ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 3,May2024,pp. 699－709
DOI:10. 16337/j. 1004-9037. 2024. 03. 017

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

# 融合多特征和表情情感词典的性别对立言论识别方法

马子晨[1,2]，张顺香[1,2]，刘云朵[1,2]，朱广丽[1,2]

(1. 安徽理工大学计算机科学与工程学院，淮南 232001；2. 合肥综合性国家科学中心人工智能研究院，合肥 232088)

**摘　要**：为识别相关极端言论，提出了一种融合多特征和表情情感词典的性别对立言论识别方法。首先，使用BERT(Bidirectional encoder representation from transformer)提取输入文本的字符特征，并使用Word2Vec提取输入文本中五笔、郑码以及拼音3个方面的特征；然后，将这4个方面的特征进行融合，再输入到Bi-GRU(Bi-directional gated recurrent unit)网络中学习更深层次的语义信息；最后，通过全连接层加SoftMax函数计算出情感极性概率，并融合表情情感词典判别输入文本是否为性别对立言论。通过在自行收集的中文性别对立数据集上进行实验，与未加入特征和表情情感词典的方法相比，在$F_1$值上有5.19%的提升。同时，在公开中文情感分析数据集Weibo_senti_100k上进行验证，证明了本方法的泛化性。

**关键词**：性别对立；表情情感词典；多特征；BERT；Bi-GRU；Word2Vec

**中图分类号**：TP391　　**文献标志码**：A

## Gender Opposition Speech Recognition Method of Fusing Multi-feature and Emoji Sentiment Lexicon

MA Zichen[1,2], ZHANG Shunxiang[1,2], LIU Yunduo[1,2], ZHU Guangli[1,2]

(1. School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China;2. Institute of Artificial Intelligence Research, Hefei Comprehensive National Science Center, Hefei 232088, China)

**Abstract**: To identify relevant extreme speech, a gender opposition speech recognition method of fusing multi-features and emoji sentiment lexicon is proposed. Firstly, BERT (Bidirectional encoder representation from transformer) is used to extract the character features of the input texts, and Word2Vec is used to extract the Wubi, Zhengma and Pinyin features of the input texts. Then, these features are fused and fed into the Bi-GRU (Bi-directional gated recurrent unit) network to obtain the deeper semantic information. Finally, the sentiment polarities are calculated with the full-connected layer and SoftMax function combining the emoji sentiment lexicon to determine whether the input texts are related gender opposition. Compared with the method without adding multi-features and emoji sentiment lexicon, the experiments on the self-collected Chinese gender opposition dataset show that the proposed model is improved on the $F_1$ value by 5.19%. In addition, the generalization of the proposed method is verified by experiments on the public Chinese sentiment analysis dataset Weibo_senti_100k.

**Key words:** gender opposition; emoji sentiment lexicon; multi-feature; BERT; Bi-GRU; Word2Vec

---

**基金项目**：国家自然科学基金面上项目(62076006)；安徽省高校协同创新项目(GXXT-2021-008)。
**收稿日期**：2023-07-18；**修订日期**：2023-10-10

## 引　言

新浪微博作为我国最大的社交媒体平台之一，每天会产生大量的数据，是各类研究的信息来源。由于人们可以在网络中自由发表言论，这使得一些负面信息混入其中，严重地影响了其他网民的心理健康[1]。微博用户主要由年轻群体组成，年轻群体激进的言论导致微博平台中的性别对立现象尤为严重，干扰互联网的健康发展，并对社会造成负面的影响，甚至导致恐婚、恐育现象[2]。目前，性别对立已经成为各大社交媒体重点关注的问题，虽然有关性别对立言论检测的研究越来越多，但言论多以英语或西班牙语为主，且以Twitter为研究平台，关于中文性别对立言论检测的研究仍然很少。如何利用中文具有的特征，快速有效地从海量微博文本中发现性别对立相关的内容，成为当务之急。

用户在发表涉及性别对立的言论时，往往会使用一些网络语去躲避社交媒体对于非法评论的检测，这些网络语和用户所要表达的文字存在一定的联系，如：字形相似、读音相似等等。虽然，人类可以轻松的识别这些文字所表达的意思，但模型仍难以识别；同时，微博评论中存在一定数量的表情符号，这些表情符号大多和其编码所表达的情感不同[3-4]，如果将这些表情符号直接删除或使用，会造成一定的语义混乱。表1给出了关于性别对立言论的两个例句。其中，句1属于性别对立中“女拳”发言，对于人类来说，“蛹”是对男性的一种侮辱，而对于主流模型来说，“蛹”只是一个向量，难以与“男”进行联系，此时，需要通过捕获拼音特征，将“蛹”与“男”进行联系；句2属于性别对立中“男拳”发言，如果直接使用主流模型进行识别，通常会得到积极的结果，而后面表情“🙂”在网络中具有消极、嘲讽的意味，此时，需要通过表情情感词典对分类结果进行修正，得到用户表达的真实含义。因此，为了设计一个可以有效识别性别对立言论的分类模型，需要考虑以下两个方面：(1)如何使模型有效地识别微博评论中与性别对立相关的网络语，从而获得用户的真实意图；(2)如何使模型有效地识别用户在使用微博表情时的真实含义，使学习到的用户情感更加准确。

**表1　性别对立文本例句**

**Table 1　Examples of gender opposition texts**

| 序号 | 性别对立文本 |
|---|---|
| 1 | 原来是国蛹啊，那就不奇怪了！ |
| 2 | 大胆！敢说不好看，明明都是小仙女🙂 |

基于以上考虑，本文提出一种融合多特征的性别对立识别模型。在该模型中，首先使用BERT(Bi-directional encoder representations from transformer)[5]预训练模型对该领域数据进行预训练，得到字符向量；然后，使用Word2Vec[6]模型学习与字形相关的五笔和郑码特征，以及与读音相关的拼音特征，来获取多个特征向量；接着，将字符向量和特征向量进行融合，通过双向门控循环单元Bi-GRU[7]对融合后的向量进行更深层次的学习；最后，使用全连接层加SoftMax函数来计算情感极性概率，并融入表情情感词典，得到最终的分类结果。模型的整体框架如图1所示，主要由预处理层、特征处理层以及分类层组成。(1)预处理层：首先，对收集的数据集进行初步筛选，去除无用部分后采用人工的方式输入标签；然后，对数据集进行去停用词处理，并将其中存在的繁体字转换为简体字。(2)特征处理层：通过BERT和Word2Vec获得字符向量和特征向量后，将其进行融合。其中，特征向量包括拼音、五笔和郑码特征。再使用Bi-GRU对融合后的特征进一步进行学习。(3)分类层：采用全连接层加SoftMax函数的方式计算句子的情感极性概率，并结合表情情感词典得到最终的分类结果。

## 1　相关工作

性别对立言论是恶意言论的一种，在前期处理工作上与其他恶意言论相似。因此，本节分别回顾了针对恶意言论检测和性别对立言论检测的研究工作。

### 1.1 恶意言论检测工作

Warner等[8]使用支持向量机(Support vector machine, SVM)对反犹太言论进行识别,准确率达到了94%。Gao等[9]利用大规模无标记数据的弱监督方法进行在线仇恨言论检测,在性能上显著优于使用人工注释数据进行监督训练的方法。Burnap等[10]选择种族、残疾和性取向3方面对网络仇恨进行分类,使用文本解析来提取类型依赖,取得了较好的效果。Zhang等[11]将卷积神经网络(Convolutional neural network, CNN)和门控循环单元(Gated recurrent unit, GRU)结合,用于划分细粒度仇恨言论。Fan等[12]使用BERT预训练模型,收集Twitter中关于英国脱欧的相关评论,并改进BERT预训练模型进行微调测试,结果表明,该模型能够有效地检测Twitter中的恶意评论。Alotaibi等[13]提出一个多通道的深度学习模型,并在3个公开的恶性言论数据集上进行测试,取得了良好的效果。

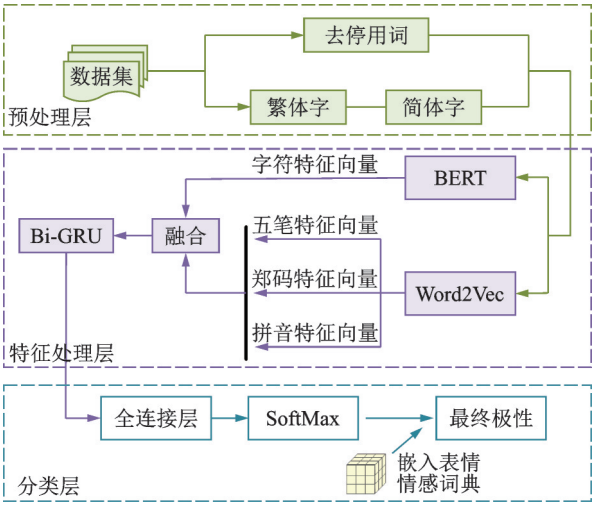


图1 模型的框架图

Fig.1 Framework of the proposed model

### 1.2 性别对立检测工作

Jha等[14]将推文分为“敌意的性别歧视”、“善意的性别歧视”以及“其他”3种类别,使关于性别歧视文本的分类更加完善。Karlekar等[15]专注于性骚扰的报道,探索使用CNN-RNN模型在识别相关报道时的可行性。Yan等[16]受量子力学的启发,使用密度矩阵编码器对关于性骚扰的个人故事进行分类。Garcia-Diaz等[17]收集了一个针对西班牙语的厌女症平衡语料库,并使用3种机器学习方法进行评估,取得了良好的效果。Rodriguez-Sanchez等[18]开发并发布了Twitter上第一个西班牙语性别歧视表达和态度的数据集,并研究了使用机器学习技术自动检测不同类型性别歧视行为的可行性。Pitsilis等[19]结合与用户信息相关的特征,提出了一种基于循环神经网络(Recurrent neural network, RNN)模型的监测方案,可以成功地从普通文本中区分出种族主义和性别歧视信息。

虽然已有许多学者在对性别对立等恶意言论进行研究,但针对中文文本的相关研究还很少,且中文与英语、西班牙语等语言在表达上有很大差异。因此,本文结合表情符号以及中文性别歧视文本的特征,对文本进行分析,最终提升识别的效果,为检测中文性别歧视言论提供理论支持。

## 2 性别对立识别模型

为解决网络中性别对立言论难以识别的问题,本文提出一种融合多特征和表情情感词典的性别对立言论识别模型,来对性别对立相关语句进行判别。图2以“因为警察大部分都是男的🙂……说不定他们也一样”进行举例,展示模型的判别过程。

### 2.1 预处理层

预处理层主要做与处理数据相关的任务。将标注好的数据输入到预处理层后,首先需要判断文本中是否包含表情符号,若存在表情则先将其单独存储,并从原文本中删除;然后将处理过表情之后的文本进行去停用词、繁-简字体转换等操作,以减少文本序列中不相关因素对最终分类结果产生的负面影响。做完一系列处理后,将处理后的文本序列作为特征处理层的输入部分。

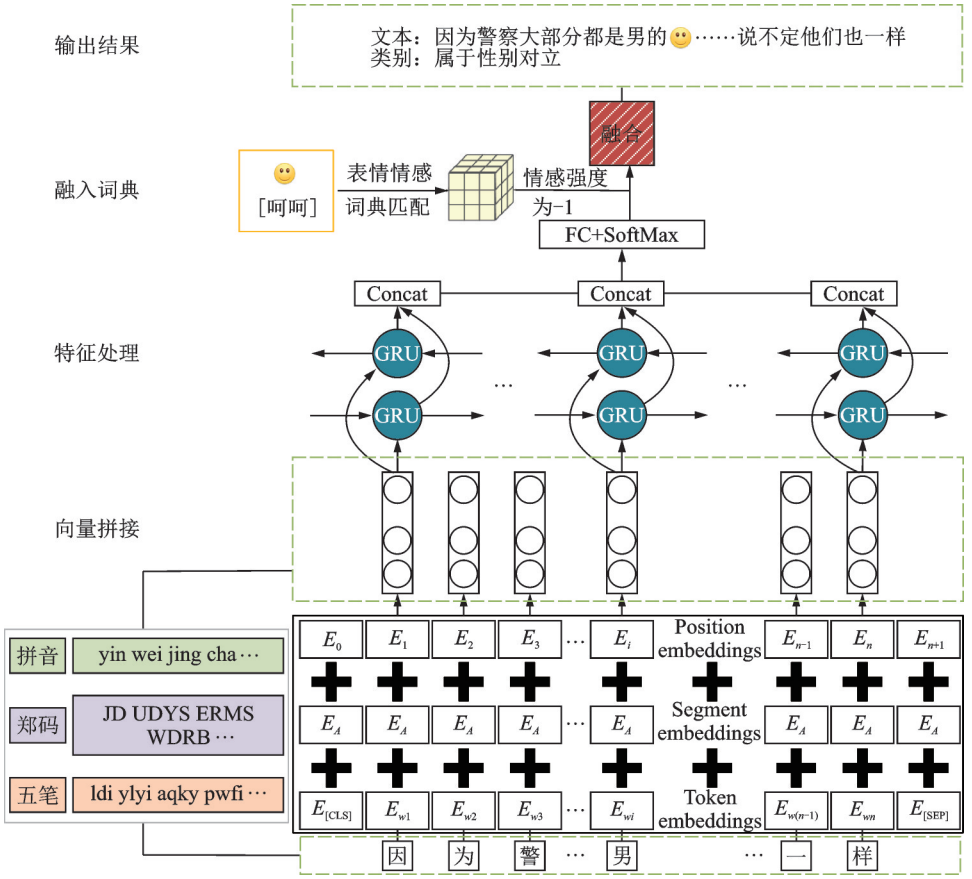


图2 模型的判别过程图

Fig.2 Recognition process of the proposed model

## 2.2 特征处理层

2018年，Devlin等[5]提出了BERT预训练语言模型，使用Transformer作为主要框架，用来捕获语句中的双向关系，并结合下一句预测（Next sentence prediction, NSP）和掩码语言模型（Masked language model, MLM）进行多任务训练。BERT在情感分析任务中应用广泛，可以挖掘出文本语句中通用领域非结构化文本的深层语义信息。因此，本文选用BERT作为性别对立文本中用于提取基础的字级别语义模型，以提高基础语义信息的质量。

Word2Vec是由Google的Mikolov等[6]提出用于生成词嵌入的模型，其由两层神经网络组成，将语料库作为输入，并生成一个向量空间，语料库中的每个不同单词都在空间中分配一个向量，向量之间的距离越近，说明单词的相似程度越高，该模型在通用领域上具有优秀的文本向量化效果。本文通过Word2Vec模型，对性别对立相关文本中的五笔、郑码以及拼音特征进行学习，获取稠密向量作为特征，与BERT模型提取的embeddings表示进行拼接。

### 2.2.1 输入序列向量化

在输入序列向量化模块中，BERT使用WordPiece作为分词器，由输入的原始序列得到词嵌入为：$\boldsymbol{S}=[\boldsymbol{w}_0,\boldsymbol{w}_1,\cdots,\boldsymbol{w}_i,\cdots,\boldsymbol{w}_n,\boldsymbol{w}_{n+1}]$，其中$S$为输入序列，$n$为句子长度，$\boldsymbol{w}_0$为句子分类标记“[CLS]”（Classfication）向量，$\boldsymbol{w}_{n+1}$为句子分隔符/结束符“[SEP]”（Separation）向量，然后使用多层双向Trans-

former网络进行编码。向量化过程可以表示为

$$\boldsymbol{V}_i^{\text{bert}}=\text{BERT}(\boldsymbol{w}_i)\quad i\in\mathbf{Z}\cap i\in[0,n+1] \tag{1}$$

式中:$\boldsymbol{w}_i$表示序列中的第$i$个字符,且$i$为0到$n+1$之间的整数;BERT(•)表示通过BERT预训练模型进行编码;$\boldsymbol{V}_i^{\text{bert}}$表示输入序列中第$i$个字符通过BERT编码后得到的对应向量。

2.2.2 五笔特征

五笔字形输入法简称五笔,是一种纯字形的编码方案,至多需要4个字母即可完成一个汉字或者一个词组的输入。文献[20-22]中,在研究过程中考虑到中文特有的字形特征,对五笔编码进行提取,从而提高模型对文本语义的识别效果。

86版五笔字形输入法对应的字根表键盘布局如图3所示。其中,每个颜色区域表示一种笔画类型,蓝色、黄色、粉色、绿色和橙色区域分别表示从右到左向下、点或从左到右向下、水平、垂直和钩状笔画,此外,五笔输入法中,将“Z”键用作通配符。

如图4所示,第1列是字形相似且含义相似的中文字符。第4列是字形相似的中文字符,它们虽然有着不同的含义,但在性别对立文本中都被用作“男”的替换。使用Word2Vec对五笔特征进行向量化,可以缓解文本中使用相似字形替代的问题。

图3 字根表键盘布局

Fig.3 Keyboard layout of the radical table

| 字符 | 字根 | 编码 | 字符 | 字根 | 编码 |
|---|---|---|---|---|---|
| 跳 | 跳跳跳 | khi | 男 | 男男 | ll |
| 跃 | 跃跃跃跃 | khtd | 偻 | 偻偻偻 | wlln |
| 蹦 | 蹦蹦蹦蹦 | khme | 娄 | 娄娄娄 | vlln |

图4 五笔编码实例

Fig.4 Examples of Wubi encoding

本文使用Python中的pywubi库,将输入序列转换为对应的五笔编码,之后使用Word2Vec模型对其进行训练,获得每个字符对应五笔编码的上下文特征,学习到在五笔特征下每个字符对应的向量。五笔特征的向量化过程如下

$$\boldsymbol{T}=f_{\text{wb}}(S) \tag{2}$$

$$\boldsymbol{W}_{\text{vec}}=\text{Word2Vec}(\boldsymbol{T}) \tag{3}$$

$$\boldsymbol{V}_i^{\text{wb}}=\boldsymbol{W}_{\text{vec}}(T_i)\quad i\in\mathbf{Z}\cap i\in[0,n+1] \tag{4}$$

式中:$f_{\text{wb}}$表示将输入序列按照pywubi库转换为对应的五笔编码;Word2Vec()表示使用Word2Vec模型按照五笔特征对序列进行向量化;$\boldsymbol{V}_i^{\text{wb}}$表示与输入序列$T_i$对应的五笔特征向量。

2.2.3 郑码特征

郑码与五笔输入法都是字形编码,但编码规则不同。郑码特征可以缓解五笔编码与拼音编码重复的问题,例如“恨”的五笔编码是“nv”,而“nv”也可以作为汉字的拼音使用,但其郑码编码为“uxo”。使用郑码特征,可以与五笔特征互相校正,提高识别的准确率。本文利用郑码和汉字的对应表将输入序列转换为对应的郑码编码,利用与训练五笔特征类似的方式,对郑码特征进行训练。郑码特征的向量化过程为

$$\boldsymbol{T}=f_{\text{zm}}(\boldsymbol{S}) \tag{5}$$

$$\boldsymbol{W}_{\text{vec}}=\text{Word2Vec}(\boldsymbol{T}) \tag{6}$$

$$\boldsymbol{V}_i^{\text{zm}}=\boldsymbol{W}_{\text{vec}}(T_i)\quad i\in\mathbf{Z}\cap i\in[0,n+1] \tag{7}$$

式中:$f_{\text{zm}}$表示按照郑码与汉字的对应表,将输入序列转换为对应的郑码编码;Word2Vec()表示使用Word2Vec模型按照郑码特征对序列进行向量化;$\boldsymbol{V}_i^{\text{zm}}$表示与输入序列$T_i$对应的郑码特征向量。

2.2.4 拼音特征

在微博评论中,有许多利用谐音词代替的网络词语,这些词作为新的表达方式,用于规避网络中的敏感词屏蔽机制,使得当前的主流模型不能充分的学习到用户在使用这些文字时所表达的真实含义,加入拼音特征可以有效地识别利用谐音代替的网络词语。文献[23-25]中,通过将拼音字符以及汉字字符等特征相结合,在模型中融入文字发音的特点,增强模型对语义的学习能力。

表2为性别对立文本中,常用谐音字替换原文字的3组词语。其中,“国男”一词指的是性别对立中“网络女拳”侮辱中国男性用词;“母人”一词指的是性别对立中“网络男拳”侮辱女性用词,和“国男”一样,带有辱骂、歧视性质;“一眼丁真”为“一眼真”的反义词,即看一眼就知道是假的,延伸后表达为幽默或嘲讽的含义。

**表2 谐音网络词实例**

**Table 2 Examples of homophonic internet words**

| 词语1 | 词语2 | 词语3 |
|---|---|---|
| 国男 | 母人 | 一眼丁真 |
| 国蝻 | 幕刃 | 一眼顶针 |
| 蝈蝻 | 牧妊 | 一眼顶真 |
| 蜾蝻 | 姆妊 | 义眼丁真 |

本文使用Python中的pypinyin库,将输入序列转换为对应的拼音编码,之后使用Word2Vec模型对其进行训练,获得每个字符对应拼音编码的上下文特征,学习到在拼音特征下每个字符对应的向量。拼音特征的向量化过程为

$$T = f_{\mathrm{py}}(S) \tag{8}$$

$$W_{\mathrm{vec}} = \mathrm{Word2Vec}(T) \tag{9}$$

$$V_i^{\mathrm{py}} = W_{\mathrm{vec}}(T_i) \quad i \in \mathbf{Z} \cap i \in [0, n+1] \tag{10}$$

式中:$f_{\mathrm{py}}$表示通过Pypinyin库,将输入序列转换为对应的拼音编码;Word2Vec()表示使用Word2Vec模型按照拼音特征对序列进行向量化;$V_i^{\mathrm{py}}$表示与输入序列$T_i$对应的拼音特征向量。将获取的特征进行拼接处理,得到最终的嵌入层向量。特征拼接过程为

$$V_i^{\mathrm{e}} = \mathrm{Contact}\left(V_i^{\mathrm{bert}}, V_i^{\mathrm{wb}}, V_i^{\mathrm{zm}}, V_i^{\mathrm{py}}\right) \tag{11}$$

## 2.3 分类层模块

在得到编码层模块的输出向量后,使用全连接层组合特征并将输出向量映射到样本标记空间,再使用SoftMax激活函数计算出每种极性对应的概率,最后融入表情情感词典,得到最终的分类结果。

结合表情符号,可以使模型对文本语义的学习更加充分,本文选取微博平台常被用户使用的131个默认表情来构建表情情感词典,如图5所示,覆盖了微博用户常用的多数表情。和文本相比,表情所包含的情感更加直观。在微博中,表情是以“[表情]”的格式存储,例如“🌞”在微博中存储为[太阳]。但微博中的大部分表情并没有承载最初想表达的意思,例如“🙂”表情,在网络中并不是最初表达的“一个愉快的微笑”的含义,而是带有一种消极态度,用户可以通过此表情表达讽刺、蔑视以及带有攻击性的情感,有时甚至可以逆转句子的情绪。本文在构建表情情感词典时,为每个表情设置对应的情感强度。部分表情情感词典如图6所示。

图5 微博默认表情

Fig.5 Default emojis on Weibo

表情情感词典中部分表情的情感强度为0,这是因为用户在使用这些表情时,或表达幽默的情感,或表达的正负向情感极性难以判断,难以给男女对立识别模型带来正向收益。如果文本序列中存在微博默认表情,则先通过表情词典对该表情进行加权计算,并与通过SoftMax函数计算出的概率相结合,得到最终分类结果,有

$$C=\mathrm{Max}\left(\lambda*E+P_{\mathrm{pos}},P_{\mathrm{neg}}\right) \tag{12}$$

式中:$P_{\mathrm{pos}}$和$P_{\mathrm{neg}}$分别为Bi-GRU进行编码后,通过全连接层加SoftMax计算得到的正向情感极性和负向情感极性的概率;$E$为此表情在词典中的情感强度;$\lambda$为表情在性别对立识别模型中的权重,经初步数据分析,本文将其定为0.17;$C$为模型最终的分类结果。

融合多特征和表情情感词典的性别对立文本识别流程如算法1所示。

| 表情 | 编码 | 情感强度 | 表情 | 编码 | 情感强度 |
|---|---|---|---|---|---|
| | [吐] | -1 | | [月亮] | 1 |
| | [黑线] | -1 | | [太阳] | 1 |
| | [弱] | -1 | | [二哈] | 0 |
| | [打脸] | -1 | | [doge] | 0 |
| | [鲜花] | 1 | | [吃瓜] | 0 |
| | [心] | 1 | | [哈欠] | 0 |

图6 部分表情情感词典

Fig.6 A part of emoji sentiment lexicon

**算法1 性别对立文本识别流程**

输入:待处理的原始文本

输出:最终的分类结果

(1) Dataset = Pre_process(Dataset) /* 对原始数据集进行预处理 */

(2) $V_{\mathrm{bert}}$=BERT(Dataset) // 字符特征向量

(3) $V_{\mathrm{wb}}$=WuBi(Dataset) // 五笔特征向量

(4) $V_{\mathrm{zm}}$=ZhengMa(Dataset) // 郑码特征向量

(5) $V_{\mathrm{py}}$=PinYin(Dataset) // 拼音特征向量

(6) $V$=Contact($V_{\mathrm{bert}}$, $V_{\mathrm{wb}}$, $V_{\mathrm{zm}}$, $V_{\mathrm{py}}$) // 将特征向量拼接

(7) $V_{\mathrm{GRU}}$=BiGRU($V$) // 通过BiGRU网络对各个特征进一步学习

(8) $V_{\mathrm{fc}}$=FC($V_{\mathrm{GRU}}$) // 通过全连接层计算得到特征向量

(9) $P_{\mathrm{s}}$=SoftMax($V_{\mathrm{fc}}$) // 通过SoftMax计算情感极性概率

(10) $P$=Fusion($P_{\mathrm{s}}$, $E$) // 融合情感词典得到最终极性概率

(11) Category=Max($P$) // 得到最终分类结果

End

将数据集进行向量化一般需要$O(n)$的时间;对向量进行融合以及再训练的时间复杂度为$O(n)$;由于算法中未出现显式的循环,算法的总时间复杂度为$O(n)$。

## 3 实验对比及结果分析

### 3.1 实验数据集

本文分别在微博性别对立数据集和公开情感分析数据集Weibo_senti_100k上进行实验。其中,微博性别对立数据集采用爬虫技术自行收集,实验数据主要来源于2020年中共共青团发布的反对极端女权相关微博。再对其进行人工筛选、标注与核对,最终保留数据9 300条,包含4 618条正向文本、4 682条负向文本。其中,正向文本指的是不包含性别对立倾向的文本,负向文本指的是包含性别对立倾向的文本。

为了验证该方法的泛化能力,本文选用公开的情感分析数据集Weibo_senti_100k进行验证实验,数据集包含11万多条带情感标注的微博评论,其中正负向评论各约5万多条。为了规避掉验证集的选择偏差,在划分验证集时采用$k$折交叉验证的方法。其中,$k$取10,如图7所示。将数据集按照9:1的比例分为训练集和测试集,

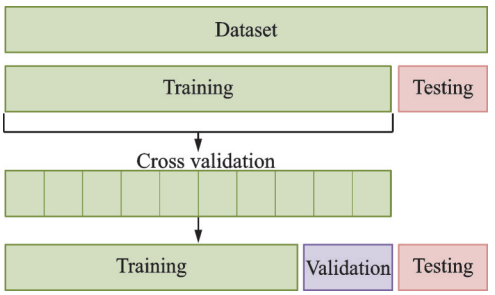


图7 交叉验证法

Fig.7 Cross-validation method

然后将训练集再分为10份(0～9),依次选择0～9作为验证集,从而得到10份不同的数据划分。最后,记录模型在10个不同验证集上的平均表现,从而解决验证集选择有偏差的问题。

### 3.2 实验配置

实验主要环境配置如下:(1)主要硬件环境为:CPU(12th Gen Intel Core i5-12400F six-core)、GPU(NVIDIA GeForce RTX 3060 12 GB);(2)主要软件环境为:编程语言(Python-3.7)、深度学习框架(Pytorch-1.11.0)、编程工具(JetBrains PyCharm)。

实验主要参数配置如下:

(1)BERT预训练模型:padsize设置为64,对于长度不足部分进行padding补全,超出部分进行cut剪切;BERT_embedding设置为768;Dropout率设置为0.5,隐藏层数量设置为500;学习率和衰减率均设置为1e－5,使用批正则化方式降低过拟合,优化器使用Adam;epoch设置为100,且若超过1 000个batch后模型的学习效果未提升,则提前结束训练;(2)Word2Vec模型:使用gensim函数库进行Word2Vec训练,由于数据量小,实验中将特征向量维度设置为100,窗口移动大小设置为5,sg设置为1;(3)Bi-GRU网络:隐藏层维度设置为500,激活函数选择默认的tanh函数。

### 3.3 评价指标

本文选用精确率$P$(Precision)、召回率$R$(Recall)以及$F_1$值作为模型的评价指标,有

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{13}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{14}$$

$$F_1=\frac{2*P*R}{P+R} \tag{15}$$

式中:TP表示真正例,即预测为正类实际结果也是正类;TN表示真反例,即预测为反类实际结果也是反类;FP表示假正例,即预测为正类实际结果为反类;FN表示假反例,即预测为反类实际结果为正类。

### 3.4 实验结果

本文以BERT＋Bi-GRU作为基线模型,通过对比实验,证明该方法的有效性。为了说明不同特征以及各特征的组合给模型带来的增益,设计了以下消融实验,特征组合介绍如表3所示,对应的实验结果如表4所示。由表4中可知,$F_1$值最低的为没有融合任何特征的实验1,而实验2～4相对于实验1在表现上均有一定的提升,证明了融合五笔、郑码以及拼音特征的有效性;实验5相对于实验1基线模型,在精确率$P$、召回率$R$以及$F_1$值上分别有1.12%、1.01%以及1.07%的提升,加入表情情感词典后有超过1%的提升,证明了加入表情情感词典的有效性;实验8相对于实验1基线模型,在精确率$P$、召回率$R$

**表3 对比实验的特征组合形式**

**Table 3 Feature combination forms for comparative experiments**

| 编号 | 特征组合 | 含义 |
|---|---|---|
| 1 | BERT＋Bi-GRU | 使用BERT＋Bi-GRU组合 |
| 2 | 五笔 | 在编号1的基础上加入五笔特征 |
| 3 | 郑码 | 在编号1的基础上加入郑码特征 |
| 4 | 拼音 | 在编号1的基础上加入拼音特征 |
| 5 | 表情情感词典 | 在编号1的基础上加入表情情感词典 |
| 6 | 五笔＋郑码 | 在编号2的基础上加入郑码特征 |
| 7 | 五笔＋郑码＋拼音 | 在编号6的基础上加入拼音特征 |
| 8 | 五笔＋郑码＋拼音＋表情情感词典 | 在编号7的基础上加上表情情感词典 |

**表4 性别对立数据集上的对比实验结果**

**Table 4 Comparative experimental results on gender-opposition dataset** %

| 编号 | 特征组合 | $P$ | $R$ | $F_1$ |
|---|---|---|---|---|
| 1 | BERT+Bi-GRU | 82.56 | 82.61 | 82.58 |
| 2 | 五笔 | 85.21 | 85.25 | 85.23 |
| 3 | 郑码 | 83.17 | 82.98 | 83.07 |
| 4 | 拼音 | 83.52 | 83.59 | 83.55 |
| 5 | 表情情感词典 | 83.68 | 83.62 | 83.65 |
| 6 | 五笔+郑码 | 86.90 | 86.97 | 86.93 |
| 7 | 五笔+郑码+拼音 | 87.21 | 87.18 | 87.19 |
| 8 | 五笔+郑码+拼音+表情情感词典 | 87.72(↑5.16) | 87.83(↑5.22) | 87.77(↑5.19) |

以及$F_1$值上分别有5.16%、5.22%以及5.19%的提升,这说明本文提出的性别对立识别方法可以有效地提高模型的性能。为了可以更清楚地看到各个组合的效果,用条形图对结果进行表示,结果如图8所示。由图8可知,实验8的结果明显优于其余7组实验。

为了使该方法更具有说服力,本文选用公开的情感分析数据集Weibo_senti_100k进行验证实验,结果如表5所示。从表中可以看到,实验1和实验5以及实验7和实验8结果相同,这是因为在Weibo_senti_100k数据集中,表情符号已经清洗干净,使得数据中不存在表情符号,因此,加入表情情感词典并不能提升实验的结果,但实验7是组合中效果最好的一个,证明本文提出的方法在情感分析数据集上进行实验仍然可以产生正向收益,体现了该方法的泛化性。

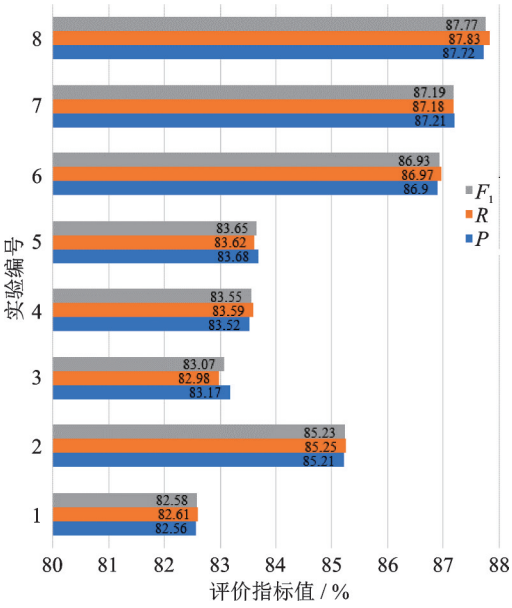


图8 实验结果图

Fig.8 Chart of experimental results

**表5 Weibo_senti_100k数据集上的对比实验结果**

**Table 5 Comparative experimental results on the Weibo_senti_100k dataset** %

| 编号 | 特征组合 | $P$ | $R$ | $F_1$ |
|---|---|---|---|---|
| 1 | BERT+Bi-GRU | 87.18 | 85.36 | 86.26 |
| 2 | 五笔 | 89.02 | 88.79 | 88.90 |
| 3 | 郑码 | 87.94 | 87.63 | 87.78 |
| 4 | 拼音 | 88.25 | 88.07 | 88.16 |
| 5 | 表情情感词典 | 87.18 | 85.36 | 86.26 |
| 6 | 五笔+郑码 | 89.56 | 89.22 | 89.39 |
| 7 | 五笔+郑码+拼音 | 90.21 | 89.93 | 90.07 |
| 8 | 五笔+郑码+拼音+表情情感词典 | **90.21** | **89.93** | **90.07** |

## 4 结束语

本文提出一种融合多特征和表情情感词典的性别对立言论识别方法,可以增强对性别对立言论的识别效果,为缓解网络中存在的性别对立现象提供了有力支持。该方法在BERT、Bi-GRU以及Soft-Max的协同下,融入了五笔、郑码、拼音这些特征,并在最后加入了表情情感词典,来提升模型的识别效果。实验结果表明,本文提出的性别对立言论识别方法具有良好的效果。未来工作的重点包括:(1)进一步扩充性别对立数据集;(2)细化表情情感词典并探索将其向量化的方法;(3)将该文的方法应用到实际应用当中。

**参考文献:**


[1] 郑晓雪, 刘理, 胡蝶,等. 微博暴力对合肥市大学生心理健康的影响[J]. 医学与社会, 2018, 31(9): 63-65,84.
ZHENG Xiaoxue, LIU Li, HU Die, et al. Influence of micro-blog's cyberbullying on mental health of college students in Hefei city[J]. Medicine and Society, 2018, 31(9): 63-65,84.

[2] 汪永涛. 新生代青年的婚恋实践及其影响因素分析[J]. 中国青年研究, 2021(12): 15.
WANG Yongtao. Analysis on marriage practice of the new generation youth and its influencing factors[J]. China Youth Study, 2021(12): 15.

[3] LI L, WANG X T. Nonverbal communication with emojis in social media: Dissociating hedonic intensity from frequency[J]. Language Resources and Evaluation, 2023, 57(1): 323-342.

[4] KRALJ N P, SMAILOVIC J, SLUBAN B, et al. Sentiment of emojis[J]. PloS One, 2015, 10(12): e0144296.

[5] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of NAACL-HLT. Minneapolis, Minnesota, USA: NAACL, 2019: 4171-4186.

[6] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL].(2013-01-16) [2024-05-19]. https://doi.org/10.48550/arXiv.1301.3781.

[7] CHO K, VAN M B, BAHDANAU D, et al. On the properties of neural machine translation: Encoder-decoder approaches [C]//Proceedings of 8th Workshop on Syntax, Semantics and Structure in Statistical Translation, SSST 2014. [S. l.]: Association for Computational Linguistics (ACL), 2014: 103-111.

[8] WARNER W, HIRSCHBERG J. Detecting hate speech on the world wide web[C]//Proceedings of the second workshop on Language in Social Media. USA: ACL, 2012: 19-26.

[9] GAO L, KUPPERSMITH A, HUANG R. Recognizing explicit and implicit hate speech using a weakly supervised two path bootstrapping approach[EB/OL]. (2017-10-20) [2024-05-19]. https://doi.org/10.48550/arXiv.1710.07394.

[10] BURNAP P, WILLIAMS M L. Us and them: Identifying cyber hate on Twitter across multiple protected characteristics[J]. EPJ Data Science, 2016, 5: 1-15.

[11] ZHANG Z, LUO L. Hate speech detection: A solved problem? The challenging case of long tail on twitter[J]. Semantic Web, 2019, 10(5): 925-945.

[12] FAN H, DU W, DAHOU A, et al. Social media toxicity classification using deep learning: Real-world application uk brexit[J]. Electronics, 2021, 10(11): 1332.

[13] ALOTAIBI M, ALOTAIBI B, RAZAQUE A. A multichannel deep learning framework for cyberbullying detection on social media[J]. Electronics, 2021, 10(21): 2664.

[14] JHA A, MAMIDI R. When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data[C]//Proceedings of the Second Workshop on NLP and Computational Social Science. Vancouver, Cannda: ACL, 2017: 7-16.

[15] KARLEKAR S, BANSAL M. Safecity: Understanding diverse forms of sexual harassment personal stories[EB/OL]. (2018-09-13)[2024-05-19].https://doi.org/10.48550/arXiv.1809.04739.

[16] YAN P, LI L, CHEN W, et al. Quantum-inspired density matrix encoder for sexual harassment personal stories classification [C]//Proceedings of 2019 IEEE International Conference on Intelligence and Security Informatics (ISI). [S.l.]: IEEE, 2019:

218-220.
[17] GARCIA-DIAZ J A, CANOVAS-GARCIA M, COLOMO-PALACIOS R, et al. Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings[J]. Future Generation Computer Systems, 2021, 114: 506-518.
[18] RODRIGUEZ-SANCHEZ F, CARRILLO-DE-ALBORNOZ J, PLAZA L. Automatic classification of sexism in social networks: An empirical study on twitter data[J]. IEEE Access, 2020, 8: 219563-219576.
[19] PITSILIS G K, RAMAMPIARO H, LANGSETH H. Effective hate-speech detection in Twitter data using recurrent neural networks[J]. Applied Intelligence, 2018, 48(12): 4730-4742.
[20] 刘宇瀚,刘常健,徐睿峰,等.结合字形特征与迭代学习的金融领域命名实体识别[J].中文信息学报, 2020, 34(11): 74-83.
LIU Yuhan, LIU Changjian, XU Ruifeng, et al. Utilizing glyph feature and iterative learning for named entity recognition in finance text[J]. Journal of Chinese Information Processing, 2020, 34(11): 74-83.
[21] 王铭涛,方晔玮,陈文亮.基于中文字形的ELMo在电商事件识别上的应用[J].中文信息学报, 2021, 35(12): 94-102.
WANG Mingtao, FANG Yewei, CHEN Wenliang. E-commerce event detection with Chinese character glyph based ELMo [J]. Journal of Chinese Information Processing, 2021, 35(12): 94-102.
[22] 米健霞, 谢红薇. 面向招标物料的命名实体识别研究及应用[J]. 计算机工程与应用, 2023, 59(2): 314-320.
MI Jianxia, XIE Hongwei. Research and application of named entity recognition for bidding materials[J]. Computer Engineering and Applications, 2023, 59(2): 314-320.
[23] 周昊,沈庆宏.基于改进音形码的中文敏感词检测算法[J].南京大学学报(自然科学),2020,56(2): 270-277.
ZHOU Hao, SHEN Qinghong. Chinese sensitive words detection algorithm based on improved sound-character code[J]. Journal of Nanjing University(Natural Sciences), 2020, 56(2): 270-277.
[24] 王艳,王胡燕,余本功.基于多特征融合的中文文本分类研究[J].数据分析与知识发现,2021,5(10): 1-14.
WANG Yan, WANG Huyan, YU Bengong. Chinese text classification with feature fusion[J]. Data Analysis and Knowledge Discovery, 2021, 5(10): 1-14.
[25] 李瀚臣,张顺香,朱广丽,等.基于拼音相似度的中文谐音新词发现方法[J].计算机应用,2023,43(9): 2715-2720.
LI Hanchen, ZHANG Shunxiang, ZHU Guangli, et al. Chinese homophonic neologism discovery method based on Pinyin similarity[J]. Journal of Computer Applications, 2023, 43(9): 2715-2750.

**作者简介:**

**马子晨**(1998-),男,硕士研究生,研究方向:Web挖掘、情感分析。

**张顺香**(1970-),**通信作者**,男,教授,博士生导师,研究方向:Web挖掘、语义搜索和复杂网络,E-mail: sxzhang@aust.edu.cn。

**刘云朵**(1995-),女,硕士研究生,研究方向:自然语言处理。

**朱广丽**(1971-),女,副教授,硕士生导师,研究方向:Web智能信息处理、文本挖掘。

(编辑:刘彦东)