基于多文本特征融合的中文微博的立场检测

奠雨洁,金 琴,吴慧敏
DIAN Yujie, JIN Qin, WU Huimin

中国人民大学 信息学院,北京 100872
School of Information, Renmin University of China, Beijing 100872, China

DIAN Yujie, JIN Qin, WU Huimin. Stance detection in Chinese microblogs via fusing multiple text features. Computer Engineering and Applications, 2017, 53(21):77-84.

Abstract: Stance detection aims to automatically determine whether the author of a text is in favor of the given target, against the given target, or neither. This paper presents a stance detection system based on multiple text feature representations. Firstly, five different feature representations are explored including statistic-based features(BoW, synonym-based BoW, sVariance) and deep text features(word vectors and character vectors). Support Vector Machine(SVM), Random Forest and Gradient Boosting Decision Tree(GBDT) are applied as classifiers. Finally, late fusion is conducted to combine different feature representations. Experiment results show that the proposed feature representations can achieve significant improvement over traditional BoW feature. Moreover, statistic-based features and deep features provide complementary information for stance detection, which leads to the wining system in the Chinese Microblog Stance Detection Evaluation by Natural Language Processing and Chinese Computing(NLPCC 2016).

Key words: stance detection; sentiment analysis; text feature representations; Chinese Microblogs; text classification

摘 要: 微博立场检测是判断微博作者对某一个话题的态度是支持、反对或中立。在基于监督学习的分类框架上,扩展并提出基于多文本特征融合的中文微博的立场检测方法。首先探究了基于词频统计的特征(词袋特征(Bag-of-Words,BoW)、基于同义词典的词袋特征、考虑词与立场标签共现关系的特征)和文本深度特征(词向量、字向量)。之后使用支持向量机,随机森林和梯度提升决策树对上述特征进行立场分类。最后,结合所有特征分类器进行后期融合。实验表明,文中提出的特征对于不同话题下的微博立场检测的结果都有提升,且文本深度特征和基于词频统计的特征能够捕捉到文本的不同信息,在立场检测中是互补的。基于本文方法的微博立场检测系统在2016年自然语言处理与中文计算会议(NLPCC2016)的中文微博立场检测评测任务中取得了最好的结果。

关键词: 立场检测;情感分析;文本特征表示;微博;文本分类
文献标志码: A **中图分类号:** TP391.1 doi:10.3778/j.issn.1002-8331.1702-0292

1 引言

随着互联网的飞速发展和多样化网络交流工具的广泛使用,越来越多的互联网用户在博客、论坛等平台上围绕社会事件、消费产品等话题发表自己的观点、态度,表达自己的立场、情绪。这些评论,对商业智能、舆情分析、政府决策等都具有重要的研究价值。文本情感分析就是研究如何用计算机来分析这些评论信息[1-3]。“微博”是当今互联网最流行的社交媒体之一,其用户基数大、传播速度快等特点使得微博成为社会热点和舆论传播的重要平台。针对微博数据的情感分析,近年来也引起了广泛的关注[4]。由于微博文本句子简短,因此对微博的情感分析主要集中在句子层面的情感倾向性分析(即判断文本的情感是积极还是消极)[5-7]。

微博用户对热点事件的立场(或态度)通常能够反映热点事件的舆情走向,因此,对微博用户的立场分析,有广泛的应用前景。微博用户的立场检测,通常是判断

基金项目: 国家重点研发计划项目(No.2016YFB1001202)。
作者简介: 奠雨洁(1992—),女,硕士,研究领域为多媒体信息处理,E-mail:dianyujie-blair@ruc.edu.cn;金琴(1972—),女,博士,副教授,研究领域为多媒体智能分析;吴慧敏(1990—),女,硕士,研究领域为自然语言处理。
收稿日期: 2017-02-27 **修回日期:** 2017-04-21 **文章编号:** 1002-8331(2017)21-0077-08
CNKI **网络优先出版:** 2017-07-19, http://kns.cnki.net/kcms/detail/11.2127.TP.20170719.1050.022.html

微博作者对于某个话题的态度是支持、反对或中立。它与传统的文本情感倾向性分析很接近,但不同之处在于,立场检测是分析文本针对某一特定话题的情感倾向,而传统的情感倾向性分析不考虑任何话题。例如:“最反感这些拉客的!还有在机动车道上行驶的!”,这条微博不考虑任何目标话题时,它的情感极性是消极,但是考虑“深圳禁摩限电”这个话题后,这条微博就是“支持”这个话题的。

本文提出的方法基于有监督的机器学习算法,并利用多种文本特征的融合来实现中文微博的立场检测。通过分析微博文本上不同话题的特点,首先选取了五种不同的文本特征表示,包括:词袋特征(BoW)、基于同义词典的词袋特征(S-BoW)、考虑话题主题词和立场标签共现关系的特征(sVariance)以及从word2vec中提取的词和字向量。其中S-BoW使用同义词典对一元文法进行扩展,有效地扩充了词汇表,能够更好地处理词汇表外词语;sVariance结合不同话题下的主题词和立场类别标签的共现情况,能够更有针对性地处理立场检测问题。之后本文采用支持向量机(Support Vector Machine,SVM),随机森林(Random Forest)和梯度提升决策树(Gradient Boosting Decision Tree,GBDT)作为分类器进行立场分类。最后,将不同特征的分类器进行后期融合得到最终微博用户的立场。实验结果表明,本文提到的方法能够有效检测中文微博中的作者立场。系统在NLPCC2016(The Fifth Conference on Natural Language Processing and Chinese Computing)的中文微博立场检测的评测任务中取得了最优的比赛结果[8]。

本文的组织结构如下:

第2章介绍立场检测相关工作。第3章介绍本文的立场检测方法,包括特征提取、立场分类和特征后期融合。第4章介绍本文的方法在中文微博数据集上的验证结果并对结果进行分析。第5章对本文研究工作做出总结以及对未来的研究做出展望。

2 相关工作

立场检测是一项特殊的情感分析任务,因此常见的情感分析的方法可以用于立场检测,如基于特征分类的方法,利用机器学习模型,通过学习大量有意义的特征来完成分类任务。文献[9]首次将机器学习的方法应用于篇章级(英文)的情感分类任务中。他们使用了n-gram词语特征和词性特征,并对比了朴素贝叶斯(Naïve Bayes)、最大熵和SVM这三种分类模型,发现Unigram特征效果最好。在中文情感分类研究中,文献[10]使用了三种机器学习算法(SVM、贝叶斯和n元文法)、三种特征选取算法(信息增益、CHI统计和文档频率)以及三种特征项权重计算方法(二值、词频和TF-IDF)对微博进行了情感分类的实证研究。研究发现,采用SVM作为分类器、信息增益作为特征选取算法,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。

基于话题的情感分析与立场检测任务较为相似,二者都是对特定的话题进行情感倾向分析。文献[11]中提出一种基于LDA话题模型与How-net词典的中文博客多方面话题情感分析方法,该方法首先利用数据语料训练LDA话题模型对博客文本进行话题识别与划分;在此基础上,基于How-net词典对划分后的话题段落进行情感倾向计算。文献[12]在话题情感模型中,将文档中情感词看作一个马尔可夫链,考虑局部上下文中情感词之间的相互依赖关系,利用该依赖关系进行情感分析。这类方法在基于篇章级的文本情感分析中取得了较好的成果,但是这类方法有如下问题:第一,模型较为复杂,求解比较困难。第二,多数方法在篇章级文本上进行验证。与篇章级文本相比,微博文本较为简短,上下文信息稀缺,且句子结构和词语表达不规范。因此上述方法很难直接用于微博文本的立场检测。

立场检测的研究工作最早开始于政治辩论,国外对于立场检测的研究工作也主要集中在政治辩论、在线辩论等平台。研究方法主要利用这些平台用户之间的关系和对话等特征进行立场检测[13-14]。在SemEval2016的评测任务(http://alt.qcri.org/semeval2016/task6/)中,有大量的研究队伍参与了英文微博(Twitter)的立场检测。例如:文献[15]将立场检测问题归结为多分类问题,抽取了二元、三元文法、语义特征(否定词的数量,感叹号的数量等)以及立场词典特征,使用层次分类器的框架检测微博用户的立场。实验表明在立场检测任务中,不同的文本特征的表现跟具体的话题类别相关。文献[16]使用了一元文法和词向量特征,利用随机森林、SVM和GBDT三种分类器进行特征分类,并在后期融合了不同特征。实验结果表明,不同特征的融合对实验结果的提升较为显著。

本文的研究任务与文献[8]中定义的立场检测任务一致,本文提出的方法借鉴并扩展了[16]在Twitter数据集上的工作,将其应用到中文微博数据的立场检测。中文微博和英文微博的共同点在于文本较简短,用词不规范,不同之处在于中文的处理更复杂,例如:中文微博中可能包含英文等外语文字;中文的语法结构更复杂等。 因此,除了使用一元文法和词向量特征外,本文根据中文微博上不同话题的特点,探索了其他语义特征。第一,本文使用同义词典对一元文法进行扩展,有效地扩充了词汇表,能够更好地处理词汇表外词语。第二,本文提出了一种新的权重计算方法,该方法结合不同话题下的主题词和立场类别标签的共现情况,更有针对性地处理立场检测问题。第三,由于微博文本比较简短,每个字在微博中相较于词语可能更重要,因此本文还抽

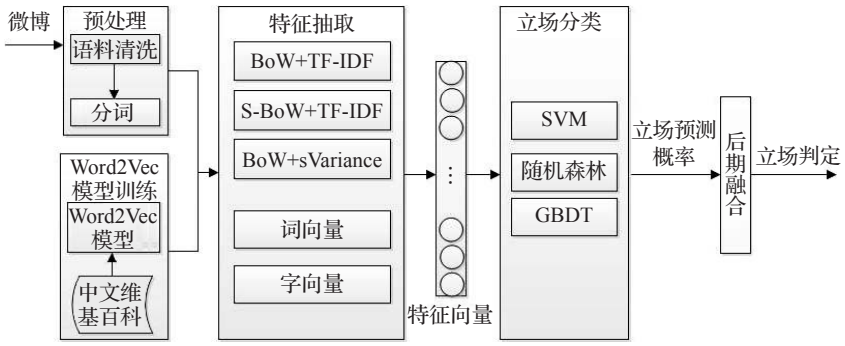

图1　系统框架

取了字向量特征。基于上述特征,使用随机森林、SVM和GBDT进行立场分类并使用特征分类器的后期融合来预测最终立场。

3　系统说明

如图1所示,本文提出的基于监督学习的立场检测方法,共分为4个阶段:数据预处理、特征抽取、立场分类和后期融合。

3.1　数据预处理

数据预处理包括清洗语料、汉字简繁体转化、分词和去除停用词。在原始的微博中有很多新闻标题,表情符号和URL链接,这些信息会增加微博正文的噪音,因此,在语料清洗阶段,使用正则表达式去除这些信息。例如:

原始微博:【禁摩限电:国家权力"内卷化"之弊 - FT中文网】从3月21日开始,深圳交警实施"禁摩限电"专项集中整治活动,在街头巷尾对摩的、电动车围追堵截。根据其发布的消息,截止3月3……(分享自 @FT中文网) http://t.cn/Rq4oQ6N

清洗后微博:从3月21日开始,深圳交警实施"禁摩限电"专项集中整治活动,在街头巷尾对摩的、电动车围追堵截。根据其发布的消息,截止3月3……

如果一条微博在清洗过后不包含任何内容,那么认为这条微博是没有立场的。语料清洗过程中去除的文本和对应的示例如表1所示。语料清洗结束后,使用开源工具zhtools(https://github.com/skydark/nstools/blob/master/zhtools/)进行汉字简繁体转换,再使用结巴(jieba)(https://github.com/fxsjy/jieba)对微博进行分词。

表1　语料清洗

文本类型	示例
URL	http://t.cn/8FjOHnX
新闻标题	【春节放鞭炮 是"限"还是"禁"?】
话题	#新观察#
引用博文	发表了博文《春节》
第三方软件	秒拍视频(使用#秒拍#录制)
表情符号	[呵呵]

3.2　特征抽取

基于有监督的机器学习的方法,通过选取大量有意义的特征来完成分类任务。在文本分类中,通常选取的特征是基于词频的特征,例如*n*-gram等。本文抽取了两类特征:基于词频统计的特征和文本深度特征。基于词频统计的特征包括:基于Unigram的词袋特征、基于同义词典的词袋特征、基于主题词和立场类别标签共现关系的特征。文本深度特征是来自word2vec模型的词向量和字向量。

3.2.1　BoW

词袋模型(Bag-of-Words-BoW)[17]是最常见的文本表示方法。它在处理文本时,通常只考虑文本中是否出现词汇表中的词语,而不考虑词语顺序、句子语法或语义结构。BoW表示的特征值,可以有很多权重方法,其中最常见的是TF-IDF[18]。本文的基本文本表示方法是基于Unigram的BoW模型和对应的TF-IDF权重,该特征将作为基本特征用来与文中其他特征进行分类性能的比较。

3.2.2　S-BoW

微博用户对相同事物的表达方式各异,微博上因此存在大量的新词和生僻词。由于微博文本较为简短,且表达方式口语化、不规范化,传统的BoW一方面会非常稀疏,另一方面不能很好地处理词汇表外词语。因此本文提出基于同义词典的词袋特征,利用同义词典对关键词做同义词替换,对词义进行扩展,这样可以有效处理词汇表外词语。

本文使用的同义词典来自哈工大《同义词林》。《同义词林》按照树状的层次结构(共5层)收录了53 859条词。本文使用以下三种类别的同义词:(1)在不同上下文环境中词义都相同的词语,如"喜欢"和"爱";(2)词义可能有差别,但隶属于同一类别的词语,如"鸡"和"鸭";(3)单独词,即不包含在上述两类词的词语。

在生成词汇表的过程中,利用同义词典对每个词语进行同义词扩充。具体来说,首先利用同义词典得到每个词语的类别ID,将类别ID放入词汇表中。如果这个词语在同义词典中没有找到任何信息,则将这个词语放入词汇表。对扩展后的词汇表再使用词袋模型和TF-IDF作为文本特征。在文中后面的实验描述中,用S-BoW代表该类基于同义词典的BoW特征。

3.2.3 sVariance

微博作者表达对某个话题的立场时,很少会直接发表“支持”或“反对”这样的立场词,更多的是发表具体支持或反对的理由。例如:微博“广州的也给全部禁了吧,特别是摩托车,容易出事!”表达了对话题“深圳禁摩限电”的支持。虽然每个话题相互独立,且微博作者的表达方式千差万别,但对于相同的话题,支持者(或反对者)所谈论的核心观点是接近的。例如:“深圳禁摩限电”的话题支持者谈论的核心是电摩不遵守交通规则所带来的不安全因素,而反对者谈论的核心是政府此举为一刀切,不考虑电摩是底层老百姓的生活必需品。因此,如果能捕捉到支持(或反对)某个话题的核心内容,就能够有效地区分微博作者的立场。

为此,借鉴文献[19]中的eVector(用于进行情感分类的情感向量;通过统计词语在不同情感类别中出现的频率计算词语的权重),提出了一种新的权重计算方式sVariance。根据词语和不同立场标签的共现情况,一个词语 i 的sVariance值的计算方式如公式(1)所示,其中,$F_{i,S}$,$F_{i,A}$ 和 $F_{i,N}$ 分别是词语 i 在支持(Support/Favor)、反对(Against)、中立(None)的微博中出现的次数,$F_{i,\text{avg}}$ 是在三个类别中的平均出现次数。

$$sVar(i)=\frac{1}{3}\Big[(F_{i,S}-F_{i,\text{avg}})^2+(F_{i,A}-F_{i,\text{avg}})^2+(F_{i,N}+F_{i,\text{avg}})^2\Big]^{\frac{1}{2}} \tag{1}$$

由公式(1)可知,如果一个词在支持(或反对)的数据中出现频率较高而在中立数据中出现较少,那么它对应的sVariance值会较高,这个词语很可能是一个能表达明确立场的词语。如果一个词语在三个立场类别中出现频率都较高,那么它对应的sVariance值会较小,这个词很可能是一个中立词。因此,对每个话题,根据词语的sVariance值,可以将词语分为两类:该话题的观点词和中立词。观点词可以有效地区分微博作者的立场。表2分别列出了“深圳禁摩限电”中sVariance最高和最低的前5个词语。从表2可以看出,sVariance较高的词语都是和话题紧密相关的词语,而sVariance较低的词语是和话题无关的词语。sVariance给予中立词更小的权重,能够减少这些词语在立场分类中带来的噪声。因此,利用sVariance值基本可以区分观点词和中立词,再结合BoW形成的文本特征表示能够有效地区分作者的立场。本文使用BoW和sVariance作为第三种特征表示,在文中后面的实验描述中,用sVariance代表该类特征。

表2 话题“深圳禁摩限电”词语的sVariance举例

	词语	sVariance值
sVariance值最高	交警	0.312 629
	支持	0.231 549
	汽车	0.189 035
	政府	0.172 308
	老百姓	0.143 463
sVariance值最低	几乎	0.001 707
	最终	0.001 707
	全面	0.001 707
	书记	0.001 707
	进入	0.001 707

3.2.4 词向量

Word2vec[20]是Google在2013年开源的一款将词表示为实数值向量的工具,它利用深度学习的思想,把对文本内容的处理简化为 K 维向量空间中的向量运算,而向量空间上的相似度可以用来表示文本语义上的相似度。Word2vec的词向量可以获取词语之间的相似度信息,因此可以处理数据稀疏的问题。近年来,已经有研究利用这个工具进行情感分析[21-22]。

本文利用Gensim[23]和维基百科中文语料(https://dumps.wikimedia.org/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2),训练了一个400维的word2vec模型。通过这个模型,可以得到每一个词语的向量表示。这些词向量已经包含了词语的一些语义信息。表3列出了在word2vec模型中,与“老百姓”的词向量最相似的5个词语以及它们之间词向量的余弦距离。从表3可以看出,余弦距离较近的词语所表达的词义信息也相近,说明通过word2vec得到的词向量已经从大规模语料中学习到了语义信息。这些语义信息可以有效地解决微博文本稀疏问题,更好地处理表外词语,尤其是对话题中的观点词的扩展,能够更高效地判定微博作者的立场。

表3 词向量相似度示例

相似词	词向量余弦距离
百姓	0.691 456 556
平民百姓	0.601 972 401
民众	0.574 062 586
农民	0.572 304 368
穷人	0.560 522 556

将一条微博中所有词语的向量进行平均,作为这条微博文本的词向量表示。

3.2.5 字向量

字符层面的特征在短文本的处理中通常能发挥更有效的作用[24]。微博文本的总长度不超过140个字符,因此,本文使用字层面的特征来表示微博文本。在文本预处理时不进行分词,而进行字切分。接着使用上文提到的word2vec模型,得到每个字的向量。将一条微博的所有字的向量进行平均,作为这条微博的字向量表示。

3.3 立场分类

本文基于上述的文本特征,使用支持向量机(SVM)、随机森林(Random Forest)以及梯度提升决策树(GBDT)作为立场检测的分类器。SVM是常见的文本分类算法,广泛用于基于词频特征的文本分类。随机森林和GBDT都是基于决策树的组合模型。随机森林在学习

时，采用随机的方式建立一个由若干决策树构成的森林，决策树之间相互独立，在对新样本进行分类时，每一棵决策树对样本单独判断，根据多数投票原则决定样本最终的类别。GBDT在学习时，每一次新模型的建立依据前一次模型的残差，并朝着前一次模型的残差梯度下降的方向产生新的模型。决策树的优点是训练时间较短，预测过程快速。随机森林有一个显著优点是能够处理高维度的数据而不用做特征选择，而GBDT更擅长处理连续特征。

每一个话题的微博文本，共分析了五种特征表示。在立场分类时，BoW，S-BoW和sVariance分别使用SVM，随机森林进行分类，词向量和字向量使用SVM，随机森林和GBDT进行分类。

3.4　后期融合

利用不同的特征和分类算法进行立场检测时，会产生不同的子分类器。在进行后期融合时，使用不同子分类器对样本预测概率的加权平均作为最终的立场判断的预测概率，并选择预测概率最大的立场类别作为最终的立场标签。

设特征 f_i 对样本 x 属于类别 c 的预测概率为 $P_{f_i}(c|\chi)$，它在参与融合时的权重为 a_{f_i}，共有5类特征参与融合，最终预测概率如公式(2)所示。若特征 f_i 不参与融合，则对应权重为 a_{f_i} 为0。

$$P(c|\chi)=\sum_{i=0}^{5}a_{f_i}\times P_{f_i}(c|\chi) \tag{2}$$

其中，$\sum_{i=0}^{5}a_{f_i}=1$。

$P_{f_i}(c|\chi)$ 是特征 f_i 对样本 x 属于类别 c 的预测概率，是该特征结合不同分类器产生的平均预测概率，BoW，S-BoW和sVariance对应的 $P_{f_i}(c|\chi)$ 的预测概率如公式(3)中(a)所示，词向量和字向量对应的 $P_{f_i}(c|\chi)$ 如公式(3)中(b)所示。

$$P_{f_i}(c|\chi)=\begin{cases}\dfrac{1}{3}(P_{s_i}(c|\chi)+P_{r_i}(c|\chi)+P_{g_i}(c|\chi)) & \text{(a)}\\ \dfrac{1}{2}(P_{s_i}(c|\chi)+P_{r_i}(c|\chi)) & \text{(b)}\end{cases} \tag{3}$$

其中 $P_{s_i}(c|\chi)$，$P_{r_i}(c|\chi)$，$P_{g_i}(c|\chi)$ 分别是SVM、随机森林和GBDT对样本 x（基于特征 f_i）属于类别 c 的预测概率。由于原始的SVM不能产生样本所属类别的概率输出，使用Softmax函数将SVM的决策值转化为概率输出，如公式(4)所示，将决策函数值 $z=f(x)$ 单调映射到区间[0，1]，因此，可把 $\sigma_{\text{softmax}}(z)$ 当作概率处理。$P_{r_i}(c|\chi)$ 是随机森林对样本的预测概率，如公式(5)所示，其中 T 为森林中树的棵数，$P_j(c|\chi)$ 为森林中每棵树对样本 x 划分到类别 c 的预测概率。$P_{g_i}(c|\chi)$ 是GBDT对样本的输出概率，与决策树的输出概率(即测试样本所划分到的叶子节点的样本个数所占的比例)一致。

$$P_{s_i}(c|\chi)=\sigma_{\text{softmax}}(z)=\frac{1}{1+e^{-2z}} \tag{4}$$

$$P_{r_i}(c|\chi)=\frac{1}{T}\sum_{j=0}^{T}P_j(c|\chi) \tag{5}$$

4　实验与分析

4.1　数据

本文实验的数据集来自NLPCC2016中文微博立场检测的评测任务[8]。数据来源是新浪微博，一共包含5个话题的微博及其立场标签，共4 000条数据。每条数据提供了微博ID，微博对应的话题，文本信息和立场标签，例如：

<ID>4
<Target>深圳禁摩限电
<Text>只要骑电动车，拘不拘还不是警察说的算
<Stance> AGAINST

数据集涉及的话题分别是："iPhone SE"、"俄罗斯叙利亚反恐行动"、"开放二胎"、"春节放鞭炮"、"深圳禁摩限电"。立场标签分别是："FAVOR(支持)"、"AGAINST(反对)"、"NONE(中立)"。每个话题共800条数据，实验中每个话题划分了120条数据作为验证集用来选择分类器最优参数，200条作为测试集用来评测分类算法。

4.2　评测指标

在立场检测中通常更关心有明确立场的内容，因此实验选取的评测指标是"支持"和"反对"的平均 F 值，计算如公式(6)所示，其中 F_{favor} 和 F_{against} 分别是"支持"和"反对"的 F 值。

$$F_{\text{avg}}=\frac{F_{\text{favor}}+F_{\text{against}}}{2} \tag{6}$$

4.3　实验参数设置

实验中主要参数有：SVM核函数及代价值；随机森林中树的棵树和树深；GBDT的迭代次数和树深；后期融合时各特征分类器的权重。使用Scikit-learn[25]中实现的SVM，随机森林和GBDT进行实验。根据经验，将分类器的参数设置如下：

SVM：RBF核函数，代价值在 2^{-2} 到 2^{10} 之间。

随机森林：100棵决策树，每棵树的树深在2到30之间。

GBDT：100轮迭代，学习率为0.2，树深在2到30之间。

其中SVM的代价值，随机森林、GBDT中的树深，通过优化验证集的 F_{avg} 确定。

后期融合时，对不同特征的分类器设置相同权重参与融合，即平均各子分类器预测概率作为最终后期融合的预测概率。例如，设S-BoW，sVariance和词向量对某话题下的样本 x 属于"支持"该话题的预测概率分别为0.6，0.4，0.7，则三个特征参与融合时，各自的权重为0.33，

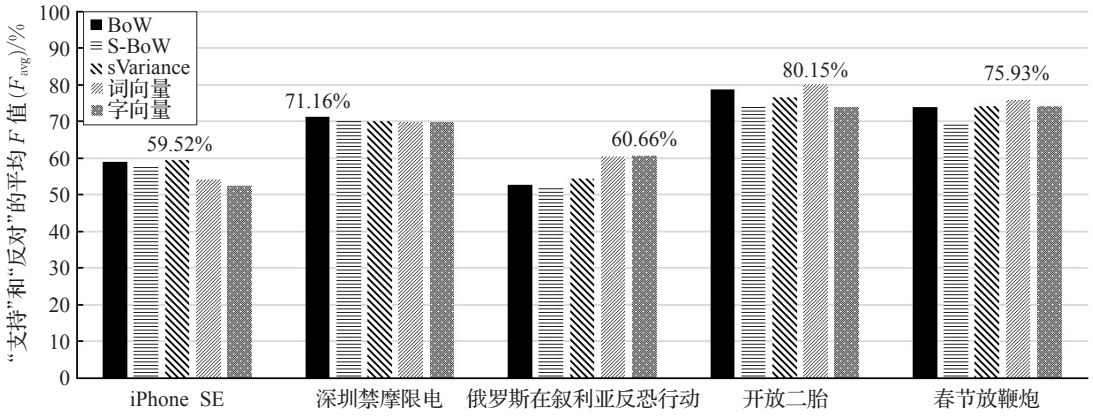

图2　单特征立场分类结果

最终样本 x 属于“支持”该话题的预测概率为0.51。

4.4　立场分类的结果与分析

由于话题相对独立,实验中为每个话题单独抽取特征并建立分类模型。

图2展示了5个话题的单特征的分类情况。如图2所示,不同的特征在不同话题上的表现不同。为了更好地分析每个话题上特征的表现,表4还列出了每个话题下最好特征的 F_{avg} 值以及与基本特征之间的比较。

表4　不同话题下表现最好的单一特征

话题	最好特征	F_{avg}/%	F_{avg}(BoW)/%
iPhone SE	sVariance	59.52	58.87
深圳禁摩限电	BoW	71.16	71.16
俄罗斯在叙利亚反恐行动	字向量	60.66	52.76
开放二胎	词向量	80.15	78.85
春节放鞭炮	词向量	75.93	73.87

4.4.1　文本深度特征性能分析

整体而言,来自word2vec的特征表现比较突出。具体来说,对于“俄罗斯叙利亚反恐行动”,字向量表现最好,达到60.66%的 F_{avg} 值,比BoW提升了14.97%。表5列出了不同话题的微博平均长度。由表5可以看到,这个话题的微博非常简短,每条微博平均长度是49个字,训练语料中的单字比例占了词汇表的10.13%,因此每个字在一条微博中的作用相比词更重要。另一方面,这个话题属于新闻话题的范畴,在维基百科上有大量的相关语料提供,因此给词语的向量表示提供了更多有意义的信息。对于“开放二胎”和“春节放鞭炮”,词向量表现最好,分别比BoW提升了2.6%和2.82%。因为在维基百科中,相关话题的语料比较丰富,word2vec学习到了更丰富的词汇表示。例如,在word2vec模型中,通过词向量的相似度计算,与“二胎”最相似的词语是:“独生子女”,“二孩”,“三胎”,“买房”;与“鞭炮”最相似的词语有:“爆竹”,“烟花爆竹”,“燃放”,“春节”,“放炮”,“过年”。这些词语都是和话题紧密相关并且微博作者讨论较多的。从这里可以看出,这些词向量已经包含了词语的语义信息(如同义词信息等),一方面扩充了词汇表,另一方面也提供了额外的有助于文本分析的信息。但“iPhone SE”的词向量和字向量的表现并不够好,猜测与训练word2vec的语料有关,维基百科提供的语料是比较通用领域的,而“iPhone SE”这个话题的讨论偏向电子产品的用户反馈,因此维基百科并不能提供额外的帮助。

表5　不同话题的微博平均长度

话题	微博平均长度
iPhone SE	68
深圳禁摩限电	65
俄罗斯在叙利亚反恐行动	49
开放二胎	82
春节放鞭炮	102

4.4.2　S-BoW性能分析

对5个话题而言,S-BoW并没有明显改善性能。基于数据观察,表6列出了5个话题中话题支持者和反对者的谈论中心。如表6所示,这些话题的支持者(或反对者)讨论的内容比较近似,因此使用的一些有代表性的、能显示作者立场的词语都很接近,使得产生的新词数量不多,同义词典并不能发挥作用。另一方面,微博上充满了大量的网络用语和微博作者个人特色的发言,而同义词典相对比较书面化,很难捕捉到微博上词语的同义词信息。

表6　不同话题的支持者反对者谈论核心

话题	支持者	反对者
iPhone SE	手机的属性	手机的属性
深圳禁摩限电	交通秩序混乱	政府一刀切
俄罗斯在叙利亚反恐行动	无	无
开放二胎	独生子女	女性生育权、经济压力
春节放鞭炮	传统习俗	环境污染

4.4.3　sVariance性能分析

sVariance在“iPhone SE”这个话题上表现最好,F_{avg} 值达到59.52%,比BoW特征提高了1.1%。这个话题的微博类似于产品评论,微博作者描述更多的是手机的属性,如“屏幕”、“尺寸”、“配置”、“外观”等。在对比

这个话题词语的TF-IDF值和sVariance值时发现，这些属性词的TF-IDF值比较低，而sVariance值比较高。这说明TF-IDF和sVariance都能够将属性词区分出来。但是sVariance更能够把描述属性词的观点词区分出来。例如：在sVariance值最高的前20个词语中包含“喜欢”、“好”、“小”、“不错”、“好看”，这些词语能够比较明确判断微博作者的立场。除了“iPhone SE”，sVariance在其他话题上并没有明显改善性能。通过分析这些话题的具体词语的sVariance值发现，sVariance较高的词语通常是在有立场（支持或反对）的数据样本中出现，而在没有立场（中立）的数据样本中出现的词语sVariance值较低，因此sVariance可以较好地区分微博是否有立场。但是，正如之前提到，这些话题中，微博用户表达支持或者反对的观点比较集中，支持者和反对者各自所站的角度不同，因此他们讨论的内容很少有交集，使得支持者和反对者使用的有代表性的词语之间的交集也比较少。在支持（或反对）的数据样本中出现的词语只在支持（或反对）的数据样本中出现，例如“横冲直撞”只出现在支持“深圳禁摩限电”的话题中。因此，sVariance在具体区分支持和反对上表现不够好。

4.4.4　特征与话题适应度分析

在前面的分析中提到，“iPhone SE”这个话题的微博，更像是产品评论。对产品评论的立场分析与情感分析类似，重要的是区分出文本中产品的属性和描述属性的部分。在本文提取的特征中，sVariance能够有效地提取属性词和部分属性描述词。由于大部分属性描述词是形容词，同义词典中包含的形容词的同义信息也较为充足，因此在更规范的产品评论文本中，S-BoW也能对属性描述词的扩充有帮助。

“俄罗斯叙利亚反恐行动”这个话题，与新闻政治评论比较类似。由于缺乏背景知识，即使人为分析，这个话题的立场检测都比较困难。这个话题的微博比较简短，单纯地统计词频信息并不能对文本进行有效分析。因此需要借助其他语料来丰富词汇和语义信息，尤其是新闻政治领域的语料。因此，本文提取的特征中，使用维基百科语料训练的word2vec模型所提供的词和字向量，在这个数据集上能够表现出良好性能。

“深圳禁摩限电”、“春节放鞭炮”、“开放二胎”，这三个话题的微博有相似的特点。首先，这三个话题与人们的生活息息相关，是日常热点话题，微博作者愿意表达更多内容，因此这三个话题的词汇信息比较丰富。其次，微博作者在表达立场观点时讨论的内容比较集中，通过分析一些词汇信息，就能大致判断出作者立场。因此，在本文提取的特征中，BoW在这三个话题上的表现都不错。而文本深度特征在这三类话题上的良好表现说明word2vec在通用语料上学习到的大量词汇和上下文信息对于这类日常讨论较多的话题的分析也是有帮助的。

4.5　后期融合的结果与分析

实验发现，后期融合都能对分类结果带来极大的提升。表7给出了后期融合中每个话题表现最好的特征组合以及与该话题下表现最好的单特征的比较。从表7可以看出，第一，5个话题的最优特征组合中，都包含了文本深度特征。文本深度特征，不仅在单独的立场分类中表现优异，在与其他基于词频统计的特征融合后，也对结果有显著提升。第二，在单特征分类中表现一般的S-BoW和sVariance，在与其他特征结合后也对分类结果有提升。经过后期融合，5个话题的 F_{avg} 值相较于每个话题表现最好的单特征平均提升了4.63%。实验证明，本文所抽取的特征能够捕捉到微博中不同的信息，这些信息对于微博作者的立场判断是互补的，能在不同程度上提升立场分类的表现。

5　结束语

本文在基于监督学习的文本分类框架上，扩展并提出基于多文本特征融合的中文微博的立场检测方法。方法包括预处理、特征抽取、立场分类和后期融合四个阶段。在特征抽取阶段，本文探究了五种文本表示方法，包括：词袋特征（BoW）、基于同义词典的词袋特征（S-BoW）、考虑词与立场标签共现关系的特征（sVariance）以及利用word2vec得到的文本深度特征（词向量和字向量）。之后，使用SVM，随机森林和GBDT对上述特征进行立场分类。最后，结合所有特征分类器进行后期融合，并预测立场标签。实验结果表明，对于不同的话题，文中提取的特征相较于传统的BoW特征，都有明显提升。并且在所有话题上，特征的后期融合对立场分类结果都有明显改善。说明文本深度特征和基于词频统计的特征能够捕捉到文本的不同信息，在立场检测中是互补的。此外，考虑实验中的5个话题的特性与本文提取的特征之间的适应关系，本文介绍的文本分类方

表7　后期融合的分类表现

话题	特征组合	F_{avg}/%	F_{avg}（最好单特征）/%
iPhone SE	BoW+S-BoW+词向量+字向量	61.49	59.52
深圳禁摩限电	S-BoW+词向量+字向量	78.16	71.16
俄罗斯在叙利亚反恐行动	S-BoW+词向量+字向量	61.96	60.66
开放二胎	sVariance+词向量	84.69	80.15
春节放鞭炮	BoW+sVariance+词向量+字向量	77.61	75.93

法对于其他文本分类任务也是有效的,例如产品评论的情感分析、产品属性词的抽取、新闻评论的情感倾向性分析等。使用本文介绍的方法在NLPCC2016的中文微博的立场检测评测任务中取得了第一名[8]。

本文探究了五种话题,并针对这五种话题进行问题分析和立场检测,但微博上的话题种类繁多且具有不同的特点,在今后的工作中,将致力于探索更普适的方法进行立场检测。另外,文本深度特征在立场检测任务中的有效性已经得到证实,在今后的工作中,还将继续探索更多的文本深度特征及其在立场检测中的应用。

参考文献:

[1] Pang B,Lee L.Opinion mining and sentiment analysis[J].Foundations & Trends in Information Retrieval,2008,2(1/2):1-135.

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

[3] 周立柱,贺宇凯,王建勇.情感分析研究综述[J].计算机应用,2008,28(11):2725-2728.

[4] 周胜臣,瞿文婷,石英子,等.中文微博情感分析研究综述[J].计算机应用与软件,2013,30(3):161-164.

[5] 高凯,李思雨,阮冬茹,等.基于微博的情感倾向性分析方法研究[J].中文信息学报,2015(4):40-49.

[6] 刘全超,黄河燕,冯冲.基于多特征微博话题情感倾向性判定算法研究[J].中文信息学报,2014,28(4):123-131.

[7] 刘龙飞,杨亮,张绍武,等.基于卷积神经网络的微博情感倾向性分析[J].中文信息学报,2015,29(6):159-165.

[8] Xu R,Zhou Y,Wu D,et al.Overview of NLPCC shared task 4:Stance detection in Chinese Microblogs[M]//Natural language understanding and intelligent applications.[S.l.]:Springer International Publishing,2016.

[9] Pang B,Lee L,Vaithyanathan S.Thumbs up? Sentiment classification using machine learning techniques[C]//Isabelle P.Proc of the EMNLP 2002.Morristown:ACL,2002:79-86.

[10] 刘鲁,刘志明.基于机器学习的中文微博情感分类实证研究[J].计算机工程与应用,2012,48(1):1-4.

[11] 傅向华,刘国,郭岩岩,等.中文博客多方面话题情感分析研究[J].中文信息学报,2013,27(1):47-56.

[12] Li F,Huang M,Zhu X.Sentiment analysis with global topics and local dependency[C]//Twenty-Fourth AAAI Conference on Artificial Intelligence,AAAI 2010,Atlanta,Georgia,USA,2010:1371-1376.

[13] Murakami A,Raymond R.Support or oppose? Classifying positions in online debates from reply activities and opinion expressions[C]//Proceedings of the International Conference on Computational Linguistics(ACL),2010:869-875.

[14] Sridhar D,Getoor L,Walker M.Collective stance classification of posts in online debate forums[C]//Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media,2014:109-117.

[15] Wojatzki M,Zesch T.ltl.uni-due at SemEval-2016 task 6:Stance detection in social media using stacked classifiers[C]//NAACL Hlt,2016.

[16] Liu C,Li W,Demarest B,et al.IUCL at SemEval-2016 task 6:An ensemble model for stance detection in twitter[C]//International Workshop on Semantic Evaluation,2016.

[17] Harris Z S.Distributional structure[J].Word,1954,10(2/3):146-162.

[18] Jones K S.A statistical interpretation of term specificity and its application in retrieval[J].Journal of Documentation,1972,60(1):493-502.

[19] Li C,Wu H,Jin Q.Emotion classification of Chinese microblog text via fusion of bow and evector feature representations[M]//Natural language processing and Chinese computing.Berlin Heidelberg:Springer,2014:217-228.

[20] Mikolov T,Chen K,Corrado G,et al.Efficient estimation of word representations in vector space[J].Computer Science,2013.

[21] Xue B,Fu C,Shaobin Z.A study on sentiment computing and classification of Sina Weibo with Word2vec[C]//2014 IEEE International Congress on Big Data,2014:358-363.

[22] Zhang D,Xu H,Su Z,et al.Chinese comments sentiment classification based on Word2vec and SVM perf[J].Expert Systems with Applications,2015,42(4):1857-1863.

[23] Rehurek R,Sojka P.Software framework for topic modelling with large corpora[C]//Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,2010.

[24] Zhang X,LeCun Y.Text understanding from scratch[J].Computer Science,2015.

[25] Pedregosa F,Varoquaux G,Gramfort A,et al.Scikit-learn:Machine learning in python[J].Journal of Machine Learning Research,2012,12(10):2825-2830.