文章编号: 1001-4543(2024)01-0079-09 DOI: 10.19570/j.cnki.jsspu.2024.01.010

融合对抗训练与 ERNIE 的短文本情感分析模型

刘 婷[a,b], 杜 奕[a,b], 曹晓夏[a,b], 侯淏文[a,b]

(上海第二工业大学 a. 计算机与信息工程学院; b. 人工智能研究院, 上海 201209)

摘 要: 使用深度学习技术进行文本情感分类是近年来自然语言处理领域的研究热点, 好的文本表示是提升深度学习模型分类性能的关键因素。由于短文本蕴含情感信息较少、训练时易受噪声干扰, 因此提出一种融合对抗训练的文本情感分析模型 PERNIE_RCNN。该模型使用 ERNIE 预训练模型对输入文本进行向量化, 初步提取文本的情感特征。随后在 ERNIE 预训练模型的输出向量上添加噪声扰动, 对原始样本进行对抗攻击生成对抗样本, 并将生成的对抗样本送入分类模型进行对抗训练, 提高模型面临噪声攻击时的鲁棒性。实验结果表明, PERNIE_RCNN 模型的文本分类性能更好, 泛化能力更优。

关键词: 短文本情感分析; 深度学习; 对抗训练; 文本分类

中图分类号: TP391 **文献标志码**: A

A Short Text Affective Analysis Model Combining Adversary Training and ERNIE

LIU Ting[a,b], DU Yi[a,b], CAO Xiaoxia[a,b], HOU Haowen[a,b]

(School of Computer and Information Engineering; b. Institute for Artificial Intelligence, Shanghai Polytechnic University, Shanghai 201209, China)

Abstract: Text sentiment classification using deep learning techniques is a hot research topic in the field of natural language processing in recent years, and good text representation is a key factor in improving the classification performance of deep learning models. A text sentiment analysis model PERNIE_RCNN that includes adversarial training is proposed, as short texts contain little sentiment information and are susceptible to noise interference during training. The model uses the ERNIE pre-trained model to vectorize the input text and initially extract the sentiment features of the text. The model then adds noise perturbations to the output vector of the ERNIE pre-training model to generate adversarial samples against the original samples, and feeds the generated adversarial samples into the classification model for adversarial training to improve the robustness of the model against noise attacks. The experimental results show that the PERNIE_RCNN model has better text classification performance and better generalisation ability.

Keywords: short text sentiment analysis; deep learning; adversarial training; text classification

0 引言

随着网络技术的不断发展, 社交网络逐渐兴起, 各种社交平台的用户数量也在不断增加。用户在网络平台上对各种热门事件发表看法, 产生了许多蕴含情感信息的短文本数据。有关部门可根据这些文本中蕴含的情感信息对相关事件进行针对性的处理和问题的改进。但是, 仅凭借人工力量对这些文本数

收稿日期: 2023-02-15

通信作者: 杜 奕 (1977–), 女, 江苏吴江人, 教授, 博士, 主要研究方向为人工智能, 智能数据分析。E-mail: duyi@sspu.edu.cn

基金项目: 国家自然科学基金 (41672114, 41702148), 中国教育部科发中心产学研创新基金 (2021ZYA03008) 资助

据进行收集和处理非常困难。为了解决这个难题, 引入了文本情感分析方法 [1], 该方法可以帮助人们更高效地对网络文本数据进行情感挖掘和了解事件的舆情走向, 为有关部门制定相关政策提供数据支持。

文本情感分析实质上是将文本按照不同的情感极性进行分类, 属于自然语言处理 (NLP) 领域 [2] 任务之一。现有的文本情感分析方法可分为: 基于情感词典的情感分析方法 [3]、基于传统机器学习的情感分析方法 [4] 和基于深度学习的情感分析方法 [5-7]。基于深度学习的情感分析方法无需人工提取特征, 而是通过构建深度神经网络来分析文本并提取特征。这种方法具有较强的语义表达能力, 因为深度神经网络可以学习和理解语言的复杂结构和含义, 从而对文本进行更加准确和全面的情感分析。但若想从根本上提高模型的分类性能, 除了要有复杂高效的分类网络外, 好的文本表示也非常重要。近年来, 迁移学习取得了迅速发展, 它可以利用预训练模型已经学习到的知识进行迁移。通过这种方式可以在新任务上有效地提高模型性能, 尤其是在数据量较少的情况下。2018 年谷歌提出了一个名为 BERT[8] 的预训练模型来提高文本表示的质量。在 BERT 模型的基础上, Sun 等 [9] 提出了知识增强的语义表示 (enhanced representation through knowledge integration, ERNIE) 模型, 相较于 BERT, ERNIE 使用了大量的中文数据集来进行训练, 使模型训练结果更加符合中文文本特性。

为了提高深度学习模型的抗干扰能力, 一些研究者提出了使用对抗训练的方法来优化模型。该方法通过在训练中引入对抗样本来使模型更好地抵御干扰和攻击。对抗训练最初被广泛运用于图像领域, Goodfellow 等 [10] 在 2015 年最早提出了对抗训练的概念, 并成功运用在了计算机视觉领域。Miyato 等 [11] 首次将对抗训练运用到了文本分类领域, 提升了文本分类模型的泛化能力。张晓辉等 [12] 2020 年提出了一种基于对抗训练的文本分类算法, 该算法在传统词向量上添加扰动噪声, 实验结果表明, 该算法可以在多个文本分类数据集上取得优秀的分类性能。Wang 等 [13] 提出了一种快速文本对抗攻击方法, 称为快速梯度投影法 (filtered gaussian process method, FGPM), 该方法基于同义词替换, 能够有效提高深度学习模型的鲁棒性并阻止对抗性示例的可迁移性。Li 等 [14] 提出了一种虚拟对抗训练方法, 引入了基于 token 级别的累积扰动词汇, 更好地初始化扰动, 从而提高模型对抗性攻击的防御能力。Chen 等 [15] 提出了一种名为特征级对抗训练 (feature-level adversarial training, FLAT) 的新型特征级对抗训练方法, 它在神经网络中结合了变分词掩码来学习全局词的重要性, FLAT 可以有效地提高神经网络模型的性能, 并降低其遭遇对抗性攻击的风险。陈立潮等 [16] 为了提升短文本分类模型的健壮性, 将训练集按不同比例进行分类对抗训练, 该方法使模型拥有了更强的防过拟合能力。尽管对抗训练在文本分析领域已经取得了一定的研究成果, 但还有许多值得深入研究的问题, 例如对抗训练可以与其他技术相结合, 如迁移学习、元学习等, 本文将对抗训练与预训练模型相结合, 以进一步提高整体模型的鲁棒性和泛化能力。

基于上述调查与研究, 为了提高短文本情感分析模型文本表示的质量以及在遭遇对抗攻击时提升模型的鲁棒性, 本文提出一种融合对抗训练的文本情感分析模型 PERNIE_RCNN, 以提升模型在面临噪声攻击时的正确分类能力。PERNIE_RCNN 模型使用 ERNIE 预训练模型代替传统的词向量表示, 对输入的短文本进行向量化, 初步获取文本的情感信息。随后在 ERNIE 预训练模型的输出向量上, 采用小步多走的策略进行对抗攻击, 生成对抗样本。并将对抗样本送入分类模型进行对抗训练, 以提高模型在面对对抗攻击时的鲁棒性。通过对抗训练与 ERNIE 预训练模型的融合, PERNIE_RCNN 模型不仅获得了更好的文本表示, 还提升了整体模型的泛化性能。

1 ERNIE 预训练模型

1.1 ERNIE 整体结构

ERNIE 预训练模型是一种基于知识增强的语义理解模型, 它能够学习文本数据中的词法结构、句法结构及语义信息, 从而使模型获取到高质量的词向量。ERNIE 模型在结构上可以分为 transformer[17] 编码和知识整合两个部分。前者用来生成词向量, 后者整合短语或实体知识获得文本整体的语义表示。ERNIE 的模型结构图如图 1 所示。

由图 1 可知, $\boldsymbol{E}[\boldsymbol{E}_{[\mathrm{CLS}]}, \boldsymbol{E}_1, \cdots, \boldsymbol{E}_n]$ 为输入文本向量, 经过 transformer 编码跟知识整合后, 得到

ERNIE 模型的输出向量 $[\boldsymbol{C}, \boldsymbol{T}_1, \cdots, \boldsymbol{T}_n]$。其中, "$\boldsymbol{C}$" 包含学习到的整体的语义信息, 可将其运用到各种下游任务, 如文本情感分析任务。

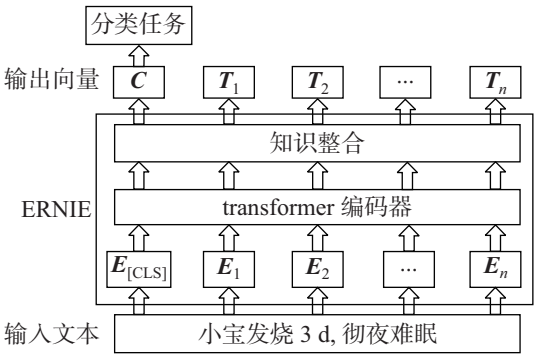

图 1　ERNIE 模型结构图

Fig. 1　Structure of the ERNIE model

1.2　transformer 编码器

ERNIE 预训练模型使用多层双向 transformer 作为基础编码器。Vaswani [18] 等在 2017 年首次提出了 transformer 架构。transformer 采用注意力机制来压缩和提取文本信息, 使得文本中重要的单词或短语具有更大的权重, 从而避免模型忽略这些关键信息。transformer 编码器中的多头注意力机制由多个自注意力机制组成, 能够从不同的角度对句子中的重要信息进行学习和处理。自注意力机制的输出表示:

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\Big(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\Big)V \quad (1)$$

式中: $\boldsymbol{Q}$ 为查询矩阵; $\boldsymbol{K}$ 为被查询向量矩阵; $\boldsymbol{V}$ 为值矩阵; d_k 为 $\boldsymbol{Q}$、$\boldsymbol{K}$ 的维度, 并通过点积运算获得 $\boldsymbol{Q}$ 矩阵和所有 $\boldsymbol{K}$ 矩阵之间的相似度。softmax 函数用来进行归一化处理, 使得模型更容易处理长远距离的特征依赖。

1.3　知识整合

ERNIE 的掩码采用的是多阶段的遮蔽 (MASK) 策略, 它通过整合短语或实体知识来获取文本整体的语义表示。BERT 与 ERNIE 模型遮蔽策略对比图如图 2 所示。

从图 2 可以看出, BERT 模型将字作为一个语义单元, 它在训练中学习到更多的是字与字之间的关系, 例如, "胞" 与 "双" "胎" 之间的关系, 属于局部关系。而 ERNIE 模型修改了遮蔽的粒度大小, 它会遮蔽掉一些连续的字, 即一些实体词, 这样不仅可以学习到如 "双胞胎" "共同" 等实体语义单元, 还能学习到 "姐妹" 与 "共同" 之间的关系, 使得模型可以学习到完整概念的语义表示。

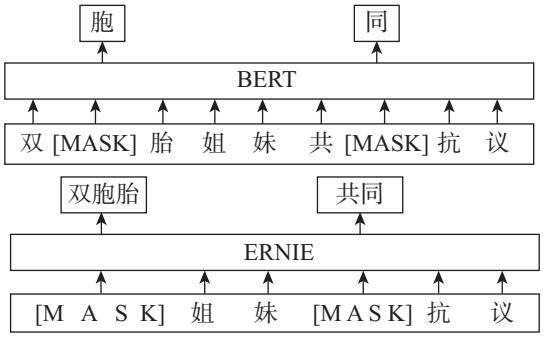

图 2　BERT 与 ERNIE 模型遮蔽策略对比图

Fig. 2　Comparison figure of BERT and ERNIE model masking strategies

2　PERNIE_RCNN 短文本情感分析模型

为了提高短文本表示的质量、提升模型的抗干扰能力, 并提取文本更深层次的情感语义, 本文提出一种融合对抗训练的短文本情感分析模型 PERNIE_RCNN, 模型整体结构如图 3 所示。

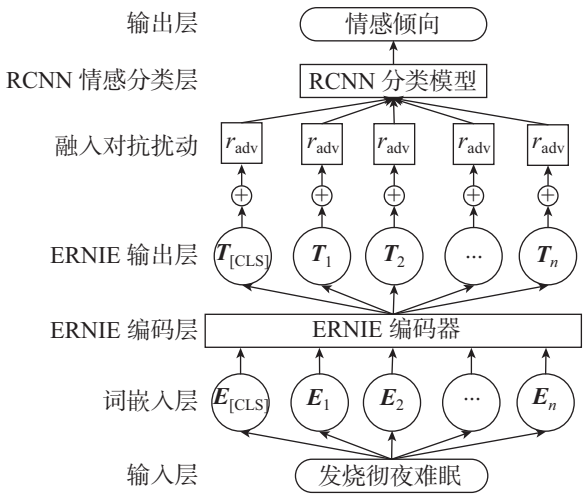

图 3　PERNIE_RCNN 模型整体结构图

Fig. 3　Overall structure of the PERNIE_RCNN model

由图 3 可知, 在 PERNIE_RCNN 模型中, 首先将文本数据送入 ERNIE 预训练模型进行文本表示, 利用其输出向量进行对抗训练, 以增强模型的鲁棒性和泛化能力, 从而提升整体模型的分类准确性。随后使用 RCNN 模型进行文本分类, 提取文本更深层次的语义特征, 并输出模型的预测结果。其中, $\boldsymbol{E}$ 为文本的词嵌入表示; $\boldsymbol{T}$ 为 ERNIE 的输出向量; r_{adv} 为添加的对抗扰动; [CLS] 为句子的开始标记。

2.1 PERNIE 文本表示模型

2.1.1 PERNIE 模型整体结构

为了提高模型在处理文本数据时的鲁棒性和泛化能力, 本文提出一种融合对抗训练和 ERNIE 预训练模型的文本表示模型 PERNIE。PERNIE 模型的整体结构如图 4 所示。

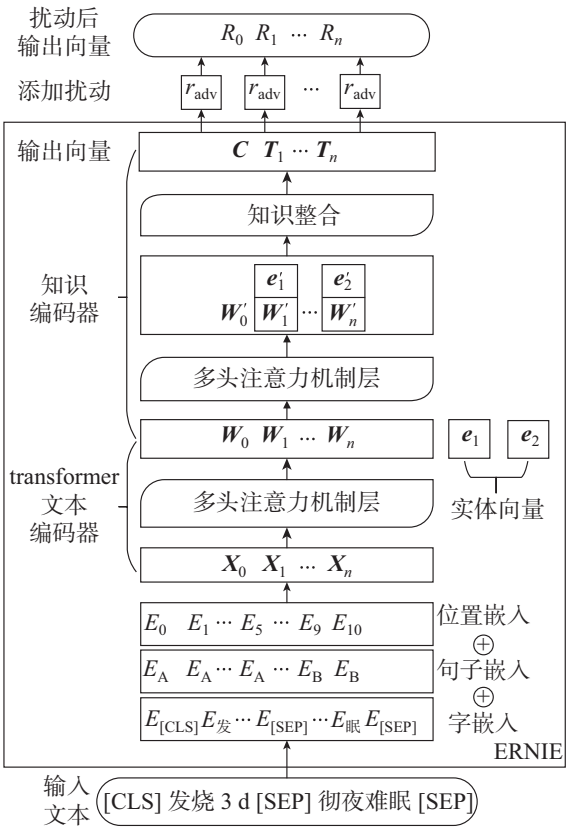

图 4 PERNIE 模型的整体结构图

Fig. 4 Overall structure of the PERNIE model

由图 4 可知, PERNIE 模型首先使用 ERNIE 预训练模型对输入文本进行向量表示。ERNIE 模型在预训练时除了使用 BERT 模型的中文数据集外, 还加入了百度的中文数据集, 因此很适用于中文文本的处理。ERNIE 的输入由两个句子组成, [CLS] 为开始标记, [SEP] 为一个句子的结束标记。输入向量 $\boldsymbol{X}$ 由 3 种不同的向量拼接而成, 其中 “位置嵌入” 代表词的位置信息、“字嵌入” 代表词向量编码、“句子嵌入” 代表句子的位置编码。输入向量经过 transformer 编码后, 得到学习后的表示向量 $\boldsymbol{W}$。随后将表示向量 $\boldsymbol{W}$ 和经过 transformer 编码后的实体向量 $\boldsymbol{e}$ 共同传入知识编码器进行融合, 生成 ERNIE 模型的输出向量 $\boldsymbol{T}[\boldsymbol{C}, \boldsymbol{T}_1, \cdots, \boldsymbol{T}_n]$。其中, 向量 “$\boldsymbol{C}$” 包含了学习到的整体的情感语义信息。最终在 ERNIE 模型输出向量上添加噪声扰动 r_{adv}, 生成扰动后的输出向量 $\boldsymbol{R}[\boldsymbol{R}_0, \boldsymbol{R}_1, \cdots, \boldsymbol{R}_n]$。

2.1.2 对抗训练

采用对抗训练来增强文本分类模型的抗干扰能力, 实际上是一种数据增强的手段, 有研究者会想到用同义词替换来进行文本数据的增强。例如: 在进行文本向量化时, 输入的文本为: (共同战疫双胞胎姐妹都是在读研究生), 经过分词后的内容为 (共同, 战疫, 双胞胎, 姐妹, 都是, 在读, 研究生), 随后在词表中寻找到对应词的词向量组合在一起, 就得到了整个句子的向量表示。进行同义词替换操作, 只需要将句子向量中 “共同” 的词向量替换为 “一起” 的词向量, 就会得到新的输入向量。如果直接在某些词上加入细微扰动, 也能得到新的输入向量, 这种新向量的获得方式称为噪声扰动。同义词替换与噪声扰动对比图如图 5 所示。

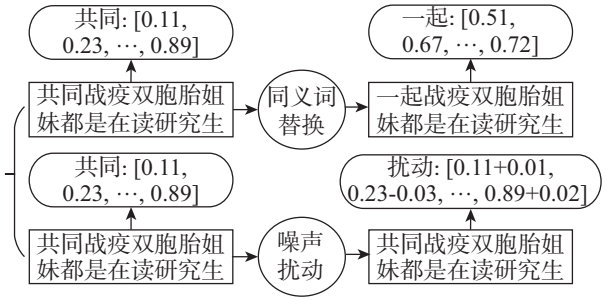

图 5 同义词替换与噪声扰动对比

Fig. 5 Comparison of synonym substitution and noise perturbation

在文本数据上添加噪声扰动, 有离散型和连续型两种类型。离散型扰动是在将文本送入文本表示模型前添加, 此时的噪声添加在字符层面。连续型扰动是在文本经过文本表示模型后, 在其输出向量中添加。PERNIE 模型使用的是连续型扰动方式。

PERNIE 模型进行对抗训练的流程可简述为: 在模型训练时, 往损失函数值上升的方向, 对连续的 ERNIE 模型输出向量发起对抗攻击, 添加噪声扰动 r_{adv}, 生成扰动后的输出向量 $\boldsymbol{R}[\boldsymbol{R}_0, \boldsymbol{R}_1, \cdots, \boldsymbol{R}_n]$, 所有的扰动向量组成最终的对抗样本 R_{adv}(由原始样本 x 和扰动项 r_{adv} 共同组成)。

$$\min_{\theta} -\log P(y|x + r_{\mathrm{adv}}; \theta) \tag{2}$$

式中, x 为原始数据样本; y 为样本标签; r_{adv} 为添加的噪声扰动; θ 为模型参数。

本文 r_{adv} 的计算采用 PGD [19] 计算公式, 其思想为: 在限定扰动的范围前提下, 将扰动沿着不同梯

度迭代多次, 缓慢寻找扰动的最优点。若迭代过程中超出扰动范围, 则将扰动投影回正常范围, 即:

$$\boldsymbol{x}_{t+1} = \Pi_{x+S(x_t+\alpha\cdot g(x_t)/\|g(x_t)\|_2)} \tag{3}$$

$$g(x_t) = \nabla_x L(\theta, x_t, y) \tag{4}$$

式中, $S = r \in R^d$ 且 $\|r\|_2 \leqslant \epsilon$; $\boldsymbol{x}_t$ 为原始样本词向量矩阵; g 为输入模型梯度; ϵ 为扰动半径; α 为迭代步长; $\|\cdot\|_2$ 为 L_2 正则; S 为约束空间; $\boldsymbol{x}_{t+1}$ 为扰动后的输出向量。对抗训练的实现过程如表 1 所示。

表 1　对抗训练的实现过程
Tab. 1　The realization process of adversarial training

输入: 对抗训练轮数 T, 扰动半径 ϵ, 迭代步长 α, 数据批次 N、PGD 扰动步数 K, 网络模型 f_θ (θ 为模型参数)
输出: 模型参数更新

1. for t in range(T):
2. for i in range(N):
　　$\delta = 0$ //随机初始化
3. for j in range (K):
4. $\delta = \delta + \alpha * \text{sign}(\nabla\delta L(f_\theta(xt+\delta), yi))$
5. $\delta = \max(\min(\delta, \epsilon), -\epsilon)$
6. $\theta = \theta - \nabla\delta L(f_\theta(xt+\delta), yi)$ //更新模型参数

2.2　RCNN 文本分类模型

文本分类任务除了要有好的文本表示, 文本特征的提取也很重要。虽然已使用 ERNIE 预训练模型对短文本进行了初步情感提取, 但由于短文本蕴含的情感信息量有限, 还需要提取出更深层次的情感特征, 以提升模型情感识别的准确性。RCNN 是 Lai 等 [20] 提出的一种名为循环卷积神经网络的融合神经网络分类模型, 其使用双向循环神经网络代替 TextCNN 进行特征提取。RCNN 缓解了传统 RNN 模型训练速度慢且无法解决长距离依赖问题, 并且在卷积层使用了大小不同的卷积核来提取大小不同的特征值, 能提取更深层次的特征信息。RCNN 的模型结构如图 6 所示。

由图 6 可知, RCNN 分类模型结构包含 BiLSTM 层、最大池化层、全连接层 3 个部分, 后接 softmax 函数输出最终的分类结果。

BiLSTM 层可细分为两个小部分: 首先得到词的左上文与右下文的信息, 即:

$$C_l(w_i) = f(\boldsymbol{W}^{(l)}C_l(w_{i-1}) + \boldsymbol{W}^{(sl)}e(w_i - 1)) \tag{5}$$

$$C_r(w_i) = f(\boldsymbol{W}^{(r)}C_r(w_{i-1}) + \boldsymbol{W}^{(sr)}e(w_i - 1)) \tag{6}$$

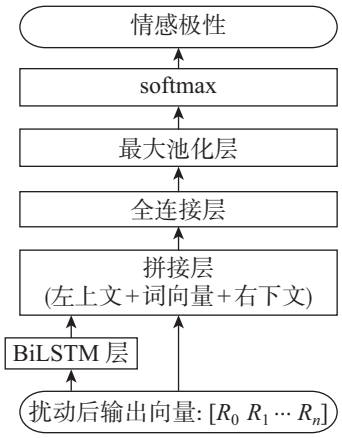

图 6　RCNN 模型结构图
Fig. 6　RCNN model structure diagram

式中, $C_l(w_i)$ 为词 w_i 的左上文; $C_r(w_i)$ 为词 w_i 的右下文; $e(w_i)$ 为词 w_i 的词向量。$\boldsymbol{W}^{(l)}$ 为权重矩阵。$\boldsymbol{W}^{(sl)}$ 目的是将当前词的语义传递到下一个词的左上文信息中。

其次计算词 w_i 的潜在语义向量, 其表达式为:

$$x_i = [C_l(w_i); e(w_i); C_r(w_i)] \tag{7}$$

$$\boldsymbol{y}_i^{(1)} = \tanh(\boldsymbol{W}^{(1)}x_i + b^{(1)}) \tag{8}$$

式中, x_i 为词 w_i 的左上文、词向量、右下文拼接起来的结果。$\boldsymbol{y}_i^{(1)}$ 将拼接后的向量非线性映射到低维, 计算词 w_i 的潜在语义向量。

最大池化层的公式为

$$\boldsymbol{y}^{(2)} = \max_{i=1}^{n} \cdot \boldsymbol{y}_i^{(1)} \tag{9}$$

它将每一个 $\boldsymbol{y}^{(1)}$ 中的最大值提取出来, 组成新的组合向量 $\boldsymbol{y}^{(2)}$, 最大池化可帮助模型找到文本句子中最重要的潜在语义。全连接层的公式表示为

$$\boldsymbol{y}^{(3)} = \boldsymbol{W}^{(3)}\boldsymbol{y}^{(2)} + b^{(3)} \tag{10}$$

将 $\boldsymbol{y}^{(2)}$ 进行全连接, 并通过 softmax 函数输出分类结果, softmax 函数公式为

$$P_j = \frac{\exp(\boldsymbol{y}_j^{(3)})}{\sum_{k=1}^{n}\exp(\boldsymbol{y}_k^{(3)})} \tag{11}$$

RCNN 学习时可以大范围保留输入文本的词序, 并自动判断分类过程中最重要的特征, 提取更深层次的语义信息, 使得模型分类结果更加准确。

3　实验与分析

3.1　实验数据

本文使用 DataFountain 竞赛平台公开的共计 10

万条微博文本数据 nCoV_100k_labled [21] 进行实验, 数据已由人工进行标注, 分为 3 类: 1(积极), 0(中性) 和 −1(消极)。

为了使模型能够更好的训练, 需要对数据进行预处理。首先使用 Python 中的 Pandas 工具包对数据进行异常值处理, 剔除情感分类不属于 1(积极), 0(中性) 和 −1(消极) 的数据, 再对数据进行去重以及剔除无意义符号等操作, 数据预处理前后对比如表 2 所示。

表 2 数据预处理前后对比
Tab. 2 Before and after data preprocessing

处理前	处理后
#新年心愿#魔幻, 2020 年第一天感冒发烧躺了一天…… 哪儿都没有去	新年心愿魔幻, 2020 年第一天感冒发烧躺了一天哪儿都没有去
//@孕事: 宝宝发烧怎么办? 这些应对方法一定要学会!	孕事: 宝宝发烧怎么办? 这些应对方法一定要学会!

预处理后剩余 99 913 条数据, 并将情感标签修改为 0(消极), 1(中性) 和 2(积极)。将所有数据按比例 6: 2: 2 切分为训练集、验证集、测试集。训练集样本分布见图 7, 文本长度分布见图 8。

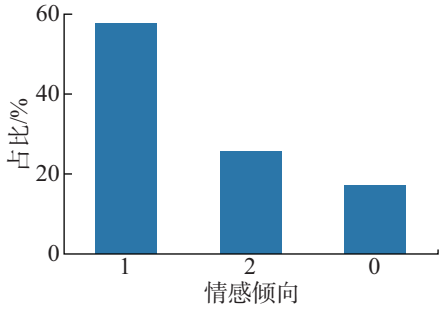

图 7 训练集样本分布图
Fig. 7 Distribution of samples in the training set

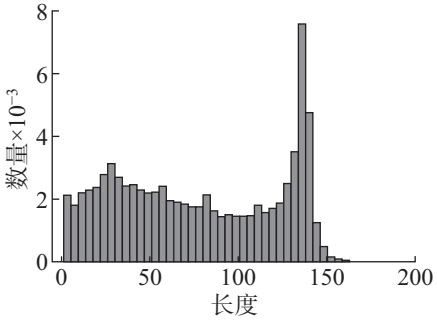

图 8 文本长度分布图
Fig. 8 Text length distribution

由图 7 可以看出, 训练集的样本分类别并不平衡, 其中, 1(中性) 类别占比最大。从整体上看, 大概有 60% 持中立态度, 25% 持积极态度, 还有约 15% 持消极态度。因此, 在模型测试时, 需考虑类别不均衡对测试结果的影响。

由图 8 可以看出, 训练集中文本长度几乎在 0∼160 词之间, 因此在模型训练设定词嵌入维度时, 可以设置为 160, 在保证信息的真实性前提下避免空间的浪费。

3.2 实验环境

实验环境如表 3 所示。

表 3 实验环境
Tab. 3 The experimental environment

实验环境	配置
操作系统	Windows10(64 位)
开发工具	Pycharm
开发语言	Python3.9
深度学习框架	Pytorch-gpu-1.7.1

3.3 实验过程

本文在文本表示部分融合了对抗训练和 ERNIE 预训练模型, ERNIE 预训练模型选用 “ERNIE 1.0-Base-Chinese” 版本, 并使用 GPU 设备进行训练。ERNIE 模型重要训练参数如表 4 所示。

表 4 ERNIE 模型重要训练参数
Tab. 4 Important training parameter

ERNIE 模型参数	参数值
每批次训练数据大小	32
隐藏层数量	768
Self-attention 头数	12
transformer 编码器层数	12
学习率	10^{-5}
最大句长	160

由表 4 可知, 每批次训练数据大小设置为 32, 设置过小容易收敛慢, 设置过大容易超过显存极限。隐藏层数量、Self-attention 头数、transformer 编码器层数均采用标准模型的默认值 768、12、12。学习率根据模型收敛情况调整为 10^{-5}。如图 8 所示, 训练集样本中文本长度几乎在 0∼160 之间, 因此最大句长设置为 160, 可节省空间。

3.4 评价标准

由图 8 可知, 数据集样本分布并不平衡, 因此使用模型在测试集上的加权平均后的精确率 (precision, P)、召回率 (recall, R) 和 $F1$ 值 ($F1_{\text{-score}}$) 来作

为评判标准, 加权平均考虑了每个类别样本的数量在总样本中的占比:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{12}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

$$F1_{\mathrm{score}} = \frac{2PR}{P + R} \tag{14}$$

式中, TP 为实际分类正确且预测分类正确; FP 为实际分类错误但预测分类正确; TN 为实际分类错误且预测分类错误; FN 为实际分类正确但预测分类错误。

加权平均操作是对每个类别的 P、R 和 $F1$ 相加后的和求平均值, 再乘以每个类别占总样本的比例:

$$\mathrm{Pro}(i) = \frac{\mathrm{count}(i)}{\mathrm{count(all)}} \tag{15}$$

$$P_{\mathrm{weighted\_avg}}(i) = \frac{\sum_{i=0}^{i=n} P(i)}{n} \cdot \mathrm{Pro}(i) \tag{16}$$

$$P_{\mathrm{weighted_{avg}}}(i) = \frac{\sum_{i=0}^{i=n} R(i)}{n} \cdot \mathrm{Pro}(i) \tag{17}$$

$$F1_{\mathrm{weighted\_avg}}(i) = \frac{\sum_{i=0}^{i=n} F1_{\mathrm{score}}(i)}{n} \cdot \mathrm{Pro}(i) \tag{18}$$

式中, 式 (15) 为每个类别占总体样本数量的比例; count(i) 为类别 i 的数量; count(all) 为总样本数量。式 (16)∼(18) 分别为 P、R、$F1$ 的加权平均计算公式; n 为类别数。

3.5　实验结果及对比分析

为了综合评估 PERNIE_RCNN 模型的性能, 本文设置了 4 组对比实验。第 1 组对比实验使用不同的情感分类模型验证 PERNIE_RCNN 模型的有效性; 第 2 组对比实验验证不同的词向量模型对模型分类性能的影响; 第 3 组对比实验验证对抗训练的有效性; 第 4 组对比实验验证模型在不同数据集上的性能。

3.5.1　不同情感分析模型的对比实验

为了验证本文提出的模型 PERNIE_RCNN 的有效性, 与 TextCNN、BiLSTM、RCNN、BERT、ERNIE、ERNIE_BiGRU、ERNIE_TextCNN、ERNIE_RCNN 模型进行了对比实验。实验结果如表 5 所示, 其中 P、R 及 $F1$ 值均为加权平均后的结果。

由表 5 可知, PERNIE_RCNN 模型的 P、R 和 $F1$ 值均优于其他模型。其中, BERT 的 $F1$ 值分别高于 TextCNN、BiLSTM 模型 11.74% 和 5.6%, 说明较传统文本表示模型使用预训练模型进行文本表示对文本分类效果有较大提升。ERNIE 的 3 个指标均优于 BERT, 说明 ERNIE 对 BERT 掩码机制的改进提高了对中文文本的分类性能。ERNIE_RCNN 的 3 个指标均优于 ERNIE_BiLSTM、ERNIE_TextCNN 模型, 说明 ERNIE 模型后接 RCNN 模型较后接 BiLSTM、TextCNN 模型提取情感特征的效果更好, 模型的分类准确率更高。PERNIE_RCNN 的 3 个指标均优于 ERNIE_RCNN, 其中 R 提高了 0.33%, 说明模型的泛化能力得到了提升, 证明了对抗训练对提高整体模型分类性能的有效性。

表 5　不同情感分析模型的测试结果
Tab. 5　Test results of different sentiment analysis models

模型	P/%	R/%	$F1$
TextCNN	69.21	56.74	0.594 8
BiLSTM	66.30	67.34	0.656 2
RCNN	71.28	71.77	0.706 4
BERT	71.36	72.02	0.712 2
ERNIE	72.03	72.53	0.720 2
ERNIE_BiLSTM	75.36	75.64	0.754 0
ERNIE_TextCNN	75.55	75.68	0.756 0
ERNIE_RCNN	75.61	75.85	0.756 5
PERNIE_RCNN	76.00	76.18	0.760 4

3.5.2　不同词向量模型的对比实验

为了验证不同词向量模型对整体模型分类性能的影响, 本文分别使用以下 3 个模型进行对比实验:

(1) Word2Vec_RCNN: 使用词嵌入模型 (Word2Vec) 进行文本的向量表示, 然后送入 RCNN 模型进行分类训练。

(2) BERT_RCNN: 使用 BERT 预训练语言模型进行文本的向量表示, 然后送入 RCNN 模型进行分类训练。

(3) ERNIE_RCNN: 使用 ERNIE 预训练语言模型进行文本的向量表示, 然后送入 RCNN 模型进行分类训练。

对比实验结果如表 6 所示。

表 6　不同词向量模型的测试结果
Tab. 6　Test results for different word vector models

模型	P/%	R/%	$F1$
Word2Vec_RCNN	71.28	71.77	0.706 4
BERT_RCNN	73.14	73.65	0.730 4
ERNIE_RCNN	**75.61**	**75.85**	**0.756 5**

由表 6 所示, BERT_RCNN 与 ERNIE_RCNN 模型的 3 个指标均优于 Word2Vec_RCNN 模型, 说明使用预训练语言模型进行文本的向量表示较使用传统的 Word2Vec 模型分类效果更优, 预训练语言模型得到的动态词向量较 Word2Vec 模型得到的静态词向量更能

体现相同词向量在不同语境下的含义。ERNIE_RCNN 模型的 3 个指标均优于 BERT_RCNN 模型, 说明 ERNIE 对 BERT 掩码机制的改进以及对中文数据集的扩充提高了模型处理中文文本的性能。

3.5.3 验证对抗训练的有效性

(1) 分类性能的比较。由于样本类别分布不平衡, 为验证对抗训练对 ERNIE_RCNN 模型分类性能的影响, 对比了 ERNIE_RCNN 与 PERNIE_RCNN 模型的 ROC 曲线和 PR 曲线。ROC 曲线在不限制阈值的前提下, 同时考虑了分类模型对各个类别的分类能力。PR 曲线在样本不平衡的情况下可以对模型的性能做出更合理的评价。ROC 曲线如图 9 所示。

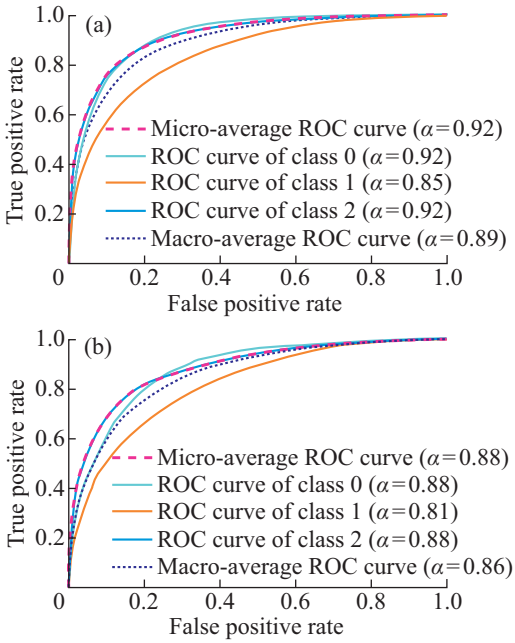

图 9 PERNIE_RCNN (a) 与 ERNIE_RCNN (b) 的 ROC 曲线对比

Fig. 9 ROC curves for PERNIE_RCNN (a) versus ERNIE_RCNN (b)

由图 9 可以看出, PERNIE_RCNN 模型 3 个类别的 ROC 曲线均比 ERNIE_RCNN 更靠近左上角, 且 PERNIE_RCNN 模型 3 个类别的 AUC 面积 (ROC 曲线下面积) 均高于 ERNIE_RCNN 模型, 说明加入对抗训练后的模型的准确性更高, 模型更理想。PR 曲线如图 10 所示。

由图 10 可以看出, PERNIE_RCNN 模型的 AP 值 (PR 曲线下面积) 高于 ERNIE_RCNN 模型, 说明加入对抗训练后, 模型的分类性能更优。

(2) 泛化性能及鲁棒性的比较。为验证对抗训练对模型泛化性能及鲁棒性的影响, 对比了 ERNIE_RCNN 与 PERNIE_RCNN 模型在 SMP2020-EWECT 数据集 [22] 上的迁移泛化能力, 并使用其中的通用微博数据集作为测试集。测试集的标注类别有 0(消极)、1(中性)、2(积极)3 个类别, 共 5 000 条数据。测试结果如表 7 所示, 其中 P、R 及 $F1$ 值均为加权平均后的结果。

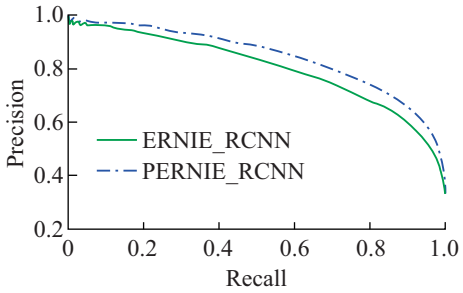

图 10 PERNIE_RCNN 与 ERNIE_RCNN 的 PR 曲线对比

Fig. 10 PR curves of PERNIE_RCNN versus ERNIE_RCNN

表 7 模型在 SMP2020-EWECT 上的迁移泛化能力测试结果

Tab. 7 Migration generalisation ability test results of the model on SMP2020-EWECT

模型	P/%	R/%	$F1$
ERNIE_RCNN	80.21	68.94	0.699 8
PERNIE_RCNN	**80.43**	**69.18**	**0.703 1**

由表 7 可知, 在 SMP2020-EWECT 测试集上, PERNIE_RCNN 模型的 P、R 和 $F1$ 值均高于 ERNIE_RCNN 模型, 说明将添加对抗训练后的模型迁移运用到其他数据集上, 模型的鲁棒性更优、泛化能力更强, 证明了对抗训练对提升模型泛化能力及鲁棒性的有效性。

3.5.4 在不同数据集上的对比实验

为了验证模型在不同数据集上的性能, 采用公开数据集 weibo_senti_100k 进行实验, weibo_senti_100k 包含 10 万多条带情感标注的新浪微博文本数据, 正负向评论约各 5 万条。对文本进行数据预处理后, 将所有数据按比例 6: 2: 2 切分为训练集、验证集、测试集。各模型在该数据集上的表现如表 8 所示。

表 8 不同模型在 weibo_senti_100k 数据集上的测试结果

Tab. 8 Test results of different models on the weibo_senti_100k dataset

模型	P/%	R/%	$F1$
TextCNN	94.88	94.88	0.948 8
BiLSTM	95.38	95.37	0.953 7
BERT	97.92	97.83	0.978 3
ERNIE	98.12	98.08	0.980 8
ERNIE_TextCNN	98.03	98.00	0.980 0
ERNIE_BiLSTM	98.22	98.17	0.981 7
ERNIE_RCNN	98.38	98.33	0.983 3
PERNIE_RCNN	**98.45**	**98.42**	**0.983 9**

由表 8 所示, 各个模型在 weibo_senti_100k 数据集上的表现与在 nCoV_100k_labled 数据集上具有相同的规律。说明 PERNIE_RCNN 模型在不同数据集上依然具有更优的性能, 且具有更强的泛化性。

4 结　语

本文提出一种融合对抗训练的短文本情感分析模型 PERNIE_RCNN。通过实验证明了对抗训练与 ERNIE 预训练模型的融合, 不仅获得了更佳的文本表示, 同时也提升了整体模型的泛化性能及鲁棒性。并将 PERNIE_RCNN 模型成功运用在了微博文本情感分析任务中。政府可使用此模型在其他类似公共事件中及时掌握大众的主要情感倾向, 为有关部门制定相关政策提供数据支持。本文目前只对网络平台短文本进行了情感分析, 而用户除了使用文字进行情感表达外, 发表的图片、表情、视频等也蕴含大量的情感信息。因此, 后续工作将考虑将文字、图片、表情等融入模型进行多模态分析, 使情感分析结果更加准确。

参考文献：

[1] 王婷, 杨文忠. 文本情感分析方法研究综述 [J]. 计算机工程与应用, 2021, 57(12): 11-24.

[2] 赵京胜, 宋梦雪, 高祥. 自然语言处理发展及应用综述 [J]. 信息技术与信息化, 2019(7): 142-145.

[3] 罗浩然, 杨青. 基于情感词典和堆叠残差的双向长短期记忆网络的情感分析 [J]. 计算机应用, 2022, 42(4): 1099-1107.

[4] 戚天梅, 过弋, 王吉祥, 等. 基于机器学习的外汇新闻情感分析 [J]. 计算机工程与设计, 2020, 41(6): 1742-1748.

[5] 邢长征, 李珊. 文本情感分析的深度学习方法 [J]. 计算机应用与软件, 2018, 35(8): 102-106.

[6] 刘春磊, 武佳琪, 檀亚宁. 基于 TextCNN 的用户评论情感极性判别 [J]. 电子世界, 2019(3): 48.

[7] 方悦, 张琨, 张云纯, 等. 基于特征融合深度学习网络的情感分析模型 [J]. 计算机与数字工程, 2022, 50(6): 1239-1245.

[8] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2019-09-16]. https://arxiv.org/abs/1810.04805.

[9] SUN Y, WANG S H, LI Y K, et al. ERNIE: Enhanced representation through knowledge integration [EB/OL]. [2019-04-19]. https://arxiv.org/abs/1904.09223.

[10] GOODfELLOW I J, SHLENS J, SZEGEDY C. Explaining and Harnessing Adversarial Examples [EB/OL]. [2015-02-25]. https://arxiv.org/abs/1412.6572.

[11] MIYATO T, DAI A M, GOODfELLOW I J. Virtual Adversarial Training for Semi-Supervised Text Classification [EB/OL]. [2016-05-25]. https://arxiv.org/abs/1605.07725.

[12] 张晓辉, 于双元, 王全新, 等. 基于对抗训练的文本表示和分类算法 [J]. 计算机科学, 2020, 47(S1): 12-16.

[13] WANG X, YANG Y, DENG Y, et al. Adversarial training with fast gradient projection method against synonym substitution based text attacks [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Online: Association for the Advancement of Artificial Intelligence, 2021, 35(16): 13997-14005.

[14] LI L, QIU X. Token-aware virtual adversarial training in natural language understanding [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Online: Association for the Advancement of Artificial Intelligence, 2021, 35(9): 8410-8418.

[15] CHEN H, JI Y. Adversarial training for improving model robustness? look at both prediction and interpretation [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver Canada: Association for the Advancement of Artificial Intelligence, 2022, 36(10): 10463-10472.

[16] 陈立潮, 秦杰, 陆望东, 等. 自注意力机制的短文本分类方法 [J]. 计算机工程与设计, 2022, 43(3): 728-734.

[17] 杜朋, 卢益清, 韩长风. 基于 Transformer 模型的商品评论情感分析 [J]. 中文信息学报, 2021, 35(2): 125-132.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need [C]// NIPS. Long Beach: Conference and Workshop on Neural Information Processing Systems, 2017: 1706- 3762.

[19] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2017-06-19]. https://arxiv.org/abs/1706.06083.

[20] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification [C]//National Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2015: 2267-2273.

[21] 北京市政务数据资源网. 疫情期间网民情绪识别 [DB/OL]. [2020-03-03]. https://www.datafountain.cn/competitions/423/datasets.

[22] 哈尔滨工业大学社会计算与信息检索研究中心. SMP2020 微博情绪分类技术评测 [DB/OL]. [2020-06-19]. https://smp2020ewect.github.io.