基于位置和词性特征的藏文情感三元组抽取模型

斯曲卓嘎[1,2,3],拥　措[1,2,3*],赛鸣宇[1,2,3]
(1. 西藏大学 信息科学技术学院,西藏 拉萨 850000;
2. 西藏自治区藏文信息技术人工智能重点实验室,西藏 拉萨 850000;
3. 藏文信息技术教育部工程研究中心,西藏 拉萨 850000)

摘　要:藏文情感三元组(方面词、情感词、情感极性)是细粒度情感分析的核心任务,对于深入理解藏族情感表达和趋势至关重要。但藏文的独特语言结构和文化背景导致其情感表达方式与其他语言不同,从而增加了细粒度情感分析的复杂性。为了提高藏文情感三元组的提取能力,该文提出了 OpinionNet-OTE-MTL 模型,该模型融合了词性信息、Word2Vec 词向量和绝对位置向量,并通过双向长短时记忆网络(BILSTM)进行特征提取。其中,由于藏文词性种类较多,该文分析了大量的情感数据集并从中提取出 11 种词性辅助模型识别。最后,为了验证 OpinionNet-OTE-MTL 模型的有效性,在自构建的 2 000 句藏文细粒度情感分析数据上进行了对比实验和消融实验。消融实验表明,词性较位置信息对模型的影响更大,其三元组抽取 F1 值提高了 3.06 百分点;对比实验结果表明将词性和位置特征融入进模型后,在情感三元组提取(Triple)任务上的精确率、召回率和 F1 值较基线实验提高了 4.73 百分点、6 百分点、6.14 百分点,融入词性和绝对位置信息使模型能更精确地理解藏文的语法结构和语义规则,从而提升了情感三元组分类任务的准确度。

关键词:藏文;Word2Vec;词性;位置特征;情感三元组
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2025)02-0130-08
doi:10.20165/j.cnki.ISSN1673-629X.2024.0300

Tibetan Sentiment Triplet Extraction Model Based on Location and Part-of-Speech Features

SI Qu-zhuo-ga[1,2,3], YONG Tso[1,2,3*], SAI Ming-yu[1,2,3]
(1. School of Information Science and Technology, Tibet University, Lhasa 850000, China;
2. Key Laboratory of Tibetan Information Technology and Artificial Intelligence in Tibet Autonomous Region, Lhasa 850000, China;
3. Engineering Research Center of Tibetan Information Technology, Ministry of Education, Lhasa 850000, China)

Abstract: The extraction of Tibetan emotional triplets (aspect words, emotional words, emotional polarity) is a core task in fine-grained sentiment analysis, which is crucial for a deep understanding of Tibetan emotional expressions and trends. However, the unique language structure and cultural background of Tibetan make its emotional expression different from other languages, thereby increasing the complexity of fine-grained sentiment analysis. To enhance the capability of extracting Tibetan emotional triplets, we propose the OpinionNet-OTE-MTL model, which integrates part-of-speech information, Word2Vec word vectors, and absolute position vectors. Feature extraction is performed through Bidirectional Long Short-Term Memory networks (BiLSTM). Given the diverse types of parts of speech in Tibetan, we analyze a large amount of emotional datasets and extract 11 types of parts of speech to assist in model recognition. Finally, to validate the effectiveness of the OpinionNet-OTE-MTL model, comparative experiments and ablation experiments were conducted on a self-constructed dataset of 2 000 Tibetan sentences for fine-grained sentiment analysis. The ablation experiments indicated that part-of-speech information had a greater influence on the model than positional information, resulting in a 3.06 percentage points increase in the F1 score of triplet extraction. Comparative experimental results showed that after integrating part-of-speech and positional features into the model, the precision, recall, and F1 score of the emotional triplet extraction task increased by 4.73 percentage points, 6 percentage points, and 6.14 percentage points respectively compared to the baseline experiment. Integrating part-

收稿日期:2024-06-14　　**修回日期**:2024-10-16
基金项目:西藏自治区科技计划项目(XZ202401JD0010)
作者简介:斯曲卓嘎(1997-),女(藏),硕士研究生,研究方向为自然语言处理;通信作者:拥　措(1974-),女(藏),教授,博导,研究方向为自然语言处理、藏医药古籍智能信息处理。

of-speech and absolute positional information enables the model to better understand the grammatical structure and semantic rules of Tibetan, thereby improving the accuracy of emotional triplet classification tasks.

Key words: Tibetan; Word2Vec; part of speech; positional features; emotional triplet

0 引　言

随着藏文信息的快速增长,众多藏文文本在互联网平台上变得日益丰富,对其进行情感分析有助于深入理解藏族的传统文化。张俊等研究者[1]使用了基于情感词典的方法来识别情感倾向。与此同时,袁斌等研究者[2]利用藏文的语法和语义特点,建立了一个特征空间进行情感分析。另外,孙本旺等研究者[3]运用深度学习技术,通过 Word2Vec 创建词向量模型,并结合 CNN 和 LSTM 模型对藏文微博进行了情感分析,取得了良好的效果。

藏文情感分析虽然取得了初步进展,但目前的研究多集中在粗粒度情感分类的任务上,对于细粒度的情感分析尚未完全实现。而情感三元组抽取作为情感细粒度分析中最重要的任务之一,它通过识别文本中的方面词、情感词和情感极性,为藏文情感分析提供了更为详尽和精确的信息。不仅有助于准确识别藏文文本讨论的具体方面、相关联的观点以及相应的情感倾向,而且对于理解藏族文化中的情感表达模式和细微差别具有重要意义。然而,与中英文情感分析相比,藏文细粒度情感分析的研究起步较晚,面临诸多挑战。首先,目前缺乏足够规模的标注藏文细粒度数据集,严重限制了情感分析研究的深入和模型训练的有效性。其次,藏语文本的复杂句法结构、多样的修辞手法和丰富的词性,为情感分析带来了额外的难题。

因此,针对上述问题,该文做了以下工作。首先,针对目前没有藏文细粒度情感分析语料导致该研究空白的情况,构建了 2 000 句藏文细粒度情感语料,为后续研究提供了基础数据支持;其次,为了提高词嵌入捕捉藏文词义的丰富信息,使用了 430 MB 数据训练了 Word2Vec 词向量模型;再次,针对 OTE-MTL 模型识别藏文情感三元组效果不佳的情况,提出了融合词性、Word2Vec 向量及绝对位置向量的 OpinionNet-OTE-MTL 模型,增加模型对藏文情感三元组的识别和提取能力;最后,在该文构建的语料上使用 OpinionNet-OTE-MTL 模型进行了对比及消融实验,证明了词性以及位置编码对藏文情感细粒度分析具有良好效果。

1 相关研究

随着自然语言处理技术的进步,藏文情感分析得到了显著推动,为低资源语言的文本分析提供了新的视角和技术支撑。李海刚和于洪志[4]最早开发了基于情感词库的文本分类系统,使用相似度算法进行情感判断。2018 年拥措教授和史晓东[5]总结了藏文短文本研究方法,并分析了情感分析的趋势。公保加羊等人[6]提出了基于集成学习的藏文情感分析算法,结合现有藏文文本情感分析算法的优势,提升藏文情感分析模型的精度和效率。同年,朱宇雷等人[7]构建了大规模的情感分析数据集,并采用 Graph-SAGE 网络提取特征。孟祥和等人[8]则利用多核卷积神经网络和 BiLSTM (Bidirectional Long Short - Term Memory network)模型提取文本特征。这些方法都有效提升了情感分类的准确度,但均在粗粒度情感分类上进行研究,并未涉及细粒度分析,而细粒度分析对于深入理解和挖掘藏文文本中的情感信息具有重要作用。

在自然语言处理(NLP)领域,词性和位置特征对于提升情感分析的准确率起着至关重要的作用。在 2017 年,Wang Wenya 等人提出了 CMLA[9]模型,该模型利用多层注意力机制共同提取文本中的方面项和观点项。2018 年,Li Xin 团队提出了 HAST[10]模型,通过融合多种技术手段提升了方面项提取的准确性;同一时期,赵富团队[11]提出了 CWPAT-Bi-LSTM 模型,通过结合词性和注意力机制,进一步提高了情感分析的准确率。2019 年,姚艳秋[12]采用了 LS-SO 算法,考虑情感词的上下文信息来提高分类的准确性;同时,王行甫[13]等进一步优化了细粒度情感分析,开发了一种结合词性、位置和单词情感的内存网络模型,并引入了 POSP-CNNAM 机制,显著提升了情感分析的准确性。到了 2020 年,Lim 团队[14]结合文本和地理位置信息,运用 CNN 和 BiLSTM 网络,成功提升了 Twitter 情感分析的准确度。2021 年 Xu Lu 等人[15]提出新的位置感知标记方案和 JET 模型,该模型使用特征因式分解来捕获三元组元素间的交互作用;同年,薛芳团队[16]通过双层词性感知技术,强化了情感分析模型的能力。在 2023 年,Wang Junlang 团队[17]提出了一个融合词性信息的预训练语言模型框架,有效提升了在缺乏明显情感词汇时的文本情感分析性能;而杜孟洋团队[18]则运用自注意力机制和句法依赖,进一步提升了情感分类的准确率。同年,周雨婷等人[19]提出了基于语义增强和指导路由的方法,通过键值对网络动态学习文本特征,优化了情感三元组的抽取。到了 2024 年,李增伟[20]提出的 SSES-SPAN 模型,通过特征编码和图卷积网络技术,不仅提高了情感分类的准确率,还为理解复杂语言表达提供了新的视角。同时,赵园春[21]提出了基于细粒度标记的抽取算法,通过卷积注意力机

制增强了语言学特征的提取。此外,黄梓芃等人[22]提出的 LFE 模型通过双仿射注意力和图卷积网络,强化了语义和句法特征,提升了模型的鲁棒性和准确性。因此,词性和位置特征的结合对于深入理解句子结构和情感倾向,捕捉文本中的情感细节至关重要。为此,该文提出了一种融合藏文词性和词性位置特征的模型,能够进一步提升细粒度情感分析的准确度。

2 模型与方法

2.1 词向量

Word2Vec[23]是由 Google 的研究人员在 2013 年提出的词嵌入技术,它通过在大量文本上执行无监督学习,将每个词转换成一个固定长度的数值向量。这些向量捕捉了词的语义特征,并通过向量空间模型的原理,使得语义上相似的词在该空间中的距离更近。Word2Vec 包括两种主要的模型架构:连续词袋(CBOW)和 Skip-gram。CBOW 模型致力于根据上下文预测目标词,而 Skip-gram 模型则利用单个目标词预测其上下文,这两种模型在不同的应用场景下各有优势。

Word2Vec 的特性使其在自然语言处理(NLP)的多个领域中得到广泛应用,包括文本分类、情感分析、机器翻译等。在藏文情感分析的背景下,由于缺乏藏文的大量开源语料,且训练资源有限,词向量的训练具有挑战性。因此,该文选择 Word2Vec 作为词向量模型。

2.2 词 性

藏文作为一种音节文字,其语法结构和词性系统与汉语相比有显著的不同,特别是在情感信息的传递上这些差异尤为突出。在藏文中,名词、形容词和动词等不同的词性扮演着至关重要的角色,它们在情感表达中各司其职,共同构成了情感交流的复杂网络。在藏文情感三元组分析中,准确识别和判断方面词的情感倾向对于信息的提取和模型的训练至关重要。方面词大部分是名词,是构建句子和传达情感信息的基础。而情感词,通常用来明确表达情感倾向,它们在藏文中往往采用形容词和动词的形式,为文本情感的表达增添了丰富的层次。名词、形容词和动词等协同工作,共同构成了藏文细粒度情感分析的坚实基础。该文通过研究大量的细粒度情感分析案例,得出细粒度情感分析所需的词性分类,如表 1 所示。

表 1 细粒度的词性

词性	藏文表示	英文缩写	例词
名词	མིང་གི་ངོ་བོ་ཚིག	n	རྩོམ་པ་ལ་གྱི་（文章） བརྙན་འཕྲིན་（视频）
形容词	ཁ་དོག་ཚོམ་གཟུ་ཤིང	a	མཛེས་པ་མཉེ་གཡ་དེ་དེ་བོ་（美丽的） མ་མཚན་བཟང་ཅན།（善良的）
及物动词	བྱ་དོ་པ་པོ་འི་བྱ་དོ་ཨ་ཁ	vt	བཙོང་（买） བཙུ་རུ་（转动）
不及物动词	བྱ་དོ་པའི་བྱ་དོ་ཨ་ཁ	vi	འགྲོ་（去） གསལ་（清楚）
代词	ཚབ་ཚེ་གི	r	འདི་（这个） ཁོ་（你）
程度副词	དང་མ་ཚད་མེ་ཅད་པ	dc	ཏ་ཅང་（非常） 稍微（ཙུང་ཟད་）
转折词	མེ་མཐུན་པའི་ཚེ་གི་རྒྱུ་ན	cj	འོན་ཀྱང་（然而） དེ་ལང་（但是）
属格词	པ་དོ	gz	གི་（连词） དང་（和）
时频副词	དུ་མ་ཚད་མེ་ཅད་པ	dp	མཆམས་འགའ།（偶尔） གཟུ་བུ་རང་རུ།（突然）
否定副词	དགག་པ་མེ་ཅད་པ	df	མེན་（不） མེད་（没有）
其他类别	རེ་གམ་གཞན་དགུ	h	འདྲ་（像） རེད་（是）

通过词性标注,有助于更准确地定位和区分情感词和方面词,以表 2 中的句子“ལག་ཁུག་གི་ཇི་མཐའི་དག་ཅིང་ཤོས་ཡོད་པ། འོན་ཀྱང་ཁ་ཤས་ཆུང་ངུ་ཡིན་པ་མ་ཡིན་ནོ།(译:这个手提包的设计很漂亮,但容量不大。)”为例,能够看到代词““འདི་/r(译:它)”用来指特定的手提包,帮助模型理解文本中的指代关系,从而更准确地识别出情感词和方面词所指向的具体对象。该句中的方面词为:“ཇི་མའི་/nn(译:设计)”、“ཁྱོ ཚད་/n(译:容量)”均为名词;情感词:“ཏ ཅང མཛེས་/a(译:漂亮)”、“ཆེན པོ མ་/ad(译:不大)”均为形容词,因此对句子进行词性标注后可以帮助模型更好地定位情感词和方面词,另外句子中的“འོན་ཀྱང་/cj(译:但是)”为转折词,可以帮助模型识别情感词和方面词之间的关系,以及它们之间的情感极性。同理在第二个句子中“ཁོ་ཚོ་རང་གི་ཁྱོ་བོ་ཚང་གི་ཐེ་ཁའམ་ལ་སེང་ཁ་ཚོད་ཅིག་ཅང་དལ་མོ་ཞིག།(译:他们经常去的餐厅做饭的速度很慢,但饭菜的味道很好)”,方面词与情感词也分

别为名词和形容词,且句子中的时频副词“རྒྱག་ཏུ་(经常)”暗示了“他们”对餐厅可能有了解和偏好。及物动词“འགྲོ་(去)”表明他们习惯性访问该餐厅。虽无情感色彩,却为理解主体对餐厅的积极情感倾向提供了重要线索。因此不同的词性在句子中有特定的角色并共同相互协作。引入词性分析的情感三元组方法不仅能够关注单个词语所带有的情感色彩,还能够综合考量整个句子所表达的情感态度。有助于深入挖掘文本的深层情感含义,准确理解和分析隐藏在文本情感层面的内容。

表 2　示例句子词性

例句	译文	词性标注	(方面词，情感词，情感极性)
ལག་ཁྱེག་གི་ནང་འཇམ་འགོད་དེ་ཅང་སྐེམ་མ་ཤོད། འོན་ཀྱང་ཁ་ཤས་ཆུང་ངུ་ཡིན་ཏེ་མི་ཆོག་	手提包的设计很漂亮，但它的容量不大	ལག་ཁྱེག/n གི་/gz ནང་འཇམ་/n ཅ་ཅང་/dz སྐེམ་/a འོན་ཀྱང་/cj ཁ་ཤས་/n ཆུང་ངུ་/a ཡིན་/df	(设计，漂亮，积极) (容量，不大，消极)
གོང་ཆོང་ར་ཆུག་ཅུ་འགར་ཟེར་བ་འཚེ་ཁང་ཁག་ཏུ་ཟ་ཐང་ཡམ་པའི་སྒང་རྣམས་ཆོད་ཅུ་ཅང་དལ་སོང་། འོན་ཀྱང་ཁ་སའི་ཏུ་ཟ་ཐང་ཕྱི་རོལ་	他们经常去的餐厅做饭的速度很慢，但饭菜的味道很好	གོང་ཆོང་/dz ར་ཆུག་ཅུ་/dp འགར་ཟེར་/vi བ/h འི་/gz ཟ་ཐང་/n གི་/gz ཟ/n ས/n ཡམ/vt པ/h འི་/gz ས་ཕྱོགས་ཆོད་/n ཅུ་ཅང་/dz དལ/a སོང་/h འོན་ཀྱང་/cj ཟ་ས/n འི་/gz བཟོ་བ/n ཅུ་ཅང་/dz བཟང་/a	(速度，很慢，消极) (味道，很好，积极)

2.3　位　置

在藏文的细粒度情感分析中,精准地识别和区分文本中的“方面词”和“情感词”的位置对于构建情感三元组至关重要。情感三元组由情感主体、情感词和情感极性三个要素构成,它们共同构成了情感表达的完整框架,反映了文本中的情感倾向和评价对象。在藏文语境中,“方面词”一般出现在句子的起始位置或中间,它们指明讨论的主题或对象,为理解整个情感语境提供了出发点。紧随方面词之后的是用于评价该方面词的情感词,它们承载着与该主题或对象相关的情感态度,是情感表达的核心部分。为了提升情感分析的精确度,特别是在自动化地识别情感三元组的组成部分方面,该文在模型中添加具体位置信息,并采用 BIO 标记方法。不仅增强了模型对情感倾向的精确识别能力,而且清晰地揭示了目标与其对应观点之间的直接联系。方面词的标记方式如式 1 和式 2 所示。

$$AP = T^i - Senti, Senti = POS \mid NEG \tag{1}$$

$$OP = S^i \tag{2}$$

式 1 中当变量 $i = 1$ 时,表示第一个方面词,被标记为“T”。如果与这个方面词相应的情感词是积极的,则标记为“T-POS”;如果是消极的,则标记为“T-NEG”。同理,当 $i = 2$ 时,表示第二个方面词,被标记为“TT”,与这个方面词相应的情感词是积极的,则标记为“TT-POS”;如果情感词是负面的,则标记为“TT-NEG”;以此类推,第三个方面词为消极时,标记为“TTT-NEG”。式 2 中当 $i = 1$ 时,是第一个方面词相关联的情感词被标记为“S”,当 $i = 2$ 时而与第二个方面词相关联的情感词被标记为“SS”。通过这种标记策略,能够更有效地识别和分析文本中的情感三元组,包括方面词、情感词以及它们的情感极性。使得模型能够更好地理解和处理藏文中的复杂情感信息。如表 3所示,在示例一中,位置为(4,5)的方面词被标记为

表 3　位置编码句子示例

示例	项目	内容
示例一	例句	ཁ་ཤས་མའི་ཐུན་ཚེ་ལ་ཅ་ཅང་ལགས་སོ་འི་ནང་ཚང་ཁྱིམ་ཁ་ཏོན་ཏེ་སྡོད་
	位置信息	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
	方面词标注	O O O T-POS T-POS O O O O O O TT-NEG TT-NEG O O
	情感词标注	O O O O O S S S O O O O O SS SS
	译文	包包的外表很漂亮，但是价格比较高
示例二	例句	ཡག་པོ་ཟླ་བ་རེ་པོ་ཏ་རྣམས་ཀྱི་ཁ་ཏོན་ཏེ་ཐོབ་ཐུབ་པ་ཡིན། འོན་ཀྱང་ཁ་པའི་ཐུན་ཚེ་ལ་ཡག་པོ་ཟླ་བ་རེ་པོ་ཏ་ཡིས་ཁྱོད་ཀྱི་ཁ་ཐག་ཤིན་ཏུ་ཡངས་པ་ཞིག་ཡོད།
	位置信息	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
	方面词标注	O O O O O O T-NEG T-NEG O O O O O O O TT-POS TT-POS O O O O O O O O
	情感词标注	O O O O O O O S S O O O O O O O O O O SS SS SS SS O
	译文	这款护肤品的价格偏高,但护肤效果很明显

“T-POS”,表示正面情感;而紧随其后的情感词位置为(6,8)则被标记为“S”。同理,在示例二中,位置为(7,8)方面词被标记为“T-NEG”,表示负面情感,位置为(4,5)情感词标记为“S”。位置为(16,17)方面词被标记为“TT-POS”,表示正面情感,位置为(20,23)情感词被标记为“SS”,通过以上示例可以判断出藏文的方面词一般出现在句中起始或中间位置而紧随其后的是情感词,即位置信息的融入使方面词、情感词关系更明确有助于判断情感三元组。

2.4 模 型

OTE-MTL 模型[24]是由 Zhang Chen 等人在 2020 年提出的情感三元组识别模型,它通过多任务学习框架有效地抽取文本中的方面和观点,并准确解析它们之间的情感依赖关系。该模型采用的多头部架构和先进算法提高了处理的灵活性和准确性。但是,由于模型缺乏对藏文词义的理解,对藏文细粒度分析存在着挑战。为了克服这一难题,该文提出将词性和位置特征与 OTE-MTL 模型相结合,以更深入地理解藏文情感三元组,即 OpinionNet-OTE-MTL 模型。该模型通过两个主要步骤实现情感三元组的提取,首先是预测阶段,其次是解码阶段。在预测阶段,将位置嵌入、词嵌入与词向量嵌入相结合构建句子的编码表示,再输入至双向长短期记忆网络(BiLSTM)中进行学习。提取方面词和情感词的特征表示,并通过两个序列标记器进行方面和情感的标记,同时使用双向评分器预测词对之间的情感依赖。接着通过联合训练最小化标记损失和依赖解析损失进行多任务学习。在解码阶段,利用启发式规则从提取的方面词、情感词和情感极性生成最终的三元组输出。模型框架如图 1 所示。

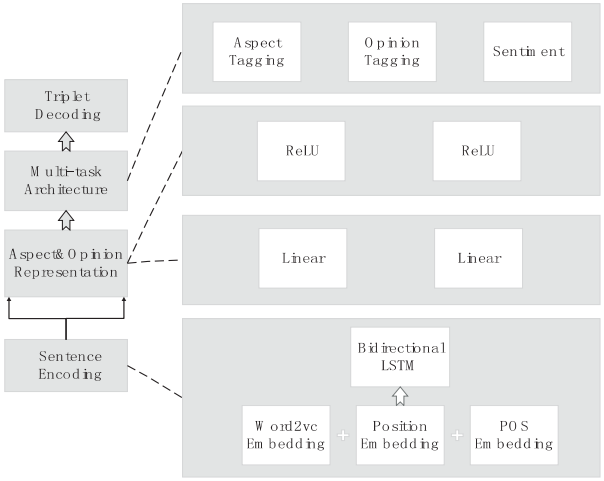

图 1 OpinionNet-OTE-MTL 模型框架

OpinionNet-OTE-MTL 模型主要使用双向长短期记忆网络(BiLSTM)来处理输入文本,以捕获每个词的上下文信息及其在句中的位置特征。其中,双向 LSTM 通过以下操作进行特征提取:

$$\text{Embedding} = e_i \oplus p_i \oplus t_i \tag{3}$$

$$h_i = [\text{Lstm}(\text{Eembedding})][\text{Lstm}(\text{Eembedding})] \tag{4}$$

式中,e_i 是由 Word2Vec 获得的词嵌入,p_i 是词的位置向量,t_i 是词的词性表示,Eembedding 是融合了位置、词性和词信息的上下文表示,h_i 是利用 Lstm 对 Eembedding 进行特征提取获得的更丰富的语义信息,使得生成的上下文向量序列更为精确。其次,模型通过线性降维和 ReLU 激活函数提取藏文方面和情感词的关键特征。这一步旨在去除对后续计算不必要的信息,减少过拟合风险,并剔除与方面标记和情感词标记无关的特征。图 2 为 OpinionNet-OTE-MTL 模型的输入输出样例图,通过在模型的输入添加词性和位置信息,可以辅助模型对句子的方面词和情感词的词性和位置编码以及情感态度进行判断,从而能够准确抽取出文本的情感三元组。结合 Word2Vec、词性及位置编码输入至 BiLSTM 模型,还可以增强模型对藏文文本情感内容的理解。

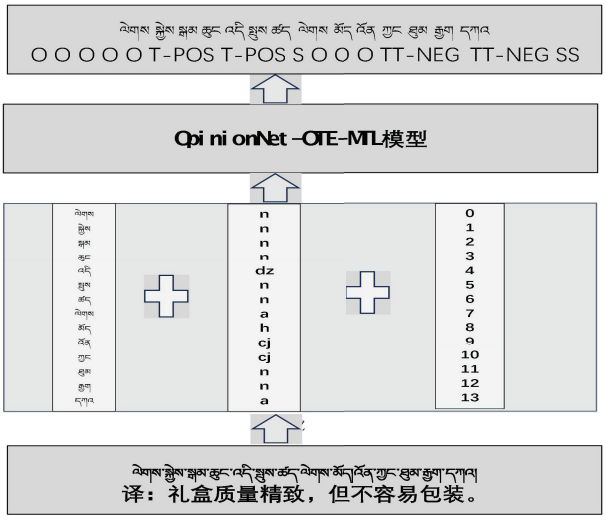

图 2 输入输出样例图

3 实验数据与评价指标

3.1 实验数据

在藏文细粒度情感分析研究领域,由于缺乏公开的语料库导致无法进行细粒度情感分析,因此,为了评估 OpinionNet-OTE-MTL 模型的有效性,该文构建了一个新的藏文细粒度文本语料库,用于情感三元组识别测试。构建步骤如下:首先,从中文社交媒体平台(如微博、电商和淘宝)收集大量产品的用户评价;其次,为了确保翻译的情感准确性,选择了微软翻译工具。基于微软翻译在精确捕捉情感极性和深入理解上下文方面的显著优势,因此使用微软翻译,将从多个中文社交媒体平台收集的数据集转换成了藏文;最后,对构建的藏语文料库进行了详尽的人工标注和校对,确

保了数据集的准确性和质量。经过以上步骤,该文构建了一个包含两千条数据的语料库,并以 8∶1∶1 的比例将其划分为训练集、测试集和验证集。数据集样式如表 4 所示。

表 4　数据集样式

例句	ཨཁྱེའུ་ཁག་ཅན་འདི་ཧ་ཅང་ཚུ་ཁ་ཤས་ཆེ་ཞིང་ཧ་ཅང་ཉུང་ཆུང་ཡིན།
方面词标注	OO O O T-POS T-POS O O O TT-NEG TT-NEG O
情感词标注	OO O O O O S O O O O SS
例句	བཀྲ་ཤིས་བདེ་ལེགས་པ་ནི་ཁྱོད་ཀྱི་ཐུགས་དགོངས་དམངཔ
方面词标注	OO O O T-POS T-POS O O O TT-NEG TT-NEG O O
情感词标注	OO O O O O S O O O O SS SS
例句	ཚོང་ཁང་ཁ་གི་ཨ་ཁ་ཤང་འདི་ནི་ཧ་ཅང་ཐ་མཁ་ཅང་ཡིན།
方面词标注	OO O T-POS T-POS O O O TT-NEG TT-NEG O O
情感词标注	OO O O O S O O O O SS SS

3.2　评价指标

使用 F1 值(F1)作为评估指标,F1 值是精确率(P)和召回率(R)的调和平均。计算公式如下所示:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

其中,TP 是真正例的数量,FP 是假正例的数量。

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

其中,FN 是假负例的数量。

$$\mathrm{F1} = \frac{P \times R}{P + R} \tag{7}$$

F1 分数是精确率和召回率的调和平均数,它在两者之间取得平衡。

4　实验结果及分析

4.1　实验设置

该文从新闻网站等爬取了 430 MB 数据用于训练 Word2Vec 词向量模型,其训练参数如表 5 所示。

表 5　Word2Vec 超参数

参数名称	参数选择
sg(model)	1(skip-gram)
window	5
min_count	10
Embed_dim	300
data size	430 MB

对 OTE-MTL 文中使用 HAST、CMLA 作为对比模型,而该文在 OTE-MTL 模型的基础上进行改进,因此选择 HAST、CMLA 和 OTE-MTL 三种模型进行基线实验,并在构建的数据集上进行了训练。另外,为了验证词性和位置编码对模型的影响,实现了 OpinionNet-OTE-MTL 模型的消融实验,基线、消融实验的参数设置如表 6 所示。

表 6　模型超参数

参数名称	参数选择
Batch size	8
Learning rate	0.05
Log_step	5
epoch	100
Embed_dim	300
Hidden size	300

4.2　基线实验

基线实验的实验结果如表 7 所示,其中 ap 是方面词提取任务,op 是情感词提取任务,triple 为三元组提取任务。CMLA 模型利用注意力机制来提取特征,专门用于识别文本中的方面词和观点词。HAST 模型通过优化识别和提取描述事物特征词性的能力。在方面词(ap)提取层面 HAST 模型表现最佳,其精确率和 F1 值比 OTE-MTL 模型提高了 1.38 百分点和 1.55 百分点。OTE-MTL 模型则在三元组提取(triple)方面,表现最佳,F1 值相比 HAST 和 CMLA 分别提高了 2.73

表 7　不同模型在情感三元组任务上的实验结果对比　%

模型	ap(方面词)			op(情感词)			triple(情感三元组)		
	精确率	召回率	F1 值	精确率	召回率	F1 值	精确率	召回率	F1 值
HAST	**39.46**	**44.49**	**41.83**	61.90	64.99	63.41	35.15	22.99	27.43
CMLA	30.27	41.24	34.92	32.69	42.99	37.14	16.49	11.74	13.72
OTE-MTL	38.08	42.74	40.28	**62.38**	**66.74**	**64.49**	**43.80**	**22.99**	**30.16**

百分点和 16.44 百分点。通过实验结果分析发现,虽然方面词提取层面 HAST 模型略优于 OTE-MTL,但对整体情感三元组提取和情感词提取的效果在 OTE-MTL 模型上表现最佳。因此后文实验在 OTE-MTL 上融合词性与绝对位置信息。

4.3 消融实验

为了验证提出的融合词性和位置向量的 OpinionNet-OTE-MTL 模型的合理性,设计了消融实验评估词性和位置对模型的影响效果,实验结果如表 8 所示。与 OTE-MTL 模型相比,仅融入词性的 POS-OTE-MTL 在三元组提取任务(triple)上表现出了一定的提升,其中 F1 值提高了 3.06 百分点。这表明词性信息对于识别文本中的三元组关系具有一定的帮助。在方面词提取任务时,POS-OTE-MTL 模型相比 OTE-MTL 模型,F1 值提升了 3.09 百分点,因此词性对提取方面词有关键作用,有助于句子结构识别、语义角色揭示、情感表达关联及上下文理解。其次,与 OTE-MTL 模型相比,在三元组(triple)提取任务上,融入位置信息的 Position-OTE-MTL 模型 F1 值提升了 2.83 百分点,与仅融入词性相比减少了 0.23 百分点。因此,结果证明了融入词性信息对三元组识别的影响大于位置信息。最后,当 OpinionNet-OTE-MTL 模型融合了词性和位置信息时,其性能提升更为显著。在三元组提取(triple)任务上,与仅融入词性的 POS-OTE-MTL 模型相比,OpinionNet-OTE-MTL 模型的精确率、召回率和 F1 值分别提高了 4.68 百分点、2.25 百分点和 3.08 百分点。而在与仅融入位置信息的 Position-mtl 模型相比时,在方面词(ap)提取任务上,OpinionNet-OTE-MTL 模型的精确率、召回率和 F1 值分别提高了 4.76 百分点、1.5 百分点和 3.33 百分点。消融实验的结果证实了 OpinionNet-OTE-MTL 模型通过整合词性和位置信息,相辅相成增强了模型各子模块的有效性,并在提升情感三元组任务的性能上发挥了重要作用。

表 8 消融实验结果 %

模型	ap(方面词)			op(情感词)			triple(三元组)		
	精确率	召回率	F1 值	精确率	召回率	F1 值	精确率	召回率	F1 值
OTE-MTL	38.08	42.74	40.28	**62.38**	**66.74**	**64.49**	43.80	22.99	30.16
Position-mtl	40.25	46.99	43.36	59.16	63.74	61.37	39.64	28.24	32.99
POS-OTE-MTL	40.16	47.99	43.37	61.35	65.74	63.36	43.85	26.74	33.22
OpinionNet-OTE-MTL	**45.01**	**48.49**	**46.69**	57.00	58.99	57.98	**48.53**	**28.99**	**36.30**

5 结束语

藏文的情感三元组提取任务作为新兴的细粒度情感分析领域,具备巨大的发展潜力,相对于传统的情感分类,它提供了更为深入和详细的情感理解。针对目前藏文情感三元组研究较少,并且藏文细粒度情感抽取困难的原因,该文提出了 OpinionNet-OTE-MTL 模型,该模型通过结合词性和其在文本中的位置信息,更全面地捕获了文本的情感信息,词性的融入可以更好地定位方面词和情感词,位置的融入则增强了模型对情感倾向的精确识别能力。通过对比和消融实验均证明了该模型的有效性。

此外,藏文作为一种语法结构复杂的语言,该文仅考虑了词性和位置信息,且标注为人工标注,因此存在着主观性、成本高和忽略上下文的问题,未来将继续探索并融合更多的藏文语法结构和语义信息,以增强藏文情感分析的细致度和全面性。同时,在数据构建方法上采用机器标注结合人工校正的方法,以及期望在后续研究中,通过深入挖掘藏文文法,并将依存关系分析和注意力机制等先进技术应用于模型,进一步提升情感三元组提取的准确度。

参考文献:

[1] 张 俊,李应兴. 基于情感词典的藏文微博情感分析研究[J]. 硅谷,2014,7(20):220.

[2] 袁 斌,江 涛,于洪志. 基于语义空间的藏文微博情感分析方法[J]. 计算机应用研究,2016,33(3):682-685.

[3] 孙本旺,田 芳. 基于深度学习算法的藏文微博情感计算研究[J]. 计算机技术与发展,2019,29(10):55-58.

[4] 李海刚,于洪志. 藏文文本情感分类系统设计[J]. 甘肃科技纵横,2011,40(1):106-107.

[5] 拥 措,史晓东,尼玛扎西. 短文本情感分析的研究现状——从社交媒体到资源稀缺语言[J]. 计算机科学,2018,45(s1):46-49.

[6] 公保加羊,拉玛杰,官却多杰,等. 基于深度学习的藏文舆情分析研究[J]. 青海科技,2023,30(1):56-60.

[7] 朱宇雷,德吉卡卓,群 诺,等. 基于图神经网络结合预训练模型的藏文短文本情感分析研究[J]. 中文信息学报,2023,37(2):71-79.

[8] 孟祥和,于洪志. 融合音节和词条特征的藏文文本情感分类研究[J]. 中文信息学报,2023,37(2):80-86.

[9] WANG Wenya, PAN S J, DAHLMEIER D, et al. Coupled

multi-layer attentions for co-extraction of aspect and opinion terms[C]//Proc of AAAI conference on artificial intelligence. Palo Alto:AAAI,2017:273-284.

[10] LI Xin, BING Lidong, LI Piji, et al. Aspect term extraction with history attention and selective transformation[C]//Proc of the 27th international joint conference on artificial intelligence. Palo Alto:AAAI,2018:4194-4200.

[11] 赵　富,杨　洋,蒋　瑞,等. 融合词性的双注意力 Bi-LSTM 情感分析[J]. 计算机应用,2018,38(S2):103-106.

[12] 姚艳秋,郑雅雯,吕妍欣. 基于 LS-SO 算法的情感文本分类方法[J]. 吉林大学学报:理学版,2019,57(2):375-379.

[13] 王行甫,王　磊,苗付友,等. 结合词性、位置和单词情感的内存网络的方面情感分析[J]. 小型微型计算机系统,2019,40(2):383-389.

[14] LIM W L, HO C C, TING C Y. Sentiment analysis by fusing text and location features of geo-tagged tweets[J]. IEEE Access,2020,8:181014-181027.

[15] XU Lu, LI Hao, LU Wei, et al. Position-Aware tagging for aspect sentiment triplet extraction[C]//Proc of conference on empirical methods in natural language processing. Stroudsburg:ACL,2020:2339-2349.

[16] 薛　芳,过　弋,李智强,等. 基于双层词性感知和多头交互注意机制的方面级情感分析[J]. 计算机应用研究,2022,39(3):704-710.

[17] WANG J, LI X, HE J, et al. Enhancing implicit sentiment learning via the incorporation of part-of-speech for aspect-based sentiment analysis[C]//China national conference on chinese computational linguistics. Singapore:Springer,2023:382-399.

[18] 杜孟洋,王红斌,普祥和. 融入词性自注意力机制的方面级情感分类方法[J]. 吉林大学学报:理学版,2023,61(6):1375-1386.

[19] 周雨婷,代金鞘,刘嘉勇,等. 一种基于语义增强和指导路由机制的方面级情感三元组抽取方法[J]. 四川大学学报:自然科学版,2023,60(5):112-120.

[20] 李增伟,刘　帅. 语义和句法依赖增强的跨度级方面情感三元组抽取[J]. 计算机系,2024,33(6):201-210.

[21] 赵园春,韩　虎,徐学锋. 细粒度标记的结点自适应方面情感三元组抽取[J/OL]. 计算机工程与应:1-11[2024-08-16]. http://kns.cnki.net/kcms/detail/11.2127.tp.20240626.1435.006.html.

[22] 黄梓芃,曾碧卿,陈鹏飞,等. 基于语言特征增强的方面情感三元组抽取[J/OL]. 计算机工程:1-12[2024-09-01]. https://doi.org/10.19678/j.issn.1000-3428.0069260.

[23] 郑文超,徐　鹏. 利用 word2vec 对中文词进行聚类的研究[J]. 软件,2013,34(12):160-162.

[24] ZHANG Chen, LI Qiuchi, SONG Dawei, et al. A multi-task learning framework for opinion triplet extraction[C]//Proc of conference on empirical methods in natural language processing. Stroudsburg:ACL,2020:819-828.