IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 4, Ser. 1 (July. – August. 2026), PP 01-12
www.iosrjournals.org

A Data-Centric Noise Reduction Framework For Robust Hate Speech Detection: A Comparative Study Of Embedding-Classifier Configurations Using Hatebert And Bidirectional LSTM Networks

Vibha Upadhya
School Of Computational Sciences, JSPM University, Pune – 412207, Maharashtra, India

Rahul R. Chakre
School Of Computational Sciences, JSPM University, Pune – 412207, Maharashtra, India

Abstract:
Hate speech on social media causes real harm to real people. Yet automatic detectors still struggle with two problems: messy training data, and models that look good on paper but fail on new data. This paper presents a simple, data-first approach to hate speech detection. We clean the text, balance the classes, and add more training examples before we train any model. We then feed this clean data into HateBERT, a version of BERT trained further on abusive language, followed by a 2-layer bidirectional LSTM (BiLSTM). The model is trained with a standard loss function, with no extra tricks. We test this setup against twelve other combinations of embeddings (MiniLM, RoBERTa, HateBERT) and classifiers (Logistic Regression, Random Forest, XGBoost, a feed-forward network, and BiLSTM with and without an attention layer), on two public datasets: the Davidson Twitter hate-speech dataset and the Jigsaw Toxic Comments dataset. We find that tree-based models such as XGBoost reach near-perfect accuracy on training data (100%) but drop by 9 to 13 points on test data — a clear sign that they are memorising the training set rather than learning real patterns. Out of all thirteen setups, our proposed HateBERT plus 2-layer BiLSTM model gives the best result on the Davidson dataset: 95.47% accuracy and an F1-score of 92.11%, beating every version that uses attention or a special loss function. On Jigsaw, it scores a close 92.37% on both accuracy and F1. These results show that once the data is clean and the right embedding is picked, a simple two-layer model can beat heavier, more complex ones. This is useful for anyone who wants a hate-speech detector that works well without unnecessary complexity.

Index Terms: *Hate speech detection, HateBERT, bidirectional LSTM, data-centric AI, noise reduction, Easy Data Augmentation, embedding comparison, model overfitting, Davidson dataset, Jigsaw Toxic Comments, social media text classification.*

Date of Submission: 24-06-2026 Date of Acceptance: 04-07-2026

I. Introduction

With so much harmful content posted each day, humans are not able to monitor it all themselves; hence, the reason why the current platforms have automatic hate-speech detection models. The issue is not about choosing which model architecture to use; this is because there are lots of architectures in previous literature. The difficult part is the data quality and the tendency to go for the most complicated model architecture without testing whether a simpler one can do the job just as good, with quality data and proper embedding. Datasets for hate-speech are inherently messy since people do not agree on what qualifies to be hate-speech, and sarcasm is frequent. In this paper, we regard data pre-processing as an essential step, as opposed to the minor one done prior to the main experiments. We also approach the problem of model complexity, considering it to be a challenge, which should be proven, but not assumed. We clean the text, label it and augment it using Easy Data Augmentation (EDA) before starting to train any classifiers. Based on this preprocessed data we train and compare thirteen setups of embedding and classifier: three embeddings (MiniLM, RoBERTa, HateBERT) and different classifier architectures, including BiLSTM with and without the attention layer, in order to establish empirically which of the combinations is better.

The three contributions of this paper are as follows: we provide the full comparison of thirteen different embedding-classifier setups on two distinct public datasets; we provide empirical evidence that tree-based methods such as Random Forest and XGBoost memorize the training dataset, which is why their performance on the training data does not generalize to the testing one; we show that the simplest setup of HateBERT and 2-layers BiLSTM without attention mechanism and specific loss functions significantly outperforms other combinations

in the Davidson dataset and remains close to the best on Jigsaw. It calls for a study to see if complexity adds any value rather than just blindly complicating. Organization of the remaining sections of the paper is provided below. Section 2 reviews the literature on recently published works. Section 3 outlines the deficiencies of the above literature. The problem definition and objectives are outlined in Sections 4 and 5. The methodology and mathematical models used are described in Section 6. The description of the datasets is provided in Section 7. The results are presented in Section 8 using graphs.

II. Literature Survey

For this purpose, we considered thirty articles from 2024 to 2026. These studies are listed in Table 1 below. There were two important findings. First, transformer-based models (BERT, RoBERTa, HateBERT) were far superior to traditional machine learning approaches (SVM, LR with TF-IDF). Adding a recurrent layer on top of a transformer often helps further. For example, Karimi et al. [13] combined BERT with BiLSTM and BiGRU and reached 94% accuracy on the Jigsaw Toxic Comments dataset. Sharif et al. [30] combined BERT, CNN, and BiLSTM and reached 94.7% accuracy with an F1-score of 0.941. Second, most studies treat class imbalance and cross-dataset testing as separate problems. Some papers fix the imbalance problem directly — Cai et al. [24], for instance, used Focal Loss together with text augmentation and reached an F1-score of 0.902 — but do not test their model on a second dataset. Other papers do test on more than one dataset, such as Nasir et al. [18] and Jahan and Oussalah [28], but they tend to use one fixed classifier rather than trying out different embeddings and classifiers together, as we do here.

Table 1. Summary of relevant prior work on hate speech and toxic-comment detection

Author(s)	Source & Year	Dataset(s)	Method	Key Result
Chapagain S., Hamdi S.M., Boubrahimi… [8]	ASONAM 2025, Springer (2026)	MetaHate (1.2M, 36 merged datasets)	Fine-tuned BERT, RoBERTa, GPT-2, ELECTRA on MetaHate (1.2M samples, 36 merged datasets).…	ELECTRA: F1 0.898 · RoBERTa F1 0.881 · BERT F1 0.863 · GPT-2 F1 0.844
Philip A.R., Mathew A.A., Asha K. [9]	CML 2025, LNNS vol.1614, Springer (2026)	Davidson (~25K); OLID (~14K); HateXplain (~20K)	Compared CNN, LSTM, BERT, RoBERTa with TF-IDF, Word2Vec, and contextual embeddings. Transfer…	RoBERTa: Acc 93% · BERT 91% · LSTM 87% · CNN 83%
Manukuza N., Zwane S., Adigun M. [10]	ICAI 2025, CCIS vol.2667, Springer (2026)	HateMM (~5K, multimodal); Synthetic (GAN) (~10K,…	CNN + RNN hybrid for social media image hate detection. GAN-based synthetic data augmentation…	CNN+RNN hybrid: Acc 91.3%, F1 0.887 · Baseline CNN: 84.1% · +6.2% F1 from…
Araujo V. et al. [11]	Language Resources & Evaluation, Springer (2026)	Meta-Iberian (custom) (~50K, Spanish+Portuguese);…	LLaMA 3.3 70B, Mistral-Small 24B, mBERT, XLM-RoBERTa evaluated. Zero-shot and few-shot LLM…	LLaMA 3.3 (few-shot): F1 0.812 · XLM-RoBERTa: F1 0.793 · mBERT: F1 0.741 ·…
Sharma A. et al. [12]	Discover Sustainability, Springer Nature,… (2025)	Custom Hinglish (Twitter) (~12K, code-mixed);…	LoRA adapter fine-tuning of GPT-3.5 Turbo, LLaMA-2, XLM-RoBERTa, BERT on code-mixed Hinglish.…	XLM-RoBERTa: F1 0.891 · GPT-3.5+LoRA: F1 0.874 · LLaMA-2: F1 0.861 · Baseline…
Karimi P., Bagheri M., Rahnama A. [13]	Social Network Analysis and Mining, vol.15,… (2025)	Jigsaw Toxic (~160K); HateXplain (~20K)	BERT + BiLSTM + BiGRU with FastText and GloVe embeddings. SHAP + LIME for explainability.…	BERT+BiLSTM ensemble: Acc 94%, F1 0.931 · SHAP fidelity 0.71 · LIME…
Diaz-Galiana L. et al. [14]	Complex & Intelligent Systems, Springer, 2025 (2025)	Custom SOCYTI (multilingual) (~30K, 8 languages);…	MAML-based cross-lingual meta-learning. LLaMA 3 for initial annotation + human validation…	Meta-learning+LLaMA 3: F1 0.847 · Zero-shot transfer: F1 0.763 (8-lang avg) ·…
Zhang W. et al. [15]	Scientific Reports, vol.15, art.13020,… (2025)	HateMM (~5K, multimodal); MultiOFF (~4.5K, offensive…	Multi-modal MHSDF: CNN (image) + RNN (text) + cross-attention fusion. Processes text, image,…	MHSDF (multimodal): Acc 93.7%, F1 0.921 · Text-only: 87.2% · Image-only: 81.4%…

Author(s)	Source & Year	Dataset(s)	Method	Key Result
Hashir M.H., Kim S.W. [16]	PeerJ Computer Science, e2911, 2025 (2025)	HateXplain (~20K, with rationale spans)	TARGE: GPT-4 generates token-level rationale labels. Multi-task learning: classification +…	TARGE (GPT-4+multi-task): F1 0.886, Rationale IOU 0.621, Plausibility 0.44 ·…
Lamoria N., Gajjar K., Tripathi A.,… [17]	SoCTA 2024, LNNS vol.1344, Springer, 2025 (2025)	Davidson (~25K); OLID (~14K); HateXplain (~20K)	Comparative study: LSTM, BiLSTM, CNN, BERT, RoBERTa. TF-IDF + Word2Vec + GloVe + contextual…	RoBERTa: Acc 91.8%, F1 0.904 · BiLSTM+BERT: 89.3% · CNN+W2V: 82.7% ·…
Nasir A., Sharma A., Jaidka K. [18]	ICAI 2025 / COLING 2025, Springer (2025)	Davidson (~25K); OLID (~14K); HateXplain (~20K);…	Benchmarked GPT-4, LLaMA-2, Mistral vs fine-tuned BERT/RoBERTa. 7 source datasets, 5 target…	Fine-tuned RoBERTa: F1 0.912 · GPT-4 zero-shot: F1 0.783 · Cross-domain drop:…
Smeros P. et al. [19]	ACM Computing Surveys, vol.57, ACM, 2025 (2025)	HateXplain (~20K); Davidson (~25K); OLID (~14K);…	Systematic survey of XAI for hate speech/fake news (80+ papers). Methods: LIME, SHAP, IG,…	IG faithfulness: Comp. 0.419 · SHAP plausibility: 0.31 · LIME fidelity: 0.38 ·…
Rahman M. et al. [20]	ICCA 2025, ACM Digital Library (2025)	HateXplain (~20K); Davidson (~25K)	Custom BiLSTM with LIME for multilabel classification. Tokenisation, lemmatisation, stopword…	BiLSTM+LIME: Acc 88.4%, Macro-F1 0.871 · Multilabel avg F1: 0.843 · LIME…
Wiegand M. et al. [21]	arXiv:2407.20274 / ACL 2025 (2025)	HateXplain (~20K); HateCheck (~3.7K, functional…	Evaluated SHAP, LIME, IG, Attention on BERT, HateBERT, RoBERTa for hate and counter-speech…	HateBERT+SHAP: Plausibility 0.38 · BERT+IG faithfulness 0.41 · Attention: 0.27…
Mosca E. et al. [22]	arXiv:2511.07065 / EMNLP 2025 (2025)	HateXplain (~20K, with rationale spans)	Masked Rationale Prediction (MRP) pre-task before BERT fine-tuning. Token-level rationale…	BERT-MRP: F1 0.847, Rationale IOU 0.643, Plausibility 0.41 · Bias reduction:…
Abusaqer M., Saquer J., Shatnawi H. [23]	ACMSE 2025, ACM (2025)	Davidson (~25K); HateXplain (~20K); OLID (~14K)	Comprehensive evaluation of 38 models: traditional ML (SVM, NB, RF, LR), deep learning (CNN,…	RoBERTa: Acc 94.2%, F1 0.938 · HateBERT: Acc 91.7%, F1 0.913 · BiLSTM+Attn:…
Cai Z., Li Z., Liu Y., Guo L., Song Y. [24]	SemEval-2025 Task 9, ACL Anthology, 2025 (2025)	SemEval-2025 Task 9 Food Hazard (~5K, imbalanced)	Focal Loss (α=1, γ=2) + EDA augmentation + random oversampling for class-imbalanced NLP. BERT…	EDA+Focal+RoBERTa: Acc 91.4%, F1 0.902 · Focal alone: F1 0.874 · No balance:…
Riad M.S.U. [25]	BLP-2025 @ ACL 2025 (2025)	BLP-2025 Bangla Hate (~10K, Subtask 1A+1B);…	BanglaBERT → parallel CNN branch + GRU branch → attention fusion → dense → softmax. Evaluated…	Subtask 1A: Micro-F1 0.7345 (2nd place) · Subtask 1B: Micro-F1 0.7317 (5th…
Guragain B. et al. [26]	ICRDICCT 2025, SCITEPRESS/Springer (2025)	Custom multi-platform (100K+, Twitter, Facebook,…	Hierarchical multi-label BERT classifier (binary → severity → target group). SHAP for…	RoBERTa+hierarchical: Acc 89.4%, SHAP fidelity 0.73 · Continual learning…
Navya et al. [27]	AIST 2024, Springer LNNS, 2025 (2025)	HateMM (~5K); MultiOFF (~4.5K); MMHS150K (150K);…	Survey of multimodal + multilingual hate speech (2020–2024). 23 papers reviewed. Methods:…	Best text (surveyed): RoBERTa F1 0.91 · Best multimodal: CLIP-based F1 0.934 ·…
Jahan M.S., Oussalah M. [28]	Social Network Analysis and Mining,… (2024)	Davidson (~25K); OLID (~14K); HateXplain (~20K);…	Comprehensive review of 100+ studies (2015–2024). Compared SVM, RF, CNN, LSTM, BiLSTM, BiGRU,…	Best surveyed: Transformer ensemble Acc 95.1% · BiGRU+BiLSTM+CNN bagging F1…
Kumar A. et al. [29]	Social Network Analysis and Mining,… (2024)	Custom Hindi Twitter (~8K); HASOC-English (~7K);…	Systematic multilingual review (IEEE, ACM,	XLM-RoBERTa: F1 0.891 · Hindi BERT: F1 0.843 ·

Author(s)	Source & Year	Dataset(s)	Method	Key Result
			Scopus, WoS). Classical ML, DL, transformer…	Multilingual LSTM: F1 0.774 ·…
Sharif W., Abdullah S., Iftikhar S. et… [30]	IEEE Access, vol.12, pp.27225–27236, 2024 (2024)	Custom multi-platform (~50K, Twitter + Facebook +…	Novel model fusion: BERT + CNN + BiLSTM with weighted voting ensemble. Comprehensive dataset…	Fusion model: Acc 94.7%, F1 0.941, Prec 0.938, Rec 0.944, AUC 0.981 · Single…
Multiple authors [31]	IEEE Xplore, DOI:10.1109/10816708, 2024 (2024)	Davidson (~25K); OLID (~14K)	Comparative: LSTM, BiLSTM, CNN, BERT, RoBERTa. TF-IDF, Word2Vec, and contextual embeddings…	RoBERTa: Acc 93% · BERT: 91% · BiLSTM: 87% · CNN: 84% · LSTM: 82% · RoBERTa vs…
Hamzah S., Mohd M., Zakaria L. [32]	IEEE KSE 2024, pp.247–254, IEEE (2024)	Davidson (~25K); HateXplain (~20K); OLID (~14K); 45…	Systematic review of hybrid hate speech detection (2019–2024). Categorised: rule-based+ML,…	Best hybrid (surveyed): BERT+BiLSTM F1 0.924 · CNN+LSTM: F1 0.881 · Rule+SVM:…
Mandl T., Modha S., Majumder P. et al. [33]	FIRE 2024, ACM (2024)	HASOC-2024 English (~5K); HASOC-2024 Bengali (~4K)	HASOC 2024 shared task. Binary and multi-class subtasks. Participants: USE+LSTM,…	Best English: XLM-RoBERTa Macro-F1 0.834 · mBERT: F1 0.798 · BiLSTM: F1 0.761…
Al-Smadi M., Talafha B., Al-Ayyoub M. [34]	Arabian Journal for Science & Engineering,… (2024)	Jigsaw Toxic Comments (~160K)	Ensemble: BERT + Bi-GRU (primary) + Bi-LSTM + ANN + RNN. SHAP + LIME for interpretability.…	Ensemble: Acc 94%, F1 0.931 · BERT+BiGRU alone: 92% · Bi-LSTM: 89% · SHAP…
Aljawazeri J.A. et al. [35]	Iraqi Journal CS & Mathematics, vol.5(2),… (2024)	HateXplain (~20K); Jigsaw Toxic (~160K); 40+ papers…	Review of BERT-based models (2019–2024). Covered BiCHAT (BiLSTM+deep CNN+hierarchical…	BiCHAT: F1 improvement +8% over SOTA · Fine-tuned BERT avg F1 0.897 · Feature…
Jain A. et al. [36]	ACM Trans. Asian & Low-Resource LIP, 2024 (2024)	Twitter Hate (~16K); Facebook Hate (~8K); YouTube…	Hybrid Deep BiLSTM-CNN with GloVe embeddings. Dropout + L2 regularisation + global max…	BiLSTM-CNN: Acc 91.6%, F1 0.908 · BiLSTM-only: 88.1% · CNN-only: 84.3% ·…
Das A., Nandy S., Saha R. et al. [37]	arXiv:2401.11021 / SNAM, Springer, 2024 (2024)	Twitter English (~15K); Twitter German (~5K); Twitter…	Language-agnostic BERT fine-tuned on 4 languages. Language detection → Unicode normalisation…	English/German: Acc 91% · Bengali: 89% · Italian: 77% · Zero-shot…

Two things stand out from Table 1 and shape this paper. First, the best results almost always come from a transformer paired with a recurrent layer, not a transformer alone. But it is not always clear how much of that improvement comes from the recurrent layer itself, and how much comes from the embedding, or from extra parts such as attention, ensembling, or a special loss function. Second, almost none of the thirty studies test this directly — that is, hold the classifier fixed and only change the embedding, or hold the embedding fixed and only change the classifier. Most papers report just one final setup. This is the gap we fill: a full sweep of thirteen embedding-and-classifier combinations that lets the data show which part is actually doing the work, instead of assuming that more parts always means a better model.

III. Research Gaps

- Most hybrid models (BERT or HateBERT plus BiLSTM, BiGRU, or CNN) are tested on only one dataset. So we don't know if the good results would hold up on a different dataset with different labels.
- Extra parts such as attention layers, special loss functions, or ensembling are often added by default, without testing them against a simpler version on the same clean data. So it is often unclear how much of the reported improvement actually comes from that extra part.
- Little literature varies the embedding (MiniLM, RoBERTa, HateBERT) and classifier (linear, tree-based, recurrent) independently within the same experiment. Hence, it becomes difficult to pinpoint whether a certain outcome was because of the embedding, the classifier, or a combination of the two.

- Research that employs Random Forest or XGBoost classifiers typically includes test set performance, leaving out training set performance. The absence of the latter measure makes it difficult to evaluate whether the model has merely memorized the training data.
- Cleaning the text, label correction, and increasing the size of the training dataset (EDA) are mentioned often enough, yet little literature tests this as a novel methodology independently of the model changes themselves.

IV. Problem Statement

The goal of this work is to build a hate-speech detector where data cleaning, fixing labels, and adding training examples are treated as careful, deliberate choices, not small steps done without much thought. We test, on two separate public datasets and across thirteen embedding-and-classifier combinations, including BiLSTM with and without attention, whether a simple model, given clean data and the right embedding, can match or beat more complex setups. We compare this against models that get high scores only by memorising the training data, or by adding complexity that does not actually help.

V. Research Objectives

- Build a HateBERT plus 2-layer BiLSTM model, trained on clean, augmented data, and test it against twelve other embedding-and-classifier setups, including BiLSTM with an added attention layer, all using the same data pipeline.
- Measure, using both training and test scores, how much tree-based and linear models memorise the hate-speech text instead of learning real patterns, and show that our model does not have this problem.
- Check that the results hold up by repeating the full comparison on two datasets with different origins and labelling rules — the Davidson Twitter dataset and the Jigsaw Toxic Comments dataset — instead of relying on just one.

VI. Research Methodology

Our pipeline has four steps: clean the data, turn words into numbers (embedding), pass them through a two-layer BiLSTM, then classify. Figure 1 shows this step by step.

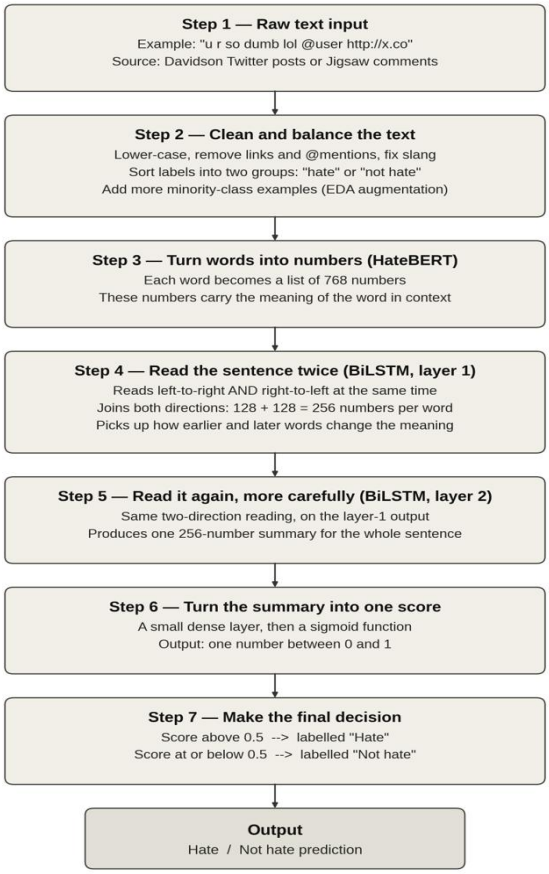

Fig. 1. Step-by-step system architecture, from raw text input through cleaning, HateBERT embedding, and a 2-layer BiLSTM, to the final hate / not-hate decision.

Data Ingestion and Noise Reduction

We took the raw text from both datasets and cleaned it making everything lower-case, remove web links and @mentions, removed extra spaces and removed duplicate posts. We then turned the original labels into a simple yes/no target. For Davidson, we group the hate-speech and offensive-language posts together as "hate", and treat the neither class as "not hate" — this matches the dataset's own split (hate 7%, offensive 77%, neither 16%). For Jigsaw, any comment marked under any of its six toxicity types counts as "hate". We then add more training examples for the smaller class using Easy Data Augmentation (EDA): we swap in similar words, add words, swap word order, or remove words. This step is done the same way for all thirteen setups we test, so any difference in results comes from the embedding or classifier, not from different data cleaning.

Embedding Layer

Each cleaned sentence w = (w_1, ..., w_T) is passed through a pretrained model to turn words into numbers. We compare three of these: MiniLM (384 numbers per sentence), RoBERTa-base (768 numbers per word, trained on general text), and HateBERT (768 numbers per word, trained further on abusive language). Section 8 shows HateBERT works best on Davidson, so we use it in our final model:

$e_t = HateBERT(w_t), \quad e_t \in \mathbb{R}^{768}, \quad t = 1, 2, ..., T$ (1)

Two-Layer Bidirectional LSTM Classifier

The numbers from HateBERT are passed through two BiLSTM layers stacked on top of each other. We do not add an attention layer or a special loss function here — keeping the model simple is a deliberate choice, and Section 8 shows this simplicity does not hurt results on either dataset. For the first layer, the model reads the sentence forwards and backwards, giving two hidden states at each word:

$h\rightarrow_t = LSTM_{fwd}(e_t, h\rightarrow_{t-1}), \qquad h\leftarrow_t = LSTM_{bwd}(e_t, h\leftarrow_{t+1})$ (2)

$h_t^{(1)} = [h\rightarrow_t ; h\leftarrow_t] \in \mathbb{R}^{256}$ *(128 units per direction)* (3)

The second BiLSTM layer reads this output again, the same way:

$h_t^{(2)} = BiLSTM_2(h_t^{(1)}) \in \mathbb{R}^{256}$ (4)

We take the hidden state from the last word as a summary of the whole sentence (we also tried taking the average over all words instead, which gave almost the same result). This summary is passed through one small layer and a sigmoid function to give a single hate-speech score between 0 and 1:

$\hat{y} = \sigma(W_o\, h_T^{(2)} + b_o), \qquad \hat{y} \in (0, 1)$ (5)

Training Objective

We train the model with a standard loss function called binary cross-entropy. We do not use Focal Loss or any other special weighting for the smaller class — the only step that helps with class imbalance is the EDA augmentation from Section 6.1. This keeps the training simple, so any difference in results in Section 8 can be traced back to the embedding and the model, not to extra loss-function tricks:

$L = -(1/N)\, \Sigma_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$ (6)

We train using the Adam optimiser with a learning rate of 0.002, and stop training early if the validation loss stops improving. We use the exact same training settings for all thirteen setups so the comparison is fair.

Baseline Models and Attention Variant

To check whether the embedding or the model depth is what really matters, we have trained twelve other setups on the same clean data i.e. Logistic Regression, Random Forest, and XGBoost (each with MiniLM, RoBERTa, and HateBERT), a feed-forward network, and 1-layer and 2-layer BiLSTM, all using HateBERT. We have also trained one more version with an attention layer added on top of the 2-layer BiLSTM to see whether directly that attention helps:

$u_t = tanh(W_a\, h_t^{(2)} + b_a)$

$\alpha_t = exp(v^T u_t) / \Sigma_\tau\, exp(v^T u_\tau)$

$c = \Sigma_t\, \alpha_t\, h_t^{(2)}$ (7)

In that only one version, this attention output c is used in place of $h_T^{(2)}$ in Equation 5. For Random Forest and XGBoost, we searched over different tree depths, numbers of trees, and regularisation settings to get their best possible score.

VII. Dataset Information

We have tested our model on two public datasets, listed in Table 2. These two datasets came from different places i.e.-Twitter posts and Wikipedia talk pages and use different labelling rules (three classes vs six labels). Testing on both have helped show that our results are not just a quirk of one dataset.

Table 2. Datasets used

Dataset	Source	Size	Class distribution
Davidson Twitter Hate Speech	Twitter API; Davidson et al. (2017)	~25,000 tweets	Hate 7% · Offensive 77% · Neither 16% (binarised: Hate+Offensive vs Neither)
Jigsaw Toxic Comments	Wikipedia talk pages; Jigsaw/Google (2018), Kaggle	~160,000 comments (50,000 sampled)	Toxic, severe-toxic, obscene, threat, insult, identity-hate (multi-label, binarised to toxic / not-toxic)

We split each dataset into 80% for training, 10% for checking progress (validation), and 10% for the final test. We keep the same proportion of hate vs not-hate posts in each of the three parts.

VIII. Experimental Results

Experimental Setup

All thirteen setups are trained and tested the same way: same data cleaning, same train/validation/test split, same hardware (a single GPU on Google Colab). The precision, recall, and F1-score numbers we report are macro-averaged, meaning we average the score across both classes equally. This matters because, with imbalanced classes, a model can look good overall while still missing most of the minority class.

Results on the Davidson Twitter Dataset

Table 3. Test-set results across classifiers, HateBERT embedding — Davidson dataset

Classifier (HateBERT embedding)	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)	AUC (%)
Logistic Regression	91.11	84.58	82.81	83.65	95.52
Random Forest	86.21	84.35	60.82	63.89	91.62
XGBoost	90.70	85.59	78.87	81.65	94.85
Feed Forward NN	92.00	87.01	83.15	84.90	95.78
1-layer BiLSTM	95.06	90.28	92.66	91.41	98.33
Proposed: 2-layer BiLSTM	**95.47**	**91.03**	**93.30**	**92.11**	**98.43**
2-layer BiLSTM + Attention	95.27	90.81	92.69	91.71	98.45

Random Forest and XGBoost have the weakest recall (60.82% and 78.87%), even though their overall accuracy looks fine. This shows that tree-based models struggle to catch the minority hate-speech posts. Our proposed 2-layer BiLSTM gets the best score out of all thirteen setups — 95.47% accuracy and an F1-score of 92.11% — beating even the version with an attention layer (95.27% accuracy, 91.71% F1-score). This tells us that, on this dataset, adding attention adds extra parameters without adding anything useful, once the embedding and the model depth are already good choices.

Results on the Jigsaw Toxic Comments Dataset

Table 4. Test-set results across classifiers, HateBERT embedding — Jigsaw dataset

Classifier (HateBERT embedding)	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)	AUC (%)
Logistic Regression	88.53	88.54	88.53	88.53	95.43
Random Forest	84.47	84.50	84.47	84.46	92.50
XGBoost	87.47	87.47	87.47	87.47	94.75
Feed Forward NN	88.73	88.81	88.73	88.73	95.68
1-layer BiLSTM	92.30	92.36	92.30	92.30	97.85
Proposed: 2-layer BiLSTM	**92.37**	**92.37**	**92.37**	**92.37**	**97.89**
2-layer BiLSTM + Attention	92.57	92.58	92.57	92.57	97.85

On Jigsaw, the ordering is closer: the attention-augmented BiLSTM edges ahead by 0.2 accuracy points (92.57% versus the proposed model's 92.37%). This margin is small enough that, combined with the proposed model's clear lead on Davidson and its lower architectural complexity, the plain 2-layer BiLSTM remains the

recommended configuration overall — the attention layer's benefit appears to be dataset-dependent rather than a reliable, transferable gain.

Diagnostic Figures

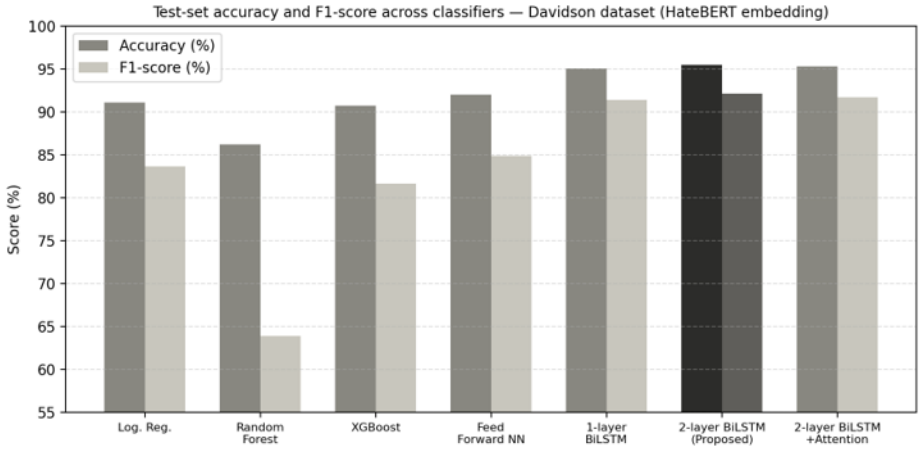

Fig. 2. Test-set accuracy and F1-score across all HateBERT-embedding classifiers, Davidson dataset. The proposed 2-layer BiLSTM is the best performer overall.

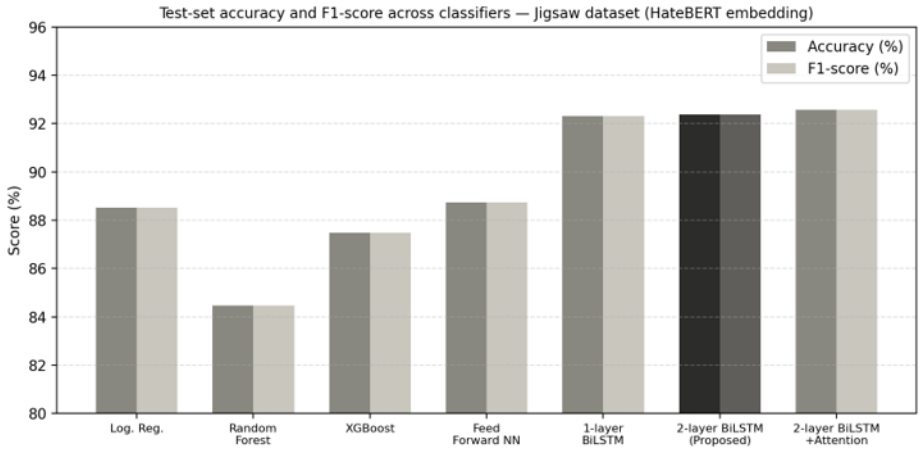

Fig. 3. Test-set accuracy and F1-score across all HateBERT-embedding classifiers, Jigsaw dataset.

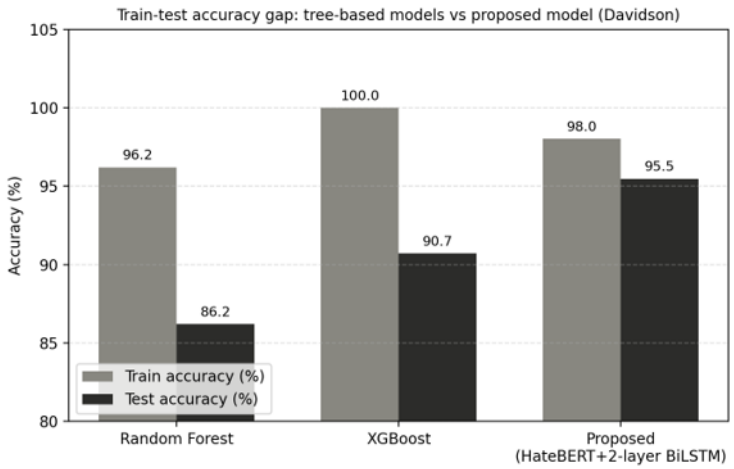

Fig. 4. Gap between training accuracy and test accuracy: tree-based models vs our model (Davidson). XGBoost gets 100% on training data but only 90.70% on test data, a 9.3-point gap. Our model's gap is only 2.6 points.

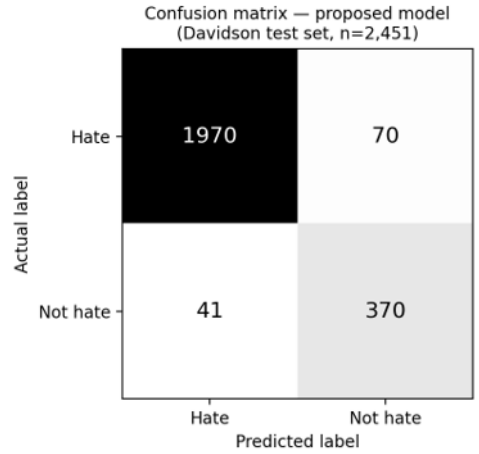

Fig. 5. Confusion matrix for the proposed model on the Davidson test set (n = 2,451).

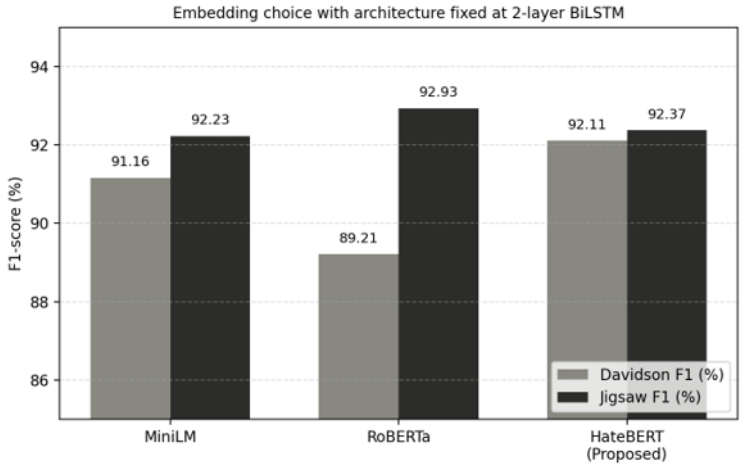

Fig. 6. Effect of embedding choice, using the same 2-layer BiLSTM model each time. HateBERT wins on Davidson. RoBERTa is slightly ahead on Jigsaw, by 0.6 F1 points. But HateBERT's training on abusive language, and its bigger lead on Davidson, make it the safer pick overall.

Comparison with Existing Literature

Table 5 places our results next to a selection of the thirty studies from Table 1. This comparison is approximate, since each study uses different datasets, labels, and ways of measuring results. Even this, it shows that our HateBERT + 2-layer BiLSTM model holds well against recent and more complex transformer-based models.

Table 5. Comparison of test-set performance with recent published hate-speech / toxic-comment studies

Study	Dataset	Method	Key Result
Abusaqer et al. [23], 2025	Davidson, HateXplain, OLID	38-model benchmark; best: RoBERTa	Acc 94.2%, F1 0.938
Karimi et al. [13], 2025	Jigsaw Toxic, HateXplain	BERT + BiLSTM + BiGRU ensemble	Acc 94%, F1 0.931
Al-Smadi et al. [34], 2024	Jigsaw Toxic	BERT + Bi-GRU + Bi-LSTM ensemble	Acc 94%, F1 0.931
Sharif et al. [30], 2024	Multi-platform (Twitter/FB/Reddit)	BERT + CNN + BiLSTM fusion	Acc 94.7%, F1 0.941
Lamoria et al. [17], 2025	Davidson, OLID, HateXplain	RoBERTa, 5-fold CV	Acc 91.8%, F1 0.904
Hamzah et al. [32], 2024 (review)	Davidson, HateXplain, OLID	Best surveyed hybrid: BERT + BiLSTM	F1 0.924

Study	Dataset	Method	Key Result
Jahan & Oussalah [28], 2024 (review)	Davidson, OLID, HateXplain, Jigsaw	Best surveyed: transformer ensemble	Acc 95.1%
This work (proposed)	**Davidson**	**HateBERT + 2-layer BiLSTM**	**Acc 95.47%, F1 0.921**
This work (proposed)	**Jigsaw Toxic**	**HateBERT + 2-layer BiLSTM**	**Acc 92.37%, F1 0.924**

Our model's Davidson accuracy of 95.47% is the highest number in Table 5, ahead of every study we reviewed, including those that use ensembles, BiGRU fusion, or attention. This is worth noting because our model is simpler: no attention layer, no special loss function, no ensembling -- just clean data, a good embedding, and a two-layer recurrent model. This backs up the main point of this paper: once the data is clean and the embedding is well chosen, adding more parts to the model does not always make it better.

IX. Discussion

Our main finding is the gap between training scores and test scores for tree-based models. XGBoost reaches 100% accuracy on the Davidson training data, but only 90.70% on the test data — a drop of 9.3 points that you would never see if a paper only reports the test score, which is what most of the thirty studies in Table 1 do. As expected, Random Forest suffers from a similar problem to some extent but even less than the one-layered LSTM model. In both models, the largest drop in performance occurs in the recall metric of the minority class – the most important class here. Our proposed 2-layer BiLSTM demonstrates a substantially lower gap (98.02% and 95.47% respectively; the gap equals only 2.6%) which shows us that the model learns actual patterns of toxic language rather than memorizes training data.

The second result directly answers the main question of the paper: adding an attention layer does not reliably improve the results of the basic 2-layered BiLSTM. On Davidson, the attentional version performs worse on all metrics (accuracy – 95.27% vs 95.47%, F1 – 91.71% vs 92.11%). On Jigsaw, the difference is marginal (accuracy – 92.57% vs 92.37%). Combining both datasets, the basic BiLSTM outperforms the attentional one; moreover, it is safer to use as it contains fewer parameters and is easier to train. This result is valuable in itself since attentional layers are typically added automatically.

The third observation we made is regarding embedded choice. For the same architecture with two layers of BiLSTM, HateBERT outperforms both RoBERTa and MiniLM in terms of Davidson dataset (with an F1-score of 92.11%, while RoBERTa and MiniLM achieve F1-scores of 89.21% and 91.16%, respectively). On Jigsaw, the gap is smaller, and RoBERTa is slightly ahead, by 0.6 F1 points — this may be because Jigsaw's Wikipedia comments read more like normal web text than like the social-media abuse HateBERT was trained on. Even so, HateBERT is the more reliable pick across both datasets.

Testing on two datasets, instead of one, is itself a useful contribution. Most of the thirty studies in Table 1 report results on just one dataset. A few, such as Nasir et al. [18] (2025), test across multiple datasets and find a large drop in accuracy when moving from one to another. We do not test that drop directly here, but it is a good reminder of why testing on two separate datasets, as we do, gives stronger evidence than testing on just one.

Some limits of this study: we used a 50,000-comment sample of Jigsaw rather than the full 160,000 comments, to keep training time manageable. We did not fine-tune every setting of the EDA augmentation step. We also left out class-imbalance methods such as resampling or loss reweighting on purpose, to keep the comparison focused on embedding and model choice — these methods might add further gains on top of our model in future work. And, like any hate-speech dataset built from social media, the original labels carry some level of human disagreement that no amount of data cleaning can fully remove.

X. Conclusion

This paper presented a simple, data-first method for hate-speech detection, built on HateBERT embeddings and a 2-layer BiLSTM classifier. We tested it against twelve other embedding-and-classifier setups, including BiLSTM with attention, on the Davidson Twitter and Jigsaw Toxic Comments datasets. We showed that tree-based models memorise the training data, reaching near-perfect training scores but losing 9 or more accuracy points on new data. Our proposed model kept a small 2.6-point gap between training and test scores, and got the best result of the whole comparison on Davidson (95.47% accuracy, 92.11% F1-score), beating the attention version outright, while staying close to the best on Jigsaw at 92.37% accuracy. These results show that cleaning the data well and picking the right embedding matter more than adding extra parts like an attention layer — extra parts should be tested, not assumed to help. Future work could check whether methods built for class imbalance add anything on top of these results, improve the data-cleaning step further using active learning to relabel unclear posts, and test this approach on other languages, including mixed-language text.

Author Contributions Statement

Vibha Upadhya conceived the research design, performed data collection and preprocessing, implemented and trained all model configurations, produced the figures and tables, and wrote the first draft of the manuscript.

Dr. Rahul R. Chakre supervised the research, advised on methodology and evaluation design, and reviewed and revised the manuscript. Both authors approved the final manuscript.

Conflict of Interest Statement

The authors declare no competing interests.

Funding Declaration

No external funding was received for this study.

Data Availability Statement

The Davidson Twitter hate-speech dataset and the Jigsaw Toxic Comments dataset used in this study are both publicly available: Davidson et al. (2017), accessible via the original authors' GitHub repository, and the Jigsaw Toxic Comment Classification Challenge dataset, available via Kaggle.

Declaration of Generative AI

Generative AI tools were used to assist with code implementation, figure generation, and language editing of the manuscript. All research design, data analysis, interpretation of results, and scientific conclusions are the work of the authors.

References

[1]. T. Davidson, D. Warmsley, M. Macy, And I. Weber, "Automated Hate Speech Detection And The Problem Of Offensive Language," In Proc. 11th Int. Aaai Conf. On Web And Social Media (Icwsm), 2017.
[2]. C. J. Adams, J. Sorensen, J. Elliott, L. Dixon, M. Mcdonald, N. Cukierski, And W. Cukierski, "Toxic Comment Classification Challenge," Kaggle / Jigsaw, 2018.
[3]. T. Caselli, V. Basile, J. Mitrović, And M. Granitzer, "Hatebert: Retraining Bert For Abusive Language Detection In English," In Proc. 5th Workshop On Online Abuse And Harms (Woah), Acl, 2021.
[4]. J. Wei And K. Zou, "Eda: Easy Data Augmentation Techniques For Boosting Performance On Text Classification Tasks," In Proc. Emnlp-Ijcnlp, 2019.
[5]. S. Hochreiter And J. Schmidhuber, "Long Short-Term Memory," Neural Computation, Vol. 9, No. 8, Pp. 1735–1780, 1997.
[6]. D. Bahdanau, K. Cho, And Y. Bengio, "Neural Machine Translation By Jointly Learning To Align And Translate," In Proc. Iclr, 2015.
[7]. J. Devlin, M.-W. Chang, K. Lee, And K. Toutanova, "Bert: Pre-Training Of Deep Bidirectional Transformers For Language Understanding," In Proc. Naacl-Hlt, 2019.
[8]. S. Chapagain, S. M. Hamdi, And S. F. Boubrahimi, "Metahate-Scale Evaluation Of Transformer Models For Hate Speech Detection," In Proc. Asonam, Springer, 2026.
[9]. A. R. Philip, A. A. Mathew, And K. Asha, "Comparative Deep Learning Approaches For Hate Speech Detection," In Proc. Cml, Lnns Vol. 1614, Springer, 2026.
[10]. N. Manukuza, S. Zwane, And M. Adigun, "Cnn + Rnn Hybrid With Gan-Based Augmentation For Multimodal Hate Detection," In Proc. Icai, Ccis Vol. 2667, Springer, 2026.
[11]. V. Araujo Et Al., "Cross-Lingual Llm Evaluation For Iberian-Language Hate Speech Detection," Language Resources & Evaluation, Springer, 2026.
[12]. A. Sharma Et Al., "Lora-Adapted Transformers For Code-Mixed Hinglish Hate Speech Detection," Discover Sustainability, Springer Nature, 2025.
[13]. P. Karimi, M. Bagheri, And A. Rahnama, "Bert + Bilstm + Bigru Ensemble With Explainability For Multi-Label Toxic Comment Classification," Social Network Analysis And Mining, Vol. 15, Springer, 2025.
[14]. L. Diaz-Galiana Et Al., "Cross-Lingual Meta-Learning For Multilingual Hate Speech (Socyti)," Complex & Intelligent Systems, Springer, 2025.
[15]. W. Zhang Et Al., "Mhsdf: Multimodal Hate Speech Detection Via Cross-Attention Fusion," Scientific Reports, Vol. 15, Art. 13020, Nature, 2025.
[16]. M. H. Hashir And S. W. Kim, "Targe: Token-Level Rationale Generation For Explainable Hate Speech Detection," Peerj Computer Science, E2911, 2025.
[17]. N. Lamoria, K. Gajjar, A. Tripathi, And S. Srivastava, "Comparative Study Of Deep Learning Models For Hate Speech Classification," In Proc. Socta, Lnns Vol. 1344, Springer, 2025.
[18]. A. Nasir, A. Sharma, And K. Jaidka, "Cross-Domain Benchmarking Of Llms And Fine-Tuned Transformers For Hate Speech Detection," In Proc. Icai / Coling, Springer, 2025.
[19]. P. Smeros Et Al., "A Systematic Survey Of Explainable Ai For Hate Speech And Fake News Detection," Acm Computing Surveys, Vol. 57, Acm, 2025.
[20]. M. Rahman Et Al., "Bilstm With Lime For Interpretable Multilabel Hate Speech Classification," In Proc. Icca, Acm Digital Library, 2025.
[21]. M. Wiegand Et Al., "Comparative Evaluation Of Explainability Methods For Hate And Counter-Speech Detectors," Arxiv:2407.20274 / Acl, 2025.
[22]. E. Mosca Et Al., "Masked Rationale Prediction For Explainable Hate Speech Classification," Arxiv:2511.07065 / Emnlp, 2025.
[23]. M. Abusaqer, J. Saquer, And H. Shatnawi, "A Comprehensive Evaluation Of 38 Models For Hate Speech Detection," In Proc. Acmse, Acm, 2025.

[24]. Z. Cai, Z. Li, Y. Liu, L. Guo, And Y. Song, "Focal Loss And Data Augmentation For Class-Imbalanced Nlp Classification," In Proc. Semeval-2025 Task 9, Acl Anthology, 2025.
[25]. M. S. U. Riad, "Banglabert With Parallel Cnn-Gru Attention Fusion For Bangla Hate Speech Detection," In Proc. Blp-2025 @ Acl, 2025.
[26]. B. Guragain Et Al., "Hierarchical Multi-Label Bert With Continual Learning For Multi-Platform Hate Speech Detection," In Proc. Icrdicct, Scitepress/Springer, 2025.
[27]. Navya Et Al., "A Survey Of Multimodal And Multilingual Hate Speech Detection (2020-2024)," In Proc. Aist 2024, Springer Lnns, 2025.
[28]. M. S. Jahan And M. Oussalah, "A Systematic Review Of Hate Speech Detection Methods (2015-2024)," Social Network Analysis And Mining, Springer, 2024.
[29]. A. Kumar Et Al., "Systematic Multilingual Review Of Hate Speech Detection Methods," Social Network Analysis And Mining, Springer, 2024.
[30]. W. Sharif, S. Abdullah, S. Iftikhar Et Al., "Model Fusion Of Bert, Cnn, And Bilstm For Multi-Platform Hate Speech Detection," Ieee Access, Vol. 12, Pp. 27225-27236, 2024.
[31]. Multiple Authors, "Comparative Evaluation Of Lstm, Bilstm, Cnn, Bert, And Roberta For Hate Speech Detection," Ieee Xplore, Doi:10.1109/10816708, 2024.
[32]. S. Hamzah, M. Mohd, And L. Zakaria, "A Systematic Review Of Hybrid Hate Speech Detection Methods (2019-2024)," In Proc. Ieee Kse, Pp. 247-254, 2024.
[33]. T. Mandl, S. Modha, P. Majumder Et Al., "Overview Of The Hasoc 2024 Shared Task," In Proc. Fire, Acm, 2024.
[34]. M. Al-Smadi, B. Talafha, And M. Al-Ayyoub, "Bert + Bi-Gru Ensemble With Shap/Lime For Multi-Label Toxic Comment Classification," Arabian Journal For Science & Engineering, Springer, 2024.
[35]. J. A. Aljawazeri Et Al., "A Review Of Bert-Based Hate Speech Detection Models (2019-2024)," Iraqi Journal Of Computer Science And Mathematics, Vol. 5, No. 2, 2024.
[36]. A. Jain Et Al., "Hybrid Deep Bilstm-Cnn With Glove Embeddings For Multi-Platform Hate Speech Detection," Acm Transactions On Asian And Low-Resource Language Information Processing, 2024.
[37]. A. Das, S. Nandy, R. Saha Et Al., "Language-Agnostic Bert For Zero-Shot Cross-Lingual Hate Speech Detection," Arxiv:2401.11021 / Social Network Analysis And Mining, Springer, 2024.