Expert Systems With Applications

KAN-Boosted Chinese Online Abuse Detection Framework with Sentiment and Toxicity Fusion through Global-Local-Differential Attention

--Manuscript Draft--

Manuscript Number:	
Article Type:	Full length article
Section/Category:	5.1 NLP / Large language model
Keywords:	Online Abuse Detection, Sentiment Analysis, Global-Local-Differential Attention, Kolmogorov-Arnold Networks, Feature Fusion
Corresponding Author:	Zhongfeng Kang Lanzhou University Lanzhou, CHINA
First Author:	Yutong Wang
Order of Authors:	Yutong Wang
	Zhongfeng Kang
	Jiaxue Yang
	Xiaopeng Fan
	Zijin Wu
	Shantian Yang
	Qinghua Zhao
	Zichen Song
Abstract:	Detecting online abuse on digital platforms is a critical yet challenging task, particularly in Chinese social media, where colloquial expressions, emojis, and abstract semantics complicate automated identification. Existing large language models (e.g., OpenAI's Omni-moderation) often exhibit limited effectiveness in Chinese contexts, yielding low precision in nuanced scenarios, while smaller task-specific models struggle to capture subtle emotional cues and implicit toxicity. Moreover, most existing architectures emphasize toxicity alone, overlooking sentiment features and thereby constraining overall performance. To address these challenges, we propose SenTox-GLDA, a KAN-boosted framework for Chinese online abuse detection that fuses sentiment and toxicity features through a Global-Local Differential Attention (GLDA) mechanism. Specifically, SenTox-GLDA employs: (i) a BERT-based encoder fine-tuned on the Weibo-100k dataset for sentiment analysis, and (ii) a RoBERTa-based encoder fine-tuned on the augmented ToxiCN dataset for toxicity detection. An adaptive fusion module integrates the encoder outputs, followed by pseudo-sequence projection and the GLDA module, which jointly capture both contextual and fine-grained semantic dependencies. The fused representation is then classified via a Kolmogorov-Arnold Network (KAN), enabling precise and interpretable decision boundaries. Extensive experiments on the COLDataset demonstrate that SenTox-GLDA achieves state-of-the-art performance, reaching an F1-score of 82.56% and AUC of 92.04%, surpassing strong baselines by 3.26% (F1) and 2.87% (AUC). These results confirm that explicitly integrating emotional cues with toxicity features, together with hierarchical attention mechanisms, substantially enhances detection robustness. Overall, our framework offers a scalable and interpretable solution for abuse detection in linguistically complex environments, with particular efficacy in Chinese social media. Codes will be made available at: https://github.com/kanglzu/SenTox-GLDA.

Cover Letter

Dear Editor,

I am submitting this revised revision of the manuscript titled " *KAN-Boosted Chinese Online Abuse Detection Framework with Sentiment and Toxicity Fusion through Global-Local-Differential Attention*" for consideration for publication in Expert Systems With Applications.

In this work, we propose the SenTox-GLDA**,** a KAN-boosted Chinese online abuse detection framework with sentiment and toxicity fusion via global-local differential attention. SenTox-GLDA employs: (i) a BERT-based encoder fine-tuned on the Weibo-100k dataset for sentiment analysis, and (ii) a RoBERTa-based encoder fine-tuned on the augmented ToxiCN dataset for toxicity detection. An adaptive fusion module integrates the encoder outputs, followed by pseudo-sequence projection and the GLDA module, which jointly capture both contextual and fine-grained semantic dependencies. The fused representation is then classified via a Kolmogorov-Arnold Network (KAN), enabling precise and interpretable decision boundaries. Extensive experiments results confirm that explicitly integrating emotional cues with toxicity features, together with hierarchical attention mechanisms, substantially enhances detection robustness. Our framework offers a scalable and interpretable solution for abuse detection in linguistically complex environments, with particular efficacy in Chinese social media.

Original Article Statement:

a. This manuscript is the authors' original work and has not been published nor has it been submitted simultaneously elsewhere.

b. All authors have checked the manuscript and have agreed to the submission.

Thank you for your valuable time and consideration. I look forward to hearing from you.

Sincerely,

Zhongfeng Kang

School of Information Science and Engineering, Lanzhou University, China.

KAN-Boosted Chinese Online Abuse Detection Framework with Sentiment and Toxicity Fusion through Global-Local-Differential Attention

Yutong Wang[a], Zhongfeng Kang[a,*], Jiaxue Yang[a], Xiaopeng Fan[a], Zijin Wu[a], Shantian Yang[b], Qinghua Zhao[c] and Zichen Song[d]

[a] *School of Information Science and Engineering, Lanzhou University, Lanzhou, 730000, China*
[b] *School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu, 611130, China*
[c] *School of Artificial Intelligence and Big Data, Hefei University, Hefei, 230601, China*
[d] *Department of Applied Argificial Intelligence, Sungkyunkwan University, Seoul, 16419, Korea*

[*]Corresponding author: Zhongfeng Kang (kangzf@lzu.edu.cn)

Email:
Yutong Wang: yutongwang2023@lzu.edu.cn
Zhongfeng Kang: kangzf@lzu.edu.cn
Jiaxue Yang: yjiaxue2023@lzu.edu.cn
XiaoPeng Fan: fanxp2023@lzu.edu.cn
Zijin Wu: wuzj2023@lzu.edu.cn
Shantian Yang: yangst@swufe.edu.cn
Qinghua Zhao: zhaoqh@hfuu.edu.cn
Zichen Song: sls530@skku.edu (Z. Song)

KAN-Boosted Chinese Online Abuse Detection Framework with Sentiment and Toxicity Fusion through Global-Local-Differential Attention

Yutong Wang[a], Zhongfeng Kang[a,*], Jiaxue Yang[a], Xiaopeng Fan[a], Zijin Wu[a], Shantian Yang[b], Qinghua Zhao[c] and Zichen Song[d]

[a] School of Information Science and Engineering, Lanzhou University, Lanzhou, 730000, China
[b] School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu, 611130, China
[c] School of Artificial Intelligence and Big Data, Hefei University, Hefei, 230601, China
[d] Department of Applied Argificial Intelligence, Sungkyunkwan University, Seoul, 16419, Korea

ARTICLE INFO

Keywords:
Online Abuse Detection
Sentiment Analysis
Global-Local-Differential Attention
Kolmogorov-Arnold Networks
Feature Fusion

ABSTRACT

Detecting online abuse on digital platforms is a critical yet challenging task, particularly in Chinese social media, where colloquial expressions, emojis, and abstract semantics complicate automated identification. Existing large language models (e.g., OpenAI's Omni-moderation) often exhibit limited effectiveness in Chinese contexts, yielding low precision in nuanced scenarios, while smaller task-specific models struggle to capture subtle emotional cues and implicit toxicity. Moreover, most existing architectures emphasize toxicity alone, overlooking sentiment features and thereby constraining overall performance. To address these challenges, we propose SenTox-GLDA, a KAN-boosted framework for Chinese online abuse detection that fuses sentiment and toxicity features through a Global-Local Differential Attention (GLDA) mechanism. Specifically, SenTox-GLDA employs: (i) a BERT-based encoder fine-tuned on the Weibo-100k dataset for sentiment analysis, and (ii) a RoBERTa-based encoder fine-tuned on the augmented ToxiCN dataset for toxicity detection. An adaptive fusion module integrates the encoder outputs, followed by pseudo-sequence projection and the GLDA module, which jointly capture both contextual and fine-grained semantic dependencies. The fused representation is then classified via a Kolmogorov-Arnold Network (KAN), enabling precise and interpretable decision boundaries. Extensive experiments on the COLDataset demonstrate that SenTox-GLDA achieves state-of-the-art performance, reaching an F1-score of 82.56% and AUC of 92.04%, surpassing strong baselines by 3.26% (F1) and 2.87% (AUC). These results confirm that explicitly integrating emotional cues with toxicity features, together with hierarchical attention mechanisms, substantially enhances detection robustness. Overall, our framework offers a scalable and interpretable solution for abuse detection in linguistically complex environments, with particular efficacy in Chinese social media. Codes will be made available at: https://github.com/kanglzu/SenTox-GLDA.

Disclaimer: The samples presented by this paper may be considered offensive or vulgar.

1. Introduction

Online platforms have become indispensable communication infrastructures in contemporary digital society, enabling unprecedented opportunities for self-expression, information dissemination, and social interaction [34]. However, these benefits are increasingly undermined by the proliferation of abusive content—ranging from overt toxicity to subtle hostility—which threatens platform integrity, distorts public discourse, and inflicts psychological harm [16, 9]. This issue is particularly pronounced in Chinese social media, where linguistic idiosyncrasies such as implicit sarcasm, homophone substitutions, emoji-laden expressions, and culture-specific colloquialisms pose significant challenges to conventional detection methods [7].

Existing moderation paradigms remain limited. Keyword filtering [30] and rule-based pattern matching [43] often fail against semantic obfuscation tactics. Similarly, single-stream transformer architectures like BERT [19], typically treat toxicity as a monolithic construct, overlooking the intricate interplay between emotional valence and malicious intent. Consequently, these models struggle to generalize to nuanced or culturally embedded abuse, such as sarcastic compliments or emotionally charged but non-explicitly toxic remarks.

Detecting online abuse in Chinese differs significantly from English due to linguistic and cultural complexity. Sarcasm, homophones, emojis, and colloquial expressions are frequently used to obscure toxic intent; for example, homophones bypass filters, and emojis can dilute or distort harmful messages [40]. Models pre-trained primarily on English thus have limited generalization to Chinese social media, failing to capture subtle, context-dependent toxicity [2, 33].

*Corresponding author: Zhongfeng Kang (kangzf@lzu.edu.cn)
yutongwang2023@lzu.edu.cn (Y. Wang); kangzf@lzu.edu.cn (Z. Kang); yjiaxue2023@lzu.edu.cn (J. Yang); fanxp2023@lzu.edu.cn (X. Fan); wuzj2023@lzu.edu.cn (Z. Wu); yangst@swufe.edu.cn (S. Yang); zhaoqh@hfuu.edu.cn (Q. Zhao); sls530@skku.edu (Z. Song)
ORCID(s): 0009-0000-0722-630X (Y. Wang); 0000-0001-9025-0748 (Z. Kang); 0009-0003-3153-5323 (J. Yang); 0009-0000-5029-6768 (X. Fan); 0009-0005-9903-2663 (Z. Wu); 0000-0003-2436-0580 (S. Yang); 0000-0003-4906-7049 (Q. Zhao); 0000-0003-4155-2410 (Z. Song)

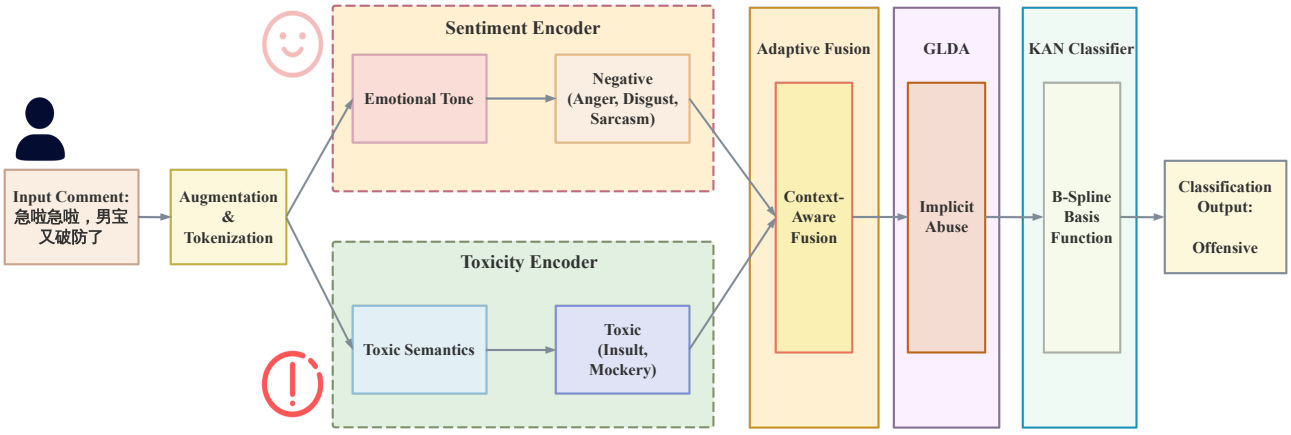

Figure 1: Workflow of the proposed SenTox-GLDA framework. The model first applies data augmentation and tokenization, then encodes comments in parallel with a sentiment encoder (capturing emotional tone) and a toxicity encoder (capturing abusive semantics). The representations are adaptively fused, refined by the Global-Local-Differential Attention (GLDA) module, and finally classified with a Kolmogorov–Arnold Network (KAN) into offensive or non-offensive categories.

This motivates the research question: *Can a multi perspective, multi-scale architecture that integrates both sentiment and toxicity cues enable more accurate and interpretable detection of abusive content in Chinese social media?* We posit that abusive language often arises from the combination of negative sentiment and toxic intent [23], rather than either dimension in isolation. For instance, the sarcastic comment " 急啦急啦，男宝又破防啦 " ("Awww, is the widdle mansplainy getting triggered again?") simultaneously conveys hostility, sarcasm, and implicit gender insult. Capturing both emotional tone and semantic toxicity is therefore crucial for robust detection, yet existing single-stream models lack the capacity to jointly model these dimensions.

To address this gap, we propose SenTox-GLDA, a dual-encoder architecture that explicitly models sentiment and toxicity in parallel. Sentiment and toxicity features are encoded separately, enabling more effective capture of emotional tone and malicious intent. These representations are fused via an adaptive fusion module that dynamically adjusts feature weights according to context, yielding a more accurate and context-sensitive representation. Additionally, we introduce a Global-Local-Differential Attention (GLDA) module to capture both local syntactic patterns and global semantic trends. For classification, we employ Kolmogorov-Arnold Networks (KANs) [21, 20], which offer strong non-linear modeling capacity with fewer parameters, enhancing both generalization and efficiency. We further explore KAN variants with diverse basis functions to improve performance. Figure 1 illustrates the processing workflow on the sample comment.

Extensive experiments on augumented COLDataset [15] validate the effectiveness of SenTox-GLDA. Our model achieves an F1 score of 82.56% and an AUC of 92.04%, outperforming strong transformer baselines and commercial moderation APIs. These results affirm that *fusing sentiment and toxicity signals at multiple scales enhances both accuracy and interpretability in detecting abusive Chinese comments*. By capturing emotional tone and malicious intent simultaneously, SenTox-GLDA effectively handles nuanced forms of abuse, including sarcasm and irony.

The main contributions of this paper are summarized as follows:

1. A dual-encoder framework that separately encodes emotional and toxic signals for nuanced abuse detection.
2. GLDA, a novel attention module that efficiently models multi-scale context via multi-order differential operations and global-local integration.
3. KAN-based classification, which improves non-linear modeling capacity while maintaining computational efficiency.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed SenTox-GLDA framework. Section 4 presents experimental results, including comparisons with state-of-the-art methods, ablation studies, and interpretability analyses. Finally, Section 5 concludes the paper, discusses practical implications, and outlines directions for future research.

2. Related Work

Toxic comment classification has drawn increasing attention across both English and non-English languages. Early approaches primarily relied on keyword spotting and rule-based filtering. While simple and efficient, these methods struggle to capture subtle emotional cues and indirect expressions such as sarcasm or homophone-based substitutions [14]. Moreover, they fail to account for the linguistic creativity of Chinese internet culture, where the frequent use

of emojis and colloquial innovations often obscures the true intent behind messages.

Large Language Models (LLMs) like omni-moderation-latest and text-moderation-latest[1] developed by OpenAI have set new benchmarks in toxic content detection [3, 4]. These models, trained on massive corpora, have significantly advanced the state-of-the-art in text classification for both English and non-English languages [13]. However, when applied to Chinese social media, they face significant challenges. While these models have performed well on Western datasets, they are typically not pre-trained on data that reflects the unique linguistic characteristics of Chinese online abuse. For example, they often misinterpret sarcasm, homophone-based expressions, and emoji usage—frequent features of Chinese internet communication [11, 26]. Even for many Chinese-based LLMs like DeepSeek[2] and Qwen[3], identifying abuse expressed through indirect or implicit language, such as backhanded compliment or sarcasm is also difficult [41], which are common in Chinese social media interactions.

Recognizing these limitations, recent research has shifted toward specialized pre-trained models tailored for Chinese online discourse. A representative example is COLDET [15], a BERT-based model pre-trained on the COLDataset, specifically designed for toxicity detection in Chinese contexts. Beyond BERT variants, deep learning architectures such as CNNs [27], LSTMs [6, 24], transformer-based models [1], and hybrid frameworks [34, 29, 31, 18] have also been widely adopted. These models, when fine-tuned on datasets like ToxiCN [22], demonstrate strong performance in detecting explicit toxic expressions. Nevertheless, they often struggle to generalize to the rapidly evolving landscape of internet slang, emoji-based communication, and implicit abuse [17]. In particular, they fall short in capturing emotional intent in non-literal expressions such as sarcasm or emotionally charged but non-explicit remarks.

To address these gaps, we propose SenTox-GLDA, a dual-encoder architecture designed to model sentiment and toxicity in parallel. By explicitly capturing both the emotional tone and the toxic intent of messages, our framework provides a more holistic understanding of abusive language. In addition, we employ advanced data augmentation strategies—including emoji substitution, homophone replacement, and synonym transformations—to enhance robustness in real-world Chinese social media scenarios. By integrating sentiment and toxicity cues within a unified framework, SenTox-GLDA overcomes the limitations of both large-scale LLMs and smaller pre-trained models, offering a more accurate and interpretable solution for abusive language detection.

[1] OpenAI offers official API of these two models through https://platform.openai.com/docs/guides/moderation. According to OpenAI, omni-moderation-latest is newer and is the better choice for application, thus this paper mainly use omni-moderation-latest instead of text-moderation-latest in the following research.

[2] https://www.deepseek.com/

[3] https://qwen.ai

3. Methodology

3.1. Overall Framework

The SenTox-GLDA architecture (Figure 2) presents a unified framework for Chinese online abuse detection, systematically integrating multiple innovative modules into a coherent pipeline. The design emphasizes parallel feature extraction and hierarchical semantic processing. It begins with an input transformation layer, which employs linguistically motivated data augmentation strategies—such as homophone substitution and emoji replacement—to enhance robustness against adversarial variations in social media text.

A dual-encoder backbone is then applied:

- A BERT-based sentiment encoder, fine-tuned on Weibo-100k, captures emotional dimensions.
- A RoBERTa-based toxicity encoder, trained on the emoji-augmented ToxiCN dataset, extracts toxicity-specific cues.

Their outputs are adaptively fused through a gating mechanism that assigns context-sensitive weights at both token and sequence levels. The fused representation is subsequently reshaped into a structured pseudo-sequence, enabling multi-scale semantic modeling.

To capture both localized abusive patterns and discourse-level aggression, we introduce a Global-Local Differential Attention (GLDA) module. This module integrates:

- Local Differential Attention (LDA): sliding-window differentials to capture short-range syntactic and lexical cues.
- Global Differential Attention (GDA): downsampled attention to detect sarcasm, context-driven toxicity, and long-range dependencies.

Finally, classification is performed by a Kolmogorov–Arnold Network (KAN), which approximates nonlinear decision boundaries in the fused sentiment–toxicity space. Unlike conventional fully connected layers, KAN offers both computational efficiency through basis function sharing and enhanced interpretability via function decomposition.

3.2. Data Augmentation

To enhance the robustness and generalization of the model, particularly in handling informal, emoji-rich, and ambiguous abusive content, we design a multi-strategy data augmentation pipeline(Figure 3). This pipeline introduces linguistic diversity and controlled noise, helping the encoders learn more resilient representations. This pipeline includes:

- **Emoji augmentation**: Emojis are frequently used in social media to convey emotions, intensify tone, or circumvent moderation filters while retaining toxic intent [5, 38]. Despite existing attempts to incorporate emojis into modeling, performance on Chinese datasets

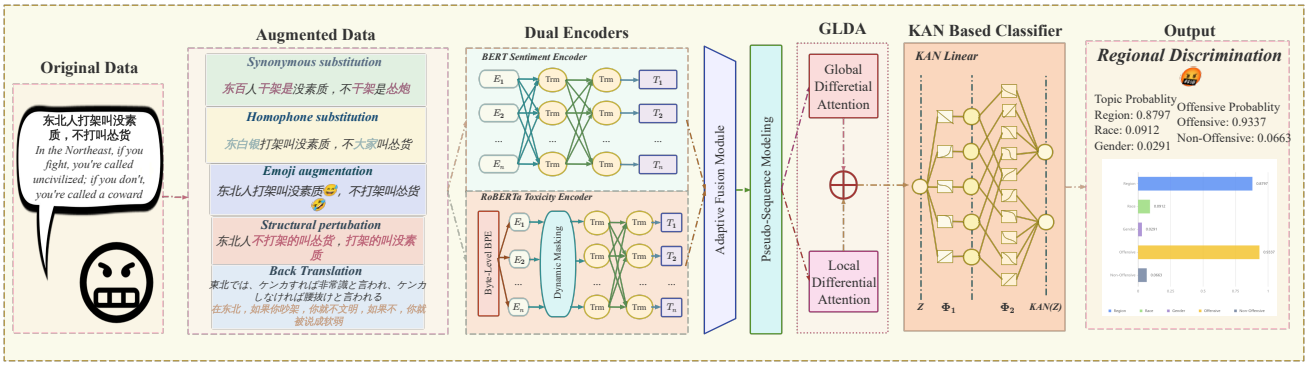

Figure 2: Overall architecture of the proposed SenTox-GLDA framework. Raw Chinese comments are first augmented through synonym replacement, homophone substitution, emoji insertion, structural perturbation, and back translation to simulate diverse linguistic variations. The processed text is then tokenized and encoded in parallel by a sentiment encoder (capturing affective polarity) and a toxicity encoder (capturing offensive semantics). Their hidden representations are combined via an adaptive fusion module and further enhanced by the Global-Local Differential Attention (GLDA) mechanism, which integrates multi-frequency features at both discourse and token levels. Finally, a Kolmogorov–Arnold Network (KAN) classifier yields interpretable predictions, including fine-grained discrimination across categories such as region, race, and gender.

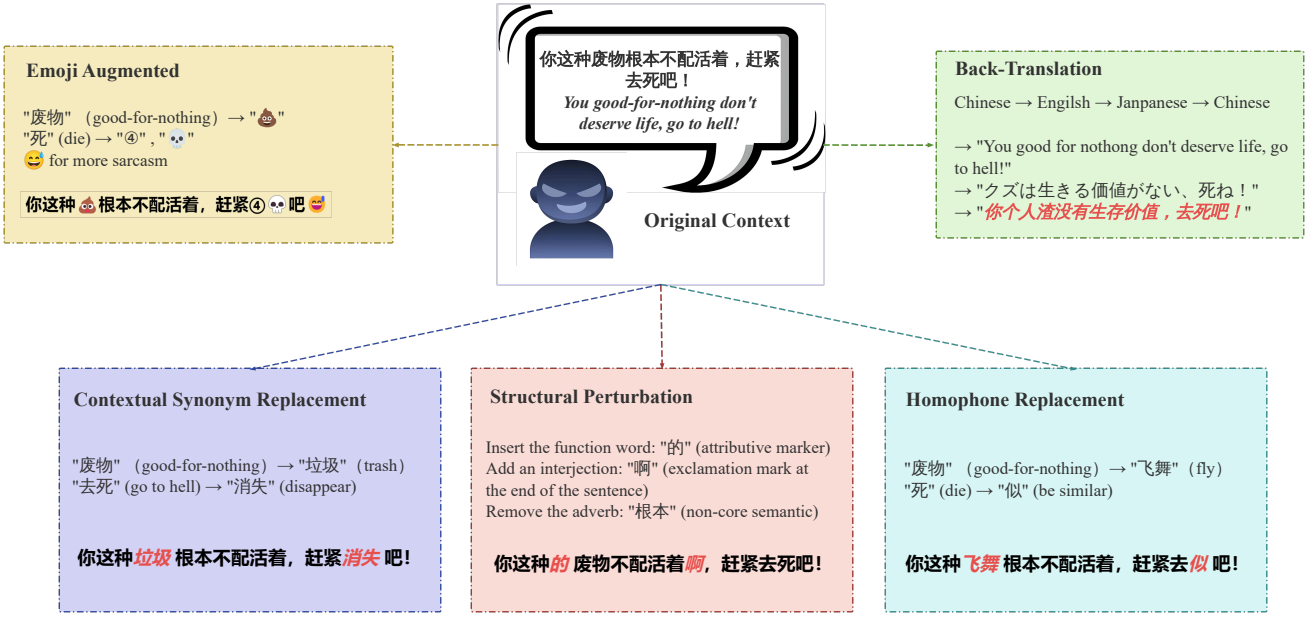

Figure 3: Multi-strategy data augmentation pipeline for Chinese abusive text. Starting from an original comment, the framework generates diverse variants through five complementary strategies: contextual synonym replacement, homophone substitution, emoji augmentation, structural perturbation, and back translation. These transformations simulate realistic linguistic variations such as sarcasm, slang, and cross-lingual rephrasing, thereby improving model robustness against implicit and adversarial expressions of abuse.

remains limited. Inspired by the NMSL project[4], our emoji augmentation replaces words with semantically or phonetically similar emojis. Three augmentation modes are designed: (i) *light mode*—direct replacement with basic emoji equivalents; (ii) *deep mode*—phonetic transformations where words are replaced with emojis sounding similar; (iii) *mixed mode*—a combination of both strategies for nuanced replacements. Unlike NMSL project, this method expands the emoji-Chinese mapping dictionary and adds a large number of new expressions that are

[4]https://github.com/THUzhangga/NMSL

more in line with contemporary Chinese social media comments. These augmentations reflect the sarcastic, indirect, and colloquial styles prevalent in Chinese social media, enabling the model to generalize across diverse emoji-laden toxic expressions [49, 39].

- **Contextual synonym replacement**: We adopt the NLPCDA (Natural Language Chinese Data Augmentation) open-source library, which provides curated synonym dictionaries tailored for Chinese NLP tasks. Synonyms are sampled contextually from semantic lexicons such as TongYiCi Cilin and HowNet-inspired mappings. This preserves syntactic roles while diversifying lexical surface forms, thereby enhancing resilience against word-level adversarial attacks.

- **Structural perturbation**: We modify function words and discourse particles while preserving the abusive core semantics. For instance, users often omit particles or insert fillers like " 啊 " or " 呢 " in natural online interactions. Such perturbations introduce surface-level diversity without altering toxicity [28].

- **Homophone replacement**: This strategy exploits the phonetic ambiguity of Chinese, replacing characters with identical pinyin but different glyphs. Such substitutions mimic real-world evasion tactics, where toxic intent is preserved while bypassing keyword filters.

- **Back-translation**: We perform iterative translation through English and Japanese, producing linguistically valid paraphrases. This exposes the model to structurally diverse yet semantically equivalent forms of abuse, improving generalization across varied phrasing styles.

Figure 3 illustrates the application of these five augmentation methods to the toxic sentence " 你这种废物根本不配活着，赶紧去死吧! " (You good-for-nothing don't deserve life, go to hell!). Each augmentation branch showcases both the linguistic operation and its practical output:

(i) Contextual synonym replacement substitutes derogatory nouns (废物 → 垃圾) and violent verbs (去死 → 消失) while preserving syntax.

(ii) Homophone replacement employs phonetic similarity (废物 → 飞舞; 死 → 似) to simulate filter evasion.

(iii) Emoji augmentation substitutes terms with visual tokens (shit for 废物, ④ for 死).

(iv) Structural perturbation introduces particles (的, 啊) and removes adverbs (根本), producing naturalistic surface variations.

(v) Back-translation yields culturally adapted paraphrases (人渣没有生存价值, "Scum has no right to exist").

3.3. Pretrained Dual-Encoders

The proposed dual-encoder framework integrates two pretrained transformer models, each specialized for complementary aspects of abusive language:

- BERT-base-Chinese [12], fine-tuned on the Weibo-100k corpus, is used to extract emotional sentiment representations. The dataset captures casual, expressive, and sentiment-rich user-generated text, making the encoder particularly effective in modeling emotional tones.

- RoBERTa-wwm-ext-large [44], fine-tuned on the Toxic-CN dataset augmented with the NMSL project (both light and deep emoji modes), is employed to capture toxic semantics, including slang, emoji-driven expressions, and abusive linguistic patterns common in Chinese social media.

Both encoders output a [CLS] embedding that serves as a sentence-level abstraction. The final-layer 1024-dimensional [CLS] embedding of RoBERTa provides a high-capacity representation of toxic semantics. This enhanced dimensionality and deeper architecture are particularly important for resolving context-dependent toxicity, such as distinguishing reclaimed slang in in-group interactions from malicious usage in out-group contexts. Such fine-grained modeling enables the framework to detect semantic shifts, where superficially neutral terms acquire toxicity depending on cultural or conversational context.

3.3.1. BERT Sentiment Encoder

The sentiment analysis module employs a fine-tuned BERT-base-Chinese model[5] to extract emotional features from social media text. At the core of BERT's transformer architecture lies the multi-head self-attention mechanism, which models contextual relationships between words:

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

where Q, K, and V represent the query, key, and value matrices, respectively, and d_k denotes the dimensionality. Multi-head attention allows each head to capture distinct linguistic dependencies, enabling layered contextual analysis of emotional expressions. Model optimization was conducted on the Weibo-100k dataset using cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{c} y_i \log(p_i) \tag{2}$$

Training use the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$), with a learning rate of 2×10^{-5} , linear warmup, batch size of 32, sequence length of 256, and 25 epochs. The fine-tuned model achieved 90.27% Macro-F1, 93.04% AUC

[5] https://huggingface.co/google-bert/bert-base-chinese

and 92.81% accuracy on the testset, demonstrating robust performance in detecting three key sentiment indicators: (1) explicit lexical markers, (2) implicit cues such as sarcastic phrasing, and (3) emoji-based expressions.

The selection of BERT is motivated by three architectural advantages critical for Chinese sentiment analysis: (1) Bidirectional context modeling (via masked language modeling) for accurate interpretation of word-order-dependent meaning; (2) hierarchical representation learning across 12 transformer layers, enabling automatic feature extraction from lexical to sentence-level semantics; and (3) transfer learning efficiency allows effective knowledge transplantation from general Chinese corpora to specialized sentiment tasks [12].

The Weibo-100k dataset was chosen for its authentic, user-generated content with natural emotional expressions, balanced sentiment coverage (59,993 positive vs. 59,995 negative samples), and inclusion of linguistically complex phenomena such as emojis, slang, and informal phrasing. These characteristics present real-world challenges like sarcasm and implicit sentiment, making it well-suited for fine-tuning.

Finally, the encoder's 768-dimensional [CLS] embedding provides a compact yet expressive sentence-level representation of emotional content. This representation is particularly effective for disambiguating cases where surface lexical signals contradict underlying emotional intent, thereby establishing a robust foundation for downstream sentiment-aware abuse detection.

3.3.2. RoBERTa Toxicity Encoder

The toxicity detection module employs a RoBERTa-wwm-ext-large model[6] optimized for identifying harmful content in Chinese social media text. Compared with standard BERT, RoBERTa enhances pretraining by removing the next-sentence prediction objective and employing dynamic masking, leading to richer contextual representations essential for nuanced toxicity detection.

The model architecture consists of 24 transformer layers with 1024-dimensional hidden states, processing text sequences through the multi-head self-attention mechanism. With 16 attention heads, the expanded dimensionality (d_k = 1024) allows the model to capture subtle lexical, syntactic, and pragmatic cues associated with implicit abuse. Hierarchical layers progressively aggregate signals from word-level indicators to sentence-level toxicity patterns, supporting fine-grained discrimination of nuanced online abuse.

Fine-tuning was performed on the ToxiCN dataset, which contains 12,011 comments with a roughly balanced distribution of offensive and hateful content. Using our multi-strategy data augmentation pipeline, the dataset was expanded to 36,033 samples, including both light and deep emoji-augmented variants. Training employed cross-entropy loss with weighted sampling to mitigate class imbalance, optimized using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a peak learning rate of 5×10^{-6} , linear warmup over 10% of steps, batch size 16, and maximum sequence length 512 tokens to accommodate long, slang-rich abusive comments. After 25 epochs, the fine-tuned RoBERTa model achieved state-of-the-art performance on ToxiCN (Table 1 and Figure 4), with Macro-F1 = 81.08%, precision = 81.21%, and recall = 80.97%.

[6] https://huggingface.co/hfl/chinese-roberta-wwm-ext

Model	Precision (%)	Recall (%)	Macro-F1 (%)
BiLSTM*	73.46	72.95	73.20
BiLSTM	75.44	74.82	75.13
BERT*	80.01	79.92	79.96
BERT	81.03	80.55	80.79
RoBERTa*	80.37	79.61	79.98
RoBERTa	**81.21**	**80.97**	**81.08**

Table 1
Performance comparison on augmented vs. non-augmented ToxiCN (* indicates training on original dataset).

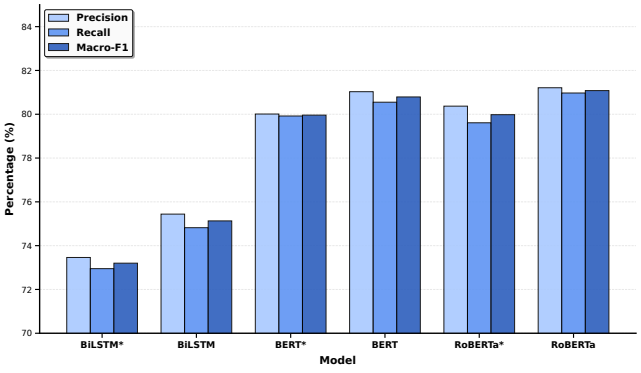

Figure 4: Performance comparison on augmented and non-augmented ToxiCN.

These results demonstrate that data augmentation significantly enhances model robustness, particularly for RoBERTa. It effectively captures subtle toxicity features, including sarcasm, emoji-based expressions, and slang. In comparison, simpler models like BiLSTM benefit less from augmentation (Macro-F1 improving only from 73.20% to 75.13%), highlighting their limitations in leveraging complex contextual cues. BERT shows moderate improvement (Macro-F1 = 80.79%) but still underperforms RoBERTa, reflecting the latter's superior ability to model complex contextual dependencies crucial for detecting nuanced abusive language.

3.4. Adaptive Fusion Module

To effectively leverage both sentiment and toxicity representations, we introduce an adaptive fusion module that dynamically balances the contribution of each feature type. Leth_s and h_t denote the [CLS] embeddings from the BERT sentiment encoder and RoBERTa toxicity encoder, respectively. The fused representation h_f is computed as:

$$g = \sigma(W_g[h_s; h_t] + b_g) \quad h_f = g \cdot h_s + (1 - g) \cdot h_t \quad (3)$$

where, $W_g \in \mathbb{R}^{d \times 2d}$ and $b_g \in \mathbb{R}$ are trainable parameters, and $\sigma(\bullet)$ is the sigmoid activation. The gating vector g learns to adaptively weight sentiment and toxicity signals based on

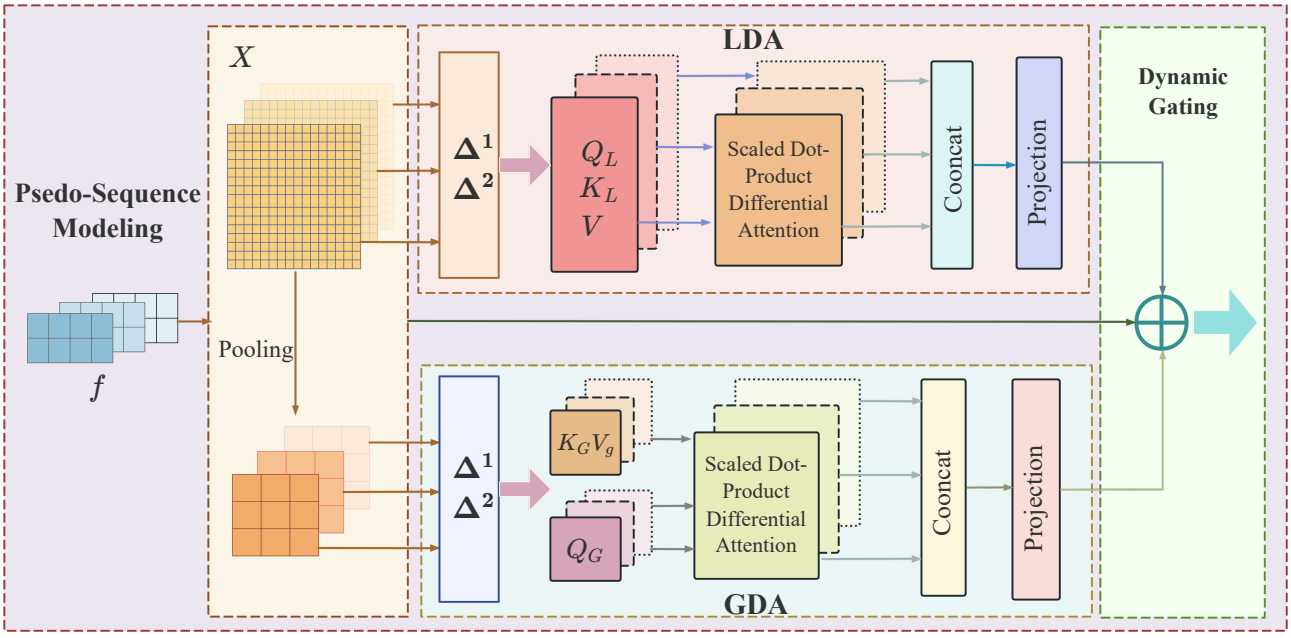

Figure 5: Overall architecture of the proposed Global-Local Differential Attention (GLDA). The module first transforms input features into pseudo-sequences and extracts multi-scale representations through pooling. Local Differential Attention (LDA) captures fine-grained token-level variations, while Global Differential Attention (GDA) models discourse-level dependencies. Their outputs are dynamically integrated via gating and projection to produce a fused representation that balances local detail with global contextual information.

context, allowing the model to emphasize emotional cues in nuanced cases or prioritize toxic patterns in explicitly harmful content.

3.5. Global-Local-Differential Attention

As illustrated in Figure 5, the Global-Local-Differential Attention (GLDA) mechanism represents a significant advancement over conventional attention models by jointly capturing local variations and global dependencies within sequences. Unlike standard attention, which focuses on pairwise token interactions, GLDA incorporates differential operations to explicitly model subtle shifts in semantics, enabling enhanced sensitivity to nuanced patterns—critical in applications such as sentiment and toxicity analysis where small tonal changes can alter meaning.

The GLDA framework consists of two complementary components: Local Differential Attention (LDA) and Global Differential Attention (GDA).

- **Local Differential Attention (LDA)**: LDA captures short-range semantic variations by computing first- and second-order differences across token embeddings. These differential features encode subtle shifts in local context, such as changes in sentiment or intensity over adjacent words.
- **Global Differential Attention (GDA)**: GDA operates on a downsampled representation of the input sequence to model long-range dependencies. This component captures overarching trends and high-level semantic patterns that may span the entire sequence, complementing the fine-grained insights from LDA.

To integrate these signals, GLDA employs a dynamic fusion mechanism with a learnable weighting factor α, producing the final output as a weighted combination of local and global features. Additionally, residual connections preserve baseline information and prevent loss of critical semantic content during processing.

The procedural logic of GLDA is summarized in Algorithm 1, where $\mathbb{DiffAttn}$ computes differential attention scores and P_{pool} handles the downsampling for global context extraction. This design allows GLDA to flexibly adapt to sequences of varying length and complexity, maintaining high fidelity in capturing both micro-level shifts and macro-level structures in the data.

3.5.1. Psedo-Sequence Modeling

o facilitate the joint modeling of local and global semantics, the fused embedding f is first transformed into a structured pseudo-sequence through a learnable linear projection:

$$X = T(f) = W_p f^T \in \mathbb{R}^{N \times d_k}, W_p \in \mathbb{R}^{N \times d} \tag{4}$$

where N denotes the pseudo-sequence length and d_k is the hidden dimension for the subsequent GLDA module. his

Algorithm 1 GLDA Core Logic

function GLDA($X, W_q, W_k, W_v, P_{pool}, \alpha, M$)
  Inputs: $X \in \mathbb{R}^{\text{batch}\times N\times d}$, W_q, W_k, W_v, $P_{pool} \in \mathbb{R}^{M\times N}$, α, M
  Outputs: $H \in \mathbb{R}^{\text{batch}\times(N-2)\times d}$
  $Q \leftarrow XW_q$; $K \leftarrow XW_k$; $V \leftarrow XW_v$
  LDA:
  $H_{local} \leftarrow$ DIFFATTN(Q, K, V, d)
  GDA:
  $W_{pool} \leftarrow \text{softmax}(P_{pool}, \text{dim} = -1)$; $X_p \leftarrow W_{pool}X$
  $Q_g \leftarrow X_pW_q$; $K_g \leftarrow X_pW_k$; $V_g \leftarrow X_pW_v$
  $H_{global,down} \leftarrow$ DIFFATTN(Q_g, K_g, V_g, d)
  $H_{global} \leftarrow \text{upsample}(H_{global,down}, \text{target\_len} = N-2)$
  Fusion:
  $H \leftarrow \alpha H_{local} + (1-\alpha)H_{global} + X[:, 2:, :]$
  Return H
end function
function DIFFATTN($Q_{in}, K_{in}, V_{in}, d$)
  $\Delta_1 Q \leftarrow Q_{in}[:, 1:, :] - Q_{in}[:, :-1, :]$
  $\Delta_2 Q \leftarrow \Delta_1 Q[:, 1:, :] - \Delta_1 Q[:, :-1, :]$
  $\Delta_1 K \leftarrow K_{in}[:, 1:, :] - K_{in}[:, :-1, :]$
  $\Delta_2 K \leftarrow \Delta_1 K[:, 1:, :] - \Delta_1 K[:, :-1, :]$
  $Q_L \leftarrow \text{concat}(\Delta_1 Q[:, :-1, :], \Delta_2 Q, \text{ax} = -1)$
  $K_L \leftarrow \text{concat}(\Delta_1 K[:, :-1, :], \Delta_2 K, \text{ax} = -1)$
  $scores \leftarrow (Q_L K_L^T)/\sqrt{2d}$
  $H_{out} \leftarrow \text{softmax}(scores, \text{dim} = -1)V_{in}[:, 2:, :]$
  Return H_{out}
end function

transformation constructs a set of orthogonal basis vectors $\{\varphi_i\}_{i=1}^N$ such that each projected component is given by:

$$X_i = \langle f, \varphi_i \rangle, \quad \varphi_i \in \mathbb{R}^d \tag{5}$$

This projection serves two critical purposes: first, it establishes a consistent feature space for differential operations within GLDA, ensuring compatibility between local and global computations; second, it generates positional embeddings that preserve semantic structure while enabling precise gradient computations along the pseudo-sequence dimension.

By converting a single fused representation into a pseudo-sequence, the model effectively bridges the gap between sentence-level abstraction and sequence-level differential reasoning, enabling GLDA to exploit both fine-grained local cues and broad global dependencies.

3.5.2. Local Differential Attention

Local Differential Attention (LDA) is designed to capture short-range, fine-grained semantic shifts within a pseudo-sequence by incorporating differential representations of the input. Unlike conventional attention mechanisms that compute attention weights solely from direct query–key similarity, LDA explicitly leverages first- and second-order differences to highlight local semantic shifts.

The first step in LDA is to compute the differential representations for the input sequence. Given a pseudo-sequence $X \in \mathbb{R}^{N\times d}$, its query and key matrices $Q, K \in \mathbb{R}^{N\times d_h}$ are first transformed into differential signals.

First-order differences capture semantic changes between adjacent tokens:

$$\Delta^1 Q_t = Q_t - Q_{t-1} \tag{6}$$

$$\Delta^1 K_t = K_t - K_{t-1} \tag{7}$$

Second-order differences capture curvature, i.e., higher-order shifts over longer spans:

$$\Delta^2 Q_t = Q_t - 2Q_{t-1} + Q_{t-2} \tag{8}$$

$$\Delta^2 K_t = K_t - 2K_{t-1} + K_{t-2} \tag{9}$$

The enhanced queries and keys are constructed by concatenating the original signals with their differential forms:

$$\vec{Q}_t = \begin{bmatrix} Q_t \\ \Delta^1 Q_t \\ \Delta^2 Q_t \end{bmatrix}, \quad \vec{K}_t = \begin{bmatrix} K_t \\ \Delta^1 K_t \\ \Delta^2 K_t \end{bmatrix} \tag{10}$$

This augmentation enriches the representational space, enabling the model to attend not only to the token itself but also to its local semantic trajectory.

The second step in LDA is to compute the local attention weights.To preserve locality, attention is restricted to a sliding window of width w centered at each position i. The attention score between token i and a neighbor $j \in [i - w/2, i + w/2]$ is computed as:

$$A_{i,j}^{\text{loc}} = \frac{\exp\left(\widehat{Q}_i \cdot \widehat{K}_j/\sqrt{d_h}\right)}{\sum_{k=i-w/2}^{i+w/2} \exp\left(\widehat{Q}_i \cdot \widehat{K}_k/\sqrt{d_h}\right)} \tag{11}$$

The output representation at position i is then given by the weighted sum of local values:

$$H_i^{\text{loc}} = \sum_{j=i-w/2}^{i+w/2} A_{i,j}^{\text{loc}} V_j \tag{12}$$

Through multi-order differential modeling, LDA enhances the encoder's sensitivity to subtle semantic fluctuations, making it particularly effective in detecting sarcasm, nuanced insults, and context-dependent emotional shifts that are common in abusive online discourse.

3.5.3. Global Differential Attention

Global Differential Attention (GDA) is designed to capture long-range dependencies and global semantic trends across a sequence. In contrast to Local Differential Attention (LDA), which emphasizes short-range variations, GDA compresses the sequence into a lower-resolution representation, enabling the model to identify broad, high-level semantic patterns that span the entire input. This makes GDA especially effective for tasks requiring long-term context, such as document-level sentiment analysis or discourse modeling.

The input pseudo-sequence $\mathbf{X} \in \mathbb{R}^{N\times d}$ is first projected into a shorter sequence $X_{pool} \in \mathbb{R}^{M\times d}(M \ll N)$ using a learnable pooling operator:

$$X_{\text{pool}} = \text{Softmax}(P_{\text{pool}}) \cdot X \tag{13}$$

where $P_{\text{pool}} \in \mathbb{R}^{M\times N}$ is a trainable pooling matrix ensuring that the compressed sequence preserves essential semantic information.

As in LDA, first- and second-order differences are computed for queries, keys, and values to capture the changes between consecutive elements in the compressed sequence. This augmentation enriches the global representation by encoding both semantic baselines and their evolving trajectories.

Attention scores are then computed using scaled dot-product attention across the entire compressed sequence:

$$A^{\text{glb}}_{i,j} = \frac{\exp\left(\vec{Q}_i \cdot \vec{K}_j/\sqrt{d_h}\right)}{\sum_{k=1}^{M}\exp\left(\vec{Q}_i \cdot \vec{K}_k/\sqrt{d_h}\right)} \tag{14}$$

where $\vec{Q}_i$ and $\vec{K}_i$ are the projected queries and keys, and d_h is the dimensionality of the attention space.

The aggregated values are obtained as:

$$H^{\text{glb}}_{\text{pool}} = A^{\text{glb}} \cdot \widetilde{V} \tag{15}$$

Finally, the global representation is expanded back to the original sequence length N using an interpolation matrix $U \in \mathbb{R}^{M\times N}$:

$$H^{\text{glb}} = U \cdot H^{\text{glb}}_{\text{pool}} \tag{16}$$

By integrating sequence compression, differential augmentation, and global attention, GDA provides a coarse-to-fine representation of long-range semantic dependencies. Together with LDA, it enables the model to balance fine-grained local cues with broad global context, a capability crucial for detecting nuanced toxic expressions in social media discourse.

3.5.4. Dynamic Gating With Residuals

The dynamic fusion mechanism integrates the outputs from Local Differential Attention (LDA) and Global Differential Attention (GDA) by adaptively balancing short-range and long-range dependencies. This is achieved through a learnable gating weight α, which is computed contextually from both local and global features as well as the original input sequence.

Formally, the fused representation is given by:

$$H_{\text{fused}} = \alpha \cdot H^{\text{glb}} + (1-\alpha)\cdot H^{\text{loc}} \tag{17}$$

Here, $\alpha \in [0,1]$ determines the relative contribution of global and local information. Unlike a fixed weight, α is dynamically estimated through a learnable function that aggregates global, local, and raw input statistics:

$$\alpha = \sigma\left(w^T\left[\text{Pool}\left(H^{loc}\right);\text{Pool}\left(H^{glb}\right);\text{Pool}(X)\right)\right. \tag{18}$$

where w is a learnable weight vector and σ is the sigmoid activation function and $\text{Pool}(\bullet)$ denotes a global pooling operation. This formulation allows the fusion gate to automatically emphasize local differential cues (e.g., sarcasm or subtle lexical shifts) or global semantic trends (e.g., long-range context) depending on the input.

To prevent information loss and stabilize training, a residual connection is applied by adding the original sequence embedding X to the fused representation, followed by layer normalization:

$$H_{\text{final}} = \text{LayerNorm}\left(H_{\text{fused}} + X\right) \tag{19}$$

This residual-enhanced fusion ensures that baseline semantic information is preserved while enriching it with context-aware global–local differentials.

3.5.5. Noise Variance Analysis and Fusion Derivation

To evaluate the robustness of the proposed GLDA framework under noisy conditions, we conduct a formal analysis of noise propagation and its impact on variance across the local, global, and fusion pathways.

Let the input sequence $\mathbf{X} \in \mathbb{R}^{N\times d}$, corrupted by additive noise $\mathbf{E}_X \in \mathbb{R}^{N\times d}$ with the following properties:

$$\mathbb{E}[\mathbf{E}_{X,ij}] = 0, \quad \text{Var}[\mathbf{E}_{X,ij}] = \sigma_X^2 \tag{20}$$

$$\mathbb{E}[\mathbf{E}_{X,ij}\mathbf{E}_{X,ts}] = \delta_{ij}\delta_{ts}\sigma_X^2 \tag{21}$$

where the noise is assumed i.i.d. with zero mean and variance σ_X^2.

Noise amplification in Local Differential Attention (LDA): For a noisy observation $\widetilde{\mathbf{X}} = \mathbf{X}_t^* + \mathbf{E}_{X,t}$, the first-order difference and is:

$$\Delta^1\widetilde{\mathbf{X}} = (\mathbf{X}t^* - \mathbf{X}t - 1^*) + (\mathbf{E}X,t - \mathbf{E}X,t-1) \tag{22}$$

yielding variance:

$$\text{Var}[\Delta^1\widetilde{\mathbf{X}}] = 2\sigma_X^2 \tag{23}$$

Since LDA concatenates original, first-order, and second-order differences, its effective noise variance is:

$$\text{Var}[\widetilde{\mathbf{Q}}] \approx \sigma_X^2 + 2\sigma_X^2 + 6\sigma_X^2 = 9\sigma_X^2 \tag{24}$$

indicating that local differencing amplifies noise variance.

Noise attenuation in Global Differential Attention (GDA): For GDA, pooling reduces the sequence length. With pooling matrix $\mathbf{W} = \text{Softmax}(\mathbf{P}_{\text{pool}}) \in \mathbb{R}^{M\times N}$, the pooled representation is:

$$\mathbf{X}\text{pool} = \mathbf{W}\mathbf{X} \tag{25}$$

and its noise variance:

$$\text{Var}[\mathbf{E}\text{pool}, ij] = \sigma_X^2\sum_{k=1}^{N} w_{ik}^2 \tag{26}$$

For nearly uniform weights $w_{ik} \approx \frac{1}{N}$:

$$\mathrm{Var}[\mathbf{E}_{\mathrm{pool},ij}] \approx \frac{\sigma_X^2}{N} \tag{27}$$

showing that pooling strongly suppresses noise. First- and second-order differences of pooled sequences further yield:

$$\mathrm{Var}[\Delta^1 \widetilde{\mathbf{X}}] \approx 2 \cdot \frac{\sigma_X^2}{N} \tag{28}$$

$$\mathrm{Var}[\Delta^2 \widetilde{\mathbf{X}}] \approx 6 \cdot \frac{\sigma_X^2}{N} \tag{29}$$

demonstrating that GDA is inherently more robust to noise compared to LDA.

Fusion error analysis: For the used representation $\mathbf{H} = \alpha\mathbf{G} + (1-\alpha)\mathbf{L}$ where $\mathbf{L} = \mathbf{W}_L\mathbf{D}$, with ideal output $\mathbf{H}^* = \alpha^*\mathbf{G}^* + (1-\alpha^*)\mathbf{L}^*$, the error $\delta_{\mathbf{H}} = \mathbf{H} - \mathbf{H}^*$ has first-order Taylor expansion:

$$\delta_{\mathbf{H}} \approx \alpha\delta_{\mathbf{G}} + (1-\alpha)\delta_{\mathbf{L}} + (\mathbf{G}-\mathbf{L})\delta_\alpha \tag{30}$$

leading to the variance upper bound:

$$\mathrm{Var}[\delta_{\mathbf{H}}] \le 3\left[\alpha^2\sigma_G^2 + (1-\alpha)^2\sigma_L^2 + \|\mathbf{G}-\mathbf{L}\|_2^2\sigma_\alpha^2\right] \tag{31}$$

where $\sigma_G^2 = \frac{\sigma_D^2}{k}$ (noise suppression in GDA) and $\sigma_L^2 = \sigma_D^2$ (baseline LDA variance).

This analysis reveals a noise–robustness tradeoff: LDA amplifies noise due to differential operations, while GDA attenuates it via pooling. The dynamic fusion mechanism adaptively balances these pathways, ensuring stability under noisy conditions. Thus, GLDA achieves both fine-grained sensitivity (via LDA) and robust global consistency (via GDA), maintaining resilience to real-world noisy inputs.

3.6. Kolmogorov–Arnold Networks for Classification

The output representation from GLDA is fed into a Kolmogorov–Arnold Network (KAN), which replaces the conventional MLP classifier. According to the Kolmogorov–Arnold representation theorem, any multivariate continuous function f defined on a bounded domain can be represented as a finite composition of univariate continuous functions:

$$y = \sum_{i=1}^{k} \theta_i \cdot \phi_i(h) \tag{32}$$

where ϕ_i are learnable basis functions, and θ_i are trainable coefficients. This formulation endows KAN with three key advantages:

- The ability to approximate highly non-linear decision boundaries with fewer parameters.
- Smooth, differentiable, and interpretable decision functions enabled by localized spline interpolation.

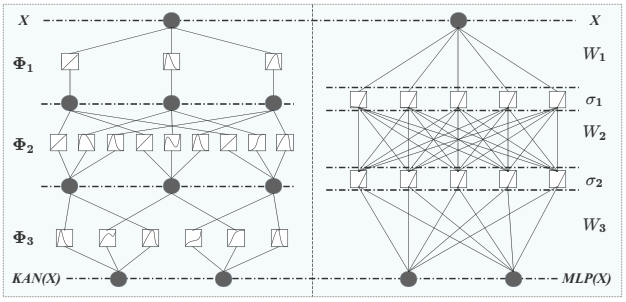

Figure 6: Comparison between KAN and MLP. Unlike MLPs that rely on stacked linear projections with nonlinear activations, KAN introduces functional bases at each layer, enabling more expressive feature transformations and improved interpretability while maintaining comparable architectural simplicity.

- Improved generalization and faster convergence under limited data, compared to traditional MLPs [46].

As shown in Figure 6, the classifier employs three stacked KANLinear layers with the following dimensional transformations:

$$[768 \to 512] \to [512 \to 256] \to [256 \to 2] \tag{33}$$

Key implementation specifications:

- **B-spline basis**: Spline order 3 with grid size 5
- **Normalization**: LayerNorm after each KAN layer for stability
- **Basis functions**: Defined recursively B-spline formulation

The B-spline basis functions used in KAN are formally defined as:

$$B_{i,0}(x) = \begin{cases} 1, & \text{if } t_i \le x \le t_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

$$B_{i,p}(x) = \frac{x - t_i}{t_{i+p} - t_i} B_{i,p-1}(x) + \frac{t_{i+p+1} - x}{t_{i+p+1} - t_{i+1}} B_{i+1,p-1}(x) \tag{35}$$

This recursive formulation enables smooth function approximation with localized control, allowing the classifier to better capture fine-grained semantic differences in toxic comment detection.

To comprehensively assess the effectiveness of SenTox-GLDA, we conduct extensive experiments on real-world abusive comment detection datasets, as described in the following section.

Table 2
Model Training Parameters

Parameter	Value
GPU	NVIDIA RTX 4090 × 2
Epochs	100 (early stopping at 15)
Batch size	128
Learning rate	1×10^{-5} (cosine decay to 1×10^{-7})

4. Experimental Study

4.1. Dataset

We evaluate our model on the **COLDataset** [15], a manually annotated corpus containing 37,480 Chinese online comments collected from Weibo[7], Zhihu[8] and other Chinese social media platforms. Among them, 18,041 samples are offensive and 19,439 are non-offensive. The dataset is cleanly split into training, validation, and test sets. After augmentation, it expands to 449,760 samples with a balanced offensive rate of 48%. The augmented version reflects greater linguistic diversity, incorporating emojis, homophones, and sarcasm.

4.2. Experimental Setup

The experimental environment is summarized in Table 2. We adopt mixed-precision training (FP16/FP32 hybrid) with gradient accumulation (steps = 4) and data preloading into RAM. To mitigate overfitting and stabilize training, the first six layers of both BERT and RoBERTa encoders are frozen. All BERT and RoBERTa models mentioned henceforth are consistent with the dual-encoder backbone of SenTox-GLDA.

4.3. State-of-the-Art Models

COLDET [15]: A BERT-based fine-tuned model optimized for Chinese social media offensive language detection. Trained on COLDataset, it leverages bidirectional attention to capture both implicit biases and explicit insults, achieving strong macro-F1 performance.

BAIDUTC[9]: A proprietary moderation system developed by Baidu that combines rule-based filtering with deep learning. While it demonstrates higher precision than keyword-based approaches, its sensitivity to implicit toxic content remains limited.

XLM-R large[10]: A multilingual BERT variant pretrained on 100+ languages with cross-lingual masked objectives. By exploiting shared subword embeddings and parallel corpus alignment, it achieves state-of-the-art transfer learning, particularly effective in low-resource scenarios.

omni-moderation-latest[11]: A multimodal moderation framework developed by OpenAI, incorporating ensemble classifiers with real-time policy updates. It employs cross-modal consistency checks and adaptive thresholding based on contextual risk scoring.

4.4. Evaluation Metrics

We assess model performance using five widely adopted metrics, each offering complementary insights.

Accuracy measures the overall correctness of predictions:

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

Although intuitive, it may be misleading in imbalanced datasets, as majority classes can dominate the score.

Precision evaluates the reliability of toxic predictions:

$$\text{Precision} = \frac{TP}{TP + FP}$$

High precision is crucial for minimizing false alarms and ensuring fairness in moderation.

Recall quantifies the ability to capture all toxic instances:

$$\text{Recall} = \frac{TP}{TP + FN}$$

It is critical for comprehensive detection, though overly high recall may introduce excessive false positives, especially in nuanced contexts such as sarcasm or slang.

Macro-F1 balances precision and recall by computing their harmonic mean:

$$\text{Macro-F1} = \frac{2 \cdot (P \cdot R)}{P + R}$$

This is particularly useful for evaluating models under class imbalance.

ROC-AUC measures the threshold-independent ranking capability:

$$\text{AUC} = \int_0^1 \text{TPR}(FPR)\, d(FPR)$$

It reflects the model's robustness across different operating points, which is especially important when applying moderation to platforms with varying toxicity tolerance.

4.5. Comparison Study

We compare SenTox-GLDA against a range of state-of-the-art baselines, including BERT [12], RoBERTa [44], XLM-R large, COLDET, BAIDUTC[12], MuDA [36], and OpenAI's omni-moderation-latest, as well as several Chinese-specific large models such as DeepSeek-R1[13], DeepSeek-V3[14], Qwen-Plus-Latest and Qwen-Turbo-Latest[15]. As shown in Figure 7 and Table 3, SenTox-GLDA consistently outperforms all competitors across evaluation metrics, achieving a new state-of-the-art on the augmented COLDataset with a Macro-F1 of 82.56% and an AUC of 92.14%.

[7] https://weibo.com
[8] https://www.zhihu.com
[9] https://ai.baidu.com/tech/textcensoring
[10] https://huggingface.co/FacebookAI/xlm-roberta-large
[11] https://platform.openai.com/docs/guides/moderation
[12] https://ai.baidu.com/tech/textcensoring
[13] https://github.com/deepseek-ai/DeepSeek-R1
[14] https://github.com/deepseek-ai/DeepSeek-V3
[15] Qwen models could be visited through https://bailian.console.alibabacloud.com/.
Official APIs could be viewed in https://bailian.console.aliyun.com

Table 3
Comparison of SenTox-GLDA and baselines on the augmented COLDataset. Best results in **bold**, second best underlined.

Model	Precision (%)	Recall (%)	Accuracy (%)	Macro-F1 (%)	AUC (%)
BERT	81.07	78.33	80.37	79.68	88.21
RoBERTa	82.11	79.61	81.90	80.84	89.43
COLDET	83.07	81.33	82.47	81.53	89.07
XLM-R large	78.21	80.00	81.87	79.09	88.43
BAIDUTC	60.66	25.89	63.74	36.29	51.74
omni-moderation-latest	63.71	81.49	75.11	71.51	85.13
MuDA [36]	69.94	72.45	70.28	71.17	80.91
DeepSeek-R1	67.72	82.14	74.41	74.24	86.03
DeepSeek-V3	68.34	**83.33**	76.12	75.09	86.38
Qwen-Plus-Latest	71.29	80.11	76.34	75.44	87.83
Qwen-Turbo-Latest	66.37	79.79	72.88	72.46	85.02
Ours	**83.12**	82.01	**83.24**	**82.56**	**92.14**

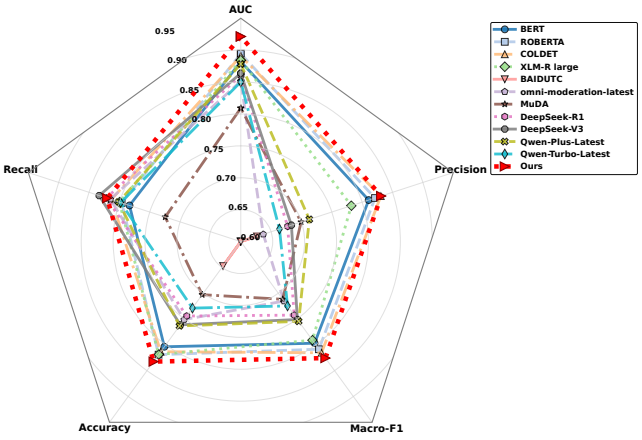

Figure 7: Performance comparison of different models on the augmented COLDataset.

Compared with transformer baselines such as BERT and RoBERTa, our dual-encoder design substantially improves representation learning by explicitly disentangling sentiment and toxicity signals. This enables SenTox-GLDA to better capture subtle abuse forms—sarcasm, implicit insults, and homophone-based toxicity—that are prevalent in Chinese social media. Even against strong multilingual models like XLM-R and the BERT-based COLDET, SenTox-GLDA demonstrates superior generalization on Chinese content, showcasing the benefits of its emotion–toxicity fusion, hierarchical differential attention, and expressive KAN classifier.

DeepSeek-R1 and DeepSeek-V3 show competitive recall, with DeepSeek-V3 achieving the highest recall (83.33%). However, both lag behind SenTox-GLDA in precision and F1. DeepSeek-R1 tends to over-rely on lexical patterns, sacrificing precision, while DeepSeek-V3 improves recall at the cost of false positives. In contrast, SenTox-GLDA's adaptive fusion balances emotional and toxic cues, yielding stronger precision–recall trade-offs.

OpenAI's omni-moderation-latest achieves moderate performance (F1 = 71.51%, AUC = 85.13%), but struggles with Chinese-specific phenomena such as slang, sarcasm, and emoji-based toxicity, reflecting a domain mismatch. BAIDUTC, reliant on handcrafted rules and lexicons, performs poorly (F1 = 36.29%, AUC = 51.74%), failing to capture implicit abuse.

MuDA underperforms (F1 = 71.17%), as auxiliary tasks (e.g., user profiling, topic prediction) dilute its focus on abuse detection, leading to unstable gradients and weaker generalization.

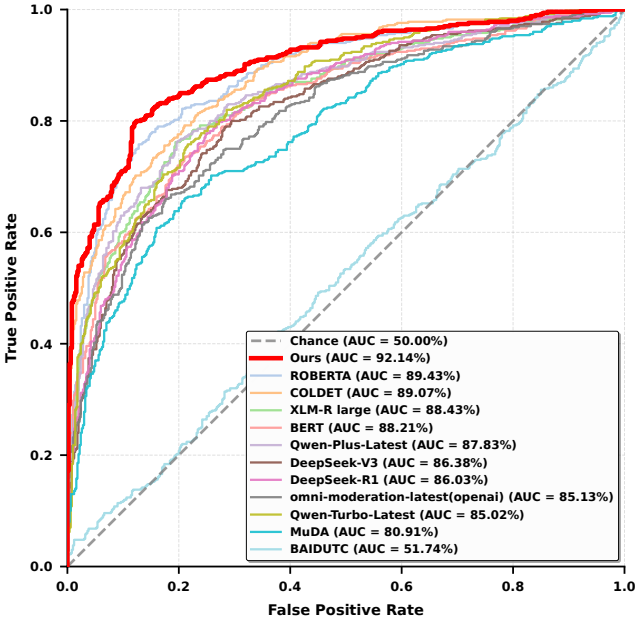

Figure 8: ROC curves of different models on the augmented COLDataset.

Overall, SenTox-GLDA achieves the best balance across precision, recall, and AUC, confirming the effectiveness of its dual-encoder, adaptive fusion, and KAN-based classifier. To further validate its design, we next conduct ablation studies to isolate the contribution of each component.

Table 4
Impact of dual-encoder architecture. Best results in **bold**.

BERT Encoder	RoBERTa Encoder	Precision (%)	Recall (%)	Accuracy (%)	Macro-F1 (%)	AUC (%)
✓	✓	**83.12**	**82.01**	**83.24**	**82.56**	**92.04**
×	✓	82.87	81.34	82.75	82.10	90.61
✓	×	70.65	76.81	73.44	73.60	82.45

Table 5
Impact of encoder type assignment. Best results in **bold**.

Sentiment Encoder	Toxicity Encoder	Precision (%)	Recall (%)	Accuracy (%)	Macro-F1 (%)	AUC (%)
BERT	**RoBERTa**	**83.12**	**82.01**	**83.24**	**82.56**	**92.04**
BERT	BERT	83.00	81.86	83.12	82.43	91.50
RoBERTa	RoBERTa	83.06	81.94	83.17	82.50	91.61
RoBERTa	BERT	83.08	81.99	83.21	82.53	91.72

4.6. Ablation Study

4.6.1. Ablation of Dual-Encoder Architecture

SenTox-GLDA relies on two complementary encoders: a BERT-based sentiment encoder and a RoBERTa-based toxicity encoder. Table 4 shows that removing either encoder substantially reduces performance. In particular, discarding the toxicity encoder lowers the Macro-F1 score from 82.56% to 73.60%, confirming that toxicity features are critical for detecting abusive language. Similarly, excluding the sentiment encoder also degrades performance, highlighting the importance of emotional cues in providing contextual grounding. The complete dual-encoder configuration consistently achieves the best results, demonstrating that both sentiment and toxicity signals are indispensable for robust abuse detection.

To further explore the effect of encoder types, we switch the roles of BERT and RoBERTa. As shown in Table 5, using RoBERTa for sentiment and BERT for toxicity yields performance close to the default configuration (F1 = 82.53%, AUC = 91.72%). Employing the same encoder type for both roles (BERT–BERT or RoBERTa–RoBERTa) also results in strong but slightly reduced performance. These findings suggest that SenTox-GLDA is robust to encoder permutations, provided that sentiment and toxicity signals are preserved. Nonetheless, the original configuration—BERT for sentiment and RoBERTa for toxicity—remains optimal.

The superior performance of the default setup can be attributed to two factors:

- RoBERTa's larger capacity and whole-word masking improve its ability to capture coarse-grained toxic expressions.
- BERT's finer-grained token masking better models subtle affective cues in sentiment-bearing Chinese texts.

These results validate the complementarity of pretraining objectives and dataset alignment in optimizing the dual-encoder design.

4.6.2. Ablation of Adaptive Fusion Module

The adaptive fusion module is responsible for dynamically integrating sentiment and toxicity features. As shown in Table 6, replacing this mechanism with static fusion strategies—such as weighted averaging or simple concatenation—results in a performance degradation of nearly 4% in Macro-F1. This indicates that fixed strategies fail to capture the complex, context-dependent interactions between sentiment and toxicity signals. In contrast, the adaptive module learns to modulate their relative contributions, enabling a more flexible and effective representation of abusive language.

Overall, these results confirm that dynamic feature integration is crucial for capturing the nuanced interplay between sentiment and toxicity. The adaptive fusion design thus plays a central role in enhancing the robustness and accuracy of SenTox-GLDA.

4.6.3. Ablation Of GLDA

Table 7 and Figure 9 present the ablation results comparing GLDA with six alternative attention mechanisms. GLDA consistently achieves the highest precision and Macro-F1 score, demonstrating its ability to balance class performance. In contrast, traditional multi-head attention [37] and sparse attention [10] exhibit very high recall, but their lower precision results in reduced F1 scores. Notably, sparse attention prioritizes toxic content, leading to excessive false positives and degraded overall precision.

Differential attention [45] offers a more balanced trade-off between precision and recall, reaching a Macro-F1 of 81.52%. Although slightly below GLDA, this highlights the value of explicitly modeling both local and global differentials. Other variants, such as tensor product attention [48], multi-head latent attention [25], and naive concatenation, underperform, underscoring their limited ability to capture multi-scale semantic shifts.

From an efficiency perspective, concatenation is the fastest and most memory-efficient, but its poor accuracy makes it unsuitable for high-stakes moderation. Tensor

Table 6
Effectiveness of the adaptive fusion module. Best results are highlighted in **bold**.

Feature Fusion Type	Precision (%)	Recall (%)	Accuracy (%)	Macro-F1 (%)	AUC (%)
Adaptive Module	**83.12**	**82.01**	**83.24**	**82.56**	**92.04**
Weighted Average	79.66	78.54	80.15	79.09	88.73
Concatenation	79.79	78.51	81.09	79.10	88.33

Table 7
Comparison of attention mechanisms. Time refers to average training time per epoch, Memory refers to training memory usage. Best results in **bold**.

Attention Type	Precision (%)	Recall (%)	Macro-F1 (%)	Time (s/epoch)	Memory (GB)
GLDA (Ours)	**83.12**	82.01	**82.56**	1654	4.612
Multi-Head Attention [37]	78.84	84.32	81.49	1730	4.788
Multi-Head Differential Attention [45]	81.71	81.34	81.52	1717	4.791
Sparse Attention [10]	69.44	**90.27**	78.50	1667	4.035
Tensor Product Attention [48]	80.13	80.12	80.12	1431	**2.985**
Multi-Head Latent Attention [25]	73.35	86.24	79.27	1796	4.978
Concatenation	76.02	74.81	75.14	**1398**	**2.275**

product attention provides slight efficiency gains, but at the cost of a 2.5% drop in F1. GLDA, while not the most efficient, achieves the best trade-off between resource usage and classification performance. In contrast, multi-head and latent attention incur higher computational overhead without meaningful performance improvements.

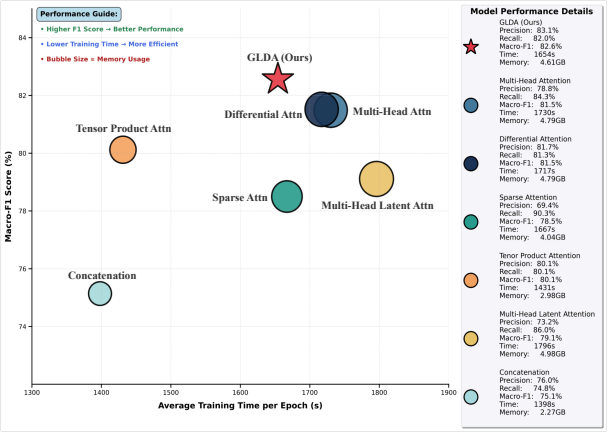

Figure 9: Performance comparison across attention mechanisms.

Dual-path contribution. Table 8 compares GLDA against its individual components, LDA and GDA. GLDA achieves the best balance with precision (83.12%), recall (82.01%), and Macro-F1 (82.56%). Removing GDA yields higher recall but much lower precision, suggesting that while LDA captures local variations effectively, it also amplifies noise in the augmented COLDataset. Conversely, removing LDA reduces Macro-F1, underscoring the importance of local sensitivity for precise detection. Although LDA and GDA individually require less time and memory, their performance lags behind GLDA, confirming the advantage of the dual-path design.

Effect of differential orders. To further assess the role of multi-order differentials, we tested several variants: GLA (no differentials), GLFA (only first-order), GLSA (no first-order), and GLTA (third-order). Results in Table 9 show that GLA performs worst, validating the necessity of differential operations. GLFA and GLSA improve over GLA but remain inferior to GLDA, confirming that both first- and second-order differentials are essential. GLTA, which adds third-order differentials, boosts recall but incurs heavy computational overhead, leading to diminishing returns in F1 relative to resource consumption.

In summary, GLDA's superior performance derives from its dual-path structure and balanced use of first- and second-order differentials, which together enable robust and efficient handling of noisy, nuanced abusive language.

4.6.4. Ablation Of KAN

Table 10 presents the ablation results of different Kolmogorov–Arnold Network (KAN) variants. The B-spline basis KAN achieves the best trade-off between performance and efficiency. Implemented with cubic splines (order = 3), 10 basis functions per activation, and a grid size of 5, it improves Macro-F1 by 0.74% over MLP while requiring less memory.

The Fourier Basis KAN achieves comparable accuracy but consumes 64% more memory, highlighting the superior efficiency of B-splines. Compared with Fibonacci and GRBF KANs [32, 8], the B-spline variant shows consistent advantages: both Fibonacci and GRBF yield lower recall and higher memory usage. GRBF offers slightly higher recall than Fibonacci, but its memory overhead makes it less practical.

These differences stem from the characteristics of the basis functions. B-splines, being piecewise polynomials, excel at modeling sharp and localized transitions—patterns

Table 8
Dual-path comparison of GLDA. Time refers to average training time per epoch, Memory refers to training memory usage. Best results in **bold**.

Attention Type	Precision (%)	Recall (%)	Macro-F1 (%)	Time (s/epoch)	Memory (GB)
GLDA (Ours)	**83.12**	82.01	**82.56**	1654	4.612
Only LDA	72.53	**87.71**	79.40	**1577**	**2.367**
Only GDA	79.63	81.44	80.52	1627	2.371

Table 9
Impact of multi-order differentials in GLDA. Time refers to average training time per epoch, Memory refers to training memory usage. Best results in **bold**.

Variant	Precision (%)	Recall (%)	Macro-F1 (%)	Time (s/epoch)	Memory (GB)
GLDA (Ours)	**83.12**	82.01	**82.56**	1654	4.612
GLA (No differential)	78.75	79.54	79.14	**1278**	**3.419**
GLFA (No 2nd-order)	80.44	80.17	80.30	1526	4.436
GLSA (No 1st-order)	81.17	80.56	80.86	1598	4.612
GLTA (With 3rd-order)	79.52	**85.41**	82.36	2676	9.747

that are common in abusive or toxic expressions. Fourier bases [42, 47], with global frequency components, are less effective for non-periodic linguistic structures and demand more resources. Fibonacci bases, while mathematically structured, struggle to capture the irregular boundaries of abusive language. GRBF functions capture non-linearities effectively, but their dense representation inflates memory usage.

In summary, the B-spline KAN not only achieves the highest Macro-F1 but also provides the best memory efficiency. Its ability to capture localized transitions aligns well with the irregular, context-sensitive nature of toxic expressions in Chinese social media, making it the most suitable classifier for this task.

4.7. Interpretability Analysis

To further elucidate how SenTox-GLDA makes classification decisions, we conduct an interpretability analysis. The Kolmogorov–Arnold Network (KAN) inherently provides stronger transparency than conventional MLPs, owing to its use of univariate B-spline basis functions. These basis functions localize the model's response to specific input ranges, allowing us to directly inspect activation patterns per input dimension. In contrast, dense MLP weights are highly entangled, obscuring interpretability.

Figure 10 shows t-SNE [35] projections of 3,000 test samples from the penultimate layer, comparing SenTox-GLDA with KAN to an MLP-based variant. The KAN representation produces three clearly distinguishable semantic regions:

- **Gender-related content** (blue) forms a compact cluster in the upper-right quadrant, with Offensive (dark blue) and Non-Offensive (light blue) samples moderately separated.
- **Race-related content** (green) appears in the lower-left quadrant, with greater dispersion and partial overlap between Offensive and Non-Offensive samples.
- **Region-related references** (red) occupy the upper-left quadrant, demonstrating the clearest separation between Offensive (dark red) and Non-Offensive (light red) instances.

By contrast, the MLP embedding yields more diffuse clusters, with substantial overlap between Offensive and Non-Offensive samples across all categories—particularly in race- and region-related instances. This highlights the weaker feature separation and reduced interpretability of MLPs compared to KAN.

Overall, these results confirm that SenTox-GLDA's architectural design contributes not only to improved predictive performance but also to enhanced transparency. The clear sub-cluster separations achieved with KAN provide an interpretable decision structure, offering a promising foundation for explainable moderation in real-world applications.

4.8. Analysis of Emoji Augmentation

Leveraging the flexibility of our emoji augmentation method, which supports both light and deep augmentation, we constructed five datasets by combining emoji-based strategies with the four augmentation techniques introduced in Section 3.2. Specifically:

- **Dataset V1**: Original COLDataset without any augmentation.
- **Dataset V2**: Augmented with contextual synonym replacement, structural perturbation, homophone replacement, and back-translation, but no emoji augmentation (149,920 comments).
- **Dataset V3**: Light emoji augmentation combined with the four strategies (299,840 comments).

Table 10
Comparison of KAN variants. Best results in **bold**.

Classifier Type	Precision (%)	Recall (%)	Macro-F1 (%)	Memory (GB)
B-Spline Basis KAN	**83.12**	82.01	**82.56**	4.612
Fourier Basis KAN	82.83	82.31	82.55	12.821
Fibonacci Basis KAN	81.71	81.27	81.49	6.368
GRBF Basis KAN	82.42	**82.64**	82.53	9.284
MLP	82.71	80.66	81.67	4.823

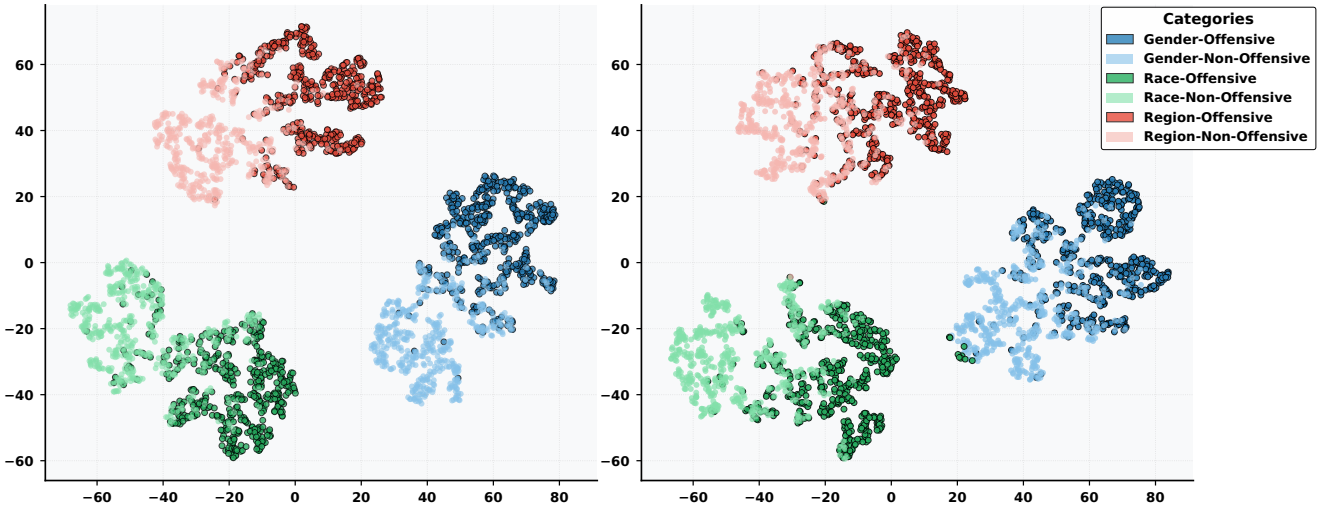

Figure 10: t-SNE[35] visualization of 3,000 test samples projected from the penultimate layer of SenTox-GLDA with KAN (left) and with a standard MLP classifier (right). The proposed KAN yields more compact and well-separated clusters across sentiment–toxicity categories, highlighting its stronger discriminative power and improved decision boundary clarity compared to the baseline.

- **Dataset V4**: Deep emoji augmentation combined with the four strategies (299,840 comments).
- **Dataset V5**: Mixed emoji augmentation (light + deep) with the four strategies (449,760 comments).

For all datasets, the positive-to-negative ratio is consistent with the original COLDataset. The augmented versions capture linguistic diversity characteristic of Chinese online discourse, including emojis, homophones, and sarcasm.

As shown in Table 11 and Figure 11, SenTox-GLDA consistently outperforms other models, particularly on emoji-rich datasets. While BERT and RoBERTa exhibit minor improvements with augmentation, their overall performance remains relatively stable, suggesting limited adaptability to the nuanced, emoji-driven expressions common in Chinese social media. By contrast, SenTox-GLDA achieves clear gains in both precision and recall, highlighting its ability to capture sentiment-emotion interactions masked by emojis.

The results also underscore the challenge of detecting toxicity when users creatively disguise intent with emojis or indirect expressions. Models such as COLDET and XLM-R show slight improvements with emoji augmentation but still fall short of SenTox-GLDA, reflecting their weaker capacity to model emotionally charged, context-sensitive language.

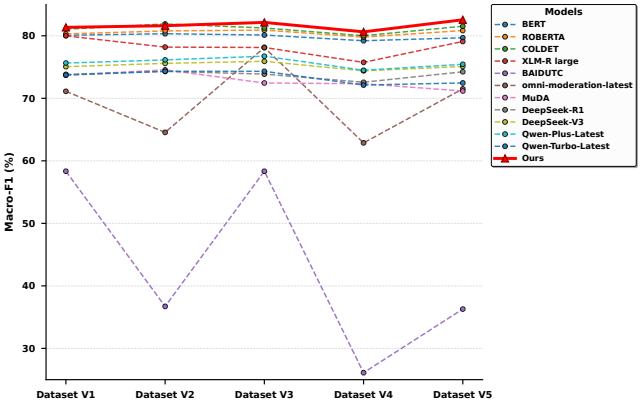

Figure 11: Macro-F1 performance trends across different augmentation strategies.

Rule-based or hybrid systems like BAIDUTC and omni-moderation-latest perform poorly under emoji augmentation, confirming their inability to generalize to dynamic, evolving abuse patterns.

In contrast, SenTox-GLDA's dual-encoder design and adaptive fusion mechanism enable it to maintain high precision without sacrificing recall, even in the presence of

Table 11
Performance comparison across datasets. The **best results** in each column are in bold, and the second-best are underlined. P = Precision, R = Recall, F1 = Macro-F1.

Model	Dataset V1			Dataset V2			Dataset V3			Dataset V4			Dataset V5		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
BERT	80.12	80.05	80.08	80.87	80.33	80.60	80.45	80.12	80.28	79.21	78.95	79.08	81.07	78.33	79.68
ROBERTA	80.33	80.27	80.30	81.02	80.78	80.90	80.91	80.45	80.68	79.87	79.12	79.49	82.11	79.61	80.84
COLDET	80.57	**81.53**	81.05	81.33	**81.89**	**81.61**	81.25	81.55	81.40	80.12	80.01	80.06	83.07	81.33	81.53
XLM-R large	79.87	80.12	79.99	80.45	78.19	79.30	78.13	77.59	77.86	76.08	75.49	75.78	78.21	80.00	79.09
BAIDUTC	62.99	27.54	38.31	61.87	26.12	36.72	58.34	22.93	32.91	56.87	26.11	35.79	60.66	25.89	36.29
omni-moderation-latest	65.12	78.33	71.12	64.55	77.89	70.62	65.01	78.12	70.95	62.87	75.45	68.57	63.71	81.49	71.51
MuDA	72.33	75.12	73.69	71.89	74.55	73.20	72.45	75.01	73.71	70.12	72.33	71.21	69.94	72.45	71.17
DeepSeek-R1	70.24	77.83	73.82	70.59	78.27	74.31	71.37	79.08	75.04	68.91	76.49	72.57	67.72	82.14	74.24
DeepSeek-V3	71.53	78.91	75.06	72.07	79.41	75.59	72.89	80.15	76.38	70.35	77.85	73.94	68.34	83.33	75.09
Qwen-Plus-Latest	73.15	78.37	75.64	73.58	78.85	76.14	74.35	79.47	76.82	72.03	77.14	74.49	71.29	80.11	75.44
Qwen-Turbo-Latest	70.57	77.14	73.72	71.05	77.87	74.32	71.89	78.47	75.07	69.14	75.35	72.12	66.37	79.79	72.46
Ours	**80.89**	81.15	**81.02**	**81.87**	81.35	**81.61**	**82.14**	**81.96**	**82.04**	**81.36**	**79.92**	**80.63**	**83.12**	**82.01**	**82.56**

heavily augmented, emoji-laden content. This advantage is most evident in Dataset V5, where combined light and deep emoji augmentation yields the best results: SenTox-GLDA achieves the highest Macro-F1 and demonstrates superior robustness across all evaluation metrics.

Taken together, these findings confirm that SenTox-GLDA not only achieves state-of-the-art performance but also demonstrates strong adaptability to diverse augmentation strategies, offering a reliable solution for toxicity detection in real-world, emoji-rich online environments.

5. Conclusion

In this paper, we introduced SenTox-GLDA, a dual-encoder framework that fuses sentiment and toxicity representations through an adaptive fusion module, models multi-scale textual patterns with GLDA, and employs a lightweight yet expressive Kolmogorov-Arnold Network (KAN) for classification. Comprehensive experiments on the Chinese COLDataset demonstrate that SenTox-GLDA establishes a new state-of-the-art, achieving an F1 score of 82.56% and an AUC of 92.04%, surpassing eleven competitive baselines. Ablation studies confirm the complementary value of sentiment cues, toxic semantics, hierarchical attention, and the KAN classifier, while interpretability analyses reveal that the model attends to intuitive linguistic and emoji-based indicators of abuse. These results highlight the potential of SenTox-GLDA for large-scale content moderation, enhancing detection accuracy and promoting safer online environments.

Our study emphasizes the importance of integrating emotional and semantic perspectives in abuse detection and provides a scalable, extensible framework for practical deployment. Future work will focus on real-world deployment trials on social media platforms, assessing latency, throughput, and user trust, as well as developing an ethics-guided feedback loop for handling contested cases.

Acknowledgments

This work was supported by the Talent Scientific Fund of Lanzhou University under Grant 561120212.

References

[1] M. M. Abdelsamie, S. S. Azab, and H. A. Hefny. The dialects gap: A multi-task learning approach for enhancing hate speech detection in arabic dialects. *Expert Systems with Applications*, 295:128584, 2026.

[2] S. Aggarwal and D. K. Vishwakarma. Exposing the achilles' heel of textual hate speech classifiers using indistinguishable adversarial examples. *Expert Systems with Applications*, 254:124278, 2024.

[3] N. AlDahoul, M. J. T. Tan, H. R. Kasireddy, and Y. Zaki. Advancing content moderation: Evaluating large language models for detecting sensitive content across text, images, and videos. *arXiv preprint arXiv:2411.17123*, 2024.

[4] H. Assoudi. A comparative benchmark of a moroccan darija toxicity detection model (typica. ai) and major llm-based moderation apis (openai, mistral, anthropic). *arXiv preprint arXiv:2505.04640*, 2025.

[5] Q. Bai, Q. Dan, Z. Mu, and M. Yang. A systematic review of emoji: Current research and future perspectives. *Frontiers in psychology*, 10:476737, 2019.

[6] A. Bleiweiss. Lstm neural networks for transfer learning in online moderation of abuse context. In *ICAART (2)*, pages 112–122, 2019.

[7] S. A. Castaño-Pulgarín, N. Suárez-Betancur, L. M. T. Vega, and H. M. H. López. Internet, social media and online hate speech. systematic review. *Aggression and violent behavior*, 58:101608, 2021.

[8] C. Chen, Z. Xu, Y. Liu, Q. Wu, T. Ji, H. Ji, J. Tang, Z. Sun, L. Fan, J. Liang, et al. Kolmogorov-arnold network for efficient equalization in short-reach im/dd systems. *Optics Express*, 33(16):33139–33152, 2025.

[9] N. Chetty and S. Alathur. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118, 2018.

[10] R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[11] L. Cui. Construction and implementation of knowledge enhancement pre-trained language model for text sentiment analysis. *Systems and Soft Computing*, page 200293, 2025.

[12] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021.

[13] F. M. P. del Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120, 2021.

[14] F. M. P. del Arco, M. D. Molina-González, L. A. Ureña-López, and M.-T. Martín-Valdivia. Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowledge-Based Systems*, 258:109965, 2022.

[15] J. Deng, J. Zhou, H. Sun, C. Zheng, F. Mi, H. Meng, and M. Huang. COLD: A benchmark for chinese offensive language detection. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[16] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

[17] M. Fernández-Gavilanes, E. Costa-Montenegro, S. García-Méndez, F. J. González-Castaño, and J. Juncal-Martínez. Evaluation of online emoji description resources for sentiment analysis purposes. *Expert Systems with Applications*, 184:115279, 2021.

[18] P. Kapil and A. Ekbal. A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, 210:106458, 2020.

[19] Z. Li, S. Cao, M. Zhai, N. Ding, Z. Zhang, and B. Hu. Multi-level semantic enhancement based on self-distillation bert for chinese named entity recognition. *Neurocomputing*, 586:127637, 2024.

[20] Z. Liu, P. Ma, Y. Wang, W. Matusik, and M. Tegmark. Kan 2.0: Kolmogorov-arnold networks meet science. *arXiv preprint arXiv:2408.10205*, 2024.

[21] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljacic, T. Y. Hou, and M. Tegmark. KAN: Kolmogorov–arnold networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

[22] J. Lu, B. Xu, X. Zhang, C. Min, L. Yang, and H. Lin. Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16235–16250, Toronto, Canada, July 2023. Association for Computational Linguistics.

[23] Q. Lu, X. Sun, Y. Long, X. Zhao, W. Zou, J. Feng, and X. Wang. Multimodal dual perception fusion framework for multimodal affective analysis. *Information Fusion*, 115:102747, 2025.

[24] E. Mahajan, H. Mahajan, and S. Kumar. Ensmulhatecyb: Multilingual hate speech and cyberbully detection in online social media. *Expert Systems with Applications*, 236:121228, 2024.

[25] F. Meng, P. Tang, X. Tang, Z. Yao, X. Sun, and M. Zhang. Transmla: Multi-head latent attention is all you need. *arXiv preprint arXiv:2502.07864*, 2025.

[26] M. A. Mersha, M. G. Yigezu, A. L. Tonja, H. Shakil, S. Iskander, O. Kolesnikova, and J. Kalita. Explainable ai: Xai-guided context-aware data augmentation. *Expert Systems with Applications*, 289:128364, 2025.

[27] Y. Mu, J. Yang, T. Li, S. Li, and W. Liang. Ha-gcen: Hyperedge-abundant graph convolutional enhanced network for hate speech detection. *Knowledge-Based Systems*, 300:112166, 2024.

[28] R. Qian, C. Ross, J. Fernandes, E. Smith, D. Kiela, and A. Williams. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*, 2022.

[29] E. B. Ramezani. Sentiment analysis applications using deep learning advancements in social networks: A systematic review. *Neurocomputing*, page 129862, 2025.

[30] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *Proceedings of the International AAAI Conference on Web and social media*, volume 5, pages 297–304, 2011.

[31] D. Sultan, M. Mendes, A. Kassenkhan, and O. Akylbekov. Hybrid cnn-lstm network for cyberbullying detection on social networks using textual contents. *International Journal of Advanced Computer Science and Applications*, 14(9), 2023.

[32] H.-T. Ta, D.-Q. Thai, A. B. S. Rahman, G. Sidorov, and A. Gelbukh. Fc-kan: Function combinations in kolmogorov-arnold networks. *arXiv preprint arXiv:2409.01763*, 2024.

[33] T. H. Teng, K. D. Varathan, and F. Crestani. A comprehensive review of cyberbullying-related content classification in online social media. *Expert Systems with Applications*, 244:122644, 2024.

[34] N. Tran, P. Ta, H. Nguyen, H. D. Nguyen, and A.-C. Le. Hybrid contextual and sentiment-based machine learning model for identifying depression risk in social media. *Expert Systems with Applications*, 291:128505, 2025.

[35] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[36] F. Vargas, F. Benevenuto, and T. Pardo. Toward discourse-aware models for multilingual fake news detection. In S. Djabri, D. Gimadi, T. Mihaylova, and I. Nikolova-Koleva, editors, *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 210–218, Online, Sept. 2021. INCOMA Ltd.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[38] E. D. Wahyuni, T. L. M. Suryanto, and H. Arviani. Deep learning multimodal sarcasm detection in social media comments: The role of memes and emojis. *Journal of Artificial Intelligence and Technology*, 5:192–201, 2025.

[39] B. Wang, S. Huang, B. Liang, G. Tu, M. Yang, and R. Xu. What do they "meme"? a metaphor-aware multi-modal multi-task framework for fine-grained meme understanding. *Knowledge-Based Systems*, 294:111778, 2024.

[40] Z. Wang, D. Huang, J. Cui, X. Zhang, S.-B. Ho, and E. Cambria. A review of chinese sentiment analysis: subjects, methods, and trends. *Artificial Intelligence Review*, 58(3):75, 2025.

[41] J. Wen, P. Ke, H. Sun, Z. Zhang, C. Li, J. Bai, and M. Huang. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338, 2023.

[42] J. Xu, Z. Chen, J. Li, S. Yang, W. Wang, X. Hu, and E. C.-H. Ngai. Fourierkan-gcf: Fourier kolmogorov-arnold network–an effective and efficient feature transformation for graph collaborative filtering. *arXiv preprint arXiv:2406.01034*, 2024.

[43] L. Xu, R. Mao, C. Zhang, Y. Wang, X. Zheng, X. Xue, and F. Xia. Deep transfer learning model for semantic address matching. *Applied Sciences*, 12(19):10110, 2022.

[44] Z. Xu. Roberta-wwm-ext fine-tuning for chinese text classification. *arXiv preprint arXiv:2103.00492*, 2021.

[45] T. Ye, L. Dong, Y. Xia, Y. Sun, Y. Zhu, G. Huang, and F. Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024.

[46] R. Yu, W. Yu, and X. Wang. Kan or mlp: A fairer comparison. *arXiv preprint arXiv:2407.16674*, 2024.

[47] J. Zhang, Y. Fan, K. Cai, and K. Wang. Kolmogorov-arnold fourier networks, 2025.

[48] Y. Zhang, Y. Liu, H. Yuan, Z. Qin, Y. Yuan, Q. Gu, and A. C. Yao. Tensor product attention is all you need. *arXiv preprint arXiv:2501.06425*, 2025.

[49] Y. Zhou, P. Xu, X. Wang, X. Lu, G. Gao, and W. Ai. Emojis decoded: Leveraging chatgpt for enhanced understanding in social media communications. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 2302–2316, 2025.

ORCID Information

Yutong Wang: 0009-0000-0722-630X

Zhongfeng Kang: 0000-0001-9025-0748

Jiaxue Yang: 0009-0003-3153-5323

Xiaopeng Fan: 0009-0000-5029-6768

Zijin Wu: 0009-0005-9903-2663

Shantian Yang: 0000-0003-2436-0580

Qinghua Zhao: 0000-0003-4906-7049

Zichen Song: 0000-0003-4155-2410

KAN-Boosted Chinese Online Abuse Detection Framework with Sentiment and Toxicity Fusion through Global-Local-Differential Attention

Highlights:

- Dual-encoder architecture for disentangled sentiment-toxicity representation learning.
- GLDA mechanism unifying local syntactic patterns and global semantic trends via differential attention.
- KAN-based classification, outperforming MLPs in modeling complex feature interactions.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:
