A Dual Contrastive Learning Framework for Hate Speech Detection based on Deep Learning

Jie Wang, Rui Lv, Xuan Liu, Lirong Chen[†]

College of Computer Science (College of Software), College of Artificial Intelligence, Inner Mongolia University, Hohhot 010000, Inner Mongolia, China

Abstract

In recent years, the rapid advancement of information technology has facilitated the widespread dissemination of hate speech on social media, severely disrupting the online ecosystem. Detecting hate speech has become a critical task in maintaining a healthy internet environment. This paper presents a deep learning framework that integrates dual contrastive learning, dynamic text embedding, sentiment analysis, TextCNN, BiLSTM, and GRU Gated Recurrent Unit to enhance the accuracy of hate speech detection, the model is called DeepDuCo. We conduct experiments on the publicly available Davidson dataset,the HateEval dataset from SemEval-2019 Task 5, and the Twitter White Supremacy dataset. The experimental results demonstrate that our contrastive learning-enhanced deep learning model, DeepDuCo, significantly outperforms existing approaches in terms of both accuracy and F1 score, especially in addressing data imbalance challenges.

Keywords: hate speech; deep learning; contrastive learning; data imbalance; text classification

1. Introduction

The rapid growth of social media has brought the issue of online hate speech into sharp focus. Hate speech not only undermines the integrity of the online environment but also poses significant threats to societal stability. Hate speech refers to malicious language directed at specific individuals or groups based on characteristics such as race, religion, gender, sexual orientation, disability, or other attributes.[1]The proliferation of hate speech can exacerbate social division and discrimination, incite violence, and inflict lasting psychological and social harm on victims[2]. According to a 2023 survey by the Anti-Defamation League(ADL), 41% of Americans reported experiencing some form of online harassment, with 35% subjected to offensive name-calling, 13% stalked, and 12% experiencing sexual harassment[1]. Detection of hate speech is complicated by the diverse expressions, nuanced semantics, and pronounced class imbalance inherent in hate speech content. Effectively detecting and curbing hate speech has thus emerged as a key challenge in the field of Natural Language Processing. Early detection techniques were primarily based on keyword matching[3], which are overly simplistic and require frequent manual updates. With advancements in deep learning and the emergence of large-scale pre-trained language models, architectures such as LSTM, CNN, and BERT have become standard. For example, Fazil et al.[4]

[†]Corresponding author: Lirong Chen (Email: lrchen10@126.com; ORCID: 0000-0003-4516-209X)

[1]https://www.adl.org/resources/report/online-hateand-harassment-american-experience-2023

proposed a multi-channel attention-based CNN-BiLSTM network(AMC-CNN-BiLSTM) for hate speech detection on social platforms, but their method relies heavily on explicit lexical cues. Riyadi et al.[5] developed a hybrid CNN-RNN model with an additional LSTM layer to capture long-range dependencies and mitigate data imbalance. However, the Indonesian tweet dataset used in their study exhibits only slight imbalance(a hate to non-hate ratio of 1.36), and the use of random oversampling increases the risk of overfitting by duplicating minority samples. Lu et al.[6] proposed a dual contrastive learning framework to address class imbalance, but their augmentation method, injecting noise via dropout, introduced only minor perturbations in Transformer hidden layers, thus limiting representational diversity and robustness.

To address the challenges arising from the complexity of hate speech and severe data imbalance, we propose a comprehensive detection framework, DeepDuCo, consisting of six interrelated modules, data augmentation, which employs synonym replacement and Gaussian noise to enhance data diversity; dynamic embedding generation: which utilizes RoBERTa-large[7] for end-to-end encoding of both original and augmented texts; sentiment analysis: which extracts sentiment intensity scores using a twitter-roberta-base-sentiment model[8]; semantic feature extraction: which integrates BiLSTM for global contextual modeling and TextCNN for capturing local semantic information; GRU Gated Recurrent Unit fusion: which adaptively integrates sentiment and semantic features to enhance emotional representation; and dual contrastive learning: which aims to optimize the embedding space and improve classification performance by pulling semantically similar instances closer and pushing dissimilar ones apart, thereby enhancing inter-class discriminability. Additionally, a consistency regularization term is applied to ensure prediction stability under input perturbations. This unified architecture enables our model to robustly handle datasets with diverse distributions, achieving improved generalization and minority class recognition.

Research Contributions:

We propose a framework that integrates dynamic contextual embedding, deep semantic modeling, Gated Recurrent Unit and dual contrastive learning. The model generates embeddings dynamically during training and combines BiLSTM and TextCNN to capture both global and local features. Gated Recurrent Unit adaptively fuses sentiment feature with semantic feature, enhancing the model's sensitivity to emotionally driven hate speech. The contrastive learning strategy, along with consistency regularization, further improves embedding discriminability.

We validate the model's adaptability and transferability on datasets with diverse label distributions, achieving consistently strong performance. Moreover, our augmentation and contrastive learning strategies require no additional annotation, making the framework broadly applicable and easily deployable for hate speech detection in varied social media contexts.

2. Related Work

This section is divided into two parts: hate speech detection methods and contrastive learning approaches.

2.1. Hate Speech Detection Methods

Automatic hate speech detection is a critical task in the field of Natural Language Processing (NLP). In recent years, machine learning techniques have been applied to identify and mitigate hate speech[9]. However, traditional machine learning methods often struggle to capture contextual information effectively, resulting in suboptimal performance[10]. Alzamzami et al.[11] proposed a BERT-based feed-forward network that integrates sentiment analysis, demonstrating a strong

correlation between hateful behavior and negative emotions. Roy et al.[12] adopted a deep convolutional neural network combined with LSTM and ten-fold cross-validation to address imbalanced datasets; however, the method heavily relies on the quality of the data. Cao et al.[13] introduced HateGAN, a reinforcement learning model that applies data augmentation to address class imbalance. Shakir et al.[14] developed BICHAT, a hybrid model combining BERT, CNN, BiLSTM, and hierarchical attention for hate speech detection on social media. Zhang et al.[15] evaluated ten strategies for addressing class imbalance in abusive language detection, encompassing data-level approaches such as random oversampling, undersampling, and textual augmentation, as well as model-level techniques including weighted cross-entropy and focal loss. Recently, Qin et al.[16] proposed MS-FSLHate, an innovative framework that combines prompt-based learning and adversarial augmentation to improve model performance in low-resource settings by addressing data scarcity and limited generalization.

These studies suggest that although deep learning models are effective at capturing contextual semantic features, they often fall short in tasks involving highly complex semantic structures, such as hate speech detection, particularly under conditions of severe data imbalance.

2.2. Contrastive Learning Approaches

Contrastive learning is a metric-based approach designed to learn a representation space in which similar pairs (x, x^+) are pulled closer and dissimilar pairs (x, x^-) are pushed apart. It has achieved remarkable success in various computer vision tasks, such as person re-identification[17] and object detection[18]. Inspired by its effectiveness in vision domains, researchers have extended contrastive learning to natural language processing tasks, including text classification[19] and hate speech detection[20]. Nan et al.[21] proposed a dual contrastive learning framework to align textual and video content. Notably, Gao[22] leveraged dropout-induced noise from BERT encoders for supervised contrastive learning, providing a simple yet effective textual augmentation strategy that optimizes the embedding space using labeled data. Chen et al.[19] introduced a label-aware dual contrastive learning method for text classification, combining supervised contrastive loss with cross-entropy loss into a novel training paradigm. Dehghan et al.[23] adopted a similar framework with their BERTurk-DualCL model, which integrates cross-entropy and supervised contrastive losses for hate speech detection in Turkish. Lu et al.[6] further examined the applicability of dual contrastive learning across various English text datasets.

Although contrastive learning has shown strong performance in both vision and language tasks, it also exhibits certain limitations. Its effectiveness depends heavily on the quality of positive and negative sample pairs and the representativeness of the data distribution. Without support from deeper modeling architectures, contrastive learning may result in overly simplified embedding spaces and fail to capture complex semantic relationships. This challenge is particularly evident in hate speech detection, where semantic complexity and expressive diversity are prevalent, making contrastive learning alone insufficient to ensure stable performance.

Therefore, we propose a Dual contrastive learning framework for hate speech detection based on deep learning to optimize the embedding space and better capture rich semantic features while enhancing the distinction between different classes. This combined strategy helps reduce class confusion and mitigates distributional shifts, leading to more stable performance in hate speech detection tasks.

3. Methodology

This section systematically presents the overall design and implementation of the proposed hate speech detection model DeepDuCo. To address the challenges of complex semantics, diverse expressions, and severe class imbalance, the model adopts a unified architecture that combines deep semantic modeling with dual contrastive learning strategies to improve representation quality and class separability. Specifically, the model first applies a lightweight data augmentation strategy based on synonym replacement using the Easy Data Augmentation(EDA) technique. In parallel, a fine-tuned twitter-roberta-base-sentiment model is employed to assign sentiment intensity scores to the original text. To generate robust contextual embeddings, we construct a dynamic embedding generation module that performs end-to-end encoding for both original and augmented texts, and introduces Gaussian noise to enhance model robustness.

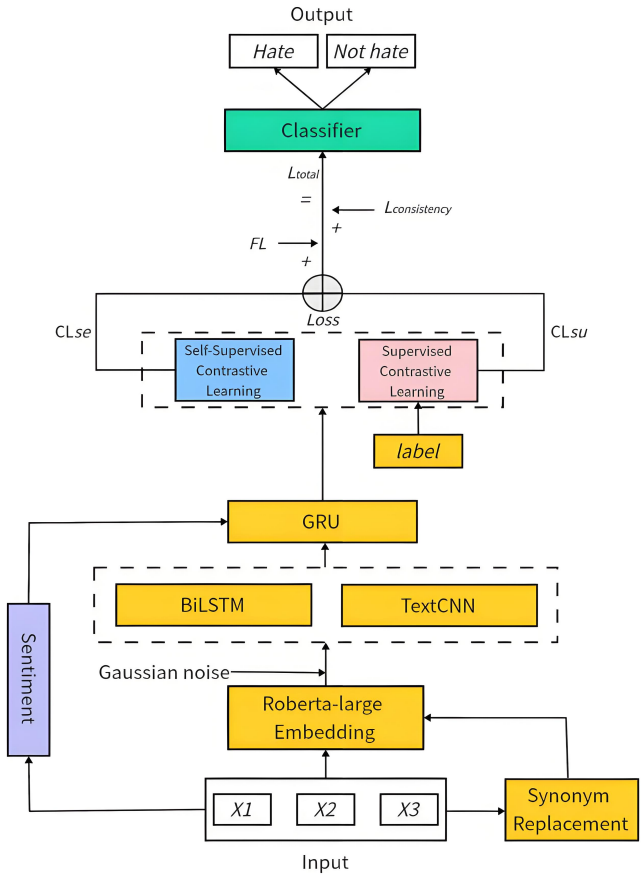

Figure 1: DeepDuCo:Self-Supervised Contrastive Learning: Leverages the original text and its augmented counterpart as positive sample pairs to construct diverse views, thereby enhancing the quality of semantic representations under augmentation consistency. Supervised Contrastive Learning: Utilizes label information to further reinforce the clustering of samples within the same class and the separation of samples across different classes in the representation space, explicitly optimizing class boundaries and improving inter-class discriminability.

In the feature extraction stage, the model employs BiLSTM and TextCNN networks in parallel

to capture semantic features, with BiLSTM modeling global semantic dependencies and TextCNN focusing on local semantic dependencies. To further incorporate sentiment feature into semantic modeling, GRU Gated Recurrent Unit is used to adaptively fuse sentiment feature with semantic features. For the classification and representation learning process, the model introduces a dual contrastive learning strategy that optimizes the semantic structure of sample embeddings from both self-supervised and supervised perspectives. This approach improves the discriminability and generalizability of representations while iteratively refining the distribution of the embedding space during training. Additionally, a consistency regularization term is incorporated to constrain output stability under perturbations, thereby improving model robustness. Finally, the loss function integrates classification loss and contrastive loss to facilitate joint optimization. A fully connected classifier is employed to conduct binary classification between hate and non-hate speech. The overall architecture of the proposed model is depicted in Figure 1.

3.1. Data Augmentation

In this study, we employ Easy Data Augmentation(EDA)[24] to enrich the training data, specifically utilizing the Synonym Replacement(SR) strategy tailored for the hate speech detection task. EDA is a straightforward yet effective text augmentation technique comprising four core operations: synonym replacement(SR), random insertion(RI), random swap(RS), and random deletion(RD). These operations are designed to enhance model performance by increasing data diversity.

For our task, we apply only the synonym replacement operation to maintain semantic consistency, as hate speech detection demands a high level of semantic precision. Operations such as random insertion, swap, or deletion may introduce grammatical inconsistencies or semantic distortions, potentially causing label shifts, such as misclassifying hateful expressions as non-hateful or vice versa. In contrast, synonym replacement substitutes words with semantically similar alternatives, increasing data variability while preserving the original sentence meaning, thereby improving the model's generalization capability.

3.2. Sentiment Feature

We utilize a sentiment analysis model based on the RoBERTa architecture(twitter-roberta-base-sentiment), to extract sentiment scores from the original texts. This model is pre-trained on a large-scale Twitter corpus and fine-tuned for sentiment analysis tasks on social media, making it highly effective and well-adapted to informal and emotionally rich online language.

For each input text, the model outputs a sentiment score denoted as s, which is subsequently used as an input feature in our model pipeline. By incorporating sentiment features, the model can more precisely capture the emotional polarity and intensity embedded in the text, thereby enhancing its ability to distinguish between hateful and non-hateful content.

3.3. Dynamic Embedding Generation

We employ a fine-tuned RoBERTa-large model to perform encoding on the input texts, producing representations for both the original text X_{ori} and its augmented version X_{var}. Average pooling is applied to obtain the corresponding semantic vectors h_{ori} and h_{var}. To enhance robustness during training, Gaussian noise is injected into these embedding vectors, and the perturbed vectors are then used as inputs to subsequent modules.

3.4. Semantic Feature Extraction Module

This module adopts a parallel architecture that integrates BiLSTM and TextCNN to capture semantic features. The BiLSTM component models long-range dependencies and complex semantic structures within the text, effectively capturing contextually implied hate intentions. Meanwhile, the TextCNN component excels at identifying local dependence of phrases, thereby improving the recognition of short or fragmentary expressions. The outputs from both networks are pooled and concatenated to form a unified semantic representation vector f_{text}, which is then used in the downstream fusion and classification modules.

3.5. Gated Fusion Mechanism

To adaptively regulate the contribution of sentiment feature to textual representations, we introduce a gated fusion mechanism. Specifically, the gating weight is computed based on semantic features as follows:

$$g = \sigma(Wgate \cdot ftext + b) \tag{1}$$

Where W_{gate} is a learnable linear transformation matrix, b is the bias term, and σ denotes the sigmoid activation function. The sentiment score s is then scaled by the gating weight $g \cdot s$ and concatenated with the text feature vector f_{text} to form the final fused representation f_{fused}. This mechanism enhances the model's ability to detect emotionally driven hate speech while reducing the risk of being misled by irrelevant sentiment cues.

3.6. Dual Contrastive Learning

Unlike traditional classification losses that focus solely on the discrepancy between predicted labels and ground truths, contrastive learning exploits semantic relationships between samples to shape a structured embedding space by encouraging semantically similar samples to cluster and dissimilar ones to diverge. Meanwhile the classification performance has also been optimized. This promotes better feature representation and improved generalization.

3.6.1. Self-Supervised Contrastive Loss

To bring the original text and its augmented variant closer in the semantic space, we define the self-supervised contrastive loss as:

$$\mathcal{L}_{se} = -\sum_{j=1}^{2N} log \frac{e^{sim(z_j, z_j^+)/\tau_{se}}}{\sum_{k=1}^{2N} 1_{[j \neq k]} \cdot e^{sim(z_j, z_k)/\tau_{se}}} \tag{2}$$

Where τ_{se} is the temperature coefficient, and (z_j, z_j^+) represent the embeddings of the original and augmented texts, forming a positive pair. All other samples in the batch are treated as negative examples.

3.6.2. Supervised Contrastive Loss

To leverage label information for guided sample clustering, we define the super-vised contrastive loss as:

$$\mathcal{L}_{su} = -\sum_{i=1}^{N} \frac{1}{N_{y_i} - 1} \sum_{j=1}^{N} 1_{[i \neq j]} \cdot 1_{[y_i = y_j]} \cdot log \frac{e^{sim(z_i, z_j)/\tau_{su}}}{\sum_{k=1}^{N} 1_{[i \neq k]} \cdot e^{sim(z_i, z_k)/\tau_{su}}} \tag{3}$$

Where τ_{su} is the temperature coefficient used in the supervised setting. The pair (z_i, z_j) represents two samples sharing the same label (a positive pair), while (z_i, z_k) denotes a pair where z_k is randomly selected. y_i and y_j denote the labels corresponding to z_i and z_j, respectively. N_{y_i} denotes the number of samples in the batch that share the same label as z_i.

3.7. Consistency Regularization

We incorporate a KL divergence loss to penalize significant discrepancies in the model's predictions when subjected to different perturbations of the same input:

$$\mathcal{L}_{consistency} = D_{KL}\left(\text{Softmax}(\mathbf{z}_{ori}) \| \text{Softmax}\left(\mathbf{z}_{var}\right)\right) \tag{4}$$

This regularization promotes prediction stability and enhances the model's robustness against input perturbations.

3.8. Classifier Design and Joint Loss Optimization

The final classifier is implemented as a two layers fully connected network with a sigmoid activation function, producing the probability that the input expresses hate. During training, the model jointly optimizes three types of loss functions: focal loss, contrastive loss, and consistency regularization:

$$\mathcal{L}_{total} = \mathcal{L}_{focal} + \lambda(\mathcal{L}_{se} + \mathcal{L}_{su}) + \lambda_c \mathcal{L}_{consistency} \tag{5}$$

Where $\mathcal{L}_{\text{focal}}$ denotes the primary classification loss, and λ and λ_c are hyperparameters that control the relative weights of the contrastive and consistency losses, respectively.

4. Experiment

This section presents the experimental evaluation of our proposed model. We begin by introducing the datasets used, followed by a description of the experimental settings, and conclude with a comparative analysis against several baseline models.

4.1. Dataset

We conduct experiments on three widely used benchmark datasets for hate speech detection: the Davidson dataset[25], the HateEval dataset from SemEval-2019 Task 5[26], and the Twitter White Supremacy dataset[27]. Detailed descriptions of these datasets are provided below:

Davidson Dataset: The Hate Speech and Offensive Language Dataset, compiled by Thomas Davidson et al. in 2017, consists of 24,783 English tweets intended to support the automatic detection of hate speech on social media. It uses three labels: hate speech, offensive language, and neither. Following the approach taken by many other researchers who use this dataset for a binary hate speech classification task[28][29], we group both "offensive language" and "neither" under the single category non-hate speech. As a result, only 1,430 are labeled as hate speech,

while the remaining 23,353 are non-hate. This dataset exhibits a significant class imbalance, with hate speech forming a small minority class.

HateEval (SemEval-2019 Task 5) Dataset: This multilingual hate speech detection task focuses on identifying hateful content targeting immigrants and women on Twitter. We use the English subset, which contains a total of 11,971 tweets, with 5,035 labeled as hate speech and 6,936 as non-hate speech. The data is split into 9,000 training samples, 1,000 validation samples, and 2,971 test samples. This dataset is relatively balanced.

White Supremacy Dataset: This dataset, released by Ona de Gibert et al. in 2018, originates from a white supremacy forum and aims to investigate hate speech within extremist online communities. We use a perfectly balanced subset, which includes 2,304 training sentences (1,152 hate and 1,152 non-hate) and 475 test sentences.

4.2. Experimental Settings

As detailed in Section 3.1, our data augmentation strategy is configured as follows:Number of Augmentations: 1 (one-to-one augmentation per original sentence), Alpha for Synonym Replacement: 0.05 (5% of non-stop words are replaced with their synonyms), Alpha for Random Deletion / Insertion / Swap: 0.0 (these operations are deactivated). We utilize the WordNet lexicon to identify synonyms for non-stop words and randomly replace selected tokens in each sentence. This method increases lexical diversity while preserving semantic consistency, thereby enhancing the model's robustness to varied expressions.

During model training, we adopt the AdamW optimizer with an initial learning rate of 1×10^{-4} and a batch size of 128. A cosine annealing learning rate scheduler is employed to dynamically adjust the learning rate across epochs. To improve generalization and prevent overfitting, a dropout rate of 0.5 is applied. For the loss functions, we implement a dual contrastive learning mechanism with the following configurations: Temperature for Self-supervised Contrastive Loss: $\tau_{\text{se}} = 0.1$, Temperature for Supervised Contrastive Loss: $\tau_{\text{su}} = 0.05$, Loss Weights: $\lambda = 2$ for the contrastive loss, and $\lambda_c = 0.05$ for the consistency regularization term. All training and evaluation procedures are conducted on an NVIDIA A100 GPU platform.

4.3. Evaluation Metrics

To evaluate model performance, we use Accuracy, Precision, Recall and F1-score as the primary metrics. Specifically: For the SemEval-2019 Task 5 and White Supremacy datasets, we report macro-F1 to reflect performance across balanced classes. For the Davidson dataset, due to its severe class imbalance, we report weighted-F1 to better evaluate performance on the minority (hate speech) class.

The macro-F1 score is calculated by averaging the F1-score of each class without considering class imbalance, The weighted-F1 score is the weighted average of F1-scores for all classes, weighted by the number of instances in each class.

$$\text{macro} - \text{F1} = \frac{1}{C} \sum_{i=1}^{C} \text{F1}_i \tag{6}$$

Where C denotes the total number of classes, F1_i is the F1-score for class i.

$$\text{weighted-F1} = \sum_{i=1}^{C} \frac{n_i}{N} \cdot \text{F1}_i \tag{7}$$

Where n_i is the number of samples in class i, $N = \sum_{i=1}^{C} n_i$ is the total number of samples across all classes.

4.4. Comparative Experiment and Results Analysis

We will compare the proposed model with several baseline methods applied in all three datasets, including traditional baseline models such as CNN, LSTM, BERT and RoBERTa. Then, we include a comparison with a general-purpose large language model, DeepSeek. In addition, we evaluate our model against several advanced approaches.

4.4.1. Comparative Model

The compared models cover some common methodologies for hate speech detection. CNN+ LSTM [28] combines convolutional neural networks with long short-term memory networks to extract local semantic patterns and model long-range dependencies. ADASYN[30] introduces an adaptive oversampling strategy to address class imbalance, paired with traditional classifiers for text classification. SKS[29] leverages external sentiment lexicons to extract affective signals and integrates them with semantic features for enhanced emotional understanding. DCL[6] employs a dual-channel contrastive learning framework—both self-supervised and supervised—to optimize the structure of the embedding space based on sample similarity. The UA[31] system uses traditional machine learning algorithms with shallow textual features such as TF-IDF and Bag-of-Words to perform classification. LSTM-ELMo+BoW[32] combines contextualized embeddings from ELMo, sequential modeling from LSTM, and lexical information from bag-of-words to capture both deep semantics and word-level features. MC-CNN[33] introduces a multi-channel convolutional architecture to extract features from different receptive fields, improving the modeling of short social media texts. Finally, Opt-1.3b[34] utilizes lightweight large language models, fine-tuned via parameter-efficient techniques like LoRA to specialize in hate speech detection tasks.

4.4.2. Results Analysis

Model	**Accuracy**	**Precision**	**Recall**	**weighted-F1**
CNN	0.901	0.732	0.673	0.853
LSTM	0.864	0.692	0.574	0.837
BERT	0.926	0.763	0.754	0.900
RoBERTa	0.885	0.878	0.924	0.897
DeepSeek	0.482	0.097	**0.964**	0.596
CNN+LSTM	0.908	0.902	0.908	0.905
ADASYN	0.942	0.937	0.934	0.945
SKS	0.951	-	-	0.963
DCL	0.959	-	-	0.956
DeepDuCo	**0.961**	**0.955**	0.823	**0.964**

Table 1: The Davidson dataset results

We begin with the Davidson datasetas, as shown in Table 1, among the baseline models, CNN and LSTM achieve weighted-F1 scores of 0.853 and 0.837, respectively, reflecting their basic capability in modeling hate speech. BERT and RoBERTa, benefiting from large-scale pre-trained

language representations, significantly improve the weighted-F1 scores to 0.900 and 0.897, respectively, demonstrating superior semantic understanding. DeepSeek obtains a weighted-F1 of 0.596, showing strong recall of 0.964 but extremely low precision, indicating a tendency to overpredict hate speech under severe class imbalance. Furthermore, CNN+LSTM, which combines local and global features, achieves a weighted-F1 of 0.905, outperforming single-architecture models. ADASYN, designed to handle imbalanced data, attains a weighted-F1 of 0.945, highlighting its effectiveness in addressing class imbalance. SKS, which incorporates sentiment-aware semantic modeling, further improves performance to a weighted-F1 of 0.963. DCL leverages contrastive learning to enhance representational discriminability, yielding a weighted-F1 of 0.956. Our proposed multi-module fusion model, which integrates dynamic embedding, semantic modeling, and discriminative optimization, achieves the highest weighted-F1 of 0.964, demonstrating a comprehensive advantage in scenarios with severe data imbalance.

Next, we evaluate the performance on the SemEval-2019 Task 5 HateEval dataset using the same baseline models: CNN, LSTM, BERT, RoBERTa and DeepSeek. We also include comparisons with recent advanced approaches. Since SemEval-2019 Task 5 emphasizes Accuracy and macro-F1 as the primary evaluation metrics, our analysis focuses on these two indicators. The results are summarized in Table 2.

Model	**Accuracy**	**macro-F1**
CNN	0.565	0.462
LSTM	0.550	0.530
BERT	0.598	0.586
RoBERTa	0.621	0.617
DeepSeek	0.675	0.668
SKS	0.659	0.652
DCL	0.678	0.672
UA	0.728	0.720
LSTM-ELMo+BoW	0.743	0.738
MC-CNN	**0.838**	0.780
Opt-1.3b	0.800	0.810
DeepDuCo	0.803	**0.817**

Table 2: HateEval dataset results

As shown in Table 2, on the HateEval dataset—which features a relatively balanced label distribution—basic models such as CNN, LSTM, BERT, and RoBERTa exhibit gradual improvements in macro-F1 scores. However, their overall performance is constrained by limited feature modeling capabilities, making it difficult to fully capture the complex semantics of hate speech. DeepSeek achieves a macro-F1 of 0.668 surpassing traditional models, indicating a strong semantic understanding ability, but there is a considerable gap compared with deep learning frameworks that focus on this field SKS and DCL. For SKS and DCL, although they perform exceptionally well on the Davidson dataset, their results here are weaker, with macro-F1 scores of 0.652 and 0.672 respectively. This indicates that their advantages are less pronounced under relatively balanced data conditions.

In contrast, UA and LSTM-ELMo+BoW, which integrate both shallow features and contextual representations, achieve macro-F1 scores of 0.720 and 0.738, respectively. MC-CNN further enhances local expression modeling through its multi-channel architecture, reaching a macro-F1 of 0.780. Opt-1.3b, as a lightweight large-scale model, demonstrates strong generalization and

discriminative capability, achieving a macro-F1 of 0.810. Our proposed model further improves macro-F1 to 0.817, demonstrating solid adaptability to diverse datasets and nuanced emotional cues.

Lastly, we evaluate our model on the White Supremacy dataset. Unlike the SemEval-2019 Task 5 dataset, which is relatively balanced, this experiment is designed to examine whether our model maintains high performance under strictly balanced label distributions. For this purpose, we use a balanced subset of the White Supremacy dataset. It is important to note that all baseline and comparative results are derived from this subset, however, we haven't found the comparison model for conducting experiments with this subset. Nevertheless, the results effectively demonstrate the generalizability of our model across datasets with different characteristics. The performance comparison is summarized in Table 3.

Model	**Accuracy**	**Precision**	**Recall**	**macro-F1**
CNN	0.660	0.663	0.652	0.672
LSTM	0.667	0.665	0.654	0.653
BERT	0.692	0.671	0.678	0.700
RoBERTa	0.702	0.661	0.675	0.715
DeepSeek	0.821	0.755	**0.954**	0.819
DeepDuCo	**0.822**	**0.844**	0.792	**0.822**

Table 3: White Supremacy dataset results

It is worth noting that DeepSeek's overall performance is close to that of DeepDuCo, which indicates that on balanced datasets with smaller data size and relatively concentrated semantic patterns, large language models can fully leverage their contextual understanding capabilities. However, its very high Recall, similar to the metrics observed on the Davidson dataset, also suggests that DeepSeek tends to classify a large portion of instances as hate speech. This behavior may be attributed to a risk avoidance bias inherent in general-purpose LLMs when detecting harmful content[35].

4.5. Ablation Study

To better assess the contribution of each component in our proposed model, we conduct a series of ablation experiments. Beginning with the full model configuration A1, we systematically remove individual modules to construct the following comparative variants: A2 excludes the dual contrastive learning module, A3 removes the supervised contrastive learning, A4 removes the sentiment feature, A5 eliminates the synonym-based data augmentation strategy, and A6 omits the consistency regularization term.

All ablation experiments are performed on the Davidson dataset, each group of experiments was conducted under the same data and training Settings to ensure fair comparison. This experimental setup allows us to isolate and evaluate the impact of each component on model robustness, semantic representation capacity, and performance in handling class imbalance.

As shown in the ablation results in Table 4, each component of the proposed architecture contributes meaningfully to the model's overall performance. The full model A1 achieves the highest scores across all metrics, confirming the effectiveness of the integrated design. Removing the dual contrastive learning module A2 leads to a substantial drop in weighted-F1, yielding a score of 0.912, which highlights its central role in enhancing embedding discriminability and improving the classification of minority-class instances. Excluding the supervised contrastive

Model Variant	Accuracy	Precision	Recall	weighted-F1
A1: Full Model	**0.961**	**0.955**	**0.823**	**0.964**
A2: w/o DCL	0.910	0.903	0.774	0.912
A3: w/o SupCon	0.931	0.928	0.765	0.927
A4: w/o Sentiment	0.954	0.948	0.777	0.958
A5: w/o Synonym Repl.	0.958	0.951	0.781	0.952
A6: w/o Consistency Loss	0.953	0.946	0.773	0.956

Table 4: Ablation experiments based on the Davidson dataset

learning module A3 leads to a clear performance drop, with a weighted-F1 score of 0.927, highlighting its role in improving feature discriminability and confirming the advantage of the dual contrastive learning strategy. Excluding the sentiment feature A4 causes a moderate decrease in performance, resulting in a score of 0.958, indicating its utility in identifying emotionally driven hate speech. Omitting the synonym-based data augmentation strategy A5 slightly reduces the model's adaptability to linguistic variations, with a resulting score of 0.952. Meanwhile, removing the consistency regularization term A6 results in a minor performance decline, reflected in a score of 0.956, demonstrating its auxiliary role in enhancing robustness under perturbations. Overall, these findings demonstrate that all modules contribute synergistically to the model's robustness, semantic expressiveness, and generalization ability, with the contrastive learning strategy being the most influential component.

5. Conclusion and Future Expectations

This study addresses key challenges in hate speech detection on social media, including complex expression patterns, diverse semantic structures, and extreme class imbalance. We propose a multi-module detection framework, DeepDuCo, that employs RoBERTa-large for dynamic text embedding, utilizes TextCNN and BiLSTM for local and global semantic feature extraction, and integrates emotional and semantic features via a GRU Gated Recurrent Unit fusion mechanism. Furthermore, a dual contrastive learning strategy—combining self-supervised and supervised objectives—is introduced to significantly enhance the discriminability of the embedding space and improve classification of minority-class samples. Extensive experiments on multiple public datasets validate the effectiveness and strong generalization ability of the proposed approach, which maintains robust performance even under highly imbalanced conditions.

Future Work: To promote interpretablity and applicability in public governance, we will investigate attention-based visualization and causal inference mechanisms to provide transparent, human-understandable explanations of model predictions.

Author Contributions

This research was jointly conducted under the guidance of the first author, Jie Wang, and the corresponding author, Lirong Chen. Jie Wang was primarily responsible for proposing the research idea, designing and implementing the methodology, organizing the experiments, analyzing the results, and drafting the manuscript. Lirong Chen oversaw the overall direction of the study, provided constructive feedback, optimized the research design, critically revised the manuscript, and secured funding for the project. Rui Lv contributed to data preprocessing and verification

of the experimental outcomes. Xuan Liu assisted with statistical analysis of the results and contributed to the discussion and formulation of the conclusions. All authors have read and approved the final version of the manuscript.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 71862027, and in part by the Natural Science Foundation of Inner Mongolia Autonomous Region under Grant No. 2025MS07016.

References

[1] J Nockleyby. hate speech in encyclopedia of the american constitution. *Electronic Journal of Academic and Special Librarianship*, 2000.

[2] Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95, 2017.

[3] Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*, 2017.

[4] Mohd Fazil, Shakir Khan, Bader M Albahlal, Reemiah Muneer Alotaibi, Tamanna Siddiqui, and Mohd Asif Shah. Attentional multi-channel convolution with bidirectional lstm cell toward hate speech prediction. *IEEE Access*, 11:16801–16811, 2023.

[5] Slamet Riyadi, Annisa Divayu Andriyani, and Siti Noraini Sulaiman. Improving hate speech detection using double-layers hybrid cnn-rnn model on imbalanced dataset. *IEEE Access*, 2024.

[6] Junyu Lu, Hongfei Lin, Xiaokun Zhang, Zhaoqing Li, Tongyue Zhang, Linlin Zong, Fenglong Ma, and Bo Xu. Hate speech detection via dual contrastive learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2787–2795, 2023.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[8] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*, 2020.

[9] Sindhu Abro, Sarang Shaikh, Zahid Hussain Khand, Ali Zafar, Sajid Khan, and Ghulam Mujtaba. Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11(8), 2020.

[10] Karimah Mutisari Hana, Said Al Faraby, Arif Bramantoro, et al. Multi-label classification of indonesian hate speech on twitter using support vector machines. In *2020 International Conference on Data Science and Its Applications (ICoDSA)*, pages 1–7. IEEE, 2020.

[11] Fatimah Alzamzami and Abdulmotaleb El Saddik. Monitoring cyber sentihate social behavior during covid-19 pandemic in north america. *Ieee Access*, 9:91184–91208, 2021.

[12] Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das, and Xiao-Zhi Gao. A framework for hate speech detection using deep convolutional neural network. *IEEE Access*, 8:204951–204962, 2020.

[13] Rui Cao and Roy Ka-Wei Lee. Hategan: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6327–6338, 2020.

[14] Shakir Khan, Mohd Fazil, Vineet Kumar Sejwal, Mohammed Ali Alshara, Reemiah Muneer Alotaibi, Ashraf Kamal, and Abdul Rauf Baig. Bichat: Bilstm with deep cnn and hierarchical attention for hate speech detection. *Journal of King Saud University-Computer and Information Sciences*, 34(7):4335–4344, 2022.

[15] Yaqi Zhang, Viktor Hangya, and Alexander Fraser. A study of the class imbalance problem in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 38–51, 2024.

[16] Zhenkai Qin, Dongze Wu, Yuxin Liu, and Guifang Yang. Few-shot hate speech detection based on the mindspore framework. *arXiv preprint arXiv:2504.15987*, 2025.

[17] Khadija Khaldi and Shishir K Shah. Cupr: Contrastive unsupervised learning for person re-identification. In *VISIGRAPP (5: VISAPP)*, pages 92–100, 2021.

[18] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8392–8401, 2021.

[19] Qianben Chen, Richong Zhang, Yaowei Zheng, and Yongyi Mao. Dual contrastive learning: Text classification via label-aware data augmentation. *arXiv preprint arXiv:2201.08702*, 2022.
[20] Youngwook Kim, Shinwoo Park, and Yo-Sub Han. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th international conference on computational linguistics*, pages 6667–6679, 2022.
[21] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2765–2775, 2021.
[22] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.
[23] Somaiyeh Dehghan and Berrin Yanıkoğlu. Multi-domain hate speech detection using dual contrastive learning and paralinguistic features. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11745–11755, 2024.
[24] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.
[25] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.
[26] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63, 2019.
[27] Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018.
[28] Ayush Madhukar, Aparnay Madhukar, Sushama Nagpal, et al. An ensemble based approach to detect hate speech. In *2024 IEEE Region 10 Symposium (TENSYMP)*, pages 1–6. IEEE, 2024.
[29] Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. 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)*, pages 7158–7166, 2021.
[30] Inderdeep Kaur Aulakh, Raj Kumari, et al. The impact of employing sampling-based strategies to address the issue of an imbalanced dataset in hate speech and offensive language categorization. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2024.
[31] Carlos Perelló, David Tomás, Alberto Garcia-Garcia, Jose Garcia-Rodriguez, and Jose Camacho-Collados. Ua at semeval-2019 task 5: setting a strong linear baseline for hate speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 508–513, 2019.
[32] Juan Manuel Pérez and Franco M Luque. Atalaya at semeval 2019 task 5: Robust embeddings for tweet classification. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 64–69, 2019.
[33] Alon Rozental and Dadi Biton. Amobee at semeval-2019 tasks 5 and 6: Multiple choice cnn over contextual embedding. *arXiv preprint arXiv:1904.08292*, 2019.
[34] Tanmay Sen, Ansuman Das, and Mrinmay Sen. Hatetinyllm: hate speech detection using tiny large language models. *arXiv preprint arXiv:2405.01577*, 2024.
[35] Kailing Peng and Lukas Derungs. *Hate Speech Detection with LLM*. PhD thesis, OST Ostschweizer Fachhochschule, 2024.

Author Biography

Jie Wang is currently pursuing his Master's degree in the School of Computer Science at Inner Mongolia University. His research interests lie primarily in the field of natural language processing, with a particular emphasis on hate speech detection and related text classification tasks. He is also interested in machine learning, sentiment analysis, and computational linguistics methods for understanding and processing human language.

Rui Lv, a master's student at the College of Computer Science, Inner Mongolia University. His research interests focus on natural language processing, especially multimodal hate speech classification involving both images and text.

Xuan Liu, a master's student at the School of Computer Science, Inner Mongolia University. Her research mainly focuses on natural language processing, in particular on hate speech detection.

Lirong Chen, Associate Professor at the School of Computer Science, Inner Mongolia University. She graduated with a Ph.D. from Dalian University of Technology. Her research interests primarily include e-commerce reputation and trust; fake information detection and hate speech detection on social media platforms; the application of large language models (LLMs) in cross-border e-commerce, etc. She has published multiple high-impact papers in the fields of Natural Language Processing (NLP) and Information Systems.