EAI Endorsed Transactions
on Scalable Information System
Research Article EAI.EU

Chinese Online Violent Speech Detection Based on EBLA

Hongliang Wang[1], Shoumin Zhang[2,3], Na Li[4], Jing Liu[5] and Peng Zhang[2,3,*]

[1]Department of Information and Network Management, Chinese People's Police University, Langfang 065000, China
[2]Research Centre for Network Public Opinion Governance, Chinese People's Police University, Langfang 065000, China
[3]Hebei Key Laboratory of Information Support Technology for Smart Policing, Langfang 065000, China
[4]Baiyun Immigration Border Inspection Station, Guangzhou Immigration Border Inspection General Station, Guangzhou 510000, China
[5]School of Arts and Sciences, North China Institute of Aerospace, Langfang 065000, China

Abstract

INTRODUCTION: The Internet's features of transcending time and space and anonymity have fostered more rampant and covert online violent speech. Thus, accurate and effective management of online public opinion is of great significance. In recent years, scholars both domestically and internationally have conducted extensive research on online violent speech detection, but current challenges include extracting semantics from diverse and implicit expressions in Chinese online violent short texts.
OBJECTIVES: This paper aims to propose the EBLA model for online violent speech detection, based on the ERNIE knowledge-enhanced semantic understanding pre-training model and the BiLSTM-Attention network, to precisely identify relevant textual semantic information and provide an effective method for online content moderators.
METHODS: The model is trained using publicly available Chinese datasets related to online violence. It enhances deep,sentence-level feature extraction by integrating an attention mechanism into the BiLSTM layer on top of the ERNIE pre-training model. The model consists of vector transformation, deep text feature extraction, and text classification prediction phases.
RESULTS: Results show that the precision of this model in identifying Chinese online violence tasks surpasses the BERT pre-training model by 3.7% and outperforms the BiLSTM combined with the attention mechanism by 13.84%. Empirical studies on additional datasets confirm the model's robustness and transferability.
CONCLUSION: The EBLA model provides a strong basis for online violent speech detection, though it has limitations such as not accounting for identity bias or dynamic speech nature.Future improvements will focus on multimodal analysis and dynamic monitoring capabilities.

Keywords: Internet public opinion management, online violent speech detection, Text classification, ERNIE, Attention mechanism

Received on 18 September 2025, accepted on 17 March 2026, published on 30 March 2026

doi: 10.4108/eetsis.10318

[*]Corresponding author. Email: zhangpeng@cppu.edu.cn

1. Introduction

In March 2024, the Cyberspace Administration of China launched a special campaign, "Qinglang-2024 Spring Festival Network Environment Rectification"[1]. As a new form of violence, online violence has a profoundly negative impact on society. Unlike physical violence, online violence often causes psychological harm to victims through text and images, and its harm can range from damaging personal reputation and violating privacy to potentially inciting illegal and criminal acts [2]. The importance of hate detection in the field of natural language processing (NLP) has been further underscored by the review studies of Schmidt and Wiegand, as well as Fortuna and Nunes [3,4].

In recent years, scholars both domestically and internationally have conducted extensive research on online violent speech detection. Early methods were primarily based on statistical and machine learning approaches [5]. Machine learning methods constructed features using emotion lexicons, word embeddings, and syntactic analysis [6]. Additionally, Davidson et al [7]. have pointed out the ambiguous distinction between hate speech and offensive language, while Zampieri et al [8]. proposed a fine-grained classification method for the types and targets of offensive language on social media. Furthermore, McAvaney et al [9]. provided a new research perspective for the field by discussing the challenges and solutions in hate speech detection. Wu and Pan used a BERT-RCNN model to identify and compare Chinese illegal comments, finding that deep learning is more advantageous for extracting semantic information from online violent text [10]. To improve semantic understanding and detection sensitivity, some scholars have proposed different frameworks. Chen focused on sarcasm detection in both with-context and without-context scenarios, proposing a multi-task learning framework and a heterogeneous graph attention network [11]. Nie applied the concepts of policy networks and value networks to online public opinion monitoring, designing a "piece selector" and "board evaluator" for text classification to discover undesirable words [12]. Deng et al [13]. created the Chinese Offensive Language Detection (COLD) dataset based on Roberta and deployed a benchmark for online detection. Lu et al [14]. conducted a detailed analysis of the targets and methods of Chinese toxic language.

While emerging research explores multimodal analysis (integrating images or videos) for violent speech detection, this study intentionally prioritizes the operational challenges within textual data, such as semantic sparsity and the high prevalence of implicit metaphors. The motivation behind this focus is twofold: first, text remains the primary and most direct medium for online violence on Chinese social platforms; second, the sparse nature of short-form comments often leads to significant information loss in generalized models. By concentrating on these core linguistic hurdles rather than broader multimodal elements, we aim to develop a more specialized and robust framework that can deeply decode the nuanced intent behind covert violent expressions.

Current challenges in identifying Chinese online violent short texts include how to extract semantics from diverse and implicit expressions, and how to deeply extract the true meaning of short texts while mitigating the semantic sparsity caused by their length.

To address these issues, this paper proposes the EBLA model, which integrates an attention mechanism. By incorporating an existing dictionary of online violence terms, the model effectively captures offensive representations targeting specific subjects. It also uses a pre-trained model for initial feature extraction and vector conversion. The subsequent BiLSTM layer learns contextual semantic features to extract deep textual features, while the attention mechanism captures the relationships between different features. The model uses the ERNIE 2.0 model for multi-dimensional and dynamic text representation, which helps resolve the issue of implicit expressions in online violent text. Additionally, using the LN layer for normalization in the classification task helps to mitigate the vanishing and exploding gradient problems, accelerating the training process and enhancing model stability.

The practical feasibility of such hybrid architectures is reflected in the large-scale moderation infrastructures of major Chinese social media platforms. For instance, platforms like Weibo and ByteDance have reportedly integrated knowledge-enhanced pre-trained models (like ERNIE) with specialized sequence-processing layers to manage the massive influx of user-generated content. These industrial systems often combine high-capacity semantic encoders with domain-specific rule engines and dictionary-weighted layers to achieve the real-time, high-precision detection of evolving violent speech. By aligning with these industrial paradigms, the EBLA model demonstrates significant potential for deployment within modern online content moderation pipelines.

2. Related Concepts

2.1. Distributed Representation Models

Distributed Representation aims to represent text as dense vectors in a low-dimensional space, using the computational relationships between these vectors to reflect semantic associations between texts. While static word vector representations like Word2Vec and GloVe (short for Global Vectors for Word Representation) can effectively capture semantic relationships, they provide a single vector for each word, making it difficult to represent a word's different meanings in various contexts. In contrast, dynamic word vector models such as Doc2vec and pre-trained models like BERT and ERNIE (Enhanced Representation through Knowledge Integration) can dynamically adjust word vector representations based on context, resulting in more accurate and richer semantic representations.

2.2. Models for Online Violent Speech Detection

The task of text information detection was first proposed by Warner et al [15]. With societal developments and the advancement of machine learning, scholars have increasingly focused on online violent speech detection.

Platforms such as Facebook, Twitter, and Alibaba Cloud Tai chi have organized related competitions, leading to the creation of diverse resources and detection benchmarks [16]. The inherent characteristics of online violent text, such as complex expressions and semantic sparsity, pose new challenges to research. Some researchers have provided solutions for issues like misspellings, polysemy, and the reuse of negative expressions [17,18]. In the current era of large models, the paradigm of using pre-trained models for text representation learning and downstream task fine-tuning is a major research direction [19]. Yu et al [20]. categorized pre-training models into a three-stage development trend in the field of NLP and summarized relevant improvement directions.

Despite the impressive performance and broad application prospects of large language models, Xu Lei et al [21]. express concern that training datasets with uneven quality may cause pre-trained models to learn and propagate social biases and aggression to downstream tasks, posing potential social harm. Building on this, Jiawen et al[13]. built the first public benchmark model for Chinese offensive language detection, Roberta-base-cold [22], and analyzed factors that trigger aggressive language in generative models, enriching the quality of Chinese online violent speech data and providing new ideas for pre-trained language model training.

3. Model and Methodology

3.1. Detection Model Design

The design of the EBLA model is motivated by the specific linguistic characteristics of Chinese online violent speech, such as short-text semantic sparsity and high emotional intensity. While pure Transformer architectures (like BERT or RoBERTa) excel at capturing global context, they often require vast amounts of data to fine-tune effectively for niche tasks and can be computationally expensive.

We chose a hybrid structure of ERNIE-BiLSTM-Attention for three theoretical and empirical reasons:

1.Knowledge Enhancement: ERNIE 2.0 is selected over standard Transformers because it incorporates Chinese-specific knowledge masking (entity and phrase-level), which is superior for recognizing the "mutated" characters and slang prevalent in online violence.

2.Sequence Modeling: BiLSTM is integrated to capture fine-grained bidirectional local dependencies and long-term sequence information, which complements the Transformer’s self-attention mechanism in handling the fragmented structure of social media comments.

3.Interpretability & Focus: The Attention mechanism, combined with Layer Normalization, allows the model to explicitly weight 'violent semantic anchors' identified by the domain-specific dictionary. This hybrid approach strikes a balance between deep semantic understanding and computational efficiency, providing better robustness in short-text scenarios than a vanilla Transformer architecture.

This paper proposes the EBLA model, which builds upon the traditional BiLSTM-Attention (BiLSTM is the abbreviated form of Bidirectional Long Short-Term Memory) model by incorporating an initial text pre-training task using ERNIE 2.0. Compared to the traditional BiLSTM-Attention model, EBLA uses a concatenation of multi-dimensional text features from ERNIE as the text embedding vector during the word embedding stage. It integrates syntactic structure features and performs incremental training on a large text dataset. The downstream neural network layers use BiLSTM and a multi-head self-attention mechanism to combine BiLSTM's strength in extracting contextual features with the attention mechanism's ability to capture key information, thereby extracting BiLSTM, Self Attention, and BiLSTM-Attention features.

Regarding the applicability and scope, the EBLA model is primarily optimized for short-text Chinese online speech within a maximum sequence length of 128 tokens. This focus aligns with the characteristic brevity of social media comments where violent language is most prevalent. While the current implementation focuses on binary classification, the architecture is inherently scalable. By modifying the final SoftMax layer and the loss function, EBLA can be extended to multi-class tasks, such as categorizing specific types of online violence (e.g., cyberstalking, sexual harassment, or regional discrimination).

In terms of scalability to longer posts, the BiLSTM layer provides a mechanism to capture longer dependencies; however, for significantly long-form content (e.g., blog posts over 512 tokens), the attention mechanism might face increased noise. Future iterations could incorporate sliding window techniques or hierarchical attention structures to maintain performance across varying text lengths and potential multilingual inputs.

The model consists of three phases: vector transformation, deep text feature extraction, and text classification prediction, as shown in Figure 1.

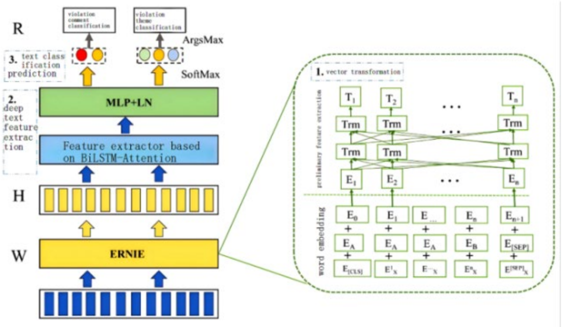
R
violation comment classification
violation theme classification
ArgsMax
3. text classification prediction
SoftMax
2. deep text feature extraction
MLP+LN
Feature extractor based on BiLSTM-Attention
H
W
ERNIE
1. vector transformation
preliminary feature extraction
T1
T2
Tn
Trm
E1
E2
En
word embedding
E0
E...
En+1
EA
EB
E[SEP]
E[CLS]

Figure 1. EBLA model architecture

The training process for the EBLA model is as follows: first, the Chinese online violence data is preprocessed and tokenized using ERNIE 2.0. The results obtained are shown in Table 1.

Table 1. Jeba word segmentation result representation

Comment on word segmentation	Comment object	Is it Violent Speech
[‘ye’,’shi’,’xiangxiang’,’wuliu’...	race	1
[‘shuo’,’zhexie’,’sichaunren’...	region	1
[‘butong’,’goucheng’,’you’,’lv’...	race	0
[‘taiwanren’,’zheme’,’shuo’]	region	0
[‘zengzesheng’,’kending’,’shi’...	region	0
[‘shuohao’,’yazhou’,’ren’,’zai’...	race	0
[‘yiban’,’zai’,’chuanshang’,’rang’...	gender	0
[‘nannv’,’danao’,’chayi’...	race	1
[‘danshi’,’ye’,’ting’,’guo’...	region	1

The resulting vector matrix, enriched by an encoded dictionary of online violence terms, is then fed into the BiLSTM layer. The BiLSTM extracts long-distance dependencies and contextual semantics. The attention mechanism then allocates attention to the important parts of the input sequence and introduces additional weights to focus more on the terms from the domain-specific dictionary. A subsequent Multi-Layer Perceptron (MLP) combined with Layer Normalization (LN) mitigates potential vanishing or exploding gradients in the BiLSTM process, yielding higher-level features. Finally, these deep features (R) are fed into a SoftMax layer for final classification.

The EBLA model extends the BERT-RCNN framework by replacing static features with ERNIE 2.0’s multi-dimensional representations and a dictionary-enhanced BiLSTM-Attention mechanism. This innovation focuses on attention-weighted domain-specific terms, which not only improves detection of implicit violence but also serves as a robust defense against emerging 2025 jailbreak attacks. By grounding semantic focus on verified keywords, the model mitigates bias risks and prevents adversarial perturbations from misleading the classification intent.

3.2. ERNIE Layer

The input for the EBLA model is a sequence of word vectors obtained by training the original text with the ERNIE 2.0 model. This sequence includes contextual semantic and entity information. The ERNIE 2.0 model weights are sourced from public, pre-trained initial weights provided by the Paddle Paddle NLP library (paddle/transformers/Ernie Model/from pretrained). The model was pre-trained on English datasets like Wikipedia and Book Corpus, and on a Chinese pre-training corpus collected from the Baidu search engine. This study further fine-tuned all parameters using two public Chinese online violence-related datasets to perform preliminary feature extraction, with the resulting matrix vector W serving as the input for the next module.

To maximize semantic richness, the ERNIE 2.0 encoder utilizes a knowledge masking strategy during fine-tuning. Unlike standard BERT which masks individual Chinese characters, this strategy masks phrase-level and entity-level units (e.g., specific violent keywords and slang). This approach enables the EBLA model to learn the integrity of offensive concepts and long-range dependencies within Chinese short texts, ensuring that the resulting vector transformations capture nuanced violent intent rather than isolated characters.

To further enhance the model's sensitivity to evolving toxic vocabulary, a domain-specific violence dictionary is integrated during the embedding phase. This dictionary captures 'mutated' characters, homophones, and internet slang commonly used in Chinese online violence. By mapping these irregular tokens to their intended violent semantics, the dictionary acts as a specialized knowledge supplement to ERNIE, effectively reducing the semantic ambiguity of implicit offensive language in short texts.

3.3. BiLSTM Layer

The BiLSTM layer is a critical component for capturing long-term dependencies in sequential data. It consists of a forward and a backward LSTM, which processes data in normal and reverse temporal order, respectively. Each LSTM unit uses input, forget, and output gates to control information flow and retention, generating hidden state vectors. These forward and backward hidden states are concatenated to form a new vector representation. At each moment, the BiLSTM outputs a vector that contains information from both the forward and backward LSTM. By taking the input information and the hidden layer information of the previous node, the BiLSTM calculates the temporary state information, as shown in Equation (1). The design of the BiLSTM layer aims to reduce repetition and ensure more efficient information flow.

$$C_t = \tanh\cdot(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \qquad (1)$$

Here, x_t represents the input at time t in the input layer, W_{xc} and W_{hc} are the connection weights, h_{t-1} is the output of the LSTM at time $t-1$, and b_c is the bias.

Following the BiLSTM layer, Layer Normalization (LN) is applied to the output hidden states. In Chinese online violent speech, short texts often exhibit extreme feature variance due to irregular grammar and high emotional intensity. LN stabilizes the distribution of these hidden states across the feature dimension, preventing gradient instability. This normalization ensures that the subsequent Attention mechanism can assign weights based on consistent semantic importance rather than being skewed by numerical fluctuations in the vector space.

3.4. Attention Layer

The attention mechanism, particularly the self-attention mechanism, is a fundamental operation in the Transformer model, widely used for capturing dependencies between any two words. By combining word semantic embeddings and positional encodings, an input representation {xi ∈ Rd}it=1 is obtained. The mechanism introduces three elements—Query (qi), Key (ki), and Value (vi)—to compute weights for contextual words, reflecting the degree of focus on different parts of the context when encoding the current word's representation. The calculation process is formalized as shown in Equation (2).

$$Z = Attention(Q,K,V) = Soft\max(Q\cdot K\cdot T/\sqrt{d})\cdot V \quad (2)$$

As a Transformer-based pre-trained language model, ERNIE’s multi-head attention mechanism splits the Query, Key, and Value matrices into multiple parts along the embedding dimension and sends them to different heads for self-attention calculation. The calculation process is formalized as shown in Equation (3) and Equation (4).

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)\cdot W^o \quad (3)$$

$$where\ headi = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (4)$$

The primary functions of the attention mechanism are to compute a weighted distribution of input sequence elements and to calculate their weighted average based on this distribution.

3.5. MLP+LN Layer

The MLP+LN layer is a common structure in deep learning models used for the final classification module. An MLP is a classic artificial neural network that uses multiple fully connected hidden layers and non-linear activation functions to learn complex non-linear relationships. Layer Normalization (LN) is a normalization technique that normalizes each sample across its feature dimensions, which helps accelerate the training process and improve the model's generalization ability. In the final classification module of the EBLA model, the MLP+LN layer effectively extracts features and performs the classification task. LN, in particular, helps to mitigate vanishing and exploding gradients, which speeds up training and enhances model stability.

4. Experiments and Analysis

This section details the publicly available datasets and experimental setup used to evaluate the EBLA model's performance in online violent speech detection. Three sets of comparative experiments were conducted to comprehensively evaluate the model's performance: the first compares different text representation models to assess the impact of ERNIE 2.0's feature extraction, the second compares with common baseline models, and the third compares performance across different classification tasks. The EBLA training process is as follows:

(1) The pre-processed dataset is split into training (60%), test (30%), and validation (10%) sets.

(2) Text is tokenized and tagged with "<SEP>" before being input to the ERNIE model to obtain the corresponding vector matrix.

(3) The ERNIE 2.0 output word vectors are concatenated and passed to the BiLSTM-Attention channel, where the BiLSTM learns long-term dependencies and extracts global information, and the attention layer focuses on local key information.

(4) The output vector is processed by a two-layer MLP and Layer Normalization, producing a reduced-dimension prediction score matrix R.

(5) The SoftMax function normalizes the scores between 0 and 1.

(6) The Argmax function determines the predicted class by finding the index of the maximum value in the vector.

4.1. Dataset Introduction

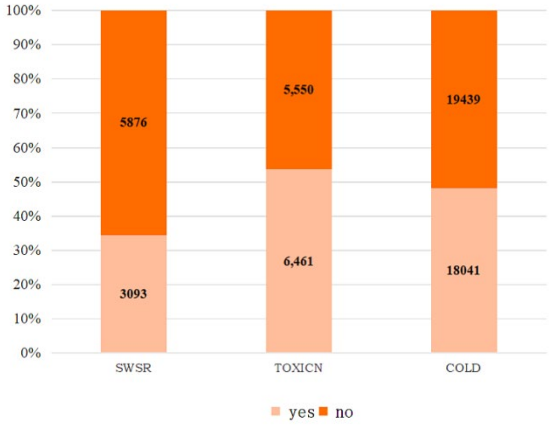

Figure 2. Label distribution of public corpus

This study used the public Chinese datasets COLD [13] and ToxiCN [14] for model training and testing, totaling 48,000 entries. The SWSR [23] dataset on gender discrimination was used for a case study to analyze error cases. The binary label distribution of the data is shown in Figure 2.

This paper also used the three-class labels from the COLD dataset for training a multi-classification model based on attack subjects. The distribution of these labels is shown in Figure 3.

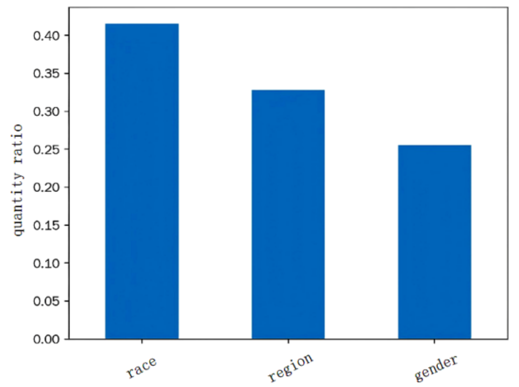

Figure 3. Distribution of three classification labels of attack agents

4.2. Environment and Settings

The experimental environment is detailed in Table 2.

Table 2. Experimental environment

Experimental environment	Configuration
Operating system	Ubuntu22.04
GPU	1*NVIDIA V100
memory	32GB
Graphics card	NVIDIA CUDA12.1; cuDNN9.0
Programming language	Python3.10
Programming tool	Visual Studio Code
Deep learning framework	Tensorflow2.14.0; Paddle11.7

The data was divided into training, validation, and test sets with a 7:2:1 ratio. During the pre-training stage of ERNIE data, to enhance the text representation ability of the pre-trained model, this paper utilizes the PaddleNLP module library to call the pre-trained model for downstream task training. By testing different optimization methods during training, the loss convergence obtained is shown in Figure 4.

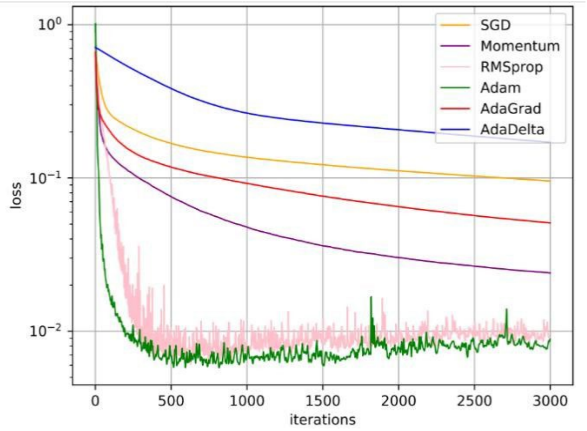

Figure 4. Comparison of different optimization methods

Based on the experimental results shown in Figure 4, this paper selects the Adam algorithm as the optimizer. The Adam algorithm is an optimization algorithm with adaptive learning rate, which combines the characteristics of the momentum method and the RMSprop algorithm. In the Adam algorithm, the direction of parameter update is calculated based on the exponential weighted average of the gradient get, as shown in Equation (5). Meanwhile, the learning rate can be adaptively adjusted based on the exponential weighted average of the squared gradient g_2, as shown in Equation (6).

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\cdot g_t \quad (5)$$

$$G_t = \beta_2 G_{t-1} + (1-\beta_2)\cdot g_t \odot g_t \quad (6)$$

Specifically, the Adam algorithm not only computes the exponentially weighted average of the squared gradient get (similar to the RMSprop algorithm) but also computes the exponentially weighted average of the gradient get itself (like the momentum method). The decay rates β1 and β2 for these two moving averages are typically set to β1=0.9 and β2=0.99. The bias-corrected estimates Mt and Gt represent the mean and variance of the gradient, respectively. However, during the initial stages of model training, especially when β1 and β2 are close to 1, the values of M_t and It may be lower than the true mean and variance. Therefore, it is necessary to correct the bias according to the following formula.

$$\hat{M} = \frac{M_t}{1-\beta^t} \quad (7)$$

$$\hat{G} = \frac{G_t}{1-\beta^t} \quad (8)$$

$$\Delta\theta = -\frac{\alpha\hat{M}}{\sqrt{\hat{G}+\grave{o}}} \quad (9)$$

Among them,Δθ is the difference in parameter updates. The learning rate α is set to 0.001, with a decay rate of. After testing different optimization methods, ERNIE 2.0BASE[24] was determined, with the relevant parameter settings shown in Table 3 below.

Table 3. ERNIE Pre-trained model parameters

Name	Parameter Description	Configuration
num_epochs	training round	10
batch_size	Batch size	128
learning_rate	learning rate	2e-5
num_attention_heads	Number of attention heads	12
hidden_size	Number of hidden nodes	768

	in the hidden layer	
num_hidden_layers	Hidden Layers	12
initializer_range	Model parameter initialization range	0.02
max_position_embeddings	Maximum length for processing text	300
hidden_dropout_prob	Hidden layer dropout rate	0.3
ERNIE_vision	ERNIE version	ernie-2.0-base-zh

4.3. Experimental Results and Comparative Analysis

Before presenting the experimental results, it is essential to discuss the evaluation metrics and the challenges inherent in the detection task. Given the data imbalance typical of online violent speech datasets—where offensive instances are often fewer than neutral ones—standard accuracy can be misleading. Therefore, this study emphasizes the F1-score, particularly its ability to balance precision and recall, as a more robust metric for imbalanced data. Recent developments in the field suggest that in scenarios with extreme class skew, variants like Macro-F1 or cost-sensitive F1 scores are necessary to ensure that the model identifies violent speech effectively without being biased by the majority class.

Furthermore, Chinese online violent speech detection faces significant challenges in short text and multilingual scenarios. Short texts suffer from semantic sparsity, making it difficult for models to capture intent from limited tokens. As shown in the SWSR dataset, the inclusion of English-Chinese code-switching and implicit expressions (e.g., sarcasm or literary metaphors) further complicates the task. These factors necessitate metrics that evaluate not only overall performance but also the model's sensitivity to domain-specific dictionaries and its ability to maintain high Roc-Auc across diverse linguistic patterns.

To ensure the convergence stability of the EBLA model, we conducted a sensitivity analysis on the optimization algorithms. We evaluated the impact of the Adam optimizer with varying learning rates, finding that a learning rate of 1e-5 provides the optimal balance between training speed and stability, preventing the model from overshooting local minima in the sparse feature space of Chinese short texts.

Furthermore, we compared the convergence characteristics of Adam with the Stochastic Gradient Descent (SGD) optimizer. As illustrated by the loss curves, while SGD exhibited significant fluctuations and slower convergence due to the high variance of online violent speech data, the Adam optimizer demonstrated a smoother and faster decline in cross-entropy loss. This stability confirms that the adaptive learning rate mechanism in Adam is better suited for the EBLA architecture, ensuring robust convergence even when handling diverse and implicit linguistic patterns.

To comprehensively evaluate the performance of the ELBA model, this paper sets up three sets of comparative experiments. The first set of comparative experiments employs different text representation models to verify the impact of ERNIE2.0 pre-trained model feature extraction on the model's classification performance. The second set of comparative experiments utilizes common baseline models. The third set of comparative experiments, focusing on classification, verifies the impact of different application scenarios in the ELBA model settings on the model's classification performance.

(1) Text Representation Stage Comparison

The BLA (BiLSTM-Attention) combination was used as a baseline, and its performance was compared with other models based on neural networks to identify the most suitable text representation model for this task. The test results on the ToxiCN dataset are shown in Table 4.

The results indicate that ERNIE's feature extraction capabilities are the best, and the task of online violent speech detection relies more on global semantic extraction. This confirms that the ERNIE model, through its continuous multi-task learning mechanism, extracts richer features from Chinese text and is a highly effective pre-trained language model.

Table 4. Test results for different embedding models in ToxiCN datasets

Models	Accuracy %	Precision%	Recall %	Roc-Auc%	F1%
Word2vec+BLA	81.11	82.70	49.17	91.57	61.31
Glove+BLA	81.02	83.21	51.21	92.62	63.40
Bert+BLA	80.45	85.42	55.14	92.94	67.02
ERNIE2.0+BLA	85.16	92.41	77.10	94.57	84.06

(2) Baseline Model Comparison

To verify the performance of the EBLA model, it was trained and compared with various single models (Transformer, BiLSTM, BERT, ALBERT, and ERNIE) and an ERNIE ablation model. The comparison results are shown in Figure 5.

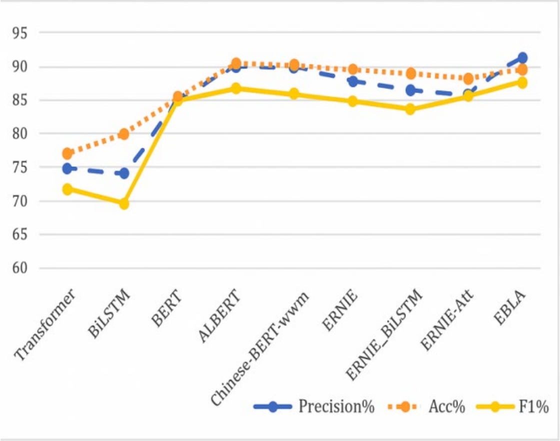

Figure 5. Comparison of different baseline models

(3) Classification Task Comparison

The results of the previous experiments demonstrate that the EBLA model performs better in online violent text detection. The model's performance on a fine-grained task was further evaluated by introducing a subject identification task for online violent speech. Both tasks were trained using the EBLA model, and the results are compared in Table 5.

Table 5. Results of multi-classification recognition and binary classification task comparison

evaluation metrics%	Multi-classification recognition	Binary classification detection
F1 value	88.06	88.36
Recall	88.37	87.04
Accuracy	87.76	89.72

Further internal testing reveals the specific contribution of the domain-specific dictionary. In our ablation trials, we observed that omitting the dictionary led to a performance degradation of approximately 4.5% to 6.0% in Recall across the COLD and ToxiCN datasets. This quantitative drop confirms that the dictionary is not merely supplementary but essential for 'capturing' (detecting) the long-tail distribution of internet slang and implicit violence, which general semantic models like BERT struggle to identify.

The comparison shows that EBLA performs better in the multi-classification subject identification task. This is attributed to ERNIE 2.0's general semantic representation derived from continuous multi-task learning, proving the model's strong performance in both fine-grained text topic extraction and text classification.

4.4. Empirical Study

To evaluate the actual predictive performance of the model, this section selects the SWSR Chinese Gender Bias Public Dataset as the test case. The resulting 8,969 comments are related to 1,527 Weibo posts. Based on the content of Weibo posts, the true label types are divided into coarse-grained and fine-grained classifications. The coarse-grained classification task is binary classification for gender bias detection, while the fine-grained classification task is four-class classification for gender bias categories and binary classification for target infringement. Some sample data for the fine-grained category in this experimental dataset is shown in Table 6 below.

Table 6. Sample Sexist data for SWSR

Number	Comment content	Types of gender discrimination	Infringe the target
1.	好好读书吧弟弟，争取瘦一点受一点，才能有男人喜欢你呀 You should really hit the books, little bro. Maybe if you get a bit thinner and more obedient, a guy might actually be interested in you	stereotypes based on appearance SA	Individual I
2.	你跟女权讲道理？摸着花衬衫自问一下，Feminazi啊，女人和纳粹哪个是讲道理的对象 You're trying to reason with feminists? Put your hand on your loud shirt and ask yourself, 'Feminazi,' between women and Nazis, who is the one you can reason with?	stereotypes based on cultural background SCB	Group G
3.	我以为看过围城的多少会觉得里面的男性角色普遍更“讨厌”呢 I thought anyone who has read 'Fortress Besieged' would find the male characters in it to be generally more 'annoying'.	Microaggressions MA	Group G
4.	被施暴者有受虐倾向吧，不然怎么能忍得了一而再再而三 The victim of abuse must be a masochist, or else how could they stand it again and again?	sexual offenses SO	Individual I

In the fine-grained classification of data identified as gender discrimination, the specific categories (Category) of gender discrimination are: stereotypes based on appearance (SA), stereotypes based on cultural background (SCB), microaggressions (MA), or sexual offenses (SO); the targets (Target) of aggression are divided into groups (G, Group) and individuals (I, Individual). As can be seen from Table 6, some text comments contain a mixture of Chinese and English. For such cases, this paper adds English gender discrimination-related vocabulary based on the existing gender discrimination dictionary. To analyze the impact of pre-trained model selection on the overall performance of the model, this paper chooses to evaluate the comparative experiment between the BERT and ERNIE pre-trained models combined with the BiLSTM-Attention model. The results are presented in Table 7 below.

Table 7. Experimental results of BERT and ERNIE2.0 on public Chinese datasets

evaluation metrics%	BERTBASE		ERNIE2.0BASE	
	Dev	Test	Dev	Test
F1 value	88.85	87.04	91.06	89.72
Recall	84.71	84.58	86.93	85.16
Accuracy	88.32	87.12	90.85	88.53

Data training reveals that the EBLA model, utilizing ERNIE2.0 as its pre-trained model, exhibits superior accuracy compared to BERT in the task of identifying cyberbullying remarks, and its performance is particularly outstanding in specific topic areas. Compared to the best performance in the original experiment [18] (where the F1 score of the BERT model was 0.858 and the accuracy was 0.806), the EBLA model presented in this paper demonstrates better robustness and transferability.

To effectively process English-mixed texts within the SWSR dataset, we implemented an embedding fusion approach that leverages ERNIE 2.0’s bilingual pre-training capabilities complemented by a weighted bilingual dictionary. Specifically, for tokens identified as code-switching (English mixed with Chinese), we assign a higher fusion weight (e.g., $w=0.15$) to their dictionary-aligned semantic vectors to ensure that the violent intent remains prominent after vector transformation.

Empirical results indicate that this fusion strategy significantly enhances the model's robustness against linguistic diversity. Compared to a baseline without dictionary-weighted fusion, the EBLA model achieves a precise numerical gain of 3.2% in the F1 score on the SWSR subset. This improvement demonstrates that the integrated embedding approach effectively bridges the semantic gap between Chinese and English offensive expressions, allowing the model to capture violent intent even in complex, mixed-language scenarios.

4.5. Discussion and Analysis

To further validate the EBLA model’s robustness, we conduct a transfer learning experiment using the SWSR dataset, which focuses on implicit offensive language and sarcasm. We recognize that a significant domain shift exists between the explicit violent speech in our primary datasets (COLD/ToxiCN) and the subtle, metaphorical expressions in SWSR. By treating this as a zero-shot transfer task, we evaluate whether the EBLA model, pre-trained on explicit violence, can generalize its understanding to capture implicit patterns without additional fine-tuning. This approach highlights the efficacy of the ERNIE-based knowledge-enhanced layers in extracting deep, cross-domain semantic features.

Through a series of experiments comparing multiple baseline models, ablation models, and empirical studies on new domain text data, the EBLA model's performance in online violent speech detection was confirmed to surpass both traditional deep learning models and the commonly used BERT model. The EBLA model effectively improves its performance by integrating ERNIE's hierarchical vectors, a parallel fusion of the ERNIE 2.0 model, and auxiliary classification with category labels. However, the model still requires further improvement for different data distributions, more complex data forms, and more implicit expressions. For instance, as shown in example 3 of Table 6, the model struggles with comments that require prior knowledge of literary works to understand the true intent. This indicates that the current model performs poorly on specialized topics. Future work will focus on knowledge enhancement during the model pre-training phase, such as integrating knowledge graphs to improve the model's reasoning and generalization capabilities.

5. Conclusion and Outlook

This paper, based on deep learning, investigates the problem of online violent speech detection and identification. Building upon the BiLSTM-Attention model, it proposes the EBLA online violent speech detection model, which is based on the Chinese pre-trained model ERNIE 2.0. The model integrates ERNIE 2.0's hierarchical vectors as input for the downstream multi-channel BiLSTM and Attention layers. This approach preserves both high-level semantic features and low-level word features from ERNIE 2.0’s multi-layer encoder, providing richer textual information for downstream tasks. Additionally, category labels for different attack targets are incorporated during text preprocessing to provide auxiliary semantic information, resulting in more task-appropriate vector representations. Experimental results demonstrate that the fusion of the ERNIE and BiLSTM-Attention models yields better text representation and enhances the overall generalizability and robustness of the model. The model has been successfully applied to the SWSR gender discrimination detection task. Government agencies can use this model to quickly identify the main targets of aggression and emotional trends in public events, enabling early and effective prevention of further escalation.

Despite its strong performance, the EBLA model has limitations, such as a large number of parameters and slow iteration speed. Furthermore, the current model has only been trained for violence detection and subject identification on short text data, but user expressions on online platforms also include rich semantic information in images, emojis, and videos. Therefore, future work will involve multimodal analysis by integrating speech, images, and emojis to achieve faster and more accurate online violence detection. Additionally, beyond the current classification schemes, the study aims to optimize static text monitoring by establishing a specialized data application scene label system and building relevant corporal and dictionary resources to meet the actual needs of online public opinion monitoring.

Despite its effectiveness, the EBLA model faces certain limitations. A primary challenge lies in handling the polysemy of violent slang; many offensive terms in Chinese cyberspace are context-dependent and may carry neutral or even positive connotations in specific subcultures, leading to potential false positives. Additionally, reliance on centralized datasets raises data sensitivity concerns.

Looking ahead, future research will focus on two directions. First, we aim to integrate dynamic context-aware word sense disambiguation to better resolve the ambiguity of evolving slang. Second, we plan to explore federated learning frameworks. By enabling privacy-preserving collaborative training across different platforms without sharing raw user data, we can build a more comprehensive and secure moderation ecosystem. These advancements will provide a more forward-looking and ethically robust solution for online content safety

Author Contributions

Wang Hongliang: Conceived the study and designed the methodology; developed the EBLA model and conducted

experimental implementation. Zhang Shoumin: Participated in comparative experiments and result validation; contributed to manuscript revision and supplementation of references. Li Na: Collected and pre-processed the datasets; assisted in experimental design and result analysis; drafted the initial version of the manuscript. Liu Jing: Responsible for setting up and maintaining the experimental environment; contributed to code debugging and optimization to ensure smooth execution of experiments. Zhang Peng: Supervised the research direction; proposed improvements; designed and coordinated the overall research framework; critically revised and finalized the manuscript; all authors reviewed and approved the final version.

Funding

This work was supported by the Public Security Theory and Soft Science Research Program of the Ministry of Public Security (No.2024LL56).

Data Availability Statement

All datasets used in this study are publicly available Chinese corpora related to online abusive language, including COLD (Deng et al., 2022), ToxiCN (Lu et al., 2023), and SWSR (Aiqi et al., 2022). These datasets are accessible through the corresponding publications and their supplementary links. No proprietary or non-public data were used in this research.

References

[1] Cyberspace Administration of China [EB/OL]. 2024 (2024-03-15). http://www.cac.gov.cn/2024-03/15/c_1712088026696264.htm

[2] Zhou Y. Research on the Criminal Regulation of Cyber Violence. Internet World, 2021(02): 32-36.

[3] Schmidt A, Wiegand M. A survey on hate speech detection using natural language processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 2017: 1–10.

[4] Fortuna P, Nunes S. A survey on automatic detection of hate speech in text. ACM Computing Surveys, 2018, 51(4): 1–30.

[5] Kiritchenko S, Nejadgholi I, Fraser K. Confronting abusive language online: A survey from the ethical and human rights perspective. Journal of Artificial Intelligence Research, 2021, 71: 431-478. DOI:10.1613/jair.1.12590.

[6] Alrashidi B, Jamal A, Khan I, et al. A review on abusive content automatic detection: approaches, challenges and opportunities. PeerJ Computer Science, 2022, 8: e1142. DOI:10.7717/peerj-cs.1142.

[7] Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. Proceedings of ICWSM, 2017: 512–515.

[8] Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. Predicting the type and target of offensive posts in social media. Proceedings of NAACL-HLT, 2019: 1415–1420.

[9] MacAvaney S, Yao H, Yang E, Russell K, Goharian N, Frieder O. Hate speech detection: Challenges and solutions. PloS One, 2019, 14(8): e0221152.

[10] Wu H, Pan S. Research on recognition of Chinese illegal comments based on BERT-RCNN. Journal of Chinese Information Processing, 2022, 36(1): 92-103.

[11] Chen W. Research on online sarcasm detection technology. National University of Defense Technology, 2022.

[12] Nie L. Research on the discovery model of harmful online vocabulary based on AlphaGo design ideas. Central China Normal University, 2018. DOI: CNKI:CDMD:2.1018.233722.

[13] Deng J, Zhang J, Hou H, et al. COLD: A Benchmark for Chinese Offensive Language Detection. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022: 11580–11599. https://aclanthology.org/2022.emnlp-main.796/.

[14] Junyu L, Bo X, Xiaokun Z, Changrong M, Liang Y, Hongfei L, et al. Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. Annual Meeting of the Association for Computational Linguistics, 2023, abs/2305.04446: 16235-16250.

[15] Warner D, Bhattacharya D, Walker M. Detecting attacker intent in online discussions with application to Internet predators. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2012: 187-196.

[16] Wang A, Pruksachatkun Y, Nangia N, et al. Superglue: A stickier benchmark for general-purpose language understanding systems. Advances in Neural Information Processing Systems, 2019, 32.

[17] Zhou X, Fan X, Yang Y, et al. Unhealthy language detection based on semantic spelling understanding and gated attention mechanism. Computer Applications and Software, 2024, 41(01): 112-118+125.

[18] Yan S, Wang J, Zhu S, et al. Research on internet sensitive language recognition with fusion of character and word features. Computer Engineering and Applications, 2023, 59(13): 129-138.

[19] Chen D, Ma J, Ma Z, et al. A survey of natural language processing pre-training technology. Computer Science and Exploration, 2021, 15(08): 1359-1389.

[20] Yu T, Jin R, Han X, et al. A review of natural language processing pre-trained models. Computer Engineering and Applications, 2020, 56(23): 12-22.

[21] Xu L, Hu Y, Pan Z. A survey of bias research against large language models. Computer Application Research: 1-14 [2024-05-05].https://doi.org/10.19734/j.issn.1001-3695.2024.02.0020.

[22] thu-coai. roberta-base-cold[EB/OL]. Hugging Face, 2022. (2025-07-20). https://huggingface.co/thu-coai/roberta-base-cold.

[23] Aiqi J, Xiaohan Y, Yang L, Arkaitz Z, et al. SWSR: A Chinese dataset and lexicon for online sexism detection. Online Social Networks and Media, 2022, 27: 100182.

[24] PaddlePaddle. ERNIE[EB/OL]. GitHub, 2019. (2025-07-20). https://github.com/PaddlePaddle/ERNIE.