结合语言知识和深度学习的中文文本情感分析方法

徐康庭，宋威
北方工业大学信息学院，北京 100144

摘要
在目前的中文文本情感分析研究中，基于语义规则和情感词典的方法通常需要人工设置情感阈值；而基于深度学习的方法由于未能运用语义规则和情感词典等语言知识，不能充分提取情感特征。针对这两种方法的缺点，提出了一种将语言知识和深度学习结合的文本情感分析方法。该方法首先根据语义规则提取文本中的关键情感片段，再根据情感词典从关键情感片段中抽取出情感更加明确的情感词来构建情感集合，然后利用深度学习模型分别从原始文本、关键情感片段、情感集合中抽取深层次特征，最后对提取的特征进行加权融合，并利用分类器实现情感极性的判断。实验结果表明，与未引入语言知识的深度学习模型相比，该方法的情感极性分类能力有明显提升。

关键词
文本情感分析；语言知识；深度学习

中图分类号：TP391.1 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022026

A Chinese text sentiment analysis method combining language knowledge and deep learning

XU Kangting, Song Wei
School of Information Science and Technology, North China University of Technology, Beijing 100144, China

Abstract
At present, in the research of Chinese text emotion analysis, the method based on semantic rules and emotion dictionary usually needs to set the emotional threshold manually. However, the method based on deep learning can't fully extract emotional features because it fails to use language knowledge such as semantic rules and emotional dictionary. As to shortcomings of two methods, a text emotion analysis method combining language knowledge and deep learning was proposed. Firstly, the key emotional segments in the text were extracted according to the semantic rules. Secondly, more explicit emotion words were extracted from the key emotional segments according to the emotional dictionary to construct the emotion set. Thirdly, the deep level features were extracted from the original text, key emotional segments and

emotional set by using the deep learning model. Finally, the features were weighted and fused, and the classifier was used to judge the emotional polarity. The experimental results show that compared with the deep learning model without language knowledge, this method has significantly improved the ability of emotional polarity classification.

Key words
text sentiment analysis, language knowledge, deep learning

0 引言

情感分析(sentiment analysis)是指通过分析、归纳、推理等过程自动地对具有感情色彩的文本进行情感极性的判断[1]。随着Web2.0的到来，越来越多的人成为互联网的参与者，并通过博客、在线门户网站、电商平台等产生大量具有感情色彩的文本。对这些文本进行分析挖掘对于舆情分析、政府决策、产品分析具有重要意义。

目前情感分析的方法可以大致分为3类：基于词典和规则的方法、基于传统机器学习的方法、基于深度学习的方法。

基于词典和规则的方法通过词典分析、句法分析、句型分析等方法对文本的情感极性进行判断。Wu J S等人[2]通过构建情感词典、否定词词典、程度副词词典等多部词典提出词语级情感判断方法，并对文本进行句法分析、句型分析，从而实现了对中文微博的情感判断。赵妍妍等人[3]通过构建大规模情感词典实现了中文微博的情感分析。Xu G X等人[4]通过扩展现有词典的方法实现了对评论文本的情感分析。KESHAVARZ H等人[5]为了改善微博情感分类的性能，通过将语料库和词典结合的方式构建自适应词典。李继东等人[6]通过扩展词典，并对句间规则和句型规则进行分析，提高了中文微博情感分析的性能。Zhang S X等人[7]首先对情感词典进行构建和扩充，然后通过计算权重得到微博文本的情感值，实现了对微博文本的情感分类。

基于传统机器学习的方法一般先对文本利用词袋(bag of word)模型进行编码，然后利用朴素贝叶斯(naive Bayes，NB)、支持向量机(maximum entropy，ME)、决策树等传统机器学习模型进行情感分类。Pang B等人[8]基于词袋模型分别利用朴素贝叶斯模型、最大熵模型和支持向量机(support vector machine，SVM)实现了对电影评论的情感分类。苏莹等人[9]将朴素贝叶斯模型和潜在狄利克雷分布(latent Dirichlet allocation，LDA)结合，并引入合适的情感词典，实现了对网络评论的篇章级别和句子级别的情感倾向性分析。

基于词典和规则的方法需要人工预先对每个情感词和语义规则设定情感极性值，而针对不同领域的情感任务共用通用的词典，必然会带来人工误差，而且对情感词和规则进行情感极性标注需要耗费大量的人力。基于传统机器学习的方法一般基于词袋模型，忽略了上下文语义，而且需要做特征工程。基于深度学习的方法能够自动实现端到端的学习和推理过程，鉴于此，近年来该方法成为研究的热点。基于深度学习的方法得到的特征可以直接用于预测概率，也可以使用支持向量机等浅层分类器进行分类[10]。Kim Y[11]将预训练的词向量作为输入，利用卷积神经网络提取文本特征，从而实现了文本分类任务，并取得了不错的效果。胡荣磊等人[12]将预训练的词向量作为输入，利用长短期记忆(long short-term memory，LSTM)网络学习文本的语义特征和序列特征，并与注意力模型相结合，有效提高了文本情感分析

任务的性能。李洋等人[13]将卷积神经网络（convolutional neural network，CNN）和双向长短期记忆（BiLSTM）网络结合，充分利用了CNN提取局部特征的能力和双向长短期记忆网络提取文本序列特征的能力，提高了文本情感分析的性能。宋婷等人[14]通过区域卷积神经网络提取文本局部特征以及不同句子的时序关系，并利用改进的分层长短期记忆网络获取句子内部和句子间的情感特征，从而提高了方面级情感分析的性能。

虽然，基于数据驱动的深度学习模型能够有效弥补基于词典和规则的方法以及基于传统机器学习的方法的不足。但是，仅仅依靠数据进行深度学习模型训练，忽略了情感词典和语义规则等语言知识，导致模型不能充分学习文本特征，进而无法突破深度学习模型的性能瓶颈。近年来，融合情感词典或语义规则的深度学习模型逐渐成为热点。谢润忠等人[15]将情感集合和深度学习模型进行融合，得到了不错的结果，但其未考虑语法规则。邱宁佳等人[16]将语义规则和深度学习融合并建立三通道模型，提高了文本情感分析的性能，但其未考虑情感集合。鉴于此，本文提出了一种结合语言知识和深度学习的中文文本情感分析的新方法CLKDL（the combination of language knowledge and deep learning），充分将语言知识和深度学习模型结合，通过数据和知识共同驱动模型学习，以提高情感极性分类模型的性能。

本文主要贡献如下。

- 为了解决中文语义多样性问题，降低语义的复杂性，突出关键情感信息对模型的贡献，提出了CLKDL方法。
- CLKDL方法首先利用词典和规则抽取出情感倾向明确的情感集合信息；然后，为了防止出现由词典维护不及时造成的情感集合信息缺失的问题，利用语义规则抽取出情感倾向更加明确的关键情感片段；最后，为了防止出现由抽取情感集合和情感片段造成的文本序列特征缺失的问题，从原始文本中抽取出序列特征，三者相辅相成。
- 构建深度学习模型，分别从原始文本、关键情感片段、情感集合3个部分中抽取深层次特征，从而将语言知识与深度学习结合，并完成中文文本情感分析任务。然后，利用酒店评论数据集ChnSentiCorp和O2O商铺食品安全评论相关数据对所提方法进行有效性实验。实验结果表明，所提方法的情感极性分类能力有明显提升。

1 基于语义规则和情感词典的信息抽取

1.1 关键情感片段的抽取

相对英文而言，中文语义规则具有较高的复杂性。对于使用不同语义规则描述的文本，其表达的情感倾向以及情感强度也不相同。因此，需要根据语义规则把能够改变情感倾向以及情感强度的关键情感片段从原文本中剥离出来，以降低中文文本语义的复杂度，为后续深度学习特征的提取加入语义知识，进而提高情感分析的性能。本文从句间规则和句型规则两种角度抽取关键情感片段。

1.1.1 基于句间规则的抽取

标点符号是划分句间关系的重要标准，首先利用标点符号“？”“！”“。”以及“；”将原始文本划分为若干个复句，用集合$\{C_0,C_1,\cdots,C_n\}$表示。每个复句又可以划分为若干个子句，用集合$\{S_0,S_1,\cdots,S_m\}$表示。

对句间规则的分析即对复句中若干个子句之间相互关系的分析。会对情感分析造成影响的句间关系有转折关系、递进关系和假设关系。根据分析，定义以下规则。

（1）转折关系规则

当句间出现转折时，情感也随之改变。一般转折前后的情感极性是相反的，值得注意的是，真实表达的情感极性在转折词后的情感片段中。具体规则如下。

● 如果复句C只出现单一转折后接词（如“但是”“然而”）且该单一转折后接词出现在S_k中，则取出子集$\{S_k,S_{k+1},\cdots,S_m\}$作为关键情感片段。

● 如果复句C只出现单一转折前接词（如“虽然”“尽管”）且该单一转折前接词出现在S_k中，则取出子集$\{S_0,S_1,\cdots,S_k\}$作为关键情感片段。

● 如果复句C中出现连续完整转折词（如“虽然……但是……”）且该转折后接词出现在S_k中，则取出子集$\{S_k,S_{k+1},\cdots,S_m\}$作为关键情感片段。

（2）递进关系规则

当句间出现递进关系时，一般后句的情感表达程度要明显强于前句。当复句C中出现递进关系词（如“更加”“更有甚者”），并且该递进关系词出现在S_k中时，取出子集$\{S_k,S_{k+1},\cdots,S_m\}$作为关键情感片段。

（3）假设关系规则

当句间出现假设关系时，一般情感表达的重心在前句，如果出现否定假设，那么情感表达的极性会相反。当复句C中出现假设后接词（如“那么”）且该假设后接词出现在S_k中时，取出子集$\{S_0,S_1,\cdots,S_k\}$作为关键情感片段。

以上3种句间关系都会对情感分析造成一定的影响，因此需要根据语义规则把关键情感片段抽取出来做进一步的分析。至于其他句间关系，比如并列关系、因果关系、一般关系等并不会对情感倾向和情感程度造成影响，因此本文不做特殊处理。

1.1.2 基于句型规则的抽取

中文常用的句型有疑问句、反问句、感叹句、陈述句。其中疑问句和反问句会使得情感极性变反；感叹句虽然不影响情感极性，却会改变情感表达的程度；陈述句一般不会对情感极性和情感程度造成影响。基于上述分析，定义以下关键情感片段的抽取规则。

（1）感叹句

如果复句C是感叹句，即以“！”或多个“！”（如“！！！”“！！！！！！”）结尾，则将整个复句C（即$\{S_0,S_1,\cdots,S_m\}$）作为关键情感片段。

（2）反问句及疑问句

如果复句C是反问句或者疑问句，即以“？”或多个“？”（如“？？？”“？？？？？？”）结尾，则将整个复句C（即$\{S_0,S_1,\cdots,S_m\}$）作为关键情感片段。

（3）陈述句

如果复句C是陈述句，其不会对情感极性和情感强度造成影响，因此本文不做特殊处理。

1.2 情感集合的抽取

情感词是主体对客体的情感偏离度的直接表达，程度副词、否定词等修饰词会对情感词的情感极性和情感强度造成影响。为了进一步降低中文语义的复杂度，以进一步提高后续深度学习模型提取特征的准确性，进而提高情感分析的性能，本文引入情感词典、程度副词词典、否定词词典等已知的语言学知识。从上述根据句间规则抽取的关键情感片段中抽取出情感更

加明确的情感词来构建情感集合。定义如下规则。

规则1：如果当前词为情感词，则将其加入情感集合中。

规则2：如果当前词为程度副词，且下一个词为情感词，则将当前的程度副词和情感词组合成新词加入情感集合。如果情感集合中有当前情感词，则将其从情感集合中删除。

规则3：如果当前词为程度副词，且下一个词为否定词、下面第二个词为情感词，则将当前的程度副词、否定词、情感词组合在一起并加入情感集合。如果当前情感词或者当前否定词和情感词的组合已经在情感集合中，则将其从情感集合中删除。

规则4：如果当前词为否定词，且下一个词为情感词，则将当前的否定词和情感词组合在一起并加入情感集合。如果当前的情感词在情感集合中，则将其从情感集合中删除。

规则5：如果当前词为否定词，且下一个词为程度词、下面第二个词为情感词，则将当前的否定词、程度副词、情感词组合在一起并加入情感集合。如果当前情感词或者程度副词和情感词的组合在情感集合中，则将其从情感集合中删除。

2 深度学习相关技术

2.1 BERT预训练模型

基于Transformer的双向编码器表征（bidirectional encoder representation from Transformer，BERT）技术是一个多任务模型，通过遮蔽语言模型（masked language model，MLM）和下一句话预测（next sentence prediction，NSP）分别捕获词语和句子级别的向量表示[17]。BERT模型结构如图1所示。BERT整体处理流程为：首先对输入文本进行字向量编码、文本向量编码、位置向量编码，然后通过双向Transformer模块得到文本的向量化表示。对于文本向量静态嵌入而言，与传统的word2vec相比，使用BERT作为向量化工具能够根据下游任务对文本的向量表示进行动态调整，从而解决一词多义的问题。

2.2 卷积神经网络

卷积神经网络利用一个滤波器在一个文本数据上上下滑动以探知不同位置的特征，从而提取文本局部特征。假设 $\boldsymbol{x}\in\mathbb{R}^{L\times d}$ 表示文本中第i个字向量。$\boldsymbol{x}\in\mathbb{R}^{L\times d}$ 表示一个输入文本的向量，其中L表示输入文本的长度。$\boldsymbol{m}\in\mathbb{R}^{k\times d}$ 表示卷积操作的滤波器，其中k表示滤波器的长度，d表示词嵌入的维度。文本中的每个位置j都有包含k个连续字向量的窗口向量$\boldsymbol{w}_j$，如式（1）所示：

$$\boldsymbol{w}_j=[\boldsymbol{x}_j,\boldsymbol{x}_{j+1},\cdots,\boldsymbol{x}_{j+k-1}] \quad (1)$$

其中，逗号表示行向量连接，滤波器$\boldsymbol{m}$通过逐一滑动窗口产生一个特征映射（feature map）$\boldsymbol{c}\in\mathbb{R}^{L-k+1}$，$\boldsymbol{c}=[c_1,c_2,\cdots,c_{L-k+1}]$。特征映射$\boldsymbol{c}$是提取的文本局部特征。

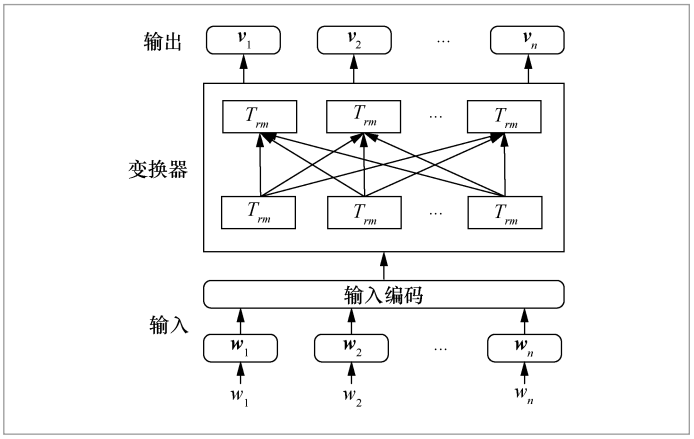


图1 BERT模型结构

2.3 长短期记忆网络

LSTM是由Hochreiter S等人[18]在1997年第一次提出，随后经过Graves A[19]改良推广的模型。其能够有效解决循环神经网络（recurrent neural network，RNN）中长期依赖的问题，在很多任务中取得了不错的表现，LSTM的结构如图2所示。LSTM在t时刻的转换函数定义如下。

$$\boldsymbol{i}_{(t)}=\sigma(\boldsymbol{W}_{xi}^{\top}\cdot\boldsymbol{x}_{(t)}+\boldsymbol{W}_{hi}^{\top}\cdot\boldsymbol{h}_{(t-1)}+\boldsymbol{b}_{i}) \quad (2)$$

$$\boldsymbol{g}_{(t)}=\tanh(\boldsymbol{W}_{xg}^{\top}\cdot\boldsymbol{x}_{(t)}+\boldsymbol{W}_{hg}^{\top}\cdot\boldsymbol{h}_{(t-1)}+\boldsymbol{b}_{g}) \quad (3)$$

$$\boldsymbol{f}_{(t)}=\sigma(\boldsymbol{W}_{xf}^{\top}\cdot\boldsymbol{x}_{(t)}+\boldsymbol{W}_{hf}^{\top}\cdot\boldsymbol{h}_{(t-1)}+\boldsymbol{b}_{f}) \quad (4)$$

$$\boldsymbol{c}_{(t)}=\boldsymbol{i}_{(t)}\otimes\boldsymbol{g}_{(t)}+\boldsymbol{f}_{(t)}\otimes\boldsymbol{c}_{(t-1)} \quad (5)$$

$$\boldsymbol{o}_{(t)}=\sigma(\boldsymbol{W}_{xo}^{\top}\cdot\boldsymbol{x}_{(t)}+\boldsymbol{W}_{ho}^{\top}\cdot\boldsymbol{h}_{(t-1)}+\boldsymbol{b}_{o}) \quad (6)$$

$$\boldsymbol{y}_{(t)}=\boldsymbol{h}_{(t)}=\boldsymbol{o}_{(t)}\otimes\tanh(\boldsymbol{c}_{(t)}) \quad (7)$$

其中，σ为Sigmoid激活函数，tanh表示双曲正切函数，$\otimes$表示矩阵相乘，$\boldsymbol{x}_{(t)}$表示当前时刻的输入向量，$\boldsymbol{i}_{(t)}$、$\boldsymbol{f}_{(t)}$、$\boldsymbol{o}_{(t)}$分别表示输入门、遗忘门和输出门，$\boldsymbol{g}_{(t)}$表示当前细胞的候选状态，$\boldsymbol{W}_{xi}$、$\boldsymbol{W}_{xf}$、$\boldsymbol{W}_{xo}$、$\boldsymbol{W}_{xg}$表示每层连接到输入向量$\boldsymbol{x}_{(t)}$的权重矩阵，$\boldsymbol{W}_{hi}$、$\boldsymbol{W}_{hf}$、$\boldsymbol{W}_{ho}$、$\boldsymbol{W}_{hg}$表示每层连接到前一个隐藏状态$\boldsymbol{h}_{(t-1)}$的权重矩阵，$\boldsymbol{b}_{i}$、$\boldsymbol{b}_{f}$、$\boldsymbol{b}_{o}$、$\boldsymbol{b}_{g}$表示每层的偏置，输入门$\boldsymbol{i}_{(t)}$用于控制被存储在当前记忆细胞中的新信息量，$\boldsymbol{f}_{(t)}$用于控制旧记忆细胞中信息被丢弃的程度，$\boldsymbol{o}_{(t)}$根据当前记忆细胞$\boldsymbol{c}_{(t)}$用于控制被输出的信息，$\boldsymbol{h}_{(t)}$表示当前时刻细胞状态，$\boldsymbol{y}_{(t)}$表示输出，$\boldsymbol{c}_{(t-1)}$表示上一时刻的细胞状态。

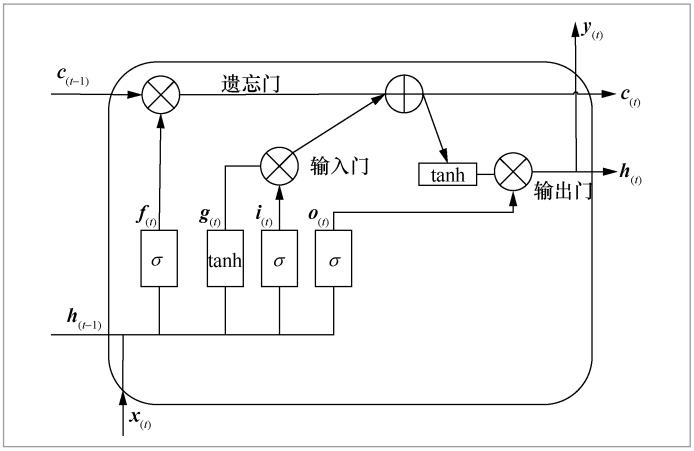


图 2 LSTM 结构

2.4 注意力机制

注意力机制为模型的输入赋予不同的权重，根据任务的具体情况，为更加关键和重要的信息设置更高的权重；反之，设置更低的权重，以此来提升模型的性能。

3 结合语言知识和深度学习的CLKDL方法

本文提出的CLKDL方法由3个部分组成，分别是原始文本部分、语义规则部分、情感集合部分。每个部分又由5个层次组成，分别是语义规则及预处理层、词嵌入层、特征提取层、加权融合层和输出层。CLKDL结构如图3所示。

3.1 语义规则及预处理层

该层主要根据第1节定义的基于语义规则和情感词典的信息抽取方法，从原始文本中抽取关键情感片段和情感集合，以降低中文语义的复杂度，加强深度学习模型的特征提取能力。

对于一个文本，首先利用语义规则抽取出关键情感片段$S=\{S_0^0,S_1^0,\cdots,S_k^0,S_0^1,S_1^1,\cdots,S_k^1,\cdots,S_k^n\}$，其中上标$n$表示该文本中的第$n$个复句，下标$k$表示从第$n$个复句中抽取的第$k$个关键情感片段。将该集合元素进行拼接作为语义规则部分的输入。

然后利用情感词典、程度副词词典、

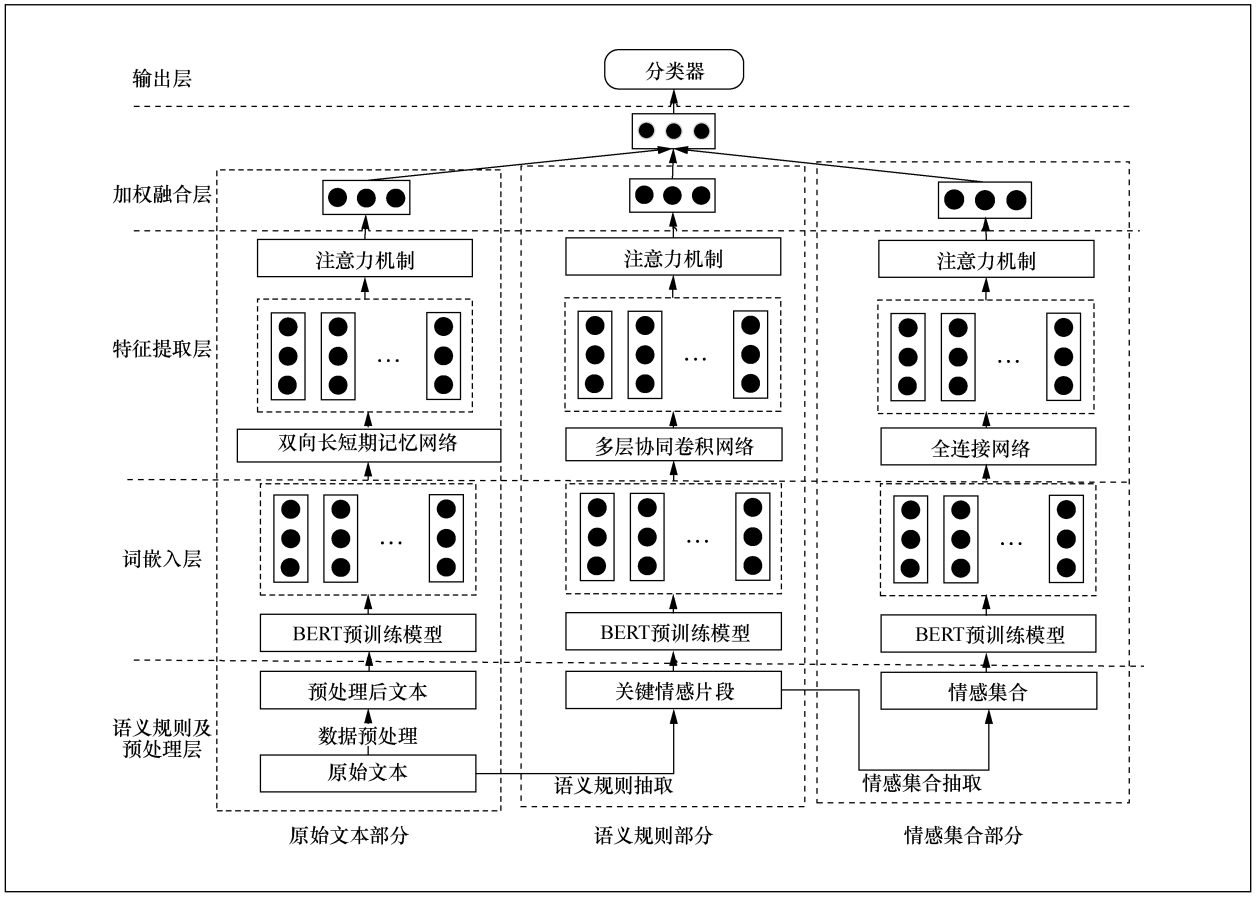


图 3 CLKDL 结构

否定词词典从关键情感片段中抽取情感集合$E=\{E^n_{k1},E^n_{k2},\cdots,E^n_{km}\}$，其中$E^n_{km}$表示该文本第$n$个复句中的第个$k$关键情感片段中的第$m$个情感词。将该集合元素进行拼接并作为情感集合部分的输入。

最后将原始文本进行去停用词、去特殊字符等预处理后得到的文本作为原始文本部分的输入。

3.2 词嵌入层

利用BERT预训练模型分别对原始文本部分、语义规则部分、情感集合部分的输入进行词向量化，得到3个部分的文本向量。BERT词嵌入能够对词向量进行动态调整，解决一词多义的问题，从而将真实的语义嵌入词向量中。3个部分的词向量分别用$\boldsymbol{x}\in\mathbb{R}^{L^T\times d}$、$\boldsymbol{s}\in\mathbb{R}^{L^S\times d}$、$\boldsymbol{e}\in\mathbb{R}^{L^E\times d}$表示。其中$L^T$表示原始文本长度，$L^S$表示关键情感片段的长度，$L^T$表示情感集合的长度。

3.3 特征提取层

3.3.1 原始文本特征提取

为了防止文本序列特征的丢失并充分考虑上下文特征，使用BiLSTM网络从原始文本中提取深层次特征。

在某一时刻，BiLSTM的输出$\boldsymbol{h}_t$由前向输出向量$\overrightarrow{\boldsymbol{h}_t}$和反向输出向量$\overleftarrow{\boldsymbol{h}_t}$组合而成。计算方式如下。

$$\vec{h}_t = \overrightarrow{\mathrm{LSTM}}(x_t), t \in (0, L^T) \quad (8)$$

$$\overleftarrow{h}_t = \overleftarrow{\mathrm{LSTM}}(x_t), t \in (0, L^T) \quad (9)$$

$$h_t = [\vec{h}_t, \overleftarrow{h}_t], t \in (0, L^T) \quad (10)$$

其中，x_t为在t时刻BiLSTM的输入，L^T表示文本序列总长度。

为了增加关键特征对情感分析任务的贡献，降低无效信息对模型的干扰，在BiLSTM提取特征后引入注意力机制。首先生成目标注意力权重u_t，然后将目标注意力权重向量化，生成权重向量a_t，最后将生成的权重向量配置给隐层状态语义编码h_t，生成包含注意力权重的原始文本特征向量V，计算过程如下。

$$u_t = \tanh(W_w \cdot h_t + b_w) \quad (11)$$

$$a_t = \frac{\exp(u_t^\top u_w)}{\sum_t \exp(u_t^\top u_w)} \quad (12)$$

$$V = \sum_t a_t h_t \quad (13)$$

其中，h_t为经过BiLSTM得到的原始文本特征向量，W_w、b_w、u_w为注意力网络的可调节参数。

3.3.2 关键情感片段特征提取

为了进一步从关键情感片段中提取情感特征，本文采用多个不同大小的卷积核并行地对关键情感片段进行多层次的特征抽取。不同大小的卷积核能够提取多种N-gram特征，并且使用并列结构在一定程度上能够解决深度学习模型过深导致的信息丢失和梯度消弱的问题。多尺寸卷积神经神经网络（multi-scale convolutional neural network，MCNN）[20]结构如图4所示。

卷积层使用尺寸为$r \times k$的卷积核对关键情感片段向量s进行上下卷积，以提取关键情感片段的局部特征c_i。

$$c_i = f(m \cdot s[i:i+r-1] + b) \quad (14)$$

其中，m表示尺寸为$r \times k$的卷积核；$s[i:i+r-1]$表示情感片段向量s中从第i行到$i+r-1$行的r行向量；b表示偏置。f表示激活函数，常用的激活函数有Sigmoid、tanh、ReLU，本文将ReLU函数作为卷积层的激活函数。

卷积核在情感片段向量s中从上向下以步长1进行卷积，最终得到当前卷积核提取的全部局部特征向量C。

$$C = \{c_1, c_2, \cdots, c_{L^S - r + 1}\} \quad (15)$$

为了减少模型参数，提高模型的收敛速度，一般在卷积后进行池化操作。常用的池化操作有平均池化和最大池化。本文选取最大池化提取C中的最大值来表示局部特征。

$$d_i = \max(C) \quad (16)$$

最后，将所有池化后的局部特征进行拼接，形成经过MCNN抽取的深层次特征d。

$$d = \{d_1, d_2, \cdots, d_n\} \quad (17)$$

为了突出情感片段的深层次特征d中的关键信息对模型的贡献，减少无效特征对模型的干扰，在MCNN后引入注意力机制。

$$a_d = \mathrm{softmax}(W_d \cdot d + b_d) \quad (18)$$

$$U = a_d \cdot d \quad (19)$$

其中，W_d、b_d为注意力网络可调节参数，d为经过MCNN抽取后的特征，a_d为注意力权重，U为包含注意力权重的关键情感片段特征。

3.3.3 情感集合特征提取

为了进一步抽取情感集合的情感特征，本文采用全连接神经网络从情感集合向量e中进行抽取。为了增大关键信息的权重，在全连接网络后引入注意力机制。

$$h = f(W \cdot e + b) \quad (20)$$

$$a_h = \mathrm{softmax}(W_h \cdot h + b_h) \quad (21)$$

$$P = a_h \cdot h \quad (22)$$

其中，h为抽取后的情感特征；W为全连接网络权重；b为全连接网络偏置；$f(\cdot)$为激活函数，本文将ReLU函数作为激活函数；a_h为注

意力权重；W_h、b_h为注意力网络可调节参数；P为包含注意力权重的情感集合特征。

3.4 加权融合层

加权融合层将从原始文本中提取的特征向量V、关键情感片段中提取的特征向量U、情感集合中提取的特征向量P三者进行拼接，形成最后的全局情感特征向量g。

$$g=[V,U,P] \quad (23)$$

从原始文本部分、语义规则部分、情感集合部分提取的特征重要程度各不相同，为了突出从关键部分提取的特征对情感分析任务的影响，在全局情感特征向量g后引入注意力机制。

$$a_g=\text{softmax}(W_g \cdot g+b_g) \quad (24)$$

$$G=a_G \cdot g \quad (25)$$

其中，W_g、b_g为注意力网络可调节权重和偏置，a_g为注意力权重，G为包含注意力权重的全局情感特征向量。

3.5 输出层

将包含注意力权重的全局情感特征G输入分类器，从而得到输入文本最终的所属类别。

$$p=\text{softmax}(w_p \cdot G+b_p) \quad (26)$$

其中，w_p为权重系数，b_p为偏置，p为预测的所属类别概率。

3.6 模型训练

本文使用反向传播最小化交叉熵损失函数的方式进行模型的训练。

$$L=-\sum_{i=1}^{D}\sum_{j=1}^{C} y_i^j \log \hat{y}_i^j \quad (27)$$

其中，L表示交叉熵损失，D表示训练集，C表示情感分析任务中的类别集合，y_i^j表示第i个样本的真实标签，$\hat{y}_i^j$表示模型预测的第i个样本的概率。

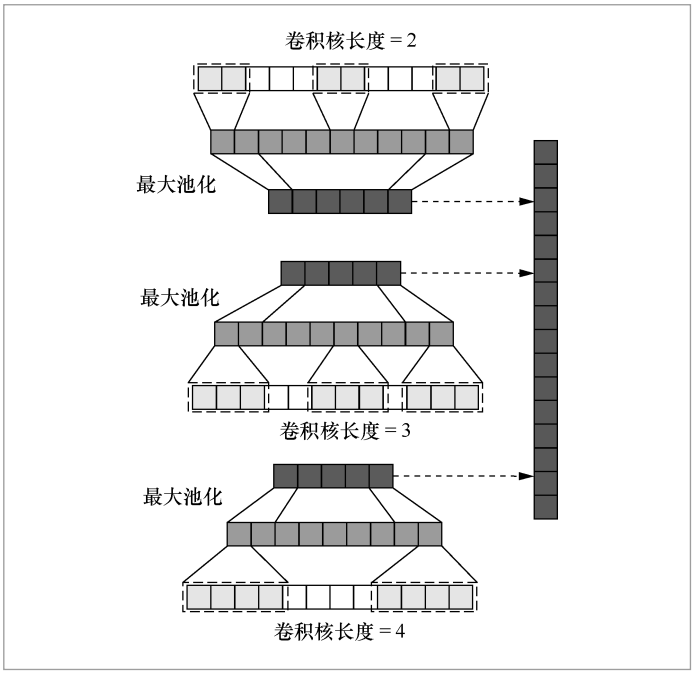


图 4 MCNN 结构

4 实验与分析

4.1 实验环境

本文的实验环境为：操作系统是Ubuntu 16.04，CPU是Intel Core i7-8750，GPU是GeForce GTX 1060，内存是DDR4 16 GB，显存大小是6 GB，深度学习框架是TensorFlow 2.4.0、Keras 2.4.3，开发工具是PyCharm 2020.1.1。

4.2 实验数据

数据集：本文使用两种数据集对所提方法进行验证，数据集1为中国科学院的谭松波整理的酒店评论数据集ChnSentiCorp，其共有10 000篇评论语料，分为4个子数据集。本文选用ChnSentiCorp-Htl-ba-6000进行实验，该语料正样本和负样本（即正面评论和负

面评论）各3 000篇，示例见**表1**。数据集2采用Data Fountain的O2O商铺食品安全相关评论数据，示例见**表2**。

词典：本文使用的情感词典为大连理工大学的中文情感词本体库，程度副词词典和否定词词典使用知网中文词库HowNet，见**表3**。使用的连词词典为人工梳理而得，见**表4**。

4.3 实验参数

超参数的设置会直接影响CLKDL方法的性能，按**表5**进行调参后本文所提方法达到最优。

4.4 评价指标

使用精准率precision、召回率recall以及两者的调和平均测度F1来衡量模型的性能。计算式如下。

$$\text{precision}=TP/(TP+FP) \quad (28)$$

$$\text{recall}=TP/(TP+FN) \quad (29)$$

$$F_1=\frac{2\cdot\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}} \quad (30)$$

其中，TP表示标记为正例、模型预测也为正例的样本数，FP表示标记为负例、模型预测为正例的样本数，FN表示标记正例、模型预测为负例的样本数。

4.5 实验结果及分析

为了验证使用CLKDL方法进行中文文本情感分析的有效性，本文使用如下3种方法在酒店评论和商铺食品评论两种数据集上进行实验，分别为基于语言知识的方法、基于深度学习的方法、CLKDL方法。实验结果见**表6**和**表7**，并对结果进行分析。

基于语言知识的方法：该方法首先利用句型规则、句间规则对原始文本进行语义分析，并计算其在语义规则角度上的得分；然后利用情感词典、程度副词词典、否定词词典对原始文本中的词进行分析，并计算其在情感词典角度上的得分；最后综合计算出原始文本的情感得分，完成中文文本的情感分析的任务。

基于深度学习的方法：该方法首先利用BERT对原始文本进行词嵌入以提取原始文本的语义表达，然后分别利用

表1 ChnSentiCorp 数据集示例

积极	消极
酒店环境、房间档次都很不错，服务水平专业。住了两天，感觉确实不错，二楼的自助烧烤给人宾至如归的感觉	房间装修陈旧，下水管堵塞，晚上折腾了两个多小时还是没有修好
非常好的酒店，四星级的标准，完全超值的享受，服务非常好	环境一般，住了之后让人感觉价格和服务不成正比
地点不错，邻近师范大学，到步行街走路15 min。房间很大很舒适	我肯定不会再住在这里了，太陈旧了，霉味太重，感觉不好

表2 O2O 商铺食品安全相关评论数据示例

积极	消极
一如既往地好吃，希望可以开到其他城市	跟平时吃的味道一样，本来吃得很开心的，但是突然吃到一只虫，把我吐惨了，太恶心了，再也不吃了
老板很热情，店虽然小，但是味道非常好，很实惠	无良商家，排骨不是排骨就算了，肉还是臭的，臭得熏天，太恶心了
菜品不是特别多，但是这个价位已经很划算了	好咸，不好吃，非常不好吃

BiLSTM提取原始文本的序列特征，利用MCNN提取原始文本的多层次语义特征，利用全连接神经网络提取原始文本的全局特征，接着利用注意力机制突出关键信息对模型的贡献，最后将含有注意力权重的3种特征进行融合，并利用sotfmax分类器完成中文文本的情感分析任务。

CLKDL：该方法与基于深度学习方法的网络结构相同，不同的是该方法将语言知识融合进基于深度学习的方法中。具体地，首先利用句间规则、句型规则从原始文本中抽取出关键情感片段，然后利用情感词典、程度副词词典、否定词词典从关键情感片段中抽取出情感集合，最后将原始文本、关键情感片段、情感集合作为模型的输入，完成中文文本的情感分析任务。

对上述实验结果进行分析发现，在上述两种数据集上，CLKDL的性能均高于基于语言知识的方法和基于深度学习的方法。这是因为单纯基于语言知识的方法忽略了文本蕴含的深层次特征，单纯基于深度学习的方法忽略了语法规则等语言知识信息。而CLKDL不仅考虑了语法规则等语言知识信息，降低了中文文本的复杂性，分析出关键情感信息，又利用深度学习模型提取了文本的深层次特征，从而得到了较好的性能，验证了所提方法的有效性。

5 结束语

目前深度学习模型大多是基于数据驱动的方法，其忽略了语义规则、情感集合等语言知识，导致无法充分提取文本特征。针对这一问题，本文提出了一种结合语言知识和深度学习的情感分析方法CLKDL。首先利用语义规则和情感集合将原始文本分为3个部分，即原始文本部分、关键情感片段部分、情感集合部分，

表3 情感语言库

词典种类	数量/个	实例
情感词	27 466	索然寡味、一塌糊涂、杰出、宝贝
否定词	59	不必、徒然、毋庸、不
程度副词	219	极其、极度、非常、很

表4 连词词典

词性	数量/个	实例
转折	10	不料、但是、偏偏、岂知
递进	9	并且、更有甚者、况且、甚至
假设	12	那么、若、倘若、假如
因果	14	因此、所以、以致、以便
让步	10	尽管、虽然、纵使、即使

表5 模型参数设置

参数	值
BiLSTM隐层节点数	128
MCNN隐层节点数	256
DENSE隐层节点数	128
Kernel数量	3
Kernel_size	3、4、5
优化器	Adam
学习率	2×10^{-5}
Batch_size	32
词嵌入维度	768
激活函数	ReLU

表6 ChnSentiCorp 数据集实验结果

方法	precision	recall	F1
基于语言知识的方法	70.26%	87.25%	77.84%
基于深度学习的方法	90.88%	91.81%	91.34%
CLKDL	96.12%	95.53%	95.82%

表7 O2O 商铺食品安全评论相关评论数据实验结果

方法	precision	recall	F1
基于语言知识的方法	95.92%	82.96%	88.97%
基于深度学习的方法	95.30%	97.66%	96.47%
CLKDL	98.33%	98.12%	98.22%

然后分别利用深度学习模型抽取文本特征并进行加权融合，最后利用分类器进行情感极性判断。实验结果表明，所提方法能够有效提高情感分析的性能。本文所提方法是从中文文本的角度进行的建模，下一步计划将该方法应用到其他语言，并进行文本情感分析。

参考文献:

[1] HATZIVASSILOGLOU V, MCKEOWN K R. Predicting the semantic orientation of adjectives[C]//Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1997: 174–181.

[2] WU J S, LU K, SU S Z, et al. Chinese micro-blog sentiment analysis based on multiple sentiment dictionaries and semantic rule sets[J]. IEEE Access, 2019, 7: 183924–183939.

[3] 赵妍妍, 秦兵, 石秋慧, 等. 大规模情感词典的构建及其在情感分类中的应用[J]. 中文信息学报, 2017, 31(2): 187–193.
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.

[4] XU G X, YU Z H, YAO H S, et al. Chinese text sentiment analysis based on extended sentiment dictionary[J]. IEEE Access, 2019, 7: 43749–43762.

[5] KESHAVARZ H, ABADEH M S. ALGA: adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs[J]. Knowledge-Based Systems, 2017, 122: 1–16.

[6] 李继东, 王移芝. 基于扩展词典与语义规则的中文微博情感分析[J]. 计算机与现代化, 2018(2): 89–95.
LI J D, WANG Y Z. Sentiment analysis of Chinese microblog based on expand-dictionary and semantic rule[J]. Computer and Modernization, 2018(2): 89–95.

[7] ZHANG S X, WEI Z L, WANG Y, et al. Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary[J]. Future Generation Computer Systems, 2018, 81: 395–403.

[8] PANG B, LEE L, VAITHYANATHAN S. Thumbs up?: sentiment classification using machine learning techniques[C]// Proceedings of the ACL-02 Conference on Empirical methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2002: 79–86.

[9] 苏莹, 张勇, 胡珀, 等. 基于朴素贝叶斯与潜在狄利克雷分布相结合的情感分析[J]. 计算机应用, 2016, 36(6): 1613–1618.
SU Y, ZHANG Y, HU P, et al. Sentiment analysis research based on combination of naive Bayes and latent Dirichlet allocation[J]. Journal of Computer Applications, 2016, 36(6): 1613–1618.

[10] 刘栋军, 王宇涵, 凌文芬, 等. 基于脑机协同智能的情绪识别[J]. 智能科学与技术学报, 2021, 3(1): 65–75.
LIU D J, WANG Y H, LING W F, et al. Emotion recognition based on brain and machine collaborative intelligence[J]. Chinese Journal of Intelligent Science and Technology, 2021, 3(1): 65–75.

[11] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1746–1751.

[12] 胡荣磊, 芮璐, 齐筱, 等. 基于循环神经网络和注意力模型的文本情感分析[J]. 计算机应用研究, 2019, 36(11): 3282–3285.
HU R L, RUI L, QI X, et al. Text sentiment analysis based on recurrent neural networks and attention model[J]. Application Research of Computers, 2019,

36(11): 3282-3285.

[13] 李洋, 董红斌. 基于CNN和BiLSTM网络特征融合的文本情感分析[J]. 计算机应用, 2018, 38(11): 3075-3080.
LI Y, DONG H B. Text sentiment analysis based on feature fusion of convolution neural network and bidirectional long short-term memory network[J]. Journal of Computer Applications, 2018, 38(11): 3075-3080.

[14] 宋婷, 陈战伟, 杨海峰. 基于分层注意力网络的方面情感分析[J]. 大数据, 2020, 6(5): 82-91.
SONG T, CHEN Z W, YANG H F. Aspect sentiment analysis based on a hierarchical attention network[J]. Big Data Research, 2020, 6(5): 82-91.

[15] 谢润忠, 李烨. 基于BERT和双通道注意力的文本情感分类模型[J]. 数据采集与处理, 2020, 35(4): 642-652.
XIE R Z, LI Y. Text sentiment classification model based on BERT and dual channel attention[J]. Journal of Data Acquisition and Processing, 2020, 35(4): 642-652.

[16] 邱宁佳, 王晓霞, 王鹏, 等. 融合语法规则的双通道中文情感模型分析[J]. 计算机应用, 2021, 41(2): 318-323.
QIU N J, WANG X X, WANG P, et al. Analysis of double-channel Chinese sentiment model integrating grammar rules[J]. Journal of Computer Applications, 2021, 41(2): 318-323.

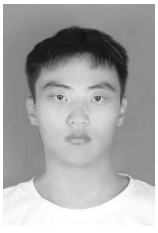
[17] CHUNG Y A, ZHU C G, ZENG M. SPLAT: speech-language joint pre-training for spoken language understanding[C]// Proceedings of 2021 Conference of the North American Chapter of Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2021: 1897-1907.

[18] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

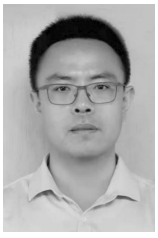
[19] GRAVES A. Supervised sequence labelling with recurrent neural networks[M]. Heidelberg: Springer, 2012.

[20] CUI Z C, CHEN W L, CHEN Y X. Multi-scale convolutional neural network for time series classification[J]. arXiv preprint, 2016, arXiv:1603.06995.

作者简介



徐康庭(2000-), 男, 北方工业大学信息学院在读, 主要研究方向为机器学习、自然语言处理。



宋威(1980-), 男, 博士, 北方工业大学信息学院教授、博士生导师, 主要研究方向为数据挖掘、推荐系统。

收稿日期: 2021-07-30
通信作者: 徐康庭, bjxukangting@163.com
基金项目: 国家自然科学基金资助项目(No.61977001)
Foundation Item: The National Natural Science Foundation of China (No.61977001)