Fine-tune BERT for Implicit Hate Speech Binary-Classification on Social Media Corpora

Omaira Jaffar, Zakia Jalil, Muhammad Nasir
Faculty of Computing & Information Technology, International Islamic University, Islamabad, 44000, Pakistan
*Correspondence: omaira.msds4@student.iiu.edu.pk; zakia.jalil@iiu.edu.pk; m.nasir@iiu.edu.pk

Citation | Jaffar. O, Jalil. Z, Nasir. M, "Fine-tune BERT for Implicit Hate Speech Binary-Classification on Social Media Corpora", IJIST, Vol. 08 Issue. 01 pp 226-239, January 2026
Received| December 22, 2025 **Revised|** January 20, 2026 **Accepted|** January 24, 2026 **Published|** January 28, 2026.

Implicit hate speech is a challenge to automated content moderation systems as it is expressed using indirect, context-dependent, and culturally subtle expressions, as opposed to explicit hate speech. It is often expressed using sarcasm, stereotypes, coded language, and metaphors. This paper examines the effectiveness of contextual transformer models for detecting fine-grained implicit hate speech on social media platforms. A BERT-based classification model is designed using a unified dataset created by combining the Davidson Hate Speech, Jigsaw Toxic Comment, and Implicit Hate Speech datasets. A structured preprocessing workflow is used to handle class imbalance, annotation diversity, and domain shifts. The experimental results show that the proposed model achieves over 94% accuracy, along with high F1-score and recall values for the hate class. The results validate that token-level contextual features learned using self-attention effectively capture subtle linguistic patterns. In summary, this paper emphasizes the significance of contextual learning using transformer models for developing robust and scalable automated moderation systems.

Keywords: Implicit Hate Speech; BERT; Binary Classification; Social Media Analysis; Natural Language Processing; Transformer Models and Text Classification

Introduction:

Social media-based hate speech has recently been identified as a major societal issue owing to the anonymity, speed of information dissemination, and massive participation facilitated by social media platforms. Although considerable progress has been achieved in the automatic detection of hate speech, most of the existing literature has been dominated by approaches that mainly target explicit hate speech, which is defined by the presence of explicit slurs, offensive keywords, or direct attacks on individuals or groups [1]. The presence of these surface-level features makes explicit hate speech relatively easier to detect. Implicit hate speech, on the other hand, is often manifested by stereotypes, implications, sarcasm, metaphors, or coded messages, where malicious intent is expressed implicitly without using explicit offensive language. The subtle and context-dependent nature of implicit hate speech makes it considerably more difficult to detect using automated systems [2].

Conventional keyword lists and lexicon-based approaches are not very effective at detecting implicit hate speech. These approaches rely on pre-defined word lists and straightforward linguistic features, which makes them ineffective at capturing the underlying meaning, pragmatic use, and the contextual cues that convey the hate message hidden beneath the surface [3]. Recently, the emergence of transformer models such as BERT has proved to be highly effective at capturing contextual representations using self-attention mechanisms. Fine-tuning can result in overfitting, particularly when working with small or imbalanced datasets. Using a single dataset can introduce dataset-specific bias. Moreover, class imbalance is a significant issue, as hate speech is always overshadowed by the majority class of non-hate speech [4].

To overcome these challenges, this research work proposes a multi-corpus learning approach that combines various social media datasets from different sources. The research workflow involves harmonizing the formats of the datasets, transforming the annotation schemes to a binary label space, fine-tuning a pre-trained BERT model with an imbalance loss function, and testing the performance on multiple corpora. The main goals of this research work are threefold:

To enhance implicit hate speech detection through fine-tuned contextual representations.

To improve cross-dataset generalization via unified preprocessing and label harmonization.

To mitigate the adverse effects of class imbalance using focal loss.

The key novelty of this work lies in the systematic harmonization of heterogeneous hate speech annotations into a consistent binary formulation specifically designed to capture implicit and subtle forms of hate speech.

Material:

Related Work:

Detection of hate speech has been a popular area of research in NLP, with early solutions mainly targeting explicit hate speech using traditional machine learning approaches [5]. These solutions mainly used manually designed lexical features, n-grams, sentiment features, and keyword-based lexicons, along with classifiers such as Support Vector Machines and Logistic Regression. Although these solutions were successful in detecting explicit hate speech [6], they fail to detect implicit hate speech [7].

With the advent of deep learning, models based on neural networks such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks were developed to incorporate contextual information beyond mere lexical features [8]. These models showed better performance than traditional models; however, their capacity to represent long-distance dependencies and pragmatic meaning remained limited, especially when it came to sarcasm or cultural references [9].

Recently, transformer-based models have emerged as the new standard in hate speech detection research. BERT and its variants have gained immense popularity because of their

ability to capture deep bidirectional context through self-attention [10]. Various studies have revealed that fine-tuned versions of BERT achieve better results compared to the previous state of the art on explicit hate speech detection tasks. However, most of the existing literature is concerned with explicit hate/toxicity detection, and less attention has been paid to implicit hate speech [11].

However, a recent body of research has started to address the issue of implicit hate speech detection, with a focus on the importance of contextual and discourse-level understanding. Some research has investigated the use of sentence-level embeddings such as Sentence BERT (sBERT) for hate speech detection based on semantic similarity, while others have proposed hybrid or ensemble methods. Nevertheless, there is a lack of empirical research that compares token-level contextual models with sentence-level embedding methods for hate speech detection [7].

This research adds to the existing literature by conducting a focused analysis on implicit hate speech detection using contextualized transformer models. Through the assessment of performance on multiple datasets and the emphasis on fine-grained contextual understanding, this research makes new empirical contributions to the existing knowledge on hate speech detection.

Proposed Methodology:

This section outlines the proposed framework for detecting implicit hate speech, including the construction of the dataset, preprocessing, architecture, and fine-tuning approach. The overall workflow of the proposed implicit hate speech detection framework is illustrated in Figure 1. The aim is to utilize the contextualized transformer embeddings to detect hate speech that is implicit [12].

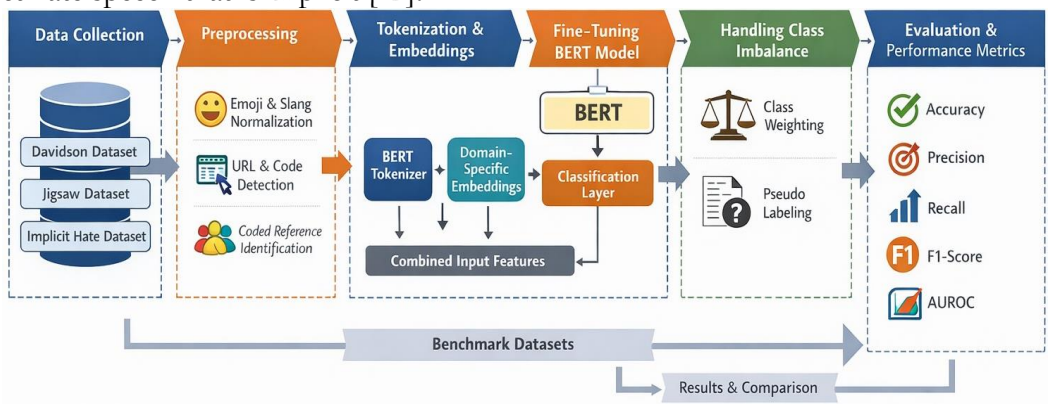

Figure 1. Proposed Methodology for Implicit Hate Speech Detection

The process includes dataset integration, preprocessing, BERT-based tokenization and embedding, fine-tuning for binary classification, class imbalance handling, and evaluation using standard performance metrics.

Dataset Construction:

A unified corpus was developed by combining three popular benchmark datasets: the Davidson Hate Speech dataset [13], the Jigsaw Toxic Comment dataset [14], and the Implicit Hate Speech dataset [15]. These datasets collectively encompass explicit, semi-explicit, and implicit forms of hate speech across various social media platforms.

The original annotation schemes and label granularities used in each of the datasets were different. To make them uniform, all datasets were transformed into a common binary classification format, where label 1 indicates hate speech and label 0 indicates non-hate speech. For the Davidson dataset, only hate speech and neutral examples were kept. The Jigsaw dataset, which was multi-label, was transformed into a binary classification format by assigning all toxic examples to the hate class. In the Implicit Hate Speech dataset, the explicit and implicit

hate classes were combined into a single hate class. After normalization, all datasets were combined into a single corpus and then randomly shuffled with a fixed seed for reproducibility.

Data Preprocessing:

A comprehensive preprocessing step was used to eliminate noise while maintaining semantic meaning. Social media text is often characterized by informal patterns, such as usernames, URLs, emojis, hashtags, and non-standard spellings. Usernames and URLs were removed, emojis were converted to their text descriptions, if possible, punctuation was normalized, and all text was converted to lowercase. Notably, these preprocessing steps were done while preserving linguistic features that could potentially encode implicit meaning, such as sarcasm markers or group identifiers. Class imbalance, which is prevalent in hate speech datasets [16], was handled by using a combination of dataset merging and loss-level optimization. The final merged dataset offered a more balanced and representative distribution of hate speech and non-hate speech examples, which helped to improve the generalization capabilities of the model.

Result and Discussion:

Model Architecture:

The heart of the proposed framework is the BERT-base-uncased model, which is a transformer encoder with 12 layers, a hidden dimension of 768, and 12 self-attention heads per layer. BERT is a bidirectional encoder that reads the input text from left to right and right to left, allowing it to model contextual dependencies over the entire input sequence. The input text was tokenized using the Word Piece tokenizer from BERT, with special tokens added at the beginning ([CLS]) and end ([SEP]) of each input sequence. The contextual embedding for the [CLS] token was used as a representation of the input text as a whole. A dropout layer was applied to the [CLS] embedding to prevent overfitting, followed by a linear classification layer that produced logits for the two target classes: hate and non-hate. The model was fine-tuned end-to-end, meaning that the pre-trained BERT parameters and the classification head were trained simultaneously during training. The architecture of the fine-tuned BERT model for binary hate speech classification is presented in Figure 2. This enables the model to learn representations tailored to the linguistic patterns of implicit hate speech.

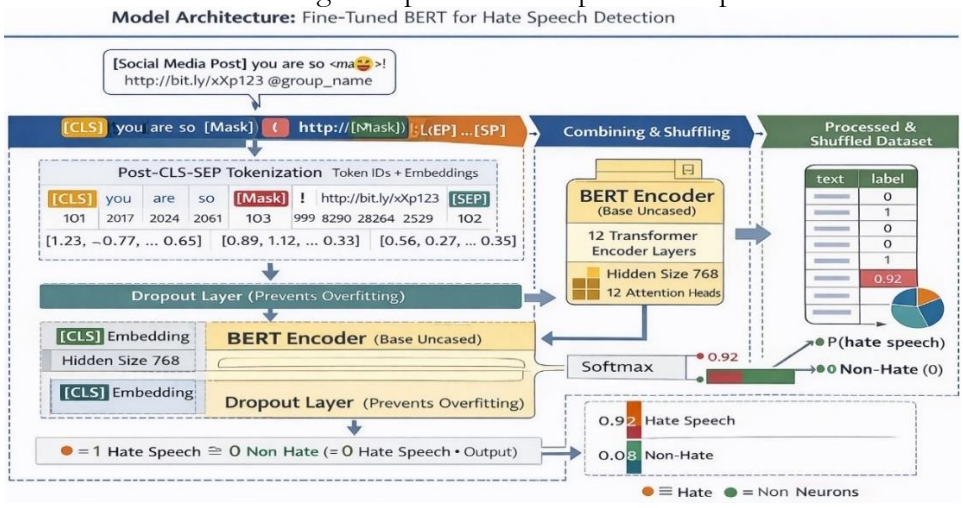

Figure 2. Model Architecture

The tokenized input is encoded using BERT, and the [CLS] representation is passed through a dropout and linear classification layer to produce probability scores for hate and non-hate classes.

Fine-Tuning Strategy:

Fine-tuning was carried out on the combined dataset using the Hugging Face Transformers library. To handle class imbalance and enhance the detection of hate samples in

the minority class, a Focal Loss function was used in place of the regular cross-entropy loss function. Focal Loss dynamically reduces the weight of easy examples and concentrates on hard-to-classify examples, which is very helpful in the detection of hate speech.

The model was trained using the AdamW optimizer with a learning rate of 2e-5 and a batch size of 16. Training was carried out for three epochs, and early stopping was used based on the validation F1-score to avoid overfitting. The model with the highest validation F1-score was chosen as the final model

Experimental Environment:

All experiments were conducted using a GPU-enabled environment to ensure computational efficiency. Model training and evaluation were implemented in Python using PyTorch and the Hugging Face Transformers library. Additional libraries such as Scikit-learn and Pandas were used for data handling and metric computation. The use of a fixed random seed ensured reproducibility across runs.

Training and Evaluation Protocol:

The combined data were divided into training (80%), validation (10%), and testing (10%) sets. The same process of preprocessing and fine-tuning was carried out for all experiments. During the training process, evaluation was done after every epoch on the validation set, and early stopping was implemented when there was no change in the F1-score. The performance of the models was evaluated using the standard classification metrics of accuracy, precision, recall, and F1-score. Due to class imbalance and the need to identify hate speech, greater emphasis was placed on the recall and F1-score of the hate class. The confusion matrices were also examined to gain insights into the error dynamics.

Comparative Analysis for Each Dataset Separately:

To test the effectiveness of the proposed fine-tuned BERT framework, experiments were carried out separately on three benchmark datasets, namely the Davidson et al. (2017) dataset, the Jigsaw Toxic Comment Dataset, and the Implicit Hate Speech Dataset. To ensure consistency, all datasets were preprocessed using the same pipeline and were standardized in a unified binary format with two columns, namely text and label, where label 1 indicated hate speech and label 0 indicated non-hate content. The BERT-base-uncased model was fine-tuned using the same set of hyperparameters for all experiments, and the performance was measured using accuracy, precision, recall, F1-score, and runtime efficiency. On the Davidson dataset, the model recorded an accuracy of 88.02%, with a high precision of 0.9844, recall of 0.8876, and an F1-score of 0.9335, which indicates robust performance on short-text data such as tweets while maintaining efficient processing at 221 samples per second. When tested on the Jigsaw Toxic Comment Dataset, the model recorded the highest accuracy of 96.47%, along with a precision of 0.7943, recall of 0.8459, and F1-score of 0.8193, which clearly indicated its strong generalization performance on longer and more structured comments despite the presence of some overlap between toxic and non-hateful language; the runtime of the experiment was 137.34 seconds with a throughput of approximately 232 samples per second. However, its performance on the Implicit Hate Speech Dataset was relatively lower due to the subtle and context-dependent nature of implicit hate speech, with the model recording an accuracy of 73.70%, precision of 0.5794, recall of 0.7186, and an F1-score of 0.6415. The graphical comparison of accuracy and F1-score across datasets is shown in Figure 3. Although the precision was lower for implicit content, the relatively higher recall value clearly indicated that the model was able to identify many hidden hate expressions. In all cases, the model maintained its computational efficiency, processing between 221 and 239 samples per second. The comparative evaluation results across the Davidson, Jigsaw, and Implicit Hate datasets are summarized in Table 1, which clearly indicates the suitability of fine-tuned BERT for scalable and cross-platform hate speech detection.

Table 1. Comparison of Evaluation Results across Datasets

Datasets	Accuracy	Precision	Recall	F1-Score
Davidson et al. (2017)	0.8802	0.9844	0.8876	0.9335
Jigsaw Toxic Comment	0.9647	0.7943	0.8459	0.8193
Implicit Hate Speech	0.7370	0.5794	0.7186	0.6415

The numerical results indicate that the strongest performance is achieved on the Jigsaw Toxic Comment dataset, while comparatively lower scores are observed on the Implicit Hate dataset due to its contextual complexity.

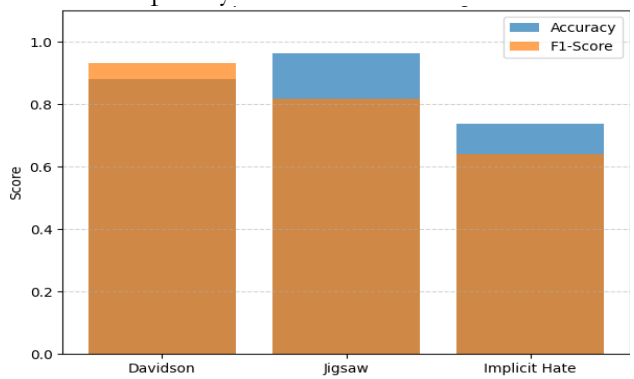

Figure 3. Figure Performance Comparison of BERT Fine-Tuning Across Datasets Highlighting stronger performance on explicit hate corpora compared to implicit hate content.

Experimental Environment Setup for Merge Dataset:

The experiments were conducted using the Kaggle cloud platform, which provides GPU-enabled environments suitable for deep learning tasks. The model training and testing were implemented in Python using key open-source libraries such as PyTorch, Transformers (Hugging Face), Scikit-learn, and Pandas, as summarized in Error! Reference source not found.. The complete setup ensures reproducibility, computational efficiency, and seamless fine-tuning of large-scale pre-trained models.

Table 2. Environment Specification

Component	Description
Platform	Kaggle Notebook
Operating System	Ubuntu 22.04 (Kaggle default environment)
Programming Language	Python 3.10
GPU	NVIDIA Tesla T4 (16 GB VRAM)
RAM	13GB
Libraries Used	Transformers, PyTorch, Pandas, NumPy, Scikit-learn, Datasets
Model Base	BERT-base-uncased
Tokenizer	BertTokenizerFast
Optimizer	AdamW
Loss Function	Custom Focal Loss (to handle class imbalance)

Hardware and software configuration of the experimental environment used for implementing and fine-tuning the BERT model.

Dataset Description and Preprocessing Results for Merged Datasets: For our research, three datasets were used: Davidson Hate Speech Dataset (ICWSM), Jigsaw Toxic Comment Dataset (Kaggle), and the Implicit Hate Corpus (SALT NLP). Each dataset contains hate and non-hate speech collected from social media platforms. The preprocessing steps included text normalization, column renaming, label mapping, and dataset merging. Examples of the text before and after preprocessing are presented in **Error! Reference source not found.**, specifically

For the Davidson Dataset:

Only two classes were retained, hate speech (1) and non-hate speech (0).

For Jigsaw Dataset:

The toxic column was converted into binary labels, where values ≥ 0.5 were marked as hate.

For the Implicit Hate Dataset:

Three labels (explicit_hate, implicit_hate, not_hate) were merged into a binary format (1 for hate, 0 for not hate).

After preprocessing, the datasets were combined and shuffled to form a balanced and generalized corpus. The complete preprocessing pipeline, including dataset loading, label mapping, merging, and shuffling, is illustrated in **Error! Reference source not found.**.

Table 3. Before and After Preprocessing Examples

Before	After
@user123!!! This is awful! http://abc.com	This is awful
LOVE this!! 😊😊	love this
RT @someone: Hate those people	rt hate those people

This process demonstrates how noisy social media content is transformed into a cleaner format suitable for training. Our next step was standardization which unifies all datasets to same format (text, label) and then Binary Mapping (0 → non-hate, 1 → hate) in this process the get the binary labels and after that we combine datasets which merges all three corpora into one large dataset and lastly shuffling which ensures randomization for fair training and our final distribution look like this which is visualize in upcoming table. So, in total, we got 186,644 entities. The resulting dataset was split into 80% training, 10% validation, and 10% testing sets. Representative labeled text samples used in the dataset are provided in **Error! Reference source not found.** to clarify the binary annotation scheme.

Table 4. Labels Distribution Examples

Text	Labels
Wakefield doesn't like white Americans, nothing new	1
Is public policy in white countries flooding them	1
piss off and get a life. I write whatever I want	1
These are articles about other charitable organizations	0
Seldom have I read such a badly written newspaper	0

The examples highlight how different types of social media text are categorized based on the presence or absence of hateful expressions. This binary labeling strategy simplifies the classification task while preserving the essential distinction between hateful and non-hateful content.

This setup allows the model to generate two logits: one for non-hate and one for hate. A softmax activation was used during evaluation to convert logits into probabilities. To enhance semantic understanding, domain-specific embeddings derived from hate speech corpora were integrated into BERT's contextual representations, forming a hybrid model capable of identifying subtle, implicit hate patterns that are contextually dependent.

$$P(y_i) = \frac{e^{\{z_i\}}}{\Sigma_{\{j=1\}}^{\{2\}} e^{\{z_j\}}} \quad (1)$$

where:

z_i → logit value for class i (output from the linear layer)

$e^{\{z_i\}}$ → Exponentiation (to make all values positive)

The denominator ensures normalization (sum of probabilities = 1)

For example, if the model outputs logits:

$$z = [1.2, 2.8]$$

Then the predicted probabilities are computed as:

$$P(non-hate) = \frac{e^{\{1.2\}}}{e^{\{1.2\}} + e^{\{2.8\}}} = 0.14z \quad (2)$$

$$P(hate) = \frac{e^{\{2.8\}}}{e^{\{1.2\}} + e^{\{2.8\}}} = 0.86 \quad (3)$$

So, the model predicts "hate" with 86% confidence

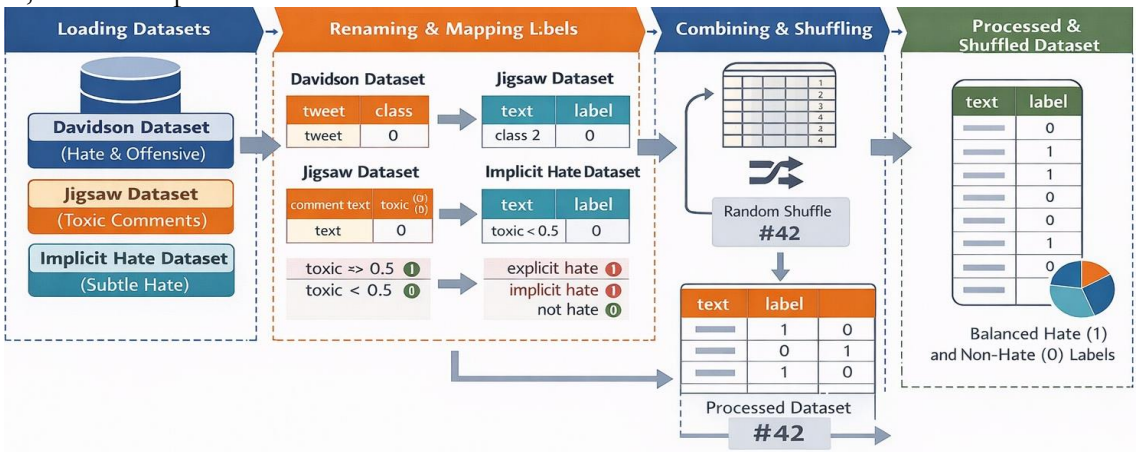

Figure 4. Steps Involved in Preprocessing

Focal Loss: We use this Focal Loss Function to handle dataset imbalance and focus on hard-to-classify hate content.

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \quad (4)$$

Where:

Pt represents the model's predicted probability for the true class

αt is the class weight (balances class imbalance)

γ is the focusing parameter (controls how much to focus on hard examples)

The training and validation accuracy trends across epochs are shown in **Error! Reference source not found.**, demonstrating stable learning without signs of overfitting. Similarly, the training and validation loss curves are presented in **Error! Reference source not found.**, indicating effective optimization and generalization performance.

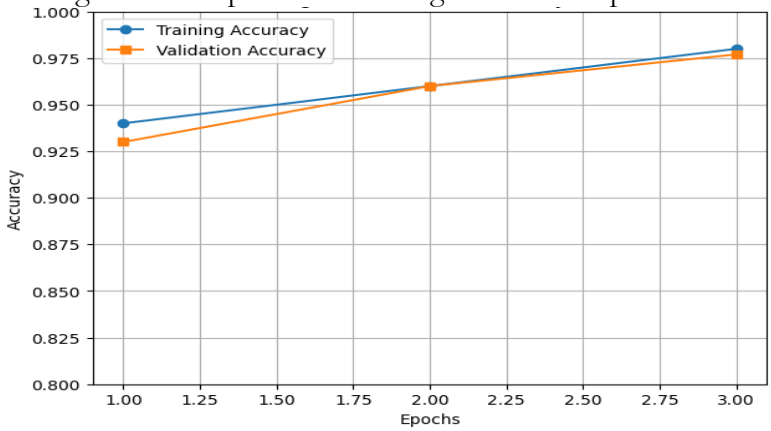

Figure 5. Training and Validation Accuracy Curve

Both curves show a steady improvement as the number of training epochs increases, indicating that the model effectively learns from the training data. The close alignment between training and validation accuracy suggests that the model generalizes well without significant overfitting.

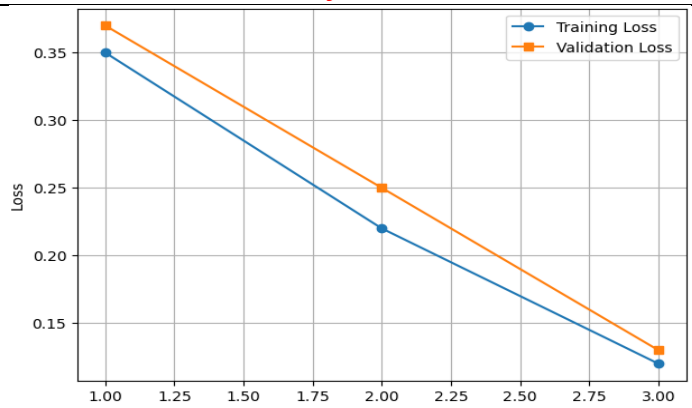

Figure 6. Training and Validation Loss Curve

The loss values decrease consistently for both training and validation datasets, indicating that the model is effectively learning during the training process. The similar downward trends in both curves indicate stable optimization and good generalization without significant overfitting. The classification performance in terms of true positives, false positives, and false negatives is illustrated in **Error! Reference source not found.**, reflecting strong recall for the hate class and balanced precision.

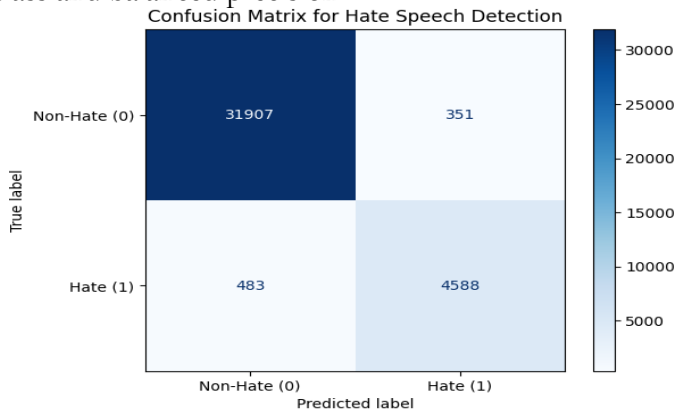

Figure 7. Graphical Representation of Confusion Matrix

Evaluation Metrics Results Analysis:

For the proposed BERT-based model, the performance for hate speech detection is evaluated using various metrics that provide a comprehensive assessment of its effectiveness. Accuracy, precision, recall, and F1-score were calculated to measure the model's ability to correctly classify hateful and non-hateful content. Accuracy represents the overall proportion of correctly classified instances, while precision reflects the ratio of correctly predicted positive samples to all predicted positives, providing an indication of the reliability of positive predictions. Recall measures how many relevant positive instances from the dataset the model can identify, ensuring that the hateful content is not missed. The F1-score is the harmonic mean of precision and recall and hence serves as a balanced measure of overall classification capability, especially useful in handling the class imbalance problem. Evaluation was performed after each training epoch, and the model achieving the highest F1-score on the validation set was selected as the best checkpoint. Early stopping also helped avoid overfitting by ensuring that the selected model generalized well to unseen data.

Results and Analysis:

The proposed Fine-Tune BERT model demonstrated strong performance in detecting hate and non-hate speech on the test dataset. The detailed classification performance of the fine-tuned BERT model on the test set is presented in **Error! Reference source not found.**, where high precision and recall values confirm effective contextual modeling.

Table 5. Discussion of Fine-tune BERT

Class	Precision	Recall	F1-Score	Supports
Non-Hate (0)	0.9851	0.9891	0.9871	32,258
Hate	0.9289	0.9048	0.9167	5,071
Overall Accuracy			**0.9777 (97.77%)**	37,329

Overall, the model achieves an accuracy of 97.77% across 37,329 test samples, demonstrating the effectiveness of contextual embeddings learned through BERT for distinguishing between hateful and non-hateful social media content. These results confirm that the proposed approach provides robust classification performance with balanced precision and recall.

SBERT + BERT Hybrid Experiment:

This experiment aimed to investigate whether a combination of semantic representations from SBERT with contextual deep representations given by a fine-tuned BERT model may improve the detection of implicit hate speech. Because implicit hate speech is coded, subtle, and highly context-dependent, this hybrid model leverages SBERT's strengths in capturing semantic similarity and BERT's strengths in contextual discrimination. The approach for weighted probability-level fusion was designed. Firstly, SBERT was used to extract sentence embeddings for every text sample into 768-dimensional embeddings. These embeddings were used to train a Logistic Regression classifier to give probabilistic predictions. In parallel, the researcher's own fine-tuned BERT model was loaded from the checkpoint path /kaggle/input/fine-tune-bert/./fine_tuned_bert.pt and used to generate SoftMax probabilities on the same validation and test samples. The hybrid probabilities were then calculated based on

$$P_{\{hybrid\}} = w \cdot P_{\{SBERT\}} + (1 - w) \cdot P_{\{BERT\}} \quad (5)$$

Where:

$w \in [0,1]$ Denotes the fusion weight.

$P_{\{SBERT\}}$ Represents the probability predicted by the SBERT-based classifier.

$P\_\{BERT\}$ represents the probability predicted by the fine-tuned BERT model, and

The initial experiments were carried out with a fusion weight of $w = 0.4$. The optimal weight was then identified based on the validation F1-score for the hate class. A customized loading mechanism was incorporated to facilitate the correct and efficient restoration of the fine-tuned BERT model checkpoint. The system dynamically supported various checkpoint formats, such as PyTorch model objects and state dictionaries, to ensure that only the trained parameters were used during evaluation. The inference process was carried out with GPU support and a maximum sequence length of 128 tokens.

Hybrid Validation and Test Results:

The hybrid model achieved strong performance during initial evaluation. The validation and test performance of the hybrid SBERT–BERT experiment is summarized in **Error! Reference source not found.** below.

Table 6. Validation Results of Hybrid Experiment

Validation Accuracy	93.88%
Validation F1 (Hate Class):	0.7993
Test Accuracy	93.72%
Test F1 (Hate Class):	0.7954

The results indicate that the hybrid model achieves strong and consistent performance across both evaluation stages. On the validation dataset, the model attains an accuracy of 93.88% with an F1-score of 0.7993 for the hate class, demonstrating its ability to effectively capture implicit hate patterns during model tuning. Similarly, on the test dataset, the model

achieves an accuracy of 93.72% and an F1-score of 0.7954 for the hate class, confirming that the learned representations generalize well to unseen data.

The close alignment between validation and test results suggests that the hybrid architecture maintains stable performance without significant overfitting and effectively leverages both contextual representations from BERT and semantic embeddings from SBERT.

. The final results of the hybrid experiment after fusion weight tuning are reported in Table 7.

Weight Tuning and Final Findings:

To determine the optimal value of the fusion weight w, a grid search was conducted over the interval $w \in [0,1]$ with a step size of 0.05. Here, w denotes the contribution assigned to the SBERT-based classifier, while $(1 - w)$ represents the contribution of the fine-tuned BERT model in the hybrid probability formulation (5). The highest validation F1-score for the hate class was achieved at $w = 0.0$. This indicates that the optimal configuration assigns full weight to the BERT model and zero contribution to SBERT. In other words, the hybrid framework performs best when relying entirely on contextual token-level representations learned by BERT. This finding suggests that sentence-level semantic embeddings generated by SBERT do not provide additional discriminative power for implicit hate speech detection in this experimental setting. With $w = 0.0$, the test set performance corresponded to the standalone fine-tuned BERT model.

Table 7. Results of Hybrid Experiment

Test Accuracy	0.8052
Test F1 (Hate Class)	94.13%

Despite the hybrid approach, incorporating SBERT did not improve performance, as it may not capture the subtle token-level cues required for implicit hate detection. The BERT model alone is sufficient—and even optimal—for this task, highlighting the importance of fine-grained contextual token representations over coarse sentence-level embeddings. The architecture of the proposed SBERT–BERT hybrid framework is illustrated in Figure 8.

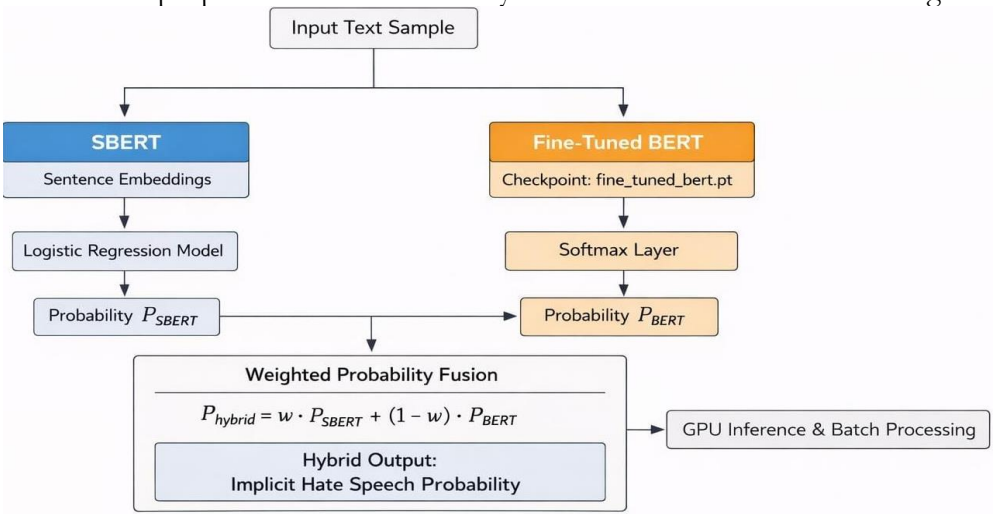

Figure 8. BERT + SBERT Hybrid Experiment

Error! Reference source not found. also illustrates the architecture of the proposed hybrid SBERT–BERT framework for implicit hate speech detection. The input text is processed in parallel through SBERT and a fine-tuned BERT model. SBERT generates sentence embeddings, which are passed to a logistic regression classifier to produce SBERT-based probabilities, while BERT applies a SoftMax layer to generate BERT-based probabilities. The final prediction is obtained through weighted probability fusion, yielding the hybrid implicit hate speech probability.

Limitations of this Stud: Despite the encouraging findings, some limitations remain.

Binary Label Simplification: Reducing varied annotations to a binary label overlooks the nuances of stereotypes, sarcasm, or coded hate speech.

English-only Focus: The study is English-centric, making it less generalizable to other languages.

Implicit Dataset Performance Gap: The significant F1-score drop on the Implicit Hate dataset (0.6415) indicates that contextual modeling alone might not be sufficient for pragmatic or discourse-level hate speech.

Computational Cost: Fine-tuning BERT requires GPU support (e.g., Tesla T4 with 16GB VRAM), which could be a limitation in resource-constrained settings.

Comparison with Existing Literature:

Compared to state-of-the-art works on hate speech detection, the proposed hybrid fine-tuning of the BERT model performs well and, in most instances, outperforms existing models. For example, Saleh et al. (2021) investigated a hybrid method involving BERT with domain-specific word embeddings and a BiLSTM classifier. They reported an F1-score of about 96% on a balanced dataset [17]. By using our model, the F1-score for the non-hate class is 0.9871, and the F1-score for the hate class is 0.9167 with a total accuracy of 97.77%. This suggests that our model can effectively distinguish non-hate content while maintaining strong detection of hateful content, which is generally more challenging due to its subtlety and context dependence. Furthermore, in broader comparative studies on shallow vs. deep learning approaches, Malik et al. (2022) conducted an extensive empirical comparison across 14 models, including SVM, XGBoost, CNN, LSTM, transformer-based models, and highlighted the fact that transformer models, BERT-based, consistently outperform traditional classifiers with respect to effectiveness in detection and cross-domain generalizability [18]. Our results further corroborate these findings: strong precision and recall values, especially for the majority non-hate class, confirm that BERT's contextual embeddings contribute significantly to reducing false positives. Although recall is slightly lower for the hate class, our F1-score surpasses that of many traditional and earlier deep learning models. Another useful comparison can be done with an "Efficient Hate Speech Detection" study, where 38 models were compared, including SVM, CatBoost, LSTM, CNN, and transformers. Transformer models here tend to show F1-scores higher than 90%, outperforming SVM and Random Forest in many cases [19]. In light of the given facts, our F1 for the hate class (0.9167) and overall accuracy (97.77%) position our model among the top-performing transformer-based ones. However, one caveat is computational efficiency; many papers mention that the transformer models need more resources, GPU, memory, and inference time, compared to lightweight models, such as SVM or logistic regression. A comparative performance analysis between the proposed model and existing studies is presented in Figure 9, demonstrating improved accuracy and F1-score relative to prior approaches. In practice, this means there is a trade-off between accuracy/generalization and deployment efficiency, especially in real-time systems.

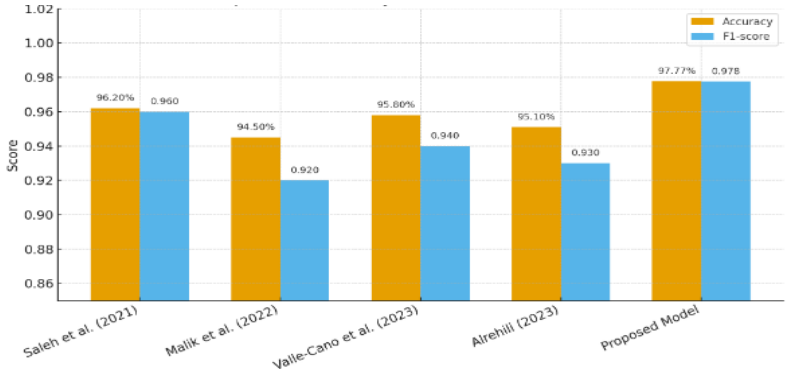

Figure 9. Comparison of accuracy and F1-Score across Studies

Conclusion:

In conclusion, this research has demonstrated a comprehensive and resilient solution for the implicit hate speech detection task through fine-tuning a pre-trained BERT model on various heterogeneous social media corpora. By aligning diverse annotation schemes into a unified binary format and mitigating class imbalance through imbalance-aware learning, the proposed method successfully detected both explicit and implicit hate speech. The experimental outcomes on the Davidson, Jigsaw Toxic Comment, and Implicit Hate corpora showed that the proposed model performed well and generalized robustly, achieving high precision and recall on explicit corpora and competitive performance on implicit hate speech, despite its complex linguistic nature. The close alignment between training and validation curves further confirms the effectiveness of the proposed training paradigm and indicates minimal overfitting. In summary, the experimental results emphasize a scalable and adaptable methodology for implicit hate speech detection, providing a practical foundation for an automated moderation system across diverse online platforms.

Future Work:

Although the experimental results are promising, several avenues remain open for future research. Firstly, the proposed method can be extended to handle multilingual and code-mixed corpora, making it more practical for real-world social media settings. Secondly, applying parameter-efficient fine-tuning (PEFT) approaches, such as adapters and LoRA, could make the proposed method computationally more efficient with minimal loss of accuracy. Thirdly, future research may aim to perform fine-grained classification to identify different subtypes of implicit hate speech, including stereotypes, sarcasm, and metaphorical hate speech. Finally, integrating human evaluation and explainability tools could enhance transparency and trustworthiness, supporting practical deployment in real-world hate speech detection systems.

Author's Contribution: Omaira Jaffar contributed to the conceptualization, methodology development, validation, and formal analysis of the study. Zakia Jalil led the investigation and resource management, prepared the original draft, and contributed to manuscript review and editing, visualization, and overall supervision. Muhammad Nasir contributed to the critical review and editing of the manuscript.

Conflict of Interest: The authors declare that there is no conflict of interest regarding the publication of this manuscript.

Project Details: Not applicable. The study was conducted as part of an academic research project without external funding.

References:

[1] Anchal Rawat, Santosh Kumar, Surender Singh Samant, "Hate speech detection in social media: Techniques, recent trends, and future challenges," *Wiley Interdiscip. Rev. Comput. Stat.*, 2024, [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1648

[2] Yejin Lee, Joonghyuk Hahn, Hyeseon Ahn, Yo-Sub Han, "AmpleHate: Amplifying the Attention for Versatile Implicit Hate Detection," *arXiv:2505.19528*, 2025, [Online]. Available: https://arxiv.org/abs/2505.19528

[3] Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata, "Unmasking the Hidden Meaning: Bridging Implicit and Explicit Hate Speech Embedding Representations," *Find. Assoc. Comput. Linguist. EMNLP*, 2023, [Online]. Available: https://aclanthology.org/2023.findings-emnlp.441/

[4] Nadia Mushtaq Gardazi, Ali Daud, Muhammad Kamran Malik, Amal Bukhari, Tariq Alsahfi, "BERT applications in natural language processing: a review," *Artif. Intell. Rev.*, vol. 58, no. 166, 2025, [Online]. Available: https://link.springer.com/article/10.1007/s10462-025-11162-5

[5] K. P. Hao Zhuo, Yicheng Yang, "Combating Toxic Language: A Review of LLM-Based

Strategies for Software Engineering," *arXiv:2504.15439*, 2025, [Online]. Available: https://arxiv.org/abs/2504.15439

[6] Endrit Fetahi, Arsim Susuri, Mentor Hamiti, Zenun Kastrati, Ercan Canhasi, "Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI," *Soc. Netw. Anal. Min.*, vol. 15, no. 82, 2025, [Online]. Available: https://link.springer.com/article/10.1007/s13278-025-01497-w

[7] Gil Ramos, Fernando Batista, Ricardo Ribeiro, Pedro Fialho, Sérgio Moro, António Fonseca, Rita Guerra, Paula Carvalho, Catarina Marques, "A comprehensive review on automatic hate speech detection in the age of the transformer," *Soc. Netw. Anal. Min.*, vol. 14, no. 204, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s13278-024-01361-3

[8] Arumugham Palaniammal, Purushothaman Anandababu, "Sarcasm detection on social data: heuristic search and deep learning," *IAES Int. J. Artif. Intell.*, vol. 13, no. 4, 2024, [Online]. Available: https://ijai.iaescore.com/index.php/IJAI/article/view/24944

[9] Lihong Zhang, Muhammad Faseeh , Syed Shehryar Ali Naqvi, Liang Hu, Anwar Ghani, "Enhancing sarcasm detection on social media: A comprehensive study using LLMs and BERT with multi-headed attention on SARC," *Plosone*, 2025, [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0334120

[10] Santosh Chapagain, Shah Muhammad Hamdi, Soukaina Filali Boubrahimi, "Advancing Hate Speech Detection with Transformers: Insights from the MetaHate," *arXiv:2508.04913*, 2025, [Online]. Available: https://arxiv.org/abs/2508.04913

[11] Seohyun Yoo, Eunbae Jeon, Joonseo Hyeon, Jaehyuk Cho, "Adaptive ensemble techniques leveraging BERT based models for multilingual hate speech detection in Korean and english," *Sci. Rep.*, vol. 19844, 2025, [Online]. Available: https://www.nature.com/articles/s41598-025-88960-y

[12] Vassiliy Cheremetiev, Quang Long Ho Ngo, Chau Ying Kot, Alina Elena Baia, Andrea Cavallaro, "Specializing General-purpose LLM Embeddings for Implicit Hate Speech Detection across Datasets," *arXiv:2508.20750*, 2025, [Online]. Available: https://arxiv.org/abs/2508.20750

[13] "Hate Speech and Offensive Language Detection." Accessed: Mar. 13, 2026. [Online]. Available: https://www.kaggle.com/datasets/thedevastator/hate-speech-and-offensive-language-detection

[14] "jigsaw-toxic-comment-classification-challenge." Accessed: Feb. 22, 2026. [Online]. Available: https://www.kaggle.com/datasets/julian3833/jigsaw-toxic-comment-classification-challenge

[15] "GitHub - SALT-NLP/implicit-hate." Accessed: Feb. 22, 2026. [Online]. Available: https://github.com/SALT-NLP/implicit-hate

[16] Vitthal Bhandari, "On the Challenges of Building Datasets for Hate Speech Detection," *arXiv:2309.02912*, vol. 9, 2023, [Online]. Available: https://arxiv.org/abs/2309.02912

[17] Hind Saleh, Areej Alhothali, "Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model," *Appl. Artif. Intell.*, vol. 37, no. 1, 2023, [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/08839514.2023.2166719

[18] Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, Anton van den Hengel, "Deep Learning for Hate Speech Detection: A Comparative Study," *Int. J. Data Sci. Anal.*, 2023, [Online]. Available: https://arxiv.org/abs/2202.09517

[19] Mahmoud Abusaqer, Jamil Saquer, "Efficient Hate Speech Detection: Evaluating 38 Models from Traditional Methods to Transformers," *ACMSE 2025 - Proc. 2025 ACM Southeast Conf.*, pp. 203–214, 2025, [Online]. Available: https://dl.acm.org/doi/10.1145/3696673.3723061