基于SVM的高维混合特征短文本情感分类

王义真,郑 啸,后 盾,胡 昊
(安徽工业大学 计算机科学与技术学院,安徽 马鞍山 243032)

摘 要: 针对短文本具有的稀疏性、不规范性、主题不明确性等相关特点,提出一种基于SVM的高维混合特征模型。首先介绍了兼顾语义和情感的6类特征:表情符号特征、词聚类特征、词性标注特征、n-gram特征、否定特征和情感词典。其中主要介绍了该6类特征的概念、抽取方式以及输出形式;其次在第六届中文倾向性分析评测(COAE2014)为基础的数据集上,采用5折交叉的方法对该模型进行了有效性验证,其平均准确率为84.69%、平均召回率为83.13%,而平均F_1值为83.90%;接着探讨了SVM惩罚系数对实验的影响;最后将该模型与一步三分类方法、Recursive Auto Encoder、Doc2vec做了对比分析,结果表明提出的模型对短文本情感分类更有效。

关键词: 情感分类;混合特征;支持向量机;情感词典
中图分类号: TP393 **文献标识码:** A **文章编号:** 1673-629X(2018)02-0088-06
doi:10.3969/j.issn.1673-629X.2018.02.020

Short Text Sentiment Classification of High Dimensional Hybrid Feature Based on SVM

WANG Yi-zhen,ZHENG Xiao,HOU Dun,HU Hao
(School of Computer Science and Technology,Anhui University of Technology,Ma'anshan 243032,China)

Abstract: Aiming at the characteristics of short texts which are sparse,unnormative and ambiguous in subject,we present a hybrid feature model with high dimension based on SVM.Firstly,we introduce six types of feature about both semantics and emotion,involving expression symbols,word clustering symbols,part-of-speech tagging,n-gram,negation and the sentiment dictionary,which are mainly introduced in their concept,extraction and output form.Then a five-fold crossover method is used to verify the validity of the model according to the data of COAE2014.The average accuracy rate is 84.69%,the average recall rate is 83.13%,and the average F_1 value is 83.90%.Thirdly,we discuss the influence of SVM regularization parameter on experiment.Finally,the proposed model is compared and analyzed with Recursive Auto Encoder,Doc2vec and so on,which show that it is more effective for short text emotion classification.

Key words: sentiment classification;hybrid feature;SVM;emotion dictionary

0 引 言

随着移动互联网的高速发展,智能终端的普及,用户通过移动网络更容易获取和发布互联网信息。社交媒体的兴起,加速了用户自由表达对人或事的态度、观点以及情感倾向。近年来网络上涌现的短文本迅速膨胀,如商品评论、影评、移动短信、微博、论坛等,仅靠人工的方法难以应对网上海量信息的收集和处理。传统的基于关键字的检索,文本的分类,文本的聚类往往忽略了文本中的情感。因此迫切需要计算机帮助用户快速获取和整理这些情感相关信息。文本情感分类主要是通过分析用户发表的主观性文本内容,挖掘其情感倾向,从而判断其情感倾向的极性(如:正向,负向,中立)。针对文本的情感分析有利于更好地了解用户的情感观点,从中发现商业价值,增强用户体验。文本根据长度的不同可以分为长文本和短文本两类。由于短文本具有发布频率快、参与者多、长度较短、结构差异大、交互性强、口语化、省略化、特征关键词稀疏等特性,直接采用现有的情感倾向分类方法对短文本分类的准确率较低。此外,短文本在社区问答[1]、搜索引擎[2]等领域发挥了重要作用,短文本的情感分析日益受到学术界和工业界的广泛关注。

目前,国内外短文本的情感分析主要是针对微博、

收稿日期: 2017-03-04 **修回日期:** 2017-07-11 **网络出版时间:** 2017-11-15
基金项目: 国家自然科学基金(61402008,61402009);安徽省科技重大专项(16030901060);安徽省高校自然科学研究重大项目(KJ2014ZD05);安徽省高校优秀青年人才支持计划
作者简介: 王义真(1992-),男,硕士研究生,研究方向为自然语言处理;郑 啸,教授,博士,研究方向为计算机网络、服务计算与云计算。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20171115.1436.044.html

在线评论等短文本。在国外,短文本的情感分析研究主要分为主题无关的情感分析和主题相关的情感分析。情感分析的研究思路主要分为两种: 一种是基于语义的研究方法,主要利用现有情感词典或建立倾向性语义模式库,应用情感规则匹配的方式实现文本语义的理解,从而实现对文本的情感识别。文献[3]利用词典中情感词和短语的相关极性和强度,并采用集约化和否定化计算文本的情感得分。文献[4]结合词典和规则来计算文本的情感极性。重点是情感评价词语或其组合的极性判断以及极性求和的方法。另一种是基于机器学习的研究方法,将情感分析看做分类问题。Pang 等[5]将机器学习方法应用于电影评论的二分类问题; Kang 等[6]提出应用在酒店评论的朴素贝叶斯的改进算法; Liu 等[7]提出应用在 Tweet 的自适应协同训练算法。

传统的方法或只依赖情感知识(需要建设情感词典或领域性情感词库),或只侧重从大量的训练集中抽取情感特征,而大量的工作表明,这两者之间相互依赖、互为补充。虽然针对文本的情感分析研究已经取得了一定的成果,如果能将两者很好地进行融合,必将对情感分类的效果有很大的提升。基于此,文中提出基于 SVM 的高维混合特征模型。在短文本的特征提取上,兼顾了情感和语义两者,充分挖掘短文本的情感特征,并且引入了新的特征。

1　相关工作

情感分析[8](sentiment analysis),又称倾向性分析,意见抽取(opinion extraction),意见挖掘(opinion mining),情感挖掘(sentiment mining),主观分析(subjectivity analysis),它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。情感分析自从 2002 年由 Bo Pang 提出后,获得了很大程度的关注,特别是在在线评论的情感倾向性分析上获得了很大的发展,具有很大的研究和应用价值。由于短文本的特殊性,直到近些年,人们才开始关注微博等短文本情感分析任务。

一般而言,当前的短文本情感分析任务主要关注于特征提取和分类器选择两个部分。由于短文本特征非常稀疏,Flekova 等[9]通过计算同义词词典词汇语义相似度拓充 twitter 情感特征,并结合词典、n-gram 等特征训练支持向量机分类器。Kokciyan 等[10]则加入了主题标签、上下文、指示等特征构建 twitter 情感分析系统。由于中文短文本的复杂性,不少研究人员利用现有的通用词典 WordNet 或 HowNet,进行扩展来获取大量的极性词语及极性。杨超等[11]在 HowNet 和 NTUSD 的基础上进行扩展,建立了一个具有倾向程度的情感词典。基于情感词典和修饰词词典,计算句子的倾向性,最后得到一条评论的倾向性。何凤英等[12]以 HowNet 情感词语集为基准,构建中文基础情感词典,利用词典及程度副词和否定副词词典计算情感词的极性,利用词典及程度副词和否定副词词典来获取博文的情感倾向性。研究发现,综合考虑三种因素,采用支持向量机(SVM)和信息增益(IG),以及 TF-IDF(term frequency-inverse document frequency)作为特征项权重,三者结合对微博的情感分类效果最好。谢丽星等[13]针对中文微博消息展开了情感分析方面的初步调研,实验对比了三种情感分析的方法,包括表情符号的规则方法、情感词典的规则方法、基于 SVM 的层次结构多策略方法,结果证明基于 SVM 的层次结构多策略方法效果最好。

2　情感特征的构造

2.1　表情符号特征

文中通过对微博、在线评论等主流网站采集一定规模的数据后,发现短文本语料中包含丰富的表情符号。有些表情含有明显的情感倾向,利用正则表达式能够提取文本的表情符号。选择了表 1 具有代表性的带情感倾向的表情符号。

表 1　表情符号列表

正向倾向表情	负向倾向表情
[哈哈]、[嘻嘻]、[呵呵]、[鼓掌]、[偷乐]、[可爱]、[爱你]、[佩服]、[good]、^_^、^o^	[失望]、[衰]、[伤心]、[悲伤]、[鄙视]、[恶心]、[泪]、[无语]、[怒]、T^T、T_T、::>_<::

选择的依据: 一是出现频次越高,选取的机会越大; 二是根据经验知识判定表情符号情感倾向。最终在抽取特征后形成:

$$\begin{bmatrix} & [X_1 & X_2 & \cdots & X_a] \\ \begin{bmatrix} \mathrm{doc}_1 \\ \mathrm{doc}_2 \\ \vdots \\ \mathrm{doc}_m \end{bmatrix} & \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1a} \\ x_{21} & x_{22} & \cdots & x_{2a} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{ma} \end{bmatrix} \end{bmatrix}$$

其中,$x_{ij} = \begin{cases} 1, \text{表情存在} \\ 0, \text{其他} \end{cases}$; $X_{1,2,\cdots,a}$ 表示表情符号; $\mathrm{doc}_{1,2,\cdots,m}$ 表示 m 个文本。实验中最终用稀疏矩阵表示。

另外,网民在发表这些评论信息结束时会使用一个或者多个表情用于更好地表达自己的情感,而这些表情看上去是图片,实际上是由特殊符号组成。

例如,这部电影真心不错 [good]。由此可见,最后一个表情能够表达网民发表的短文文本情感倾向。在该特征的提取方面,发现正则表达式能够很好地处

理这种有特殊表情符号组成的表情。

2.2 词聚类特征

词向量具有良好的语义特性，是表示词语特征的常用方式。考虑到很多字、短语表现形式不同，但是表达的意义相同，采用词聚类的方法能够有效地避免处理相近的词语多余的计算量。word2vec[14] 是 Mikolov 等提出模型的一个实现，可以用来快速有效地训练词向量。可以把对文本内容的处理简化为向量空间中的向量运算，计算出向量空间上的相似度来表示文本语义上的相似度。word2vec 包含了 cbow 和 skip-gram 两种训练模型。文中在 skip-gram 模型完成了词聚类，构建了词聚类词典。在这个模型中，给出语料库的词 w 及其上下文 c，在给定条件概率 $p(c \mid w)$ 和语料库 Text 下，去求参数 θ 使属于语料库的概率最大化：

$$\arg\max_{\theta} \prod_{w \in \text{Text}} \left[\prod_{c \in C(w)} p(c \mid w; \theta) \right]$$

其中，$C(w)$ 表示一组单词 w 的上下文。文中使用该工具运用在收集到的语料库上聚 100 类后得到 1 533个基元。

2.3 词性标注特征

常见的分词系统的词性标注的粒度能达到：名词、动词、形容词、副词等。文中选用中科院的 ICTCLAS[15] 作为分词系统，它能将词性标注粒度更为细分。例如，名词可以分成人名、地名；形容词可以分为副形词、名形词、形容词性语素、形词词性惯用语。

例如，语句是 “又一部国产良心之作 笑点从头到尾 搞笑却不乏温情 真是让人又哭又笑，同一个道理听过太多次总觉得平淡无味没有分量，然而这一次却说到心里。”标注后：又/d 一/m 部/q 国产/b 良心/n 之/uzhi 作/ng 笑/vd 点/v 从头到尾/dl 搞/v 笑/v 却/v 不乏/v 温情/n 真/d 是/vshi 让/v 人/n 又/d 哭/v 又/d 笑/v，/wj 同/p 一个/mq 道理/n 听/v 过/uguo 太/d 多/m 次/qv 总/d 觉得/v 平淡/a 无味/a 没/d 有/vyou 分量/n，/wj 然而/c 这/rzv 一/m 次/qv 却/d 说/v 到/v 心里/s 。/wj

2.4 n-gram 特征

对于给定的文本，都可以将其看做是长度不同序列的集合。在这些序列中，相邻的 N 个字或词称为 n-gram，n-gram 算法的基本思想是通过一个大小为 N 的滑动窗口将文本内容进行切分，形成长度为 N 的片段序列，每个片段序列称为 gram。使用 n-gram 特征，尽可能地获取有限长度短文本的未登录情感词和情感信息。

例如 “乒乓球拍卖啦”，采用传统的分词技术，会被切分成 “乒乓球/拍卖/啦”或 “乒乓/球拍/卖啦”。可见传统分词技术对于短文本的分词存在明显的缺陷，甚至可能会改变原有评价对象。文中将 n-gram 作为一类特征用于短文本的情感分析。鉴于此类情况增加 n-gram 特征：对于 1-gram 是单个的字或词对于特征的选择并没有多大意义，所以选择从 2-gram 开始，但超过 4-gram 同样没什么意义。

2.5 否定特征

含有主观倾向的语句往往有很明显的否定词。与传统文本情感分类不同，“不”、”没“等否定词不再作为停顿词被删除。在句子里 “不”或 “没”的否定范围是 “不”或 “没”的全部词。一个词在不在否定范围内对正确情感分类产生了很大影响。

例如 “他一直没上班/他没一直上班；你没天天学习/你天天没学习。”文中采用否定特征是以句子出现否定词为否定特征的开始直至句子结束都加上否定标记，并且记录否定词的个数也作为否定特征的一部分。

2.6 情感词典

在对文本情感分类时，往往文本中含有的少数带有情感倾向的词汇最直接表现文本情感的倾向。如正向词汇 “高兴”和负向情感词 “难过”。由于中文词语的复杂性，情感词汇非常丰富，多为形容词、副词等。文中选择四个情感词典进行情感特征选择。其中包含整理好的 HowNet、NTUSD、大连理工大学的本体词汇以及使用 CHI 统计对情感短文语料库构建的 AHUT 词典。其中由于前两者并没有标注情感词的情感极性，所以将正向词汇的得分定为 1.0，负向词汇的得分定为-1.0。在情感词典特征上，采用下面四个规则进行情感分数的计算。

规则 1：分别计算情感文本中的正向词、负向词的数量；

规则 2：分别计算情感文本中的正向词、负向词的得分总数；

规则 3：分别计算情感文本中的得分最大正向词、负向词的分值；

规则 4：分别计算情感文本中的最后一个情感词的分值。

3 SVM 高维混合特征情感分类器

3.1 理论基础

情感短文本经过特征抽取后得到的是高维稀疏向量矩阵，直接用来作为分类器的训练和测试数据，选用适合处理大规模文本分类的 SVM 算法构建情感分类器。给定一组样本集 $\{x_i, y_i\}$，$i = 1, 2, \cdots, l$，$x_i \in R^n$，$y_i \in \{-1, +1\}$，SVM 需要解决如下无约束最优化问题：

$$\min_{w} \frac{1}{2} w^{\mathrm{T}} w + C \sum_{i=1}^{l} \xi(w; x_i, y_i) \tag{1}$$

其中，$\xi(w; x_i, y_i)$ 为损失函数；C 为惩罚系数；l 为样本总数。

通常在分类问题中使用标准 C-SVM(L1-SVM)作为有效的分类算法。L1-SVM 的损失函数是一阶范数，而二阶 L2-SVM 的损失函数增加了一个由惩罚因子对角矩阵逆的 Hessian 矩阵的双重方法。这提高了求解过程的稳定性。L1-SVM 和 L2-SVM 的损失函数公式分别如下:

$$\max(1 - y_i w^{\mathrm{T}} x_i, 0) \tag{2}$$

$$\max(1 - y_i w^{\mathrm{T}} x_i, 0)^2 \tag{3}$$

通常在 SVM 的分类问题中增加一个偏置项 b，文中处理偏置项 b 如下所示:

$$\begin{aligned} x^{\mathrm{T}} &\leftarrow [x_i^{\mathrm{T}}, B] \\ w^{\mathrm{T}} &\leftarrow [w^{\mathrm{T}}, b] \end{aligned} \tag{4}$$

其中，B 为常数。

式(1)称作 SVM 的原始形式，在求解中将其转变成对偶形式:

$$\begin{cases} \min\limits_{\alpha} f(\alpha) = \dfrac{1}{2} \boldsymbol{\alpha}^{\mathrm{T}} \bar{\boldsymbol{Q}} \boldsymbol{\alpha} - \boldsymbol{e}^{\mathrm{T}} \boldsymbol{\alpha} \\ \text{s.t. } 0 \leqslant \alpha_i \leqslant U, \forall i \end{cases} \tag{5}$$

其中，$\boldsymbol{e}$ 为全 1 矩阵；$\bar{\boldsymbol{Q}} = \boldsymbol{Q} + \boldsymbol{D}$，$\boldsymbol{D}$ 为对角矩阵，$Q_{ij} = y_i y_j x_i^{\mathrm{T}} x_j$。

在 L1-SVM 中，$U = C$，$D_{ij} = 0$；在 L2-SVM 中，$U = \infty$，$D_{ij} = 1/2C$，$\forall i$。对于式(5)中对偶问题的求解，文献[16]提供了开源的大规模线性 SVM 的工具包 LIBLINEAR，实现了 L1-SVM、L2-SVM 等损失函数。

通过实验对比表明: 在处理大规模数据时，L2-SVM 的性能优于 L1-SVM、PEGASOS、SVMperf。因此，文中同样选用 L2-SVM 作为 SVM 情感分类器的损失函数。

3.2　框架实现

情感文本特征的表示是情感分类的关键步骤，包括预处理、中文分词、特征抽取三个部分。

预处理: 目的是将原始文本中涉及到用户隐私的内容删除。其中可能会包含超链接、用户名以及一些特定话题。

中文分词: 文中使用的是 ICTCLAS 分词工具，为下一步的特征抽取提供较为准确的基元。

特征抽取: 第 2 节已经列举了实验所要用到的各种情感特征。

实验思路: 先从目标网站爬取评论、微博等数据进行标注; 然后使用 k 折交叉的方法进行训练和测试; 最后经过情感分类器输出情感极性(正向、负向、中立)，并统计实验结果。

4　实验方法及相关分析

4.1　实验数据及预处理

中文情感文本分析不同于英文，到目前为止情感评测语料库尚未完善。实验采用的语料库是由 COAE2014 评测提供的语料集和从新浪、京东等国内知名网站上采集的数据组成。文中将语料库命名为 DataSet，其中正向条数为 5 200，中立条数为 5 600，负向条数为 5 430。考虑到短文本内容可能含有用户的一些隐私信息，所以要对实验数据进行预处理。文中删除了语料库中 url 链接、用户名、话题等信息。

4.2　评价指标

使用准确率 P(precision)、召回率 R(recall)和 F_1 值(F-Score)作为评价分类器的性能指标，其具体计算公式如下:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$F_1 = \frac{2PR}{P + R} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{8}$$

其中，TP 表示分类器将输入文本正确地分类到某个类别的数量; FN 表示分类器将输入文本错误地分类到某个类别的数量; FP 表示分类器将输入文本错误地排除在某个类别之外的数量。

4.3　实验结果与分析

文本语料库经过特征筛选器处理后得到的稀疏向量矩阵，可直接作为情感分类器训练、测试以及交叉验证的数据集。

(1) 基于 5 折交叉验证的实验结果。

首先对短文本语料库进行特征抽取(约有 267 万)，然后对语料库进行 5 折交叉验证的实验。选用 Naive Bayes 作为对比的 baseline 方法，在全部特征上做 5 折交叉的实验(见图 1)，并且模型参数为默认值。

从图 1 可以看出，文中模型的分类效果明显高于 Naive Bayes，其平均准确率为 84.69%，平均召回率为 83.13%，而平均 F_1 值为 83.90%。

(2) 不同惩罚系数的实验比较。

在 5 折交叉实验的基础上，验证不同惩罚系数 C 对模型三个评价指标的影响，实验结果如图 2 所示。

对比图 2 可以发现，短文本情感分类的各评价指标的变化趋势一致，惩罚系数在 75 左右时，实验效果达到最好。

(3) 不同特征组合的实验对比。

为验证不同类特征对实验结果的影响，选用不含有 AHUT 情感词典和词性标注作为 base feature，然后依次在上一次特征的基础上加入 AHUT 字典、表情符

号、词聚类、n-gram 以及否定特征(即所有特征)来对部分语料进行实验。实验的平均准确率结果统计如图 3 所示。

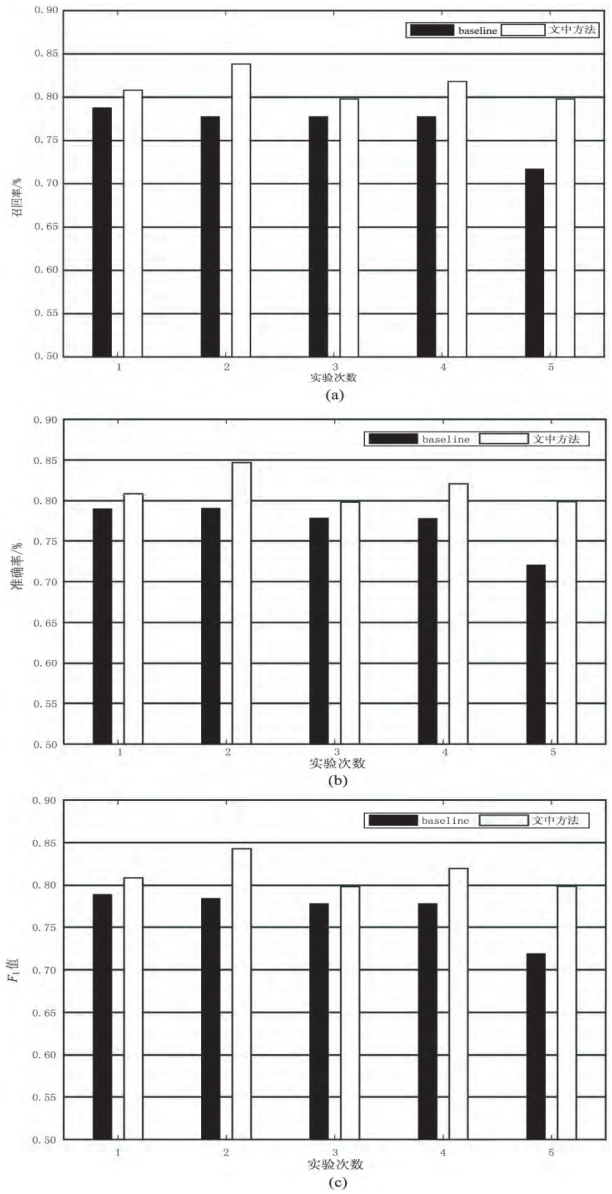

图 1 5 折交叉的实验结果

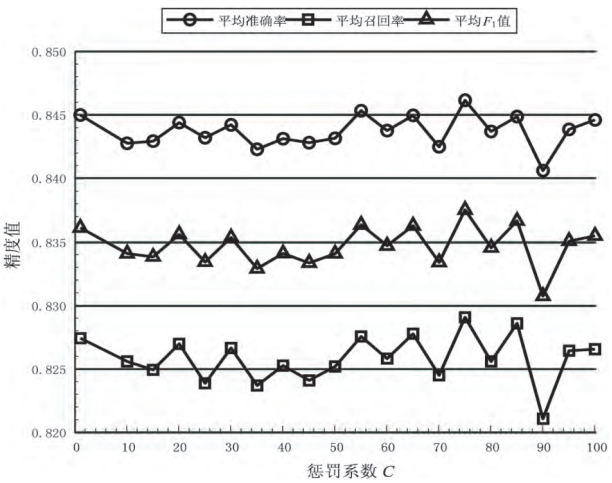

图 2 不同惩罚系数 C 在 5 折交叉验证中的分类性能

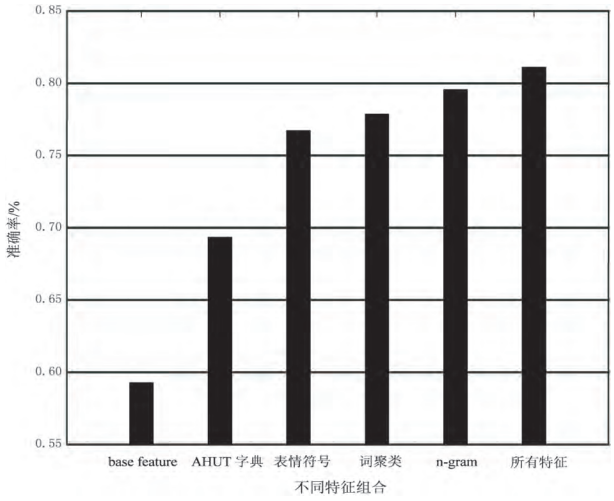

图 3 不同特征组合的实验结果对比

从图 3 可以看出,在 base feature 的基础上加入文中构建的 AHUT 字典后,分类效果提升比较明显,在加入全部特征后效果达到最好。这是由于文中方法针对短文本抽取的特征有效。但由于实验特征的组合上采用的是依次增加的方式,而不是随机选用其中几类特征的组合,故存在不足。

(4) 多种模型的对比。

最后,为进一步验证提出模型的有效性,在使用同样语料库的基础上与一步三分类方法[13]、Recursive AutoEncoder[17]、Doc2vec 方法进行对比,结果如表 2 所示。

表 2 多种模型准确率对比 %

方法	准确率/%
高维混合特征分类方法	84.69
一步三分类方法	79.33
Recursive AutoEncoder	78.83
Doc2vec	76.76

实验结果表明,提出模型的准确率优于其他几种模型,验证了模型的正确性。这是因为与一步三分类方法对比,文中的情感特征增加了词聚类、否定特征等特征,明显提高了准确率;与 Recursive AutoEncoder、Doc2vec 相比,后两者在准确率多分类上低于二分类。而且,文中在特征选取方面采取正则化手段,避免了特征的二次选择和“高维”灾难。

5 结束语

文中充分考虑短文本的特点,从多维混合特征的角度进行文本的特征抽取,做到尽可能兼顾语义和情感,并且取得了较好的实验效果,验证了该方法的有效性和鲁棒性。

文中提出了基于 SVM 的高维混合特征框架,采用正则化的手段解决维数灾难问题;弥补了传统情感

字典未标注情感强度值的不足,构建了带有情感强度值的 AHUT 情感词典;考虑到语义对短文本情感分类的正确率影响,将词聚类加入到情感分析的特征,提高了 1.4%的准确率。虽然取得了一定的成果,但也存在不足之处:对情感词典有一定的依赖;在针对不同特征的组合上,并没有随机选取几种特征的组合进行实验,可能给实验的最终结果带来偏差;无法对海量数据进行实时、并行化处理。接下来的工作将着手解决上述存在的不足之处。

参考文献:

[1] WU H,WU W,ZHOU M,et al.Improving search relevance for short queries in community question answering [C] //Proceedings of the 7th ACM international conference on web search and data mining.New York,NY,USA: ACM,2014: 43-52.

[2] TEEVAN J,RAMAGE D,MORRIS M R.#TwitterSearch: a comparison of microblog search and web search [C] //Proceedings of the fourth ACM international conference on web search and data mining.New York,NY,USA: ACM,2011: 35-44.

[3] TABOADA M,BROOKE J,TOFILOSKI M,et al.Lexicon-based methods for sentiment analysis [J] .Computational Linguistics,2011,37(2): 267-307.

[4] YUAN D,ZHOU Y,LI R,et al.Sentiment analysis of microblog combining dictionary and rules [C] //IEEE/ACM international conference on advances in social networks analysis and mining. [s.l.]: IEEE,2014: 785-789.

[5] 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-Volume 10. [s.l.]: Association for Computational Linguistics,2002: 79-86.

[6] KANG H,YOO S J,HAN D.Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews [J] .Expert Systems with Applications,2012,39(5): 6000-6010.

[7] LIU S,LI F,LI F,et al.Adaptive co-training SVM for sentiment classification on tweets [C] //Proceedings of the 22nd ACM international conference on information & knowledge management.New York,NY,USA: ACM,2013: 2079-2088.

[8] 赵妍妍,秦　兵,刘　挺.文本情感分析 [J] .软件学报,2010,21(8): 1834-1848.

[9] FLEKOVA L, FERSCHK O, GUREVYCH I.UKPDIPF: a lexical semantic approach to sentiment polarity prediction in twitter data [C] //Proceedings of the 8th international workshop on semantic evaluation.Dublin,Ireland: [s.n.],2014: 704-710.

[10] KOKCIYAN N,ARDA C,OZGUR A,et al.BOUNCE: sentiment classification in twitter using rich feature sets [C] //Proceedings of the 7th international workshop on semantic evaluation.Atlanta,Georgia: ACL,2013: 554-561.

[11] 杨　超,冯　时,王大玲,等.基于情感词典扩展技术的网络舆情倾向性分析 [J] .小型微型计算机系统,2010,31(4): 691-695.

[12] 何凤英.基于语义理解的中文博文倾向性分析 [J] .计算机应用,2011,31(8): 2130-2133.

[13] 谢丽星,周　明,孙茂松.基于层次结构的多策略中文微博情感分析和特征抽取 [J] .中文信息学报,2012,26(1): 73-83.

[14] MIKOLOV T,SUTSKEVER I,CHEN K,et al.Distributed representations of words and phrases and their compositionality [C] //Advances in neural information processing systems. [s.l.]: [s.n.],2013: 3111-3119.

[15] ZHANG H P,YU H K,XIONG D Y,et al.HHMM-based Chinese lexical analyzer ICTCLAS [C] //Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. [s.l.]: Association for Computational Linguistics,2003.

[16] FAN R E,CHANG K W,HSIEH C J,et al.LIBLINEAR: a library for large linear classification [J] .Journal of Machine Learning Research,2008,9: 1871-1874.

[17] 梁　军,柴玉梅,原慧斌,等.基于深度学习的微博情感分析 [J] .中文信息学报,2014,28(5): 155-161.