A Hybrid Deep Learning Model with Gated Representation for the Detection of Offensive Text in Noisy Social Media Environments

Praneetha Garagadakuppe Nanjundappa
Department of Computer Science and Engineering, Sapthagiri College of Engineering, Visvesvaraya Technological University, Belagavi, India | Department of Computer Science and Engineering, M S Ramaiah University of Applied Sciences, Bangalore, India
praneethaguddi@gmail.com (corresponding author)

Kamalakshi Naganna
Department of Computer Science and Engineering, Sapthagiri College of Engineering, Visvesvaraya Technological University, Belagavi, India
kamalnags@gmail.com

Received: 12 August 2025 | Revised: 17 September 2025 | Accepted: 23 September 2025

 DOI: https://doi.org/10.48084/etasr.14025

ABSTRACT

The increasing prevalence of offensive and abusive language on social networks represents a strong threat to online safety and user welfare. Existing methods, frequently constrained by the argumentative and heterogeneous characteristics of social media text, are unable to effectively deal with informalities and contextual dependencies. To address these challenges, this study proposes a novel hybrid deep learning framework that combines BERT-based contextual encoding with a novel Subword Pattern Recognizer (SPR) for extracting character-level morphological features. This study uses a gated Multi-Layer Perceptron (MLP) fusion mechanism to retain the contribution from both character-level and semantic features, enabling robust detection of objectionable content even in ambiguous or manipulated inputs. The experimental results demonstrated the effectiveness of the proposed model by achieving 91% and 98 % accuracy on the OLID and Davidson datasets. Furthermore, a comprehensive ablation study validates the complementary strengths of the dual-branch architecture and fusion mechanism to mitigate weaknesses in noisy, informal, and ambiguous offensive language.

Keywords-offensive language; hybrid deep learning; BERT; character-level features; gated MLP fusion

I. INTRODUCTION

In today's interconnected world, social media platforms such as Twitter, Facebook, Instagram, and YouTube are widely trusted and used because they are places for dialogue, autonomy, and community [1]. Individuals often use a variety of informal, formulaic, or cryptic forms of writing to express themselves. These forms of writing may include hate speech, cyberbullying, derogatory language, personal attacks, threats, and veiled slander or profanity. These expressions often target individuals or groups for various reasons, such as race, gender, religion, sexual orientation, nationality, or different political beliefs. However, while offensive language may not be considered hate speech in legal or social terms, it can cause harm, reduce the quality of conversation, and weaken the inclusiveness of online communities [2]. Offensive content can exist both in explicit and implicit manners. An explicit statement includes direct slurs or foul language, while implicit statements can include sarcasm, euphemisms, obfuscations, or creative spellings to circumvent moderation systems [3]. Previous studies have shown that exposure to this kind of language can activate pain centers in the brain and manifest psychological harm, and constant online abuse can lead to anxiety, depression, loss of reputation, or self-harm. At a societal level, unchecked offensive discourse risks normalizing harmful behavior, fueling polarization, and even contributing to real-world violence [4].

A broad range of approaches for automatic identification of offensive language have been investigated in the literature. Most previous works have used traditional Machine Learning (ML) models, such as Support Vector Machines (SVM), Random Forests (RF), and Artificial Neural Networks (ANN), using handcrafted lexical, syntactic, and semantic features [5]. Researchers have combined SVM, RF, and ANN to identify offensive language, rank its severity, and the offenders' targets in the OLID dataset. Other studies have used eXplainable AI (XAI) for interpretability [6], combining multiple embeddings [7] and word2vec-based representations in binary and multi-class classification [8].

In recent years, the popularity of Deep Learning (DL) models has led researchers to focus on using Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformers, such as the BERT model, which are effective in automating feature extraction and learning richer contextual semantics [9, 10]. Hybrid approaches based on the combination of transformers and traditional ML classifiers have shown effectiveness in multiple languages, including Arabic [11, 12], Chinese [13], and Hindi/Marathi [14]. The multilingual and code-mixed setting has been addressed in shared tasks such as SemEval OffensEval [15], with systems utilizing multilingual variants of BERT. Recent advances include dataset augmentation through back-translation [16], multi-task learning for abuse detection and emotion recognition [17], and improvements in interpretability using Large Language Models (LLMs) [18].

Despite the numerous computational schemes presented by researchers, there are still significant challenges that limit the practical applicability of existing models. Ambiguity and context dependence remain a major challenge, as offensive language is highly context-sensitive and often contains cultural references, community-specific slang, or sarcasm, which can be misinterpreted even by human interpreters. Many benchmark datasets suffer from severe class imbalance. Not much work has been done to effectively handle class imbalance and pre-processing problems to achieve robust detection of offensive language without compromising linguistic features. In addition, feature representation deficiencies exist in many studies, and existing prediction models are often based on hand-crafted features and static embeddings that struggle when the same word can have multiple meanings depending on the context.

To address these research challenges, this study presents three important contributions. First, a round-trip translation strategy is adopted for class-specific data augmentation to mitigate class imbalance. Second, a multi-modal feature extractor is developed based on a joint approach of BERT for sentence-level contextual semantic modeling with a novel parallel Attentive-CNN designed to capture character-level morphological and orthographic patterns. Third, Gated MLP Gusion (GMF) is introduced, using an adaptive attention fusion strategy to dynamically balance and integrate features from both modalities for robust classification of offensive language.

II. MATERIALS AND METHODS

A. Dataset

The proposed model was developed and evaluated considering two benchmark offensive-language datasets. The first dataset considered is the Davidson Dataset (DV) [19], which is widely utilized for detecting both offensive and hateful speech, consisting of 24,783 English tweets. The second dataset adopted is the Offensive Language Identification Dataset (OLID) [20], which was introduced for the SemEval-2019 Task 6 competition, consisting of three subtasks. This study only focused on sub-task A, identifying offensive language on Twitter. Table I highlights the data distribution of both the DV and OLID datasets.

TABLE I. DATASET STATISTICS

Dataset	Total samples	Classes
DV	24,783	Hate (5.77%), Offensive (77.43%), Neither (16.80%)
OLID	14,100	OFF (33.23%), NOT (66.77%)

Both datasets, DV and OLID, are imbalanced, which is a common problem since real-world datasets often exhibit unequal distribution of classes, especially those related to ecological and social phenomena such as hateful and offensive speech detection. However, the challenge is to ensure efficient training of the model on an imbalanced dataset without making its predictive decisions biased toward the majority classes.

B. Preprocessing

This study aimed at detecting offensive speech in social media texts, which are considered highly unstructured, noisy, and informal, as many posts often contain emojis, excessive punctuation, hashtags, slang, repetitive characters, etc. Therefore, a comprehensive offensive language-specific preprocessing operation was performed on the data before model training.

The proposed preprocessing algorithm is adaptive to both DV and OLID datasets, preserving critical linguistic and contextual attributes that are important in detecting toxic content. The algorithm retains some punctuation, such as exclamation marks (!) and quotation marks ("..."), which often signal emphasis, sarcasm, or aggression. Apart from this, digits and rare characters are also preserved to handle stylized profanity, for example, b1tch, f4gg0t, commonly used to bypass moderation filters. The algorithm considers emoji normalization to convert expressive icons into textual tokens; for example, '😡' is transformed to its angry face meaning, which preserves the user's emotional and offensive context. Moreover, the proposed algorithm avoids considering non-semantic elements such as usernames, URLs, and retweet markers to reduce noise in the input text. In addition, repetition normalization is performed to ensure that lengthened words such as noooo and soooo are consistently handled.

C. Text-Data Augmentation

For model training and evaluation, each preprocessed dataset was split into training (80%), validation (10%), and testing (10%) subsets using a stratified splitting strategy to preserve the class distribution. This study applied data augmentation only to the training set so that the validation and testing sets remain unchanged to provide a fair and unbiased evaluation. The proposed data augmentation technique adopts a Round-Trip Translation (RTT) strategy [21] to artificially generate similar sentences of the existing samples from minority classes. It uses machine translation to translate whole sentences into and back from an intermediate language (e.g., English→French→English) to obtain a dataset with paraphrastic diversity while preserving semantic labels. An example of the RTT process is:

Original: *F@@k that b!**h a** dude, he's fake af*

After RTT: *F@#k that idiot, he's a liar.*

In this example, RTT introduces syntactic and lexical changes while preserving the offensive intent. The surface form of the sentence may change, but the core intent and category labels remain unchanged. The advantage of RTT is that it can normalize or explain misspellings or exaggerations such as "f@@@@ck" and "b!!!!t@h", as well as slang abbreviations commonly found in social media posts, such as "af". In addition, by changing sentence structure and vocabulary, RTT introduces new expressions, which can help the learning model generalize beyond memorized phrases and learn deeper contextual patterns.

D. Proposed Hybrid Learning Model

This study introduces a predictive approach by developing a hybrid DL architecture with gated representation that integrates both word-level semantic context and character-level sub-word patterns for robust offensive text classification. The proposed hybrid learning system consists of three core modules, discussed below.

1) Contextual Feature Modeler

This module adopts the BERT model, which is based on the multi-level bidirectional transformer encoder architecture [22]. This model takes the cleaned input text sentences T_{raw} and tokenizes it into word-level units or tokens $T \in \{t_1, t_2, \dots, t_n\}$, and then assigns a unique integer ID based on the predefined vocabulary V on which the BERT is trained. The obtained T token sequences are then subjected to embedding layers, and the embedded sequence E is generated. These embedding sequences then serve as inputs for the encoder blocks to learn rich semantic features and extract sentence-relevant attributes by analyzing the dependencies between words from the input text. The encoder consists of a stack of $N = 12$ identical layers, and each of these layers has two sub-layers. The first one is a Multi-Head Self-Attention (MHSA) mechanism [23] for capturing contextual dependencies between tokens, enabling the model to consider relationships between all words in parallel by attending to every other word in the sentence using an attention score computed as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q denotes the query, K is the key, and V are value projections derived from E. Here, multiple heads are computed in parallel, learning to focus on different linguistic patterns such as co-reference, sarcasm, or negation. Their outputs are then concatenated and linearly transformed. The second module involves a feedforward network consisting of two linear transformations with a non-linear activation in between to process the output of the attention layer independently, followed by a layer normalization. The output of each sub-layer is computed as follows:

$$LayerNorm(x + Sublayer(x)) \quad (2)$$

where $Sublayer(x)$ is the function implemented by the sub-layer [24].

2) Subword Pattern Recognizer (SPR)

The SPR module is responsible for capturing fine-grained character-level features for detecting noisy, obfuscated, or morphologically altered offensive content common in informal text (for example, *f***k*, *daaaamn*). The SPR design is inspired by the Conformer model used for speech recognition [25]. However, the customization made in SPR is the inclusion of parallel CNN branches, each followed by an MHSA mechanism that allows the network to capture local patterns and model long-range dependencies along the input character sequence to learn sub-word features such as sarcasm, elongation, or word shape attributes that span multiple characters. Table II presents the layer-wise operations with shapes and hyperparameter values used in the proposed SPR.

TABLE II. LAYER-WISE DETAILS OF THE PROPOSED SUBWORD-ATTENTION MODEL

Layer	Values	Input	Output
Input	Char indices	[B, 300]	[B, 300]
Embedding	Vocab-size=300, embed-dim=50	[B, 300]	[B, 300, 50]
Transpose	permute(0, 2, 1)	[B, 300, 50]	[B, 50, 300]
Conv1D	kernel=3, filters=128	[B, 50, 300]	[B, 128, 298]
Conv1D	kernel=4, filters=128	[B, 50, 300]	[B, 128, 297]
Conv1D	kernel=5, filters=128	[B, 50, 300]	[B, 128, 296]
Activation	ELU()	[B, 128, L]	[B, 128, L]
Transpose	permute(0, 2, 1)	[B, 128, L]	[B, L, 128]
MHSA (each branch)	Embed-dim=128, heads=4	[B, L, 128]	[B, L, 128]
Transpose	permute(0, 2, 1)	[B, L, 128]	[B, 128, L]
Adaptive MaxPool1D	Output-size=1	[B, 128, L]	[B, 128, 1]
Squeeze	dim=2	[B, 128, 1]	[B, 128]
Concatenation	concat across branches	3 × [B, 128]	[B, 384]

The SPR module considers an input sentence $S = [w_1, w_2, \dots, w_N]$, where each word w_i is a sequence of characters with the length of the i word L_i:

$$w_i = \left[c_1^{(i)}, c_1^{(i)}, \dots, c_{L_i}^{(i)}\right] \quad (3)$$

A character embedding matrix is defined as $E_c \in \mathbb{R}^{|\mathcal{V}_c| \times d_c}$, which maps each character $c_j^{(i)} \in |\mathcal{V}_c$ to its vector:

$$e_j^{(i)} = Embed\left(c_j^{(i)}\right) \in \mathbb{R}^{d_c} \quad (4)$$

where $\mathcal{V}_c$ is the character vocabulary and d_c is the character embedding dimension. This process results in the generation of character-level embeddings $C^{(i)}$ for an input word:

$$C^{(i)} = \left[e_1^{(i)}, e_2^{(i)}, \cdots, e_{L_i}^{(i)}\right] \in \mathbb{R}^{L_i \times d_c} \quad (5)$$

The features are then extracted by implementing 1D convolutions with kernel k and filter f over the embeddings:

$$f_i = \sigma\left(W_k \times C_{i:i+k-1}^{(1)} + b\right) \quad (6)$$

where $W_h \in \mathbb{R}^{h \times d_c}$, $C_{i:i+h-1}^{(1)} \in \mathbb{R}^{h \times d_c}$, σ is the ELU [26] activation function, and b is the bias vector. This convolutional operation generates a local feature map $F^{(i)} \in \mathbb{R}^{L_i - k + 1 \times \mathrm{f}}$, which is then subjected to the attention layer MHSA to model long-range interdependencies by matching patterns between prefix and suffix, or repeated intent. The convolutional output $F^{(i)}$ is linearly projected into a set of queries (Q), keys (K), and values (V) to facilitate scaled dot-product attention:

$$Q = F^{(i)}W^{Q}, K = F^{(i)}W^{k}, V = F^{(i)}W^{V} \quad (7)$$

The scaled dot-product attention is computed as in (1), allowing the model to assign dynamic importance to different subword regions depending on context. The outputs of each branch are concatenated, followed by adaptive max-pooling to obtain fixed-size word-level vectors from variable-length character sequences. The final output of the SPR module for the input sentence S is:

$$Z = \left[z^{(1)}, z^{(2)}, \cdots, z^{(N)}\right] \epsilon \mathbb{R}^{N \times 2f} \quad (8)$$

Hence, by integrating local sensitivity via the 1D-CNN with global interpretability through MHSA, the proposed SPR module augments the feature representation capability of the proposed hybrid DL system. Figure 1 illustrates the modeling of the SPR module with three parallel 1D-CNN-attention branches and pooling operations described above.

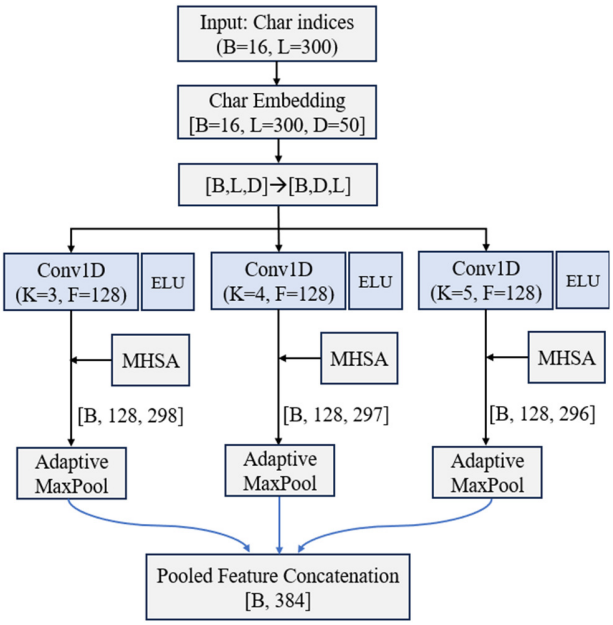

Fig. 1. Illustration of the proposed subword-attention model.

3) Gated MLP Fusion (GMF)

Unlike most existing fusion strategies that statically combine multimodal feature representations, this study uses GMF to achieve a content-aware inter-branch fusion mechanism. It consists of a learnable gating layer that balances contributions from both modalities, i.e., semantic representations from the contextual encoder (BERT) with morphological and orthographic learned attributes from the SPR module. In addition, three parallel MLP fusion blocks are included, followed by soft attention to capture deeper nonlinear interactions in the combined feature space. The proposed GMF considers pooled outputs H and Z (local feature map) from BERT and SPR, respectively, which are concatenated to form a joint representation vector $C = [H||Z]$. Then, a parameterized gating layer is modeled using linear projection and softmax activation, which produces two normalized weights $g_H, g_Z \in [0,1]$, satisfying $g_H + g_Z = 1$, that regulate the relative contribution of each branch:

$$[g_H, g_z] = softmax\left(C \times W_g + b_g\right) \quad (9)$$

$$F_{gated} = [g_H \odot H \,||\, g_Z \odot Z] \quad (10)$$

Here, W_g and b_g are the learnable parameters, $\odot$ denotes elementwise scaling, and F_{gated} is the fused gated vector that allows the network to emphasize semantics or morphology depending on the input characteristics. The gated vector is processed in three parallel MLP branches, each applying a distinct transformation with ELU activation. Their outputs are then combined via learnable attention weights α_l, normalized by softmax activation:

$$F_{fused} = \sum_{l=1}^{3} \alpha_l F^{(l)}, \quad \alpha_l = \frac{exp^{w_l}}{\sum_{j=1}^{3} exp^{w_l}} \quad (11)$$

Finally, the fused representation is passed through a fully connected softmax activated classifier to produce the predicted label distribution. By jointly optimizing the gate layer and the fusion MLP through backpropagation, the model learns to encode not only multi-modal content but also the interaction structure between word-level semantics and character-level morphology, where a single modality might be insufficient. Figure 2 illustrates the workflow of the GMF module, where contextual features H from BERT and subword features Z from SPR are concatenated, passed through a gating layer to compute modality-specific weights, fused via parallel MLP branches, aggregated through attention-weighted summation, and classified.

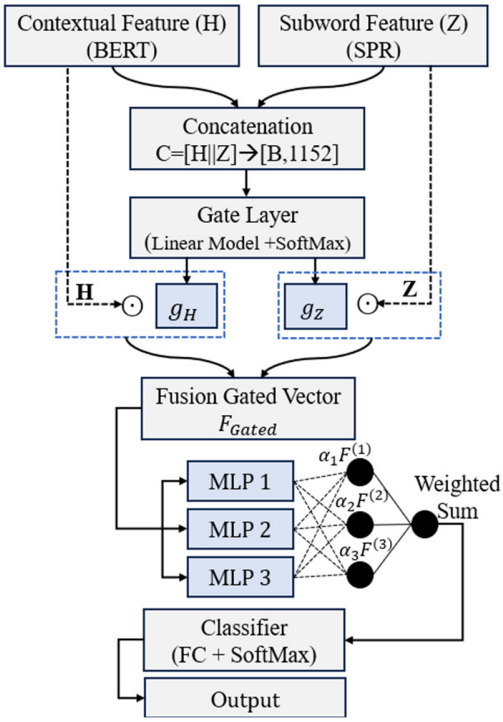

Fig. 2. Illustration of the proposed GMF.

III. RESULTS AND DISCUSSION

The development and experimental validation of the proposed hybrid DL model were carried out on a Windows 11 system with 16 GB RAM using Python 3.9. The model was trained using an NVIDIA CUDA-enabled GPU. Table II summarizes the model and training configuration parameters, selected based on empirical analysis considering the characteristics of the datasets and the model's initial performance on the validation set. It should be noted that the same configuration was used for both datasets, with the only change being the number of neurons in the output layer, determined by the number of classes in each dataset.

TABLE III. TRAINING AND MODEL CONFIGURATION

Parameter	Values
Max word tokens (BERT)	128
Max char tokens (SPR)	300
Character vocabulary size	69 (a–z, 0–9, punctuations, space)
Contextual module (BERT)	Hidden Size = 768
Subword module (SPR)	See Table II
Fusion layer	Gate output dim = 2
	MLP dim = 512, activation = ELU
Classifier layer	Fully connected + Softmax
	Output dim = 3(dataset-1), 2 (dataset-2)
Optimizer	Adam
Learning rate	2×10^{-5}
Weight decay	1×10^{-4}
Loss function	CrossEntropyLoss
Epochs	5
Batch size	16

A. Test Performance on the Davidson (DV) Dataset

Figure 3 presents a confusion matrix for the trained model tested on the DV dataset, illustrating the distribution of correctly and incorrectly classified classes. The model demonstrated a strong predictive capacity with higher true positive rates and a lower number of misclassification rates subjected to both hate and offensive classes, which is expected due to their semantic closeness in certain contexts. It can be seen that among the 117 actual hate samples, 32 were misclassified as offensive, likely because hate and offensive share some common features that confuse the model, which can be difficult even for human-based analysis. However, for other classes, the models exhibited very low false-positive rates. Table IV presents a detailed performance analysis considering Overall Accuracy (OA), precision, recall, and F1-score. The proposed model achieved 98% OA, showing strong performance in identifying the Offensive class with a recall rate of 100% and an F1-score of 99%. However, a slightly lower recall of 72% was obtained for the Hate category, but the high precision (0.97) indicates the model is cautious and conservative in predicting hate content, thereby minimizing false positives.

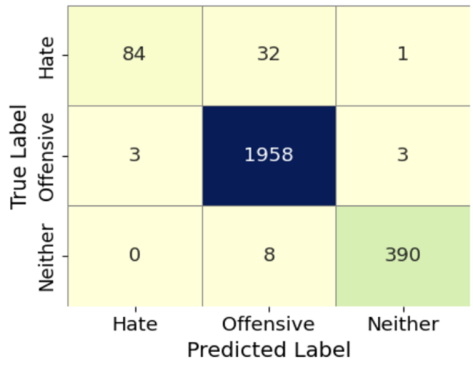

Fig. 3. Confusion matrix of predictions on the DV dataset.

TABLE IV. CLASSIFICATION REPORT ON DV

Label	Precision	Recall	F1-score	Support
Hate	97%	72%	82%	117
Offensive	98%	100%	99%	1964
Neither	99%	98%	98%	398
OA			98%	

B. Test Performance on the OLID Dataset

Figure 4 shows the confusion matrix for the model tested on the OLID dataset, which indicates an effective generalization capability for both NOT (non-offensive) and OFF (offensive) classes. Table V shows class-wise numerical results on the OLID dataset, where the model achieved OA of 91% with a precision of 0.85 and a recall of 0.89 for the OFF class. The balanced F1-scores (0.93 for NOT, 0.87 for OFF) demonstrate that the model maintains consistent performance across both categories. These results suggest that the dual-branch architecture with the GMF effectively captures both contextual semantics and subword-level patterns.

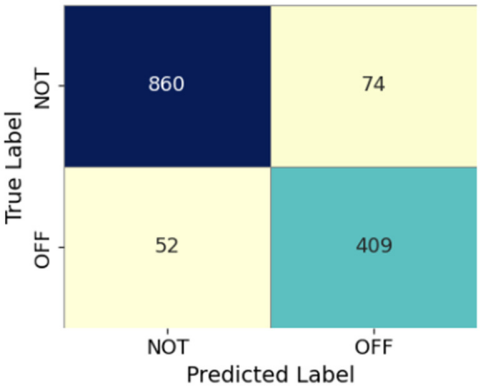

Fig. 4. Confusion matrix of predictions on the OLID dataset.

TABLE V. CLASSIFICATION REPORT ON OLID

Label	Precision	Recall	F1-score	Support
NOT	94%	92%	93%	934
OFF	85%	89%	87%	461
OA			91%	

The results show that the model is effective in identifying offensive content but faced some challenges in borderline or context-dependent cases due to high class imbalance and the heterogeneous nature of OLID, where the OFF class includes diverse subtypes, such as insults, threats, profanity, and sarcasm.

C. Comparative Analysis

Table VI presents comparative results in terms of OA and weighted F1-score to evaluate the effectiveness of the proposed approach against similar existing approaches. A closer observation of the comparative results shows that the proposed hybrid DL model with gated representation outperforms existing models with a significant margin in both OA and weighted F1-score.

TABLE VI. COMPARISON WITH EXISTING WORKS

Approach	Dataset	OA	F1-score
Transformer-ANN [7]	DV	—	90%
BiGRU [27]	OLID	78%	74%
BiLSTM-Att [28]	OLID	78%	74%
SKS [29]	DV	95.1%	96.3%
DS-LoRA [30]	OLID	82.5%	81.3%
Stacking Ensemble [31]	DV	77.6%	97.06%
XLNet [32]	OLID	85%	78%
Proposed	DV	98%	98%
Proposed	OLID	91%	91%

The reason behind achieving better performance is the adoption of a multi-stream architecture, with the novel SPR and GMF modules. Existing transformer-based models [7, 30, 32] without explicit subword handling failed to capture critical orthographic signs, whereas recurrent [27-29] and ensemble [31] approaches lack adaptive cross-modal interaction, which is important for capturing high-variance social media text. The proposed SPR enables robust character-level modeling of obfuscated, misspelled, and morphological patterns, and the GMF further enhances classification by adaptively weighting and integrating semantic features from BERT with morphological signals from SPR, thereby ensuring the most informative modality leads the decision process for each input.

D. Ablation Study

1) Hyperparameter Sensitivity

This section presents a hyperparameter sensitivity analysis considering various optimizer types, learning rates, and weight decay values to determine optimal training conditions and examine model generalizability under different configurations. Figure 5 illustrates an optimizer-wise sensitivity analysis concerning OA on both the DV and OLID datasets. The Adam optimizer yielded the best performance on both datasets. On the other hand, AdamW shows comparable performance, whereas SGD and RMSprop underperform because of their limited adaptability to dynamically adjust learning rates and handle sparse gradients. As illustrated in Figure 6, the learning rate has a significant impact on model performance. After testing standard learning rate steps commonly used in DL, such as 0.01, 0.001, and 0.0001, a consistent improvement was observed as the learning rate decreased. However, after 0.0001, the performance did not plateau, suggesting that further fine-tuning could still improve convergence. In this regard, a finer granularity was introduced in the lower learning rate, showing that 0.00002 resulted in the highest accuracy and stable convergence of the proposed architecture. Figure 7 presents an analysis of weight decay, which is a vital parameter that helps prevent overfitting by penalizing large weights. Performance initially improves from no regularization (0) to 1e-4, but lower values offered a competitive result, but slightly underperformed compared to 1e-4. However, as the weight decay increases beyond 1e-4, a clear performance degradation was observed, indicating an excessive penalization of weights.

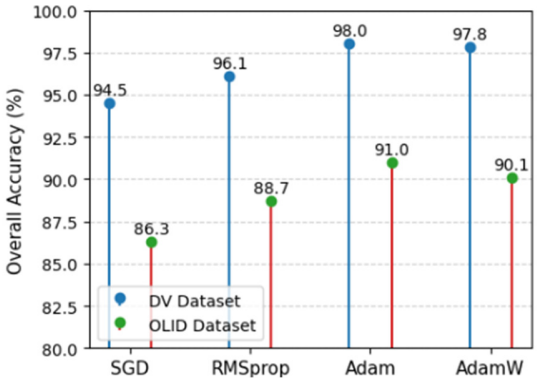

Fig. 5. Optimizer-wise sensitivity analysis.

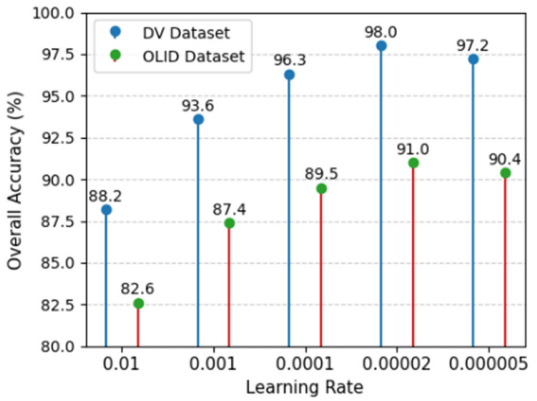

Fig. 6. Learning rate sensitivity analysis.

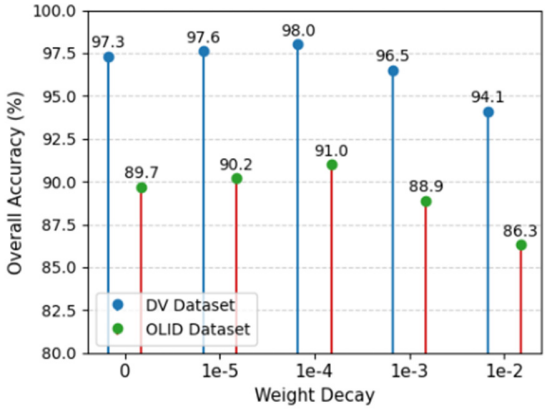

Fig. 7. Weight decay sensitivity analysis.

2) Component-Level Impact

Table VII presents an ablation analysis to examine the impact of components on classification accuracy. Experiments were carried out to isolate the contribution of each major module, removing the SPR, BERT, and GMF mechanisms. The aim was to evaluate the additive value of each component in the proposed multi-stream hybrid design. The results validate the adoption of the multi-stream architecture that contributes substantially to the model's superior performance.

TABLE VII. COMPONENT-LEVEL ABLATION STUDY

Variants	DV OA	Δ	OLID OA	Δ
Full (proposed)	0.98	—	0.91	—
w/o SPR (BERT only)	0.96	–0.02	0.84	–0.07
w/o BERT (SPR only)	0.94	–0.04	0.85	–0.06
Concat + MLP	0.96	–0.02	0.88	–0.03
Gating only	0.97	–0.01	0.89	–0.02
Parallel MLPs + avg	0.97	–0.01	0.90	–0.01
SPR w/o MHSA	0.96	–0.02	0.88	–0.03
Single kernel (k=5)	0.96	–0.02	0.88	–0.03
w/o RTT	0.97	–0.01	0.86	–0.05

IV. CONCLUSION

This study presented a novel method for detecting objectionable social media text using data augmentation, multi-modal feature extraction, and parameterized feature fusion. The proposed hybrid DL framework involves text preprocessing and data augmentation through RTT using the Google API to reduce the impact of class imbalance. The proposed multi-modal architecture uses the BERT model for contextual feature learning and a novel SPR designed based on convolutional and autoencoder mechanisms. Another important contribution is feature fusion using a gated MLP representation, which adaptively combines features from both modalities for better classification performance. Extensive experiments demonstrated that the proposed approach outperformed existing methods in terms of accuracy and weighted F1-score. Future work will explore the extension of the model to multilingual and multimodal settings and integrating explainability tools to support more transparent and trustworthy decision-making.

REFERENCES

[1] G. Bouvier, "What is a discourse approach to Twitter, Facebook, YouTube and other social media: connecting with other academic fields?," *Journal of Multicultural Discourses*, vol. 10, no. 2, pp. 149–162, May 2015, https://doi.org/10.1080/17447143.2015.1042381.

[2] S. S. Mane, S. Kundu, and R. Sharma, "A Survey on Online Aggression: Content Detection and Behavioral Analysis on Social Media," *ACM Computing Surveys*, vol. 57, no. 7, Oct. 2025, Art. no. 171, https://doi.org/10.1145/3711125.

[3] S. V. Kogilavani, S. Malliga, K. R. Jaiabinaya, M. Malini, and M. Manisha Kokila, "Characterization and mechanical properties of offensive language taxonomy and detection techniques," *Materials Today: Proceedings*, vol. 81, pp. 630–633, Jan. 2023, https://doi.org/10.1016/j.matpr.2021.04.102.

[4] C. J. Ferguson, "Does the Internet Make the World Worse? Depression, Aggression and Polarization in the Social Media Age," *Bulletin of Science, Technology & Society*, vol. 41, no. 4, pp. 116–135, Dec. 2021, https://doi.org/10.1177/02704676211064567.

[5] O. M. Alyasiri and Y. N. Cheah, "Multi-Class Text Classification using Machine Learning Techniques," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 22598–22604, Jun. 2025, https://doi.org/10.48084/etasr.9994.

[6] H. Mehta and K. Passi, "Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI)," *Algorithms*, vol. 15, no. 8, Aug. 2022, Art. no. 291, https://doi.org/10.3390/a15080291.

[7] A. Yadav, F. A. Khan, and V. Singh, "A Multi-Architecture Approach for Offensive Language Identification Combining Classical Natural Language Processing and BERT-Variant Models," *Applied Sciences*, vol. 14, no. 23, Jan. 2024, Art. no. 11206, https://doi.org/10.3390/app142311206.

[8] A. T. Azar, H. M. Noori, A. R. Mahlous, A. Al-Khayyat, and I. K. Ibraheem, "Quasi-Reflection Learning Arithmetic Firefly Search Optimization with Deep Learning-based Cyberbullying Detection on Social Networking," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17162–17169, Oct. 2024, https://doi.org/10.48084/etasr.8314.

[9] Z. Boulouard, M. Ouaissa, M. Ouaissa, M. Krichen, M. Almutiq, and K. Gasmi, "Detecting Hateful and Offensive Speech in Arabic Social Media Using Transfer Learning," *Applied Sciences*, vol. 12, no. 24, Jan. 2022, Art. no. 12823, https://doi.org/10.3390/app122412823.

[10] M. Madhavi *et al.*, "Elevating Offensive Language Detection: CNN-GRU and BERT for Enhanced Hate Speech Identification," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 5, 2024, https://doi.org/10.14569/IJACSA.2024.01505118.

[11] W. Aldjanabi, A. Dahou, M. A. A. Al-qaness, M. A. Elaziz, A. M. Helmi, and R. Damaševičius, "Arabic Offensive and Hate Speech Detection Using a Cross-Corpora Multi-Task Learning Model," *Informatics*, vol. 8, no. 4, Dec. 2021, Art. no. 69, https://doi.org/10.3390/informatics8040069.

[12] A. Alhazmi, R. Mahmud, N. Idris, M. E. M. Abo, and C. I. Eke, "Code-mixing unveiled: Enhancing the hate speech detection in Arabic dialect tweets using machine learning models," *PLOS ONE*, vol. 19, no. 7, 2024, Art. no. e0305657, https://doi.org/10.1371/journal.pone.0305657.

[13] M. Xu and S. Liu, "RB_BG_MHA: A RoBERTa-Based Model with Bi-GRU and Multi-Head Attention for Chinese Offensive Language Detection in Social Media," *Applied Sciences*, vol. 13, no. 19, Jan. 2023, Art. no. 11000, https://doi.org/10.3390/app131911000.

[14] A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, "Hate and Offensive Speech Detection in Hindi and Marathi." arXiv, Nov. 01, 2021, https://doi.org/10.48550/arXiv.2110.12200.

[15] B. T. Pham-Hong and S. Chokshi, "PGSG at SemEval-2020 Task 12: BERT-LSTM with Tweets' Pretrained Model and Noisy Student Training Method," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online), Sep. 2020, pp. 2111–2116, https://doi.org/10.18653/v1/2020.semeval-1.280.

[16] H. Nghiem and H. D. III, "HateCOT: An Explanation-Enhanced Dataset for Generalizable Offensive Speech Detection via Large Language Models." arXiv, Oct. 05, 2024, https://doi.org/10.48550/arXiv.2403.11456.

[17] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "Hate Speech and Offensive Language Detection Using an Emotion-Aware Shared Encoder," in *ICC 2023 - IEEE International Conference on Communications*, Rome, Italy, May 2023, pp. 2852–2857, https://doi.org/10.1109/ICC45041.2023.10279690.

[18] A. Joshi and R. Joshi, "Harnessing Pre-Trained Sentence Transformers for Offensive Language Detection in Indian Languages." arXiv, Oct. 03, 2023, https://doi.org/10.48550/arXiv.2310.02249.

[19] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, pp. 512–515, May 2017, https://doi.org/10.1609/icwsm.v11i1.14955.

[20] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov, "SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification." arXiv, Sep. 24, 2021, https://doi.org/10.48550/arXiv.2004.14454.

[21] T. Y. Zhuo, Q. Xu, X. He, and T. Cohn, "Rethinking Round-Trip Translation for Machine Translation Evaluation." arXiv, May 15, 2023, https://doi.org/10.48550/arXiv.2209.07351.

[22] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "Comparing BERT Against Traditional Machine Learning Models in Text Classification," *Journal of Computational and Cognitive Engineering*, vol. 2, no. 4, pp. 352–356, Apr. 2023, https://doi.org/10.47852/bonviewJCCE3202838.

[23] Y. Lin, C. Wang, H. Song, and Y. Li, "Multi-Head Self-Attention Transformation Networks for Aspect-Based Sentiment Analysis," *IEEE Access*, vol. 9, pp. 8762–8770, 2021, https://doi.org/10.1109/ACCESS.2021.3049294.

[24] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[25] A. Gulati *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition." arXiv, May 16, 2020, https://doi.org/10.48550/arXiv.2005.08100.

[26] S. Kılıçarslan, K. Adem, and M. Çelik, "An overview of the activation functions used in deep learning algorithms," *Journal of New Results in Science*, vol. 10, no. 3, pp. 75–88, Dec. 2021, https://doi.org/10.54187/jnrs.1011739.

[27] R. Ong, "Offensive Language Analysis using Deep Learning Architecture." arXiv, Mar. 19, 2019, https://doi.org/10.48550/arXiv.1903.05280.

[28] L. S. Mut Altin, À. Bravo Serrano, and H. Saggion, "LaSTUS/TALN at SemEval-2019 Task 6: Identification and Categorization of Offensive Language in Social Media with Attention-based Bi-LSTM model," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, MN, USA, Mar. 2019, pp. 672–677, https://doi.org/10.18653/v1/S19-2120.

[29] X. Zhou et al., "Hate Speech Detection Based on Sentiment Knowledge Sharing," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Dec. 2021, pp. 7158–7166, https://doi.org/10.18653/v1/2021.acl-long.556.

[30] Y. Wang *et al.*, "Dynamic Sparse LoRA: Adaptive Low-Rank Finetuning for Nuanced Offensive Language Detection." Computer Science and Mathematics, May 27, 2025, https://doi.org/10.20944/preprints202505.2020.v1.

[31] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "BERT-based Ensemble Approaches for Hate Speech Detection," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, Rio de Janeiro, Brazil, Dec. 2022, pp. 4649–4654, https://doi.org/10.1109/GLOBECOM48099.2022.10001325.

[32] R. Alothman, H. Benhidour, and S. Kerrache, "Offensive Language Detection on Social Media Using XLNet." arXiv, Jun. 26, 2025, https://doi.org/10.48550/arXiv.2506.21795.