

RESEARCH ARTICLE

Domain-Enhanced Prompt Learning for Chinese Implicit Hate Speech Detection

YAOSHENG ZHANG¹, TIEGANG ZHONG², TINGJUN YI¹, AND HAOMING LI²¹School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China²School of Information Engineering, Henan University of Animal Husbandry and Economy, Zhengzhou 450000, China

Corresponding author: Tiegang Zhong (86621816@qq.com)

ABSTRACT Hate Speech Detection, aims to identify the widespread presence of harmful speech on social networks, is a long-standing research field. Despite its significance, previous efforts almost focused on English, leading to a notable scarcity of datasets for Hate Speech Detection in Chinese. Even more, two emerging forms of hate speech under stringent regulatory environments: 1) domain specificity, manifesting itself as nuanced and harder-to-detect proprietary aggressive rhetoric within various domains; and 2) implicitness, characterized by indirect, abstract and ambiguous cold language. This evolution presents additional complexities for Multi-domain Implicit Hate Speech Detection in Chinese. To fill this gap, we construct a 20,000-large implicit hate speech detection dataset containing nine domains. Furthermore, this research introduce a Domain-enhanced Prompt Learning (DePL) approach, tailored to navigate the complexities of multi-domain and data-limited scenarios. This methodology innovatively combines domain feature fusion to effectively encode domain-specific features in hate speech with the latest advances in prompt learning, effectively tackling the dual challenges of domain diversity and data scarcity. Experimental results demonstrate that the DePL method achieves state-of-the-art (SOTA) results on our benchmark dataset in both few-shot and full-scale scenarios.

INDEX TERMS Hate speech detection, domain feature, prompt learning, few-shot.

I. INTRODUCTION

The spread of social media caused an increase in the amount of online hate speech, which hurts the level of public discussion and can even lead to violence and terrorism [1], [2]. Hate speech, often based on race, gender, religion, or other identity markers, fosters discrimination and hostility toward specific groups, negatively impacting victims' mental health [3] and contributing to social tensions and divisions [4]. Although stringent regulations have curtailed the spread of hate speech, they have also led to a surge in more covert forms of hate speech, termed "implicit hate speech", which are harder to detect and recognize due to their subtle expression [5], [6]. Therefore, developing an effective automated approach to detect hate speech, especially in implicit forms, is of significant societal and research importance [7].

High-quality labeled datasets are essential for tackling the problem of identifying hate speech. In English contexts,

considerable work contributed to hate speech datasets, initially focusing on the binary classification of texts as offensive or non-offensive [1], [8], and later expanding to a multiclass problem at the domain level, such as race and gender [9], [10]. However, the prevalence and severity of implicit hate speech have been underestimated in these studies [11], [12]. Reference [7] pioneered recognizing the importance of implicit hate speech, creating the first dataset of 22,584 English tweets. Compared to English, research on Chinese hate speech detection datasets is laggard, with a marked scarcity of open-source accessible data [10], [13], [14], [15], [16]. Despite being the most widely spoken language worldwide, several research studies have investigated the influence of various natural language processing (NLP) tasks on syntactic structures or semantic characteristics. Meanwhile, implicit hate speech is prevalent on Chinese social media platforms, but it still lacks effective detection techniques, necessitating well-annotated, real-world, large-scale datasets for Chinese implicit hate speech detection.

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera¹.

To fill this gap, we construct a multidomain Chinese implicit hate speech detection (MCIHD) dataset containing about 20,000 texts. This dataset categorizes texts into implicit hate, explicit hate, and no-hate text and extends to a multi-domain focus, encompassing nine domains, including race, gender, religion, region, education, etc. All data in MCIHD are sourced from real-life scenarios across various Chinese social media platforms, ensuring diversity. The data have been carefully examined, filtered, and anonymized, without connection to political material. This dataset is the first to explore Chinese implicit hate speech and is designed to provide a strong basis for the promotion of research in Chinese hate speech detection, particularly for the implicit forms.

Previous methods for the hate speech detection task have evolved through various stages. The early approaches were rule-based, utilizing dictionaries and similarity algorithms [17]. Subsequently, machine learning methods gained prominence, involving the construction and definition of artificial features for training different classifiers, including Naïve Bayes, Support Vector Machines (SVM), Decision Trees, Logistic Regression, etc. [2], [9], [18], [19], [20]. More recently, the advent of deep neural networks introduced methods such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks for hate speech detection [1], [11], [21], [22], [23], [24], [25], [26], [37]. The latest research trend employs pre-trained language models (PLMs) like BERT, which, through self-supervised training on large corpora, achieve deeper linguistic understanding, significantly enhancing the performance of hate speech detection [10], [14], [27], [28], [29], [30], [31], [32], [33]. Despite the progress made, there are still challenges to overcome, such as the scarcity of well-labeled data, the complexity of dealing with implicit hate speech, and the incapability of models to generalize.

To address these, we developed a Domain-Enhanced Prompt Learning (DePL) approach, which integrates an effective domain feature fusion strategy with a prompt learning method suited for low-resource scenarios. Specifically, we set up a domain embedding vector for each category. After processing through a PLM encoder, any input text undergoes domain attention interaction with the domain embeddings via a multi-head attention mechanism, and the results are directly injected into a tailor-made prompt template. This feature fusion strategy focuses on the intricacies of inter-domain differences, combining the broad semantic capabilities of PLM with prompt learning, an efficient few-shot learning. Our proposed DePL method not only effectively addresses hate speech recognition in data-limited scenarios but also demonstrates remarkable robustness in handling hate speech across various domains.

The main contributions of this paper can be summarized as follows:

We constructs the first large-scale Multi-domain Chinese Implicit Hate Speech Detection (MCIHD). Our

dataset encompasses multiple domains and reflects the complexity and diversity of implicit hate speech, significantly contributing to the advancement of research and technological development in Chinese implicit hate speech detection.

This research proposes a method based on domain feature and prompt learning is proposed for the implicit hate speech detection task. This approach, based on few-shot learning, effectively utilizes existing knowledge and task-specific data by incorporating domain knowledge and tailor-made prompts, successfully adapted to data-scarcity scenarios.

Our method exhibits notable superiority in both few-shot and full-scale settings. Our method demonstrates the effectiveness of prompt learning and the integration of domain knowledge by achieving the highest accuracy in extreme few-shot environments and in full-scale settings.

The remainder of this paper is structured as follows. Section II reviews related work on hate speech detection. Section III details the complete construction process and statistical analysis of the MCIHD dataset. Section IV introduces our DePL method for Chinese implicit hate speech, which includes task definition, domain enhancement strategy, and prompt learning method. Section V covers all experiment aspects, including experimental setup, results, ablation studies, and case analyses. Finally, Section VI concludes the paper.

II. RELATED WORK

Over the past few decades, hate speech detection has emerged as a focal research topic in Natural Language Processing (NLP). Tracing its evolutionary trajectory, research in this domain can be broadly segmented into four phases: rule-based methods, machine learning methods, deep learning methods, and pre-trained language model methods.

A. RULE-BASED AND MACHINE LEARNING METHODS

Initial efforts in detecting hate speech predominantly employed rule-based methods. Razavi et al. [17] pioneered a dictionary-based detection system capable of automatically extracting multilayered features, harnessing statistical models and rule-based patterns to match text with classified entries. Many early social media platforms filtered hate speech and offensive content based on rudimentary keyword matching and similarity computations. The subsequent rise of machine learning methods enhanced task performance to some extent. These methods heavily depended on designing feature engineering manually, defining significant textual features for training. Yin et al. [18] were among the first to identify harassment behaviours in online communities, leveraging N-gram, TF-IDF, and semantic features and feeding them into SVM models for classification. Chen et al. [2] employed N-gram combined with textual syntax for detecting varied linguistic assaults and cyberbullying. Kwok et al. [19] applied the Continuous Bag of Words (CBOW) model with the Naive Bayes classification algorithm to detect racist speech. Waseem et al. [20] confirmed the efficacy

TABLE 1. Crawling methods and specific examples.

Crawling Method	Specific Example
Keyword query	Gender discrimination, Educational discrimination, Occupational discrimination 女子座地铁不戴口罩称有病毒的人才戴
Hot event query	Woman on subway claims only those with the virus wear masks 郑州工程用路面钻机男子肇事被击毙 Man causing trouble in Zhengzhou with a road drill was shot dead (微信视频号) #年轻人为什么越来越反感“专家”?
Manual crawling	(WeChat Video) Why are young people increasingly resenting “experts”? (抖音) 男人着急看突发疾病的老人路遇“暴走团”无路可走 (TikTok) A man, in a hurry to see an elderly with sudden illness, encounters a “walking group” and can’t get through

of extracting multigram characteristics and feeding them to logistic regression or SVM for aggressive language detection. Davidson et al. [9] adopted unigram, bigram, and trigram features, as well as TF-IDF features based on LR classifiers, to achieve a three-way classification between hate speech, offensive language, and non-hateful language.

B. DEEP LEARNING METHODS

Despite the progress achieved using traditional methods, feature engineering often proved expensive and machine learning techniques did not adequately grasp the intricate context and semantics underlying texts [36]. With the advent of deep neural networks, many deep learning techniques have been used to enhance the performance of hate speech detection. Djuric et al. [1] tackled hate speech detection in comments on Yahoo Finance by employing neural language models, addressing technical challenges like high dimensionality and sparsity. Gambäck and Sikdar [21] utilized word2vec as word embeddings, subsequently leveraging CNNs for automated feature extraction, achieving a four-way classification of hate speech. Park et al. [22] proposed a two-step approach, discovering that combining character embeddings and word embeddings with a CNN model offers enhanced results. Chen et al. [23] converted one-dimensional features obtained via TF-IDF into two-dimensional features using convolutional neural networks, achieving textual classification for cyberbullying and hate speech. Badjatiya et al. [11], working on a hate speech dataset from Twitter, employed a two-phase training mechanism, comparing various embedding techniques and neural network models, discovering that the combination of LSTM with random embeddings and the Gradient-Boosted Decision Tree classifier yielded optimal results. During SemEval2019 Task 6, Zhang et al. [24] employed Glove pre-trained word vectors combined with bidirectional LSTM and attention mechanisms, achieving promising results. Kapoor et al. [25] applied transfer learning to build an LSTM-based classification model for hate speech detection in both Hindi and English. Rajamanickam et al. [26]

proposed a multitask learning framework, recognizing the association between user sentiments and abusive behaviours, sharing parameters for joint learning of emotional features, and employing bidirectional LSTM coupled with attention mechanisms as classifiers.

C. METHODS BASED ON PRE-TRAINED LANGUAGE MODELS

Recently, fine-tuning PLMs has emerged as a mainstream approach in NLP. PLMs for specific tasks rather than training models from scratch can significantly enhance performance in downstream tasks [27]. Hate speech detection is no exception. In SemEval2019 Task 6 (Sub-task A), Liu et al. [28] utilized the BERT model and achieved outstanding results with minimal fine-tuning, claiming the top position without intricate pre-processing. This underscored the robust impact of pre-trained language models on downstream tasks. Mozafari et al. [29] explored this avenue, creating a fine-tuned model with BERT, and verified its efficacy in multi-category hate speech detection across two Twitter datasets. Dai et al. [30] built upon Liu et al.’s work, constructing a multi-task learning framework and further elevating performance on the same datasets. Sohnd et al. [31] proposed a multi-channel model based on BERT and investigated the utility of translation as a supplementary input, validating its effectiveness on datasets in three distinct languages. Recognizing the distinctiveness of the Chinese language and culture, Deng et al. [10] created a benchmark dataset, COLD, for Chinese offensive language analysis to address the lack of domestic datasets. In addition, research efforts have been made to fuse PLMs with intricate and novel deep-learning structures. Kim et al. [32] introduced a novel contrastive learning method, employing potent data augmentation strategies, conducting experiments across and intra-dataset, and consistently outperforming BERT and HateBERT. Miao et al. [33], leveraging the Graph Attention Network and using BERT for text representations, considered the structure of users’ social networks, constructing an end-to-end model for abusive language detection.

TABLE 2. Detailed statistical charts of each field in the MCICD dataset.

Label	Gender	Country	Race	Region	Character	Other	Profession	Religion	Education	Total
Not Hate	-	-	-	-	-	-	-	-	-	12037
Implicit Hate	1033	703	581	390	415	353	400	419	132	4426
Explicit Hate	458	637	331	330	269	269	164	104	64	2626
Total	1494	1341	914	723	686	625	564	523	195	19089

III. DATASET CONSTRUCTION

This section details the construction process of the MCIHD dataset, which comprises about 20,000 sentences, focusing on the prevalent phenomenon of implicit hate speech and its multi-domain aspects. To collect high-quality textual resources, we extracted a substantial amount of raw corpus from various Chinese social media platforms. This corpus underwent rigorous cleaning, filtering, and annotation to form a well-annotated benchmark dataset dedicated to studying Chinese implicit hate speech. In this work, we elaborate on the collection, filtering, and annotation process of the dataset below.

A. CORPUS COLLECTION AND FILTERING

An extensive investigation of the discourses on Chinese Internet platforms revealed that implicit hate speech is a common phenomenon across various types of Chinese social media applications. Due to the strict rules against hate speech on most domestic platforms, explicit hate speech is rare, with most hostile comments being expressed in more subtle ways. This impact our data collection, leading us to source data from multiple platforms, including Weibo,¹ Zhihu,² WeChat,³ and Xiaohongshu,⁴ to ensure a comprehensive and diverse corpus, free from the constraints of individual platform moderation.

Hate speech often manifests itself in concentrated distributions, particularly around specific topics or events that trigger extensive, aggressive discourse. Adopting the methodology proposed by elsherief2021latent, we employed keyword and event hotspot queries for data collection. Additionally, this work discovered that certain video websites, due to lax moderation, harbour numerous abusive comments under misleading videos. Although sparsely distributed, these sources were manually supplemented to our data collection. Our overall corpus collection strategy encompassed three methods:

1) KEYWORD QUERY

In this investigation, we have extensively collected keywords with aggressive or hateful connotations. These keywords were chosen for their representativeness, conciseness, timeliness, information density, and high association with related terms, ensuring comprehensive coverage of trending topics

¹www.weibo.com

²www.zhihu.com

³www.wechat.com

⁴https://www.xiaohongshu.com

and themes. Keyword queries on various platforms allowed us to rapidly accumulate a large volume of hate speech, thus expanding and enriching our corpus. Terms like “女拳”, “东百人”, and “臭婊子”, known for their overt aggressiveness, were included. Furthermore, we identified certain terms that become sources of aggressive discourse in specific contexts or scenes, such as “佛祖” and “圣母”. Our keyword list encompassed these terms and the retrieved corpus was meticulously filtered.

2) HOT TOPIC QUERY

Hot topics, often centred on specific events, issues, or entities, are characterized by timeliness and reliability. Discussions on these topics frequently generate substantial hate or aggressive speech. We selected 195 hot topics from 2022 to 2023 based on user participation and likes on Weibo topics, gathering large-scale discourse from these discussions. Examples include “专家称年轻人工资低可能是能力不够” and “唐山烧烤店打人事件”.

3) MANUAL COLLECTION

While keyword and hot topic queries expanded our corpus, an imbalance was observed in the distribution of hate labels. Investigation revealed that niche social platforms, due to fewer members and lenient moderation, contained abundant hate speech, such as abusive comments in video website sections. Given the dispersed distribution, this research supplemented our dataset through manual collection, focusing on sections lacking in hate speech categories. This manual approach balanced the distribution of the dataset and enhanced its quality. Specific examples of these three collection methods are shown in Table 1.

B. DATA ANALYSIS

The finalized MCIHD dataset, after selection, filtering, and annotation, comprises 19,089 texts, encompassing implicit hate, explicit hate, and non-hate speech across multiple domains. Table 2 and Figure 1 illustrate the distribution of each label and domain within the dataset. In addition, a detailed analysis of each category is conducted to gain a deeper understanding of the dataset.

Within the realm of implicit hate speech, gender (1,033 instances), nationality (703 instances), and race (581 instances) emerge as the most prevalent domains. Additionally, occupation (400 instances), religion (419 instances) and region (390 instances) represent significant areas of focus. This diversity reflects the wide-ranging content implicated in

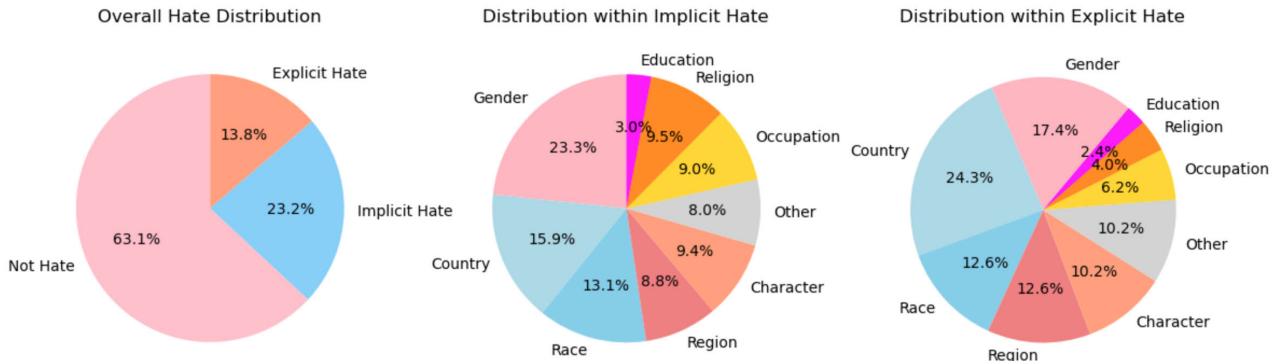


FIGURE 1. Data label category distribution statistics.

implicit hate speech, covering various aspects of social life. In contrast, for explicit hate speech, gender (458 instances) and nationality (637 instances) remain predominant, though with reduced frequency. The region (330 instances) and the race (331 instances) also commonly feature in explicit hate speech, indicating a potential trend in public discourse where explicit hate speech less frequently involves domains such as occupation and religion.

Furthermore, the dataset reveals that gender-based hate speech is the most numerous, totaling 1,494 instances, followed by nationality-based hate speech at 1,341 instances. These categories considerably outnumber others, possibly reflecting a broader trend in online environments. In particular, the total of instances of explicit hate speech (2,626) is fewer than that of implicit hate speech (4,426). This disparity suggests that there is a tendency in online settings for individuals to express hate sentiments implicitly rather than through direct attacks or insults.

In general, our data set mirrors the diversity and complexity of hate speech online, encompassing both explicit and implicit forms and spanning various domain categories. This provides a rich resource for studying online hate speech and facilitates designing and comparing different detection methodologies. In the next phase of method development, this research plans to leverage this extensive domain information to enhance model performance and enable effective handling of various types of hate speech.

C. METHOD

D. OVERVIEW OF PROMPT LEARNING

In this section, we will discuss the basics of prompt learning. Prompt learning involves constructing natural language text (i.e., a template) and feeding it into the encoder of a Masked Language Model (MLM), thereby transforming classification problems into cloze tasks [38]. A prompt learning model consists of a template and a verbalizer. Specifically, for an input text $X = \{x_1, x_2, \dots, x_n\}$ and its corresponding domain d , a textual sequence $T(X, d)$ is created using a custom template $T(\cdot, \cdot)$, as illustrated in Table 3. The template must include at least one [MASK] token. Assuming

\mathcal{M} is an MLM, \mathcal{M} outputs the probability of each word in \mathcal{M} 's vocabulary \mathcal{V} being filled in the [MASK] token $P_{\mathcal{M}}([\text{MASK}] = v | T(o, c)) | v \in \mathcal{V}$.

The process of mapping the predicted masked word to a label indicating whether o and c are synonymous is known as verbalization. This process converts the masked prediction task into a classification task. The configuration of the verbalizer significantly influences the performance of prompt learning. The verbalizer can be defined as $f : \mathcal{V}_y \mapsto \mathcal{Y}$, where \mathcal{V}_y is a carefully designed set of label words, a subset of the overall vocabulary \mathcal{V} , and \mathcal{Y} represents corresponding output labels (e.g., $y_0 : 0$ and $y_1 : 1$ for the three-class labels of hate speech). The final probability output of the classification model can be formalized as:

$$y_p = \text{Classifier}(T(X, d)) \tag{1}$$

In our task, our goal is to train our model using the given template and verbalizer to maximize the prediction accuracy between y_p and y .

E. DOMAIN-ENHANCED PROMPT LEARNING METHOD

In this section, we present our proposed method called Domain-Enhanced Prompt Learning (DePL). The overall architecture of DePL is depicted in Figure 2. The DePL method aims to fully leverage domain knowledge to enhance the performance of hate speech detection, especially in data-sparse scenarios. Research has shown that domain features are highly effective for text classification tasks, including hate speech detection [9], [10]. We found that domain representations, which capture vocabulary patterns and contexts peculiar to specific domains, significantly boost hate speech detection capabilities.

To achieve this, we designed a set of domain vectors $\mathbf{V} = \{v_1, v_2, \dots, v_{10}\}$, with each domain vector v_d corresponding to a specific domain, totaling nine domains. During model training, for each input text, the model first determines the domain $X_i \rightarrow d_i$ of the text. Then, the corresponding domain vector is selected and fed into the model, participating in computation and optimization processes alongside the text representation. Notably, an attention mechanism is employed

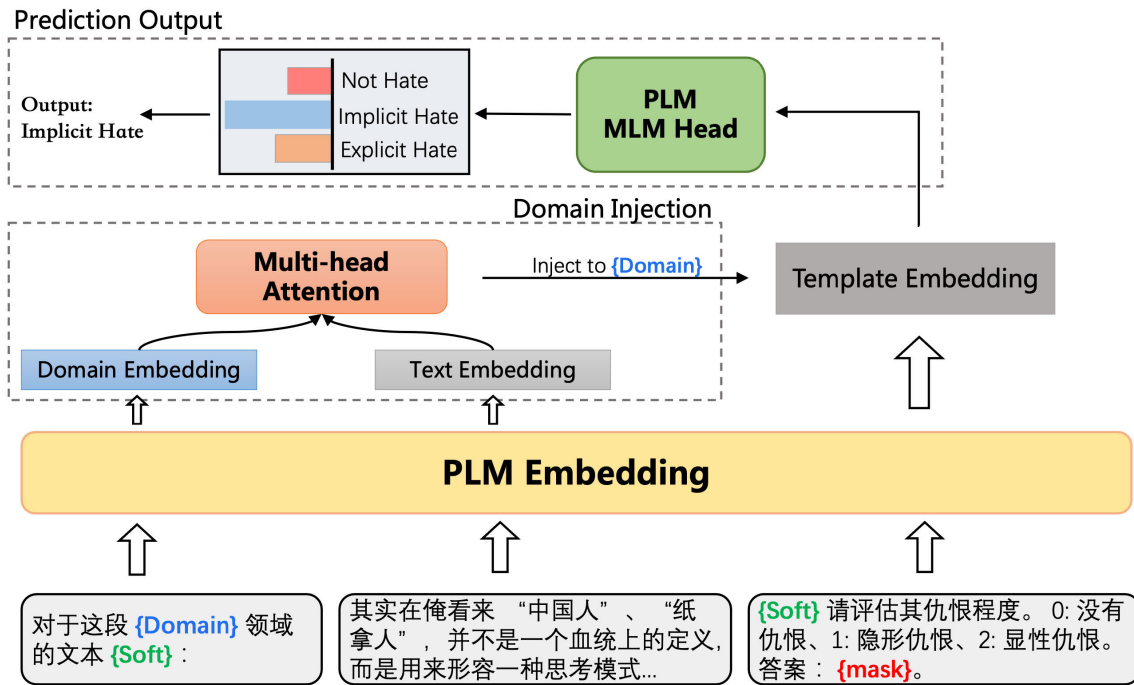


FIGURE 2. The overall architecture of our proposed domain-enhanced prompt learning method for chinese implicit hate speech detection task.

to fine-tune the interaction between the domain vector and input text, enabling the model to focus more on text sections relevant to a specific domain. Formally, the attention interaction between the domain vector and input text is represented as:

$$v'_d = \text{Attention}(v_d, \mathbf{x}), \quad (2)$$

where \mathbf{x} is the representation of the input text, Attention is the attention function, and v'_d is the domain vector adjusted by attention.

Subsequently, the attention-adjusted domain vector v'_d is directly injected into a specific position of the template $T(o, c)$. In our model, v'_d is inserted into the template as a special token, such as “This [DOMAIN] post is: [MASK]”, where [DOMAIN] is the domain marker. Input the original text entity o , candidate entity set C , and domain information d into the template, forming a new text sequence $T(o, c, d)$. The model learns the optimal representation of domain information by maximizing $\mathcal{PM}([\text{MASK}] = v \mid T(o, c, d)) \mid v \in \mathcal{V}$.

By integrating domain knowledge, the model better understands the context within the text and improves its ability to detect hate speech in specific domains. This method not only provides a straightforward and intuitive way to integrate domain information, but also allows the model to learn distinct characteristics across domains, as the domain vectors v_d are trainable.

F. TEMPLATE CONSTRUCTION

The construction of templates is a key factor that affects the performance of prompt learning methods [34], [35].

To maximize the classification capabilities of PLMs, we designed a custom template that integrates domain knowledge, effectively consolidating this knowledge within the template. Additionally, we experimented with intuitive manual templates and auto-learnable mixed templates for comparison, detailed in Table 3.

Manual templates, the most straightforward type, contain only human-readable natural text. They utilize hard coding, where token embeddings in the template sequence are fixed to their corresponding vector representations. On the other hand, mixed templates combine hard coding with soft coding, encoding tokens as dynamic, learnable word vectors. For example, the placeholders “soft:” and “soft:请” in Table 3 represent soft tokens. The embedding of the latter token is initialized to the encoding vector corresponding to “please”, while the former is randomly initialized. The embeddings of these soft tokens are optimized through gradient calculation during training. Soft tokens offer more flexibility and adaptability than hard code, guiding the model to learn task-relevant critical information and supporting more accurate classification. Studies have shown that including a special learnable token in the template makes prompt learning more effective li2021prefix, han2022ptr, liu2023xpre. Therefore, incorporating soft tokens in both the mixed and domain knowledge templates to enhance model performance.

As shown in the first row of Table 3, the domain knowledge template is a novel template we proposed based on the mixed template with added domain knowledge. Specifically, in this research insert a special domain marker “DOMAIN” into the template, whose embedding is replaced by the

TABLE 3. The three different template settings used in our prompt learning method correspond to manual templates, mixed template and domain-enhanced template respectively.

Template	Content
Domain-enhanced Template	对于这段{Domain}领域的文本{soft}: “{text}”,{soft}请评估其仇恨程度, 0: 没有仇恨、1: 隐性仇恨、2: 显性仇恨。答案: {mask}。 For the text in {Domain} field: “{text}”, {soft} please evaluate the level of hatred, 0: no hate, 1: implicit hate, 2: explicit hate. Answer: {mask}
Mixed Template	对于这段文本{soft}: “{text}”,{soft}{soft:请}评估其仇恨程度, 0: 没有仇恨、1: 隐性仇恨、2: 显性仇恨。答案: {mask}。 For the text: “{text}”, {soft} {soft:please} evaluate the level of hatred, 0: no hate, 1: implicit hate, 2: explicit hate. Answer: {mask}
Manual Template	对于这段文本: “{text}”,请评估其仇恨程度, 0: 没有仇恨、1: 隐性仇恨、2: 显性仇恨。答案: {mask}。 For the text: “{text}”, please evaluate the level of hatred, 0: no hate, 1: implicit hate, 2: explicit hate. Answer: {mask}

corresponding domain vector. This method directly integrates domain features into the template, enabling the model to consider domain characteristics in hate speech detection, especially useful in data-sparse domains.

G. MAPPING OUTPUT

A significant challenge in prompt learning for classification tasks is to extract effective classification decisions from prediction results. To address this, in this research, a template design strategy with option suffixes was proposed, which enables the model to understand the classification decision process more naturally. As shown in Table 3, each template explicitly presents the labels and meanings of each category at its end. The advantage of this design is that it does not require the model to comprehend abstract category labels; instead, it guides the model to make a selection directly from options 0, 1, and 2, leading to the corresponding classification decision.

Consequently, our verbalizer design is also intuitive and concise, with labels 0, 1, and 2 corresponding to “non-hate,” “implicit hate,” and “explicit hate” labels, respectively. In practice, the model output is the probability distribution of each word. Using the verbalizer, we can directly map the output probability distribution to the output class probability distribution of the classification task. Specifically, if the predicted word by the model matches a label word, the probability of that word is considered to be the probability of the corresponding label. This design helps to reduce the model’s difficulty in understanding abstract category labels, interpreting results that are straightforward and understandable.

Overall, this design approach offers an efficient mapping output strategy, directly mapping the model’s predicted vocabulary to category labels, thereby enhancing the model’s ability to perform classification tasks. This design enhances the effectiveness of the model and renders the interpretation of outputs more straightforward and understandable.

IV. EXPERIMENTAL ANALYSIS

This section provides a detailed analysis of the performance of various baseline methods we used, along with extensive experiments and comparative studies of our proposed model. In this investigation, we evaluated the performance of each technique on different hate speech detection tasks using the macro F1 score as our metrics. Furthermore, we compared

the effectiveness of different templates to highlight the superiority of our proposed model.

A. BASELINE METHODS

For our experiments, we selected the following five standard text classification techniques as baselines for comparison:

- **TF-IDF:** A vectorization method based on term frequency and inverse document frequency. It transforms text into high-dimensional vectors, assigning word weights by calculating the product of the term frequency and the inverse frequency of the document. Despite its simplicity and efficiency in text classification tasks, its inability to capture word order and context might limit its performance in hate speech detection.
- **SVM(Word2vec):** This method combines support vector machines with word embeddings. Specifically, they first generate pre-trained word vectors for each word using Word2vec, then obtain a textual representation through averaging or weighting with TF-IDF, followed by classification via a support vector machine. Although this method captures semantic information from words, it may fail to capture complex sentence-level context.
- **LSTM:** Long Short-Term Memory networks, a specific type of recurrent neural network, effectively handle sequence data. In our approach, each text sequence is fed into the LSTM and its final hidden state is used for classification. While LSTM can capture long-distance dependencies in the text, its sequential nature might render it less efficient for large-scale textual data.
- **TextCNN:** The Convolutional Neural Network for text is an efficient text classification technique. TextCNN extracts local features in the text using one-dimensional convolution and identifies the most salient features through max-pooling. Although TextCNN boasts computational efficiency, it might struggle to capture long-range text dependencies.
- **BERT-CLS:** BERT is a pre-trained deep bidirectional Transformer that has shown state-of-the-art performance in various NLP tasks. In our approach, text is fed into the BERT model, and the first position output (i.e., the CLS token) is used as a text representation for classification. Given its capability to capture deep semantic information, BERT may exhibit exemplary performance in our hate speech detection task. However, its substantial number of parameters might require

TABLE 4. Final experimental results of our model and baseline models on our MCIRD dataset.

Model	Accuracy				
	16-shot	64-shot	256-shot	1024-shot	All
TF-IDF	10.21	50.52	60.05	70.11	72.34
SVM(Word2vec)	12.55	55.37	63.28	72.04	74.23
LSTM	15.43	60.75	68.27	73.84	75.94
TextCNN	14.67	61.57	69.43	74.65	76.85
BERT-CLS	15.06	70.36	75.54	78.56	80.07
PL-Manual	16.83	77.40	77.91	79.82	81.82
PL-Mixed	15.47	78.27	78.85	80.07	82.04
PL-Domain	15.29	79.31	80.23	81.96	83.32

TABLE 5. Comparison of ablation results. “w/o domain” means to remove domain information from the PL-Domain model.

Model	16-shot F1	64-shot F1	256-shot F1	1024-shot F1	All(3500) F1
w/o domain	15.35	78.45	79.03	80.22	82.43
PL-Domain	15.29	79.31	80.23	81.96	83.32

significant computational resources and time for training and inference.

It’s noteworthy that while these baseline methods each have their pros and cons, they serve as crucial benchmarks against which our approach is evaluated. In the following experiments, we will investigate further their effectiveness in detecting hate speech.

B. EXPERIMENTAL SETUP

Our experiments used the BERT-based pre-trained language model,⁵ which is pre-trained in a wide range of Chinese texts. Adam optimizer uses a learning rate of 1e-5. During the training process, we ran 10 phases, with a total size of 32, and a maximum sequence length of 256 for the model input. To avoid overfitting, the probability of abandonment is set at 0.3. After each time, the model is evaluated in a validation set and the best-performing model is chosen for the final test of the test set. To evaluate the performance of models in low-data scenarios, we conducted a few shot experiments with 16, 64, 256 and 1024 training samples, imitating real-world low-data challenges. Our work also used an early suspension; training was suspended if there were no significant improvements in model performance during the validation period for two consecutive periods. To ensure the reliability of the experimental results, we conducted five independent tests in different random seed samples and averaged the results in these tests. This approach reduces the effects of randomness on experimental results and leads to more reliable evaluations.

C. MAIN RESULTS

Prompt-based learning models exhibit significant advantages in few-shot learning. Across all few-shot settings,

⁵<https://huggingface.co/bert-base-chinese>

including 16-shot, 64-shot, 256-shot, and 1024-shot scenarios, our prompt-based learning models (i.e., PL-Manual, PL-Mixed, and PL-Domain) consistently outperformed the baseline methods in accuracy. These observations validate our hypothesis that prompt-based learning approaches can harness the model’s prior knowledge and task-specific information more effectively in data-scarce scenarios, thus enhancing performance.

Manual prompts excel in extreme low-data scenarios.

In the 16-shot setting with only 16 samples, the PL-Manual model achieved the highest accuracy among all models. This suggests that manually designed prompts can more effectively guide the model to leverage its inherent semantic understanding, mainly when training samples are extremely limited.

Domain-enhanced prompt-based models show increasing superiority as the sample size increases. The PL-Domain model consistently displayed the highest accuracy in scenarios with more samples, specifically the 64-shot, 256-shot, 1024-shot, and full-sample scenarios. This observation supports our conjecture that injecting domain knowledge into prompts can improve model performance.

Prompt-based learning models maintain their edge in full-sample training. Even when trained on the entire dataset (comprising 3,500 samples), our three prompt-based learning models outperformed baselines, with the PL-Domain model exhibiting the best performance. This further confirms the efficacy of the prompt-based learning approach and the value of embedding domain knowledge within prompts.

Though BERT-based classification excels in large-sample scenarios, it underperforms in a few-shot setting. The BERT-CLS model achieved commendable accuracy when trained on the full dataset but lagged considerably behind our prompt-based learning models in few-shot scenarios. This could be attributed to BERT’s inability to fully learn task-relevant key information when data is scarce, despite its effective learning capabilities with ample data.

D. ABLATION STUDY

The ablation study robustly validates the effectiveness of our strategy to incorporate domain knowledge. In this

TABLE 6. Case study of our model's classification results.

Case	Domain	Hate Level
这些穆罕默德们与西方社会之间已经渗透到了什么程度？一个例子是... How deeply have these Muhammads infiltrated Western society? An example is...	宗教 Religion	隐性仇恨 Implicit Hate
妈宝女一个，格局没有，哪来的优越感 A "mama's girl", with no vision. Where does the superiority come from?	性别 Gender	隐性仇恨 Implicit Hate
劣等民族是没资格活着社会上的 Inferior races don't deserve to exist in society.	种族 Race	显性仇恨 Explicit Hate

experiment, we devised a version of the PL-Domain model stripped of domain information, termed the "w/o domain" model, to specifically gauge the contribution of domain knowledge to model performance. As shown in Table 5, the PL-Domain model outperformed the "w/o domain" model in 64-shot, 256-shot, 1024-shot, and full-sample training scenarios. This difference was particularly pronounced in the full-sample scenario, underscoring the value of our domain knowledge integration approach. However, in the 16-shot scenario, the PL-Domain model's accuracy was slightly inferior to that of the "w/o domain" model, potentially because of the model's limited capacity to learn and utilize domain knowledge with such few samples or due to performance fluctuations induced by data scarcity. However, even without domain information support, the "w/o domain" model outperformed the baseline models when trained on the full dataset, further attesting to the superiority of our prompt learning approach. In general, the ablation study indicates that introducing domain knowledge plays a vital role in our task, helping to enhance model performance. Though minor performance fluctuations might occur in few-shot scenarios, they don't undermine the overall efficacy of the domain knowledge introduction strategy.

E. CASE STUDY

In the three cases presented in Table 6, our model successfully classified implicit and explicit hate speech in various domains. For implicit hate speech relating to religion and gender, even when the aggressive expressions weren't direct or overt, the model could still accurately identify and classify them as implicit hate, highlighting its sensitivity and capability in handling nuanced expressions. Meanwhile, for explicit hate speech in the racial domain, the model could accurately label them as explicit hate, showcasing its robustness and versatility. These cases solidly attest to the practicality and effectiveness of our model in tackling complex hate speech classification tasks, especially when dealing with multi-domain Chinese texts involving both implicit and explicit hate, where the model's performance is particularly commendable.

V. CONCLUSION AND FUTURE WORK

This research introduced an innovative approach to detecting implicit hate speech in Chinese through a unique combination of PLMs and prompt learning methods. In particular, our model excelled in domain adaptability and detecting nuanced hate speech, outperforming established baselines. This work underscores the utility of domain features, as evidenced by a drop in performance when these are excluded, highlighting their integral role in model effectiveness.

Looking ahead, our focus will be on optimizing the balance between model complexity and efficiency, catering to diverse computational environments. In addition, we plan to explore alternative PLMs and integrate advanced features, with the aim of refining our approach for more complex hate speech detection tasks. This study contributes significantly to the field of hate speech detection and lays the foundation for future research in this critical area.

REFERENCES

- [1] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 29–30.
- [2] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *Proc. Int. Conf. Privacy, Secur., Risk Trust Int. Conf. Social Comput.*, Sep. 2012, pp. 71–80.
- [3] B. M. Tynes, M. T. Giang, D. R. Williams, and G. N. Thompson, "Online racial discrimination and psychological adjustment among adolescents," *J. Adolescent Health*, vol. 43, no. 6, pp. 565–569, Dec. 2008.
- [4] M. L. Williams and P. Burnap, "Cyberhate on social media in the aftermath of woolwich: A case study in computational criminology and big data," *Brit. J. Criminol.*, vol. 56, no. 2, pp. 211–238, Mar. 2016.
- [5] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proc. 2nd Workshop Lang. Social Media*, 2012, pp. 19–26.
- [6] H. M. Saleem, K. P. Dillon, S. Benesch, and D. Ruths, "A web of hate: Tackling hateful speech in online social spaces," in *Proc. Workshop Programme*, 2016, p. 1.
- [7] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang, "Latent hatred: A benchmark for understanding implicit hate speech," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 345–363.
- [8] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 145–153.
- [9] T. Davidson, D. Warmusley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 11, no. 1, 2017, pp. 512–515.

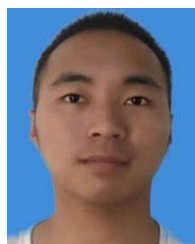
- [10] J. Deng, J. Zhou, H. Sun, C. Zheng, F. Mi, H. Meng, and M. Huang, "COLL: A benchmark for Chinese offensive language detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 11580–11599.
- [11] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion*, Kuala Lumpur, Malaysia, 2017, pp. 759–760.
- [12] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop Natural Lang. Process. Social Media*, 2017, pp. 1–10.
- [13] Linbin Liu, "Research on methods of identifying hate speech in social networks," M.S. thesis, Univ. Electron. Sci. Technol. China, Chengdu, China, 2022.
- [14] Y. Li and Y. Zhao, "A method for identifying offensive speech in social networks combining Bi-LSTM and VDCNN," *J. Fuzhou Univ.*, vol. 36, pp. 76–83, Jan. 2022.
- [15] A. Jiang, X. Yang, Y. Liu, and A. Zubiaga, "SWSR: A Chinese dataset and lexicon for online sexism detection," *Online Social Netw. Media*, vol. 27, Jan. 2022, Art. no. 100182.
- [16] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, Aug. 2023, Art. no. 126232.
- [17] A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, "Offensive language detection using multi-level classification," in *Proc. Can. Conf. Artif. Intell.*, Ottawa, ON, Canada. Cham, Switzerland: Springer, May 2010, pp. 16–27.
- [18] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kostothatis, and L. Edwards, "Detection of harassment on web 2.0," in *Proc. Content Anal. WEB*, 2009, pp. 1–7.
- [19] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013.
- [20] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Res. Workshop*, 2016, pp. 88–93.
- [21] B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in *Proc. 1st Workshop Abusive Lang. Online*, 2017, pp. 85–90.
- [22] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on Twitter," in *Proc. 1st Workshop Abusive Lang. Online*, 2017, pp. 41–45.
- [23] J. Chen, S. Yan, and K.-C. Wong, "Verbal aggression detection on Twitter comments: Convolutional neural network for short-text sentiment analysis," *Neural Comput. Appl.*, vol. 32, no. 15, pp. 10809–10818, Aug. 2020.
- [24] Y. Zhang, B. Xu, and T. Zhao, "CN-HIT-MLT at SemEval-2019 task 6: Offensive language identification based on BiLSTM with double attention," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 564–570.
- [25] R. Kapoor, Y. Kumar, K. Rajput, R. R. Shah, P. Kumaraguru, and R. Zimmermann, "Mind your language: Abuse and offense detection for code-switched languages," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 9951–9952.
- [26] S. Rajamanickam, P. Mishra, H. Yannakoudakis, and E. Shutova, "Joint modelling of emotion and abusive language detection," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4270–4279.
- [27] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Sci. China Technol. Sci.*, vol. 63, no. 10, pp. 1872–1897, 2020.
- [28] P. Liu, W. Li, and L. Zou, "NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 87–91.
- [29] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," in *Proc. Int. Conf. Complex Netw. Their Appl.* Cham, Switzerland: Springer, 2019, pp. 928–940.
- [30] W. Dai, T. Yu, Z. Liu, and P. Fung, "Kungfupanda at SemEval-2020 task 12: BERT-based multi-TaskLearning for offensive language detection," in *Proc. 14th Workshop Semantic Eval.*, 2020, pp. 2060–2066.
- [31] H. Sohn and H. Lee, "MC-BERT4HATE: Hate speech detection using multi-channel BERT for different languages and translations," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2019, pp. 551–559.
- [32] Y. Kim, S. Park, and Y.-S. Han, "Generalizable implicit hate speech detection using contrastive learning," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 6667–6679.
- [33] Z. Miao, X. Chen, H. Wang, R. Tang, Z. Yang, and W. Tang, "Detecting offensive language on social networks: An End-to-end detection method based on graph attention networks," 2022, *arXiv:2203.02123*.
- [34] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Sep. 2023.
- [35] X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun, "PTR: Prompt tuning with rules for text classification," *AI Open*, vol. 3, pp. 182–192, Jan. 2022.
- [36] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015.
- [37] Y. Ding, "Research on multi-granularity text classification of hate speech and abusive language based on RNN," M.S. thesis, Yunnan Univ., Kunming, China, 2020.
- [38] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 4582–4597.



YAOSHENG ZHANG is currently pursuing the master's degree with the School of Electronic and Information Engineering, Liaoning Technical University. His research interest includes natural language processing.



TIEGANG ZHONG received the B.S. degree in electronic science and technology and the M.S. and Ph.D. degrees in microelectronics and solid-state electronics from Jilin University, China. He is currently an Associate Professor with the School of Electronic and Information Engineering, Liaoning Technical University, China. His current research interests include natural language processing, RF microwave circuits, and semiconductor sensors.



TINGJUN YI is currently pursuing the master's degree with the School of Electronic and Information Engineering, Liaoning Technical University. His research interest includes deep learning object detection.



HAOMING LI is currently pursuing the bachelor's degree with the School of Information Engineering, Henan University of Animal Husbandry and Economy. His research interest includes natural language processing.

...