information

Article

Chinese Cyberbullying Detection Using XLNet and Deep Bi-LSTM Hybrid Model

Shifeng Chen, Jialin Wang and Ketai He *

School of Mechanical Engineering, University of Science and Technology Beijing, Beijing 100083, China; b20200288@xs.ustb.edu.cn (S.C.); m202110490@xs.ustb.edu.cn (J.W.)
* Correspondence: heketai@ustb.edu.cn

Abstract: The popularization of the internet and the widespread use of smartphones have led to a rapid growth in the number of social media users. While information technology has brought convenience to people, it has also given rise to cyberbullying, which has a serious negative impact. The identity of online users is hidden, and due to the lack of supervision and the imperfections of relevant laws and policies, cyberbullying occurs from time to time, bringing serious mental harm and psychological trauma to the victims. The pre-trained language model BERT (Bidirectional Encoder Representations from Transformers) has achieved good results in the field of natural language processing, which can be used for cyberbullying detection. In this research, we construct a variety of traditional machine learning, deep learning and Chinese pre-trained language models as a baseline, and propose a hybrid model based on a variant of BERT: XLNet, and deep Bi-LSTM for Chinese cyberbullying detection. In addition, real cyber bullying remarks are collected to expand the Chinese offensive language dataset COLDATASET. The performance of the proposed model outperforms all baseline models on this dataset, improving 4.29% compared to SVM—the best performing method in traditional machine learning, 1.49% compared to GRU—the best performing method in deep learning, and 1.13% compared to BERT.

Keywords: social media; cyberbullying detection; deep learning; language model

Citation: Chen, S.; Wang, J.; He, K. Chinese Cyberbullying Detection Using XLNet and Deep Bi-LSTM Hybrid Model. *Information* **2024**, *15*, 93. https://doi.org/10.3390/info15020093

Academic Editor: Sherali Zeadally

Received: 9 January 2024
Revised: 29 January 2024
Accepted: 5 February 2024
Published: 6 February 2024

1. Introduction

With the rise of Chinese social media platforms such as Weibo, TikTok and Little Red Book, users can post content on the platforms in a variety of forms such as short videos, texts, and photos. While the internet has brought convenience to socialization, it has also generated many problems such as hate speech, social distrust, identity impersonation and cyberstalking [1], exacerbating the occurrence of cyberbullying and creating a series of negative impacts.

Cyberbullying is a type of bullying behavior in the virtual public space on the internet, with characteristics that differ from traditional bullying [2]. So far, there is still no consensus on the definition of the term cyberbullying, and a variety of terms have been used in the related literature, such as "online-aggression", "internet harassment", "online bullying" and "electronic bullying" [3,4]. However, cyberbullying is consistently described as bullying behavior with a malicious attack initiated against another person through electronic means of communication [5]. Pieschl et al. [6] categorized five types of cyberbullying: inflammatory (online fighting using online messages with harassing and vulgar language), harassment (insulting and threatening the victim using vulgar language), defamation (spreading rumors about the victim to damage his/her reputation), impersonation (acting under the identity of the victim), exposure and deception (by publicly exposing the victim's private information), and ostracism (deliberate exclusion or isolation of the victim in group-related online activities).

Some people, who are psychologically immature and impulsive, often act as "keyboard warriors" and make vicious comments on social media platforms. Aggressive speech

involving attacks, insults and denigration often lead to the victim's mental breakdown, anxiety, loneliness, depression, and even suicidal behavior [7]. Due to the popularity of the internet and smartphones, cyberbullying happens all the time, leaving the victim feeling powerless [8]. Because of the anonymity of cyberspace, the victims often do not know the real identity of the bullies, which makes them become more fearful in real life and experience more serious psychological problems [9]. Therefore, the detection of cyberbullying becomes very important, and there is an urgent need for technological means to stop and intervene in the occurrence of cyberbullying.

A study of 63 documents from around the world showed that the incidence of cyberbullying in the surveyed populations ranged from 6.0% to 46.3%, and the rate of cyberbullying victimization ranged from 13.99% to 57.5%. The rates are still increasing year by year [10]. The seriousness of cyberbullying can be shown clearly in these data. Due to the hidden and anonymous nature of the internet, if the regulation is not in place, the tendency towards cyberbullying of some internet users will intensify. The abusers post malicious remarks containing multiple dimensions such as incitement, hatred, aggressiveness, sexism, harassment, defamation, through comments, posts, private messages and other forms, causing serious harm to the victims. Cyberbullying detection is important for timely intervention and prevention of bullying to establish a civilized social network environment.

In order to detect cyberbullying in time and reduce its harm, this research investigates the methods of cyberbullying detection on Chinese social platforms, and the main contributions are as follows:

1. A hybrid model for cyberbullying speech detection based on XLNet and deep Bi-LSTM is proposed. XLNet combines the advantages of autoregressive (AR) and autoencoding (AE) language models and overcomes their limitations, and after deep Bi-LSTM bidirectional coding, it improves the accuracy of Chinese cyberbullying detection.
2. The Chinese offensive language dataset (COLDATASET [11]) was relabeled and expanded. In total, 1.66k offensive remarks crawled from 10 real cyberbullying incidents that happened in recent years as well as one-star bad reviews from the Chinese community website Douban were added. While adding more features of cyberbullying language, the data is balanced as much as possible to avoid the problem of model bias caused by over- or under-sampling.
3. A variety of methods using traditional machine learning, deep learning and Chinese pre-trained models are used as baseline for experiments on the expanded dataset to detect whether textual speech involves cyberbullying. The detection performance between different methods is also compared.

2. Related Work

2.1. Detection of Cyberbullying

So far, most of the research on cyberbullying detection has been carried out based on textual datasets, which can be regarded as text classification, and natural language processing techniques, machine learning and deep learning are widely used. Early approaches to cyberbullying detection, mostly based on supervised machine learning, were carried out through the use of dirty word dictionary detection and manual feature extraction. The researchers used feature vector representation of TF-IDF, N-gram, bag-of-words model method feature vectors, and used algorithms such as Support Vector Machines, K-nearest neighbors, Decision Trees, Naive Bayes, Random Forests for text classification. However, such detection methods are less accurate and unable to understand the semantic information of the sentence. The detection scope is also limited. Yin et al. [12] carried out earlier research on network harassment detection, which constructed local features, emotional features and contextual features to train SVM classifiers, and used oversampling methods to overcome the imbalance of the dataset. The results show that TF-IDF with emotional contextual features outperforms the detection methods of N-gram and the dirty word dictionary, which provides a reference for the early identification of malicious network users and the detection of criminal activities.

Reynolds et al. [13] manually labeled the data from the Formspring dataset, used oversampling and increased the percentage of positive examples in the dataset to solve the problem of data imbalance, and elucidated the shortcomings of the bag-of-words model. They used machine learning algorithms such as C4.5 decision tree, JRIP, KNN, SMO for training, and the better C4.5 decision tree and JRIP achieved an accuracy of 78.5% on cyberbullying detection. Dinakar et al. [14] utilized a manually labeled YouTube comment dataset for classification using various machine learning algorithms such as Naive Bayes, JRIP, J48 and SMO. The results showed that binary classification for single specific labels used for cyberbullying detection outperforms the multi-category classifiers, JRIP achieved the best results and SMO is steadier. Sarna et al. [15] used Twitter data crawled through keywords in three categories of direct cyberbullying, indirect cyberbullying and non-cyberbullying for user behavior and trustworthiness analysis. Four machine learning algorithms, namely KNN, Decision Tree, Naive Bayes, and SVM, were used to observe the performance of cyberbullying detection. Islam et al. [16] used two feature vector representations, Bow and TF-IDF, to analyze the performance on four different machine algorithms, namely Decision Tree, Naive Bayes, Support Vector Machine, and Random Forest. The experimental results show that the feature vector representation TF-IDF achieves higher accuracy than bag-of-words Bow, as well as the excellent performance of the machine learning algorithm SVM for cyberbullying detection.

With the rise of deep learning, neural network structures have made it possible to extract deep semantic information about textual content, with higher prediction accuracies than traditional machine learning algorithms. More and more scholars are using deep learning algorithms as cyberbullying detection methods, including convolutional neural networks (CNN), recurrent neural networks (RNN), and their variants long short-term memory (LSTM), gated recurrent unit (GRU) to improve the performance of the model by incorporating bi-directionality (Bi), attention mechanisms, and network structures such as CapsNet.

Zhang et al. [17] introduced the attention mechanism into the Bi-RNN network structure to recognize cyberbullying texts in the first stage, and to detect the bullies who send offensive remarks and evaluate the severity of their attacks in the second stage. The experimental results achieved the highest score in comparison with SVM, logistic regression, and CNN. Dewani et al. [18] filled the gap of Roman Urdu in cyberbullying detection by applying data preprocessing methods such as deduplication of words and slang mapping for data cleaning, and constructed three deep learning models, namely CNN, RNN-LSTM, and RNN-BiLSTM for cyberbullying detection. The results showed the effectiveness of RNN-LSTM and RNN-BiLSTM. Eronen et al. [19] applied feature density (FD) to the field of cyberbullying detection, measured the complexity of the dataset after preprocessing the data using Pos, Tokenization, Lemmatization, NER, Stopword filtering, etc. and observed its effect on the efficiency of the classifier. Experiments are conducted on datasets in three different languages: English, Japanese and Polish. The results show that text preprocessing improves the performance of the classifier, and feature density is applied differently in different languages. The increase in feature density makes the machine learning algorithm performance decrease, but the result is opposite for CNN. Kumar et al. [20] proposed a hybrid model consisting of Bi-GRU, Attention, and CapsNet network structures, which achieved the best performance among the compared algorithms by achieving F1-scores of 94.03 and 93.89 on Formsping.me and Myspace datasets, respectively. Ablation experiments were also conducted to perform variants on Bi-GAC, replacing Bi-GRU with Bi-LSTM and CapsNet with CNN. The results showed that Bi-GAC still achieved better results than the other two variants. Yuvaraj et al. [21] designed an artificial neural network model ANN-DRL with reward and punishment decision making for cyberbullying speech classification. They used three methods of feature selection from data: information gain, chi-square X2, and Pearson correlation coefficient. The results achieved 80.69% accuracy and showed better performance in comparison with traditional machine learning algorithms.

The emergence of large pre-trained language models, represented by BERT, has set several NLP task records. Since then, pre-trained language models have turned into a mainstream method to solve natural language processing tasks, and they can also be used for cyberbullying detection. The pre-training of large unsupervised corpora enable the models to reach a higher level in understanding contextual semantics, which further improves the prediction accuracy.

Paul et al. [22] applied the BERT model to cyberbullying detection and achieved the best performance on Twitter, Wikipedia and FormSpring compared to traditional machine learning via SVM, LR and deep learning algorithms such as CNN, RNN+LSTM and Bi-LSTM. The study compressed the large pre-trained model BERT to a smaller model by knowledge distillation, which resulted in significant computational savings by greatly reducing the number of model parameters and speeding up the execution as compared to BERT. Tripathy et al. [23] achieved the best performance on the Twitter dataset by fine-tuning the ALBERT pre-trained model, beating the comparative experiments CNN+wordvec, CNN+GRU, BERT-base+CNN, GPT-2, and other cyberbullying detection methods. Zinovyeva et al. [24] investigated the possibility of using deep machine learning and natural language processing to achieve the detection of antisocial online behavior, covering such structures as bidirectional coding, attention mechanisms, and hierarchical text representation. They design comparison experiments between traditional machine learning and deep learning, comparison experiments among bidirectional coding, pooling and attention, and comparison experiments between different scales of BERT and knowledge distillation. The experimental results show the best performance of the BERT model in antisocial online behavior. It also proves that there is no unique algorithmic structure that can beat other models on all datasets.

2.2. Limitations of Existing Research

There is a large amount of literature which focuses on English language cyberbullying detection research, but there are fewer studies on Chinese language cyberbullying detection, which may be caused by the complexity of the Chinese language and the lack of Chinese language cyberbullying datasets. Jahan et al. [25] suggests the same conclusion. As the most spoken language in the world, the Chinese language has very limited research in the field of hate speech detection. He attributed one reason to the lack of Chinese datasets in the relevant competition workshops, while the multilingual tasks of the competitions supported the research efforts of researchers from other parts of the world. The API interfaces provided by social platforms such as Twitter and the establishment of the Kaggle competition platform have made it easier to construct and obtain English datasets. English cyberbullying datasets such as Formspring, Myspace, Twitter, and YouTube are currently more common and researchers mostly base their studies on them.

In the study of Chinese cyberbullying, Li [26] used Yisurvey to crawl more than 1.4 million comments on Sina Weibo, and after NLPIR segmentation, six types of Chinese offensive words involving abusing, sexual cursing, etc. were manually selected and a lexicon of cyberbullying words was constructed. Zhong et al. [27] crawled more than 43,000 posts and comments from five domains of cyberbullying incidents on Sina Weibo through Python crawler technology to analyze the linguistic features of cyberbullying incidents in five domains, including education, entertainment, society, finance, and sports. Zhang [28] selected 26 war-citing texts from WeChat subscription accounts that often published irritating articles to build a corpus of incitement behaviors in cyberbullying and identify cyberbullying vocabulary through statistical methods.

The above studies on Chinese cyberbullying detection have the following characteristics:

1. The research is mostly conducted from the perspective of bullying vocabulary. Some social media platforms have the function of keyword filtering and blocking so that bullying words cannot be displayed. However, more bullying behaviors use implicit remarks such as mockery, innuendo, rhetorical questions and denigration. Although they do not include direct bullying vocabularies, they may cause serious psychological

harm to the victims. Therefore, it is not enough to rely only on the judgment of keyword filtering for this kind of behavior, and it is necessary to dig deeper into the semantics for the judgment of bullying behavior.

2. There is still no standardized dataset for the detection of Chinese cyberbullying. For the study of cyberbullying, most of the scholars crawled from social media platforms to construct datasets, and unfortunately, none of the above studies have disclosed the datasets used.

Aiming at the above two problems, we use deep learning algorithms as well as pre-trained language models to understand the semantic information at a deeper level. This study is based on the publicly available Chinese offensive language dataset COLDATASET [11] which is expanded by adding offensive language from real cyberbullying incidents and Douban one-star bad reviews. It covers both explicit bullying and implicit bullying, through which a more comprehensive account of Chinese cyberbullying detection can be conducted.

3. Methodology

The technique of mapping words to real vectors is known as word embedding, which is a vector representation of the meaning of a word and can also be considered as a vector representation of the features of a word [29]. It is a necessary process to transform text into machine-recognizable input. Although one-hot coding implements a vector representation of words, it does not assign word weights and cannot accurately represent the similarity between different words. TF-IDF solves this problem by quantifying the importance of words with respect to other words in the document and the corpus. However, due to an oversized lexicon, both one-hot and TF-IDF suffer from the problem of excessive dimensionality and sparsity of the vector representation, and they are often used as text representations for traditional machine learning. The self-supervised pre-training model Word2vec solves the above problems and becomes more insightful for semantic understanding. Word2vec contains two models, skip-gram and CBOW, which predict the context through the center word and predict the center word through the context, respectively. It aims at maximum likelihood estimation and uses a large corpus to train a shallow neural network that projects words into a vector space, allowing the vectors to better express similar relationships between different words. It is often used as a text representation for deep learning models.

Pre-training of Deep Bidirectional Transformer (BERT) [30] drove the research boom in pre-training language models. Since then, large-scale pre-training models have been emerging and breaking records in the field of NLP. BERT adopts a transformer model structure which is completely different from that of CNN and RNN, and uses a large unsupervised corpus to pre-train the model through two tasks, masked language model (MLM) and next sentence prediction (NSP). It endows the model with the ability to learn the contextual information as well as the semantic information at the sentence level. And it can be applied to the downstream tasks by fine tuning, which allows the two-stage training strategy of pre-training + fine tuning to become the mainstream approach to solve natural language processing problems. The ELMo [31] and GPT [32] pre-training models that appeared before BERT have achieved good performance on natural language processing tasks, which shows the effectiveness of pre-training models in handling NLP tasks.

The language models in the field of NLP can be broadly classified into two categories, AR: Autoregressive Language Model and AE: Autoencoding Language Model [33]. AR language models, represented by ELMo and GPT, are trained by unidirectional encoding and cannot deeply model bi-directional contextual information. However, downstream language understanding tasks often require bi-directional contextual information, which limits the performance of AR language models in handling some non-generative NLP tasks. The AE language model, represented by BERT, adopts the MASK mechanism instead of partial tokens, and the model predicts the tokens that are masked based on the context. This empowers AE language models to utilize contextual information for bidirectional models, which compensates for the lack of unidirectional model in AR language models. Denoising Autoencoder language models like BERT introduce MASK tokens on the input side of the

pre-training phase, but do not use them in the fine-tuning phase of the downstream task. This discrepancy between the pre-training and fine-tuning phases imposes a certain loss on the performance of the model. Yang et al. [33] proposed a generalized autoregressive model, XLNet, which fully combines the advantages and overcomes the limitations of the AR and AE language models, and achieved better performance than BERT on 20 NLP tasks.

In deep learning algorithms, recurrent neural networks can handle sequential text data, but cannot handle long dependencies between words when the sequence is too long. LSTM and GRU overcomes the difficulty that RNN cannot handle long dependencies well and solves the problem of gradient vanishing and gradient explosion. Compared with GRU, LSTM has a more complex model structure as well as more parameters, and its memory cell can store more information. Bidirectionality gives the recurrent neural network the ability to simultaneously capture information from context [34], which seems to achieve good results in the field of cyberbullying detection [17,20]. For the construction of a cyberbullying detection model, this research makes attempts from two aspects:

First, can XLNet achieve better results than BERT in Chinese cyberbullying detection?

Second, can the hybrid model of XLNet and deep Bi-LSTM further improve the detection performance?

3.1. Proposed Method

Therefore, this research proposes a hybrid detection model based on XLNet and deep Bi-LSTM. The proposed model structure is shown in Figure 1. Algorithm 1 illustrates the detection model and training process. The detection model consists of three key components: representation learning, encoding and classification. The overall approach of pre-training + fine-tuning is adopted for training. The XLNet-Chinese version [35] is used as the embedding layer for representation learning to generate feature vectors for contextual information, which is obtained by pre-training the XLNet model with a large number of Chinese corpora, such as Chinese Wikipedia and Q&A. The token and location information of the detected speech are processed by tokenizer and input to the pre-trained XLNet model. After being encoded by the two-stream self-attention mechanism, the hidden state of the last layer is used as the input of the Bi-LSTM layer. The forward and backward direction sequences of the hidden state of the last layer of XLNet are input to the deeper Bi-LSTM for further encoding, and a dropout layer is added to alleviate overfitting. Next, the concatenate layer connects the LSTM hidden states of forward and backward direction sequences at time step t to obtain the bi-directional information combining the context of the input sequences. Finally, the judgment of cyberbullying speech on the internet is obtained after dimensionality reduction by the linear layer.

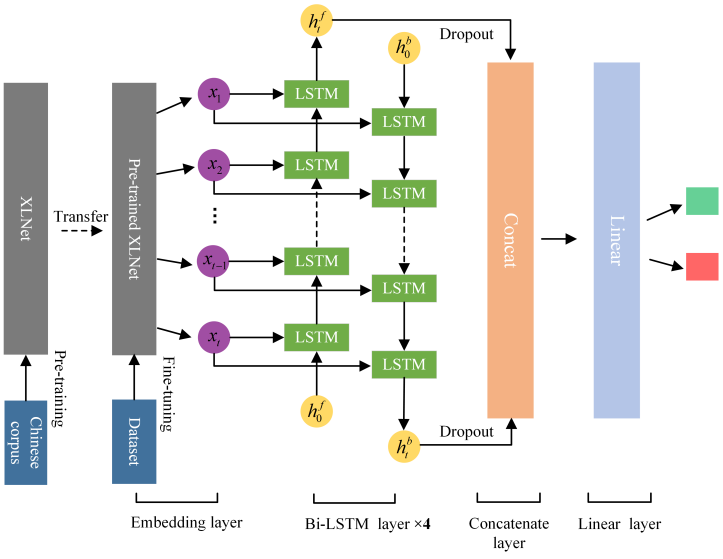

Figure 1. Proposed model.

Algorithm 1 Cyberbullying detection model and training process

Input: $\{(x_n, y_n)\}_{n=1}^{N}$, training set; θ, the initial parameters
Output: $\hat{\theta}$, the trained parameters
Parameters: nl, the number of lstm layers; cl, the number of classifications
Hyperparameters: N_{epochs}; lr, learning rate

for $i = 1, 2, \cdots, N_{epochs}$ **do**
  for $n = 1, 2, \cdots, N$ **do**
    $l \leftarrow \text{length}(x_n)$
    for $t = 1, 2, \cdots, l$ **do**
      $x_n[t] \leftarrow \text{token\_id} + \text{mask}$
    end for
    $h_n \leftarrow \text{Chinese\_xlnet}(x_n|\theta)$
    $h_n \leftarrow \text{Bilstm}(h_n|nl)$
    $h_n \leftarrow \text{Linear}(h_n|cl)$
    $Predict_n \leftarrow \text{Argmax}(h_n)$
    $\text{loss}(\theta) \leftarrow \text{Cross\_entropy}(Predict, y)$
    $\theta \leftarrow \theta - lr \cdot \text{loss}(\theta)$
  end for
end for
return $\hat{\theta} = \theta$

3.2. Embedding Layer

First, the tokenizer encodes the token and position embedding information of the detected speech and passes it as input representation to XLNet. XLNet uses the permutation language model objective as shown in Equation (1). Compared with BERT, this objective preserves the AE language model while giving the model the ability to capture bi-directional contextual information, realizing the combination of the advantages of AR and AE models. Instead of disrupting the order of the input sentences, permutation implements the order of factorization through an attention mask in the transformer architecture. Therefore, this method maintains the order of the original input sequence and avoids the discrepancy caused by the mask in the pre-training and fine-tuning stages.

$$\max_{\theta} \mathbb{E}_{z \sim \mathcal{Z}_T} [\sum_{t=1}^{T} \log p_\theta(x_{z_t} | \mathbf{x}_{z_{<t}})] \tag{1}$$

To address the lack of target position information in the transformer model, the position information of the current predicted token is introduced into the model, as shown in Equation (2), and two-stream self-attention is designed to solve the representation problem in the model by introducing two sets of hidden representations, content and query. Content hidden representation is shown in Figure 2a, which encodes the context and the content and position of the current prediction word. It is consistent with the standard transformer. The query hidden representation is shown in Figure 2b, which encodes only the context and positional information of the current prediction word, but not the content information of the current prediction word. The computation process of two-stream self-attention is shown in Figure 2c. Content hidden state is set to word embedding, while query hidden state is initialized to trainable variables, which is fed into self-attention and computed layer-by-layer. The last layer of the query hidden state is used to compute the Equation (2). For the fine-tuning stage of the downstream task, the query stream can be discarded and only the content stream is used.

$$p_\theta(X_{z_t} = x | \mathbf{x}_{z_{<t}}) = \frac{\exp(e(x)^T g_\theta(\mathbf{x}_{z_{<t}}, z_t))}{\sum_{x'} \exp(e(x')^T g_\theta(\mathbf{x}_{z_{<t}}, z_t))} \tag{2}$$

Partial prediction is used to address the problem of excessive computation due to alignments in the permutation language model. The hyperparameter K is introduced so

that the last 1/K tokens of the rearranged sequence are selected for prediction, thus saving speed and memory. In addition to this, XLNet incorporates the strengths of Transformer-XL [36], Segment Recurrence Mechanism and Relative Positional Encoding, which improve on the shortcomings of transformer's input of long sequences as well as the positional encoding problem, respectively. Regarding the training loss, XLNet only uses the loss of permutation language model but not the loss of next sentence prediction like BERT, which is more suitable for dealing with text classification problems such as cyberbullying detection. The details of the XLNet model are shown in Figure 2.

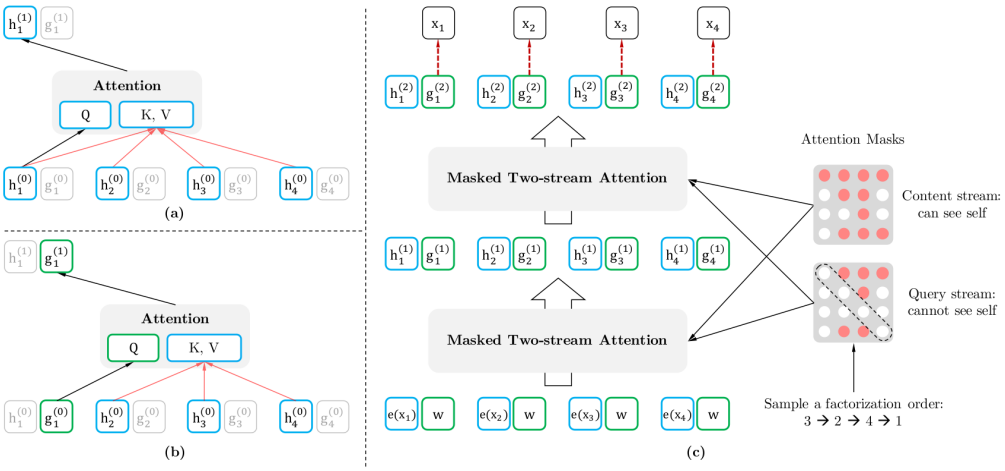

Figure 2. Two-stream self-attention and permutation language model training process [33].

3.3. Bi-LSTM Layer

Due to the shortcomings of the short-term memory of RNNs, it is difficult to transfer the information from earlier time steps to later time steps when dealing with longer sequences, which leads to the problem of losing important information at the beginning of the sequence. This is caused by the problem of gradient vanishing during backpropagation. As the length of the sequence increases, the successive products of matrices of the chain rule may make the gradient information of earlier time steps become smaller and smaller, so that learning becomes slow or even stops. The emergence of LSTM is a good solution to the problem of gradient vanishing and gradient explosion. As a variant of RNN, LSTM controls the degree of forgetting of historical information and retention of input information through a memory cell and three gate structures. The memory cell realizes the function of information transfer, which ensures that even information from earlier time steps can be transferred to subsequent time steps, thus allowing sequential information to be passed on all the time. The gate structures implement the function of adding or deleting information to the memory cell. The forget gate determines what information is retained or forgotten from the memory cell. The input gate determines what new information is added to the memory cell, and the output gate will determine the hidden state and the output value for the current time step based on the current state of the memory cell. The long short-term memory unit is illustrated in Figure 3. The formula for the LSTM cell is shown in the Equations (3)–(8) [29]. After encoding by the XLNet model, the 768-dimensional tensor is input into LSTM with a hidden layer dimension of 768.

$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i) \tag{3}$$

$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f) \tag{4}$$

$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o) \tag{5}$$

$$\widetilde{C}_t = \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c) \tag{6}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \widetilde{C}_t \tag{7}$$

$$H_t = O_t \odot \tanh(C_t) \tag{8}$$

I_t is the input gate, F_t is the update gate, O_t is the output gate, $\widetilde{C}_t$ is the candidate state at time t, C_t is the updated cell state at time t, H_t is the final hidden state at time t, X_t is the current input at time t, and H_{t-1} is the hidden state at time $t-1$. W_{xi}, W_{xf}, W_{xo}, $W_{xc} \in \mathbb{R}^{d \times h}$, W_{hi}, W_{hf}, W_{ho}, $W_{hc} \in \mathbb{R}^{h \times h}$, b_i, b_f, b_o, $b_c \in \mathbb{R}^{1 \times h}$. σ are the sigmoid activation function mapping values between 0 and 1. *tanh* is the tanh activation function mapping values between -1 and 1. $\odot$ is the Hadamard product.

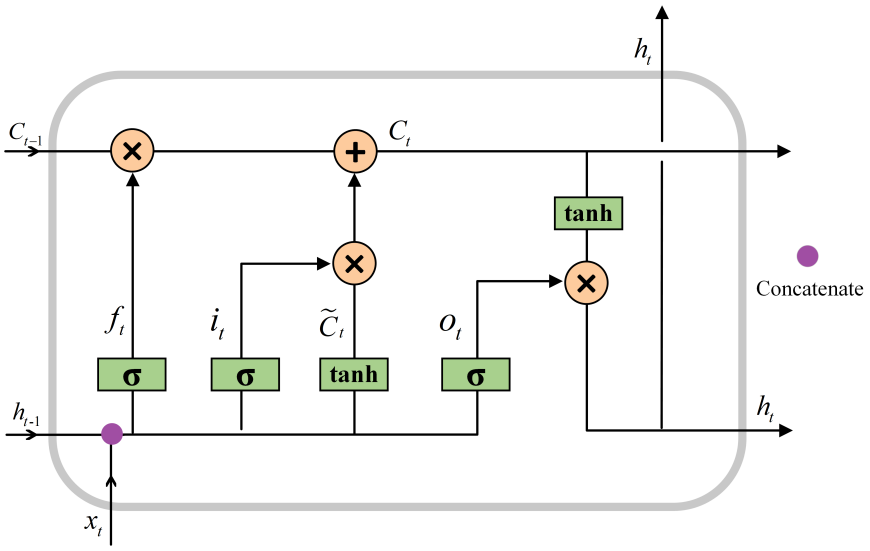

Figure 3. Long short-term memory (LSTM) cell architecture.

This work uses a four-layer bi-directional LSTM to encode the sequence from the forward and backward directions, respectively, and connects the hidden states of their last time step. The dropout value is set to 0.1, which makes all neurons stop working with a certain probability and play a regularization role. It can alleviate the occurrence of neural network overfitting to a certain extent. The connection layer connects the two 768-dimensional tensors to obtain a 1536-dimensional tensor, which gives the model the ability to capture contextual information simultaneously. It understands the context better than unidirectional LSTM when making predictions about the text. Therefore, it may achieve better performance. The formula is shown in the Equations (9)–(11):

$$\overrightarrow{h_t} = \text{LSTM}(x_t, h_{t-1}), t \in [1, T] \tag{9}$$

$$\overleftarrow{h_t} = \text{LSTM}(x_t, h_{t-1}), t \in [T, 1] \tag{10}$$

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \tag{11}$$

x_t is the current input at time t. h_{t-1} is the hidden state at time $t-1$. $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are the forward and backward hidden states at time t.

3.4. Output Layer

The last layer of the detection model is the fully connected layer. The input tensor is multiplied with the weight matrix and then added with the bias matrix to perform a linear transformation operation for dimensionality reduction of the input tensor. In this work, the 1536-dimensional tensor from the concatenate layer is reduced to two dimensions to obtain a binary classification, which ultimately categorizes the detection speech into cyberbullying and non-cyberbullying types.

4. Experiment

4.1. Dataset

The first open-source Chinese offensive language dataset COLDATASET [11] is used for cyberbullying detection. It contains three aspects of cyberbullying speech, namely gender antagonism, regional discrimination and racism and is derived from real data on two Chinese social platforms, Weibo and Zhihu. Since the dataset was constructed using a semi-automatic approach to annotation, there is still some noise in the samples. In this research, the dataset is merged, 155 duplicates are removed, and the dataset is re-labeled manually to improve the training quality. A total of four postgraduate students who regularly use Weibo and Zhihu social platforms were organized. Rules were set up: for each text, when the labeling result of all of them reached a consensus and was contrary to the result of the original dataset, the labeling of the original data was changed, otherwise it remained unchanged.

In addition, the dataset is expanded by crawling 1.66k offensive remarks from 10 real cyberbullying incidents that happened in recent years, as well as one-star movie reviews from the Chinese community website Douban. This ensures that the data is as balanced as possible as well as adding more cyberbullying language features. This helps avoid the problem of model bias due to data imbalance using oversampling or down-sampling, and helps the model to learn more implicit and explicit features of cyberbullying. The optimization of social platforms makes it difficult to collect the offensive speech of cyberbullying incidents. In this research, we collect data by searching media reports, victims' social platforms and screenshot retention of anti-violence people. Finally the data is merged, randomly disrupted and divided. The statistics of the divided dataset COLDATASET* are shown in Table 1. The training set contains 28,991 texts with an average character length of 46.7. The validation set contains 5000 texts with an average character length of 46.6. The test set contains 5000 texts with an average character length of 47.0. The cyberbullying incidents are shown in Table 2.

Table 1. Statistic of the dataset.

COLDATASET*	Cyberbullying	Non-Cyberbullying	Total	avg#char	min#char	max#char
Train	14,488	14,503	28,991	46.7	1	1217
Dev	2500	2500	5000	46.6	1	150
Test	2500	2500	5000	47.0	1	155
Total	19,488	19,503	38,991	46.7	1	1217

4.2. Experimental Settings

For the reconstructed COLDATASET*, traditional machine learning, deep learning and pre-trained language models are designed as baseline for comparison experiments. For traditional machine learning methods, the Chinese word segmentation tool JIEBA is used to segment Chinese characters, and then the TF-IDF method is used for text representation. Logistic, SVM, and Multinomial NB, which are more suitable for text categorization, and Random Forest algorithm with a decision tree of 100 are used as the classifiers for cyberbullying detection.

Table 2. Cyberbullying incidents.

Identifier	Cyberbullying Incident
Case 1	Niu Yu, a girl who survived the Wenchuan earthquake, was viciously abused
Case 2	Hangzhou Girl Zheng Linghua committed suicide due to cyberbullying over her pink hair
Case 3	Internet celebrity Guan Guan committed suicide due to cyberbullying
Case 4	100 Day Pledge Speech Girl gets cyberbullied
Case 5	Family-seeking boy Liu Xuezhou killed by cyberbullying
Case 6	Married mother Tang committed suicide due to cyberbullying
Case 7	Dr. An committed suicide due to cyberbullying
Case 8	Wuhan Sugar Water Grandpa who sold 2 Yuan sugar water suffered from cyberbullying
Case 9	A woman jumped from a building after suffering from cyberbullying because she gave the delivery boy 200 yuan to show thanks
Case 10	An oolong incident about a Tsinghua senior falsely accused a junior student of sexual harassment

For deep learning method, the pre-trained word2vec word embedding model is used to initialize the word vector representation of the split text. The Word2vec word embedding model is trained by using the Baidu encyclopedia corpus, and the weights of the word embeddings are learned through the skip-gram neural network structure. The dimensionality of the word embeddings is 300. The embedding layer is followed by using TextCNN [37], RNN and its variants including LSTM [38], GRU [39] and Bi-LSTM, Bi-GRU is used as a comparison experiment for cyberbullying detection.

For pre-trained language models, in addition to BERT and XLNet, this work also uses other Chinese pre-trained models as comparison experiments for cyberbullying detection to observe their performance, including a total of eight pre-trained language models, namely, RoBERTa [40], AlBERT [41], ERNIE [42], LERT [43], MacBERT [35], and ELECTRA [44] for. Among them, the Chinese versions of XLNet, RoBERTa and ELECTRA were trained by Cui et al. [35].

RoBERTa was released by Facebook AI, which made the following adjustments to BERT: (1) A larger batch size, more training data, and longer training time are used in model training. (2) NSP loss is removed. (3) Dynamic Masking is used, which means that each time a sequence is fed into the model, a new masking pattern is generated. These allowed RoBERTa to achieve state-of-the-art results on the GLUE, RACE and SQuAD tasks.

ALBERT is a lightweight BERT model released by Google that uses a smaller model. It reduces the overall number of parameters, speeds up training but achieves better performance. This research uses the Chinese version of ALBERT.

ERNIE was released by Baidu. Compared with BERT, ERNIE 1.0 improves two masking training strategies, one is a phrase-based masking strategy, and the other is an entity-based masking strategy. This research use ERNIE 3.0 [42] with a larger model and data size, which achieved state-of-the-art results on 54 Chinese NLP tasks.

LERT was released by HFL. Compared with BERT, it is a linguistically-informed augmented pre-training model that incorporating multiple linguistic knowledge. In addition to using the masked language model for pre-training, a linguistically informed pre-training (LIP) strategy which employs three linguistic tasks including POS, NER, and DEP for training is proposed. The effectiveness of LERT was demonstrated by conducting experiments on 10 Chinese natural language understanding tasks.

MacBERT was released by HFL and introduces an MLM as correction (Mac) language model pre-training task to alleviate the problem of discrepancy between the pre-training and downstream tasks. It uses similar words instead of mask tokens, and replaces them with random words when no similar words are available. Whole word masking and N-gram masking techniques are also introduced.

ELECTRA introduces a GAN-like learning task by replacing the MLM task with a REPLACED token-detection task. It replaces tokens instead of masks by sampling from a small generator network. The reconstructed sentences are sent to a discriminator network

to determine if they have been replaced by a generator. Compared with Roberta and XLNet models, ELECTRA achieves comparable results with less than 1/4 of the computation.

The parameters and hyperparameters are set as follows. The padding size is set to 64 because the average length of the dataset is 46.7. To improve the accuracy of the model, this work increases the length of sequences that can be processed by deep learning as much as possible within the machine's allowable limits. The batch size is set to 128 to fully utilize the GPU performance. For deep learning algorithms, the optimizer chooses Adam, with a learning rate = 1×10^{-3}, and the hidden layer neurons are set to 128 dimensions. For pre-training language model, the optimizer chooses AdamW, with a learning rate = 5×10^{-5}, and the base version is used. The dropout rate was set to 0.1 and the early stop mechanism is set to avoid overfitting. The training stops when the loss no longer decreases for 1000 iterations.

The experimental environment is as follows: CPU: 12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50 GHz. RAM: 43 GB. GPU: RTX 3090(24 GB) * 1.

4.3. Ablation Study

In addition, this research conducts ablation experiments to design a hybrid model of XLNet with TextCNN, LSTM, GRU, and Bi-GRU, respectively, to observe their performance in cyberbullying detection and the impact of increasing the number of recurrent neural network layers. The results are shown in Table 3.

Table 3. Weighted average F1-score of ablation results.

Method	Layers					
	1	**2**	**3**	**4**	**5**	**6**
+TextCNN	0.9020	-	-	-	-	-
+Bi-GRU	0.9012	0.9022	0.9016	0.8997	0.9012	0.9010
+LSTM	0.8992	0.9028	0.9018	0.9018	0.8983	0.9016
+GRU	0.9004	0.9022	0.9006	0.9036	0.9015	0.8994
Proposed	0.8986	0.9012	0.9017	**0.9043**	0.9011	0.8990

5. Results and Discussion

In this research, we use precision, recall and F1-score as evaluation metrics to assess the performance of cyberbullying detection methods in a Chinese experimental setting. The problem of model bias caused by oversampling or down-sampling methods due to data imbalance is avoided. To ensure the reliability of the results, each method is run multiple times. In view of the instability of some models, the average value of the evaluation metrics that were repeated five times was used as the final result.

Table 4 shows the results of traditional machine learning methods for Chinese cyberbullying detection. For the cyberbullying and non-cyberbullying categories, the SVM method achieves the highest F1-score values of 0.8623 and 0.8595, respectively, which exceeds other traditional machine learning algorithms. It can be seen that SVM can achieve better performance on text classification tasks.

Table 5 shows the results of deep learning methods for Chinese cyberbullying detection. The F1-scores of the deep learning methods used in this work outperform the traditional machine learning methods in Table 4 except for RNN. GRU achieves the highest F1-score in both cyberbullying and non-cyberbullying categories, and achieves the best performance among the deep learning methods. RNN performs poorly, most likely due to the problem of vanishing gradient, but GRU and LSTM solve this problem well. TextCNN also has a good performance, failing to surpass GRU but with very close results. It is worth noting that the addition of bi-directionality does not increase the performance of GRU and LSTM, but rather decreases their performance. This shows that bi-directionality does not improve the performance of the model in all cases.

Table 4. Experiment results for traditional machine learning.

Method	Non-Cyberbullying			Cyberbullying		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
NB	0.8972	0.7328	0.8067	0.7742	0.9160	0.8391
SVM	0.8517	0.8752	**0.8623**	0.8717	0.8476	**0.8595**
LR	0.8442	0.8712	0.8575	0.8669	0.8392	0.8528
RF	0.8545	0.8388	0.8466	0.8417	0.8572	0.8494

Table 5. Experiment results for deep learning.

Method	Non-Cyberbullying			Cyberbullying		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
TextCNN	0.9027	0.8720	0.8871	0.8762	0.9060	0.8909
RNN	0.5129	0.9392	0.6635	0.6398	0.1080	0.1848
GRU	0.9018	0.8740	**0.8877**	0.8778	0.9048	**0.8911**
LSTM	0.9037	0.8592	0.8809	0.8658	0.9084	0.8866
Bi-GRU	0.9052	0.8636	0.8839	0.8696	0.9096	0.8891
Bi-LSTM	0.8867	0.8736	0.8801	0.8754	0.8884	0.8819

Table 6 shows the results of the pre-trained language models for Chinese cyberbullying detection. The F1-scores of the pre-trained models used in this work outperform the deep learning methods in Table 5, except for ALBERT. To improve the model performance, these pre-trained language models are either tuned in parameters, or improved from mask strategy, network structure, or considering linguistic knowledge and using a larger corpus in the pre-training task. Although both ELECTRA and MacBERT take some approach to eliminate the discrepancy of pre-training and fine-tuning, neither of them outperform XLNet in weighted average F1-score in Chinese cyberbullying detection. The performance of the base version of the lightweight ALBERT model is suboptimal. To achieve improved detection results, it is recommended to utilize the large version. The weighted F1-score of all compared methods is shown in Figure 4.

Table 6. Experiment results for pre-trained models.

Method	Non-Cyberbullying			Cyberbullying		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
BERT	0.9015	0.8824	0.8919	0.8848	0.9036	0.8914
XLNet	0.9147	0.8792	0.8966	0.8837	0.9180	0.9005
RoBERTa	0.8955	0.8944	0.8949	0.8945	0.8956	0.8951
ALBERT	0.8685	0.8744	0.8714	0.8735	0.8676	0.8706
ERNIE3.0	0.9062	0.8732	0.8894	0.8777	0.9096	0.8933
LERT	0.9147	0.8668	0.8901	0.8734	0.9192	0.8957
MacBERT	0.8967	0.8992	0.8979	0.8989	0.8964	0.8977
ELECTRA	0.9079	0.8716	0.8894	0.8765	0.9116	0.8937
Proposed	0.9100	0.8974	**0.9037**	0.8988	0.9112	**0.9050**

Table 3 shows the results of the ablation studies. On one hand, TextCNN, LSTM, GRU and Bi-GRU are used to replace Bi-LSTM. On the other hand, the number of layers of the bidirectional recurrent neural network is deepened to observe the performance of the detection model. The results show that the proposed hybrid model achieves the best performance at a layer number of four for Bi-LSTM, outperforming all baselines. As shown in Figure 5, to some extent, the performance of the model can become better as the number of layers of the network is deepened, but it does not lead to a sustained improvement. The performance decreases when the model is too deep.

In summary, the F1-score of the proposed model reaches 90.43% on COLDATASET*. As shown in Figure 6, it improves 4.29% compared to SVM, which is the best performer in

traditional machine learning (TML) methods. It improves 1.49% compared to GRU, which is the best performer in deep learning (DL) methods. And there is a 1.13% improvement over BERT. The proposed model further improves on XLNet by 0.57% and outperforms all baseline models on this dataset with good performance.

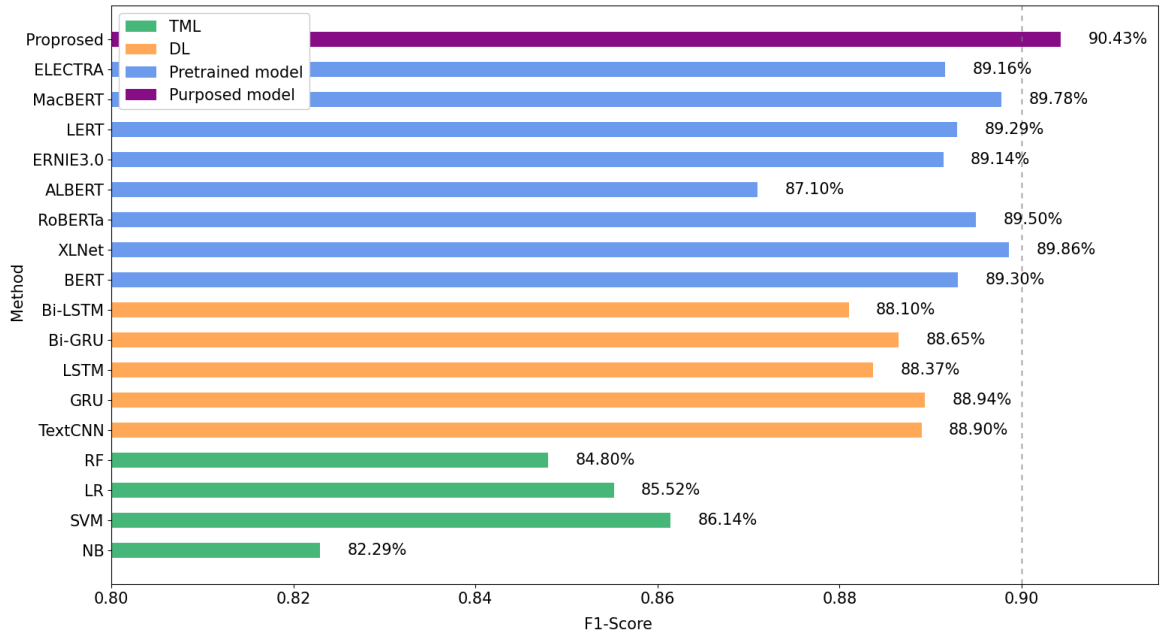

Figure 4. Comparison of weighted average F1-score for all methods.

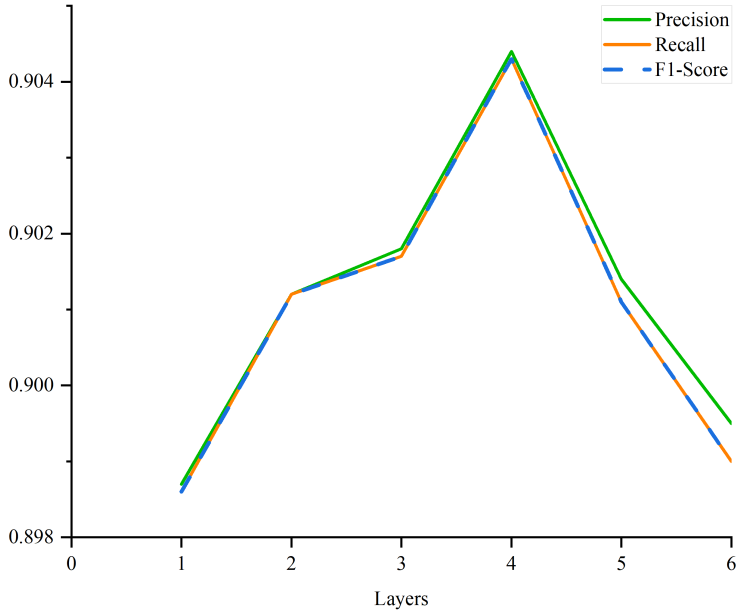

Figure 5. The effect of deepening the layers of Bi-LSTM on the results of the proposed model.

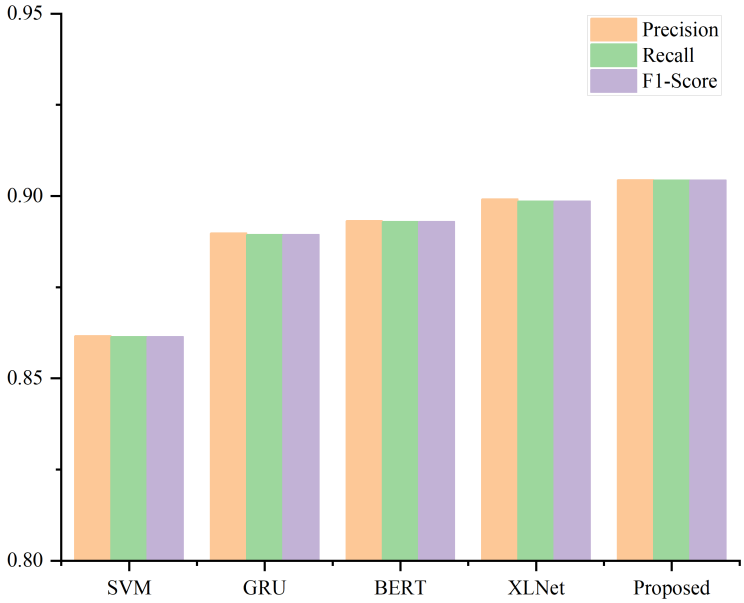

Figure 6. Comparison between the proposed model and four advanced methods.

6. Conclusions and Future Work

Cyberbullying has become a widespread social problem and may cause serious mental and psychological harm to individuals. However, the huge amount of data on social media platforms makes this behavior difficult to detect in a reasonable time, and it is far from enough to rely only on manual supervision. Advanced technology for automatic detection is required.

In this research, a hybrid model based on XLNet and deep Bi-LSTM is designed for Chinese cyberbullying detection. XLNet, which fully combines the advantages of AR and AE language models and overcomes their limitations, is used for representation learning. The accuracy of Chinese cyberbullying detection is further improved after deep Bi-LSTM bidirectional coding. The best performance is achieved in a baseline comparison with the use of traditional machine learning, deep learning and Chinese pre-trained language models. In addition, the original dataset is expanded to include bullying remarks from real cyberbullying incidents and bad Douban reviews involving personal abuse. It avoids the problem of model bias due to data imbalance using over- or under-sampling, and helps the model to learn more implicit and explicit features of cyberbullying.

Although the dataset has been supplemented, it is still insufficient for the detection of cyberbullying. In the future, more remarks related to cyberbullying incidents will be collected and accumulated, and more textual information such as harmonic stems, abbreviations and slang will be supplemented to dig deep into features of the offensive language. In addition to textual features, it is considered to try to incorporate user features into the training of the detection model to observe the prediction performance and identify the bullies. Due to constraints on machine performance, base versions of the Chinese pre-training language model are used. Large versions of the Chinese pre-trained model may achieve better performance. However, that would undoubtedly incur a greater computational cost. As suggested by Bender et al. [45], it is also important to consider the environmental and financial costs for the application of LLMs, as high costs can prevent some people from benefiting from the advances made by LLMs. In the future, energy and computational efficiency can be considered to be included in the model evaluation to achieve a balance between performance and efficiency as much as possible.

Although the application of pre-trained large language models (LLMs) has improved the performance of cyberbullying detection to a higher level, whether LLMs really have "understanding" and "consciousness" is still a question that deserves further thinking and research. Hamid [46] made an analogous experiment with ChatGPT and the Chinese Room

Argument, and put forward the view that ChatGPT and other LLMs lack true signs of consciousness. It is also argued that if LLMs have comprehension, it does so statistically rather than grammatically. LLMs are more like an abacus than a sentient mind. Hull [47] makes the same observation that AI is unconscious. Bender et al. [45] describe large language models in terms of stochastic parrots, viewing them as a system that combines information probabilistically, splicing together large amounts of data for training in the linguistic forms observed by the model without any reference to meaning. Hamid [48] points out that AI lacks the brain's innate mechanisms: fear behavior and the release of neurotransmitters, to the point that it is difficult to build natural abilities to deal with unusual situations in the face of surprises, and he suggests that this problem can be solved by combining AI with nature-inspired methods and human-machine coexistence through direct interaction between humans and algorithms.

There is still much debate regarding the impact of the rapid development of AI on human-machine coexistence, as the risks arising from the inclusion of biases in pre-training data, and the risks arising from applications that mimic humans, also require attention. More meaningful things will be explored in the future to benefit more marginalized groups. Of course, in addition to the detection of cyberbullying, the role of relevant laws and regulations are indispensable to avoid the occurrence of cyberbullying. Only by educating users about the law and enhancing their legal awareness and self-restraint can the problem be solved fundamentally.

Author Contributions: Conceptualization, S.C., J.W. and K.H.; methodology, J.W.; software, J.W.; validation, S.C.; resources, S.C.; writing—original draft preparation, J.W.; writing—review and editing, K.H.; funding acquisition, K.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Key Research and Development Program of China (No. 2022QY1403).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in reference number [11].

Acknowledgments: Thanks to the anonymous reviewers for their valuable comments on the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kumar, A.; Sachdeva, N. Cyberbullying detection on social multimedia using soft computing techniques: A meta-analysis. *Multimed. Tools Appl.* **2019**, *78*, 23973–24010. [CrossRef]
2. Smith, P.K.; Mahdavi, J.; Carvalho, M.; Fisher, S.; Russell, S.; Tippett, N. Cyberbullying: Its nature and impact in secondary school pupils. *J. Child Psychol. Psychiatry* **2008**, *49*, 376–385. [CrossRef] [PubMed]
3. Kwan, I.; Dickson, K.; Richardson, M.; MacDowall, W.; Burchett, H.; Stansfield, C.; Brunton, G.; Sutcliffe, K.; Thomas, J. Cyberbullying and children and young people's mental health: A systematic map of systematic reviews. *Cyberpsychol. Behav. Soc. Netw.* **2020**, *23*, 72–82. [CrossRef]
4. Smith, P.K.; Del Barrio, C.; Tokunaga, R.S. Definitions of bullying and cyberbullying: How useful are the terms. In *Principles of Cyberbullying Research: Definitions, Measures, and Methodology*; Routledge: London, UK, 2013; pp. 26–40.
5. Englander, E.; Donnerstein, E.; Kowalski, R.; Lin, C.A.; Parti, K. Defining cyberbullying. *Pediatrics* **2017**, *140* (Suppl. S2), S148–S151. [CrossRef] [PubMed]
6. Pieschl, S.; Porsch, T.; Kahl, T.; Klockenbusch, R. Relevant dimensions of cyberbullying—Results from two experimental studies. *J. Appl. Dev. Psychol.* **2013**, *34*, 241–252. [CrossRef]
7. Nixon, C.L. Current perspectives: The impact of cyberbullying on adolescent health. *Adolesc. Health Med. Ther.* **2014**, *5*, 143–158. [CrossRef]
8. Dooley, J.J.; Pyżalski, J.; Cross, D. Cyberbullying versus face-to-face bullying: A theoretical and conceptual review. *Z. Psychol. Psychol.* **2009**, *217*, 182–188. [CrossRef]
9. Slonje, R.; Smith, P.K.; Frisén, A. The nature of cyberbullying, and strategies for prevention. *Comput. Hum. Behav.* **2013**, *29*, 26–32. [CrossRef]

10. Zhu, C.; Huang, S.; Evans, R.; Zhang, W. Cyberbullying among adolescents and children: A comprehensive review of the global situation, risk factors, and preventive measures. *Front. Public Health* **2021**, *9*, 634909. [CrossRef]
11. Deng, J.; Zhou, J.; Sun, H.; Zheng, C.; Mi, F.; Meng, H.; Huang, M. Cold: A benchmark for chinese offensive language detection. *arXiv* **2022**, arXiv:2201.06025.
12. Yin, D.; Xue, Z.; Hong, L.; Davison, B.D.; Kontostathis, A.; Edwards, L. Detection of harassment on web 2.0. In Proceedings of the Content Analysis in the WEB, Madrid, Spain, 21 April 2009; Volume 2, pp. 1–7.
13. Reynolds, K.; Kontostathis, A.; Edwards, L. Using machine learning to detect cyberbullying. In Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops, Honolulu, HI, USA, 18–21 December 2011; Volume 2, pp. 241–244.
14. Dinakar, K.; Reichart, R.; Lieberman, H. Modeling the detection of textual cyberbullying. *Proc. Int. Aaai Conf. Web Soc. Media* **2011**, *5*, 11–17. [CrossRef]
15. Sarna, G.; Bhatia, M.P.S. Content based approach to find the credibility of user in social networks: An application of cyberbullying. *Int. J. Mach. Learn. Cybern.* **2017**, *8*, 677–689. [CrossRef]
16. Islam, M.M.; Uddin, M.A.; Islam, L.; Akter, A.; Sharmin, S.; Acharjee, U.K. Cyberbullying detection on social networks using machine learning approaches. In Proceedings of the 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 16–18 December 2020; pp. 1–6.
17. Zhang, A.; Li, B.; Wan, S.; Wang, K. Cyberbullying detection with birnn and attention mechanism. In *International Conference on Machine Learning and Intelligent Communications*; Springer International Publishing: Cham, Switzerland, 2019; pp. 623–635.
18. Dewani, A.; Memon, M.A.; Bhatti, S. Cyberbullying detection: Advanced preprocessing techniques & deep learning architecture for Roman Urdu data. *J. Big Data* **2021**, *8*, 160.
19. Eronen, J.; Ptaszynski, M.; Masui, F.; Smywiński-Pohl, A.; Leliwa, G.; Wroczynski, M. Improving classifier training efficiency for automatic cyberbullying detection with feature density. *Inf. Process. Manag.* **2021**, *58*, 102616. [CrossRef]
20. Kumar, A.; Sachdeva, N. A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media. *World Wide Web* **2022**, *25*, 1537–1550. [CrossRef]
21. Yuvaraj, N.; Srihari, K.; Dhiman, G.; Somasundaram, K.; Sharma, A.; Rajeskannan, S.M.G.S.M.A.; Soni, M.; Gaba, G.S.; AlZain, M.A.; Masud, M. Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking. *Math. Probl. Eng.* **2021**, *2021*, 6644652. [CrossRef]
22. Paul, S.; Saha, S. CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification. *Multimed. Syst.* **2022**, *28*, 1897–1904. [CrossRef]
23. Tripathy, J.K.; Chakkaravarthy, S.S.; Satapathy, S.C.; Sahoo, M.; Vaidehi, V. ALBERT-based fine-tuning model for cyberbullying analysis. *Multimed. Syst.* **2022**, *28*, 1941–1949. [CrossRef]
24. Zinovyeva, E.; Härdle, W.K.; Lessmann, S. Antisocial online behavior detection using deep learning. *Decis. Support Syst.* **2020**, *138*, 113362. [CrossRef]
25. Jahan, M.S.; Oussalah, M. A systematic review of Hate Speech automatic detection using Natural Language Processing. *Neurocomputing* **2023**, *546*, 126232. [CrossRef]
26. Li, W. A Content-Based Approach for Analysing Cyberbullying on Sina Weibo. In Proceedings of the 2nd International Conference on Information Management and Management Sciences, Chengdu, China, 23–25 August 2019; pp. 33–37.
27. Zhong, J.; Qiu, J.; Sun, M.; Jin, X.; Zhang, J.; Guo, Y.; Qiu, X.; Xu, Y.; Huang, J.; Zheng, Y. To be ethical and responsible digital citizens or not: A linguistic analysis of cyberbullying on social media. *Front. Psychol.* **2022**, *13*, 861823. [CrossRef]
28. Zhang, S. From flaming to incited crime: Recognising cyberbullying on Chinese wechat account. *Int. J. Semiot. Law-Rev. Int. SéMiotique Jurid.* **2021**, *34*, 1093–1116. [CrossRef]
29. Zhang, A.; Lipton, Z.C.; Li, M.; Smola, A.J. Dive into deep learning. *arXiv* **2021**, arXiv:2106.11342.
30. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
31. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 2227–2237; Association for Computational Linguistics.
32. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (accessed on 5 December 2023).
33. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*; 2019; Volume 32. Available online: https://papers.nips.cc/paper_files/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html (accessed on 5 December 2023).
34. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
35. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; Hu, G. Revisiting pre-trained models for Chinese natural language processing. *arXiv* **2020**, arXiv:2004.13922.
36. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv* **2019**, arXiv:1901.02860.

37. Zhang, Y.; Wallace, B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv* **2015**, arXiv:1510.03820.
38. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
39. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
40. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
41. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
42. Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv* **2021**, arXiv:2107.02137.
43. Cui, Y.; Che, W.; Wang, S.; Liu, T. Lert: A linguistically-motivated pre-trained language model. *arXiv* **2022**, arXiv:2211.05344.
44. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv* **2020**, arXiv:2003.10555.
45. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada, 3–10 March 2021; pp. 610–623.
46. Hamid, O.H. ChatGPT and the Chinese Room Argument: An Eloquent AI Conversationalist Lacking True Understanding and Consciousness. In Proceedings of the 2023 9th International Conference on Information Technology Trends (ITT), Dubai, United Arab Emirates, 24–25 May 2023; pp. 238–241.
47. Hull, G. Unlearning Descartes: Sentient AI is a Political Problem. *J. Soc. Comput.* **2023**, *4*, 193–204. [CrossRef]
48. Hamid, O.H. There Is More to AI than Meets the Eye: Aligning Man-made Algorithms with Nature-inspired Mechanisms. In Proceedings of the 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab, 5–8 December 2022; pp. 1–4.