文章编号：1003-0077(2020)08-0094-11

基于粗糙集和多通道词向量的中文文本情感特征分析

陈 波[1]，谢 珺[1]，苗夺谦[3]，王雨竹[1]，续欣莹[2]

(1. 太原理工大学 信息与计算机学院，山西 晋中 030600；
2. 太原理工大学 电气与动力工程学院，山西 太原 030024；
3. 同济大学 电子与信息工程学院，上海 201804)

摘 要：粗糙集是一种能够有效处理不精确、不完备和不确定信息的数学工具，粗糙集的属性约简可以在保持文本情感分类能力不变的情况下对文本情感词特征进行约简。针对情感词特征空间维数过高、情感词特征表示缺少语义信息的问题，该文提出了 RS-WvGv 中文文本情感词特征表示方法。利用粗糙集决策表对整个语料库进行情感词特征建模，采用 Johnson 粗糙集属性约简算法对决策表进行化简，保留最小的文本情感词特征属性集，之后再对该集合中的所有情感特征词进行词嵌入表示，最后用逻辑回归分类器验证 RS-WvGv 方法的有效性。另外，该文还定义了情感词特征属性集覆盖力，用于表示文本情感词特征属性集合对语料库的覆盖能力。最后，在实验对比的过程中，用统计检验进一步验证了该方法的有效性。

关键词：属性约简；情感特征提取；词向量；情感分类

中图分类号：TP391 **文献标识码**：A

Chinese Text Sentiment Feature Analysis Based on Rough Set and Multi Channel Word Vector

CHEN Bo[1], XIE Jun[1], MIAO Duoqian[3], WANG Yuzhu[1], XU Xinying[2]

(1. School of Information and Computer, Taiyuan University of Technology, Jinzhong, Shanxi 030600, China;
2. School of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan, Shanxi 030024, China;
3. School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: Rough set is a mathematical tool that can greatly reduce the dimension and number of text sentiment word features while keeping the ability of text sentiment classification unchanged. Aiming at the problem that the text sentiment word feature dimension is too high and the sentiment word feature representation lacks semantic information, this article proposes a novel Chinese text sentiment word feature representation method named RS-WvGv. The decision table of rough set is used to model the text sentiment word feature of the whole corpus. The Johnson attribute reduction algorithm is applied to simplify the decision table and get the minimum set of text sentiment word feature attributes. And then based on the word embedding of all the sentiment feature words in the set, the RS-WvGv method is verified with logistic regression classifier in the experiment.

Keywords: attribute reduction; sentiment feature extraction; word vector; sentiment classification

0 引言

随着互联网技术的不断发展，网络已经成为人们自由发表言论的聚集地，人们的参与感逐步增加的同时，网络也成为社会舆情的重要载体。因此，互联网产生了大量用户参与的包括对事件的观点、对物品的评论等文本信息，这些信息或多或少地表达

收稿日期：2019-11-24 **定稿日期**：2019-12-27

基金项目：山西省应用基础研究项目(201801D221190，201801D121144)

了人们的情感倾向，而情感是影响人类行为的重要因素之一。当人们需要做出决策的时候，往往会以他人的观点或者意见作为行动决策的参考。然而，评论文本信息与日俱增，仅仅依靠人工收集处理则变得十分繁琐和低效。因此，需要利用情感分析技术快速获得互联网用户的情感倾向，从而帮助决策者做出更可靠的决策行动。

情感分类是情感分析任务的一个重要组成部分。情感分类常常被当作是一个文本分类问题，是因为其根据文本所表达的情感信息，将给定的文本分为积极的情感和消极的情感[1]。目前文本情感分类算法主要分为基于监督学习、基于半监督学习、基于无监督学习的情感分类方法。基于监督学习的情感分析流程图如图 1 所示。Pang 等人[2]第一次将监督学习的方法应用到文本情感分类任务中，文中比较了支持向量机(SVM)、朴素贝叶斯(NB)和最大熵模型(ME)三种分类方法。李平等人[3]将逻辑回归(LR)用于文本情感分类任务，验证了逻辑回归在文本情感分类任务中的有效性。除了传统的情感特征构建和机器学习相结合的情感分析方法之外，深度学习模型也已经广泛应用于情感分类任务，很好地解决了传统的情感特征表示方法存在的高维稀疏的问题。Yang 等人[4]提出了一种基于卷积神经网络(CNN)的中文文本情感分析模型，实验结果表明，词向量 Word2Vec 和卷积神经网络构建的神经网络模型可以有效提高情感分类的性能。赵富等人[5]提出一种基于双注意力的双向长短时记忆网络模型(Bi-LSTM)，融合了字、词和词性的深层语义表达特征，提高了情感分类模型的准确率。

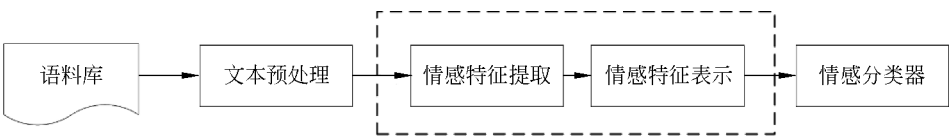


图 1　基于监督学习的情感分析流程图

对于基于监督学习的文本分类任务而言，分类效果很大程度上取决于文本特征的提取和表示。特征提取(feature extraction)又称属性约简，是训练机器学习模型的关键步骤。由于特征的数量较为庞大，且多数特征为普适特征，对特定的分类任务而言，存在较高的冗余。此外，特征与特征之间存在较强的相关性，也会增加训练模型的时间。因此，通过合适的特征提取方法可以筛选掉冗余的特征，减少特征的整体数量，减少时间花费，同时提高分类的准确率。

字词是组成文本特征的最小结构单元，而关键词是刻画文本内容的重要特征，提取关键词作为文档特征对于文本分析有很大帮助。常用的文本关键词提取方法有文档逆文档频率 TF-IDF (term frequency-inverse document frequency)法、基于图排序法 TextRank 等。TF-IDF 综合考虑了词在单个文档中出现的频率 TF 以及该词在整个文档集中的权重 IDF，按照计算得到的每个词的 TF-IDF 值进行降序排列，提取前几位的关键词。TextRank 算法是一种用于文本的基于图的排序算法，通过把文本分割成若干组成单元并建立图模型，利用局部词汇之间的关系对后续关键词进行排序，仅利用单篇文档本身的信息即可实现关键词提取[6]。

对于文本情感分类任务来说，文本的情感特征是指一个文本所能表达出的情感倾向的文本特征，一般是指情感词特征。例如，如果一条文本中的正向情感词较多，那么文本的情感倾向就为积极，反之则为消极。大多数情况下，情感倾向的表达是由几个，甚至是一个情感词特征来决定的，而不是由所有文本情感词特征之间的微小差异来决定的，提取合适的文本情感词特征十分必要。

本文针对文本情感词特征空间高维、冗余，情感词特征表示缺乏语义信息的问题，从文本数据中抽取有效的文本情感词特征并进行有效的文本表示。

1　相关工作

1.1　情感词特征提取

张克君等人[7]考虑到经典的 CBOW 模型在进行语言模型建模的时候只考虑到词本身与其上下文的联系，而未考虑到该词与整个文本语料之间的关系，该文对其进行改进和优化，引入了 TF-IDF 文本关键词计算方法，把 TF-IDF 值较高的几个词作为文本关键词和中间词的上下文来一起预测中间词，提升了词向量表示模型的质量。徐立[8]提出一种融合多特征的 TextRank 中文文本关键词(包括词频、

词长、词语位置以及词性)提取方法,并将融合后的关键词权重计算公式用于 TextRank 算法的候选关键词得分公式,提升文本关键词提取的准确率。

以上方法在进行关键词提取的过程中存在信息丢失的问题,有些低频词也可能携带一定的情感信息,对最终的情感分析也十分重要。粗糙集(rough set,RS)作为一种能有效处理不精确、不完备和不确定信息的数学工具[9],也被大量用于文本数据的特征词挖掘任务。张志飞等人[10]将粗糙集的属性约简过程和文本分类任务的特点相结合,提出了一种改进的快速约简方法(IQR),用于文本的词特征提取,结果表明粗糙集属性约简的方法可以有效降低文本词特征的空间维度。

1.2 情感特征分布式词向量表示

文本情感词特征表示是一个将情感词特征向量化的过程,与传统的浅层词向量表示模型不同,基于深度学习的 Word Embedding 表示方法很好地嵌入了词的上下文信息。Word Embedding 多数使用文本的 One-Hot 编码或者 N-gram 等浅层文本表示作为输入,利用神经网络的非线性结构映射学习词的深层语义信息,并将其映射到一个低维稠密的向量空间。孙晓等人[11]尝试将 unigrams 和 bigrams 特征作为浅层特征输入到深度信念网络 DBN 中,得到深层 unigrams 和 bigrams 特征,实验结果说明,深层特征比浅层特征包含更多的文本语义信息,分类准确率有所提高。

Hinton[12]提出的词向量(distributed representation)表示模型,将词映射到一个低维稠密的实数向量空间中,克服了传统词向量高维稀疏的缺点。Word2Vec[13]利用神经网络语言模型(包括 CBOW 和 Skip-Gram 模型),把每个词的 One-Hot 稀疏向量表示映射成一个统一长度的 K 维向量,把对文本内容的处理转换成对 K 维向量空间中的向量运算。同时,Word2vec 可以很好地利用向量空间中向量之间的距离来度量文本之间的语义相似度。唐明等人[14]在分析了传统的 One-Hot 编码方式和词袋模型在文本表示方面的高维稀疏的特点,提出了一种结合 Word2Vec 的文档向量计算方法,用于中文文档分类。

FastText[15-16]加入了 N-gram 特征,其结构和 Word2Vec 中的 CBOW 模型的结构类似,相较于 Word2Vec 输入是窗口内的局部信息,FastText 输入的是整个句子的上下文信息,而且还包括字符级别的 N-gram 子字信息,将局部的词序信息考虑在内,这样学习得到的词向量在一定程度上可以解决未登录词的向量表示问题。Joulin 等人[16]通过将整篇文档的词及子字 N-gram 向量叠加求平均之后用来表示最终的文档向量,用来处理文档分类任务,取得了较好的效果。

以上的 Word2Vec 方法在构建神经网络语言模型的时候,固定窗口的大小,仅仅是获取局部语料的语义特征,没有有效地利用全局的词汇共现统计信息。LSA 矩阵分解词向量是主题模型的一种,在进行 LSA 矩阵分解之前,需要构造词汇和文本的矩阵,然后再对词汇—文本矩阵进行奇异值分解,将文档映射到低维的潜在语义空间,最后对分解后的矩阵进行降维,得到最终的词向量表示。LSA 虽然有效地利用了全局信息,但是在词汇类比方面表现很差。同时,利用 SVD 求解构建词向量计算复杂度太高。

GloVe[17]相较于 Word2Vec 词向量来说,更加关注词语的共现统计信息,这样可以根据词语实际所处的语料场景来学习词语的语义信息。首先根据语料库构建一个共现矩阵,其中每个元素代表单词和上下文词在特定窗口内共现的次数。之后再根据词频矩阵分解得到全局的上下文语义信息。GloVe 结合了 LSA 和 Word2Vec 的优点,但是由于 GloVe 要统计共现概率,需要事先准备固定的语料信息,所以无法进行在线学习。

以上所有的词向量表示方法都是静态的词向量表示,一旦语料库固定,最终学习到的词向量就是固定不变的。但是对于情感分析任务而言,尤其是中文文本的情感分析,很多中文词语所表达的语义会随着实际的语境而发生变化,存在一词多义现象。结合特定的上下文语境对静态词向量进行微调的方法被大量研究。Peters 等人[18]提出了 Emlo 模型利用双向长短时记忆网络(BiLSTM)来生成词的上下文表示向量,在获得预先训练的词向量之后根据实际的语料数据的上下文信息进行词向量的动态调整。Devlin 等人[19]提出了基于 Transformer[20]的双向编码器模型 BERT,一定程度上缓解了一词多义的问题。

本文从文本情感词特征提取和文本情感词特征表示两个方面入手,对中文文本情感分类进行研究。针对文本情感词特征维度高、情感词特征表示稀疏的问题,结合了粗糙集特征提取模型和 Word2Vec 和 GloVe 词向量模型对酒店评论数据[21]的情感特

征进行融合表示，实验结果表明，粗糙集在进行情感词特征提取的时候可以很大程度地保留文本的情感特征词，丢弃与情感预测无关的其他特征词，之后再结合 Word2Vec 和 GloVe 词向量模型对所选择的情感词特征进行嵌入表示，解决了传统的文本情感词特征表示的高维稀疏性。

2　基于粗糙集和 WvGv 的情感词特征表示模型

本文的情感分析过程如图 2 所示，包括文本预处理、情感词特征提取与表示、情感分类三个部分。

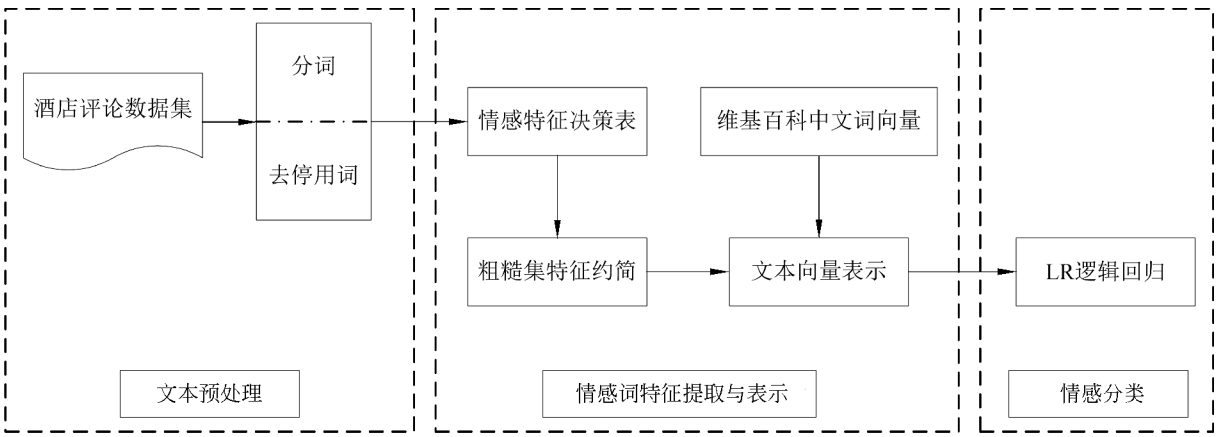


图 2　基于粗糙集和 WvGv 的情感分析流程图

2.1　基于粗糙集属性约简的情感词特征提取方法

基于粗糙集的特征提取方法主要分为文本决策表的建立与离散、基于属性重要度的条件属性约简两个部分。其优势在于粗糙集属性约简可以在无须提供问题之外的任何先验信息的情况下提取对数据分类学习的重要属性，分析隐藏在数据中的事实，同时又可以保证决策表的分类能力不改变[9]。

在粗糙集理论中，本文应用了以下几个重要定义。

定义 1　信息表[9,22]

设四元有序元组 $S=(U,A,V,f)$ 表示信息表，其中 $U=\{x_1,x_2,x_3,\cdots,x_{|U|}\}$ 是有限非空集合，称为论域或者对象全体，U 中的元素称为对象；A 也是一个有限非空集合，称为属性全体，A 中的元素称为属性；$V=\bigcup\limits_{a\in A}V_a$ 是属性的集合，V_a 是属性 a 的值域范围；$f:U\times A\to V$ 是一个信息函数，它为每个对象的每个属性赋予一个值，即 $f(x,a)\in V_a$，$x\in U$，$a\in A$。通常属性集合 A 可以划分为条件属性集 C 和决策属性集 D 两个子集，即 $A=C\cup D$，此时的信息表就称为决策表。

定义 2　等价关系[9,22]

设 $P\subset A$，$x_i,x_j\in U$，定义二元关系 IND(P)为论域 U 上的等价关系，如式(1)所示。

$$\mathrm{IND}(P)=\{(x_i,x_j)\in U\times U\mid \forall_p\in P,\ p(x_i)=p(x_j)\} \tag{1}$$

定义 3　属性约简[9,22]

属性约简是粗糙集理论中的重要一环，属性约简所得到的属性集是保证正域不变的最小属性集合。设 U 为决策表所讨论的论域，C 和 D 分别为定义在论域 U 上的条件属性：集合、决策属性集合。计算出二维决策表中决策属性集 D 相对于条件属性集 C 的正域 $\mathrm{POS}_{\mathrm{IND}(C-\{R\})}(\mathrm{IND}(D))$，如果：

$$\mathrm{POS}_{\mathrm{IND}(C)}(\mathrm{IND}(D))=\mathrm{POS}_{\mathrm{IND}(C-\{R\})}(\mathrm{IND}(D)) \tag{2}$$

则称 $C^*=\{C-\{R\}\}$ 为决策表的相对约简集。

在进行属性约简之前，首先要对现有的评论数据建立文本情感词特征决策表，设给定评论文档集合为 $S=\{s_1,s_2,\cdots,s_n\}$，经过分词等文档预处理过程，得到的情感特征候选词集合为 $C=\{c_1,c_2,\cdots,c_m\}$，在特征空间中评论文档被用特征词表示为特征向量，各维度的值为 0 或者 1，1 代表该特征词在当前文档中出现，0 代表不出现。若以文档 S 作为论域，特征词集合作为条件属性集 C，文档的情感类别集合 $D=\{d_1,d_2,\cdots,d_k\}$ 作为决策属性集 D，则文本情感分类问题的二维决策表如表 1 所示。

表 1 文本情感分类问题决策表

U/S	C					D
	c_1	c_2	c_3	…	c_m	
s_1	0	1	0	…	1	d_1
s_2	1	1	0	…	0	d_2
…	…	…	…	…	…	…
s_n	1	0	1	…	0	d_k

在构建完文本情感词特征决策表之后，本文采用粗糙集中的 Johnson 属性约简算法[23]约简特征词属性集 C，得到约简过后的特征词属性集 C^*，称为情感词特征属性集。设 C^* 表示约简过后的特征属性集，算法 1 是 Johnson 属性约简算法的具体过程。

算法 1 Johnson 属性约简算法

Step 1 令 $S=(U,C\cup D,V,f)$，其中条件属性 $C=\cup a_i$，$i=1,2,\cdots,n$；

Step 2 计算可分辨矩阵 M，$M=\{m_{ij}: m_{ij}\neq\varnothing\}$；

Step 3 计算属性 a_i 在 M 中出现的次数 $\omega_{a_i}(M)$；

Step 4 选择使 $\omega_{a_i}(M)$ 最大的属性，记为 a，计算 $C^*=C^*\cup\{a\}$；

Step 5 将 M 中包含 a 属性的全部删除；

Step 6 如果 $M\neq\varnothing$，停止计算，否则转至 Step 3。

通过采用 Johnson 算法对所选文本词特征进行属性约简，大大降低了文本情感词特征的维数。

定义 4 词特征覆盖力

设训练集文档集 S 的个数为 n，如果特征词 c_i，$i\in(1,m)$ 出现在文档 s_j，$j\in(1,n)$ 中，那么表示该条文档至少被一个特征词覆盖，如果某条文档不包含特征词集中的任何特征词 c_i，那么该条文档未被任何特征词覆盖。设未被任何特征词覆盖的文档个数为 l，$l\in(0,n)$，那么特征覆盖力 γ 表示如式(3)所示。

$$\gamma=\frac{n-l}{n} \tag{3}$$

其中，$\gamma\in(0,1)$，γ 越接近 1，说明特征提取的结果越好，反之亦然。

2.2 情感特征词向量表示

本文利用 Jieba 分词工具对训练集文档进行分词、去停用词预处理，之后利用 Wiki 官网下载的中文语料作为 Word2Vec 和 GloVe 的语料库进行中文词向量训练，训练过程如图 3 所示。

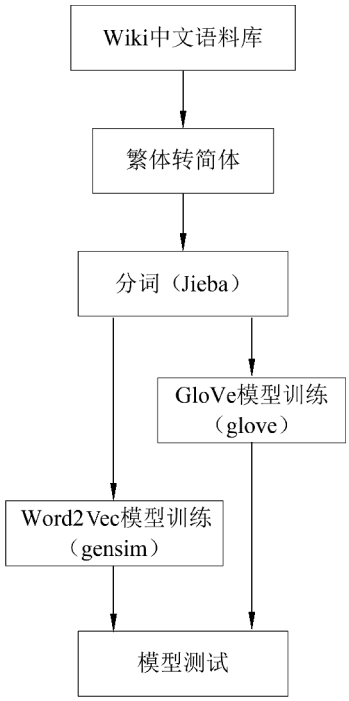


图 3 Word2Vec 和 GloVe 词向量模型训练流程图

2.3 基于 RS-WvGv 的中文文本情感分类算法

在获得训练好的中文词向量 Word2Vec 和 GloVe 之后，本文结合两种词向量在进行向量表示过程中的不同优势，对两种向量进行融合，提出一种多通道的文本情感词特征表示方法 RS-Word2Vec⊕GloVe（RS-WvGv）。首先利用粗糙集对酒店评论语料库[21]建立如表 1 所示的决策表，采用 Johnson 属性约简算法对决策表进行约简，得到情感特征词集 C^*，之后用训练好的中文词向量分别对训练集文本文档进行向量化表示，最后用得到的训练集特征向量集合构建逻辑回归情感分类器。

图 4 为 RS-WvGv 模型图，按照本模型的描述，算法 2 为详细的中文文本情感分类算法。

算法 2 基于 RS-WvGv 的中文文本情感分类算法

输入：待分类文本文档 $S=\{s_1,s_2,\cdots,s_n\}$

输出：文档类别 d_i，$d_i\in D$，$D=\{d_{\mathrm{POS}},d_{\mathrm{NEG}}\}$

Step 1 使用 Jieba 分词，去停用词，获得情感特征词候选集 $C=\{c_1,c_2,\cdots,c_m\}$；

Step 2 构建文本情感分类决策表 $\mathrm{DT}=(S,C,D)$

① for each $c_i\in C$，$i\in(1,m)$，s_j，$j\in(1,n)$

② if c_i in s_j，

③ $c_i=1$

④ else $c_i=0$

⑤ $i=i+1$，$j=j+1$

⑥ 如果 $i=m$，$j=n$，停止计算，否则重复①－③

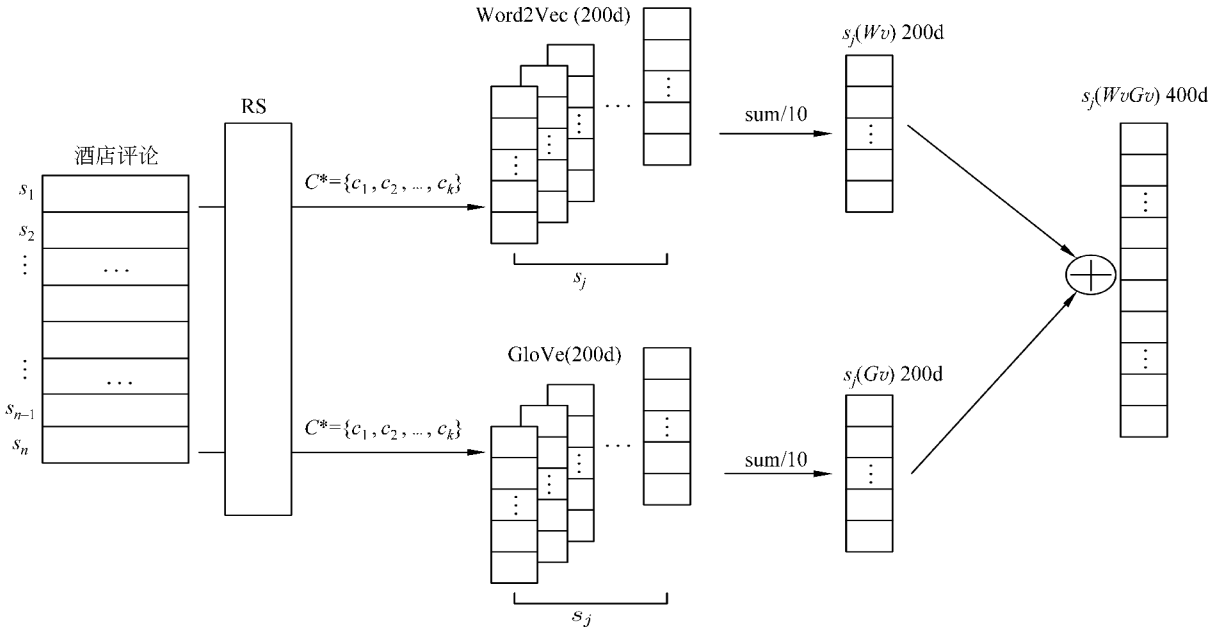


图 4　RS-WvGv 模型图

Step 3　Johnson 属性约简算法得到情感词特征属性集 $C^*=\{c_1,c_2,\cdots,c_k\}$

Step 4　文本文档向量表示

① for each $s_j\in S, j\in(1,n)$

②　if c_i in s_j, $c_i\in C^*$, $i\in(1,k)$

③　　load word2vec and glove

④　　$i=i+1, j=j+1$

⑤ 取所有特征词的词向量进行平均池化，得到文档 s_j 的两种向量表示 s_j(Wv)和 s_j(Gv)

⑥ Wv⊕Gv 向量拼接得到最后文档 s_j 的向量 s_j(WvGv)

⑦ 如果 $j=n$，停止加载，否则重复①－④

Step 5　逻辑回归情感分类器构建

将文档向量 s_j(WvGv)作为输入特征向量 x，输入到逻辑回归模型 $h_\theta(x)=g(\theta^Tx)$ 中。其中，θ 为中间参数，$g(x)$为 sigmoid 函数。

$$g(x)=\frac{1}{1+e^{-x}} \tag{4}$$

按照上述构建文本文档向量表示的过程，设定文档的统一长度为 10，对长度超出 10 的文档进行截断，长度不足 10 的文档进行补 0。在用两种词向量分别对文档进行向量化表示之后，取平均值作为分别的文档向量表示 Wv 和 Gv，在输入逻辑回归情感分类器之前，对两种文档向量进行拼接得到最终的文档向量表示 Wv⊕Gv。

3　实验结果与分析

3.1　数据集介绍

本文选取了 Tan 等人[21]整理的已经进行过正负情感极性标注的中文酒店评论文档作为实验数据，情感极性分为正向和负向。将实验数据分为 4 个数据集，每个数据集的正负评论数如表 2 所示，其中前 3 个数据集是平衡数据集，第 4 个数据集为非平衡数据集。为方便下文表示，本文用 H-1 代表 ChnSentiCorpHou-1，H-2 代表 ChnSentiCorpHou-2，H-3 代表 ChnSentiCorpHou-3，H-4 代表 ChnSentiCorpHou-4。

表 2　酒店评论数据集

数据集	正向评论数	负向评论数	是否平衡
ChnSentiCorpHou-1	1 000	1 000	是
ChnSentiCorpHou-2	2 000	2 000	是
ChnSentiCorpHou-3	3 000	3 000	是
ChnSentiCorpHou-4	7 000	3 000	否

3.2　实验参数设置

首先，本文采用如图 3 所示的词向量模型训练方法获得所需的词向量，具体的训练参数如表 3 所示。之后再用训练好的中文词向量模型对所提取的情感词特征进行向量化表示。

表 3　词向量模型训练参数表

词向量模型	Word2Vec (CBOW)	GloVe
size	200d	200d
window	5	15

续表

词向量模型	Word2Vec (CBOW)	GloVe
min_count	5	5
cbow_mean	1	NaN
hs	0	NaN
iter	15	15

其次,对于所提出的情感词特征表示方法,本文采用 scikit-learn[24] 提供的逻辑回归函数构建情感分类器,最后采用 10 折交叉验证的方式对 4 个数据集分别进行实验验证,其中留出 25%的数据作为测试集。实验过程中,采用网络搜索(GridSearchCV)的方法确定 LR 分类器的最优参数组合,待寻优参数包括正则化方式(L1 或 L2)和正则化强度 C。

3.3 评价指标

实验过程中采用情感词特征覆盖力 γ 作为情感词特征提取结果的评价标准,用情感词特征向量化的时间花费和准确率 Acc 作为分类结果的评价标准,准确率计算如式(5)所示,同时采用混淆矩阵结果(confusion matrix,CM)来直观地判别分类器的性能好坏。

混淆矩阵是一张总结分类器预测结果的情形分析表,混淆矩阵表如表 4 所示。从混淆矩阵结果中可以直观地看到分类器的预测结果,包括预测正确的结果和产生混淆的结果。另外,从混淆矩阵的结果中,可以分别计算得到正例和负例所对应的召回率(R)和精确率(P),计算如式(6)、式(7)所示。

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}} \tag{5}$$

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

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{7}$$

表 4 混淆矩阵表

真实标签	预测结果	
	Positive	Negative
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

3.4 对比实验

实验 1 情感词特征提取

训练集文本经分词过后统计共有 10 626 个词,构成情感特征候选词集,本文分别使用 TF-IDF、TextRank、RS 三种方法进行情感词特征提取。其中,TF-IDF 和 TextRank 选择前 10 000 个词作为情感特征词,粗糙集 RS 直接在整个特征词集里进行属性约简得到情感特征词。此外,与情感特征相关的词大部分为形容词(a)、副词(d)、形副词(ad)、动词(v)、副动词(vd)、名动词(va)和名形词(an),把以上词性的词作为预先定义的情感特征词,再次使用三种方法进行情感词特征提取,实验结果如表 5 所示。

表 5 特征提取结果对比表

特征词提取方法	情感特征词个数/特征词集覆盖力 r	预先定义的情感特征词个数/特征词集覆盖力
TF-IDF	10 000/0.992 7	3 858/0.978 5
TextRank	10 000/0.993 8	3 173/0.968 5
RS	**2 589**/0.997 5	**2 033**/0.996 5

实验 2 情感词特征表示

将本文提出的基于粗糙集和 WvGv 的情感词特征表示方法(RS-WvGv)与基于 TF-IDF 和 WvGv 的情感词特征表示方法(TF-IDF-WvGv)和基于 TextRank 和 WvGv 的情感词特征表示方法(TextRank-WvGv)以及基于所有词和 WvGv 的情感词特征表示方法(ALL-WvGv)的分类效果进行对比。在这之前,还直接利用了训练好的 Word2Vec 和 GloVe 中文词向量分别对三种情感词特征提取方法提取之后的结果进行向量化表示,最后使用逻辑回归分类器 LR 验证本文所提方法的有效性。

3.5 实验结果分析

实验 1 情感词特征提取结果分析

由表 5 统计的结果可知,基于粗糙集的情感词特征提取方法的词特征覆盖力 γ 接近于 1,也就是说,粗糙集在进行属性约简后,可以保证整个决策表的情感分类能力不变。同时,属性约简保留了大部分情感词特征,能用较少的情感词特征来表达文本的情感倾向信息,维度大大降低。

实验 2 情感词特征表示结果分析

用实验 2 中所介绍的 12 种方法在所用数据集上进行实验,使用准确率作为情感分类结果的评估标准。表 6、表 8、表 10 分别给出了 12 组实验在所用数据集上的准确率结果。表 7 和表 9 给出了不同情感词特征提取方法在特征向量化表示过程中所花费的时间。图 5~图 8 给出了四种方法在数据集

H-1 和 H-4 上的混淆矩阵结果图，其中正例 Positive 是类别为 0 的样本。

表 6　对比实验在酒店评论数据集上的准确率

情感词特征表示方法	H-1	H-2	H-3	H-4
ALL-Word2Vec	0.814	0.828	**0.834**	0.838
TextRank-Word2Vec	0.741	0.756	0.759	0.801
TF-IDF-Word2Vec	0.745	0.765	0.766	0.806
RS-Word2Vec	**0.830**	**0.834**	**0.834**	**0.858**

表 7　情感词特征向量化所花费的时间

情感词特征表示方法	H-1	H-2	H-3	H-4
ALL-Word2Vec/s	**2.907**	**5.148**	**8.136**	**12.948**
TextRank-Word2Vec/s	1.986	2.415	4.086	8.625
TF-IDF-Word2Vec/s	1.826	2.347	4.973	8.025
RS-Word2Vec/s	**1.854**	**2.020**	**4.667**	**8.351**

表 8　对比实验在酒店评论数据集上的准确率

情感词特征表示方法	H-1	H-2	H-3	H-4
ALL-GloVe	0.806	0.823	0.816	0.833
TextRank-GloVe	0.721	0.734	0.767	0.795
TF-IDF-GloVe	0.751	0.743	0.768	0.795
RS-GloVe	**0.810**	**0.830**	**0.832**	**0.852**

表 9　情感词特征向量化所花费的时间

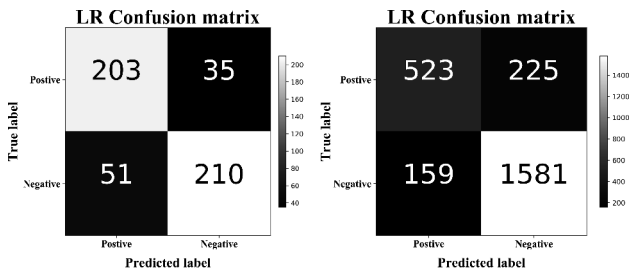
情感词特征表示方法	H-1	H-2	H-3	H-4
ALL-GloVe/s	**2.511**	**5.783**	**7.997**	**13.233**
TextRank-GloVe/s	1.884	3.431	4.793	9.864
TF-IDF-GloVe/s	1.862	3.523	4.422	9.680
RS-GloVe/s	**1.830**	**3.421**	**4.086**	**8.196**

表 10　对比实验在酒店评论数据集上的准确率

情感词特征表示方法	H-1	H-2	H-3	H-4
ALL-WvGv	0.828	0.832	0.841	0.846
TextRank-WvGv	0.751	0.774	0.771	0.805
TF-IDF-WvGv	0.761	0.769	0.787	0.810
RS-WvGv	**0.854**	**0.841**	**0.842**	**0.860**

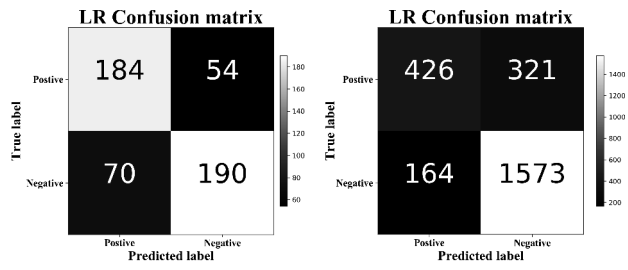
结合表 6 和表 7 的结果可以看出，基于粗糙集和 Word2Vec 的情感词特征表示方法（RS-Word2Vec）在准确率和特征向量化的时间花费上均有一定的优势，相较于未进行情感词特征提取的 ALL-Word2Vec 情感词特征表示方法在情感特征向量化表示的过程中节省将近一半的时间。在准确率方面，相较于另外两种情感词特征表示方法 TextRank-Word2Vec和 TF-IDF-Word2Vec 都有一定的提升。以上两种情感词特征提取方法均是概率优先原则，会丢失掉一些低频的情感特征词，而基于 RS 的情感词特征提取方法是基于全局的语料库进行建模，提取的结果保留了大部分低频的情感特征词，保留了更多的情感词特征。而且相对于未进行情感词特征提取的 ALL 来说，词汇的维度大大降低，为后续的情感词特征向量化表示节省时间开销。对比表 6 和表 8 可以看出，用 GloVe 对所提取的情感词特征进行向量化表示的结果准确率普遍低于用 Word2Vec 表示的结果。GloVe 词向量是基于概率模型学习得到的词向量表示，更关注相似词的语义信息，而 Word2Vec 更倾向于相邻词的语义信息，语义表达更加完整。

由表 10 可知，相较于单独的 Word2Vec 和 GloVe 向量，融合向量 Wv⊕Gv 对情感词特征的表示结果有一定的提升。三种情感词特征提取方法在用融合词向量表示之后的准确率均提高 1%到 2%。这表明基于多通道的情感词特征的向量表示既考虑了当前词在全局语料上的统计信息，也兼顾了当前词与局部语料的相似性，可以对提取过后的情感词特征进行更好的向量化表示。



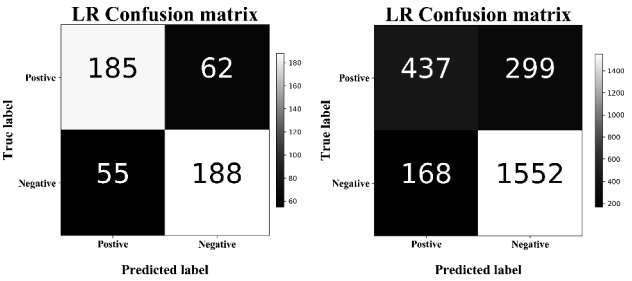
(a) H-1　　(b) H-4

图 5　基于 ALL-WvGv 的 LR 混淆矩阵图



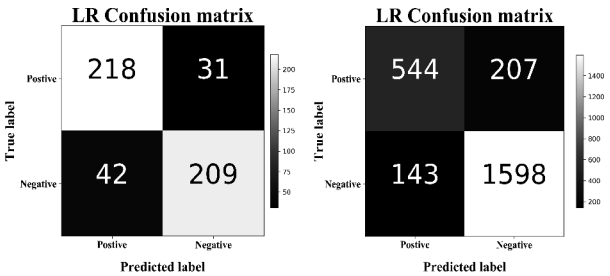
(a) H-1　　(b) H-4

图 6　基于 TextRank-WvGv 的 LR 混淆矩阵图



(a) H-1 (b) H-4

图 7 基于 TF-IDF-WvGv 的 LR 混淆矩阵图



(a) H-1 (b) H-4

图 8 基于 RS-WvGv 的 LR 混淆矩阵图

数据集 H-1 是平衡数据集，混淆矩阵图 H-1 显示，四种方法对平衡数据集的表现差别仅在于准确率高低不同。对于非平衡数据集 H-4 来说，本文采用精确率指标进行分析，从混淆矩阵图中可以利用式(6)、式(7)计算得到每个类别的召回率和精确率。混淆矩阵图 H-4 显示，本文所提方法 RS-WvGv 与另外两种情感词特征提取方法对于大样本量的类别 1 的召回率差别不大(小于 2%)，精确率差别较大(将近 5%)；而对于小样本量的类别 0，召回率和精确率都差别较大，分别差将近 13% 和 7%，结果如表 11 所示。以上结果说明在对情感词特征进行提取的过程中存在特征边界模糊的问题，所提方法可以很大程度上缓解这种问题。

表 11 对比实验在酒店评论数据集上的精确率和召回率

情感词特征表示方法	精确率 P		召回率 R	
	0	1	0	1
ALL-WvGv	0.767	0.875	0.699	0.909
TextRank-WvGv	0.722	0.831	0.570	0.906
TF-IDF-WvGv	0.722	0.838	0.594	0.902
RS-WvGv	**0.792**	**0.885**	**0.724**	**0.918**

3.6 统计分析

为了对比不同特征表示方法在不同数据集上的表现差异，本文把表 10 转换成按准确率从高到低排序的排序表，如果两种方法的差别小于 0.001，两种方法将平均排序。最后获得不同方法在不同数据集上的排序情况，结果如表 12 所示。

表 12 不同方法在不同数据集上的准确率排序表

情感词特征表示方法	H-1	H-2	H-3	H-4	平均序值 r_i
ALL-WvGv	2	2	2	2	2
TextRank-WvGv	4	3	4	4	3.75
TF-IDF-WvGv	3	4	3	3	3.25
RS-WvGv	**1**	**1**	**1**	**1**	**1**

在获得不同方法的准确率排序表之后，使用 Friedman 检验去判断不同的方法是否具有相同的表现，同时做出假设“所有的方法表现相同”。变量 τ_F 服从自由度为 $(k-1)$ 和 $(k-1)(N-1)$ 的 F 分布，计算方法如式(8)、式(9)所示。

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}} \tag{8}$$

$$\tau_{\chi^2} = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4}\right) \tag{9}$$

通过上述公式计算得到的变量值 τ_F 与临界值 $F_{\alpha=0.05}$[①] 进行比较，假设“所有的方法表现相同”被拒绝。为了进一步区分各种方法不显著差异，使用 Nemenyi 检验作为后续检验，使用式(10)计算出临界值域 CD[②] 后，画出 Friedman 检验结果图，如图 9 所示。其中，中心圆点表示每个方法的平均序值，以圆点为中心的横线段表示临界值域的大小。Friedman 检验结果表示，如果两种方法的横线段有较多重叠，则表明两种方法的差异性较小，反之则说明两种方法差异性显著。

$$\mathrm{CD} = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{10}$$

从图 9 可以看到，直线 4 与直线 3 有较多重合部分，而与直线 2 和直线 1 重叠部分较少或者无重叠部分。也就是说本文所提方法 RS-WvGv 与未经过情感词特征提取的方法 ALL-WvGv 差别较小，但是显著优于其他两种情感特征表示方法 TF-IDF-WvGv 和 TextRank-WvGv，这也验证了实验 2 的分析结果。

① $F_{\alpha=0.05}=3.863(k=4,N=4)$

② $q_{\alpha=0.05}=2.569(k=4)$

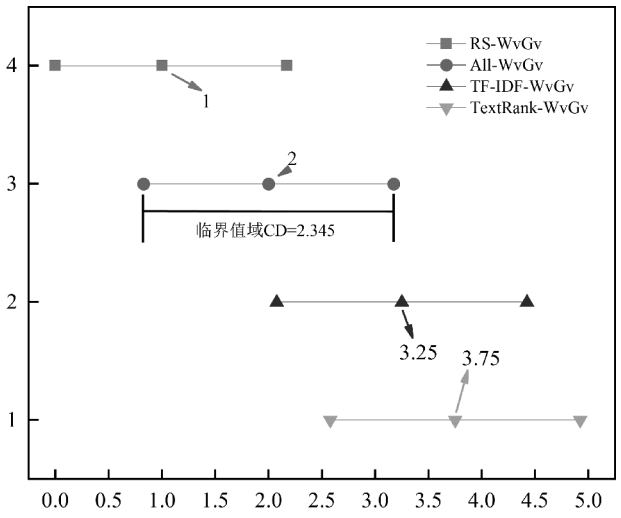


图 9　Friedman 检验结果图

4　结论与展望

本文在基于监督学习的情感分类过程中，提出了一种基于粗糙集和多通道词向量的文本情感词特征表示方法 RS-WvGv，实验结果表明，该方法在不损失分类准确度的基础上，大大降低了情感词特征的空间维数，减少了情感词特征在向量表示过程中的时间花费。

在情感分析任务中，选择合适的情感词特征是任务成败的关键，现有的情感词特征提取方法存在降维效率有限或者选择结果特征覆盖力低等问题，本文将粗糙集属性约简方法应用于特征提取任务中，提高了情感词特征子集的覆盖程度，实现情感词特征子集的精准选择，降维效果理想。下一阶段的工作中，笔者将针对中文文本情感分析任务中的一词多义现象，结合当前词位置、词性等外部文本特征信息对中文文本进行表示。另外，结合 Elmo，BERT，XLNet 等预训练语言模型，对文本情感词特征进行动态表示也是接下来的重点工作。

参考文献

[1] 赵妍妍，秦兵，刘挺. 文本情感分析[J]. 软件学报，2010，21(8)：1834-1848.

[2] 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. PA，USA：2002：79-86.

[3] 李平，戴月明，王艳. 基于混合卡方统计量与逻辑回归的文本情感分析[J]. 计算机工程，2017，43(12)：192-196.

[4] Yang S，Xia Z. A convolutional neural network method for Chinese document sentiment analyzing [C]//Proceedings of IEEE International Conference on Computer and Communications. IEEE，2016：308-312.

[5] 赵富，杨洋，蒋瑞，等. 融合词性的双注意力 BiLSTM 情感分析[J]. 计算机应用，2018，38(S2)：108-111，152.

[6] Mihalcea R，Tarau P. TextRank：Bringing order into texts[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing，2004：404-411.

[7] 张克君，史泰猛，李伟男. 基于统计语言模型改进的 Word2Vec 优化策略研究[J]. 中文信息学报，2019，33(7)：11-19.

[8] 徐立. 基于加权 TextRank 的文本关键词提取方法[J]. 计算机科学，2019，46(z1)：142-145.

[9] Pawlak Z. Rough sets[J]. International Journal of Computer & Information Sciences，1982，11(5)：341-356.

[10] 张志飞，苗夺谦. 基于粗糙集的文本分类特征选择算法[J]. 智能系统学报，2009，4(5)：453-457.

[11] 孙晓，高飞，任福继. 基于深度模型的社会新闻对用户情感影响挖掘[J]. 中文信息学报，2017，31(3)：184-190.

[12] Hinton G E. Learning distributed representations of concepts [C]//Proceedings of the 8th Conference of the Cognitive Science Society，1989.

[13] Mikolov T，Chen K，Corrado G，et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv：1301.3781，2013.

[14] 唐明，朱磊，邹显春. 基于 Word2Vec 的一种文档向量表示[J]. 计算机科学，2016，43(6)：214-217.

[15] Bojanowski P，Grave E，Joulin A，et al. Enriching word vectors with subword information[J]. arXiv preprint arXiv：1607.04606，2016.

[16] Joulin A，Grave E，Bojanowski P，et al. Bag of tricks for efficient text classification[J]. arXiv preprint arXiv：1607.01759，2016.

[17] Pennington J，Socher R，Manning C. Glove：Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha，Qatar：Association for Computational Linguistics，2014：1532-1543.

[18] Peters M E，Neumann M，Iyyer M，et al. Deep contextualized word representations[J]. arXiv preprint arXiv：1802.05365，2018.

[19] Devlin J，Chang M W，Lee K，et al. Bert：Pre-training of deep bidirectional transformers for language

understanding[J]. arXiv preprint arXiv: 1810.04805, 2018.

[20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Processings of the Advances in Neural Information Processing Systems. Long Beach, California, USA: 2017: 6000-6010.

[21] Tan S B, Zhang J. An empirical study of sentiment analysis for Chinese documents[J]. Expert Systems with Application, 2008, 34(4): 2622-2629.

[22] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001.

[23] Øhrn A. Rosetta technical reference manual[EB/OL]. [2001-05-25]. http://bioinf.icm.uu.se/rosetta/materials/manual.pdf.

[24] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python[J]. Journal of Machine Learning Research, 2011, 12(Oct): 2825-2830.

陈波(1996—),硕士,主要研究领域为自然语言处理。

E-mail: chenbobobo180@163.com

苗夺谦(1964—),博士,教授,主要研究方向为人工智能、大数据分析、粗糙集理论、粒计算等。

E-mail: dqmiao@tongji.edu.cn

谢珺(1979—),通信作者,博士,硕士生导师,主要研究领域为粗糙集、粒计算、数据挖掘和智能信息处理。

E-mail: xiejun@tyut.edu.cn

全国社会媒体处理大会(SMP 2020)召开,98 场报告日程全公开

全国社会媒体处理大会(SMP)专注于以社会媒体处理为主题的科学研究与工程开发,为传播社会媒体处理最新的学术研究与技术成果提供广泛的交流平台,旨在构建社会媒体处理领域的产学研生态圈,成为中国乃至世界社会媒体处理的风向标。会议将采取大会报告、专题研讨、张贴报告等形式进行交流。全国社会媒体处理大会创办于 2012 年,每年举办一次,现已成为社会媒体处理的重要学术活动。

第九届全国社会媒体处理大会(SMP 2020)由中国中文信息学会社会媒体处理专委会主办,浙江大学承办,将于 2020 年 9 月 4 日至 9 月 6 日线上召开。大会正在开放注册中,了解「报名详情」请参见中国中文信息学会公众号信息。

本次大会邀请到了多位著名专家和业界翘楚进行大会主题报告,另外还邀请到了计算科学、社会科学等多个领域的著名学者进行专题论坛报告。会议包含一系列学术活动,会议议程除传统的知名学者的大会报告、分论坛报告、口头报告、评测活动等之外,还将组织顶级会议论坛、企业论坛以及重大公共安全论坛等。

两场前沿讲习班,3 场特邀报告,两场青年科学家报告,19 场论坛,90 位学者的分享,9 月 4 日至 6 日,将在 Zoom 平台上精彩呈现。