Received: 8 May 2025 | Revised: 24 September 2025 | Accepted: 15 October 2025 | Published online: 12 December 2025

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2026, Vol. 5(2) 222–232
DOI: 10.47852/bonviewJCCE52026117

BON VIEW PUBLISHING

Chinese Toxic Comment Detection: A Comparative Study of Traditional ML, Deep Learning, Encoder-Based and Decoder-Based Models

Youxi Tan[1], Mingjie Fang[1], Da Shen[2], Jiayi Xu[3] and Baha Ihnaini[1,*]

[1] *Institute of Advanced Natural Language Processing, Wenzhou-Kean University, China*

[2] *Department of Mathematical Science, Wenzhou-Kean University, China*

[3] *Department of Psychology, Wenzhou-Kean University, China*

Abstract: In recent years, with the fast growth of the internet and the continuous expansion of technological applications such as social media, the health and safety of the online environment has become a matter that requires serious attention. In the Chinese context, due to the complexity and diversity of syntactic expression, the detection of toxic language in Chinese faces unique challenges. This study focuses on the performance of traditional machine learning, deep learning, encoder-based transformation models, and decoder-based transformation models (LLMs) in the identification of toxic comments in Chinese, and compares the performance characteristics of different models. The study combines two main datasets, COLD and TOCAB, into a binary classification task, using accuracy, F1-score, precision, and recall as evaluation metrics to assess all the tested models. The final results show that among the tested models, the decoder-based Qwen1.5-7B (8-bit quantization) has the highest accuracy (94.71%), the traditional machine learning models and encoder-based transformation models perform moderately, while the deep learning models have lower accuracy (77%–80%) due to the limited context understanding, indicating that decoder-based large language models have advantages in the detection of toxic comments in Chinese.

Keywords: Chinese toxic comment detection, machine learning, deep learning, encoder-based model, decoder-based model, binary classification, multilingual NLP

1. Introduction

With the rapid development of the Internet and the continuous application of social media, it has become normal for users to express their opinions through the Internet. While bringing convenience to people's communication, many harmful contents have emerged, especially toxic language. Toxic language includes hate, various forms of abuse, discrimination and malicious attacks, and other bad speech, which not only causes psychological harm to individuals but also may lead to more serious social problems, such as Internet violence and hate propaganda [1].

From a psychological perspective, several factors contribute to the prevalence of toxic language throughout the internet. Social conformity pushes individuals to follow others' words and opinions, sometimes leading to aggressive or harmful behavior that they might suppress in a situation of being alone. This psychological mechanism decreases people's self-regulation and increases impulsive actions. Plus, hostile attribution bias plays an important role in online toxic behavior. Individuals tend to interpret others' words as hostile, leading to aggressive responses. Study shows that there is a positive correlation between social media fatigue and online hostile behavior, where hostile attribution bias acts as a mediating factor [2].

***Corresponding author:** Baha Ihnaini, Institute of Advanced Natural Language Processing, Wenzhou-Kean University, China. Email: bihnaini@kean.edu

In China, because of the high-context and diverse expressions of Chinese, detecting toxic comments is particularly challenging [3], which means that understanding the psychological mechanisms behind toxic comments is very important and necessary to improve the detection methods. Combining psychological theories and advanced natural language processing techniques can improve the performance of detecting toxic comments on the internet. Therefore, identifying and processing the toxic comments is the keyway to protecting the healthy environment of the internet and society, and promoting civilized communication.

By reviewing the literature, they found that pre-trained language models (such as Bidirectional Encoder Representations from Transformers (BERT), A Robustly Optimized BERT Pretraining Approach (RoBERTa), and Cross-lingual Language Model (XLM)) showed excellent performance in natural language processing and demonstrated strong capabilities in toxic language classification tasks [4], but their detection performance in non-English languages was different [5]. At the same time, the complex context and diversified expression of Chinese comments require the construction of complex NLP models and powerful datasets for detection and classification [3].

To address these challenges, the research will combine binary label datasets to explore methods for classifying toxic comments and use them as data for model training and testing to evaluate the generalization ability of different models in different types of Chinese toxic comments. This study, by combining two large datasets, COLD

[6] and TOCAB [7], converts toxic comments from different categories into simple binary label classes.

Therefore, the research goal is to classify toxic comments by combining multi-label datasets and train the LLMs to detect the generalization ability of different models in different types of Chinese toxic comments. Through this progress, the aim is to show the accuracy and generalization ability of different models in the Chinese toxic comment classification task, which also provides support for further research on model development and detection and classification of Chinese toxic language.

The structure of this article is as follows. Section 2 is a literature review, which reviews research results related to toxic language classification. Section 3 introduces the method, including the selection and composition of datasets, the LLM used, the fine-tuning process, and the evaluation metrics (Accuracy, F1-score, precision, recall); Section 4 shows the research results; Section 5 summarizes the research results.

2. Literature Review

Toxic speech, including hate speech, insults, threats, and harassment, has become a major challenge in social media and online platforms, resulting in negative influence on mental health and social stability. The toxic comment detection in English has been widely researched; however, due to the complexity of grammar, dependency of context, and code-switching (mixing languages within a conversation), the toxic comment detection in Chinese is still challenging.

What's more, the Sarcasm, metaphors, and specific expressions of culture make detecting complicated furtherly; the existing NLP models, which are based on English, cannot handle Chinese usefully. This study aims to use traditional machine learning, deep learning, encoder-based, and decoder-based models to improve Chinese toxic comment detection to solve these challenges.

This literature review outlines the key methods, recent advancements, and current limitations, providing support for future research.

2.1. Machine learning

In the toxic comment's classification task, A New Preprocessing Technique for Toxic Comment Classification "proposed an innovative preprocessing method: Transform the multi-label problem into a multi-classification problem through label encoding." Research shows that this method significantly improves the performance of the simple model - logistic regression achieves an accuracy rate of 95.39% on the "toxic" label, outperforming Naive Bayes and decision trees. However, the recognition effect of this preprocessing on complex language phenomena (such as irony or veiled attacks) is limited. The author suggests that subsequent studies can combine neural networks to enhance semantic understanding ability [8].

By contrast, deep learning is better at capturing semantic nuances through Contextual Embedding. For example, Text Classification: How Machine Learning is Innovating Text Classification indicates that Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are sensitive to local patterns, while pre-trained models such as BERT alleviate the problem of data sparsity through transfer learning [9]. However, such models rely on large-scale data, and static word embedding is difficult to adapt to the changing speed of web terms. Therefore, the current research trend tends to favor Hybrid frameworks, such as using the preprocessing results of logistic regression as the input features of BERT to balance efficiency and accuracy [10].

2.2. Deep learning

2.2.1. Long short-term memory (LSTM)

Recent studies have explored the application of LSTM combined with multimodal features in the toxic comments' detection task. For example, a study utilized the loan scam-related conversations dataset on YouTube, integrated the voice and text features to enhance the performance of LSTM in toxic comments detection [10]. The result shows that incorporating voice features (e.g., pitch, speech rate, intensity) and text features (e.g., Bag-of-Words, TF-IDF) significantly improved the accuracy of detection, and reported the F1 score of these two kind of data, the F1 score of the text data was 92.73%, and the F1 score of the voice data was 89.09%, which was 12%–30% higher than that of the traditional LSTM model.

Although the focus of this study is Indonesian, its method is of great significance for the detection of toxic comments in Chinese. Relying solely on text-based functions, there are still certain difficulties in detecting sarcasm and implicit toxicity in Chinese. Therefore, combining speech features is helpful to improve the recognition of implicit harmful content. However, since challenges such as high data annotation costs and complex integration of Chinese speech and text features still exist, further research is needed.

2.2.2. CNN

CNN has achieved remarkable success in tasks such as image classification, object detection, and segmentation through multi-level feature extraction and automatic learning capabilities [11]. CNN is good at image classification tasks due to its advantages in multi-level feature extraction and automatic learning. Secondly, it usually automatically extracts features through input data, which improves work efficiency. However, it is not only effective in image classification but also can easily handle complex character structures in handwritten text recognition. However, as a spatial model, CNN has obvious shortcomings. It does not perform well in the context of text and complex language environments.

Sugandhi et al. [12] conducted a lot of research and concluded that CNN is effective in classifying malicious comments. Compared with the traditional BoW method, CNN has a better performance in test set accuracy and has obtained more positive feedback after interface deployment. This shows the powerful ability of CNN in word order. However, issues like data bias and difficulty in detecting vague or culture-specific expressions remain. However, problems such as data bias and difficulty in detecting ambiguous or culturally specific expressions still exist. Based on this situation, Zhang and Wang [13] proposed a new type of neural network: a character-level neural network of CNN bidirectional gated recurrent unit (Bi-GRU). It can model local features as well as sequence information at the same time. Their cw-2cbg-p model achieved an F1 score of 0.8081, outperforming other models.

2.2.3. Transformer-based models

CA study used two different datasets to evaluate hate speech detection models, with a focus on three advanced models: BERT and RoBERTa. The methodology involved essential preprocessing steps, including tokenization, lemmatization, and TF-IDF feature extraction, to prepare the data for analysis. The results indicated that RoBERTa achieved the greatest accuracy rate of 93% in detecting hate speech, demonstrating the models' remarkable ability in identifying nuanced poisonous language and outperformed the traditional model, thereby, transformer-based models have the potential of effective tool for automatic content auditing on social media platforms, also providing a feasible solution for large-scale detection tasks [14]. This study supports the view that, compared with traditional models, transformer models, especially RoBERTa, can better handle the problem of discovering complex and destructive remarks, including identifying sarcastic and harmful phrases depending on the situation.

Tarun et al.'s study [15] utilizes advanced NLP techniques and classification models, especially BERT, to classify the dataset Wikipedia comments into six types of toxicity: toxic, obscene, threat, insult, severe toxic, and identity hate. The results show that BERT achieved an ROC AUC rate of 98.23%, indicating the effectiveness of transformer-based models for toxic comments classification in text data.

2.3. Large language models

In terms of detecting hate speech, LLMs not only match well with machine learning models, but also often perform better. The researchers discovered that the prompts with thorough design can help LLMs better understand the background of hate speech by using different prompting techniques, which make LLMs more effective than the traditional methods [16]. Because LLMs possess advanced language understanding and reasoning capabilities, they can analyze texts more quickly and accurately and can extract key information, thereby reducing the need for manual preprocessing.

Additionally, LLMs skilled at handling large datasets and subtle language differences, improving the performance of tasks such as sentiment analysis, hate speech detection, and topic modeling, they can quickly learn and adapt with the least input so that achieving faster deployment and better handling of language-related challenges, their flexibility and scalability also make them become the idea choice of various text data analysis, even in low-resource or multi-language settings.

Pre-trained models like GPT-2 and RoBERTa usually show good results in detecting toxic language, among which GPT-2 performs better in English content. However, these two models sometimes generate some toxic outputs or overly rely on common words, all of which will affect their performance. Different languages can also affect the performance of the model in toxic content detection. For example, tests on small French datasets have shown the result that simpler models, such as logistic regression, are more easily influenced by biased keywords related to social groups, while tests on English and Arabic datasets are less impacted [5]. These findings indicated the importance of considering language-specific factors, dataset size, classifier bias, and cultural background in improving the accuracy of models in different languages.

2.4. Chinese toxic speech detection

Detecting malicious language in Chinese faces unique challenges due to the diversity of contextual characteristics, dialects, and cultural references of Chinese [3]. For example, the meaning of a Chinese phrase can vary greatly depending on the context, making it difficult for the detection system to accurately detect offensive language. In addition, with the rapid development of the times, the language is also evolving rapidly, including the emergence of subversive expressions and new words, as well as regional dialects or homophonic insults, which further increases the complexity of the detection task. In addition, the use of enhanced data annotation technology and cross-cultural transfer learning can make the model better adapt to the complexity and integration of the rapid development of Chinese. These changes will eventually lead to a more accurate and sensitive detection system that can better handle the evolving toxic remarks on Chinese online platforms.

The study "Chinese Offensive Language Detection Algorithm based on Pre-trained Language Model and Pointer Network Augmentation" has made breakthrough progress in the field of toxic language recognition in Chinese. It proposes a Chinese aggressive language detection algorithm enhanced by the pre-trained language model RoBERTa and pointer network based on the previous studies [17]. Meanwhile, aiming at the deficiencies of traditional Chinese toxic substance detection methods, it innovatively constructs a collaborative architecture combining the pre-trained language model and pointer network, and introduces the task information embedding mechanism through the prompt layer, which has significantly improved the model's contextual understanding ability of toxic Chinese language.

3. Research Methodology

3.1. Dataset

In this study, two Chinese toxic language datasets were used, namely, COLD dataset and TOCAB dataset. COLD dataset, which is provided by the thu-coai team at Tsinghua University and hosted on the Hugging Face platform, and is used for Chinese offensive language detection. 37,480 Chinese comments are in the COLD dataset on three topics (spanning race, geography, and gender) and binary label ("toxic", "safe"), stored in CSV format, and text sources include social media platforms and forums, with high diversity and representation. TOCAB dataset, a benchmark dataset published by Taiwan Ocean University (NTOU) for abusive language processing in Chinese, contains 121,344 social media comments using JSON format, it across six labels — gender, physical, political, racial and offensive speech, only the training set (104,002 comments in total) was allowed to be used for toxic language detection research.

3.1.1. Dataset preprocessing

Before using the dataset, we performed the following pre-processing steps to ensure the consistency of the data and the applicability of the model. First, category conversion was applied to the TOCAB dataset by merging the six multi-class labels into binary labels. All samples labeled abusive, discriminatory, or threatening were reclassified as "toxic", while unlabeled samples were considered "safe". Next, key-value adjustments were made to align both datasets with the desired input format. A new key called "instruction" was added to each record, with a uniform value: "Find from the following input whether it is toxic or not." The original text key was renamed to "input" and the label key to "output". Label values under the "output" key were converted into categorical labels, where '1' was mapped to "toxic" and '0' to "safe". All other unnecessary keys were removed, retaining only "instruction", "input", and "output". Finally, to support data fusion and further processing, all data originally in CSV format was converted to JSON format.

After preprocessing, the datasets were merged. In the COLD dataset, 19,439 samples were labeled as "safe" and 18,041 as "toxic". In the TOCAB dataset, 88,939 samples were labeled as "safe" and 15,063 as "toxic". The integrated dataset thus contained a total of 141,482 samples. The distribution of sample categories across the individual and combined datasets is shown in Figure 1.

To prepare the data for model training and evaluation, a random data splitting strategy was used. The data was divided into training, validation, and test sets in proportions of 80%, 10%, and 10%, respectively.

The complete dataset was first segmented and divided into 90% of the training data and 10% of the test data. The LLaMA-Factory [18] was used to specify the validation size as 0.1, and the 10% validation set was used to evaluate the generalization ability of the model during the fine-tuning process. To improve the reliability of the model performance, stratified sampling is used to maintain the original proportion of "toxic" and "safe" labels in the training set, validation set, and test set.

To facilitate the reproduction of subsequent experiments and error evaluation, Table 1 presents the distribution of the number of "Safe" and "Toxic" samples in different subsets of the merged dataset, reflecting the balance of each category in the three stages.

Figure 1
Safe vs. toxic sample counts across datasets

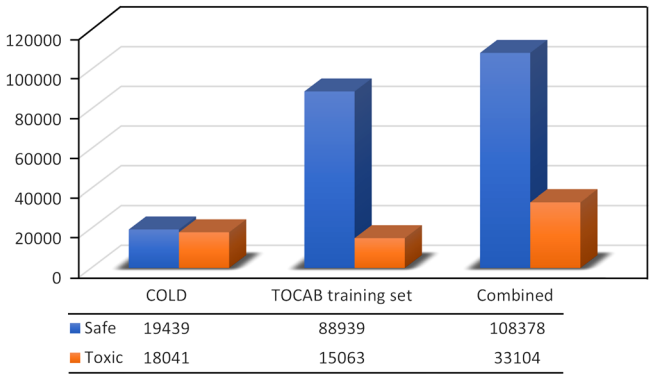


Table 2 presents an example of the combined dataset.

3.2. Model description

This study explores the performance of various models in the task of toxic speech detection, including machine learning, deep learning, encoder-based models, and decoder-based models.

Machine learning models such as logistic regression, Naive Bayes, support vector machines, and decision trees are effective for toxic comments detection due to their interpretability, rapid computing, and processing abilities of texts [19].

In deep learning, the LSTM model understands the dependencies of words over time, and CNN effectively extracts local patterns. They together improve the accuracy of identifying complex toxic languages by capturing context and semantic features [20].

Encoder-based models, such as Bidirectional Encoder Representations from Transformers (BERT-base-uncased), Decoding-enhanced BERT with disentangled attention (DeBERTa), RoBERTa, Distilled BERT (DistilBERT), and Modernized BERT (ModernBERT), improve toxic comment detection by analyzing the text in two directions and using complex attention methods. BERT versions focus on accuracy and depth of meaning. Distilled Bert balances efficiency and performance, while ModernBERT enhances training and generalization [21].

On the contrary, decoder-based models are mainly constructed to generate sequences autoregressively by utilizing decoder components of the Transformer architecture. By using hidden self-attention, they predict each token based on the previously generated token, thereby effectively capturing long-term dependencies and subtle contextual information.

3.2.1. Machine learning models

Logistic regression (LR): Logistic regression is a preferred method for binary classification and multi-class classification problems in statistical machine learning. What makes it useful in toxic speech detection is its sigmoid function, which converts the linear combination of features into probability scores indicating harmful content [22]. For example, in terms of social media auditing, platforms like Facebook have traditionally used LR to label potentially harmful content because of its interpretability and the efficiency of real-time filtering.

Naïve Bayes (NB): Naïve Bayes is a classification method based on probability. This method works on the idea that all features are independent of each other when the class is known. This isn't completely true, but it makes math much easier. Even with this simplification, it works well for text data, which is often high-dimensional and sparse. In toxic language detection, Naïve Bayes can process data quickly due to its efficient probability calculations. However, its main limitation is that it is difficult to capture complex semantic relationships and contextual meanings [23].

Support vector machines (SVM): Support vector machines are commonly used for text classification tasks and are supervised learning methods for classification, regression, and anomaly detection. In the field of data classification, SVMs are particularly effective because they have good generalization ability by maximizing the classification margin, so they are effective in detecting malicious comments. They use kernel functions to transform text features, which helps to identify sarcasm or indirect threats. However, the computational cost of training SVMs on very large datasets can be quite large, which may reduce their performance [24].

Decision tree (DT): Decision trees discover toxic comments by decomposing language patterns step by step. For example, they might sort comments based on how often certain words appear or the emotion behind the words. This helps clearly identify toxic content. Since each

Table 1
Basic statistics of combined dataset

	Safe	Toxic	Total
Train/Dev	97,540	29,793	127,333
Test	10,838	3311	14,149
Total	108,378	33,104	141,482

Table 2
Example from dataset in English

Instruction	Input	Output
Classify the following test as either 'safe' or 'toxic'. Toxic text includes hate speech, harassment, explicit content, threats, or offensive language. Safe text is free from such harmful context. Provide only the classification label ('safe' or 'toxic') without additional explanation.	Yeah, think about waiting half a month for the logistics warehouse to break up and the risk of being killed by a fat black woman at 2:00 in the morning to line up at the store, I'd rather wait, slow down, life is more important.	Toxic

decision follows simple rules, it is easier to understand why a comment is labeled toxic. However, decision trees may overfit when dealing with imbalanced data, and methods such as pruning can be used to improve the results [25].

In addition, models such as decision trees and support vector machines are sensitive to data distribution. If the proportion of offensive text in the training set is unbalanced, the model will miss "low-proportion offensive scenarios" or be overly sensitive to "high-proportion normal scenarios."

3.2.2. Deep learning models

LLSTM: LSTM is a special type of RNN designed to better capture long-term dependencies in sequential data. When detecting toxic speech, LSTM is particularly good at understanding context, especially with aggressive language [26]. Its memory cells and gating mechanism help it to avoid vanishing gradients. This way, it can better analyze the relationship between the words, improving the model's understanding of complex languages.

CNN: CNN is a type of deep learning model that is good at handling grid-like data structures. It cannot only be used for image analysis, but also in the detection of toxic text. CNN has strong abilities for feature extraction. By sliding the convolution kernels, the model can automatically identify potential offensive word combinations in the context, achieving efficient recognition and classification of toxic comments [27].

3.2.3. Encoder-based models

Encoder-based models apply multi-head self-attention and feed-forward networks to capture bidirectional context from the input sequence. Unlike traditional recurrent neural networks, these models allow each word to consider all other words in sequence at the same time, making it easier to identify long-range and complex relationships. Variants like BERT, DeBERTa, RoBERTa, DistilBERT, and ModernBERT perform well in tasks such as text classification and toxic comments detection by recognizing subtle semantic patterns [28]. This study explores how these models can extract features from Chinese toxic comments to improve detection accuracy.

BERT-base-uncased: BERT-base-uncased is a popular pre-trained language model. It understands text by looking at words from both directions using Masked Language Modeling. The model is trained on large datasets like Wikipedia and BookCorpus. It has strong language understanding. It can find complex sarcasm and hidden aggressive language in text. Studies have shown that the Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) score of this model for toxic language detection is as high as 98.396%, and its performance is even better when used together with ensemble learning [29].

DeBERTa: DeBERTa has made some improvements based on the original BERT model, such as introducing the understanding entanglement attention mechanism and relative position encoding, which makes it perform more naturally when dealing with sentence and word relations. When detecting toxic comments, it can identify some easily overlooked implicit expressions and has more advantages in understanding complex semantics compared to BERT and RoBERTa.

RoBERTa: RoBERTa made some adjustments to the BERT-optimized model, such as eliminating the prediction task of the next sentence and adopting a dynamic masking method. It can be trained for a longer time on a larger dataset. Therefore, when dealing with toxic comments, it can better capture implicit or indirect malicious expressions. It has better performance.

DistilBERT: DistilBERT is a simplified model of BERT, which uses fewer resources and runs faster, balancing the relationship between accuracy and efficiency. Its bidirectional attention function can better understand the context and effectively detect toxic remarks. Trained on large datasets, it is able to recognize subtle signs of toxic language, such as sarcasm or hidden threats [30].

ModernBERT: ModernBERT is an improved BERT model. It optimized the training method and enriched the language data. Although its size is the same as that of BERT-base, it has a faster learning speed and stronger generalization ability. It performs well when dealing with unclear or sensitive language. The study demonstrates an accuracy rate of approximately 92% in classifying toxic content, thereby proving its reliability and strong performance [31].

The general architecture of Transformer-based models is shown in Figure 2 [32]. This structure consists of a stack of encoders utilizing position encoding and multi-head self-attention, which allows the model to capture long-range dependencies and contextual relevance more effectively.

3.2.4. Decoder-based models

In contrast, decoder-based models are designed for sequence generation by using decoder components of the Transformer architecture. They predict each token in sequences, through a hidden self-attention mechanism, and only focus on the previously generated tokens. This autoregressive feature enables decoder-based models to capture complex language patterns and remote context relationships. Just as emphasized in the recent study by Acheampong et al. [28], LLMs that built on decoder structures such as Qwen and LLaMA show a strong generalization ability, which makes them particularly effective in detecting subtle, implicit, and context-related toxic speech. In this study, comparing the performance of three most advanced large-scale language models (Qwen, LLaMA, and Falcon), the result indicates that the decoder-based model performs well in identifying and interpreting subtle toxic expressions in Chinese.

1) Qwen

Qwen is a high-performance LLM developed in China. It combines a rich Chinese and multilingual corpus. In the task of toxic speech detection, Qwen can generalize powerful detection capabilities from different languages and fields through a Multi-task Learning strategy combined with Few-shot Learning. To enhance the generalization and reliability of our conclusions, the study also chooses models of different types and parameter sizes [33].

Qwen1.5-0.5B: The lightweight version of the Qwen1.5 series, with 0.5 billion parameters, improves efficient reasoning while still being able to compete in finding toxic comments.

Qwen1.5-1.8B: As a medium-sized model of the Qwen1.5 series with 1.8 billion parameters, it has shown a significant efficiency and accuracy balance in toxic comments detection tasks.

Qwen1.5-4B (8-bit): The 4 billion parameters model, with 8-bit quantization and compression, it performs well in detecting latent toxic comments and shows better context understanding.

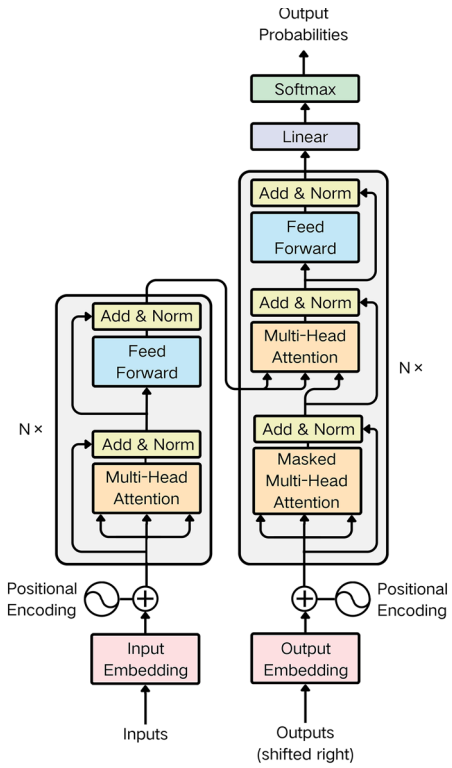
Qwen1.5-7B (8-bit): The Qwen1.5 model with 7B parameters and 8-bit quantization performed best in our study results, due to its deeper structure and better training data.

2) LLaMA

LLaMA (LLM Meta AI) is an open-source LLM model. In terms of malicious text detection, through transfer learning, LLaMA can mine semantic relationships from a large open corpus. With its rich contextual understanding ability to detect Chinese-specific malicious content evasion techniques, such as pinyin homophones and character splitting, LLaMA can detect malicious speech hidden in complex sentences and show strong scalability in multimodal tasks [34].

LLaMA2-7B (8-bit): Using transfer learning to reveal semantic relationships in malicious speech patterns. Based on a 7 billion

Figure 2
Transformer model architecture



parameter architecture, it is optimized for Chinese, and quantization technology increases inference speed by 40% and reduces memory usage by 60%.

LLaMA2-13B (8-bit): A more powerful version of LLaMA2, with 13 billion parameters, shows excellent scalability in complex sentence interpretation and significantly improves the understanding of complex Chinese sentences compared to the 7B version.

3) Falcon

Falcon is an open-source LLM developed by Hugging Face and the Big Science community. Through Zero-shot Learning and Fine-tuning, it has significant advantages in tasks with subtle cultural or linguistic differences. It is particularly good at picking up obscure speech patterns and polysemic usage scenarios [35].

Falcon-7B (8-bit): Falcon-7 B was developed by the Institute of Technological Innovation (TII) in the UAE. It is an autoregressive model with around 7 billion parameters, following the Transformer design. It was trained using a mixture of texts from various fields - approximately 1.5 trillion words - including websites, open-source materials, and some academic materials. When detecting toxic comments, Falcon-7B has shown an ability to deal with expressions that are vague or context-dependent. This may be particularly relevant when different cultural or linguistic backgrounds are involved, though the effect isn't always easy to measure. That said, it's not like the model always beats others. For example, when the task complexity is high or the sample size is insufficient, its performance may not improve significantly [36].

3.2.5. Theoretical foundation of transformer

The encoder models and decoder models used in this study are all based on the Transformer architecture, and the core mechanism of this architecture is Self-Attention, which can model the dependency relationship between any positions in the input sequence and thereby capture the context features [37]. The self-aAttention mechanism is particularly crucial for the identification of implicit harmful remarks.

In the self-attention mechanism, each word in the input sequence is respectively mapped to three vectors: Query, Key, and Value, and the similarity is calculated to determine the degree of attention. Its core calculation formula is:

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

where $Q, K, V \epsilon \mathbb{R}^{d\times r}$ respectively represent the Query, Key, and Value matrices, and d_k is the key vector dimension.

This mechanism enables the model to establish direct connections between any positions; thereby, models can effectively understand the complex context in Chinese texts.

3.3. Fine-tuning

3.3.1. Approach

In this study, using Low Rank Adaptation (LoRA) and its quantized version QLoRA fine-tune pre-trained models to complete the research tasks. These methods not only meet the efficiency of model adjustment but also reduce the requirements for memory and computing power. Therefore, LoRA and QLoRA combined become the most effective solution to complete research tasks under limited resources.

LoRA works by using smaller low-rank matrices to update only certain layers of the model, which can reduce the number of trainable parameters:

$$W + \Delta W = W + AB^T, A\epsilon\mathbb{R}^{d\times r}, B\epsilon\mathbb{R}^{r\times k} \quad (2)$$

where *r* being the rank. It has been proven that LoRA performs as well as or better than full fine-tuning, while requiring fewer parameters and data, and it has become the standard method for parameter-efficient fine-tuning (PEFT) of large language models [38].

QLoRA goes a step further and uses 4-bit quantization to reduce the model size without affecting performance, which makes fine-tuning more efficient and helps large language models (LLMs) adapt better [39].

This combination method has several advantages: first, low-rank optimization significantly speeds up the training process; second, it is highly scalable, which allows it to adapt to larger models or datasets with only minor adjustments. It performs well on tasks involving cultural or linguistic nuances. It is particularly good at recognizing rare phrases and words with multiple meanings, both when fine-tuned and for zero-shot learning [35].

3.3.2. Experimental setup

The fine-tuning experiment was conducted based on the Google Colab interface, and the experiment used an NVIDIA L4 GPU (with 22.5 GB VRAM) and 53 GB of system RAM. These computing resources provided sufficient computing power for the experiment. The learning rate was adjusted to 5e-5 to ensure effective updates, and 3 epochs of fine-tuning were selected based on cost-performance balance. Batch size was adjusted according to model architecture and memory requirement for effective use of resources. The dataset was also limited to 64,000 samples for efficient processing time and effective results.

3.3.3. Procedure

First, the pre-trained marker is used to standardize the dataset and insert task-related special markers. In the fine-tuning stage, we use LoRA technology to adapt only the attention weight matrix in the model and introduce a low-rank decomposition matrix with a rank of 8 to balance computational efficiency and model performance. At the same time, the QLoRA framework is used to quantize the model weights to 4

bits of precision to further reduce the video memory requirements. The entire training process is implemented in the Colab environment based on the Hugging Face Transformers library, which significantly reduces the number of trainable parameters by updating the low-rank adaptation matrix. After 3 training cycles, we evaluate the model performance on the validation set, closely monitor the changing trends of the loss curve and evaluation indicators, and dynamically adjust hyperparameters such as the learning rate as needed. This fine-tuning strategy can reduce video memory usage while ensuring model performance.

3.4. Evaluation metrics

In this study, a variety of evaluation indicators were used to objectively and comprehensively evaluate the model's performance in the Chinese toxic speech detection task, which includes accuracy, F1 score, precision, and recall.

Accuracy: Accuracy measures the proportion of samples the model correctly predicts. The formula is:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

In the formula, TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negative.

F1 score: F1 score is a harmonic average of accuracy and recall and is suitable for evaluating model performance on unbalanced data. The formula is:

$$F1\ score = \frac{2\times Precision\times Recall}{Precision+Recall} = \frac{2TP}{2TP+FP+FN} \tag{4}$$

Precision: The precision measures the proportion of comments that are toxic among those judged by the model to be toxic, calculated by:

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

Recall: The recall rate measures the proportion of actual toxic reviews that are correctly identified by the formula:

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

Through the comprehensive use of these evaluation metrics, the overall performance and accuracy of the model in detecting toxic text can be comprehensively analyzed.

4. Result

This study compares multiple models, which fall into four main categories: traditional Machine Learning (e.g., logistic regression), Deep Learning (e.g., CNN and LSTM), Transformer-based (e.g., DistilBERT, ModernBERT), and LLMs (e.g., Qwen, LLaMA2, and Falcon). All models were assessed using a consistent binary classification framework, evaluated with four standard metrics, and all evaluation results are presented in Table 3. Some models, particularly within the LLM category, were further optimized through 8-bit quantization to reduce computational cost while maintaining performance.

4.1. Performance of machine learning models

Among traditional machine learning models, logistic regression achieved the best overall performance, with evaluation metrics including accuracy, precision, recall, and F1 values, all of which reached 90%. This shows that logistic regression can provide reliable results in binary malicious comment classification, especially when computing resources are limited. Compared with similar models such as naive Bayes, support vector machines, and decision trees, their F1 values are slightly lower, ranging from 77% to 88%, and are more sensitive to changes in data distribution. Overall, the advantages of traditional machine learning models are fast speed and ease of interpretation, but they lack the ability to capture contextual and semantic features, which limits their effectiveness in complex Chinese text classification tasks.

4.2. Performance of deep learning models

CNN's F1 score is 73.74%, and its precision rate (83.48%) is higher than LSTM's, but the recall rate (80.13%) indicates that there is still a risk of misjudgment in some scenarios. The F1 score of LSTMS is only 67.22%, and the precision rate is only 59.54%, indicating that it is weak in recognizing harmful texts and easy to misjudge. CNN, by contrast, is more efficient in its use of resources, but still lags significantly behind LLMs. In comparison, deep learning models require fewer computing resources than LLM, but they are less capable of capturing context and processing semantic information, so deep learning performs poorly in detecting Chinese malicious text environments. CNN is more efficient than LSTM in terms of resource utilization. However, in actual experiments, CNN is less efficient than Transformer-based models.

4.3. Encoder-based models (transformer-based models)

BERT, RoBERTa, DeBERTa, ModernBERT and DistilBERT are all encoder-based models, all based on the Transformer encoder to help complete the task. The above models all optimize the bidirectional context encoding direction, allowing them to more effectively extract features from the input content. Because of the balance between interpretability and robustness, these models are widely used in the field of text classification [40–42].

Among them, from the perspective of F1 score evaluation, DistilBERT's F1 score of 83.84% is significantly better than deep learning models such as CNN (73.74%) and LSTM (67.22%). In addition, RoBERTa-base and DeBERTa-base full-size encoder models performed well, with F1 scores exceeding 93%. These results confirm that encoder models perform well in identifying structured patterns and semantic clues in Chinese malicious comments, especially when the input format is well-formed and the sequence length is moderate [40–42].

4.4. Decoder-based models (LLMs)

On the contrary, decoder-based models such as Qwen, LLaMA, Falcon, and so on, utilize Transformer decoder architecture. These large models are usually pre-trained on corpora to better adapt to various contexts. They not only have a stronger ability of context understanding, but also have high inferential capability [39, 43, 44].

In this study, Qwen1.5-7B has the highest overall performance with an F1 score of 94.72%, LLaMA2-13B and Qwen1.5-4B follow closely behind. Decoder-based LLMs' performance in all metrics is superior to that of the encoder-based model, especially in understanding the common implicit toxicity, code-mixed language, and satire in Chinese online discourse.

Although under the 8-bit quantization (e.g., Qwen1.5-4B, LLaMA2-7B), these LLMs still maintained high performance, which shows the practicality of decoder-based models' actual deployment in a restricted environment [39].

This performance advantage stems from the ability of the decoder model to learn flexible attention mechanisms and long-term dependencies, which are particularly useful for detecting subtle differences and confusing toxic expressions [44, 45].

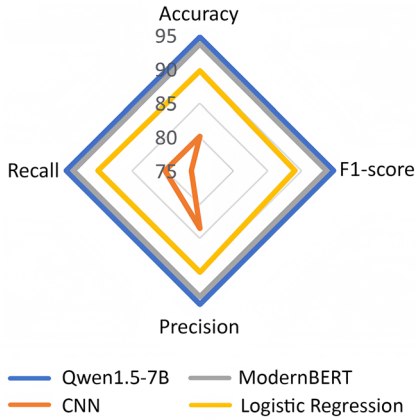
4.5. Comparison of model categories

Figure 3 provides a radar chart summarizing the best-performing models across four categories: traditional machine learning, deep learning, encoder-based models, and large language models (LLMs). Notably, Qwen1.5-7B achieved the highest scores across all four metrics, demonstrating the superior capability of LLMs in toxic comment classification. The ModernBERT-base model followed closely with an F1 score of 93.66%, demonstrating its ability to strike a good balance between performance and computational efficiency.

Logistic regression, representing traditional machine learning, showed stable results with an F1-score of 90%, though its ability to handle complex context is limited. In contrast, the deep learning model CNN performed noticeably worse, with relatively low recall and an F1-score of only 73.74%, indicating its difficulty in handling the complexity of Chinese toxic language.

The results in Figure 3 show that LLM and Transformer-based models outperform other categories in modeling Chinese semantics and contextual analysis, while traditional and deep learning models may be more suitable for applications with limited resources or lower accuracy.

Figure 3
Optimal performance for each type of model



4.6. Comparison of model parameters

As shown in Table 3, the relationship between model parameter size and performance of Chinese malicious comment detection is approximately linear. For example, LLaMA2-13B (13B parameters) achieves 94.62% accuracy, while Qwen1.5-7B (7B parameters) performs even better with 94.71% accuracy. This suggests that larger models do not always outperform smaller but optimized architectures. Instead, the architecture optimization and training strategies are more effective than simple parameter size.

Traditional models such as GPT-2 (91.59% accuracy) and BERT-base (84.30% accuracy) perform poorly in Chinese tasks due to their English-centric pre-training, which does not adapt well to the context. Their performance is affected by word segmentation mismatch and limited Chinese training data. In contrast, the Qwen series (e.g., Qwen1.5-0.5B with 94.27% accuracy) demonstrates word segmentation and training data optimization for Chinese. Although Qwen has fewer parameters than GPT, custom architectures (such as Qwen) also perform better than general models (such as GPT-2).

4.7. Summary

In summary, LLMs such as Qwen and LLaMA have shown high efficiency in Chinese malicious comment detection, significantly outperforming deep learning models (CNN, LSTM) and lightweight Transformer variants (e.g., DistilBERT). Although LLMs require more computational resources, their superior performance and understanding of context in the Chinese context make them the most suitable solution.

When system limitations prevent the use of LLMs, Transformer-based models provide a powerful alternative that maintains high accuracy with lower resource requirements. Deep learning models are more efficient, but they are limited in their ability to capture language complexity. Traditional models are fast and easy to interpret, but they perform poorly in the Chinese language context.

In addition, quantization techniques (e.g., 8-bit compression) play a vital role in enabling practical deployment of LLMs, which can not only reduce memory usage and improve efficiency but also maintain high performance.

5. Conclusion

5.1. Overall summary of findings

In this study, we compared the performance of different algorithmic models in detecting toxic Chinese comments, including traditional machine learning methods, deep neural networks, encoder-based Transformer designs, and decoder-structured large language models. The tests showed that models with decoder structures work best overall. Specifically, Qwen1.5-7B (94.71% accuracy, 94.72% F1-score) and LLaMA2-13B (94.62% accuracy, 94.63% F1-score) both exceeded 94% on key evaluation metrics, significantly surpassing other methods.

These algorithms scored above 94% on key measures — demonstrate strong capability in capturing Mandarin's unique complex grammar patterns and hidden contextual clues. Our tests confirm they're accurate at spotting subtle emotional shifts and underlying aggressive meanings in online insults. Regular neural networks showed limited performance with Chinese semantic analysis - CNN gets 80.13% accuracy and LSTM only 77.16%. They struggle to understand complicated meaning connections in text. But the encoder-based DistilBERT does much better (84.88% accuracy), giving us more accurate predictions without using too much computing power.

Besides its theoretical contributions, this study also highlights practical applications in enhancing toxic content detection on Chinese online platforms. Large language models like Qwen1.5 can be integrated into platform-level content moderation pipelines to help find and mark toxic posts automatically. This is especially helpful when dealing with complex language and fast-changing discussions. In early tests using synthetic data from sites such as Weibo and Zhihu, the Qwen1.5-7B model was able to spot toxic comments that used sarcasm or hidden meaning—types that are often missed by basic filters. These results suggest the model could be used for real-time monitoring applications. In the future, it could be linked to live data through APIs to send alerts and help mitigate the spread of offensive content.

5.2. Ethical and bias considerations

In the Chinese contexts, large language models are particularly affected by factors such as data, annotation practices, technical design, and social culture, which may result in bias. When training data include offensive material such as harmful speech or politically sensitive topics, ethical issues may arise. Models trained on this data might accidentally repeat or make these patterns stronger, and very strict filtering might wrongly mark normal expressions that have changed meaning as toxic, which could leave out certain groups to cause harm. To reduce these

Table 3
Results of all models

Module	Quantization	Accuracy	F1 score	*Precision*	*Recall*	#Params
Qwen1.5-0.5B	none	94.27	94.25	94.24	94.27	500M
Qwen1.5-1.8B	none	94.42	94.42	94.42	94.42	1.8B
Qwen1.5-4B	8-bit	94.67	94.66	94.66	94.67	4B
Qwen1.5-7B	8-bit	94.71	94.72	94.72	94.71	7B
Falcon-7B	8-bit	85.18	86.23	90.72	93.49	7B
LLaMA2-7B	8-bit	94.62	94.63	94.63	94.62	7B
LLaMA2-13B	8-bit	94.62	94.63	94.63	94.62	13B
GPT-2		91.59	91.76	92.15	91.59	124M
DistilBERT		84.88	83.84	84.01	84.88	67M
BERT-base-uncased		84.30	83.39	83.41	84.30	109M
ModernBERT		84.30	93.66	93.65	93.67	150M
RoBERTa		93.11	93.16	93.25	93.11	125M
DeBERTa		93.62	93.69	93.82	93.62	139M
CNN		80.13	73.74	83.48	80.13	1.5M
LSTM		77.16	67.22	59.54	77.16	6.6M
Logistic Regression		89.77	89.00	90.00	90.00	10K
Naïve Bayes		87.64	87.00	87.00	88.00	20K
SVM		89.72	89.00	90.00	90.00	10K
Decision Tree		81.91	77.00	84.00	82.00	10K

risks, it is important to pay attention to how data is collected, have clear labeling rules, and let human reviewers handle unclear cases.

Most of the available corpora are in Mandarin, and the way they are labeled may have biases. This can make the model perform worse in other regional dialects and might support existing stereotypes in applications. The makers of labeling rules are also limited by the social and cultural environment in which they live. When designing rules, they may unconsciously use "mainstream cultural standards" as "universal standards" and ignore the particularities of dialects or minority groups. Ways to deal with this include making tests that check for differences between dialects, showing results for different groups separately, and using strong evaluation methods, like adversarial testing. Methods like adjusting data weight or targeted fine-tuning may help, but continuous monitoring and feedback are also needed.

5.3. Limitations and future work

One of the key limitations of this study is that the model was conducted only on synthetic data generated from platforms such as Weibo and Zhihu. Although these tests showed that the model can detect irony and implicit toxicity, they cannot fully reflect the complexity of the real-world online environment. To more accurately evaluate the model's performance in dynamic and diverse environments, future validation should be carried out by directly collecting real-time data streams through social media APIs or by collaborating with platform providers.

We found two main issues in this research: First, the basic yes/no classification can't properly handle all the different kinds of toxic content online. Second, decoder-type large language models need a lot of computing power. But our tests show that when we use 8-bit quantization on both Qwen1.5-4B and 7B models, the accuracy only drops by 0.1% — barely noticeable.

Our pre-set collection of data has limitations in keeping up with how toxic language changes in real time on real websites. To make toxic content detection systems work better, future studies could try two things: First, create a multi-label tagging system to catch all the different ways harmful content appears. Second, combine voice clues (like sarcastic tones) with checking if pictures and text match — this really helps spot sneaky or cross-platform attacks.

Making algorithms that save power and run fast is useful, as they can be used in more places and work especially well on regular hardware. Building toxic content detectors that work across languages and cultures has one big problem: biased training data. Potential solutions include: (1) Use GANs to boost model performance, (2) Add knowledge bases with multicultural word meanings, (3) Run international standard tests to make sure the system works right.

Recommendations

Big language models with decoder designs (like Qwen1.5-7B) work well for spotting toxic content and should be the go-to choice. When using them, it's smart to apply parameter quantization to save computing power. To make the platform work better, we need to create custom training programs for both the coding team and community mods. Upgrading from basic yes/no classification to smart multi-tag systems is a huge step forward for finding toxic content — it can spot all different kinds of nasty stuff online much better.

Funding Support

This research was supported by Wenzhou-Kean University (Grant No. IRSPK2023005) and Wenzhou Association for Science and Technology – Service and Technology Innovation Program (Grant No. KJFW2025-044). We gratefully acknowledge the university's financial assistance, which was instrumental in the progress of this work.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in GitHub at https://github.com/TanYouxi/Chinese-Toxic-Comment-Detection.

Author Contribution Statement

Youxi Tan: Conceptualization, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Mingjie Fang:** Software, Validation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Da Shen:** Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Jiayi Xu:** Writing – review & editing. **Baha Ihnaini:** Methodology, Investigation, Resources, Supervision, Project administration, Funding acquisition.

References

[1] Shrestha, A., Kaati, L., Akrami, N., Linden, K., & Moshfegh, A. (2024). Harmful communication: Detection of toxic language and threats on Swedish. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 624–630. https://doi.org/10.1145/3625007.3627597

[2] Huang, L., Chen, L., & Ma, S. (2025). The relationship between social media fatigue and online trolling behavior among college students: The mediating roles of relative deprivation and hostile attribution bias. *Frontiers in Psychology, 15*, 1495235. https://doi.org/10.3389/fpsyg.2024.1495235

[3] Xiao, Y., Bouamor, H., & Zaghouani, W. (2024). *Chinese offensive language detection: Current status and future directions.* arXiv. https://doi.org/10.48550/arXiv.2403.18314

[4] Zhao, Z., Zhang, Z., & Hopfgartner, F. (2021). A comparative study of using pre-trained language models for toxic comment classification. In *Companion Proceedings of the Web Conference 2021*, 500–507. https://doi.org/10.1145/3442442.3452313

[5] Ousidhoum, N., Zhao, X., Fang, T., Song, Y., & Yeung, D.-Y. (2021). Probing toxic content in large pre-trained language models. 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)*, 4262–4274. https://doi.org/10.18653/v1/2021.acl-long.329

[6] Deng, J., Zhou, J., Sun, H., Zheng, C., Mi, F., Meng, H., & Huang, M. (2022). COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 11580–11599. https://doi.org/10.18653/v1/2022.emnlp-main.796

[7] Chung, I., & Lin, C.-J. (2021). TOCAB: A dataset for Chinese abusive language processing. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science*, 445–452. https://doi.org/10.1109/IRI51335.2021.00069

[8] Akshay, B. R., Pulari, S. R., Murugesh, T. S., & Vasudevan, S. K. (2024). *Machine learning: A comprehensive beginner's guide* (1st ed.) USA: CRC Press. https://doi.org/10.1201/9781032676685

[9] Allam, H., Makubvure, L., Gyamfi, B., Graham, K. N., & Akinwolere, K. (2025). Text classification: How machine learning is revolutionizing text categorization. *Information, 16*(2), 130. https://doi.org/10.3390/info16020130

[10] Gumelar, A. B., Yuniarno, E. M., Nugroho, A., Adi, D. P., Sugiarto, I., & Purnomo, M. H. (2024). An improved toxic speech detection on multimodal scam confrontation data using LSTM-based deep learning. *International Journal of Intelligent Engineering and Systems, 17*(6), 880–904. https://doi.org/10.22266/ijies2024.1231.67

[11] Wu, Z. (2024). Application of CNN classic model in modern image processing. *Journal of Advances in Engineering and Technology, 1*(3), 1–6. https://doi.org/10.62177/jaet.v1i3.25

[12] Sugandhi, K., Reddy, R. U. K., Reddy, K. R. K., & Reddy, B. B. (2024). Predicting hate words and offensive language: A machine learning approach. In *2024 8th International Conference on Inventive Systems and Control*, 205–210. https://doi.org/10.1109/ICISC62624.2024.00042

[13] Zhang, B., & Wang, Z. (2022). Character-level Chinese toxic comment classification algorithm based on CNN and BI-GRU. In *Proceedings of the 5th International Conference on Computer Science and Software Engineering*, 108–114. https://doi.org/10.1145/3569966.3570000

[14] S, S., Vathsal, V. S., & Shankar, S. R. (2024). Hate speech detection using deep learning models. In *2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions,* 1–6. https://doi.org/10.1109/CSITSS64042.2024.10816708

[15] Tarun, V. G., Sivasakthivel, R., Ramar, G., Rajagopal, M., & Sivaraman, G. (2024). Exploring BERT and Bi-LSTM for toxic comment classification: A comparative analysis. In *2024 Second International Conference on Data Science and Information System*, 1–6. https://doi.org/10.1109/ICDSIS61070.2024.10594466

[16] Guo, K., Hu, A., Mu, J., Shi, Z., Zhao, Z., Vishwamitra, N., & Hu, H. (2023). An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications*, 1568–1573. https://doi.org/10.1109/ICMLA58977.2023.00237

[17] Hou, B., Xie, X., Zhang, D., Zheng, L., & Yan, G. (2024). Chinese offensive language detection algorithm based on pre-trained language model and pointer network augmentation. In *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology,* 800–805. https://doi.org/10.1109/AINIT61980.2024.10581762

[18] Zheng, Y., Zhang, R., Zhang, J., Ye, Y., & Luo, Z. (2024). LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 400–410. https://doi.org/10.18653/v1/2024.acl-demos.38

[19] Udoeyop, A. (2024). A machine learning model for training your AI. *International Journal of Innovative Science and Research Technology, 9*(7), 3463–3474. https://doi.org/10.38124/ijisrt/IJISRT24JUL769

[20] Degadwala, S., & Vyas, D. (2024). Systematic analysis of deep learning models vs. machine learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10*(4), 60–70.

[21] Ughi, R., Lomurno, E., & Matteucci, M. (2024). Two steps forward and one behind: Rethinking time series forecasting with deep learning. In *Machine Learning, Optimization, and Data Science: 9th International Conference,* 463–478. https://doi.org/10.1007/978-3-031-53969-5_34

[22] Khan, N., Raza, M. A., Mirjat, N. H., Balouch, N., Abbas, G., Yousef, A., & Touti, E. (2024). Unveiling the predictive power: A comprehensive study of machine learning model for anticipating chronic kidney *disease. Frontiers in Artificial Intelligence, 6*, 1339988. https://doi.org/10.3389/frai.2023.1339988

[23] Risch, J., Ruff, R., & Krestel, R. (2020). Offensive language detection explained. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying,* 137–143.

[24] Alsajri, A. K. S., & Steiti, A. (2024). Intrusion detection system based on machine learning algorithms: (SVM and genetic algorithm). *Babylonian Journal of Machine Learning, 2024*, 15–29. https://doi.org/10.58496/BJML/2024/002

[25] Guo, L., Li, Z., Qian, K., Ding, W., & Chen, Z. (2024). Bank credit risk early warning model based on machine learning decision trees. *Journal of Economic Theory and Business Management, 1*(3), 24–30. https://doi.org/10.5281/ZENODO.11627011

[26] Alizadegan, H., Rashidi Malki, B., Radmehr, A., Karimi, H., & Ilani, M. A. (2025). Comparative study of long short-term memory (LSTM), bidirectional LSTM, and traditional machine learning approaches for energy consumption prediction. *Energy Exploration & Exploitation, 43*(1), 281–301. https://doi.org/10.1177/01445987241269496

[27] Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review, 53*(8), 5455–5516. https://doi.org/10.1007/s10462-020-09825-6

[28] Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artificial Intelligence Review, 54*(8), 5789–5829. https://doi.org/10.1007/s10462-021-09958-2

[29] Mazari, A. C., Boudoukhani, N., & Djeffal, A. (2024). BERT-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing, 27*(1), 325–339. https://doi.org/10.1007/s10586-022-03956-x

[30] Chabukswar, A., Shenoy, P. D., & Venugopal, K. R. (2024). A hybrid DistilBERT-BiGRU model for enhanced misinformation detection: Leveraging transformer-based pretraining language model. In *2024 IEEE Region 10 Symposium*, 1–6. https://doi.org/10.1109/TENSYMP61132.2024.10752290

[31] Irfan, A., Azeem, D., Narejo, S., & Kumar, N. (2024). Multi-modal hate speech recognition through machine learning. In *2024 IEEE 1st Karachi Section Humanitarian Technology Conference*, 1–6. https://doi.org/10.1109/KHI-HTC60760.2024.10482031

[32] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ..., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems,* 6000–6010.

[33] Zhang, S., Gao, C., Zhu, W., Chen, J., Huang, X., Han, X., ..., & Huang, X. (2024). Getting more from less: Large language models are good spontaneous multilingual learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8037–8051. https://doi.org/10.18653/v1/2024.emnlp-main.457

[34] Villate-Castillo, G., Ser, J. D., & Urquijo, B. S. (2024). *A systematic review of toxicity in large language models: Definitions, datasets, detectors, detoxification methods and challenges*. Research Square. https://doi.org/10.21203/rs.3.rs-4621646/v1

[35] Gao, X., Xie, W., Xiang, Y., & Ji, F. (2025). Falcon: Faster and parallel inference of large language models through enhanced semi-autoregressive drafting and custom-designed decoding tree. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(22), 23933–23941. https://doi.org/10.1609/aaai.v39i22.34566

[36] Rehm, G., Dietze, S., Schimmler, S., & Krüger, F. (Eds.). (2024). *Natural scientific language processing and research knowledge graphs: First international workshop*. Switzerland: Springer. https://doi.org/10.1007/978-3-031-65794-8

[37] Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., ..., & Hu, S.-M. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, *8*(3), 331–368. https://doi.org/10.1007/s41095-022-0271-y

[38] Liang, X., Jasmina Khaw, Y.-M., Liew, S.-Y., Tan, T.-P., & Qin, D. (2025). Toward low-resource languages machine translation: A language-specific fine-tuning with LoRA for specialized large language models. *IEEE Access, 13*, 46616–46626. https://doi.org/10.1109/ACCESS.2025.3549795

[39] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLORA: Efficient finetuning of quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 441.

[40] Adoma, A. F., Henry, N.-M., & Chen, W. (2020). Comparative analyses of BERT, RoBERTa, DistilBERT, and XLNet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing,* 117–121. https://doi.org/10.1109/ICCWAMTIP51612.2020.9317379

[41] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[42] He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 1–21.

[43] Gupta, A., Prince, A. A., Fredo, A. R. J., & Robert, F. (2022). Transformer-based models for supervised monocular depth estimation. In *2022 International Conference on Intelligent Controller and Computing for Smart Power*, 1–5. https://doi.org/10.1109/ICICCSP53532.2022.9862348

[44] Ocampo, N. B., Sviridova, E., Cabrio, E., & Villata, S. (2023). An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1997–2013. https://doi.org/10.18653/v1/2023.eacl-main.147

[45] Kementchedjhieva, Y., & Chalkidis, I. (2023). An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text. In *Findings of the Association for Computational Linguistics: ACL 2023*, 5828–5843. https://doi.org/10.18653/v1/2023.findings-acl.360

How to Cite: Tan, Y., Fang, M., Shen, D., Xu, J., & Ihnaini, B. (2026). Chinese Toxic Comment Detection: A Comparative Study of Traditional ML, Deep Learning, Encoder-Based and Decoder-Based Models. *Journal of Computational and Cognitive Engineering, 5*(2), 222-232. https://doi.org/10.47852/bonviewJCCE52026117