基于RoBERTa的中文仇恨言论侦测方法研究

饶晓俊[1]，张仰森[1,2]，彭爽[3]，贾启龙[1]，刘雪阳[1]
[1]北京信息科技大学智能信息处理研究所/ 北京
[2]国家经济安全预警工程北京实验室/ 北京
[3]东北师范大学文学院/ 吉林长春
{raoxiaojun588,zhangyangsen}@163.com, shuangpeng@nenu.edu.cn,
allonlon@outlook.com, luxuryshxly@bistu.edu.cn

摘要

随着互联网的普及，社交媒体虽然提供了交流观点的平台，但因其虚拟性和匿名性也加剧了仇恨言论的传播，因此自动侦测仇恨言论对于维护社交媒体平台的文明发展至关重要。针对以上问题，构建了一个中文仇恨言论数据集CHSD，并提出了一种中文仇恨言论侦测模型RoBERTa-CHSD。该模型首先采用RoBERTa预训练语言模型对中文仇恨言论进行序列化处理，提取文本特征信息；再分别接入TextCNN模型和Bi-GRU模型，提取多层次局部语义特征和句子间全局依赖关系信息；将二者结果融合来提取文本中更深层次的仇恨言论特征，对中文仇恨言论进行分类，从而实现中文仇恨言论的侦测。实验结果表明，本模型在CHSD数据集上的F1值为89.12%，与当前最优主流模型RoBERTa-WWM相比提升了1.76%。

关键词： 中文仇恨言论；文本分类；RoBERTa；TextCNN；BiGRU

Chinese Hate Speech detection method Based on RoBERTa-WWM

Rao Xiaojun[1], Zhang Yangsen[1,2], Peng Shuang[3], Jia Qilong[1], Liu Xueyang[1]
[1]Institute of Intelligent Information, Beijing Information Science &Technology University / Beijing
[2]National Economic Security Early Warning Engineering Beijing Laboratory / Beijing
[3]College of Arts, Northeast Normal University / Changchun, Jilin
{raoxiaojun588,zhangyangsen}@163.com, shuangpeng@nenu.edu.cn,
allonlon@outlook.com, luxuryshxly@bistu.edu.cn

Abstract

With the popularity of the Internet, social media provides a platform for exchanging views, but intensifies the spread of hate speech due to its virtual and anonymous nature. Therefore, automatic detection of hate speech is crucial to maintain the civilized development of social media platforms. To solve the above problems, a Chinese Hate Speech Dataset–CHSD and a RoBERTa-CHSD model which is trained on the dataset are proposed. The RoBERTa pre-trained language model is used to serialize Chinese hate speech and extract the text feature information. Then, the TextCNN model and Bi-GRU model are respectively connected to extract multi-level local semantic features and dependency information between sentences. The two results are fused to extract deeper hate speech features in the text, and Chinese hate speech is classified, so as to realize the detection of hate speech. Experimental results show that the F1 value of the proposed model on CHSD corpus is 89.12%, which is 1.76 percentage points higher than that of the current best mainstream model RoBERTa-WWM model.

Keywords: Chinese Hate Speech , Text Classification , RoBERTa , TextCNN , BiGRU

第二十二届中国计算语言学大会论文集，第501页-第511页，哈尔滨，中国，2023年8月3日至5日。

1 引言

随着互联网的快速普及，社交媒体逐渐成为人们交流观点的最重要途径。但网络空间具有虚拟性和欺骗性，一些异常用户可以借助社交平台轻而易举地发布各种歧视言论和仇恨言论，因此社交媒体常常成为仇恨言论的爆发地。与此同时，海量数据的产生也带来了社交平台难以监管的问题，因此自动侦测仇恨言论日渐成为一个急需解决的问题。

根据现有的法律规定和通用共识，联合国 (2019)将仇恨言论定义为“因为个人或群体的身份（即他们的宗教、族裔、国籍、种族、肤色、血统、性别或其他身份因素）而攻击他们或对他们使用贬损或歧视性语言的任何言论、文字或行为交流。”。仇恨言论会对目标对象造成心理伤害，排斥、分裂不同的社会群体，严重的情况可能会引发社会暴动，从而对社会秩序造成伤害。因此，自动检测仇恨言论对于净化网络环境，维护社会和平具有重要意义。

为了解决仇恨言论的自动侦测问题，一个可靠的、通用的基准是加速该方向深入研究的必要基础。目前，常用的仇恨言论数据集有Wulczyn et al. (2017)提出的WTC，Zampieri et al. (2019)提出的OLID，Xu et al. (2020)提出的BAD等，但这些工作大多针对英文领域，中文领域由于缺乏完善的数据集和可靠的检测方法，中文仇恨言论侦测问题还有待进一步研究。

针对以上问题，提出了一个中文仇恨言论数据集CHSD（Chinese Hate Speech Dataset），包含17430条文本，主题覆盖种族、性别和地域。此外为了更深层次地提取仇恨言论特征，融合BiGRU提取全局特征和TextCNN提取多层次局部特征信息的特点，提出RoBERTa-CHSD模型来对中文仇恨言论进行侦测，实验结果表明RoBERTa-CHSD模型对于中文仇恨言论侦测的有效性。我们的数据和代码开源于https://github.com/RXJ588/CHSD。

2 相关工作

2.1 仇恨言论数据集

目前英文领域对于仇恨言论的研究非常丰富，研究范围涉及二分类到多标签分类再到多级分类任务。二分类任务方面，ElSherief et al. (2018)将从Twitter收集到的27,330条语料做是否为仇恨言论的二分类，提出英文仇恨言论数据集Peer to Peer Hate。多分类任务方面，Waseem (2016)将针对仇恨言论的类型做多分类标注，涉及类别有性别主义，种族主义和其他主义。多级分类的特点是多级注释，采用更细粒度的方案，Gomez et al. (2020)从种族、性别、性取向、宗教信仰四个方面做仇恨类型标注，在此基础上进一步标注了被攻击的对象群体。Nobata et al. (2016)区分了安全语言和辱骂性语言，又将辱骂性语言标记为仇恨言论、贬损或亵渎。Basile et al. (2014)采用三层二分类对仇恨性、攻击性和目标（个人/群体）进行标注。

目前国内相关数据集主要是面向侮辱性、性别对立、社会偏见等领域的，如表1所示。Tang et al. (2020)提出了一个分类冒犯语言的数据集COLA，主要用来对侮辱性语言，反社会语言和非法语言进行分类。Jiang et al. (2022)提出了第一个中文性别主义的数据集SWSR来识别性别相关的滥用语言。Zhou et al. (2022)提出了中文对话偏见数据集CDIAL-BIASDATASET，研究了对话中对目标群体的内隐态度。Deng et al. (2022)提出了中文冒犯语言数据集COLD，主要针对中文领域的冒犯性言论做了一个二分类任务。

Table 1: 中文领域仇恨言论相关数据集

数据集	年份	研究范围	大小
COLA	2020	侮辱性语言、反社会语言、非法语言	18k
SWSR	2022	性别相关的语言滥用	16k
CDIAL-BIASDATASET	2022	对话中的社会偏见	28k
COLD	2022	性别，种族和地域的冒犯语言	37k

总体来看，中文领域相关数据集要么话题覆盖比较单一，要么仅仅在研究文本的冒犯性或者文本的偏见表达，而仇恨言论是结合某项社会偏见的冒犯性表达，当前还缺少该类中文数据集，因此针对性别、种族和地域相关话题，提出了一个仇恨言论数据集CHSD。

基金项目：国家社会科学基金重大项目（21&ZD287）；

2.2 仇恨言论侦测

根据仇恨言论的定义，其侦测识别任务大致可以分为两类：判断文本是否为仇恨言论的侦测任务，以及对仇恨言论类型的判断的识别任务。本质上，这两个任务都属于文本分类任务。近年来文本分类技术的快速发展也为仇恨言论的侦测提供了有力的支撑，目前关于仇恨言论的研究方法主要有基于传统机器学习和基于深度学习两种方法。

传统机器学习是采用特征工程的方法获取文本特征后，再利用分类器来做仇恨言论的分类。常用的分类器有线性回归模型（LR）(Ousidhoum et al. (2019))、支持向量机（SVM）(Agarwal and Sureka (2015))、朴素贝叶斯（Naive Bayes）(Abozinadah et al. (2015))等。这种做法在特征的把握上忽略了序列关系，故不能充分利用文本上下文的信息，同时在复杂文本特征的提取上效果较差。

随着深度学习和预训练大模型的快速发展，越来越多的深层神经网络被运用到仇恨言论检测任务中。Badjatiya et al. (2017)提出使用卷积神经网络和长短期记忆网络与传统机器学习相结合的方式对Twitter仇恨言论进行侦测，结果表明深度学习模型的效果明显优于传统机器学习模型。Park and Fung (2017)提出使用混合模型进行言论滥用侦测的方法。考虑到Twitter推文的特殊性，需同时对字符级别和单词级别的特征进行学习，故结合了CharCNN与WordCNN得到混合CNN模型，用以提取不同级别的特征。实验结果显示，HybridCNN模型比单独的模型效果更加优秀。卢欣 (2019)针对中文微博数据，总结了7种语言特征，并将这种特征作为CNN网络的输入进行训练，特征输出与词向量经过CNN网络后的输出相结合，最终得到侦测结果。实验结果表明深度学习框架可以大幅度提升侦测精度，加入额外文本特征信息后的效果能得到进一步的提升。

相较于传统机器学习方法可能会出现的问题，深度学习方法则展现了更强的泛化能力，故其在仇恨言论侦测相关任务中已经成为研究主流。尽管通常深度学习的效果较好，但它仍然会受到数据的限制与词向量的影响，尤其在中文领域仇恨言论资源还很匮乏，具有进一步探索研究的意义。

3 RoBERTa-CHSD模型

针对中文仇恨言论侦测的问题，提出了RoBERTa-CHSD仇恨言论侦测模型。首先利用预训练模型RoBERTa-WWM学习文本语义特征，考虑到RoBERTa-WWM只是简单地做词嵌入的工作，难以充分考虑文本的内部语义特征信息，因此将得到的词嵌入特征再分别输入到TextCNN和BiGRU中，融合TextCNN得到的多层次局部语义特征和BiGRU得到的句子间的依赖关系信息，有效提取文本中更深层次的仇恨言论特征，从而提升仇恨言论侦测模型的性能。RoBERTa-CHSD模型结构如图1所示。

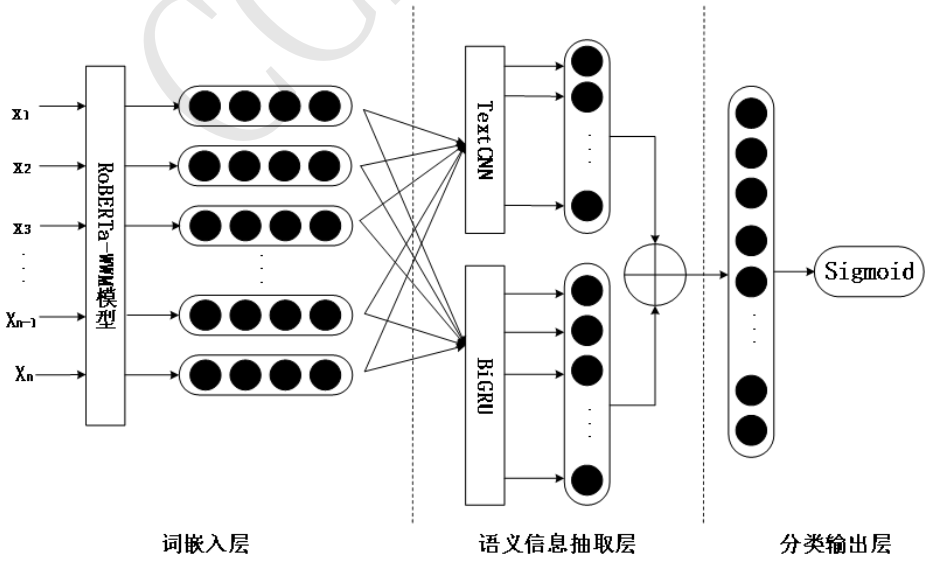

Figure 1: RoBERTa-CHSD模型结构

3.1 词嵌入层

词嵌入层是将输入的文本信息通过编码转换为相应的向量。采用RoBERTa-WWM预训练语言模型Liu et al. (2019)来对输入文本进行词嵌入表示。对于一条文本$S = [x_1, x_2, x_3, ..., x_n]$，构建文本S的词向量、句子嵌入和位置向量，将这3个向量的加和$E = [e_1, e_2, ..., e_n]$作为RoBERTa-WWM模型的输入，输入处理流程如图2所示。

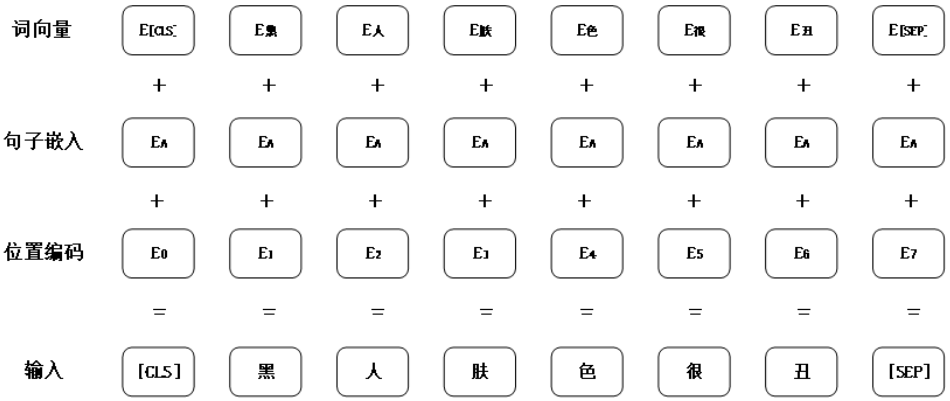

Figure 2: 输入处理流程

其中，词向量为one-hot编码后单词对应的向量表示；句子嵌入指用于分割多个句子的向量；位置向量为序列增加位置信息，保持顺序性表示。

再经过RoBERTa-WWM中的Transformer模块训练即可得到文本S的动态语义表示$V = [v_1, v_2, v_3, ..., v_n] \in R^{n \times d}$,其中$n$为输入长度，$d$为词向量维度768。RoBERTa-WWM模型结构如图3所示，中间层表示12层双向Transformer特征提取器。

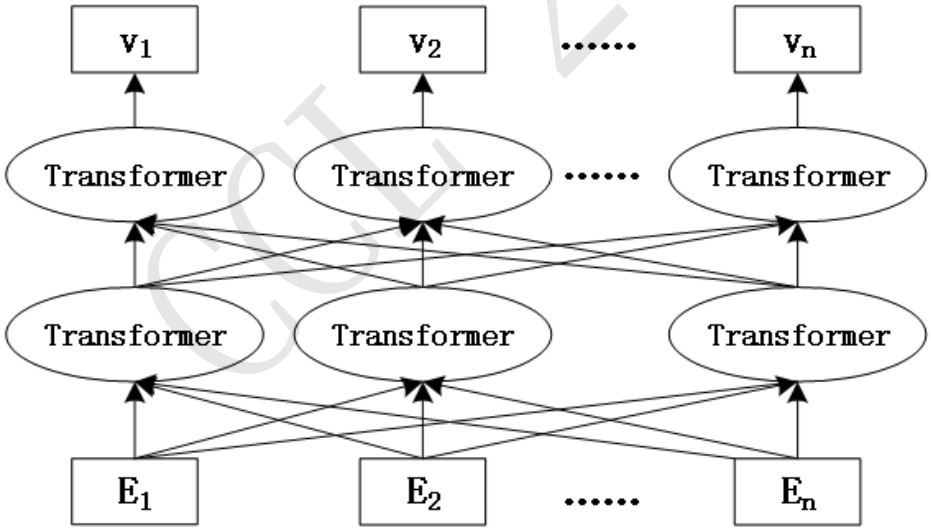

Figure 3: RoBERTa-WWM模型结构

3.2 语义信息获取层

对比词嵌入层得到的句子级表示，语义信息获取层用于获取所输入文本的更深层次的语义信息。将上一层得到句子级表示分别送入TextCNN(Kim (2014))和BiGRU(Cho et al. (2014))，中来学习文本中不同层次的局部信息和正反双向句子间的依赖关系，并将二者结果进行融合来对仇恨言论文本进行深层次的特征提取。

3.2.1 TextCNN层

TextCNN利用多个不同大小的卷积核来提取文本中的关键信息，自动对$n-gram$特征进行组合和筛选，从而能够充分捕捉文本中的局部相关信息，获得不同抽象层次的语义信息。TextCNN模型结构如图4所示。

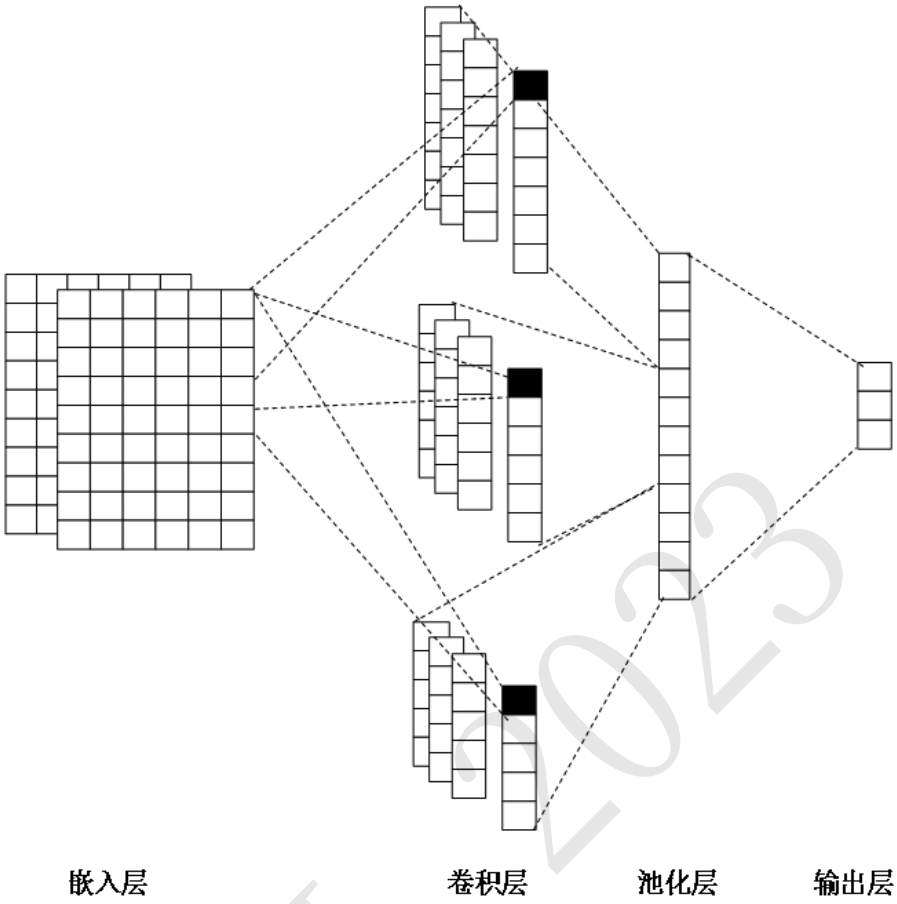

Figure 4: TextCNN模型结构图

令卷积核为$w \in R^{h \times d}$，d为卷积核的宽度，与RoBERTa-WWM层的输出维度一致，h为卷积核高度。设置三种不同的卷积核，h分别设置为3、4、5，输入通道为1，输出通道为256，保证卷积输出向量的维度一致。将卷积核w与词向量矩阵V中的第i个窗口$V_{i:i+h-1}$内的词向量进行卷积操作，得到特征c_i，计算如公式1所示。

$$c_i = f(w \cdot V_{i:i+h-1} + b) \tag{1}$$

其中，f为激活函数，b为偏置。

卷积核w与词向量矩阵V中所有窗口内词向量进行卷积操作后，得到特征图$c \in R^{n-h+1}$，如公式2所示。

$$c = [c_1, c_2, ..., c_{n-h+1}] \tag{2}$$

再进入池化层进行最大池化运算，将卷积层输出的特征进行池化操作来提取更显著的特征，卷积核w对应生成的特征图c经过池化操作得到$c^{'} = \max\{c\}$，并将池化后的特征向量进行拼接操作。通过联结所有卷积核的池化结果，得到新特征cnn_outs，计算如公式3所示。

$$cnn\_outs = [c_1^{'}, c_2^{'}, ..., c_k^{'}] \tag{3}$$

其中，k为TextCNN输出维度256。

第二十二届中国计算语言学大会论文集，第501页-第511页，哈尔滨，中国，2023年8月3日至5日。

3.2.2 BiGRU层

BiGRU层负责学习句子间双向依赖信息，其模型结构如图5所示，它是由双向的门控循环单元（Gated Recurrent Unit，GRU）组成的。可以对仇恨言论的上下文进行双向特征提取，捕获句子间双向依赖信息，以便更准确地获取上下文全局特征信息。

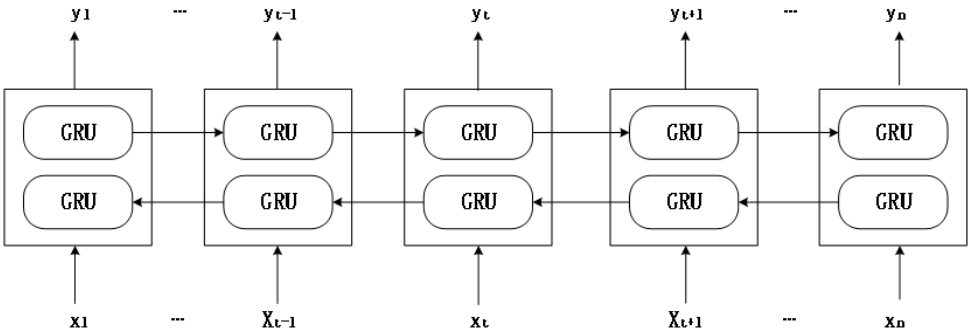

Figure 5: BiGRU模型结构

在t时刻BiGRU的双向隐状态输出分别如公式4和5所示,在t时刻的隐藏状态是双向结果地拼接，如公式6所示。

$$\overrightarrow{h_t} = \overrightarrow{GRU}(w_t, \overrightarrow{h_{t-1}}) \tag{4}$$

$$\overleftarrow{h_t} = \overleftarrow{GRU}(w_t, \overleftarrow{h_{t-1}}) \tag{5}$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{6}$$

其中，w_t为t时刻单向GRU的权重矩阵。

整段文本的隐藏状态集合gru_outs为所有时刻h_i的拼接，如公式(7)所示。

$$gru\_outs = [h_1, h_2, \ldots\ldots, h_n] \tag{7}$$

最后将两个模型得到的结果进行拼接，得到该段文本的融合语义特征信息cat_outs。计算如公式8所示。

$$cat\_outs = cnn\_outs \oplus gru\_outs \tag{8}$$

3.3 分类输出层

将融合特征表示接入全连接层，将其映射到实例标签空间，对仇恨言论进行分类。语义信息获取层的输出cat_outs与全连接层权重矩阵计算后输出M，计算如公式9所示。

$$M = \tanh(W_d \cdot cat\_outs + b_d) \tag{9}$$

其中，W_d是全连接层权重矩阵，b_d是全连接层偏置。

输出层采用$Sigmoid$函数对全连接层的输出信息M进行归一化处理，得到每个倾向类别的概率值。计算如公式10所示。

$$y = Sigmoid(W_s \cdot M + b) \tag{10}$$

其中，W_s为输出层权重矩阵，b为输出层偏置。y表示模型对每个倾向类别的概率值。

4 实验与结果分析

4.1 数据集

本节主要是对中文冒犯言论数据集COLD进行二次处理，根据仇恨言论的定义对于数据进行重新标注，分别标注仇恨言论和非仇恨言论两种类别。为了更高效地标注数据，对于训练集，采用Model-in-the-loop(Sun et al. (2021))策略来对训练数据进行标注；对于测试集，为了保证结果可靠，测试数据集依旧采用手工标注的方式。对COLD数据集进行处理后，发现仇恨言论数据只有4629条，仅占整个COLD数据集的18.23%，其中性别主题有931条，地域主题有1289条，种族主题有2409条，话题分布明显不均衡。因此为了扩充数据集规模和实现仇恨言论数据集的话题分布类型均衡，考虑从性别主义数据集SWSR（Jiang et al. (2022)）中抽取1600条性别主题相关仇恨言论，从对话偏见数据集CDIAL-BIASDATASET（Zhou et al. (2022)）中的地域主题对话中抽取1250条地域主题相关仇恨言论来对数据集做扩充，最终提出提出一个包含17430条句子，覆盖种族，性别，地域主题的仇恨言论数据集CHSD。

（1）数据收集：对于训练集的构建，基于Model-in-the-loop的设置对数据进行半自动标记。首先，通过手动标记500个样本(仇恨或非仇恨)作为训练数据。其次，对一堆未标记的数据采用分类器，并预测它们的仇恨性。然后，根据预测分数对数据进行排序，并将数据分成多个集合进行样本检查。从每个集合中抽取大约10%的数据，并使用以下策略手动标记它们:1)如果预测标签的准确率达到90%，则直接将集合中的数据加入到训练集中;否则，2)将集合完全手动重新标记，然后添加到训练集中。通过这种方法，迭代更新分类器和训练集6轮。对于测试集的构建，完全采用人工标注的策略。

（2）标注策略：邀请了3位语言安全研究者来对数据集进行手工标注。他们的性别分布为两男一女，分别来自河南，东北，北京三个地区。根据仇恨言论的定义对于数据进行标注，最终结果选取票数更高的标签作为仇恨言论标签。

（3）数据分析：基于以上流程，构建了仇恨言论数据集CHSD，数据集的详细信息如下表2、表3所示。

Table 2: CHSD数据集分布

	仇恨	非仇恨	整体
训练集	5644	7944	13588
测试集	1840	2002	3842
整体	7484	9946	17430
平均长度	53.69	44.20	48.77

Table 3: CHSD数据集样例

标签	文本
仇恨	男生就活该被女性打
非仇恨	朋友，醒醒，重男轻女时代过去了
仇恨	河南人都偷井盖
非仇恨	人家没说山东人，并不是地域黑。扣帽子可不好
仇恨	黑人肤色和五官都不好看
非仇恨	印度人数学都很好

为了保证所构建仇恨言论数据集的可靠性和可用性，采用了Kappa值来对测试集中三位标注者的结果和最终数据集结果（可以视为一个新的标注者）进行一致性检验，计算如公式11所示。其具体结果如表4所示。可以看出，3个Kappa值都高于0.6，这说明数据具有较高的一致性，构建的仇恨言论数据集质量合格。

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{11}$$

其中，p_0是每一类正确分类的样本数量之和除以总样本数。

假设每位标注者标注的仇恨和非仇恨类别的样本个数分别为a_1, a_2,模型预测到的每类样本个数分别为b_1, b_2，总体样本数为n，p_e计算如公式12所示。

$$p_e = \frac{a_1 \cdot b_1 + a_2 \cdot b_2}{n \cdot n} \tag{12}$$

Table 4: CHSD数据集的一致性检验结果

数据集	A-Result	B-Result	C-Result
CHSD	0.813	0.792	0.832

4.2 实验参数设置

为了在仇恨言论数据集上取得最优的分类结果，通过设置不同的超参数，来做多次对比试验，最后选取参数的最优值。实验的超参数设置如表5所示。

Table 5: 实验参数设置

参数	说明	最优值
batch_size	一次训练选取的样本数	32
learning rate	学习率	2e-5
epochs	训练次数	30
GRU_units	GRU输出结果的维度	128
Dropout	随机舍弃的神经元比例，防止过拟合	0.5
max_seq_len	单个语句最大长度	64

其中，batch_size与实验所使用的计算平台算力相关，综合考虑实际情况将batch_size的值设置为32，BiGRU层的隐藏单元数设置为128。

4.3 对比实验与结果分析

为了验证RoBERTa-CHSD模型在仇恨言论侦测任务中的整体性能，将其与以下几种分类基线模型进行对比分析，测试了它们的精确率、召回率和F1 值，实验结果如表6所示。

1)BERT(bert-base-chinese)：使用基于BERT的中文预训练模型做仇恨言论侦测任务。

2)ALBERT(albert-chinese-tiny)：使用基于ALBERT的中文预训练模型做仇恨言论的文本分类任务。ALBERT模型在BERT模型的基础上进行改进，使用了自监督损失函数关注构建句子中的内在连贯性。它设计了参数减少的方法，用来降低内存消耗，同时加快了BERT的训练速度。

3)RoBERTa-WWM：使用RoBERTa-WWM预训练模型做仇恨言论侦测任务。

Table 6: 各模型实验结果

模型	精确率	召回率	F1值
BERT(bert-base-chinese)	85.07%	85.23%	85.15%
ALBERT(albert-chinese-tiny)	77.47%	77.47%	77.47%
RoBERTa-WWM	87.29%	87.43%	87.36%
RoBERTa-CHSD	**89.12%**	**89.13%**	**89.12%**

从表中可以看出，RoBERTa-CHSD模型与BERT,ALBERT,RoBERTa-WWM三种基线模型中性能最好的RoBERTa-WWM对比，在精确率，召回率和F1值方面分别高出1.83%，1.7%，1.76%。值得注意的是，ALBERT的参数共享策略会导致一些特征被过度

压缩，从而导致模型性能有一定的下降；BERT模型和RoBERTa-WWM模型没有考虑向前向后的信息，因此对上下文信息的理解不够充分。综上所述，相较于典型的基线模型，提出的RoBERTa-CHSD融合模型在仇恨言论侦测任务上具有更好的性能。

以F1值为评价指标,将RoBERTa-CHSD融合模型与RoBERTa-WWM模型在单个类别的分类性能上进行对比,实验结果如图6所示。

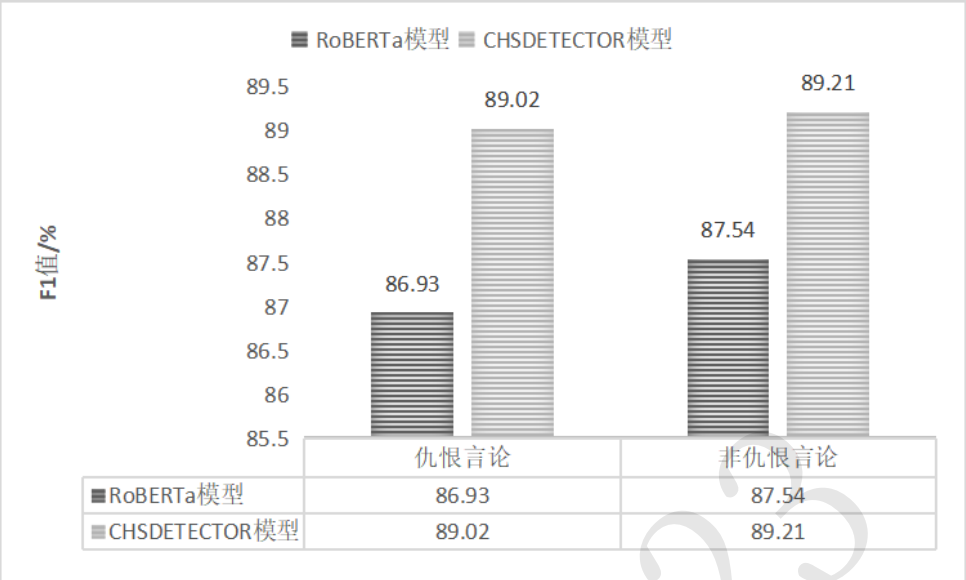

Figure 6: 不同模型的单个类别分类性能对比

相比RoBERTa-WWM模型，RoBERTa-CHSD融合模型在仇恨言论类型和非仇恨言论类型上的F1值分别提高了2.09%和1.67%，这说明融合TextCNN和BiGRU模型能够更充分地抽取文本语义特征，在各类别分类上都具有更优的性能。而且，二者都在非仇恨言论类型识别上更有优势，这是由于非仇恨言论的样本数相对而言比较多，因此模型对其语义特征学习得更充分。

4.4 消融实验设计与结果

为了进一步验证模型结构的有效性，进行消融实验，研究影响实验结果的独立因素。实验结果如表7所示。

Table 7: 不同模型实验结果对比

模型	精确率	召回率	F1值
RoBERTa-WWM	87.29%	87.43%	87.36%
RoBERTa-WWM+BiGRU	88.37%	88.06%	88.21%
RoBERTa-WWM+TextCNN	88.61%	88.50%	88.55%
RoBERTa-CHSD	**89.12%**	**89.13%**	**89.12%**

对比表中四个模型在CHSD数据集上的结果，可以发现普通基线模型RoBERTa-WWM效果最差，原因在于其只是简单地做了词嵌入的工作，未充分考虑文本的内部语义特征信息。模型RoBERTa-WWM+BiGRU利用BiGRU模型进一步提取句子间全局双向依赖信息，其F1值达到了88.21%,比单一的RoBERTa-WWM模型提升了0.85%。RoBERTa-WWM+TextCNN利用TextCNN提取多层次局部特征信息，其F1值达到了88.55%,比单一的RoBERTa-WWM模型提升了1.19%。综合来看，BiGRU和TextCNN都对文本的深层次仇恨特征提取有积极意义，因此RoBERTa-CHSD考虑融合BiGRU模型和TextCNN模型，同时提取全局信息和局部特征信息，实验结果表明，RoBERTa-CHSD模型的F1值达到了89.12%，相较于以上三种实验模型，分别提升了1.76%，0.91%，0.57%。该组对比实验体现了RoBERTa-CHSD的有效性。

5 结语

针对中文领域仇恨言论急需自动侦测的问题，提出了一个中文仇恨言论数据集CHSD，并在此基础上提出一种RoBERTa-CHSD中文仇恨言论侦测模型。该模型首先使用RoBERTa-WWM预训练语言模型捕获仇恨言论文本的语义特征，再分别使用TextCNN和BiGRU来学习文本中不同层次的局部信息和正反双向句子间的依赖关系，将二者结果进行融合来对仇恨言论文本进行深层次的特征提取，从而进一步提升了中文仇恨言论侦测模型的性能。通过实验结果可以看出，与现有的几种典型的文本分类模型相比，提出的RoBERTa-CHSD模型在仇恨言论侦测任务的整体性能上得到了有效提升。

参考文献

Ehab A Abozinadah, Alex V Mbaziira, and J Jones. 2015. Detection of abusive accounts with arabic tweets. *Int. J. Knowl. Eng.-IACSIT*, 1(2):113–119.

Swati Agarwal and Ashish Sureka. 2015. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *Distributed Computing and Internet Technology: 11th International Conference, ICDCIT 2015, Bhubaneswar, India, February 5-8, 2015. Proceedings 11*, pages 431–442. Springer.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Jiawen Deng, Jingyan Zhou, Hao Sun, Fei Mi, and Minlie Huang. 2022. Cold: A benchmark for chinese offensive language detection. *arXiv preprint arXiv:2201.06025*.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2021. On the safety of conversational models: Taxonomy, dataset, and benchmark. *arXiv preprint arXiv:2110.08466*.

Xiangru Tang, Xianjun Shen, Yujie Wang, and Yujuan Yang. 2020. Categorizing offensive language in social networks: A chinese corpus, systems and an explanation tool. In *Chinese Computational Linguistics: 19th China National Conference, CCL 2020, Hainan, China, October 30–November 1, 2020, Proceedings 19*, pages 300–315. Springer.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Frame, datasets, and benchmarks. *arXiv preprint arXiv:2202.08011*.

王素格卢欣. 2019. 融合语言特征的卷积神经网络的反讽识别方法. 中文信息学报, 33(5):31–38.

联合国. 2019. 《联合国关于仇恨言论的战略和行动计划》. https://www.un.org/zh/hate-speech/understanding-hate-speech/what-is-hate-speech.