计算机科学
COMPUTER SCIENCE

基于改进TF-IDF与BERT的领域情感词典构建方法

蒋昊达, 赵春蕾, 陈瀚, 王春东

引用本文

蒋昊达, 赵春蕾, 陈瀚, 王春东. 基于改进TF-IDF与BERT的领域情感词典构建方法[J]. 计算机科学, 2024, 51(6A): 230800011-9.

JIANG Haoda, ZHAO Chunlei, CHEN Han, WANG Chundong. Construction Method of Domain Sentiment Lexicon Based on Improved TF-IDF and BERT [J]. Computer Science, 2024, 51(6A): 230800011-9.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230800011

基于改进 TF-IDF 与 BERT 的领域情感词典构建方法

蒋昊达　赵春蕾　陈　瀚　王春东
1 天津理工大学教育部计算机视觉与系统省部共建重点实验室　天津 300384
2 天津市智能计算与软件新技术重点实验室　天津 300384
(jhd1026@163.com)

摘　要　领域情感词典的构建是领域文本情感分析的基础。现有的领域情感词典构建方法存在所筛选候选情感词冗余度高、情感极性判断失准、领域依赖性强等问题。为了提高所筛选候选情感词的领域性和判断领域情感词极性的准确程度,提出了一种基于改进词频-逆文档频率(TF-IDF)与 BERT 的领域情感词典构建方法。该方法在筛选领域候选情感词阶段对 TF-IDF 算法进行改进,将隐含狄利克雷分布(LDA)算法与改进后的 TF-IDF 算法结合,进行领域性修正,提升了所筛选候选情感词的领域性;在候选情感词极性判断阶段,将情感倾向点互信息算法(SO-PMI)与 BERT 结合,利用领域情感词微调 BERT 分类模型,提高了判断领域候选情感词情感极性的准确程度。在不同领域的用户评论数据集上进行实验,结果表明,该方法可以提高所构建领域情感词典的质量,使用该方法构建的领域情感词典用于汽车领域和手机领域文本情感分析的 F1 值分别达到 78.02%和 88.35%。

关键词:情感分析;领域情感词典;词频-逆文档频率;隐含狄利克雷分布;情感倾向点互信息算法;BERT 模型
中图分类号　TP391.1

Construction Method of Domain Sentiment Lexicon Based on Improved TF-IDF and BERT

JIANG Haoda,ZHAO Chunlei,CHEN Han and WANG Chundong
1 Key Laboratory of Computer Vision and System of Ministry of Education,Tianjin University of Technology,Tianjin 300384,China
2 Tianjin Key Laboratory of Intelligent Computing and Novel Software Technology,Tianjin 300384,China

Abstract　The construction of a domain sentiment lexicon is the foundation of domain text sentiment analysis. The existing methods for constructing domain sentiment lexicon have problems such as high redundancy of selected candidate sentiment words, inaccurate judgment of sentiment polarity, and high domain dependency. In order to improve the domain specificity of selected candidate sentiment words and the accuracy of judging the polarity of domain sentiment words, a domain sentiment lexicon construction method based on improved term frequency-inverse document frequency(TF-IDF) and BERT is proposed. This method improves the TF-IDF algorithm in the phase of selecting domain candidate sentiment words. The latent dirichlet allocation(LDA) algorithm is combined with the improved TF-IDF algorithm to perform domain corrections, improves the domain specificity of the selected candidate sentiment words. In the polarity judgment stage of candidate sentiment words, the semantic orientation pointwise mutual information(SO-PMI) algorithm is combined with BERT. By fine-tuning the BERT classification model using domain sentiment words, the accuracy of judging the sentiment polarity of domain candidate sentiment words is improved. Experiments are conducted on user comment datasets in different domains, and the experimental results show that this method can improve the quality of the constructed domain sentiment lexicon, and the F1 value of the domain sentiment lexicon constructed by this method for text sentiment analysis in the automotive field and mobile phone field reaches 78.02% and 88.35%, respectively.

Keywords　Sentiment analysis, Domain sentiment lexicon, Term Frequency-Inverse Document Frequency (TF-IDF), Latent Dirichlet allocation(LDA), Semantic orientation pointwise mutual information(SO-PMI), BERT model

1　引言

随着互联网的高速发展和社交媒体的兴起,在网上分享自己的观点已经成为人们日常生活中的一部分,微博评论、电影评论、产品评论等评论数据的数量每天都在以指数级的速度增长,这些评论能够为日常的很多决策提供参考。文本情感分析是从评论文本中挖掘情感和意见的一种重要方法[1],深度学习模型在情感分析中的应用愈发广泛,同时也面临着一些问题,如数据稀缺性、领域适应性和对情感信息的理解不足。

情感词典作为深度学习模型重要的辅助资源,在情感分析任务中经常被使用[2],将情感词典与深度学习模型相结合,

基金项目:国家重点研发计划"科技助力经济 2020"重点专项项目(SQ2020YFF0413781,SQ2020YFF0401503)
This work was supported by the Key Special Project of "Science and Technology Helps Economy 2020" of National Key R&D Program of China (SQ2020YFF0413781,SQ2020YFF0401503).
通信作者:赵春蕾(zcltjut@126.com)

不仅可以为模型提供情感标签的基准信息，还能为特定领域的情感分析任务提供准确的领域相关情感信息。

目前已经存在一些通用的中文情感词典，比较有代表性的有知网的 HowNet 情感词典[3]、清华大学李军中文褒贬义词典[4]、台湾大学 NTUSD 简体中文情感词典[5]、BosonNLP 情感词典[6]等。研究表明，与传统的情感词典相比，特定领域的情感词典可以提高情感分析的效果[7]，通用的情感词典难以覆盖特定领域中的情感表达。构建针对特定领域的情感词典，可以使深度学习模型更好地适应特定领域的情感分析任务。

本文旨在研究如何提高在领域语料库中自动构建的领域情感词典的质量。本文的主要工作如下：首先改进词频-逆文档频率（Term Frequency-Inverse Document Frequency，TF-IDF）算法，并将其与隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）算法结合以筛选候选情感词，然后将情感倾向点互信息（Semantic Orientation-Pointwise Mutual Information，SO-PMI）算法与 BERT（Bidirectional Encoder Representations from Transformer）分类模型结合判断候选情感词的情感极性，最后得到领域情感词典。

2 相关工作

领域情感词典在情感分析中具有重要作用[8]，其会对不同层次情感分析的效果产生直接影响。所以，构建领域情感词典的方法也尤其重要。目前的情感词典构建方法可以分为两类，一类是基于知识库扩展的方法，另一类是基于语料库的方法。

基于知识库扩展的方法使用带有情感极性标注的种子词集，利用知识库里词语的同义、反义、上下位、词语和义原的关联关系等来判断待判定词的情感极性。如 Neviarouskaya 等[9]提出利用已有词库的词作为种子词，与现有的知识库 WordNet 的语义关系扩充词典的方法；Hassan 等[10]提出随机行走模型，通过计算词语与已标注种子词的移动步数来判断词语的情感极性；Dragut 等[11]提出一种解决同一词语在不同词典极性下不同的自动扩展情感词典的方法；Zhu 等[12]提出计算 Hownet 中的情感词汇与候选情感词的语义相似度，从而判断词语情感类别的方法。这类方法非常依赖知识库的完整性，可扩展性较差。

基于语料库的方法是通过文本语料库中不同词语之间的共现信息、上下文约束关系、词向量关系等来构建情感词典。

利用词语共现关系的方法如 Bollegala 等[13]在语料库中根据词性提取候选词，利用点互信息算法（Pointwise Mutual Information，PMI）计算候选词与已知情感词的相关性来确定候选词的情感倾向，从而构建了情感词典；Krestel 等[14]在大型语料库中使用 LDA 提取情感词，并使用 PMI 算法判断词语的情感极性，从而构建情感词典；Deng 等[15]利用词语的先前情感信息捕捉不同主题下词语的情感属性，构建主题自适应领域情感词典；Wang 等[16]利用改进的 TF-IDF 算法计算不同词性词语的情感值，自动构建特定领域情感词典；Zhao 等[17]使用 TF-IDF 算法选择情感种子词，再利用 SO-PMI 算法对原始词集进行扩展，从而构建高等教育领域情感词典；Wang 等[18]构建细粒度种子情感词，使用同义词集对种子词扩展，构建细粒度领域情感词典；Ren 等[19]基于轻量级梯度提升机器学习（Light Gradient Boosting Machine，LightGBM）筛选候选词，使用 PMI 算法判断情感极性，构建电动汽车拆卸领域情感词典。以上方法容易出现部分词语由于共现频率不够而无法判断情感极性的问题，影响情感词典的完整性。

利用上下文约束关系的方法如 Huang 等[20]从领域语料库中定义并提取情感词之间的上下文约束关系，利用标签传播算法（Label Propagation Algorithm，LPA）构建领域情感词典；Xi[21]使用基于约束的 LPA 算法，考虑情感词之间的点互信息的同时，也考虑了上下文约束关系，构造了基本情感词典，并根据上下文信息识别领域情感词，构建领域情感词典；Li 等[22]利用正向无标记学习（Positive and Unlabeled Learning，PU Learning）筛选情感词，使用 LPA 算法判断候选词情感极性。这类方法虽然考虑了情感词的上下文关系，但对词间关系的选择要求较高。

利用词向量关系的方法如 Yang 等[23]基于 Word2Vec 在语料库中提取词向量，并选取种子词，通过计算词语与种子词间的语义距离判断其情感倾向；Zhang 等[24]使用 Word2Vec 词向量选取候选情感词，通过标签传播算法计算情感词的情感倾向，最终构建情感词典；Yang 等[25]利用 TextRank 和 Word2Vec 模型对在线课程评论进行情感词提取，构建网络课程评论情感词典；Ye 等[26]利用连续词袋模型（Continuous Bag of Words，CBOW）词向量模型和句法规则选取候选情感词，使用改进的 SO-PMI 算法判断情感词极性。以上方法对种子词的选取依赖性较强。

现有的领域情感词典构建方法主要存在两个问题，一方面，在筛选候选情感词阶段方法单一，使较多与本领域相关性较低的冗余词被选入情感词典中，从而对相关领域文本的情感分析造成干扰；另一方面，很多研究在情感极性判断阶段仅使用点互信息或余弦相似度等方法，需要依赖人工选取种子词，对种子词的选取精度要求较高，而且容易出现部分词语由于共现频率不够而无法判断情感极性的问题，造成候选情感词典极性判断失准，从而影响所构建情感词典的质量。

针对以上问题，本文提出了一种基于改进 TF-IDF 与 BERT 的领域情感词典构建方法。本文的主要贡献如下：

1）在筛选候选情感词阶段，提出了一种改进的 TF-IDF 算法，并与 LDA 算法结合，减少了情感词的冗余。

2）在情感词的情感极性判断阶段，引入 BERT 分类模型，不使用单一的 SO-PMI 算法，而是将 SO-PMI 算法与 BERT 分类模型相结合，使用 SO-PMI 算法在语料中抽取领域性情感词集，微调 BERT 分类模型后对候选情感词进行情感极性预测，使情感极性判断更准确。

3）构建了汽车领域和手机领域用户评论数据集，在两个领域数据集上进行实验，验证了本文方法的有效性和普适性。

3 基于改进 TF-IDF 与 BERT 的领域情感词典构建方法

本文提出的基于改进 TF-IDF 与 BERT 的领域情感词典构建方法方法分为 3 个部分，分别为用户评论文本预处理、基于改进 TF-IDF 算法与 LDA 算法结合的候选情感词筛选、基于 SO-PMI 与 BERT 结合的候选情感词极性判断。方法框架图如图 1 所示。

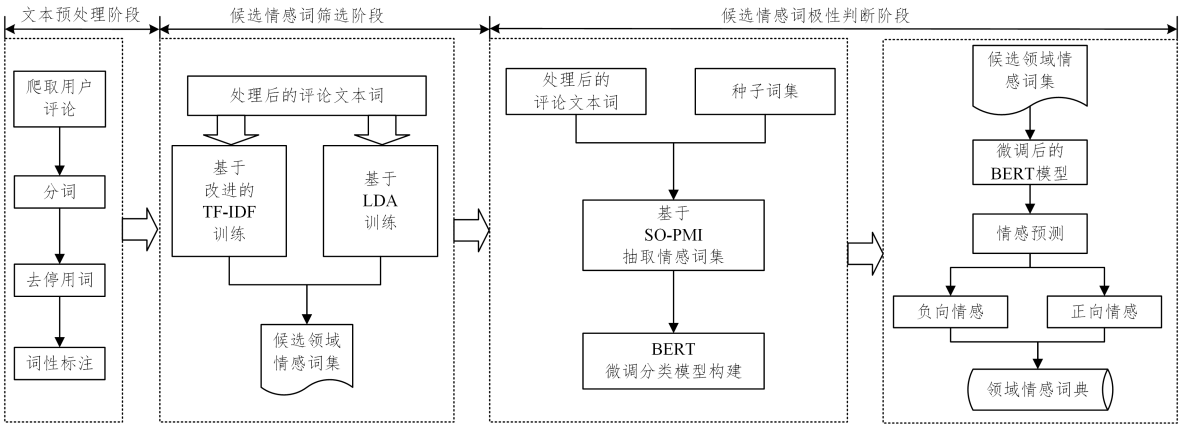

图 1　基于改进 TF-IDF 与 BERT 的领域情感词典构建方法框架

Fig. 1　Framework for constructing domain sentiment lexicon based on improved TF-IDF and BERT

1)用户评论文本预处理。该部分的主要内容是，在网络上爬取用户评论，构建领域数据集，并将评论文本进行分词、去除停用词、词性标注等预处理工作，为数据集更好地应用于领域情感词典的构建做好准备。

2)基于改进 TF-IDF 算法与 LDA 算法结合的候选情感词筛选。该部分的主要内容为对领域情感词典中的情感词进行筛选，对 TF-IDF 算法进行改进，根据不同词的词性赋予权重，将 TF-IDF 值与词性权重结合，得到带有词性特征的新权值；以新权值为依据设置阈值筛选，得到改进的 TF-IDF 词集，使用 LDA 算法对评论数据集进行主题词提取，根据词频、词性进行筛选，得到 LDA 词集，对改进的 TF-IDF 词集进行领域性修正，得到最终候选情感词。

3)基于 SO-PMI 与 BERT 结合的候选情感词极性判断。该部分的主要内容是对筛选出来的候选情感词进行情感极性判断。选取适当的种子词，利用 SO-PMI 算法从评论数据集中抽取领域特性词语，并将 SO-PMI 值作为其情感值，将这些词语输入构建的 BERT 分类模型进行领域特性微调，使用微调后的 BERT 分类模型对筛选出的候选情感词进行情感预测，生成领域情感词典。

3.1　评论数据集预处理

对于文本的预处理一般包括分词、去除停用词、词性标注、去除空行等操作。网络在线评论文本具有规范性差、随意性强、口语化严重等特点，所以容易出现很多对于研究没有意义的词汇，应该在研究开始之前，将评论文本进行分词，然后将无意义的词汇予以去除。由于 jieba 分词具有速度快、可扩展等优势，因此本文使用 jieba 作为分词工具。然而 jieba 自带的词典具有通用性，对于特定领域的一些词语分词不够准确，所以本文根据领域评论文本的特点建立自定义分词词库，先利用 jieba 自带的词库结合自定义的词库对评论文本进行分词，再利用自建的停用词表，去除分词后的评论文本中的停用词，然后利用 jieba 的词性标注模式标注词语的词性。

3.2　候选情感词筛选

3.2.1　基于改进 TF-IDF 算法的候选情感词筛选

TF-IDF[27]算法只考虑特征词在一个文档和其他文档中出现的频率，并不完全适用于候选领域情感词的筛选，领域情感词往往词性特征较明显，一般多为名词、动词、形容词，比如车领域中的“费油”为形容词，旅游领域中的“世外桃源”为名词，“逃票”为动词，图书领域中的“鸡汤文”为名词，“烂尾”为形容词，“补记”为动词。另一方面，由于传统的情感词典缺乏时效性强的网络词汇，如给力、辣鸡、绝绝子等，所以有必要在筛选候选领域情感词的阶段，将文本中的网络词汇也考虑进去。本文采用人工选取的方式从搜狗拼音词库网络流行新词中抽取网络词语加入到 jieba 自定义分词词典中，并将网络词单独作为一种词性，与名词、动词、形容词 3 种词性共同作为领域候选情感词的一个特征。

本文提出了一种适用于领域情感词典构建的改进 TF-IDF 算法，即 TF-IDF-POS 算法，在 TF-IDF 算法的基础上根据词语或短语的词性赋予其不同的权重，具体定义如式(1)所示：

$$t_{ij}=tf_{ij}\times\log\frac{N}{n_j}\times w_{posj} \tag{1}$$

其中，tf_{ij} 表示在文本 i 中词语或短语 j 的词语频率，N 表示数据集中文本的总数，n_j 表示文本数据集中包含词语 j 的文本总数，w_{posj} 表示词语或短语 j 的词性的权重，t_{ij} 表示词语或短语 j 在文本 i 中的权重。

TF-IDF-POS 算法通过引入文本中词语或短语词性的权重，可以在筛选候选情感词阶段，提高具有领域特性的情感词的权重，通过设定一定的阈值，可以减少对与领域无关的词语的误选。本文对预处理之后的词语数据集进行 TF-IDF-POS 训练，待训练的词语数据集用词集 W 表示，$W=\{w_i\}$，$i\in[1,\cdots,n]$，n 为词集中的词语数量，训练得到带有权值的词集表示为 t_j {$(key_1, weight_1)\cdots(key_j, weight_j)\cdots(key_n, weight_n)$}，其中 key_j 表示第 j 个候选词，$weight_j$ 表示第 j 个候选词的权值。对 t_j 按照权值大小进行排序，根据设定的阈值选取权值较大的词构成候选词集，表示为 $top\text{-}t_j$ {$(key_1, weight_1)\cdots(key_j, weight_j)\cdots(key_n, weight_n)$}。

3.2.2　基于 LDA 主题模型的候选情感词筛选

LDA[28]可以挖掘文本中潜在的语义信息，并通过一组词语对主题进行表征，而对于领域特性较强的评论文本而言，其蕴含的主题信息往往由该领域的情感词表示，因此，使用 LDA 主题模型对用户评论文本提取的词频较高的主题词，可以作为该领域的候选情感词。本文对预处理之后的词语数据集 W 进行 LDA 训练，训练得到的词集表示为 l_j {$(key_1, frequency_1)\cdots(key_j, frequency_j)\cdots(key_n, frequency_n)$}，其中，$key_j$ 表示第 j 个候选词，$frequency_j$ 表示第 j 个候选词的词频。对 l_j 按照词频大小进行排序，根据设定的阈值选取词频

较高的词构成候选词集，表示为 $top\text{-}l_j\{(key_1, frequency_1)\cdots(key_j, frequency_j)\cdots(key_n, frequency_n)\}$。

本文对预处理后的词语数据集先进行 TF-IDF-POS 训练，从语料库中提取领域性较强的情感词。由于 TF-IDF-POS 算法本身的特性，仍然会有非领域性的词语被选中，导致词集冗余，因此本文将文本数据集再进行 LDA 训练，将 $top\text{-}t_j$ 和 $top\text{-}l_j$ 两个训练得到的结果词集进行合并去重，得到最终的候选情感词集，表示为 $Candiword\{word_1\cdots word_n\}$。通过 LDA 词集对 TF-IDF-POS 词集进行领域性修正，可以降低词集的冗余，使得到的最终候选情感词集更具有领域代表性，能够更加准确地反映领域情感特征。

3.3 基于 SO-PMI 与 BERT 结合的候选情感词极性判断

3.3.1 基于 SO-PMI 的情感词集抽取

本文选取传统情感词典的词汇作为种子词，生成的种子词集用 Z 表示，其中包括积极种子词集 $P=\{p_i\}, i\in[1,\cdots,n]$，消极种子词集 $N=\{n_i\}, i\in[1,\cdots,m]$。预处理后的词语数据集用词集 V 表示，对于 V 中的每个词 $v_i(i=1,2,3,\cdots,m)$，利用式(2)计算其 SO-PMI 值：

$$SO\text{-}PMI(a)=\sum_{i=1}^{num(pos)}\log\frac{p(a,pos_i)}{p(a)p(pos_i)}-\sum_{i=1}^{num(neg)}\log\frac{p(a,neg_i)}{p(a)p(neg_i)} \tag{2}$$

其中，a 代表候选情感词，pos_i 代表第 i 个积极种子词，neg_i 代表第 i 个消极种子词，$num(pos)$ 代表积极种子词的总数量。$num(neg)$ 代表消极种子词的总数量，$SO\text{-}PMI(a)$ 代表候选情感词的 SO-PMI 值，如果 $SO\text{-}PMI(a)>0$，说明候选情感词 a 为正向情感词，如果 $SO\text{-}PMI(a)<0$，说明候选情感词 a 为负向情感词。

依据 SO-PMI 值的大小确定词集 V 中每个词的情感极性，得到带有极性的情感词集 S。由于特定领域语料数据集中的词语具有一定的领域特性，因此将情感词集 S 作为对 BERT 分类模型微调的数据集。

3.3.2 BERT 分类模型构建

本文将 BERT[29] 引入领域情感词典构建中，构建的词级别 BERT 微调分类模型如图 2 所示，先将输入的文本进行预处理，在文本序列的首位添加[CLS]标记，表示一个文本，将情感词集 S 中的词语文本输入模型后，词嵌入层将该文本标记化，包括词和子词的嵌入向量、词位置嵌入向量。其中，Tok_i 表示文本的第 i 个标记，E_i 表示第 i 个标记的嵌入向量，T_i 表示第 i 个标记在模型处理后生成的最终特征向量。本文选用 HuggingFace Transformers 库中的中文预训练模型 Bert-base-Chinese，在输出层后添加一个分类器层，分类器层中包含一个全连接层，通过 softmax 函数将分类器层的输出转换为两类标签的概率分布，将概率最大的标签作为模型分类的结果，使用 SO-PMI 算法生成的词集 S 对分类模型进行微调，使分类模型适应特定领域的分类任务。本文使用微调之后的 BERT 分类模型对 3.2 节得到的候选领域情感词集 Candiword 进行情感极性预测，得到带有情感极性的领域情感词集 $Senti\{(word_1, polarity_1)\cdots(word_j, polarity_j)\cdots(word_n, polarity_n)\}$，其中，$word_j$ 表示第 j 个情感词，$polarity_j$ 表示第 j 个情感词的情感极性，$Senti$ 即为最终的领域情感词典。

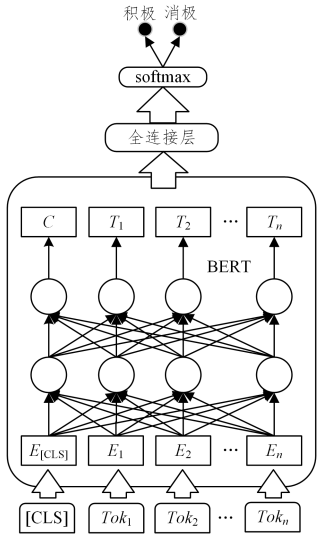

图 2 BERT 微调分类模型

Fig. 2 BERT fine-tuning classification model

本文在情感极性判断任务中将 SO-PMI 算法和 BERT 模型结合。由于 BERT 预训练模型具有双向处理单位词的能力，因此本文通过自定义参数使用从语料数据集中抽取的领域词进行模型微调，训练出适合于本领域的分类模型，解决了单一使用 SO-PMI 算法出现的部分词语由于共现频率不够而无法判断情感极性以及种子词选取精度要求较高的问题，通过提高对候选领域情感词的极性判断的准确率，提高了所构建领域情感词典的质量，降低了领域性低的词语对领域情感分类的消极影响。

3.4 算法描述

领域情感词典构建算法描述如算法 1 所示。

算法 1 领域情感词典构建算法

```
输入：特定领域评论词集 W，V，种子词集 Z
输出：特定领域情感词典 Senti
1. 语料预处理
2. T＝TF-IDF-POS(W)
3. L＝LDA(W)
4. for i in T do
5.     if i in L do
6.         Candiword. append(i)
7.     end
8. end
9. for i in V do
10.   for z in Z do
11.     if i. tag 存在 do
12.       i. tag＝SO-PMI(z,i)
13.       S. append(i)
14.     end
15.   end
16. end
17. for i in S do
18.   optimize BERT(i)
19. end
20. for i in Candiword do
21.   i. tag＝BERT(i)
22.   Senti. append(i)
23. end
```

其中 $Candiword$ 表示候选情感词集，S 表示基于 SO-

PMI 算法的情感词集抽取得到的情感词集。通过上述算法最终得到领域情感词典 Senti。

4 实验与分析

4.1 实验数据

本文选取汽车领域和手机领域两个不同领域的数据集进行实验，评论数据均取自一些用户评论网站和购物平台网站。由于太平洋汽车网和京东商城的用户评论较为丰富和完整，因此本文选择了以上两个平台的用户评论作为实验数据。语料文本样例如表 1 所列。

首先使用八爪鱼收集器在太平洋汽车网上爬取多款汽车的用户评论，其中包括用户对于该款汽车的优点和缺点的评论。通过去重降噪等处理，将汽车优点作为正向评论，将汽车缺点作为负向评论，并结合人工标注调整，得到包括 6000 余条、86000 余条、18000 余条评论数据的 3 组汽车领域数据集，分别记为数据集 A、数据集 B、数据集 C，其中正负向评论数量的比例为 1∶1。

同理，使用八爪鱼收集器在京东商城上爬取多款手机的用户评论，将好评作为正向评论，差评作为负向评论，得到包括 10000 余条、100000 余条、20000 余条评论数据的 3 组手机领域数据集，分别记为数据集 A、数据集 B、数据集 C，其中正负向评论数量的比例为 1∶1。数据集 A 用于筛选领域候选情感词阶段实验，数据集 B 用于预测领域情感词极性阶段实验，数据集 C 用于情感分析对比实验。

在种子词的选取方面，由于清华大学情感词典和台湾大学情感词典所含情感词较为通俗和通用，所以本文选取清华大学情感词典和台湾大学情感词典合并后的情感词共 21119 个词作为种子词。

表 1 语料文本样例

Table 1 Samples of corpus text

序号	语料样本	
	汽车领域	手机领域
1	外观漂亮，跑高速底盘扎实，驾驶质感好；轴距不算太长，但空间却是足够的。	手感很棒，比较轻盈，待机时间挺长的，价格优惠，运行起来并不慢；很实用的一款机器。
2	动力强、加速快、操控舒适、宽敞的空间、低油耗、经济实惠，还有一键启动等新功能。	屏幕清晰流畅，对老年人友好，侧边指纹很方便，刷新速度挺快，颜色很漂亮。
3	底盘很稳，开起来也很稳，车身钣金很厚很安全。	屏幕素质一般，某些画面显示颗粒感很明显。
4	车子的座椅调整比较麻烦，我在开车途中得停下来打开车门，才能把靠背放下去点。车子的油耗有点高，其实我不是猛踩油门族，但是在同样排量的车里，这个有点小高。	手机的外表简洁、大气，整体做工也比较精细，特别是后盖上面的磨砂做得很漂亮；手机比较轻薄，拿在手里很舒服；屏幕很清晰，反应速度也比较快；功能齐全，音质很好；价格实惠，性价比很高。
5	发动机噪音大，座椅旋钮太搞笑了，旋钮样式的，一点一点拧，而且座椅真是太硬了，没有一点点包裹感。	反应慢，刚开机使用就发现这个问题了，屏幕显示有花屏现象，会有操作失灵现象，指纹在侧边开关机键的位置操作不方便。

4.2 实验设计与分析

本文实验包括利用本文提出的 TF-IDF-POS 算法与 LDA 算法结合的筛选领域候选情感词实验、利用本文提出的 SO-PMI 与 BERT 结合的方法判断候选领域情感词的情感极性实验，以及基于情感词典和规则的对于领域用户评论的情感分析实验，同时进行一系列的对比实验，以验证利用本文方法构建的领域情感词典的有效性。

4.2.1 筛选领域候选情感词实验

使用 jieba 分词工具，对数据集 A 进行分词、根据自建的停用词表去除停用词、去空行等预处理，得到词集。首先利用改进的 TF-IDF 算法即 TF-IDF-POS 算法对预处理得到的词集进行训练，计算 TF-IDF-POS 值。本文利用 jieba 词性标注模式将词集进行词性标注，并对领域特征明显的名词(n)、动词(v)、形容词(a)、网络词(w)等几个词性的词赋予词性权重，由于网络词(w)具有较强的时效性，对于领域情感词典来说重要程度较高，因此将其权重设为最大。在实验过程中发现，网络词的权重设置为 2 可以使所有网络词被选中，所以将网络词的权重设置为 2。由于本文的实验目标是使候选情感词集中领域词汇占比最大，减少冗余，所以本文通过多次实验确定名次、动词、形容词 3 个词性的权重。在筛选阈值一定的情况下，领域词汇在筛选后的词集中所占比例随名词、动词、形容词的权重设置变化情况如图 3－图 5 所示。

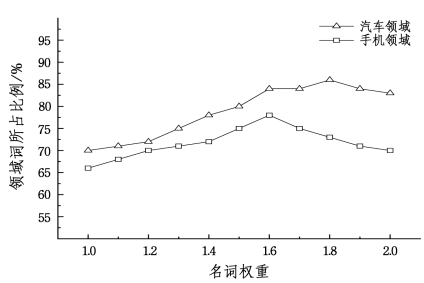

图 3 领域词占比随名词权重变化图

Fig. 3 Figure of the proportion of domain words changing with noun weights

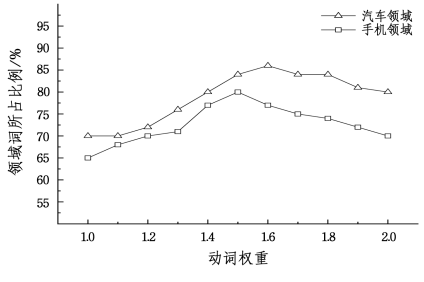

图 4 领域词占比随动词权重变化图

Fig. 4 Diagram of the proportion of domain words changing with verb weights

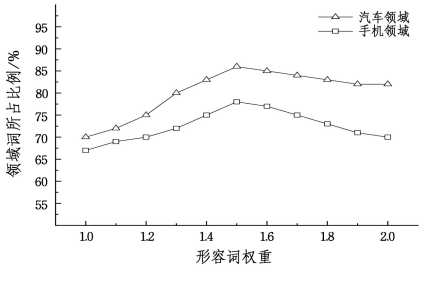

图 5 领域词占比随形容词权重变化图

Fig. 5 Diagram of the proportion of domain words changing with adjective weights

以名词为例，在汽车领域中，名词权重设为 1.8 时所筛选的候选情感词集中领域词汇占比最大，当名词权重大于 1.8 时，领域词汇占比反而下降。这是因为名词中也存在一些非领域性词汇，名词权重大于 1.8 时，一些非领域性词汇的权值也会增大，导致其被选中，增加了冗余。动词、形容词同理。词性权重赋予如表 2 所列。

表 2 词性权重

Table 2 Parts of speech weight

词性	汽车领域		手机领域	
	样例	权重	样例	权重
名词(n)	杂声、异响、胎噪	1.8	高刷、闪充、气泡	1.6
动词(v)	溜车、顿挫、干吼	1.6	掉电、闪退、掉帧	1.5
形容词(a)	耐脏、省油、肉厚	1.5	轻薄、细腻、小巧	1.5
网络词(w)	耐操、B格、蛋疼	2	辣鸡、翻车、绝绝子	2

将词性权重与 TF-IDF 值结合，形成 TF-IDF-POS 值，并将词集按 TF-IDF-POS 值降序排序，选取小于阈值的词集作为 TF-IDF-POS 词集。本文通过多次实验确定筛选阈值，领域词汇在筛选后的词集中所占比例随阈值设置的变化情况如图 6 所示。

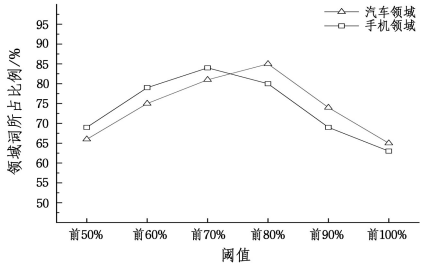

图 6 领域词占比随阈值变化图

Fig. 6 Diagram of the proportion of domain word changing with threshold

由图 6 可知，汽车领域筛选阈值为 80%时，领域词汇在筛选后的词集中占比最高，手机领域筛选阈值为 70%时，领域词汇在筛选后的词集中占比最高，所以汽车领域的筛选阈值设为前 80%，手机领域的筛选阈值设为前 70%。汽车领域的 TF-IDF-POS 词集中共得到 7900 余个词，手机领域的 TF-IDF-POS 词集中共得到 12000 余个词。

将预处理得到的词集进行 LDA 主题模型训练，每条评论视为一个文档。由于一个文档生成的若干主题中的主题词相似性较高，因此本文对于每个文档只选取第一个生成的主题中的主题词。为了减少噪声，每个主题设置生成 5 个主题词，并只考虑名词、动词、形容词和网络词 4 种词性。由于词频为 1 的词领域性较弱，所以在进行词频统计之后，去除词频为 1 的词，最终得到 LDA 词集，其中汽车领域共得到 1200 余个词，手机领域共得到 4800 余个词。

将 TF-IDF-POS 词集与 LDA 词集进行合并去重，得到最终的候选情感词集，部分候选情感词集样本如表 3 所列。

表 3 部分领域候选情感词

Table 3 Some domain candidate sentiment words

领域	候选情感词样例							
汽车领域	省油	隔音	皮实	真皮座椅	耐操	异响	漏风	胎噪
	异味	熄火	霸气	顿挫	上档次	动感	纸糊胎	推背感
	费油	桃木	太肉	风噪	漏油	硬塑料	无语	抖动
手机领域	轻薄	大气	细腻	高刷	哈曼卡顿	大电池	杜比音效	防抖
	进灰	磨砂	莱卡	火龙	划痕	死机	掉电	闪退
	掉帧	气泡	断网	闪屏	断触	翻新	漏液	杀后台

4.2.2 判断候选领域情感词的情感极性实验

选取清华大学情感词典和台湾大学情感词典的积极词和消极词共 21119 个作为种子词，使用 jieba 分词工具，对数据集 B 进行分词、去空行、根据自建的停用词表去除停用词等预处理，得到词集。使用 SO-PMI 算法计算词集与种子词集的共现信息，利用式(2)计算词集中词语的 SO-PMI 值，汽车领域得到 33000 余个情感词，其中积极词 22000 余个，消极词 11000 余个，中性词 100 余个；手机领域得到 24000 余个情感词，其中积极词 17000 余个，消极词 6800 余个，中性 200 余个，由于中性词数量太少，不具有参考价值，所以后续不采用。BERT 模型使用 Bert-base-Chinese 中文预训练模型，将 SO-PMI 算法得出的积极词与消极词合并，将总词语数量的 80% 作为 BERT 模型的训练集，10%作为验证集，10%作为测试集。对 BERT 分类模型进行微调，训练集、验证集、测试集的比例为 8∶1∶1。BERT 模型微调参数设置如表 4 所列。

表 4 BERT 模型微调参数设置

Table 4 BERT model fine-tuning parameter settings

参数	数值
max_seq_lenth	32
batch_size	64
learning_rate	2×10^{-5}
epochs	10

利用微调后的 BERT 分类模型分别对汽车领域和手机领域候选情感词进行情感预测，最终得到汽车领域情感词典和手机领域情感词典，领域情感词典部分情感词如表 5 所列。

表 5 领域情感词典部分情感词

Table 5 Some sentiment words in domain sentiment lexicon

汽车领域		手机领域	
情感词	极性	情感词	极性
省油	积极	轻薄	积极
隔音	积极	高刷	积极
皮实	积极	防抖	积极
真皮座椅	积极	莱卡	积极
耐操	积极	哈曼卡顿	积极
异响	消极	火龙	消极
漏风	消极	断触	消极
胎噪	消极	闪退	消极
异味	消极	漏液	消极
熄火	消极	杀后台	消极

4.2.3 对比实验

本文使用不同的情感词典在爬取的领域用户评论数据集 C 上进行基于规则的情感分析对比实验，从情感分类的效果角度验证本文提出的领域情感词典构建方法的有效性以及利用此方法构建的情感词典的质量。选取以下几种情感词典进行对比实验。

1) HowNet：知网情感词典。

2) TSING：清华大学李军中文褒贬义词典。

3) NTUSD：台湾大学简体中文情感词典。

4) BosonNLP：BosonNLP 情感词典。

5) LPdic：文献[19]提出的使用 LightGBM 模型筛选候选情感词，并使用 PMI 算法判断候选情感词的情感极性得到的领域情感词典。

6) PLdic：文献[22]提出的使用 PU Learning 选取候选情感词，使用 LPA 算法扩展原始词集得到的领域情感词典。

7) TSBdic：使用 TF-IDF 算法筛选候选情感词，并使用 SO-PMI 与 BERT 结合的方法判断候选情感词的情感极性构建出的领域情感词典。

8) TPSBdic：使用 TF-IDF-POS 算法筛选候选情感词并使用 SO-PMI 与 BERT 结合的方法判断候选情感词的情感极性

构建出的领域情感词典。

9）LSBdic：使用 LDA 算法筛选候选情感词并使用 SO-PMI 与 BERT 结合的方法判断候选情感词的情感极性构建出的领域情感词典。

10）TPLBdic：使用 TF-IDF-POS 算法与 LDA 算法结合筛选候选情感词，并直接使用传统情感词典词语作为种子词来微调 BERT 后对候选情感词进行情感预测构建出的领域情感词典。

4.2.4　评价指标

本文实验使用由混淆矩阵计算得到的精确率（Precision，P）、召回率（Recall，R）、F1 值（F1-measure，F1）作为情感分析实验的评价指标对各情感词典进行评价，混淆矩阵的表示如表 6 所列。

表 6　混淆矩阵表示

Table 6　Confusion matrix representation

真实类别	预测类别	
	正例	负例
正例	TP	FN
反例	FP	TN

精确率 P 表示预测正确的正样本数占总预测为正的样本数的比例，如式（3）所示：

$$P=\frac{TP}{TP+FP} \tag{3}$$

召回率 R 表示预测正确的正样本数占总真实为正的样本数的比例，如式（4）所示：

$$R=\frac{TP}{TP+FN} \tag{4}$$

F1 值表示精确率 P 和召回率 R 的调和平均值，如式（5）所示：

$$F1=\frac{2*P*R}{P+R} \tag{5}$$

4.2.5　实验结果及分析

使用各个情感词典进行情感分析的对比实验结果如表 7 所列。

根据实验结果可以看出，本文提出的领域情感词典构建方法所构建的领域情感词典在对汽车领域和手机领域的用户评论进行情感分析时，精度、召回率和 F1 值均高于其他情感词典。由于传统情感词典中缺少领域性词汇以及具有时效性的网络词汇，因此其对于特定领域用户评论的情感分析效果不理想。LPdic 在判断候选情感词极性阶段使用 PMI 算法，一些领域词汇无法判断情感极性，影响了所构建领域情感词典的完整性。而 CSdic 使用 LPA 算法扩展原始词集，造成了部分领域词汇丢失，影响了所构建词典的质量。这两种方法效果均不如本文方法构建的领域情感词典 TPLSBdic。

而在判断情感词极性阶段使用相同方法的情况下，在筛选候选情感词阶段使用本文提出的 TF-IDF-POS 算法所构建的领域情感词典的 3 个指标均高于使用 TF-IDF 算法的实验结果。由于 F1 值是精确率和召回率的调和平均值，更能客观反映词典的质量，因此实验以 F1 值为例进行分析，各情感词典情感分析实验结果的 F1 值对比如图 7 所示。可以看出，汽车领域的 TPSBdic 比 TSBdic 的 F1 值高出 0.12%，手机领域的 TPSBdic 比 TSBdic 的 F1 值高出 7.83%，原因是本文对 TF-IDF 算法的改进是提高常见领域词的词性权重，这使得筛选的候选情感词冗余减少，选中的领域词更多。

表 7　各情感词典情感分析实验结果对比

Table 7　Comparison of experimental results of sentiment analysis in various sentiment lexicons

（%）

词典	方法	汽车领域			手机领域		
		P	*R*	*F*1	*P*	*R*	*F*1
HowNet	知网 HowNet 词典	65.72	63.98	64.84	82.30	78.20	80.20
TSING	清华大学李军中文褒贬义词典	73.51	73.46	73.48	83.68	82.94	83.31
NTUSD	台湾大学 NTUSD 简体中文情感词典	72.40	71.72	72.06	85.38	84.93	85.15
BosonNLP	BosonNLP 词典	66.02	66.02	66.02	74.44	72.92	73.67
LPdic	LightGBM＋PMI	77.23	75.27	76.24	86.37	84.18	85.26
PLdic	PU Learning＋LPA	78.35	75.59	76.95	87.58	84.81	86.17
TSBdic	TF-IDF＋SO-PMI＋BERT	77.37	67.81	72.28	82.2	74.60	78.22
TPSBdic	TF-IDF-POS＋SO-PMI＋BERT	77.45	67.96	72.40	87.52	84.62	86.05
LSBdic	LDA＋SO-PMI＋BERT	76.84	72.27	74.48	87.02	85.30	86.15
TPLBdic	TF-IDF-POS＋LDA＋BERT	69.74	69.66	69.70	85.92	85.88	85.90
TPLSBdic	TF-IDF-POS＋LDA＋SO-PMI＋BERT	**80.04**	**76.09**	**78.02**	**89.15**	**87.56**	**88.35**

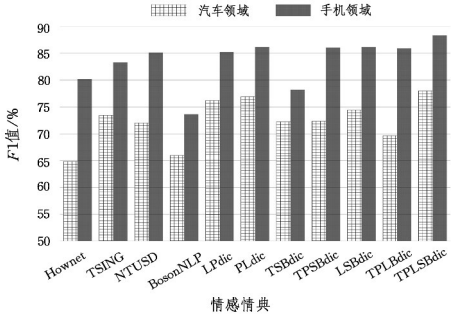

图 7　各情感词典情感分析 F1 值对比

Fig. 7　Comparison of F1 values for sentiment analysis in various sentiment lexicons

另外，在判断情感词极性阶段使用相同方法的情况下，TF-IDF-POS 算法与 LDA 算法结合筛选候选情感词在 3 个指标上均高于单独使用 TF-IDF-POS 算法筛选候选情感词和单独使用 LDA 算法筛选候选情感词。以 F1 值为例，汽车领域的 TPLSBdic 比 TPSBdic 的 F1 值高出 5.62%，比 LSBdic 的 F1 值高出 3.54%；手机领域的 TPLSBdic 比 TPSBdic 的 F1 值高出 2.3%，比 LSBdic 的 F1 值高出 2.2%。这是由于 TF-IDF-POS 算法与 LDA 算法结合筛选情感词可以更大限度地减少领域情感词的冗余，LDA 算法提取的领域词汇对 TF-IDF-POS 词集实现了领域重要程度的改善，使筛选出的情感词更加具有领域特性。

TPLSBdic 的实验指标均高于 TPLBdic，以 F1 值为例，汽

车领域的 TPLSBdic 比 TPLBdic 的 F1 值高出 8.32%，手机领域的 TPLSBdic 比 TPLBdic 的 F1 值高出 2.45%，说明在筛选候选情感词阶段使用相同方法的情况下，SO-PMI 算法与 BERT 结合的情感词极性判断方法是有效的。原因是基于 SO-PMI 算法在语料数据集中抽取的情感词集相较于传统情感词典更具有领域特性，其对 BERT 分类模型的微调使 BERT 学习到了更多领域特性，从而提高了对相应领域的情感词的极性判断的准确率。

综上所述，相比其他情感词典，利用本文方法构建的领域情感词典质量更高，更加适用于相应领域的用户评论文本情感分析。在两个领域中的实验效果相同，说明本文提出的领域情感词典构建方法具有较高的准确性以及一定的普适性。

4.2.6 情感分类错误案例分析

在情感分析对比实验中，有一些分类错误的评论文本可以反映情感词典的质量，使用本文方法构建的领域情感词典进行情感分析的部分错误案例如表 8 所列。分类错误的原因可以分为情感词多语境多义、分词问题、情感词与情感目标关联不准确 3 种，具体如下：

表 8 部分情感分析错误案例

Table 8 Some cases of sentiment analysis errors

领域	序号	评论文本	实际类别	实验类别
汽车领域	1	内饰简单，和外观形成了鲜明对比。	消极	积极
	2	这车我最看中的就是油耗，能省则省，之前的车太费油了。	积极	消极
	3	空间严密，如果关窗不开空调，那车里会比较闷。	消极	积极
手机领域	1	降价还是蛮快的，没几天又是送散热架又是降价，比起 618 真的好多了，平时价格很香。	积极	消极
	2	给老公买的，老公说，外观和界面还行，就是新手机待机电量得一天两充。	消极	积极
	3	这个红米手机很大，但是拿得住，之前用的 mix3，着实太重了点。	积极	消极

1)评论文本中的情感词存在多语境多义的情况。如汽车领域评论 1 中的“内饰简单”和评论 3 中的“空间严密”，在本文构建的汽车领域情感词典中极性均为积极，但不同的用户对内饰和空间的要求不同，有的用户喜欢内饰简单或者空间严密，有的用户不喜欢，由于语境的不同，“内饰简单”“空间严密”的极性也不同，从而导致情感分类错误；手机领域评论 1 中“降价”一词同理，在本文构建的手机领域情感词典中的极性为积极，但有的用户是降价之后买到产品，所以对于“降价”一词表现为积极情绪，有的用户是降价之前买到产品，对于“降价”一词则表现为消极情绪。

2)分词问题导致所构建情感词典中情感词汇缺失。如手机领域评论 2 中的“一天两充”一词，在手机领域代表用户的消极情绪，但此词在本文构建的手机领域情感词典中不存在，原因是在词典构建过程中，分词时将此词分为了“一天”和“两充”，从而导致情感分类错误。

3)情感词与情感目标关联不准确。如手机领域评论 3 中的“太重”一词，在手机领域代表用户的消极情绪，在本文构建的汽车领域情感词典中极性为消极，但在此评论中“太重”描述的对象是“mix3”，而不是本身评论的对象，从而导致情感分类错误，汽车领域评论 2 中“费油”一词同理。

综上所述，本文提出的领域情感词典构建方法仍有一些局限性，包括语境依赖性捕捉能力、分词的准确性以及情感词与情感目标的关联准确性等都有待进一步提高，这些也是未来改进的方向。

结束语 本文提出了一种基于改进 TF-IDF 与 BERT 的领域情感词典构建方法，利用改进的 TF-IDF 算法，即 TF-IDF-POS 算法与 LDA 模型结合，在领域评论语料中筛选领域候选情感词，将 SO-PMI 算法与 BERT 模型结合，利用领域微调后的 BERT 分类模型判断领域候选情感词的情感极性；以汽车领域和手机领域为例，验证了该方法的有效性。通过实验可知，相比于传统通用情感词典，本文提出的领域情感词典构建方法所构建的情感词典在领域评论文本的情感分析中准确率更高，说明了本文方法可以用于不同领域的情感词典构建，具有一定的普适性。

由于本文对于情感词极性的判断只考虑了积极和消极两个极性，如何进行更细粒度的极性判定也是后续研究的一个方向。

参 考 文 献

[1] ZHAO Y Y, QIN B, LIU T, et al. Sentiment Analysis[J]. Journal of Software, 2010, 21(8): 1834-1848.

[2] ZHAO Y Y, QIN B, SHI Q H, et al. Large-scale Sentiment Lexicon Collection and Its Application in Sentiment Classification [J]. Journal of Chinese Information Processing, 2017, 31(2): 187-193.

[3] DAI L, LIU B, XIA Y, et al. Measuring Semantic Similarity between Words Using HowNet[C]// 2008 International Conference on Computer Science and Information Technology, Los Alamitos, USA: IEEE Computer Society, 2008: 601-605.

[4] LI J, SUN M. Experimental Study on Sentiment Classification of Chinese Review using Machine Learning Techniques[C]// 2007 International Conference on Natural Language Processing and Knowledge Engineering. Piscataway, USA: IEEE, 2007: 393-400.

[5] KU L, CHEN H. Mining opinions from the Web: Beyond relevance retrieval[J]. Journal of the American Society for Information Science and Technology, 2007, 58(12): 1838-1850.

[6] ZHAI Y, WANG Z, ZENG H, et al. Social Media Opinion Leader Identification Based on Sentiment Analysis[C]// Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing. New York, USA: Association for Computing Machinery, 2021: 436-440.

[7] PARK S, LEE W, MOON I. Efficient extraction of domain specific sentiment lexicon with active learning[J]. Pattern Recognition Letters, 2015, 56(apr. 15): 38-44.

[8] WANG K, XIA R. A Survey on Automatical Construction Methods of Sentiment Lexicons[J]. Acta Automatica Sinica, 2016, 42(4): 495-511.

[9] NEVIAROUSKAYA A, PRENDINGER H, ISHIZUKA M. SentiFul: A Lexicon for Sentiment Analysis[J]. IEEE Transactions on Affective Computing, 2011, 2(1): 22-36.

[10] HASSAN A, ABUJBARA A, RADE V, et al. Identifying the Semantic Orientation of Foreign Words[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2011: 592-597.

[11] DRAGUT E C, WANG H, SISTLA P, et al. Polarity Consisten-

cy Checking for Domain Independent Sentiment Dictionaries[J]. IEEE Transactions on Knowledge and Data Engineering,2015,27(3):838-851.

[12] ZHU Y L,MIN J,ZHOU Y Q,et al. Semantic Orientation Computing Based on HowNet[J]. Journal of Chinese Information Processing,2006,20(1):14-20.

[13] BOLLEGALA D,WEIR D,CARROLL J. Using Multiple Sources to Construct a Sentiment Sensitive Thesaurus for Cross-Domain Sentiment Classification[C]// Meeting of the Association for Computational Linguistics:Human Language Technologies. USA: Association for Computational Linguistics, 2011: 132-141.

[14] KRESTEL R,SIERSDORFER S. Generating contextualized sentiment lexica based on latent topics and user ratings[C]// Proceedings of the 24th ACM Conference on Hypertext and Social Media. New York,USA:ACM,2013:129-138.

[15] DENG D,JING L,YU J,et al. Sentiment Lexicon Construction With Hierarchical Supervision Topic Model[J]. IEEE/ACM Transactions on Audio,Speech,and Language Processing,2019,27(4):704-718.

[16] WANG Y,YIN F,LIU J,et al. Automatic construction of domain sentiment lexicon for semantic disambiguation[J]. Multimedia tools and applications,2020,79(31/32):22355-22373.

[17] ZHAO C,ZHANG P,LIU J,et al. Research on Domain Emotion Dictionary Construction Method based on Improved SO-PMI Algorithm[C]// 2021 5th International Conference on Natural Language Processing and Information Retrieval(NLPIR). New York,USA:Association for Computing Machinery,2021:18-23.

[18] WANG Y,HUANG G,LI M,et al. Automatically Constructing a Fine-Grained Sentiment Lexicon for Sentiment Analysis[J]. Cognitive Computation,2022,15(1):254-271.

[19] REN W,ZHANG H W,CHEN M. A Method of Domain Dictionary Construction for Electric Vehicles Disassembly[J]. Entropy. 2022,24(3):363.

[20] HUANG S,NIU Z,SHI C. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation [J]. Knowledge-Based Systems,2014,56(jan.):191-200.

[21] XI Y H. Construction of Domain-specific Sentiment Lexicon in Product Reviews[J]. Journal of Chinese Information Processing,2016,30(5):136-144.

[22] LI C,YAN X,XU G,et al. Khmer Sentiment Lexicon Based on PU Learning and Label Propagation Algorithm[J]. ACM Transactions on Asian and Low-Resource Language Information Processing,2023,22(3):1-18.

[23] YANG X P,ZHANG Z X,WANG L,et al. Automatic Construction and Optimization of Sentiment Lexicon Based on Word2Vec [J]. Computer Science,2017,44(1):42-47.

[24] ZHANG P,WANG J X,WANG Y H. Sentiment Lexicon Construction Method Based on Label Propagation[J]. Computer Engineering,2018,44(5):168-173.

[25] YANG S Q,XU C J. Research on Constructing Sentiment Dictionary of Online Course Reviews based on Multi-source Combination[C]// Proceedings of the 2019 2nd International Conference on Data Science and Information Technology. New York,USA:ACM,2019:71-76.

[26] YE X,CAO J B,XU FEI X,et al. Sentiment dictionary adaptive learning method in Chinese domain[J]. Computer Engineering and Design,2020,41(8):2231-2237.

[27] LIU H,CHEN X,LIU X. A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis[J]. IEEE Access,2022,10:32280-32289.

[28] BLEI D M,NG A Y,JORDAN M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research,2003,3:993-1022.

[29] DEVLIN J,CHANG M,LEE K,et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg,USA:Association for Computational Linguistics,2019:4171-4186.

JIANG Haoda, born in 1997, master, is a member of CCF(No. I2375G). His main research interests include natural language processing and sentiment analysis.

ZHAO Chunlei, born in 1979, Ph.D, is a member of CCF (No. 18494M). Her main research interests include natural language processing and network information security.