基于 TextCNN-Bert 融合模型的不良信息识别技术

裴卓雄[1]，杨　敏[2]，杨　婧[2]

（1. 国家计算机网络应急技术处理协调中心，北京　100032；
2. 国家计算机网络应急技术处理协调中心山西分中心，山西　太原　044400）

摘　要：敏感领域的不良信息具有极强的迷惑性和欺骗性，腐蚀人们的思想，影响人们的价值观和判断能力，危害社会安全，研究敏感领域不良信息的识别技术具有深远意义。通用的识别技术忽略了背景知识和隐喻问题，直接应用于敏感领域不良信息识别效果较差。提出一种基于 TextCNN-Bert 的融合模型，通过敏感领域主题识别和情感隐喻识别，实现对敏感领域不良信息的文本识别。实验结果表明，该模型在准确率、*F*1 评分等指标方面取得了良好的结果，相较于现有模型有显著提高。

关键词：敏感领域；TextCNN；Bert；融合模型

中图分类号：TP399　　**文献标识码**：A　　**DOI**：10. 19358/j. issn. 2097 - 1788. 2023. 08. 012

引用格式：裴卓雄，杨敏，杨婧. 基于 TextCNN-Bert 融合模型的不良信息识别技术［J］. 网络安全与数据治理，2023，42(8)：72 - 76.

Bad information identification technology based on TextCNN-Bert fusion model

Pei Zhuoxiong[1]，Yang Min[2]，Yang Jing[2]

(1. National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC), Beijing 100032, China; 2. National Computer Network Emergency Response Technical Team/Coordination Center of China (Shanxi), Taiyuan 044400, China)

Abstract: The bad information in sensitive areas is extremely confusing and deceptive, corrodes people's thinking, affects people's values and judgment, and endangers social security. Research on the identification technology of bad information in sensitive areas has far-reaching significance. The general recognition technology ignores background knowledge and metaphor problems, and the effect of direct application to sensitive areas is poor in the recognition of bad information. This paper proposes a fusion model based on TextCNN-Bert, which realizes the text recognition of bad information in sensitive areas through topic recognition and emotional metaphor recognition. The experimental results show that the proposed model achieves good results in terms of accuracy, *F*1 score and other indicators, which are significantly improved compared with the existing models.

Key words: sensitive areas; TextCNN; Bert; fusion mode

0　引言

随着互联网行业蓬勃发展，网络上不良信息的泛滥引发了诸多社会问题，特别是历史、时政新闻等敏感领域的不良信息，通过编排、篡改、杜撰、伪造的方式，具有极强的迷惑性和欺骗性，腐蚀人们的思想，影响人们的价值观和判断能力，危害社会安全[1]。文本作为主要传播方式，研究敏感领域不良信息的识别技术具有深远意义。

自然语言处理技术（Natural Language Processing，NLP）能够对文本进行深入分析和理解，从而实现文本的分类和识别。Kim[2]提出一种用于文本分类的卷积神经网络模型 TextCNN，可以在一定程度上避免梯度消失的问题，而且在处理短文本和固定长度文本时表现良好。Lai[3]提出了文本分类模型 RCNN，同时结合了卷积神经网络和循环神经网络的优点。Wang[4]比较不同循环神经网络模型在文本分类任务中的性能，表明了 LSTM 模型在文本分类的优势。Devlin[5]提出了 BERT 模型，该模型是一种基于 Transformer 网络的预训练模型，可用于自然语言处理任务，如文本分类、语言推断等。Chen[6]提出了一种基于双向情感表情符号嵌入和基于注意力的 LSTM 的

Twitter 情感分析方法，该方法使用双向 LSTM 来学习句子中的上下文信息，使用注意力机制来加强对重要信息的关注，使用情感表情符号来增强情感分类的精度。李志杰[7]提出一种基于 LSTM 和 TextCNN 的联合模型，捕捉文本中的上下文关系和局部特征，提高短文本分类的准确性。Sanagavarapu[8]提出 BiLSTM 和人工神经网络 ANN 组成的混合模型，通过上下位词的概念获取新闻的语义并映射到 ANN 模型上，提升对新闻文章分类的准确性。Rehman[9]提出了一种基于 CNN-LSTM 的混合模型，用于提高电影评论情感分析的准确性。该模型利用 CNN 提取局部特征，LSTM 则用于学习序列信息，从而结合了两种模型的优点。

敏感领域属于专业领域，不良信息的识别技术研究十分有限，通用的识别技术可以直接应用于识别，但存在以下问题：一是领域特定语言和术语问题。敏感领域具有丰富的领域特定语言和术语，这些语言和术语可能对于通用模型不易理解，从而导致文本识别准确率下降。二是背景知识问题。敏感领域涉及敏感事件、人物和背景等方面的知识，这些知识对于模型来说可能是未知的，需要进行特殊的处理才能进行识别和理解。三是文本复杂性的问题。敏感领域文本非常复杂，包含大量的隐喻、比喻和引申意义，这些都需要模型具备识别和理解的能力。

因此，本文将敏感领域不良信息的识别问题转化为敏感领域主题识别任务和情感隐喻识别任务，提出一种基于 TextCNN-Bert 融合模型，既利用 TextCNN 对关键词和局部特征更加敏感的优势，准确识别敏感领域的特定语言和术语；又能利用 Bert 的预训练能力和自注意力机制，提升对隐喻、比喻和引申意的识别。实验结果表明，本模型在准确率、召回率、精确率等方面识别效果良好。

1 词向量

词向量技术是一种将文本中的单词或短语表示为向量的技术。基于 NLP 技术实现文本分类的第一步就是利用词向量表示文本。传统的 NLP 方法是基于离散符号表示的，即将每个单词表示为一个唯一的标识符或索引。这种方法没有考虑到单词之间的语义关系，因此无法捕捉到单词之间的相似性和相关性。而词向量技术通过将每个单词表示为一个向量，使得语义上相似的单词在向量空间中距离较近，从而可以更好地捕捉到单词之间的语义关系，如 Word2Vec[10]、GloVe[11]、ELMo[12] 等。Word2Vec 模型核心思想是将每个词表示为一个向量，通过计算词向量之间的余弦相似度来衡量词之间的相似度。GloVe 是一种基于全局词频统计的词向量学习方法，将单词的共现信息转化为向量空间中的距离关系。ELMo 的核心思想是通过训练深度双向语言模型来生成上下文相关的词向量表示，优点在于能够捕捉单词在不同上下文中的语义和语法信息，从而提高自然语言处理任务的性能。

2 敏感领域识别模型

本文提出的 TextCNN – Bert 融合模型如图 1 所示，模型输入为经过预处理的文本序列 $X = \{x_1, x_2, x_3, \cdots, x_n\}$ $(n > 0)$，预处理过程包括分词、词性标注和去除停用词，输出为敏感领域的判定结果。识别模型包含敏感领域主题识别和情感隐喻识别两个模块。若敏感领域识别为假，则判定与敏感领域无关，为非敏感文本。若识别为真，则作为情感隐喻识别的输入进行判定。若情感隐喻识别为真，则判定为不良信息，若判定为假，则判定为一般信息。

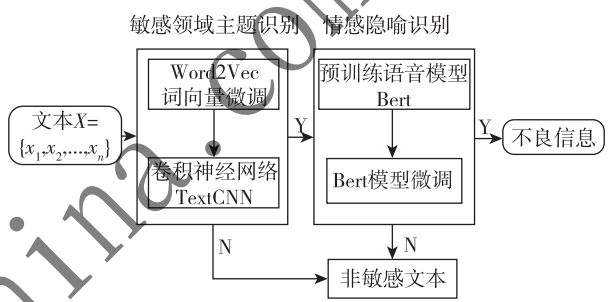


图 1 敏感文本识别模型

2.1 敏感领域主题识别模型

2.1.1 Word2Vec 词向量微调

Word2Vec 特征领域词库微调是指在特定领域的词库上对已经训练好的 Word2Vec 模型进行微调，以得到更适合该领域的词向量。

如图 2 所示，首先准备敏感领域语料和公开的大规模语料；其次使用大规模的语料库训练通用的 Word2Vec 模型，得到通用的词向量表示；然后获得敏感领域的专业术语和常用词汇，构建领域词库；最后对领域相关的词向量进行微调更新。

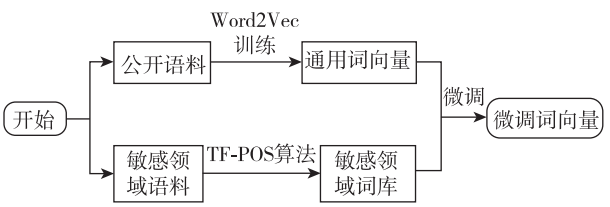


图 2 Word2Vec 词向量微调

本文结合敏感领域词汇特点，提出基于 TF-POS 算法的敏感领域词库构建算法。通过统计词频和词性分析的方式获取领域词汇。

一个词汇在敏感领域文本中出现的频率是判断其与该敏感领域相关性的重要特征，统计词频的公式如下：

$$TF_i = \frac{x_i}{\sum_k X_k} \tag{1}$$

领域词性 POS = {nr，ns，nt，nz，j}，其中 nr 表示人名，ns 表示地名，nt 表示机构团体名称，nz 表示其他专有名词，j 表示缩略语。人物、机构、事件、时间、地点等信息在敏感领域具有特殊意义。

2.1.2 卷积神经网络 TextCNN

如图 3 所示，TextCNN 第一层为输入层，用于接收输入的文本序列，将其转化为词嵌入向量，每个单词对应一个向量，并将这些向量按序列顺序组成一个矩阵。第二层为卷积层，通过多个不同大小的卷积核对输入的文本矩阵进行卷积操作，从而提取文本的局部特征。第三层为池化层，用于压缩特征图的维度和提取重要的特征。第四层为全连接层，将池化层的输出连接到一个或多个全连接层，用于学习特征之间的关系和进行最终的分类。最后一层为输出层，输出结果为敏感领域和非敏感领域两个类别。

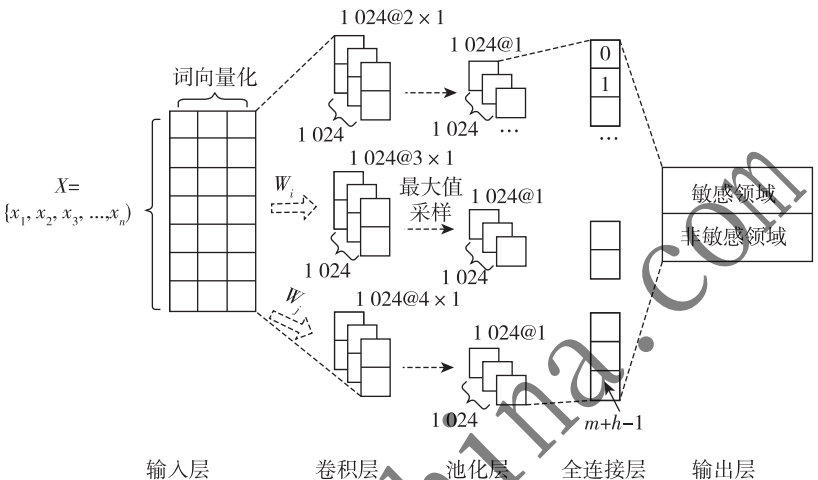


图 3 卷积神经网络 TextCNN

2.2 情感隐喻识别模型

敏感领域不良信息表达内容隐晦，具有隐蔽性和迷惑性外套的包装，与正常内容具有极强的混淆性，因此，准确识别出敏感领域不良信息的关键在于能否识别语义的隐喻。BERT（Bidirectional Encoder Representations from Transformers）是一种预训练的自然语言处理模型，适用于语义隐喻的识别和理解。

如图 4 所示，本模型的输入为敏感领域的预处理文本 $X = \{x_1, x_2, x_3, \cdots, x_n\}(n > 0)$，输出为判定结果。第一步是将输入的文本序列进行词向量处理。第二步是经过 Transformer 编码层提取文本中的语义信息，该层由多个 Transformer block 组成，每个 block 由多头自注意力机制和前馈神经网络组成。第三步经过 Bert 预训练任务层，提取深层次的语义信息。最后经过 Softmax 函数实现文本的分类，输出不良信息和一般信息两种标签。

本文提出的情感隐喻识别模型，需经过 Bert 预训练和 Bert 模型微调两个步骤得到。

2.2.1 Bert 预训练语言模型

Bert 的预训练过程分为两个阶段，分别是掩码语言建模（Masked Language Model，MLM）和下一句预测（Next Sentence Prediction，NSP）。MLM 阶段中，Bert 输入一段文本，并将其中的部分单词替换为 Mask 或其他随机

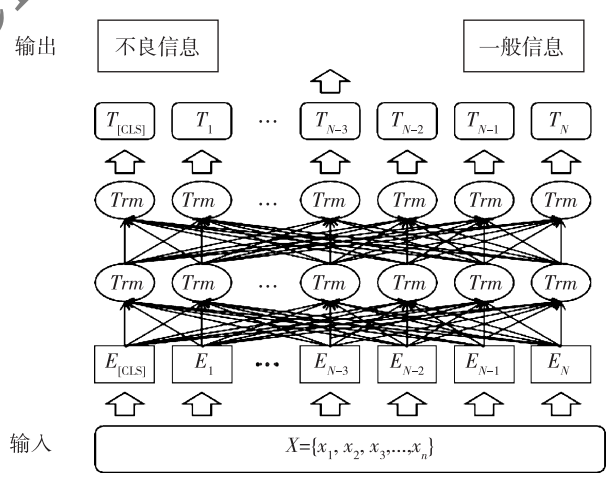


图 4 情感隐喻识别模型

单词，模型的目标是预测这些被替换的单词。NSP 阶段中，Bert 输入两个句子，并预测这两个句子是否是连续的。该任务的目的是让模型理解两个句子之间的关系。

本文采用 Hugging Face 发布的开源预训练模型 bert-base-chinese，这是基于 uncased 数据集训练的 Bert 模型，包含 12 层、768 个隐藏单元和 12 个注意力头，适用于中文文本分类等任务。

2.2.2 Bert 模型微调

Bert 模型微调是指在预训练阶段基础上，将模型进

一步训练以适应具体任务的过程。本文将敏感领域的一般信息和不良信息作为训练集和测试集进行输入，根据损失函数和评价指标来对模型进行训练和调优。模型微调时需要用到交叉熵损失函数：

$$\text{Loss} = \frac{1}{N}\sum_{i} -[y_i \times \log(p_i) + (1-y_i) \times \log(1-p_i)] \tag{2}$$

其中，y_i 表示样本 i 的标签，正类为 1，负类为 0；p_i 表示样本 i 预测为正类的概率。

3 实验及分析

3.1 实验数据

本文数据集分为三个部分：第一部分实验数据是非敏感领域数据，数据来源于搜狗实验室的全网新闻数据，本文从中筛选出汽车、科技、健康、体育、房产、教育、旅游、文化、IT、时尚共 10 个类别，每个类别约包含 2 000篇文本。第二部分实验数据是敏感领域一般信息数据，第三部分是敏感领域不良信息数据。

经过人工处理和标注，数据集分布情况为非敏感领域数据 82 万条语句，敏感领域一般信息数据 73 万条语句，敏感领域不良信息数据 78 万条语句。同时，按 6 ∶ 2 ∶ 2的比例将标注数据集划分为训练集、验证集和测试集。

3.2 实验设计

3.2.1 基线模型

为了验证基于 TextCNN-Bert 融合模型的不良信息识别方法的有效性，选取 TextCNN、LSTM、Bert 作为基线模型。

3.2.2 实验环境与模型参数设置

本文应用的深度学习框架为 PyTorch，服务器操作系统为 Windows10，使用深度学习框架 Keras 开发，且其底层支持为 TensorFlow。模型参数设置如表 1、表 2 所示。

表 1 TextCNN 模型参数

参数	值
词向量维度	300
Dropout_ pro（丢弃率）	Adam Optimezer
batch_ size	64

表 2 Bert 模型参数

参数	值
词向量维度	768
Dropout_ pro（丢弃率）	Adam Optimezer
batch_ size	64

3.2.3 评估方法

实验采用的评价指标为准确率（Accuracy）、精确率（Precision），召回率（Recall）和 $F1$ 值。混淆矩阵如表 3 所示。

表 3 混淆矩阵

实际值	预测值	
	正类	负类
正类	TP	FN
负类	FP	TN

准确率是指所有预测为正类占总数的比例。

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

召回率是指所有正确预测为正类占全部实际为正类的比例。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

精确率是指预测为正类的样本中，实际为正类的样本所占的比例。

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

$F1$ 值综合了精确率和召回率，把 Precision 和 Recall 的权重看作是一样的，是基于两者的调和平均，通常作为一个综合性的评价指标，$F1$ 值越高，代表模型的性能越好。

$$F1 = \frac{2 \times \frac{\text{TP}}{\text{TP} + \text{FP}} \times \frac{\text{TP}}{\text{TP} + \text{FN}}}{\frac{\text{TP}}{\text{TP} + \text{FP}} + \frac{\text{TP}}{\text{TP} + \text{FN}}} \tag{6}$$

3.3 实验结果

如表 4 所示，本文提出的 TextCNN－Bert 融合模型在评价指标方面优于 TextCNN、LSTM、Bert 等分类模型。TextCNN、LSTM 的 Precision 明显低于其他指标，原因在于模型无法理解深层次语义，导致将敏感领域一般信息判定为不良信息，Bert 模型指标低于本文提出模型，原因在于其对敏感领域专有词汇不敏感，导致将网友吐槽等不相关内容判定为不良信息。

表 4 各模型识别效果对比

Model	Accuracy	Recall	Precision	$F1$
TextCNN	0. 754 6	0. 743 2	0. 607 8	0. 668 6
LSTM	0. 775 3	0. 728 8	0. 643 9	0. 683 7
Bert	0. 813 7	0. 789 6	0. 693 7	0. 738 5
TextCNN-Bert	0. 855 9	0. 810 2	0. 769 7	0. 842 8

4 结论

本文提出一种基于 TextCNN-Bert 融合模型的识别方法，相比传统方法，能够更准确地识别敏感领域的术语和隐喻内容，大幅提升识别效果。未来的研究可以探索如何引入更强大的大语言模型，例如 GPT－3 或 T5 等，这些模型在文本生成和理解任务上表现出了卓越的性能。通过引入这些最新的大语言模型，可以为敏感领域不良信息识别效果带来更大的提升和改进。

参考文献

[1] 郑博熙，程达王. 网络空间意识形态斗争的特征分析 [J]. 网络安全技术与应用，2015 (10)：106－107.

[2] KIM Y. Convolutional neural networks for sentence classification [J]. arXiv preprint arXiv，2014：1408. 5882.

[3] LAI S，XU L，LIU K，et al. Recurrent convolutional neural networks for text classification [C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence，2015：2267－2273.

[4] WANG Z，YANG J，LI X，et al. A comparative study of LSTM and GRU for text classification [C]//2018 IEEE International Conference on Big Data，2018：3643－3648.

[5] DEVLIN J，CHANG M W，LEE K，et al. Bert：pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv，2018：1810. 04805.

[6] CHEN Y，YUAN J，YOU Q，et al. Twitter sentiment analysis via Bi-Sense emoji embedding and attention-based LSTM [C]// Proceedings of the 26th ACM International Conference on Multimedia，2018：117－125.

[7] 李志杰，耿朝阳，宋鹏. LSTM-TextCNN 联合模型的短文本分类研究 [J]. 西安工业大学学报，2020，40 (3)：299－304.

[8] SANGAVARAPU S，SRIDHAR S，CHITRAKALA S. News categorization using hybrid BiLSTM-ANN model with feature engineering [C]//2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC). IEEE，2021：134－140.

[9] REHMAN A U，MALIK A K，RAZA B，et al. A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis [J]. Multimedia Tools and Applications，2019，78 (18)：26597－26613.

[10] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv，2013：1301. 3781.

[11] PENNINGTON J，SOCHER R，MANNING C. Glove：Global vectors for word representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)，2014：1532－1543.

[12] PETERS M E，NEUMANN M，IYYER M，et al. Deep contextualized word representations [C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，2018：2227－2237.

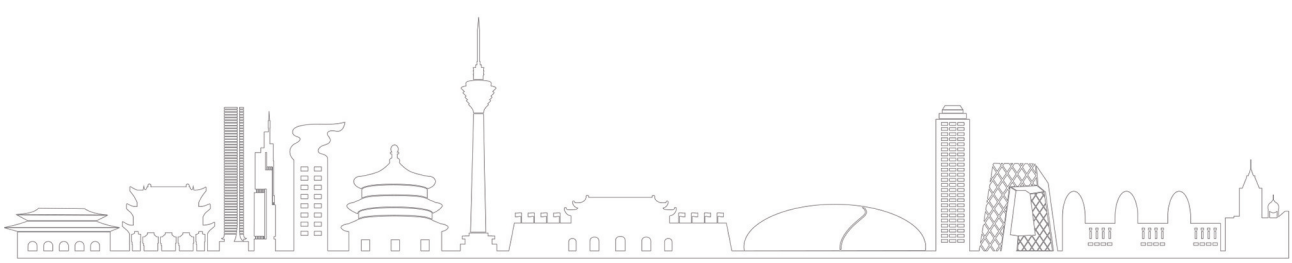
（收稿日期：2023－04－27）

作者简介：

裴卓雄（1993－），男，硕士，工程师，主要研究方向：信息安全、自然语言处理。

杨敏（1995－），男，工程师，主要研究方向：信息安全。

杨婧（1988－），女，硕士，工程师，主要研究方向：信息安全。



版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部