基于多源知识融合的领域情感词典表示学习研究

祁瑞华[1,2] 魏佳[1] 邵震[1] 郭旭[1,2] 陈恒[1,2]†
1.大连外国语大学软件学院/ 辽宁省大连市
2.大连外国语大学语言智能研究中心/ 辽宁省大连市
rhqi@dlufl.edu.cn WeiJ0417@163.com JKL4131@126.com
guoxu@dlufl.edu.cn chenheng@dlufl.edu.cn

摘要

本文旨在解决领域情感词典构建任务中标注数据资源相对匮乏以及情感语义表示不充分问题，通过多源数据领域差异计算联合权重，融合先验情感知识和Fasttext词向量表示学习，将情感语义知识映射到新的词向量空间，从无标注数据中自动构建适应大数据多领域和多语言环境的领域情感词典。在中英文多领域公开数据集上的对比实验表明，与情感词典方法和预训练词向量方法相比，本文提出的多源知识融合的领域情感词典表示学习方法在实验数据集上的分类正确率均有明显提升，并在多种算法、多语言、多领域和多数据集上具有较好的鲁棒性。本文还通过消融实验验证了所提出模型的各个模块在提升情感分类效果中的作用。

关键词： 知识融合 ； 领域情感词典 ； 表示学习

Domain Sentiment Lexicon Representation Learning Based on Multi-source Knowledge Fusion

Ruihua Qi[1,2] Jia Wei[1] Zhen Shao[1] Xu Guo[1,2] Heng Chen[1,2]
1.School of Software Engineering of Dalian University of Foreign Languages / Dalian, Liaoning
2.Research Center for Language Intelligence of Dalian University of Foreign Languages / Dalian, Liaoning
rhqi@dlufl.edu.cn WeiJ0417@163.com JKL4131@126.com
guoxu@dlufl.edu.cn chenheng@dlufl.edu.cn

Abstract

This paper is aiming at the problems of lack of annotated data and inadequate sentiment semantic representation in existing domain sentiment lexicon construction methods. In this paper, the joint weight is calculated by multi-source data. Combining prior emotional knowledge and Fasttext word vector representation learning, the sentiment semantic knowledge is mapped to a new word vector space, and the domain sentiment dictionary is automatically constructed from unlabeled data to adapt to the multi-domain and multi-language environment. The comparative experiments on Chinese and English multi-domain public data sets show that, compared with sentiment dictionary and pretrained language model, the proposed multi-source knowledge fusion method of domain sentiment dictionary representation learning has significantly improved the classification accuracy on public data sets, and has good robustness on various algorithms, multi-language, multi-domain and multi-data sets. This paper also verifies the role of each module of the proposed model in improving the effect of sentiment classification through ablation experiments.

Keywords: knowledge fusion , domain sentiment lexicon , representation learning

第二十一届中国计算语言学大会论文集，第684页-第693页，南昌，中国，2022年10月14日至16日。

1 引言

情感词汇是文本情感表达的主要途径，由情感词汇构成的情感词典能够明显提升情感分析效果的同时具有很好的可解释性，是社交网络情感分析、商品评论观点挖掘等系统中的重要技术手段，已经成为是无监督情感分析的主要依据[1]。当前大数据多语言环境下，情感分析任务主要面临两个挑战：一是网络文本情感词汇语义内涵变化快、表达方式微妙，难以准确捕捉情感倾向；二是情感分析方法具有领域依赖性，在面向特定领域情感分析任务中，通用情感词典起到一定作用，但通用情感词典无法准确判断新词和领域特有情感词，覆盖率和极性判断准确率也难以满足领域变化各异的情感分析需求，通用情感词典或某个领域的情感词典应用于另一个领域时情感分析性能往往下降明显，新兴领域虽然有海量数据但缺乏先验情感知识的指导，因此迫切需求领域情感词典的自动构建方法。

除了情感词典，目前情感知识的来源主要包括领域内规模有限的有标注数据和无标注数据，此外，大量的领域外数据也隐含着对情感知识的有益的情感信息。为充分利用领域内及领域外的有标注和无标注数据中的情感知识，本文提出基于多源知识融合的领域情感词典表示学习方法，融合多源数据语义信息和情感信息弥补先验知识的不足，从无标注数据中抽取情感信息，结合领域情感知识对比方法自动构建适应大数据多领域、多语言环境的领域情感词典。

2 情感词典研究现状

2.1 语义扩展法

语义扩展法基于专家标注的情感知识库，首先人工选定少量的种子词，在情感知识库中查找每个种子词的同义词、反义词等词间关系进行扩展，经过多轮迭代生成新的情感词典。如Westgate等[2]从Thesaurus.com和WordNet语义知识库递归获取单词同义词构成词的同情感极性图，然后对优化路径中词汇的极性值加权平均决定目标词的极性。SAĞLAM等[3]基于同义词反义词数据集构建的词汇图扩展了土耳其语情感词典。Shaukat等[4]利用Vadar和Senticnet情感词典构成领域情感词典，但每个领域只有5至30个情感词汇。语义扩展方法依赖于人工标注的情感词典，一般规模较小，难以适应词义变化和网络新词的出现，通常作为辅助方法。

2.2 词频共现法

词频共现法包括词频法和词共现法，词频法计算词汇频率筛选情感词，如贺飞艳等[5]结合TF-ID和方差统计提出面向微博短文本的情感特征抽取的计算方法。词频共现法假设共现频率越高的词其语义关联越紧密，如Turney等[6]通过点互信息PMI计算候选词与情感种子词的距离，识别候选词的情感倾向。Mullen 等[7] 过PMI计算形容词的情感倾向值，Liu等[8]针对中文情感词典覆盖率低的问题，通过CHI卡方检验与改进的SO-PMI算法关联计算发现新的情感词。词频共现法单纯地依赖词共现统计信息无法有效表示自然语言的复杂语义，人工选择种子词也增加了不确定性。词频共现法的局限在于构建的词表规模太大导致效率低，同时没有充分利用文本的语义信息。

2.3 启发规则法

启发规则法主要通过观察总结自然语言的语法规则和语言学模式建立情感词典，语法规则如连词规则、否定词规则、双向传播规则以及人工定义的其他规则。如Qiu等[9]提出双向传播算法，定义了四类句法依存关系规则通过迭代路径抽取情感词和目标词，Wu [10]加入一致性连词和否定连词等语法规则改进了双向传播算法情感词极性检测。Hutto 等[11]提出基于简单规则的情感分析模型，使用群智方法人工打分选出情感特征集。启发规则法的局限在于需要专家参与人工定义规则，无法概括日新月异的语言现象，通常与其它方法结合应用。

2.4 词向量表示学习

词向量表示学习方面，Li等[12]面向旅游评论领域通过Word2Vec计算候选词与种子词的语义相似度，并用Interior Point Algorithm 内点算法计算候选词的情感值。杨小平等[13]基于Word2Vec算法提出转换约束集多维情感词典构建方法和基于词分布密度的情感类别及强度计

基金项目：大连外国语大学研究创新团队“计算语言学与人工智能创新团队”(2016CXTD06)；大连外国语大学科研基金项目(2021XJYB16)

算和消歧方法。张璞等[14]选择与种子词具有连词关系的词语作为候选情感词，基于种子词和候选情感词之间的Word2Vec词向量相似度构建语义关联图，使用标签传播算法计算情感词的极性构建情感词典，局限在于种子集基于人工选择，增加了成本和不确定性，例如真正的情感词未必与种子词通过连词连接。方法集成方面，Li等[12]面向旅游领域利用集合互信息AMI发现领域新词，结合人工情感评分值、Wordvec词向量与种子词的语义相似度和PMI相似度构建领域情感词典，改善了情感词典构建，但过程中需要人工参与情感词评分过程。蒋翠清等[15]面向社交媒体中的汽车评论，分别利用PMI和Word2Vec 算法识别新词情感极性，根据集成规则对二者识别结果综合判定构建领域情感词典。现有词向量情感词典将语义信息看作情感信息，存在着局限。

3 基于多源知识融合的领域情感词典表示学习

本文提出基于无标注数据的多源知识融合领域情感词典表示学习方法（Multi-source knowledge Fusion based Domain Sentiment Lexicon representation Learning，MFDSL），表示学习框架如图1所示，主要分为四个模块：多源数据融合领域差异联合权重计算模块、情感知识融合模块、Fasttext表示学习模块和情感词典表示学习模块。

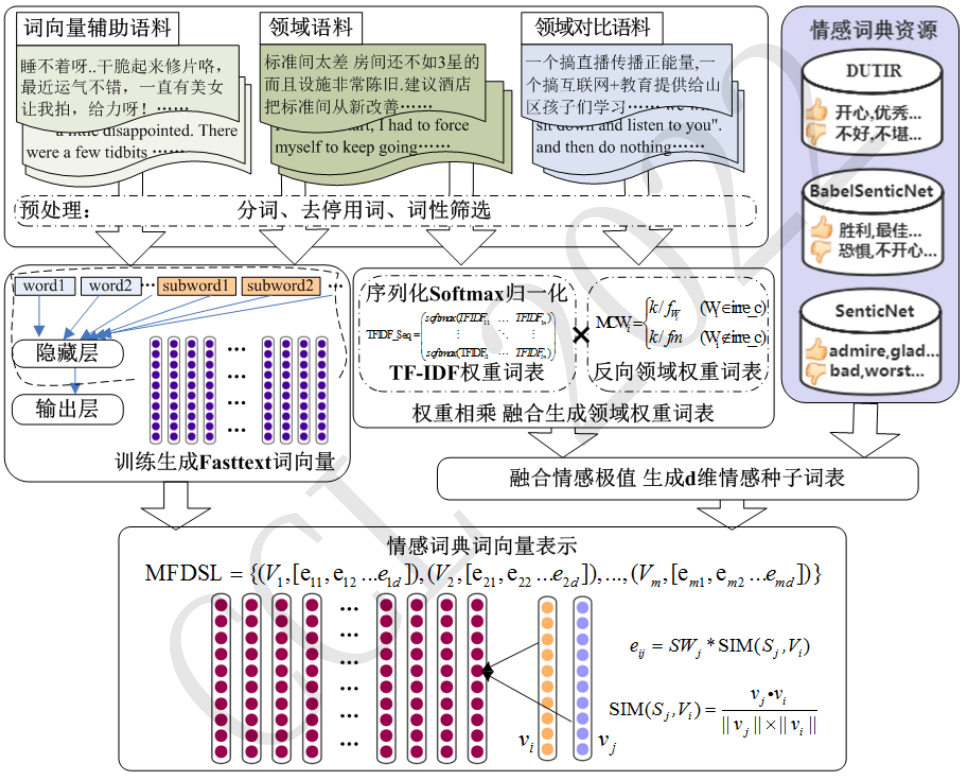

Figure 1: 基于多源知识融合的领域情感词典表示学习框架

3.1 基于多源数据和领域差异权重的情感种子词选取

3.1.1 领域差异联合权重计算

领域情感词提取的首要问题是词汇权重的计算，现有研究大多根据领域语料资源与已有情感知识库结合产生情感词表，计算权重也依赖于领域语料，本研究提出领域差异联合权重计算，引入非领域信息增强领域专有情感词的权重。首先，改进TF-IDF算法作为基础权重计算方法，赋予在少量领域样本中频率高的但总词频不高的词汇比较高的权重。TF权重采用Log标准化，IDF权重采用如(1)所示的逆向文档频率平滑计算方法，其中d为每一条样本，是词t在样本d中出现的频率，N为所有样本数，nt是出现词t的样本数。为避免出现TF-IDF数值过大的情况，对每个样本的TF-IDF值进行Softmax归一化。

$$W_{tf-idf} = log(1 + f_{t,d}) * log(\frac{N}{1 + n_t}) + 1 \tag{1}$$

情感词在不同领域语料的分布差异是发现低频领域情感词的重要线索之一。为找出领域中特有的情感词, 本研究基于统计学计算领域差异联合权重，引入外部的领域对比语料强化领域相关词的权重，降低常用的领域无关词的权重。例如在餐厅评论中，“美味可口”的领域优先级应当高于“好”。利用领域对比语料的联合权重计算思路是：计算所有本领域语料词汇在领域对比语料中的词频，以平均词频为基准设置词频和权重的反比关系。设代表每个词汇的词频，k为所有词汇在领域对比语料中的均值，为防止函数分母为0，赋予词汇一个足够小的词频值为该词汇未在领域对比语料中出现时的缺省值，如公式2所示：

$$MCW_i = \begin{cases} k/f_{w_i} & (W_i \in irre\_c) \\ k/f_m & (W_i \notin irre\_c) \end{cases} \tag{2}$$

3.1.2 融合情感知识选取种子词

对于3.1.1节生成的候选词汇联合权重，融合多个情感知识库中情感极性加权求和，再与种子词权重值相乘赋予权重情感极性值，然后通过阈值筛选情感种子词，生成情感种子词表及其对应的权重。中文情感常识知识库采用大连理工大学信息检索研究室的情感词汇本体和Babel SenticNet，英文采用Senticnet 6，多语种情感常识知识库采用Babel SenticNet。大连理工大学信息检索研究室的情感词汇本体包括7大类20小类共27466条中文情感词汇，词典中每个情感词都标注了正向、负向、中性情感极性和情感强度。SenticNet由美国麻省理工学院媒体实验室、斯特灵大学和Sitekit Solutions公司合作构建，目前由Sentic项目组和来自于新加坡南洋理工大学等多家研究机构多领域的专家学者维护，其中SenticNet 6提供了语义和情感关联的20万个英文概念级常识库，标注了情感极性和四个情感维度的情感值，Babel SenticNet是基于SenticNet借助统计翻译方法建立的40种语言的常识知识库。本研究选取上述情感知识库中正向和负向的情感词汇。

3.2 基于Fasttext的词向量表示

词向量保留的语义信息不等同于情感信息, 不同情感极性的词语在语义上可能高度相似, 例如“不错”与“不差”在词向量中有较大相似度，但这两个词的情感极性完全相反，导致词语的情感极性分类不准确。因此本文结合情感知识库和词向量的情感知识和语义信息, 结合多源数据和情感知识库自动构建情感词典，借助深度学习进行词嵌入表示,通过表示学习将Fasttext词向量和情感权重映射到新的情感语义空间，更准确地表示情感语义。

本文采用适应大数据多语言环境的Fasttext词向量进行词汇的表示学习，Fasttext词向量表示学习原理核心思想为[16]：引入子词信息丰富词汇形态学表征信息，将整篇文档的词及n-gram向量叠加平均得到文档向量，使得生僻复杂的单词也能从结构相似的其他单词获得较好的词向量表示。Fasttext突破了土耳其语、芬兰语等形态丰富语种的预训练瓶颈，支持157种语言。Fasttext预训练模型如图1所示，输入层特征向量包括词序列中的所有词、子词和n-gram，并对各个词向量进行加和平均线性变换映射到隐藏层，然后在输出层通过层次softmax函数遍历分类树的叶节点寻找最大概率的分类标签，从而提高了词向量训练速度更适合大规模数据训练。

3.3 基于多源知识融合的领域情感词典学习算法

输入：领域语料re_c、词向量辅助训练语料ft_c、非领域语料irre_c、融合情感词典sl。**输出**：情感词典表示词向量MFDSL。

步骤1：训练词向量，将领域语料与re_c与词向量辅助训练语料ft_c合并，进行分词以及去停用词处理，使用Fasttext预训练得到语义词向量；

步骤2：对领域语料re_c进行分词以及停用词和词性筛选处理，去除助词、标点符号、非语素字、介词、量词、数词、名词、动词，并进行序列化处理。

步骤3：计算各词的TF-IDF值，并对每个句子序列进行softmax处理。得到TFIDF矩阵TFIDF_Seq。然后根据TFIDF_Seq求出每个词的TF-IDF权重，得到TF-IDF值词表，其中m为语料中的词形(Type)数量。

步骤4：对非领域语料irre_c进行分词以及去停用词操作，利用公式计算得到每个词的权重MCW，其中代表的词频，参数k为公式的权重值可进行调整，默认情况下取所有情感词在非领域语料中的均值，为当不存在于irre_c中时词频的缺省值，默认情况取值为0.5。最终生成多语料权重词表。

步骤5：将TFIDF_L与MCW_L中相同词的权重相乘，构成情感候选词的融合权重词表，其中权重weight的计算公式为。

步骤6：根据sl中的情感极性，与Weight_L相结合生成维度为d的情感种子词表Seeds。首先求得各情感词的情感权重值，其中p代表sl中的情感词极值，计算方法为各情感词典的极值加权之后求和。然后将词表按照倒序排序，选取正权重值前d/2个词，负权重绝对值前d/2个词，最终得到种子词表;

步骤7：生成情感词典词向量表示，其中每个情感候选词的维度为d，表达式为，其中参数e的计算方法为，SIM函数的计算方法为：利用训练好的fasttext词向量分别得到情感词与种子词的向量表示与，然后利用与计算情感词与种子词之间的相似度。

4 实验结果及分析

情感词典是无监督情感分类任务的主要依据，因此可以通过情感词典在情感分类任务中的效果来间接评估情感词典的有效性[1]，本实验的对照实验包括：表示学习维度对照实验、中文领域情感词向量对照实验和英文领域情感词向量对照实验，同时采用不同领域语料测试本文方法对多语种和多领域的适应性。

4.1 实验数据

本文实验中的中文领域语料来源于谭松波的酒店评论公开数据集[17]，其中正向与负向情感领域语料各2000条，实验选取正负向各1000条数据作为训练数据，正负向各500条作为验证集，剩余的正负向各500条作为测试数据。中文词向量辅助语料采用NLPIR微博内容语料库中新浪微博和腾讯微博评论23万条[18]，以及谭松波整理的1万条酒店评论语料，合计24万条。对这24万条数据进行分词以及去除停用词处理，作为Fasttext词向量的训练语料。中文领域对比语料来源于SMP-EWCT2020的评测数据[19]，包含微博评论共46421条。英文领域语料采用Amazon公开评论数据集[20]，覆盖图书、DVD、电子产品、厨房用品和影像五个领域，每个领域选取标注语料6000条，实验数据随机选取各领域正负向各1000条数据，选取其中50%作为训练集。英文词向量训练辅助语料采用Blitzer收集整理Amazon评论中的无标注数据共80821条[20]，英文领域对比语料来自于纽约时报新闻评论的公开数据共49868条[21]。

Table 1: 情感词典词向量构建采用的语料

语料名称	样本总数	正向样本	负向样本	无标注样本
中文领域语料ChnSentiCorpHtlba4000	4000	2000	2000	0
中文词向量辅助语料ChnSentiCorpHtluba10000	10000	0	0	10000
中文词向量辅助语料	230000	0	0	230000
中文领域对比语料	46421	0	0	46421
英文领域语料Amazon reviews(books)	6000	3000	3000	0
英文领域语料Amazon reviews(dvd)	6000	3000	3000	0
英文领域语料Amazon reviews(electronics)	6000	3000	3000	0
英文领域语料Amazon reviews(kitchen)	6000	3000	3000	0
英文领域语料Amazon reviews(video)	6000	3000	3000	0
英文词向量辅助语料Amazon reviews	80821	0	0	80821
英文领域对比语料	49868	0	0	49868

4.2 实验参数

预处理模块中，中文语料采用结巴分词的paddle模式处理，去除助词、标点符号、非语素字、介词、量词、数词和叹词，采用哈工大的中文停用词表。英文语料通过Spacy筛选词

性，采用Spacy模块中的英文停用词表。中英文情感词向量对照实验中，选取五种对照方法与本文方法对比，分别为：(1)情感本体方法，情感词典由相应的情感知识库采用One-hot编码构成，句子向量的维度为情感本体中所有词语的个数，编码值为情感强度；(2)Word2Vec方法，词向量仅由Word2Vec预训练算法生成，词向量维度为100，迭代次数为30，词的最小出现次数为2，句向量的计算采用词向量求和平均生成，句向量的维度与词向量维度相同；(3)Fasttext方法，词向量只由Fasttext预训练算法生成，实验参数与Word2Vec相同；(4)TFIDFSenti2vec，文献[22]中的基于词向量的情感词典方法；(5)TFIDF方法，通过本文的情感种子词生成模块产生情感词典作为输入，未结合Fasttext词向量；(6) MFDSLSVM，本文提出的多源知识融合领域情感词典表示学习方法，情感分类算法采用SVM，实验平台为Sklearn，采用线性核函数Liner和概率估计；(7) BertBiLSTM，采用预训练语言模型Bert和深度学习分类算法BiLSTM；(8) MFDSLBertBiLSTM，采用本文提出的多源知识融合领域情感词典表示学习方法，结合预训练语言模型Bert和深度学习分类算法BiLSTM。

评价指标选取正负向语料的精度、召回率、F1值和总体准确率检验情感词典对文本情感分类任务的有效性。

4.3 实验结果与分析

4.3.1 表示学习维度对照实验

当采用词向量表示文本时，基本原理上是词向量的维度越大效果越好，但完成具体任务时需要达到运算速度和情感分析效果的平衡，因此进行表示学习维度对照实验，选择能达到较好情感分析效果的情感词典表示维度。对照实验中的TFIDF方法和本文的MFDSL算法在情感词典表示维度分别为20维、50维、100维、120维、150维、200维和300维时，在中文实验语料上的情感分类十折交叉验证实验的精度、召回率、F1值和准确率如表2和图2所示，当情感词典表示维度从20维增长到100维，情感分析准确率和各项指标提升明显，而增长到100维之后，准确率提升就比较少并趋于平稳，因此本文选取情感词向量的表示维度为100维。

Table 2: 表示学习维度对照实验

维度	生成方法	macro precision	macro recall	macro f1	accuracy
20	TF-IDF	81.16%	81.07%	81.07%	81.08%
	MFDSL	81.94%	81.90%	81.90%	81.91%
50	TF-IDF	82.68%	82.60%	82.60%	82.61%
	MFDSL	83.95%	83.90%	83.90%	83.91%
100	TF-IDF	83.78%	83.73%	83.72%	83.74%
	MFDSL	84.10%	84.05%	84.05%	84.06%
120	TF-IDF	83.86%	83.80%	83.80%	83.82%
	MFDSL	84.11%	84.06%	84.06%	84.07%
150	TF-IDF	84.09%	84.03%	84.04%	84.05%
	MFDSL	84.18%	84.13%	84.13%	84.14%
200	TF-IDF	84.13%	84.07%	84.08%	84.09%
	MFDSL	84.20%	84.15%	84.15%	84.16%
300	TF-IDF	84.12%	84.07%	84.07%	84.08%
	MFDSL	84.22%	84.17%	84.17%	84.18%

4.3.2 中文领域情感词典对照实验

为验证本文提出的情感词典构建方法，将生成的情感词向量MFDSL与4.2中的情感本体方法、Word2Vec方法、Fasttext方法、TF-IDF方法以及文献[22]中的TF-IDF-Senti2vec方法在中文酒店领域评论上做情感分类实验，设定领域情感词向量MFDSL维度为100维，十折交叉验证结果如表3所示：

从表3可以看出，本文提出的多源知识融合领域情感词典表示学习方法MFDSL在中文领

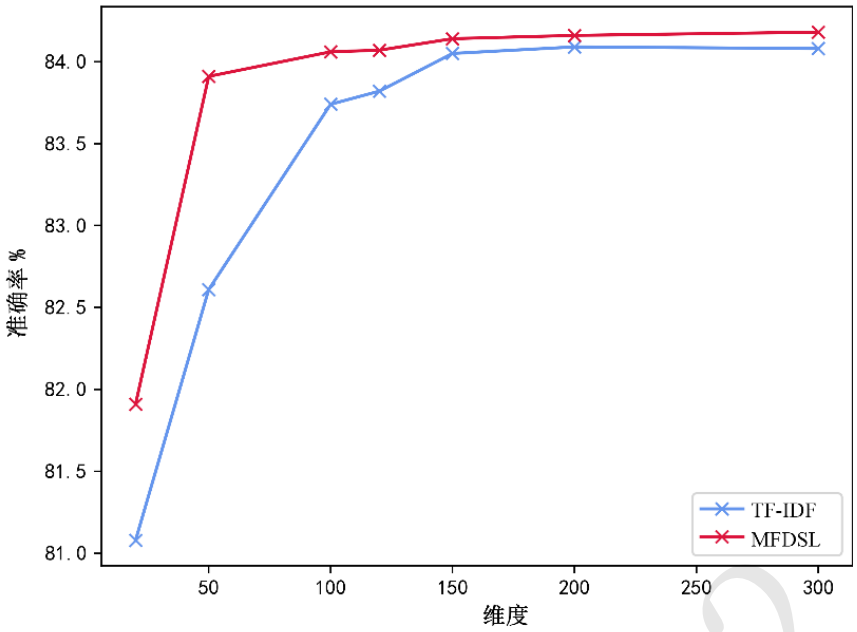

Figure 2: 表示学习维度-准确率变化关系图

Table 3: 中文情感分类对照实验结果

对照方法	macro precision	macro recall	macro f1	accuracy
情感本体-SVM	76.46%	75.21%	74.94%	75.25%
Word2vec-SVM	83.24%	83.21%	83.22%	83.24%
Fasttext-SVM	83.57%	83.55%	83.54%	83.55%
TF-IDF-Senti2vec-SVM	–	–	–	62.97%
TF-IDF-SVM	83.18%	83.10%	83.09%	83.10%
MFDSL-SVM	84.05%	83.99%	83.98%	83.99%
Bert-BiLSTM	89.15%	88.23%	87.96%	88.07%
MFDSL-Bert-BiLSTM	90.94%	90.84%	90.80%	90.82%

域情感词典应用于情感分类的对照实验中取得了比情感本体、词向量方法和TFIDF类方法更高的准确率、召回率、F1值和正确率。其中，单纯采用词向量的方法正确率高于情感本体方法，证明了词向量在语义表示上的优势。文献[22]通过TFIDF和Word2vec结合表示情感词典，效果并不理想，采用SVM分类算法时本文模型正确率比文献[22]方法提高了20.13%，证明了领域对比方法的有效性。采用SVM分类算法时，本文MFDSL结合领域对比方法和Fasttext表示学习后，比较Fasttext方法正确率提高了0.44%，比Word2vec词向量提高了0.75%，比TF-IDF方法提高了0.89%，比情感本体提高了8.74%。在与预训练语言模型Bert的对比中，增加了本文提出的MFDSL模型之后，正确率提高了2.75%，进一步证明了多源知识融合表示学习方法的有效性。

为探究领域对比方法的有效性，实验进一步比较了引入领域对比语料前后的情感词表，如表4所示，引入领域对比语料能够抽取低词频但具有领域代表性的情感词汇，正向情感词汇增加了62个，负向情感词汇增加69个，分别占引入领域对比领域前的13%和16%，更好地实现了领域情感词典构建的目标。

4.3.3 英文领域情感词典对照实验

为验证本文方法在多领域、多语言环境中的有效性，本节选取Blitzer收集整理Amazon图

Table 4: 引入中文领域对比语料前后的情感词表

	极性	数量	示例
引入领域对比语料前	正向	479	...整洁，豪华，宽敞，价廉物美，诚挚...
	负向	427	...最差，不足之处，简陋，大失所望，美中不足...
引入领域对比语料后	正向	541	...宾至如归，方便，便利，便宜，没得说...
	负向	496	...轰鸣声，坑坑洼洼，偏僻，形同虚设，置若罔闻...

书、DVD、电子产品、厨房用品和影像五个领域的英文评论语料，将本文MFDSL方法与情感本体、Word2Vec、Fasttext、TF-IDF和Bert预训练语言模型情感分类对照实验，设定领域情感词向量MFDSL维度为100维，十折交叉验证的平均正确率如表5所示，在五个领域的英文语料上，本文方法均取得了最高的正确率，比Bert预训练语言模型在五个领域分别提升了1.21%,2.2%,0.78%,1.94%和5.51%，验证了其在多语言、多领域环境中的鲁棒性。此外，实验结果表明，对比本文方法与Fasttext、Word2vec词向量在英文数据集上的正确率提升，比在中文数据集上的提升更为明显，原因是中文词向量辅助语料规模更大，并从微博评论中获得了更通用的语义知识。值得注意的是，本文提出的MFDSL领域情感词向量表示方法，需要的标注数据规模小，不仅适用于深度学习模型BiLSTM，还适用于时间复杂度较低的SVM算法，在不同的算法上也具有较好的鲁棒性。

Table 5: 英文情感分类对照实验结果

对照方法accuracy	图书	DVD	电子产品	厨房用品	影像
情感本体-SVM	70.07%	71.00%	69.80%	71.60%	75.60%
Word2vec-SVM	74.71%	76.80%	75.26%	76.70%	76.90%
Fasttext-SVM	74.81%	75.53%	74.19%	76.25%	77.26%
TF-IDF-SVM	73.42%	74.88%	74.28%	75.21%	77.12%
MFDSL-SVM	75.22%	77.57%	77.80%	78.29%	77.82%
Bert-BiLSTM	76.23%	75.74%	77.45%	79.84%	76.23%
MFDSL-Bert-BiLSTM	77.44%	77.94%	78.23%	81.78%	81.74%

为进一步探究领域对比方法的在英文语料上有效性，选取英文图书评论领域比较引入领域对比语料前后的情感词表，如表6所示。可以看出，引入领域对比语料后能够抽取诸如“gastronomic”、“machiavellian”的低频英文情感词，正向情感词汇增加了75个，负向情感词汇增加80个，并有效改进了情感分类效果。

Table 6: 引入英文图书领域对比语料前后的情感词表

	极性	数量	示例
引入领域对比语料前	正向	1804	...great，excellent，wonderful，new，easy...
	负向	1842	...bad，boring，worst，disappointing，confusing...
引入领域对比语料后	正向	1879	...gastronomic，decorative，suggestive，impressively，unforgettably...
	负向	1922	...machiavellian，sissy，unitarian，regretful，musty...

4.3.4 消融实验

本文基于多源知识融合的领域情感词典表示学习模型在词向量的基础上，主要包括领域对比模块、Tfidf模块和情感本体模块。为验证本文模型各个模块的作用，分别进行了中英文语料上的消融实验，如表7和表8所示：

从表7和表8可以看出，总体上在中英文各领域数据集上，仅采用领域对比模块、Tfidf模块和情感本体模块都比整体MFDSL模型的正确率有所下降，当三个模块都移除，仅采用Bert预训

Table 7: 中文情感分类消融实验

对照方法	macro precision	macro recall	macro f1	accuracy
MFDSL-Bert-BiLSTM	90.94%	90.84%	90.80%	90.82%
领域对比-Bert-BiLSTM	90.49%	90.33%	90.21%	90.22%
tfidf-BERT-BiLSTM	90.44%	90.41%	90.33%	90.34%
情感本体-BERT-BiLSTM	90.28%	89.93%	89.79%	89.82%
Bert-BiLSTM	89.15%	88.23%	87.96%	88.07%

Table 8: 英文情感分类消融实验结果

对照方法accuracy	图书	DVD	电子产品	厨房用品	影像
MFDSL-Bert-BiLSTM	77.44%	77.94%	78.23%	81.78%	81.74%
领域对比-Bert-BiLSTM	76.64%	75.82%	77.26%	79.03%	80.63%
tfidf-BERT-BiLSTM	75.91%	75.49%	77.23%	79.56%	79.54%
情感本体-BERT-BiLSTM	75.71%	75.20%	76.69%	79.79%	80.62%
Bert-BiLSTM	76.23%	75.74%	77.45%	79.84%	76.23%

练语言模型和BiLSTM分类算法时，正确率降到最低，证实了本文模型中三个模块的有效性。但从实验结果没有明显趋势表示具体哪一个模块在整体性能中贡献最大，正确率主要取决于各个模块之间的交互效果。

5 总结与展望

本文提出基于多源知识融合的领域情感词典表示学习方法自动从无标注数据中构建适应大数据多领域和多语言环境的领域情感词典，引入外部领域对比语料强化领域相关词的权重，融合多源数据语义信息和情感信息弥补先验知识的不足，通过表示学习将词向量和情感权重映射到新的情感语义空间更准确地表示情感语义。在中英文六个领域公开数据集上的对照实验结果表明该模型有效提高了情感词典在情感分类中的有效性，进一步的分析验证了本文方法能够有效抽取其他方法难以自动提取的低频领域情感词。在未来的工作中，将在多领域大规模语料中进一步检验模型的泛化性，进一步探究隐性情感词汇的自动抽取方法和检验标准。

参考文献

王科,夏睿.情感词典自动构建方法综述[J].自动化学报,2016,42(4): 495-511.

Westgate A, Valova I. A Graph Based Approach to Sentiment Lexicon Expansion[C]. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 2018: 530-541.

SAĞLAM F, GENÇ B, SEVER H. Extending a sentiment lexicon with synonym–antonym datasets: SWNetTR++[J]. Turkish Journal of Electrical Engineering & Computer Sciences, 2019, 27(3): 1806-1820.

Shaukat K, Hameed I A, Luo S, et al. Domain Specific Lexicon Generation through Sentiment Analysis[J]. International Journal of Emerging Technologies in Learning, 2020, 15(9).

贺飞艳,何炎祥,刘楠,刘健博,彭敏.面向微博短文本的细粒度情感特征抽取方法[J].北京大学学报(自然科学版),2014,50(01):48-54.

Turney P D, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems (TOIS), 2003, 21(4): 315-346.

Mullen T, Collier N. Sentiment analysis using support vector machines with diverse information sources[C].Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 412-418.

Liu J, Yan M, Luo J. Research on the construction of sentiment lexicon based on Chinese microblog[C]. In: International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, 2016:56-59.

Qiu G, Liu B, Bu J, et al. Expanding domain sentiment lexicon through double propagation[C].IJCAI. 2009, 9: 1199-1204.

Wu S, Wu F, Chang Y, et al. Automatic construction of target-specific sentiment lexicon[J]. Expert Systems with Applications, 2019, 116: 285-298.

Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text[C].Proceedings of the International AAAI Conference on Web and Social Media. 2014, 8(1).

Li W, Guo K, Shi Y, et al. DWWP: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain[J]. Knowledge-Based Systems, 2018, 146: 203-214.

杨小平,张中夏,王良,张永俊,马奇凤,吴佳楠,张悦.基于Word2Vec的情感词典自动构建与优化[J].计算机科学,2017,44(01):42-47+74.

张璞,王俊霞,王英豪.基于标签传播的情感词典构建方法[J].计算机工程,2018,44(05):168-173.

蒋翠清,郭轶博,刘尧.基于中文社交媒体文本的领域情感词典构建方法研究[J].数据分析与知识发现,2019,3(02):98-107.

Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

谭松波. 数据集:谭松波–酒店评论语料[DB/OL].[2020-02-08].https://blog.csdn.net/LiuKingJia/article/details/104228617.

张华平. NLPIR微博内容语料库-23万条[DB/OL].[2017-12-03].http://www.nlpir.org/wordpress/download/weibo_content_

SMP2020-EWECT.SMP2020微博情绪分类评测[DB/OL].[2020-06-19].https://smp2020ewect.github.io/.

Amazon Review Data (2018).Jianmo NiDB/OL].[2022-02-17]. https://nijianmo.github.io/amazon/index.html#complete-data.

Aashita Kesarwani. New York Times Comments.[DB/OL]. [2018]. https://www.kaggle.com/aashita/nyt-comments.

林江豪,周咏梅,阳爱民,陈锦.基于词向量的领域情感词典构建[J].山东大学学报(工学版),2018,48(03):40-47.