

陈泽众, 纪明宇, 施屹然, 等. 基于双通道的颜文字文本情感分类模型构建[J]. 智能计算机与应用, 2024, 14(12): 60-66.
DOI: 10.20169/j.issn.2095-2163.24062501

基于双通道的颜文字文本情感分类模型构建

陈泽众¹, 纪明宇², 施屹然³, 何鑫², 于海博⁴, 李尚泽¹

(1 东北林业大学 奥林学院, 哈尔滨 150006; 2 东北林业大学 计算机与控制工程学院, 哈尔滨 150036;

3 苏州大学 计算机科学与技术学院, 江苏 苏州 215006; 4 哈尔滨职业技术大学 网红产业学院, 哈尔滨 150081)

摘要: 本文提出了一种改进的用于处理包含颜文字符号和文字混合文本的新的情感分类模型。模型综合包括多层感知机、统计情感缩放、预构建词典、双通道处理和自注意力机制在内的多种技术, 具有较高的可扩展性和可解释性。同现有的情感分类模型相比, 本模型能更好地感知颜文字对文本语义的影响, 并在此基础上实现文本情感分类。实验结果表明, 该模型在融合颜文字的文本情感分类任务中表现出更好的性能。此外, 模型可用于解决其他混合结构符号和文本的自然语言处理问题, 为情感分类领域的发展提供了新思路和新方法。

关键词: 深度学习; 结构性符号; 文本分类; 情感分类模型

中图分类号: TP391.1

文献标志码: A

文章编号: 2095-2163(2024)12-0060-07

A multi-channel sentiment classification model construction based on structural symbolic text

CHEN Zezhong¹, JI Mingyu², SHI Yiran³, HE Xin², YU Haibo⁴, LI Shangze¹

(1 Aulin College, Northeast Forestry University, Harbin 150006, China;

2 College of Computer and Control Engineering, Northeast Forestry University, Harbin 150036, China;

3 School of Computer Science and Technology, Soochow University, Suzhou 215006, Jiangsu, China;

4 Netflix Industry College, Harbin Vocational and Technical University, Harbin 150081, China)

Abstract: This paper proposes the new and improved sentiment classification model for processing mixed structural symbolic and textual texts containing face characters. The model integrates various techniques including multi-layer perceptron, statistical sentiment scaling, pre-constructed dictionary, dual-channel processing and self-attentive mechanism, with high scalability and interpretability. Compared with existing sentiment classification models, this model can better perceive the influence of face text on text semantics and achieve text sentiment classification on this basis. Experimental results show that the model shows better performance in text sentiment classification tasks incorporating face characters. In addition, the model can be used to solve other natural language processing problems with mixed structured symbols and text, providing new ideas and methods for the development of the field of sentiment classification.

Key words: deep learning; structure symbol; text classification; sentiment classification model

0 引言

社交网络信息的飞速增长给文本信息分类带来了挑战。本文旨在解决混合符号与文字的文本分类问题, 以挖掘隐藏的信息和情感表示^[1-3], 对舆论把控以及用户意图分析方面有着重要意义^[4]。在实际

应用场景中, 比如社交媒体和直播平台等, 存在大量混合符号与文字的文本信息, 这些符号是遵循一定组成规则组成的具象表达, 能够反映与影响整个文本信息的情感表示^[5], 整体来看, 这种混合符号与文字的文本信息存在着口语化、内容不纯粹等特点。传统的特征工程方法难以全面感知文本的整体

基金项目: 东北林业大学大学生创新创业训练计划项目(S202410225185); 黑龙江省重点研发计划项目(2022ZX01A35)。

作者简介: 陈泽众(2002—), 男, 本科生, 主要研究方向: 自然语言处理。Email: 1241550141@qq.com; 纪明宇(1980—), 男, 博士, 副教授, 主要研究方向: 自然语言处理, 系统仿真建模; 施屹然(2003—), 男, 本科生, 主要研究方向: 多模态信息处理; 何鑫(1998—), 男, 硕士研究生, 主要研究方向: 自然语言处理; 于海博(1983—), 男, 讲师, 主要研究方向: 直播工程; 李尚泽(2004—), 男, 本科生, 主要研究方向: 自然语言处理。

收稿日期: 2024-06-25

情感^[6-8],因此本文提出一种基于符号文本特征分离、双通道处理的方法。

颜文字是由特定字符编排组合形成的表情符号,其核心在于外化情绪与表情的呈现。其中,颜文字符号更具有代表性,因其使用简单,不需要调用额外的表情系统而获得大量应用。在颜文字与普通文字混合的文本中,如果只是对文字特征进行提取,则模型会忽略颜文字对文本的影响,容易导致分类结果与作者的情感产生偏差。比如“你可真棒”本身表达正向的情感,但加上颜文字符号后,如“你可真棒(-_-|||)”,使得文本整体呈现负向情感。同样文字信息也可以影响颜文字信息的表达,有“(^v^)”表示微笑,其本身表达出正向的情感,但是在特定的情境下,比如在“放弃吧(^v^)”语句中,微笑表情具有嘲讽意思,使得整个文本呈现负向的情感。因此这些符号信息受语境与背景等方面的影响,其整体语义具有复杂多变的特点。

针对这种情况,本文提出符号文本特征分离、双通道处理的方式。具体来说,模型基于 MLP 和统计缩放的形式单独构建符号向量词典,将结构性符号信息转化为向量表示。同时,使用 BERT 预训练模型处理普通文字,将其转化为向量表示。然后,将不同通道输出的特征使用自注意力机制进行特征融合,增强结构性符号的特征表示能力。最后,将融合后的特征向量输入到基于双向简单循环网络(BiSRU)构建的情感分类网络中,输出文本情感分类结果。

1 相关工作

在短文本分类中,为提升短文本分类的准确率,Bao 等学者^[9]结合 BERT、CNN 和 BiGRU 对短文本特征进行捕捉和融合。Yu 等学者^[10]使用并行的 LSTM 网络和 GRU 网络对文本语义进行捕捉和分类。Li 等学者^[11]使用狄利克雷分布(LDA)的主题模型,调整并使用主题向量空间对短文本进行分类。Zhou 等学者^[12]采用混合注意力网络,使用 RNN 和 CNN 进行文本特征提取并融合注意力机制对文本进行分类。陈亚茹等学者^[13]对短文本词嵌入阶段加入噪声进行训练,提升模型特征提取效果,再使用 BiLSTM 网络结合注意力机制对短文本进行分类。

对于表情符号文本的处理,Kumar 等学者^[14]使用带自注意力层的噪声正则化双向门控循环单元(Bi-GRU)对含有表情符号和文本进行分类。邱全磊等学者^[15]通过构建情感词典对文字和颜文字的情感值进行累加。赵晓芳等学者^[16]使用表情符号

词典,将文本信息进行拼接,使用 CNN 和 LSTM 网络对文本进行分类。李楠等学者^[17]提出表情稳定度的概念,基于概率统计计算表情符号在不同语境下的情感稳定性。龚路^[18]使用 Word2Vec 的 Skip-gram 方法对表情符号进行词嵌入训练。以上模型能够对颜文字信息的特征进行提取,但是通过直接训练的方式得到的颜文字特征未考虑颜文字与文字、颜文字与数据集主题之间的相互影响。Xu 等学者^[19]考虑到了表情符号对文本情感分类结构造成的影响,提出使用 E-BiLSTM 模型来处理混合表情符号信息的文本数据,并结合了 Emoji2Vec 模型与 Glove 模型的计算过程,增强了表情符号嵌入层表征,优化表情符号特征的提取效果,同时使用双向长短期记忆网络对文字特征进行提取,继而将两者特征进行融合,有效实现文本情感分类,具有相对良好的效果,该模型也是基于类似使用双通道训练的思想,将表情符号和文字分开训练再融合,提升文本分类的效果。Zahra 等学者^[20]的核心思想是基于语料库与特定的权值计算公式,对表情符号进行加权,从而感知到表情符号的情感倾向,再将加权后的表情符号与文本特征输入至分类器中,最终实现融合表情符号的文本情感分类。

2 双融合训练模型

为使模型能够区分结构性符号和普通文字,实现融合符号的文本情感分类,本文设计先构建结构性符号情感词典,再根据该词典对数据集中的文本进行特征分离处理、融合,接着将融合后的特征信息传入 BiLSTM 网络中实现文本的情感分类。具体结构如图 1 所示。

模型主要分为 4 个阶段,分别为文本分离、特征提取、特征融合与文本分类。具体来说,在文本分离阶段,将弹幕文本运用基于字符串匹配的方法来判断当前文本中是否存在颜文字符号信息,如果存在则使用字符串截取的方式,对颜文字进行剥离,从而产生普通文字与颜文字两个通道的内容,为后续的特征提取阶段提供数据支持。

(1)特征提取阶段。对于文字信息,采用 BERT 模型将文字信息转换为向量表示,从而实现文字上下文的全面理解。对于颜文字信息,使用提前训练好的颜文字向量词典进行匹配,将匹配到的向量与文字向量进行拼接。

(2)特征融合阶段。旨在将上个阶段中得到的文本与颜文字的特征向量进行融合,其内部使用了

注意力算法,将文本特征与颜文字特征进行加权平均与维度对齐,以提升整体的特征融合效果。

(3)文本分类阶段。模型基于 BiLSTM 网络构建了文本分类器。该分类器接收融合后的特征向量

作为输入,通过 BiLSTM 网络进行高效计算与处理。

最终,利用映射函数对这些处理后的特征进行分类,输出最终的分类结果,从而实现对文本内容的情感分类。

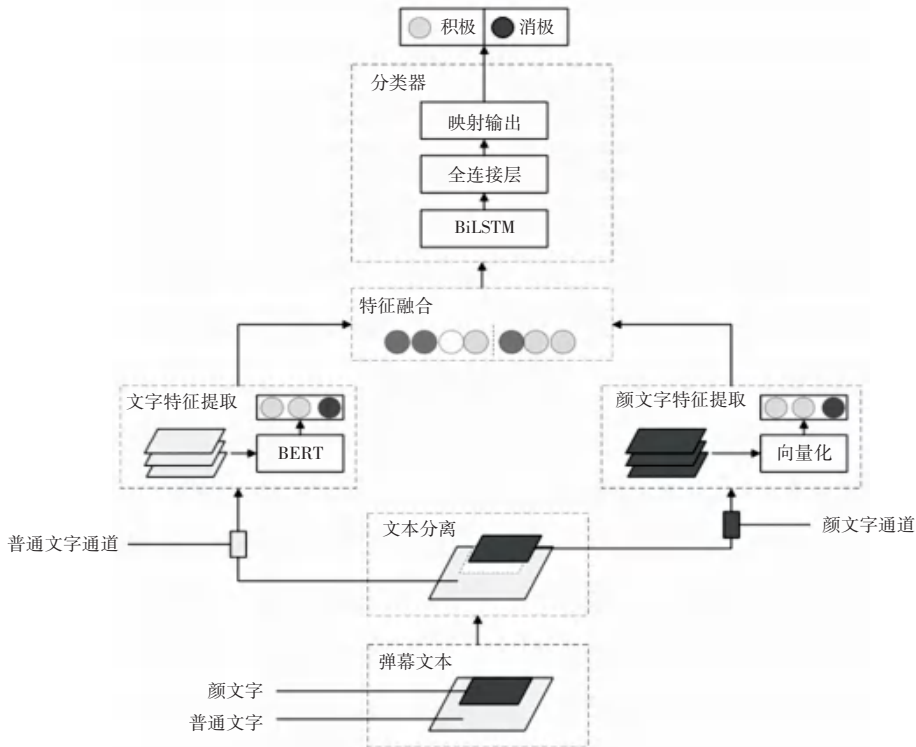


图1 模型结构图

Fig. 1 Model structure diagram

2.1 文本分离

文本分离阶段的主要内容是将文本中的颜文字与普通文字进行有效分离,并将不同类别的特征分别输送到各自的处理通道中。这一阶段是整个模型的前置驱动,为后续的数据处理奠定了基础。

本文模型基于数据集遍历和网络搜索方法来对颜文字内容进行系统化收集,并将所收集的颜文字纳入一个专门的集合中。随后,依据该集合内的颜文字信息,对当前样本执行静态匹配过程。一旦匹配成功,便将对应的颜文字单独提取并分离出来,进而将其导向颜文字专门的处理通道。在颜文字被成功提取后,剩余部分自然划分为普通文字信息。颜文字和普通文本分离过程如图2所示。

文本分离的核心逻辑在于集合运算,通过识别当前文本与颜文字集合之间的交集,以判断文本中是否包含颜文字。通过执行原文本与颜文字集合的交集运算,即可提取文本中的颜文字信息。接着,通过执行原文本与颜文字集合的差集运算,即可得到只包含普通文字的文本。最后,将颜文字与普通文

字进行分别输出,形成双通道,为特征提取提供准备工作。

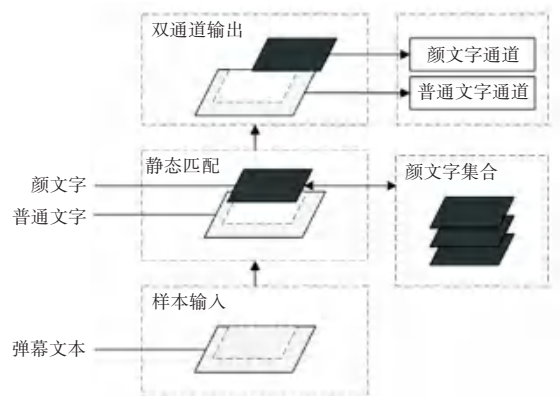


图2 文本分离示意图

Fig. 2 Text separation diagram

2.2 特征提取

对于颜文字通道的颜文字文本信息,本文模型采用 Skip-gram 模型对颜文字进行特征向量的训练。鉴于颜文字在样本空间中可能存在的语义不确定性,本研究创新性地提出了情感缩放因子。该因

子能够统计并计算颜文字在样本空间中的语义倾向,从而有效减少颜文字潜在的语义干扰问题。另一方面,对于普通文字通道,系统基于 BERT 预训练模型进行特征向量的训练,BERT 能够深层次地理解文本语义,可以显著增强文本特征提取的效果。以下将对颜文字和普通文字两个通道的特征提取过程进行详细介绍。

(1) 基于改进 Word2Vec 的颜文字向量训练。Word2Vec 模型将词汇映射为固定长度的向量,旨在捕捉词汇的语义关系和上下文信息。通过统计词频和上下文词的关系,Word2Vec 学习词汇的分布式表示,并将其映射为高维空间中的向量。颜文字特征训练过程如图 3 所示。

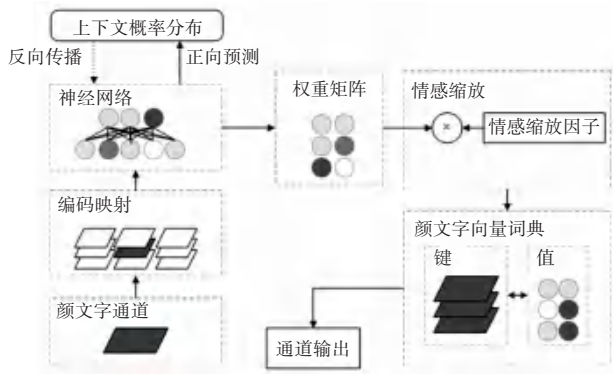


图 3 颜文字特征训练结构图

Fig. 3 Face character training diagram

为了进一步丰富颜文字的语义信息,本文模型引入了情感缩放因子,该因子能够反映颜文字在情感表达上的强度和倾向。将颜文字向量表示与情感缩放因子相乘,得到颜文字的向量表示。相对来说,该向量能更好反映当前样本空间中颜文字语义和情感特性。颜文字向量计算公式具体如下:

$$P(\text{context}(w) | w) = \prod_{c \in \text{context}(w)} P(c | w) \quad (1)$$

$$\delta = \frac{p}{p+n} + 0.5 \quad (2)$$

$$V_{\text{sign}} = W \times \delta \quad (3)$$

其中, p 表示当前颜文字在积极情感样本中的出现次数; n 表示其在消极情感样本中的出现次数; δ 表示情感因子; W 表示颜文字嵌入向量。该计算公式中的分式表示了颜文字在积极语句中的占比,其值域限定在 $[0, 1]$ 。通过增加 0.5, 该占比的值域扩展至 $[0.5, 1.5]$ 。因此,当颜文字在积极样本中出现频率较高时,其情感缩放因子的值域会变为 $(1, 1.5]$, 反之,在消极样本中频繁出现的颜文字,其情感缩放因子的值域则为 $(0.5, 1]$ 。以此得到能

够反映情感占比的因子 δ , 并将其与颜文字的初步向量相乘,以获得最终的颜文字向量表示 V_{sign} 。通过这种计算方法,对积极占比较大的颜文字向量进行放大,对消极占比较大的颜文字进行削弱,使得颜文字向量在样本空间具有有更好的特征提取效果。

(2) 基于 BERT 的文字特征提取。将普通文字文本输入到文字处理通道中,即由 BERT 对其特征向量进行训练。BERT 是由 Google AI 在 2018 年提出的预训练自然语言处理模型。BERT 是一种基于 Transformer 架构的深度学习的模型,旨在为自然语言处理任务提供高质量的特征表示。BERT 采用了 Transformer 的编码器作为其基本架构,通过预先训练, BERT 能够理解词汇之间的复杂关系和上下文依赖,从而为文本分类任务提供更丰富的语义表示。基于 BERT 的文字向量计算过程如图 4 所示。

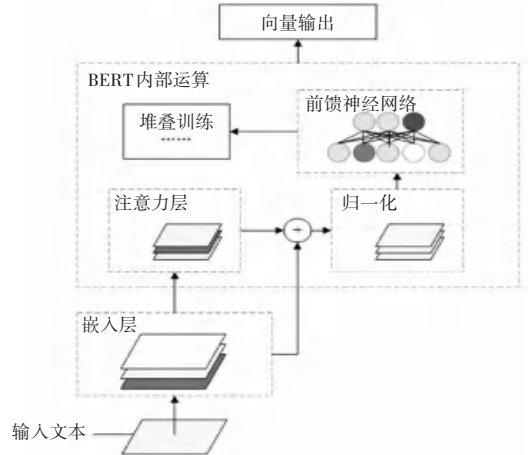


图 4 普通文字特征训练过程图

Fig. 4 General word training diagram

由此,文字通道能够成功提取出普通文字的向量特征。整个特征提取阶段的核心内容,就是基于上述方法对不同通道的文本进行特征向量的提取和转化。每个通道专注于自身内部的特征训练,旨在挖掘文本内部的特定信息。这一阶段的输出为后续的文本情感分类任务提供坚实的基础,使模型能够理解和表达文本内在的表征含义。

2.3 特征融合

在双通道成功提取不同文本的特征向量之后,为了确保与后续情感分类模块的有效对接,需要对这些特征向量进行统一的整合。直接将这 2 个具有差异性的特征向量进行拼接可能会引发一些问题,例如维度不匹配和信息冗余等,这些问题可能会对文本情感分类的效果产生不利影响,因此,本文采用了基于自注意力算法的特征融合方法,来实现对 2 个不同通道特征的更有效整合。特征融合过程如图

5所示。

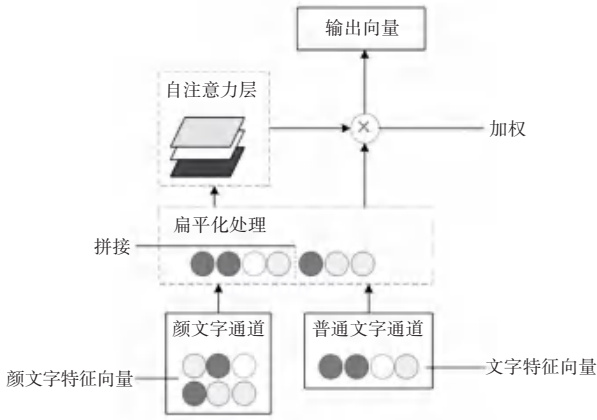


图5 特征融合过程图

Fig. 5 Feature fusion process diagram

通过上述自注意力算法的融合机制,可以成功获取到融合后的特征向量。相较于之前简单拼接的融合方法,这种特征向量展现出显著的优势。通过注意力算法的内部转化增强了特征向量的关键部分,提升了2个不同序列特征的融合效果。2个不同通道的特征融合计算的过程可用如下公式来表示:

$$\mathbf{V}_{\text{full}} = [\mathbf{V}_{\text{sign}}, \mathbf{V}_{\text{text}}] \quad (4)$$

$$\mathbf{Q} = \mathbf{V}_{\text{full}} \mathbf{W}_Q, \mathbf{K} = \mathbf{V}_{\text{full}} \mathbf{W}_K, \mathbf{V} = \mathbf{V}_{\text{full}} \mathbf{W}_V \quad (5)$$

$$A_{ij} = \frac{\mathbf{Q}_i \cdot \mathbf{K}_j^T}{\sqrt{d}} \quad (6)$$

$$\alpha_{ij} = \text{Softmax}(A_{ij}) = \frac{\exp(A_{ij})}{\sum_{k=1}^n \exp(A_{ik})} \quad (7)$$

$$\mathbf{Z}_{ij} = \sum_{j=1}^n \alpha_{ij} \mathbf{V}_j \quad (8)$$

通过注意力算法的内部转化增强了特征向量的关键部分,提升了2个不同序列特征的融合效果,这种方法为后续的情感分类任务提供了更优的文本特征表示,从而有助于提高模型的性能和预测准确性。

2.4 分类输出

本文分类基于 BiLSTM 进行搭建。LSTM 网络相较于其他循环模型,如 RNN 和 GRU,引入了更多的参数和门控机制,能够有效地捕捉序列数据上的依赖关系,具有相对更强的记忆和泛化能力,而 BiLSTM 将序列的前序和后序同时进行训练,进一步加强了网络对序列特征的理解能力。分类器结构如图6所示。

本文模型通过这种设计,较为全面地捕获到整体的结构和语义信息,从而能够更细致地理解文本的特征表征,实现情感分类任务。

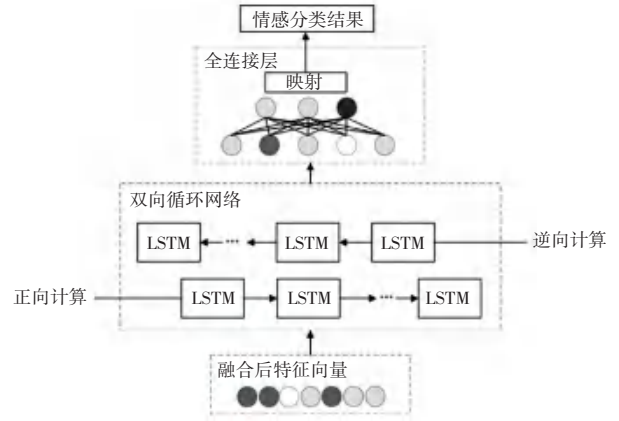


图6 分类器结构图

Fig. 6 Classifier structure diagram

3 仿真结果实验

实验部分基于 Pytorch 框架开发,开发工具为 Jupyter notebook,开发语言为 Python,使用 GPU 训练模型。

本文模型旨在应用于弹幕文本情感分类任务。数据样本采用2种方式获取:一部分来自 THUCNews 数据集。该数据集是清华大学自然语言处理实验室公开的中文文本二分类数据集,其内容基于新浪新闻 RSS 订阅频道的历史数据筛选过滤生成;另一部分为 B 站视频弹幕文本,主要内容是通过 Python 和 Request 库爬取 B 站视频的弹幕信息。2部分数据进行拼接合并,形成模型的训练和测试数据集。数据集总共 12 761 条样本,通过数据处理标注后共分为 6 405 条积极弹幕样本和 6 356 条消极弹幕样本。为了使模型的训练更贴近弹幕与颜文字的背景,整体数据集中约包括混合颜文字文本的数据样本 3 000 条,较为均匀地分布在积极弹幕和消极弹幕中。

3.1 对比实验

对比模型选取了部分机器学习与深度学习模型,调整不同结构计算其相应的指标进行对比。相关模型在当前数据集对比结果见表1。

表1 对比实验结果表

Table 1 Comparative experimental results table				%
序号	模型	Precision	Recall	F1
1	TFIDF-SVM	75.2	79.4	77.3
2	BERT	80.4	83.6	81.9
3	(单通道)BERT-SVM	81.8	84.5	83.2
4	(双通道)颜文字词典-BERT-SVM	82.9	83.7	83.3
5	(单通道)BERT-BiGRU	84.4	86.2	85.3
6	(单通道)BERT-BiLSTM	85.4	86.8	86.1
7	(双通道)颜文字词典-BERT-BiGRU	87.0	86.4	86.7
8	(双通道)本文模型	88.2	87.9	88.0

通过对比发现,本文模型比 TFIDF-SVM 模型的精确率、召回率和 $F1$ 值分别高出了 13.0%、8.5% 以及 10.7%, 相差幅度相对较大, 说明在该数据集中, 传统机器学习方法对文本特征情感分类能力以及颜文字信息捕捉能力存在一定的局限性, 深度学习技术更为适用。本文模型比 BERT 模型的精确率、召回率和 $F1$ 值分别高出了 7.8%、4.3% 以及 6.1%, BERT 模型是本文模型结构中的部分环节, 因此单独的 BERT 模型在实验表现中性能低于本文模型, 将本文模型和 BERT 模型进行对比可以用于检验模型结构堆叠的有效性。本文模型比 BERT-SVM 模型的精确率、召回率和 $F1$ 值分别高出了 6.4%、3.4% 以及 4.8%, 可以发现 BERT-SVM 模型相较于 BERT 模型的性能上有所提升, 可知在当前环境中, BERT 模型可以为下游提供较好的向量表征。本文模型比颜文字词典-BERT-SVM 模型的精确率、召回率和 $F1$ 值分别高出了 5.3%、4.2% 以及 4.7%, 该项指标可以发现在 BERT-SVM 的基础上添加颜文字词典形成双通道后, 与本文模型的指标值差距变小, 性能得到了进一步的提升。本文模型比 BERT-BiGRU 模型的精确率、召回率和 $F1$ 值分别高出了 3.8%、1.7% 以及 2.7%, BERT-BiGRU 模型的下游分类器与本文模型方向一致, 同样采用了深度学习的方式进行构建, 性能得到了进一步的提升, 由于其仅使用一个通道, 考虑颜文字带来的干扰信息, 因此在性能表现上与本文模型存在一定的差别。本文模型比 BERT-BiLSTM 模型的精确率、召回率和 $F1$ 值分别高出了 2.8%、1.1% 以及 1.9%, BERT-BiLSTM 模型同样是单通道的形式, 说明在本次实验中, 实验 BiLSTM 的性能表现会稍优于 BiGRU 模型。本文模型比颜文字词典-BERT-BiGRU 模型的精确率、召回率和 $F1$ 值分别高出 1.2%、1.5% 以及 1.3%, 序号 7 模型采用双通道的形式, 与之前的结论类似, 本次实验中, 下游分类器使用 BiLSTM 模型的效果表现会稍优于 BiGRU 模型。

3.2 消融实验

消融实验将对原模型的情感缩放因子、通道以及注意力算法部分进行消融, 因此本实验将在原模型的基础上划分为 4 个不同的子结构模型, 分别由序号 n 来表示模型。模型消融实验结果如图 7 所示。

序号 1 结构, 是在本文原模型基础上取消了颜文字向量词典, 使用 BERT 直接训练文本向量进行

分类。本次实验结果显示, 模型分类效果出现了下降, 验证了颜文字向量词典的有效性; 序号 2 结构, 是在颜文字向量词典构建中去掉了情感缩放因子, 可以发现情感因子具有一定作用, 可以提升颜文字向量的训练效果; 序号 3 的结构是在特征融合阶段取消了自注意力算法, 可以发现模型的效果大幅下降, 验证了特征融合时自注意力算法的有效性。

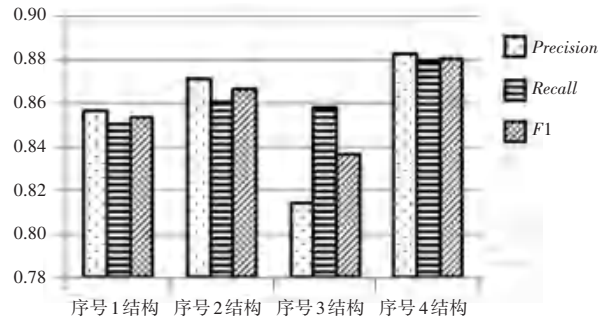


图 7 消融实验结果图

Fig. 7 Ablation experiment results diagram

对于序号 1 模型和序号 2 模型, 与本文的原模型(序号 4 模型)相比, 性能下降相对来说并不明显, 这表明颜文字向量词典虽然对模型分类效果的提升有一定的贡献, 但其相较于其他模块影响较小; 且序号 1 模型与序号 2 模型之间, 两者在性能上的差距并不显著, 主要表现在精确率上的差异, 这一发现表明, 情感缩放因子对模型识别正样本的能力有更多影响, 但其效果仍有提升的空间; 序号 3 的性能大幅下降, 这表明自注意力算法在特征融合阶段对模型的整体分类效果具有显著影响。因此说明自注意力算法在此阶段的应用是一种有效的策略。

4 结束语

本文提出了一种新的双通道处理方式, 能够更全面地感知混合符号与文字的文本信息, 实现更准确的情感分类。该方法通过符号向量词典的预构建、MLP 和统计缩放的形式对结构性符号信息进行处理, 并使用 BERT 预训练模型处理普通文字, 最终使用自注意力机制进行特征融合, 大幅提升了模型的性能。本文通过仿真对比实验, 充分证明了该模型的有效性。相较于其他常见的文本分类模型, 该模型在融合颜文字符号的情感分类任务中取得了更好的效果, 这一发现有助于更好地挖掘社交网络中隐藏的情感信息。

随着社交网络与弹幕文化社区的不断发展, 混合颜文字与普通文字的文本信息将越来越多地出现。该方法为处理这类文本信息提供了一种有效的

方式,可应用于电商评论、在线游戏聊天室、营销直播弹幕分析、信息推荐等领域,同时可以结合其他领域的知识和技术,如图像处理 and 声音识别等,探索更加全面和深入的情感分类方法,提高情感分类的准确度和可信度。

参考文献

- [1] LIU S, WANG G, LI Y, et al. A review of Chinese short text classification techniques [J]. *Journal of Information Engineering University*, 2021, 22(3): 304-312.
- [2] JIA Pengtao. A survey of text classification based on deep learning [J]. *Computer and Modernization*, 2021(7): 29-37.
- [3] 张明书, 葛晓义, 魏彬. 基于情感特征和谣言种类的谣言检测 [J]. *计算机仿真*, 2023, 40(6): 21-25.
- [4] 张苑, 祝小兰, 杨东晓. 基于深度学习的疫情情感分析 [J]. *智能计算机与应用*, 2022, 12(3): 40-45.
- [5] 陈俊, 李佳敏, 朱丽佳, 等. 多类表情符号短文本情感分析模型研究 [J]. *计算机仿真*, 2024, 41(8): 292-295.
- [6] TANG Y, HEW K F. Emoticon, Emoji, and sticker use in computer-mediated communication: a review of theories and research findings [J]. *International Journal of Communication*, 2019, 13: 2457-2483.
- [7] PEINELT N, NGUYEN D, LIAKATA M. tBERT: Topic models and BERT joining forces for semantic similarity detection [C]// *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 2020: 7047-7055.
- [8] CHAIMA M, ZAHIA G, BEN L R. A deep learning model for opinion mining in Twitter combining text and emojis [J]. *Procedia Computer Science*, 2022, 207: 2628-2637.
- [9] BAO Tong, REN Ni, LUO Ruiwang, et al. A BERT-Based hybrid short text classification model incorporating CNN and attention-based BiGRU [J]. *Journal of Organizational and End User Computing*, 2021, 33(6): 1-21.
- [10] YU Shujuan, LIU Danlei, ZHU Wenfeng, et al. Attention-based LSTM, GRU and CNN for short text classification [J]. *Journal of Intelligent & Fuzzy Systems*, 2020, 39(1): 1-8.
- [11] LI Yangyang, LIU Bo. A new vector representation of short texts for classification [J]. *The International Arab Journal of Information Technology*, 2020, 17(2): 241-249.
- [12] ZHOU Yujun, XU Jiaming, CAO Jie, et al. Hybrid attention networks for Chinese short text classification [J]. *Computación y Sistemas*, 2018, 21(4): 759-769.
- [13] 陈亚茹, 陈世平. 融合自注意力机制和 BiGRU 网络的微博情感分析模型 [J]. *小型微型计算机系统*, 2020, 41(8): 1590-1595.
- [14] MOHAN K A V, NANDAKUMAR A N. Noise-regularized bidirectional gated recurrent unit with self-attention layer for text and emoticon classification [J]. *International Journal of E-Collaboration*, 2022, 18(1): 102-123.
- [15] 邱全磊, 崔宗敏, 喻静. 基于表情和语气的情感词典用于弹幕情感分析 [J]. *计算机技术与发展*, 2020, 30(8): 178-182.
- [16] 赵晓芳, 金志刚. 融合表情符号与短文本的微博多维情感分类 [J]. *哈尔滨工业大学学报*, 2020, 52(5): 113-120.
- [17] 李楠, 张羽卉. 融合表情符号动态特征的舆情分析研究 [J]. *现代情报*, 2021, 41(8): 98-108.
- [18] 龚路. 融合表情符号的中文短文本情感分析研究 [D]. 重庆: 重庆大学, 2021.
- [19] XU A Q, JAYNE C, CHANG V. An emoji feature-incorporated multi-view deep learning for explainable sentiment classification of social media reviews [J]. *Technological Forecasting Social Change*, 2024, 202: 123326.
- [20] ZAHRA A, AKMAR M I. A multi-label emoji classification method using balanced pointwise mutual information-based feature selection [J]. *Computer Speech Language*, 2022, 73(C): 101330.