PGLHate:A Chinese Hate Speech Detection Framework Integrating Auxiliary Linguistic Representations

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

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

Abstract

The covert dissemination of hate speech on social media poses substantial challenges to content governance and the preservation of a healthy digital environment. Conventional detection methods primarily depend on semantic features and often fail to detect obfuscated expressions, including pinyin substitutions, visually similar characters, and code-mixed Chinese-English text. To address these challenges, this paper proposes PGLHate (Pinyin-Glyph-Lexicon model for Chinese Hate Speech), a detection framework that integrates auxiliary linguistic representations. The framework adopts Chinese RoBERTa as the semantic encoder, supplemented by a semantic enhancement module comprising TextCNN and BiLSTM. It incorporates three types of linguistic features—pinyin, glyph, and lexicon—and applies a gated fusion mechanism to facilitate dynamic cross-modal interaction and integration. Experimental evaluations on two benchmark datasets, COLD and TOXICN, show that the proposed framework surpasses multiple state-of-the-art baselines in terms of both accuracy and F1-score. Additional ablation studies further validate the complementary contributions of each integrated feature in addressing variational and implicit hate speech expressions.

Keywords: Chinese Hate Speech Detection; Auxiliary Linguistic Representations; Pinyin and Glyph Encoding; Semantic Enhancement; Gated Multi-Feature Fusion Mechanism

1. Introduction

With the rapid development of social media, the dissemination of information on online platforms has become increasingly swift and convenient. However, the enhanced accessibility and speed of information has also facilitated the widespread proliferation of hate speech across digital platforms. Hate speech typically refers to expressions that attack or demean individuals or groups based on attributes such as race, ethnicity, religion, disability, gender, age, or sexual orientation/gender identity[1]. Given its detrimental impact on individual dignity and social cohesion, online hate speech has increasingly been acknowledged as a pressing societal issue[2]. Studies have shown that such harmful rhetoric can intensify divisions and conflicts between social groups in the public discourse,

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

and in severe cases, may even threaten the harmony and stability of communities[3]. Consequently, the automatic detection and mitigation of hate speech within social media environments holds considerable practical significance and societal relevance.

Contemporary hate speech detection methods predominantly depend on pre-trained language models (e.g., BERT, RoBERTa) to extract semantic features from textual data and subsequently conduct classification. Although these approaches have demonstrated considerable success in English-language datasets, their effectiveness in Chinese-language scenarios remains limited. On one hand, the Chinese language is highly ideographic and polysemous, making it difficult to accurately detect offensive intent based solely on contextual semantics.On the other hand, hate speech frequently employs “variant expressions” to circumvent platform moderation, introducing semantic ambiguity and substantially increasing the difficulty of detection[4]. Table1 summarizes the common types of such variants in hate speech together with representative examples, which constitute the primary challenges addressed in this study. These variants include substitution with pinyin spelling, replacement with homophones, visually similar character, and code-mixing with English.Variants derived from pinyin information generally fall into two categories. The first involves substituting sensitive words with their pinyin forms, where the offensive term “婊” is replaced by “biao.” The second uses homophones to convey implicit derogatory meanings„ “没叠也没晾,” which alludes to “没爹也没娘.” Variants based on visual similarity transform offensive characters into non-offensive ones by substituting them with visually similar characters, such as replacing “弱智” with “蒻痣.” Another variant is code-mixing with English, where English hate-related terms are embedded in Chinese text, further complicating the detection process.

Variants	**Examples**	**Label**
Substitution with pinyin spelling	什么平权?！实际上，中国女权 **biao** 只想最大化的谋求女性权利最大化。	1
Replacement with homophones	你像那刚洗的衣服，**没叠也没晾**	1
Visually similar character	你真是个**蒻痣**。	1
Code-mixing with English	我一直觉得就算是 gay 亚裔男的地位也很…**potato queen** 感觉和 **easy girl** 差不多…	1

Table 1: Examples of Adversarial Variants in Chinese Hate Speech

In response to these challenges, it is crucial to design a novel model architecture capable of mitigating semantic ambiguity and identifying implicit attack strategies, thereby improving the detection performance of covert hate speech. To this end, we propose a hate speech detection framework named PGLHate, which incorporates auxiliary linguistic representations. The model integrates phonetic cues from Mandarin (pinyin), visual character structures (glyphs), and semantic priors derived from an English offensive lexicon to improve overall recognition accuracy in the context of Chinese hate speech.

The main contributions of this paper are as follows:

1. Development of a Chinese hate speech detection model integrating auxiliary linguistic features: We propose PGLHate, a Chinese hate speech detection framework

specifically designed to tackle covert expression strategies commonly found in Chinese online discourse, including pinyin substitution, visually similar character variants, and code-mixed English usage. The model integrates semantic, phonological, visual, and lexicon-based features to improve its sensitivity to subtle and disguised hate expressions, thereby enhancing recognition accuracy.

2. Development of a semantic enhancement module for capturing multi-granular dependencies: To enhance the model’s capacity for representing textual features, we integrate TextCNN and BiLSTM into the semantic encoding stage. TextCNN is employed to capture local n-gram patterns, while BiLSTM is used to model long-range semantic dependencies. This architecture significantly enhances the expressiveness of core semantic representations and offers a more stable and comprehensive semantic foundation for subsequent classification tasks.
3. Introduction of a gated multi-feature fusion mechanism: The proposed mechanism facilitates fine-grained interactions between the primary semantic modality and auxiliary modalities such as pinyin and glyph representations. It allows the model to dynamically integrate complementary information from multiple modalities, thereby improving its robustness and adaptability to diverse and complex linguistic patterns.

2. Related Work

2.1. Hate Speech Detection

Driven by the rapid progress of deep learning and large-scale pre-trained language models, deep neural networks have been increasingly utilized for hate speech detection. Compared to conventional machine learning approaches, deep neural networks exhibit enhanced capabilities in capturing contextual dependencies and semantic nuances in text. Georgios K. Pitsilis et al.[5] proposed an ensemble classification approach based on Long Short-Term Memory (LSTM) networks to detect hate speech on Twitter. Badjatiya et al.[6] introduced a hybrid strategy combining Convolutional Neural Networks (CNNs), LSTM, and traditional machine learning techniques, which significantly improved performance. Ziqi Zhang et al.[7] explored CNN+GRU and CNN+skipped CNN architectures to address long-tail class imbalance, thereby extending the applicability of deep models to fine-grained classification tasks.

The widespread adoption of pre-trained models has led to extensive use of BERT and its variants in hate speech detection tasks. Saleh Alatawi et al.[8] systematically compared BERT with BiLSTM and LSTM models, demonstrating the substantial benefits of pre-trained semantic representations in improving classification performance. Caselli et al.[9] retrained BERT on Reddit hate speech data to create HateBERT, a model specifically tailored for English hate speech, which outperformed standard BERT across multiple sub-tasks due to its enhanced semantic sensitivity.

However, such approaches generally focus on a single semantic stream and often fall short in detecting implicit hate expressions such as insinuation, sarcasm, and metaphorical language. To address this limitation, recent studies have proposed module-level enhancements through multi-channel semantic modeling. For instance, Plaza-Del-Arco et al.[10] incorporated BiLSTM atop BERT to capture sequential information and integrated sentiment features for improved hate speech detection in Spanish. Shakir Khan et al.[11] introduced BiCHAT, a deep learning model that combines BERT, CNN, BiLSTM, and hierarchical attention mechanisms to jointly learn semantic features at various levels of granularity, achieving state-of-the-art results on several English datasets. Arshad et al.[12] leveraged transfer learning by fusing FastText embeddings with RoBERTa, thereby enhancing semantic understanding in Urdu hate speech scenarios.

In summary, these studies indicate that multi-channel semantic enhancement frameworks have become a prominent trend in the development of hate speech detection models. In line with this trend, we propose a semantic enhancement module that integrates TextCNN and BiLSTM, establishing a robust semantic foundation for subsequent multi-feature fusion.

2.2. The Application of Pinyin and Glyph in NLP

Owing to the unique ideographic characteristics of Chinese characters, pinyin and glyphs have emerged as auxiliary linguistic features that offer phonological and visual cues, thereby garnering increasing attention in natural language processing (NLP) research. Kaiting La et al.[13] proposed Semorph, a pre-trained model based on the morphological and semantic properties of Chinese characters, for spam SMS detection. By incorporating morphological features and data perturbation techniques, their model significantly improved detection performance and robustness against adversarial inputs. Li et al.[14] introduced MFE-NER, a lightweight fusion framework that integrates glyph encodings (e.g., Wubi), pinyin representations, and pre-trained semantic vectors. Experimental results demonstrated notable improvements in robustness against character-level attacks, with minimal additional computational cost, along with modest gains in overall named entity recognition (NER) performance. Jigui Zhao et al.[15] developed CPL-NER, a model that integrates character, pinyin, and lexicon embeddings using multiple embedding and attention, yielding substantial improvements in Chinese NER. Jing Li et al.[16] further advanced Chinese sentiment analysis by applying a matrix-based fusion strategy to deeply integrate glyph and pinyin embeddings, resulting in higher accuracy and improved cross-domain generalizability compared to models that rely solely on semantic representations.

Overall, incorporating pinyin and glyph-based multimodal features has shown considerable promise in strengthening both the expressive power and robustness of models in Chinese NLP applications. In tasks such as text classification, sentiment analysis, and named entity recognition, the fusion of semantic, phonetic, and structural features has proven effective in enhancing model comprehension and discrimination of Chinese text. This research paradigm also provides valuable insights for hate speech detection, indicating that phonological and visual character cues can facilitate the identification of implicitly expressed hate speech. Despite these advances, the joint modeling of pinyin and glyph features for Chinese hate speech detection remains underexplored, especially in capturing subtle and implicit aggression. Therefore, further investigation into the complementary advantages of these modalities and the development of a tailored pinyin-glyph fusion framework represent a promising direction for future research.

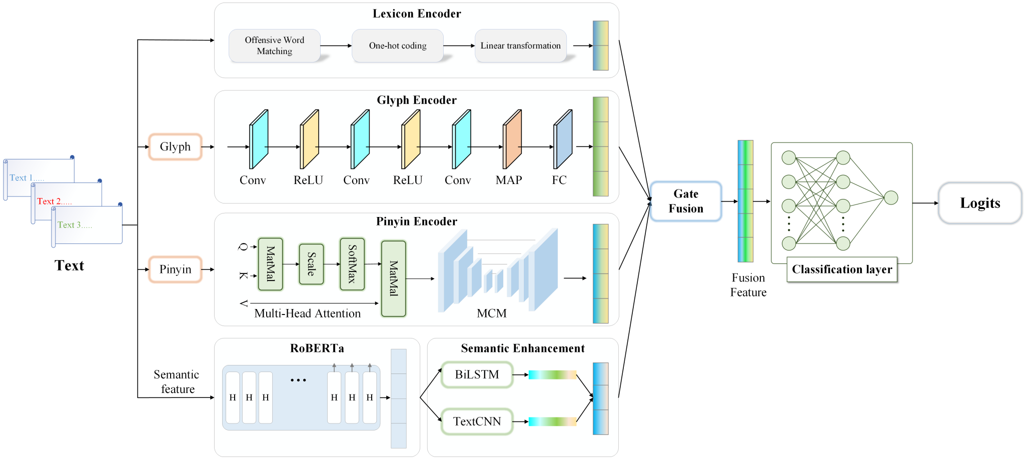

Figure 1: The overall architecture of the PGLHate model

3. Methodology

3.1. Model Overview

The objective of hate speech detection is to determine whether a given text contains offensive, discriminatory, or hostile language targeting specific individuals or social groups. This task is typically formulated as a binary classification problem. Given a Chinese text input $X = \{x_1, x_2, \ldots, x_L\}$,the model is tasked with predicting whether the input contains hate speech and subsequently assigns a binary label $y \in \{0, 1\}$。

This study introduces a Chinese hate speech detection model, PGLHate, which incorporates auxiliary linguistic representations to enhance semantic understanding. The overall architecture of the model is depicted in Figure 1. The model comprises four main components: a semantic feature extraction module, an auxiliary feature modeling module, a gated multi-feature fusion mechanism, and a prediction module. Specifically, the model employs the Chinese pre-trained language model chinese-roberta-wwm-ext as the backbone encoder and integrates TextCNN and BiLSTM in parallel to construct a multi-level semantic enhancement module, thereby effectively capturing contextual semantics, local n-gram patterns, and global dependencies within the text. In parallel, the auxiliary feature modeling branch introduces three complementary modalities: pinyin, glyphs, and an English offensive lexicon, which provide additional phonological, visual, and prior linguistic information, respectively. Pinyin features are derived through pinyin ID mapping, and further encoded using a multi-head attention mechanism coupled with residual connections to obtain phonological representations. Glyph features are generated by converting characters into grayscale images and processed through convolutional and pooling layers to capture visual structural information. Lexicon features are encoded as one-hot vectors based on an offensive word list and projected through a linear transformation to produce sentence-level prior representations. Finally, a gated fusion mechanism is employed to integrate multi-feature information, upon which hate speech classification is performed.

3.2. Semantic Feature Extraction and Enhancement

To effectively capture contextual semantics and potential aggressive expressions in Chinese texts, the model adopts the Chinese pre-trained language model chinese-roberta-wwm-ext as the primary semantic encoder. This module processes the tokenized input to generate contextualized token-level representations:

$$\mathrm{H{=}RoBERTa(x)} \tag{1}$$

On top of the RoBERTa output, the semantic representation is further enriched by integrating two parallel submodules: TextCNN and BiLSTM. Specifically, TextCNN applies multi-scale convolutional filters to capture local n-gram features and outputs a fixed-size vector h_{CNN}, while BiLSTM models long-range dependencies and extracts the final hidden state h_{LSTM}. Additionally, the [CLS] vector $h_{[\mathrm{CLS}]}$ from RoBERTa is used as a global sentence-level embedding. These three components are concatenated to form the final sentence-level semantic representation:

$$h_{\mathrm{sem}} = \left[h_{[\mathrm{CLS}]}; h_{\mathrm{CNN}}; h_{\mathrm{LSTM}}\right] \tag{2}$$

3.3. Auxiliary Features

3.3.1. Pinyin Feature

In Chinese, homophones and pinyin substitutions are commonly utilized as implicit strategies to convey hate speech. To address this challenge, we introduce a pinyin modality to capture phonological features from the text. Each Chinese character is initially mapped to its corresponding pinyin ID, which is then transformed into a dense embedding E_{pinyin}. Contextual representations are extracted via a multi-head attention mechanism, regulated using a matrix combination mechanism (MCM), and subsequently projected into the semantic space through a linear transformation:

$$H_{\mathrm{pinyin}} = W_p \cdot \mathrm{MCM}\left(\mathrm{Attn}\left(E_{\mathrm{pinyin}}\right)\right) + b_p \tag{3}$$

In this formulation, W_p and b_p represent the trainable weight matrix and bias term, respectively. H_{pinyin} denotes the phonological feature embedding enriched with contextual and residual-controlled information. This representation is subsequently integrated with the semantic modality in the feature fusion module, enhancing the model’ s capacity to detect pinyin-level implicit hate speech.

3.3.2. Glyph Feature

Hate speech in Chinese frequently employs deceptive techniques such as visually similar character substitution or character decomposition. To mitigate this, we introduce a character image modality aimed at capturing the structural characteristics of Chinese characters. Each Chinese character is rendered as a grayscale image I_{glyph} , which is fed into a convolutional neural network (CNN) to extract visual structural features. The extracted features are subsequently compressed to a fixed length via pooling and projected into the final glyph representation through a linear transformation:

$$H_{\mathrm{glyph}} = W_g \cdot \mathrm{Pool}\left(\mathrm{CNN}\left(I_{\mathrm{glyph}}\right)\right) + b_g \tag{4}$$

In this formulation,W_g and b_g represent the trainable weight matrix and bias of the linear transformation. The CNN is responsible for capturing local structural patterns in the glyph image, while the pooling layer reduces the feature dimensionality. The final vector H_{glyph} is incorporated into the multi-feature fusion module to strengthen the model' s robustness and ability to discriminate against visually perturbed or camouflaged inputs.

3.3.3. Lexicon Feature

In addition to commonly observed explicit offensive expressions in Chinese, the presence of embedded English hate-related keywords and their frequency of occurrence also serve as important prior indicators for hate speech detection. In Chinese contexts in particular, the inclusion of derogatory English expressions—such as "easy girl" —often implies semantic bias and stigmatizing undertones. Although such phrases may not appear overtly offensive, their latent hostility should not be overlooked. Drawing upon the offensive word lexicon developed by Salminen et al.[17], we construct a one-hot feature vector l to encode the presence of English offensive terms in the input text. This discrete feature is subsequently mapped into a continuous semantic space via a linear transformation and fused with the sentence-level semantic representation:

$$l' = W_l \cdot l + b_l \tag{5}$$

In this formulation, W_l and b_l denote the trainable weight matrix and bias, respectively, while l' represents the embedded vector of the lexicon-based feature.

3.4. Gated Multi-Feature Fusion Mechanism

To facilitate the effective integration of core semantic features and auxiliary representations, we propose a gated multi-feature fusion mechanism that dynamically adjusts the contribution of each modality. At the token level, pinyin features H_{pinyin} and glyph features H_{glyph} aligned with the semantic representation H from the RoBERTa encoder and then fused via a gating mechanism. For a given auxiliary modality F, the fusion is defined as:

$$\widehat{H} = \sigma(W_m[H;F]) \odot H + (1 - \sigma(W_m[H;F])) \odot F \tag{6}$$

Here, $\sigma(\cdot)$ denotes the sigmoid activation function, $\odot$ represents element-wise multiplication, and $[H;F]$ refers to the concatenation of semantic and auxiliary features. This mechanism adaptively adjusts the fusion ratio between semantic and auxiliary information based on contextual cues, thereby improving the model' s ability to selectively integrate informative features.

At the sentence level, a unified gating strategy is applied to combine sentence-level auxiliary representations with the semantic vector h_{sem}. The pinyin and glyph features are pooled into sentence-level vectors l'_{pinyin} and l'_{glyph} , respectively, while Lexicon features are derived from a hate-word dictionary, encoded as one-hot vectors, and projected into sentence-level prior representations l'_{lexicon} via a linear transformation. For any auxiliary sentence-level vector l', the fusion is computed as:

$$h_{\text{fused}} = \sigma\left(W_g[h_{\text{sem}};l']\right) \odot h_{\text{sem}} + \left(1 - \sigma\left(W_g[h_{\text{sem}};l']\right)\right) \odot l' \tag{7}$$

where $[h_{\text{sem}}; l']$ denotes the concatenation of semantic and auxiliary features, and W_g is a learnable fusion weight matrix. This unified gating approach enables context-sensitive and adaptive control of multimodal integration, improving robustness and sensitivity to complex inputs such as code-switching and implicit aggression. Compared with static concatenation or weighted averaging, this fine-grained and learnable fusion mechanism effectively suppresses redundant or noisy information and is structurally extensible to both token- and sentence-level fusion. The resulting fused representation h_{fused} serves as a multi-modal enhanced semantic vector for final hate speech prediction.

3.5. Prediction Module

After multimodal feature fusion, the resulting sentence-level representation h_{fused} is input into a linear classifier to determine whether the text contains hate speech. Specifically, a linear transformation and a bias term are used to compute the final prediction score (logit) as follows:

$$\hat{y} = W_o \cdot h_{\text{fused}} + b_o \tag{8}$$

Here, W_0 and b_o denote the weight matrix and bias term of the linear classification layer, respectively. The output $\hat{y}$represents the model' s predicted confidence (logit) that the input text constitutes hate speech. To convert the logit into a probability, the sigmoid function $\sigma(\cdot)$ is applied. The binary cross-entropy loss with logits is then used to optimize the model parameters, and is defined as:

$$\mathcal{L} = -\left[y \cdot \log \sigma(\hat{y}) + (1 - y) \cdot \log\left(1 - \sigma(\hat{y})\right)\right] \tag{9}$$

where $y \in \{0, 1\}$ denotes the ground-truth label.

4. Experiments

4.1. Dataset and Data Preprocessing

In this study, we adopt the COLD[18] and TOXICN[19] datasets as benchmark corpora for modeling and evaluating hate speech detection. The COLD dataset comprises 37,480 real-world posts collected from social platforms including Zhihu and Weibo, and addresses sensitive topics such as race, gender, and geography. It is annotated using a binary classification scheme to distinguish between offensive and non-offensive content. The training set was constructed using a model-assisted semi-automatic method, whereas the test set was manually annotated with fine-grained labels. The test set is further subdivided into four categories: individual attacks, group attacks, anti-bias, and other non-offensive cases. However, the current study focuses solely on the binary classification of hate speech.

The TOXICN dataset is sourced from two Chinese social media platforms—Zhihu and Tieba—and includes discussions on gender, race, region, and LGBTQ-related topics. It comprises 12,011 Chinese texts, of which 6,461 are labeled as hate speech and 5,550 as normal speech. Descriptive statistics for the COLD and TOXICN datasets are summarized in Table 2.

To ensure data quality and facilitate robust multimodal modeling, a comprehensive preprocessing pipeline was applied to both textual and visual modalities. For textual data, the preprocessing steps involve Unicode normalization, elimination of URLs and

Dataset	Classes	Train	Test	Total
COLD	Hate	15,934	2,107	18,041
	No Hate	16,223	3,216	19,439
	Total	**32,157**	**5,323**	**37,480**
TOXICN	Hate	5,178	1,274	6,464
	No Hate	4,413	1,137	5,550
	Total	**9,600**	**2,411**	**12,011**

Table 2: Dataset splits for COLD and TOXICN

special characters, as well as removal of emojis and kaomoji. For glyph images, all inputs are resized to a standardized shape and converted into tensor format to ensure consistency and compatibility with the model. This preprocessing pipeline significantly reduces noise and lays a robust foundation for subsequent multimodal feature integration.

4.2. Experimental setup

The pre-trained Chinese-RoBERTa-wwm-ext model is employed as the text encoder, with its parameters fine-tuned on the downstream task. The AdamW optimizer is utilized with a learning rate of 1e-5 and a weight decay of 1e-4 to improve generalization. A cosine annealing learning rate scheduler with a 5% linear warm-up is adopted to stabilize gradient updates in the early stages of training. A batch size of 16 is used for both training and validation phases. The model is trained for a maximum of 20 epochs, using the macro-averaged F1 score on the validation set as the primary evaluation metric. The best-performing model checkpoint is retained for final evaluation. An early stopping strategy is also implemented, where training is terminated if no improvement in validation F1 score is observed over five consecutive epochs, to prevent overfitting and reduce redundant computation.

4.3. Baseline Models

To evaluate the effectiveness of the proposed model, we conduct comparative experiments on two benchmark datasets for Chinese hate speech detection: COLD and TOXICN. We compare our model against several representative baselines on each dataset.

On the COLD dataset, we consider the following baselines:

COLDetector[18]: A model built upon a pre-trained Chinese BERT, fine-tuned on the COLDataset, serving as a strong contextual language baseline.

`RB_BG_MHA`[20]: This approach combines RoBERTa with a Bidirectional GRU and a Multi-Head Attention mechanism to capture sequential and contextual dependencies in offensive language.

PSCNN-RoBERTa-wwm-ext[21]: A phonetic-aware model that addresses homophonic perturbations in hate speech using phonetic substitution and a CNN-enhanced RoBERTa.

RMSF[22]: This model introduces Rich Multidimensional Sentiment Features to enhance the discrimination of hate speech.

On the TOXICN dataset, we compare against:

RMSF : As described above.

RoBERTa+TKE[19]: A hybrid framework that integrates RoBERTa with Toxicity Knowledge Enhancement (TKE), leveraging toxic lexicons and linguistic cues for improved detection accuracy.

4.4. Result

To comprehensively evaluate the effectiveness of the proposed PGLHate model, we conduct systematic comparisons with several state-of-the-art models on two Chinese hate speech detection datasets, namely COLD and TOXICN. Evaluation metrics include Accuracy, Precision, Recall, and F1-score, with detailed results summarized in Table 3.

Dataset	**Models**	**Accuracy**	**Precision**	**Recall**	**F1**
COLD	COLDetector	81.00	80.00	82.00	81.00
	RoBERTa	81.61	80.97	82.16	81.22
	RB_BG_MHA	82.93	82.26	83.44	82.84
	PSCNN-RoBERTa-wwm-ext	–	**84.63**	82.79	83.71
	RMSF	–	82.94	82.96	82.85
	PGLHate (ours)	**84.28**	83.62	**84.87**	**83.92**
TOXICN	RoBERTa	–	80.80	80.20	80.30
	RoBERTa+TKE	–	80.90	80.50	80.60
	RMSF	–	82.63	82.41	82.45
	PGLHate (ours)	**82.87**	**82.98**	**82.65**	**82.74**

Table 3: Performance Comparison with State-of-the-Art Models on the COLD and TOXICN Datasets (%)

On the COLD dataset, the proposed PGLHate model achieves the best overall performance, with an accuracy of 84.28% and an F1-score of 83.92%, outperforming all baseline models. Compared to the RoBERTa model (81.61% accuracy, 81.22% F1), PGLHate achieves a 2.7% improvement in F1-score, reflecting its superior overall detection capability. While multimodal models such as PSCNN-RoBERTa-wwm-ext (83.71% F1) and RMSF (82.85% F1) demonstrate competitive performance, they still fall slightly short of PGLHate, highlighting its strength in integrating semantic, structural, and prior knowledge features. PGLHate also achieves a recall of 84.87%, underscoring its robustness in recognizing offensive content.

On the TOXICN dataset, PGLHate likewise maintains its superiority, with an F1-score of 82.74%, surpassing RMSF (82.45%) and RoBERTa+TKE (80.6%). Although the overall performance gap among models on this dataset is relatively small, PGLHate demonstrates a well-balanced performance in both precision (82.98%) and recall (82.65%), indicating strong generalization capability.

In summary, PGLHate consistently yields the highest F1-scores across both datasets, validating its effectiveness and robustness in Chinese hate speech detection. Its high accuracy and recall make it particularly suitable for real-world applications that require comprehensive identification and filtering of Chinese hate speech.

4.5. Ablation experiment

We conducted a series of systematic ablation experiments to investigate the contribution of each modality and structural component to the overall model performance.

The analysis considers four key dimensions: the semantic augmentation module (SA, consisting of TextCNN and BiLSTM), pinyin features, glyph features, and lexicon-based features. The experiments were conducted on the COLD and TOXICN datasets, and the results are summarized in Table 4.

Models	**COLD**				**TOXICN**			
	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1
PGLHate (ours)	**84.28**	**83.62**	84.87	**83.92**	**82.87**	**82.98**	**82.65**	**82.74**
w/o Pinyin	83.81	83.45	**84.94**	83.56	82.50	82.45	82.42	82.43
w/o Glyph	83.40	82.75	84.03	83.03	82.08	82.13	81.90	81.97
w/o Lexicon	83.94	83.37	84.73	83.62	81.96	81.95	81.82	81.87
w/o SA	82.94	82.37	83.67	82.60	82.21	82.28	82.00	82.08
Static Fusion	83.58	82.85	83.96	83.17	82.22	82.30	82.05	82.17

Table 4: Ablation Study Results (%) on the COLD and TOXICN Datasets

The complete PGLHate model achieved the best performance on both datasets (COLD: Accuracy 84.28%, F1 83.92%; TOXICN: Accuracy 82.87%, F1 82.74%). Specifically, removing pinyin features led to a noticeable decline in recall (−0.94% on COLD, −0.23% on TOXICN), highlighting the importance of phonological cues in detecting disguised expressions such as homophone substitutions. Excluding glyph features led to a further decline in F1-score (−0.88% on COLD,−0.77% on TOXICN), demonstrating the complementary role of visual character representations in identifying visually confusing variants. The performance drop resulting from the removal of lexicon-based features was relatively modest but still confirmed the effectiveness of prior knowledge in improving detection. Importantly, the most pronounced performance degradation occurred when the semantic augmentation module was removed, underscoring its essential role in enhancing contextual representations and overall model effectiveness. Finally, replacing the dynamic multi-level attention fusion with static feature concatenation (Static Fusion) degrades performance (−0.75% F1 on COLD and −0.57% F1 on TOXICN), which confirms that adaptive fusion mechanisms are more effective in leveraging diverse feature representations than static fusion strategies.

In summary, the ablation study thoroughly validates the necessity and complementarity of multimodal features in PGLHate, particularly for detecting implicit and variational forms of hate speech.

5. Conclusion and Future Work

This study proposes a multimodal framework for Chinese hate speech detection, named PGLHate, which integrates auxiliary linguistic representations. The model employs RoBERTa for deep semantic encoding, augments sentence-level representations with TextCNN and BiLSTM, and integrates semantic, phonetic (pinyin), visual (glyph), and lexicon-level features. A gated fusion mechanism is adopted to facilitate dynamic interaction and integration among multiple modalities. Experimental results demonstrate that PGLHate achieves strong performance on both the COLD and TOXICN benchmark datasets, outperforming several state-of-the-art models across a range of evaluation metrics. These results confirm its effectiveness in both detection accuracy and feature fusion.

Although PGLHate achieves strong detection performance, it still faces challenges in accurately capturing implicit semantics and addressing diverse forms of hate speech variations, including emoji substitution, punctuation disruptions, and complex combinational transformations. Future research should focus on integrating sentiment and topic modeling, developing robust strategies for variant handling, and leveraging external knowledge to further improve the model’s generalization and adaptability.

Author Contributions

This research was primarily conducted by the first author, Xuan Liu, and the corresponding author, Lirong Chen. Xuan Liu was responsible for formulating the research problem, designing the framework, conducting data collection and analysis, and drafting the manuscript. Lirong Chen provided overall supervision, refined the research objectives, critically revised the manuscript, and secured funding for the project. Qiang Li contributed to the formal analysis of the experimental results and assisted with visualization. Jie Wang supported data curation and validation of the experiments. Rui Lv contributed to software implementation and maintenance of the computing environment. 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] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

[2] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 687–690, 2016.

[3] Matheus Schmitz, Goran Muric, and Keith Burghardt. Quantifying how hateful communities radicalize online users. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 139–146. IEEE, 2022.

[4] Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-wei Lee. Toxicloakcn: Evaluating robustness of offensive language detection in chinese with cloaking perturbations. *arXiv preprint arXiv:2406.12223*, 2024.

[5] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742, 2018.

[6] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760, 2017.

[7] Ziqi Zhang and Lei Luo. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945, 2019.

[8] Hind S Alatawi, Areej M Alhothali, and Kawthar M Moria. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374, 2021.

[9] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*, 2020.

[10] Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, L Alfonso Ureña-López, and María Teresa Martín-Valdivia. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489, 2021.

[11] 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.

[12] Muhammad Umair Arshad, Raza Ali, Mirza Omer Beg, and Waseem Shahzad. Uhated: hate speech detection in urdu language using transfer learning. *Language Resources and Evaluation*, 57(2):713–732, 2023.

[13] Kaiting Lai, Yinong Long, Bowen Wu, Ying Li, and Baoxun Wang. Semorph: A morphology semantic enhanced pre-trained model for chinese spam text detection. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 1003–1013, 2022.

[14] Jiatong Li and Kui Meng. Mfe-ner: multi-feature fusion embedding for chinese named entity recognition. In *China National Conference on Chinese Computational Linguistics*, pages 191–204. Springer, 2024.

[15] Jigui Zhao, Yurong Qian, Shuxiang Hou, Jiayin Chen, Kui Wang, Min Liu, and Aizimaiti Xiaokaiti. Unleashing the power of pinyin: promoting chinese named entity recognition with multiple embedding and attention. *Complex & Intelligent Systems*, 11(1):1–13, 2025.

[16] Jing Li, Dezheng Zhang, Yonghong Xie, Aziguli Wulamu, and Yao Zhang. Gp-fmlnet: A feature matrix learning network enhanced by glyph and phonetic information for chinese sentiment analysis. *CAAI Transactions on Intelligence Technology*, 9(4):960–972, 2024.

[17] Joni Salminen, Fabio Veronesi, Hind Almerekhi, Soon-Gvo Jung, and Bernard J Jansen. Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *2018 Fifth international conference on social networks analysis, management and security (SNAMS)*, pages 88–94. IEEE, 2018.

[18] Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. COLD: A benchmark for Chinese offensive language detection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[19] Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16235–16250, 2023.

[20] Meijia Xu and Shuxian 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*, 13(19):11000, 2023.

[21] Wang Yanhui, Wang Xiaolong, Zhang Shunxiang, Zhou Yuhao, and Wang Caiqin. Chinese hate speech detection method based on the replacement of homophonic noise words. *Applied Science and Technology*, 51(03):72–81, 2024.

[22] Dan Zhiping, Li Lin, Yu Xiaosheng, Lu Yujie, and Li Bitao. A chinese hate speech detection method integrating multi-dimensional sentiment features. *Data Analysis and Knowledge Discovery*, pages 1–15, 2025.

Author Biography

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.

Qiang Li, Lecturer at the School of Computer Science, Inner Mongolia University. He holds a master's degree from the National University of Defense Technology (NUDT). He is dedicated to research in distributed computing and intelligent decision-making

Jie Wang, a master's student in the School of Computer Science, Inner Mongolia University. His research interests focus on various aspects of natural language processing, especially in the area of hate speech identification.

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.

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.