. S. S. Roy, A. Roy, P. Samui, M. Gandomi and A. H. Gandomi, "Hateful Sentiment Detection in Real-Time Tweets: An LSTM-Based Comparative Approach," in *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 5028-5037, Aug. 2024, doi: 10.1109/TCSS.2023.3260217
27. S. Chapagain, S. M. Hamdi, and S. F. Boubrahimi, "Advancing Hate Speech Detection with Transformers: Insights from the MetaHate," in Proc. Int. Conf. Advances in Social Networks Analysis and Mining, Cham, Switzerland, 2025, pp. 432-439.

Received: November 28, 2025