文章编号:1671-9352(2014)11-0001-07 DOI:10.6040/j.issn.1671-9352.3.2014.194

基于中文微博语料的情感倾向性分析

罗毅,李利,谭松波,程学旗
(中国科学院计算技术研究所,北京 100190)

摘要:微博的兴起与传播使得短文本情感分类成为目前的热门研究领域。通过对中文微博语料的情感倾向性分析进行研究,提出了一种新的情感分类方法。首先构建了两级情感词典,并对不同级别情感词作不同增强;然后在情感特征方面使用N-Gram方法,尽量获取有限长度博文中的未登录情感词和情感信息。经实验验证与传统方式相比较,该方法的准确率和召回率都有所提高,在COAE2014微博情感倾向性评测任务中也取得了较好的成绩。

关键词:情感分类;倾向性分析;观点挖掘
中图分类号:TP391 **文献标志码**:A

Sentiment analysis on Chinese Micro-blog corpus

LUO Yi, LI Li, TAN Song-bo, CHENG Xue-qi
(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The rise and spread of Micro-blog make sentiment classification on short texts become a hot area. A new method was proposed for Micro-blog sentiment classification. First of all, this method will create an emotional dictionary with two-levels, and the words for different levels will get different enhancement; then in order to get features, N-gram method was used, which found new emotional words and emotional information from a short text. The experiment results show this approach has improved precision and recall rate compared to the traditional ways. This algorithm also did a very good job in COAE 2014.

Key words: sentiment classification; tendentious analysis; opinion mining

0 引言

微博是一个基于关系的信息制造、交流、传播以及获取的社交媒体,网民从原来单纯的信息接收方过渡到了接收和发布信息的完全参与方。根据新浪微博发布的最新财报统计数据显示,新浪微博的注册用户已经超过了5.36亿,用户每日新增的发博量超过1亿条。由于微博的开放和即时传播特性,信息会在极短时间内得到大范围的传播。面对海量微博信息,如何从其中提取有用数据,已经成为了现如今各领域研究的热点。文本情感倾向性分析作为相关的研究方向,其重要性日益凸显。倾向性分析是指分析作者在传达信息时所隐含的情绪状态,对作者的意见进行判断或者评估,给出作者态度是褒义、贬义还是中性的结论。企业和商家面对微博用户发布的大量情感微博信息,可以不断挖掘有倾向的数据,统计用户对自家产品的反馈,

收稿日期:2014-08-28; **网络出版时间**:2014-10-24 14:20
网络出版地址:http://www.cnki.net/kcms/doi/10.6040/j.issn.1671.9352.3.2014.194.html
基金项目:国家自然科学基金资助项目(61232010,61100083);国家重点基础研究发展计划("九七三"计划)项目(2013CB329601/02);国家高技术研究发展计划("八六三"计划)项目(2012AA011003);国家科技支撑计划项目(2012BAH39B04);国家安全专项项目(2013A140)
作者简介:罗毅(1989-),男,硕士研究生,主要研究方向为情感分类、自然语言理解. E-mail:luoyi@software.ict.ac.cn

同时也可以指导广告的精准投放,向消费者推荐潜在消费品。

1 相关研究现状

句子级情感分析研究可以应用有监督、无监督和半监督三种方法完成。Davidiv 等人[1]设计了一种有监督分类器,借鉴了 K-近邻法(K-nearest neighbors,KNN)的思想,利用表情符号和 hashtag 将 tweet 划分为多种情感类型。Barbosa,等[2]使用从三个不同的 Twitter 情感分析网站上获取到的训练数据训练标准的 SVM 分类器,实验结果表明 SVM 分类器可以达到 81.3% 的精度。Hassan 等人[3]运用监督型马尔科夫模型来确定 Usenet 上的消息极性,使用的是依存关系和词性信息。A. Meena 等[4]重点分析了连词对语句情感倾向性分析的影响,结合连词和短语分析语句情感极性,不过其系统不具有领域适应能力。Turney[5]提出一种无监督的方法,主要应用在产品和电影评论的消极和积极分类上。Socher 等[6]将句子中词语两两合并,递归的构建短语树,使用短语节点特征判断句子情感类别。赵军和王根[7]提出情感语句分级模型,缺点是会有错误逐层传播的现象。Tan 等[8]认为社交信息可以提高倾向性分析准确度,提出了基于用户社交关系和半监督学习的倾向性分析模型。Li 等[9]提出的半监督倾向性分析模型,一定情况下解决了不平衡语料问题。

为了促进文本情感倾向性分析技术的发展,一些相关评测也适时出现在国际上,例如 TREC Blog Track[10]主要检索英文文本中的观点信息;NTCIR[11]主要任务是对英、日、韩、中文文本进行情感分类以及观点要素抽取。而在国内,中文信息学会信息检索专委会主办的中文倾向性分析评测(Chinese opinion analysis evaluation, COAE)迄今已经举办了五届。从 COAE2013 开始,加入了"短文本情感倾向性分析"相关任务。这类评测极大促进了中文倾向性分析语料库的建设,推动了观点信息检索、抽取以及倾向性分析等研究的进展。

目前,对中文微博进行情感倾向性分析有着以下几点困难:(1)微博数据量巨大,且更新频繁,每天新增量级在一亿以上,这就要求提出通用、有效的情感分类方法;(2)中文微博训练语料匮乏,国际上主要研究对象为英文短文本,利用 Twitter 建立了完善的语料集,而国内微博情感语料集才刚开始建设;(3)中文微博比英文微博包含信息量更丰富;(4)微博文本中包含大量缩写词、口语和新网络用语。相比研究领域的情感倾向性分析,微博语言结构不严谨,口语、缩写居多,文本内容简洁,不停地出现新兴网络用语。因此,现阶段微博情感倾向性分析准确度不高,单纯使用国外情感分析方法很难适应微博文本的多样性。

2 微博语料的情感分析

短文本情感倾向性分析本质上是一个基于主题的文本分类问题,将微博短文本做两类分类,最终归纳到正面和负面两种情感类别中。所以该问题目前使用的方法与传统的文本分类方法有相似之处。但这两个问题也有本质区别:一是情感倾向性分析更偏重语言本身的意义,所以在分类的过程中需要很多语言、语义方面的信息;二是情感倾向性分析对情感资源敏感,往往情感词典的好坏决定了情感分类的成败。针对第一个区别,本文提出以 N-Gram 语言模型为基础,考虑了各种长度的"片段"tf-idf 值作为文本特征,并引入"情感指数"概念,从维度巨大的特征集中找出和情感最相关的语义特征,构造分类器分类。针对第二个区别,本文在构建情感词典时提出"情感词分级"思想,对情感词典进一步细分为:"核心情感词"、"普通情感词",完善情感词典。虽然 N-Gram 语言模型虽然可以囊括所有语言、语义信息,但是对所有特征都一视同仁,而情感分析过程中的主要事实是,情感分类中情感词语信息起到的作用远不是普通特征可以比拟的,所以我们在计算特征的 tf-idf 时,对情感词语特征做一定程度增强,另外不同级别的情感词也会得到不同程度的增强。通过实验对比及 COAE 2014 实际评测任务的检验,本文方法优于传统情感分析方法,而且实际应用有效性也得到了验证。以下主要介绍本文所采用的短文本情感倾向性分析方法。将从两个方面做重点阐述:情感资源构建和特征提取与选择。

2.1 情感资源构建

2.1.1 训练语料和情感词典构建

训练语料的质量和规模通常是影响分类效果至关重要的因素。在预处理过程中采用的微博语料集来自

COAE 和 NLP&&CC 两大评测,正负样本均衡。虽然同为微博语料,但该样本所包含的博文内容都太过老旧,发布时间均为 2011—2012 年,显然已经不再适合最新的 COAE 微博评测任务。基于以上事实,本文采集了 2013 下半年的最新原创微博(非转发),从中挑选出了一些倾向性明显的微博标注并加入微博语料。最终,训练集总体规模为 57 284,正负类平衡。

本文采用 Hownet 情感词语集和清华大学情感词集作为初始情感词典,但作为微博倾向性分析,传统情感词典是不能完全适用的,比如微博流行用语:“高大上”就是“高端大气上档次”的缩写。基于最新的微博新词热词榜,人工找出部分情感新词添加入情感词典,比如“给力”、“高大上”等。最终得到的情感词典规模为 24 570,其中正面情感词数目为 11 599,负面情感词数目为 12 971。

2.1.2　情感词分级

本文中的情感词典需要进行分级操作,即对现有情感词典分成两级:核心情感词典和普通情感词典。核心情感词是情感词典中常见,且情感倾向高的情感词。而普通情感词典则是对现有情感词典剔除掉核心情感词所产生的。这是一个两类分类问题,首先采用语料初选过滤掉一部分普通情感词,然后采用朴素贝叶斯分类方法,对小范围训练词语提取相关特征,训练分类器,最终完成全部词语的分类。

(1)初选过滤

在接近 6 万的微博语料中存在的情感词不一定是核心情感词,但是语料中不存在的情感词一定会是普通情感词。通过此假设,可以对情感词典中的所有词做初选过滤,找到语料中未出现的情感词并将其加入到普通情感词集中。通过这一步操作,可以过滤出两个集合:普通情感词集,规模为 6 137;未定情感词集,规模为 18 433。

(2)训练语料准备

对词语进行训练时,需要有已经标注好的训练语料,语料的样本将是分好类的词语。本文随机选择原情感词典规模的 2% 作为训练语料:在普通情感词集中选择 122 个词,将这些词标注为普通情感词,另外从未定情感词集中随机选取 368 个词进行人工标注。形成规模为 490 的训练语料,其中标注为普通情感词的有 303,核心情感词为 187。以核心情感词为准,对两类标注样本均取 187 个词,组成一个最终版本的训练样本,规模为 374,两类样本均衡。

(3)特征选择

定义情感词的“相关微博”为含有该词语的微博,“相关微博”能为词语的情感倾向性判断提供指导,比如含有问号、感叹号的微博中,如果存在情感词,那该情感词可能是一个核心情感词。另外定义“出现率”为满足条件的相关微博数/相关微博总数。如“@ 标签出现率”即为含有@ 标签的相关微博数/相关微博总数。本文从微博训练语料中抽取的词语特征如下:

- 词语本身特征。词语本身的特征是分类最重要的决策依据,而这方面能够抽取的可选项很少,最终抽取如下特征:1)词语长度;2)词语在语料中的词频;3)包含词语的负类微博数;4)包含词语的正类微博数。
- 微博内容特征。相关微博中的一些特有属性对微博的情感倾向性虽然没有多少帮助,但间接可能会与情感有关联,比如@ 标签。另外一些则是与情感关联很大的特征,比如问号和感叹号。注意到如果微博中重复出现多个“%”、“;”、“数字编号”将意味着该条微博更多的是在陈述事实,这种微博一般不会带有强烈的主观倾向性。最终本文从含有该情感词的相关微博中统计出如下特征:1)@ 标签出现率;2)地理位置出现率;3)短链出现率;4)问号出现率;5)感叹号出现率;6)省略号出现率;7)多个“%”出现率;8)多个“;”出现率;9)“数字编号”出现率。
- 微博表情特征。微博中一旦具有表情,则比较可能是带有强烈主观倾向性的相关微博,它们对于确定情感词是否是核心情感词具有重大意义。本文从语料中统计出如下六种常见表情的出现率作为特征:1)[哈哈];2)[嘻嘻];3)[嘿嘿];4)[怒];5)[发火];6)[失望]。

(4)训练和分类

使用朴素贝叶斯分类算法进行训练,学习出情感词分类器。并对未定情感词典中还未确定下来的 18 065个情感词进行分级。最终,分好类的核心、普通情感词典规模如表 1 所示。

表1 情感词典的分级
Table 1 Classification of emotion dictionary

情感词典类别	词典规模	词语举例
核心情感词典	正向:2 810	一帆风顺、公平、友爱、纯真
	负向:2 876	一蹶不振、亡命、失败、沮丧
普通情感词典	正向:8 789	得体、言之有理、没说的、不紧张
	负向:10 095	七零八落、呆愣愣、乱烘烘、紧张

注:固定搭配或词组也会加入词典,比如不紧张、令人失望等。

2.2 特征提取和选择

要提取微博文本的特征,目前的主要做法是对微博进行分词,匹配情感词典,选用其中的情感词或者情感词的相关得分作为特征。但是微博属于短文本范畴,噪声大、新词多、缩写频繁、有自己的固定搭配。对微博做分词歧义明显,往往得到的是不好的切分。比如:"我发现了一个高大上网站:去哪儿网",在该句中,"高大上网站"如果使用传统分词技术,会被切分为"高大/上/网站"或"高大/上网/站"。这样的切分无法体现句子的正确语义,甚至后一种切分还将"网站"切分导致丢失评价对象。

基于以上情况,本文不再使用分词来获取微博特征,而采用语言模型中的N-Gram,切分出所有可能的词语片段。而所有切分出的片段都会被作为特征进行下一步处理。

2.2.1 微博表示

本文使用传统向量空间模型(vector space model, VSM)中的向量来表示每一条微博,其中的特征值使用tf-idf方法计算。定义第 i 个文档 $\boldsymbol{D}_i$ 的向量为

$$\boldsymbol{D}_i = (t_{i1}, t_{i2}, \cdots, t_{in}), \tag{1}$$

其中 $t_{ij} \in \mathbf{R}$, 表示词典中第 j 个词语 T_j 在文档集中的tf-idf值。假设TS为词典集,则 n 为TS的词规模,即 $T_j \in \mathrm{TS}, n = |\mathrm{TS}|$。定义文档向量 $\boldsymbol{D}_i$ 的各维度值 tfidf_{ij} 如式(2)所示:

$$\mathrm{tfidf}_{ij} = \frac{n_{ij}}{\sum\limits_k n_{ik}} \times \mathrm{lb}\frac{|\mathrm{DS}|}{|\{\boldsymbol{D}_i | T_i \in \boldsymbol{D}_i\}|}。 \tag{2}$$

其中,DS为所有文档 $\boldsymbol{D}_i$ 组成的集合,$|\mathrm{DS}|$ 为集合DS的大小;n_{ij} 为词语 T_j 在文档 $\boldsymbol{D}_i$ 中出现的次数;$n_{ij}/\sum\limits_k n_{ik}$ 是词频(term frequency, TF)部分,表示词 T_j 在文档中出现的频率。而 $\ln(|\mathrm{DS}|/|\{\boldsymbol{D}_i | T_i \in \boldsymbol{D}_i\}|)$ 则是逆文档频率(inverse document frequency, IDF)部分,是出现词 T_j 的文档数占总文档数目比值的倒数,衡量该词是否常见。

本文通过对tf-idf的值做向量归一化来计算特征值,定义如下:

$$w_{ij} = \frac{\mathrm{tfidf}_{ij}}{\sqrt{\sum\limits_{j=0}(\mathrm{tfidf}_{ij})^2}} = \frac{n_{ij}}{\sum\limits_k n_{ik}} \times \mathrm{lb}\frac{|\mathrm{DS}|}{|\{\boldsymbol{D}_i | T_i \in \boldsymbol{D}_i\}|} \Big/ \sqrt{\sum_{j=0}\left(\frac{n_{ij}}{\sum\limits_k n_{ik}} \times \mathrm{lb}\frac{|\mathrm{DS}|}{|\{\boldsymbol{D}_i | T_i \in \boldsymbol{D}_i\}|}\right)^2}。 \tag{3}$$

其中 w_{ij} 即为最终词 T_j 在文档 $\boldsymbol{D}_i$ 中的特征值,经过归一化处理后,w_{ij} 将在[0,1]的范围内。

2.2.2 情感词增强

采用N-Gram语言模型来切分微博并计算特征值,将会导致每一个切分出来的词语片段对微博语句的情感贡献都一样。这显然与实际情况不符合,理论上来说,情感词对语句贡献最大,所以需要从语句中找出对语句情感倾向性有贡献的情感词,做特殊的处理。

在对词语片段计算tf-idf值时,如果遇到情感词,需要对tf-idf计算过程做一些小改动,即引入情感增强系数 X,表示对该特征的增强。增强后该特征在tf-idf计算方式中的维度值 t_{ij} 如式(4)所示:

$$t_{ij} = \frac{X * n_{ij}}{(\sum\limits_k n_{ik}) + (X-1) * n_{ij}} \times \mathrm{lb}\frac{|\mathrm{DS}|}{1 + |\{\boldsymbol{D}_i | T_i \in \boldsymbol{D}_i\}|}。 \tag{4}$$

其中,IDF部分加1避免除数为0;TF部分对情感词做了 X 倍的增强,其中 $X \geqslant 1$。公式(4)的物理意义为:情感词对微博情感倾向性判断极其重要,所以为其增强 X 倍。这样的假设并不影响IDF计算,但TF计算方式将发生变化,可以理解为该条微博中 T_j 出现了 $X * n_{ij}$ 次,而整个语料多出 $(X-1) * n_{ij}$ 个词。

可以有如下推导:

$$\frac{n_{ij}}{\sum_k n_{ik}} = \frac{X * n_{ij}}{X * (\sum_k n_{ik})} = \frac{X * n_{ij}}{(\sum_k n_{ik}) + (X-1) * (\sum_k n_{ik})}, \tag{5}$$

因为 $\Sigma_k n_{ik} > n_{ij}$,所以可以证明新的 TF 计算方式确实是对传统 TF 做了增强。

但对每个词的 X 值的确定需要用之前整理完成的“两级情感词典”。实际上本文的特征词分成三种:普通词、普通情感词、核心情感词。其中普通词是不需要增强计算的,X 默认为 1;而普通情感词和核心情感词可以取两个递增的数来增强,比如普通情感词的 X 设定为 2,核心情感词 X 设定为 3,最终的 X 值通过实验确定。

2.2.3　特征选取

使用 N-Gram 切分微博,不论成词与否,都需要作为特征计算 tf-idf。这将会导致文档向量空间维度很大,分类器训练时间呈几何级数增长。为了解决这个问题,首先需要对特征集进行特征选择,减少特征维度。

另外,使用 N-Gram 语言模型来选择特征存在一个与语言学上相悖的事实,即语句中的情感特征权重与普通词特征权重一致。本文在训练阶段已经使用“分级情感词”策略,能够保证在训练时情感词得到重视。但是在特征选取阶段,这个悖论仍然存在,容易漏掉重要但出现频率少的情感词对语句情感的贡献。为了尽量避免这一现象的产生,将在特征选取过程中引入一种指标——“情感指数”,用于衡量该特征值对于语句情感倾向性的贡献。

默认使用的选择方法是根据词频来排序,选择词频最高的前 N 个词语留下来进行训练,此处的词频是指整个语料中的词频。但是,这样的方法不考虑任何情感信息,只单纯地选择高频词作为特征,显然不能衡量特征与情感倾向性的相关信息。借鉴特征表示的相关思想,我们可以使用 tf-idf 类似思想,定义该特征在语料中的情感 tf-idf,定义如下:

$$\text{tfidf}_j = \frac{t_j}{\text{idf}_j} = \frac{N_j}{N} * \text{lb}\frac{|\text{DS}|}{1 + |\{\boldsymbol{D}_i | T_i \in \boldsymbol{D}_i\}|}, \tag{6}$$

其中,N_j 表示词 T_j 在整个语料中的出现词数,N 为整个语料的词数目。在加入 tf-idf 特性后,情感词所代表的的情感倾向仍然没有体现出来,需要再考虑情感词相关因素。考虑到本文引入了“情感词分级”策略,可以对 tfidf 乘上一个情感因子 $X_t/X_{\max}$,其中 X_t 是词 T_t 的情感增强因子,$X_{\max}$是所有词中的最大增强因子,公式(6)变为

$$\text{tfidf}'_j = \frac{t_j}{\text{idf}_j} * \frac{X_t}{X_{\max}} = \frac{N_j \cdot X_t}{N \cdot X_{\max}} * \text{lb}\frac{|\text{DS}|}{1 + |\{\boldsymbol{D}_i | T_i \in \boldsymbol{D}_i\}|}。 \tag{7}$$

使用 tfidf'_j 来计算所有特征的权重,最终保留高权重的词当做特征,而保留下的特征数目由实验结果决定。

3　实验及结果分析

3.1　实验数据

训练数据集包括了早年评测所公布的一些微博标注语料,总计 9 500 条,正负样本均衡。同时也包括一些最新采集并标注的微博数据,经过整理,这部分微博数目为 47 784 条,同样正负样本均衡。训练集总体规模为 57 284。

3.2　实验评价指标

准确率(precison)和召回率(recall)是两个用来评判分类问题结果好坏的两种指标,定义如下:

$$\text{precision} = \frac{\text{System. Correct}}{\text{System. Output}}, \tag{8}$$

$$\text{recall} = \frac{\text{System. Correct}}{\text{Human. Labeled}}。 \tag{9}$$

其中,System. Output 是指系统返回的总记录数,System. Correct 是指系统为该类返回的正确结果数,Human. Labeled 是指测试集中该类的总数目。准确率用于衡量分类器准确性,召回率用于衡量分类器是否能找全该类样本,这两个指标应该同时兼顾,不可偏废。使用 $F1$ 测度来均衡两方面,定义如下:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}。 \tag{10}$$

依据 COAE2014 的评价指标,实验中采用宏观 $F1$ 测度(macro $F1$)衡量实验效果。宏观评价值是指分别算出正类与负类评价值后平均所得。比如,宏观准确率是指正类准确率与负类准确率平均之后的结果,而使用宏观准确率和宏观召回率,通过公式(10)的计算可以得出宏观 $F1$ 测度。

3.3 实验准备

本文在实验准备部分,针对确定特征选择部分需要保留的特征数目做了实验1。实验1中使用 N-Gram 获取特征,N 取[1,5]范围内的整数,使用的机器学习算法为支持向量机(support vector machine, SVM),分别使用 500~5 000 范围内的特征数目,得到如图1所示变化趋势。

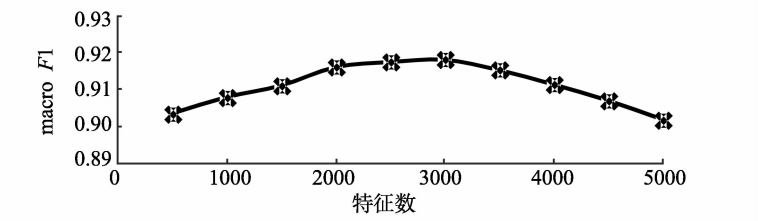

图1 特征选择数据走势

Fig. 1 Data trend of feature selection

从图1可以看出,特征数较小时,情感信息不够充沛,分类效果不好,而特征数增大到一定程度时,冗余信息增多造成识别误差,弱化了分类效果,所以 $F1$ 测度随着特征数的增加呈现先增长后降低的趋势,在选择3 000时将得到比较高的 $F1$ 测度值。

另外本文对公式(4)中的 X 值选取做了实验2和实验3,分别用以确定两级情感词典中核心情感词的增强系数 X_1 和普通情感词的增强系数 X_2,一般 $X_1 > X_2$。在实验中使用 N-Gram 获取特征,N 取[1,5]范围内的整数,使用的机器学习算法为 SVM,特征数使用实验1中确定的3 000。实验2中,X_2 取固定值1.5,对 X_1 在[2,20]之间进行调参优化,实验结果如表2所示。实验3中,将 X_1 取固定值4,对 X_2 在[1,4]之间进行参数优化,实验结果如表3所示。

表2 参数 X_1 的调优($X_2 = 1.5$)

Table 2 Tuning parameter X_1 ($X_2 = 1.5$)

X_1 取值	2	3	4	5	6	7	8	9	10	20
macro F1	0.814 92	0.815 37	0.816 14	0.812 59	0.811 70	0.810 71	0.810 03	0.809 81	0.809 59	0.803 93

表3 参数 X_2 的调优($X_1 = 4$)

Table 3 Tuning parameter X_2 ($X_1 = 4$)

X_2 取值	1	1.5	2	2.5	3	3.5	4
macro F1	0.812 92	0.813 14	0.816 14	0.812 01	0.811 70	0.807 62	0.803 47

根据表2和表3可以得出,在 $X_1 = 4$、$X_2 = 2$ 时,本文算法将取得比较好的 $F1$ 值。

3.4 实验结果

本文使用朴素贝叶斯(Naïve Bayes, NB)分类、SVM 分类作为实验的对照组,它们均采用情感词作为特征。衡量指标使用准确率、召回率和 $F1$ 测度,特征数均为3 000,情感词分级增强参数 $X_1 = 4$,$X_2 = 2$。训练集规模为两种:9 500 条微博平衡语料库 $S1$ 和60 000条微博平衡语料库 $S2$,用以说明不同数据集规模对情感分类效果的影响。三种方法在两种训练集规模下情感分类结果如表4所示。

表4 训练集 $S1$、$S2$ 下的三种情感分类方法 $F1$ 测度值

Table 4 $F1$ value for three kinds of classification methods on $S1$ and S_2 training sets

分类方法	$S1$			$S2$		
	准确率	召回率	$F1$ 测度	准确率	召回率	$F1$ 测度
Naive Bayes	0.715	0.721	0.718	0.859	0.858	0.859
SVM	0.786	0.739	0.762	0.907	0.903	0.905
本文方法	0.828	0.815	0.822	0.925	0.922	0.924

本文方法以 ICT_WDSE 为名参加了“COAE2014 任务三”评测,评测的结果如表5所示。

为了防止作弊,评测要求参赛者从 40 000 篇微博中找出自己算法认定最有可能的 10 000 篇,并且评测只对其中的 7 000 篇微博做了标注,所以准确率和召回率分母很大,结果数值偏低。综上所述,表 5 中的指标值只能用于同一评测中的同行比较,而在与其他算法比较时,要说明本算法的真实效果应该参考表 4。

表 5　COAE2014 评测结果
Table 5　The Result of COAE2014

系统	ICT_WDSE	Best	Medians
Pos_F1	0.347	0.715	0.445
Neg_F1	0.778	0.778	0.428
Macro_P	0.877	0.962	0.877
Macro_R	0.489	0.543	0.309
Macro_F1	0.562	0.677	0.443
Micro_P	0.826	0.962	0.857
Micro_R	0.467	0.547	0.305
Micro_F1	0.597	0.681	0.450

3.5　结果分析

从实验结果上来看,本文提出的方法很有效。无论是在小规模还是大规模情感语料上,其准确率和召回率较 NB 和 SVM 均有明显提高,原因如下:

- 引入情感词分级增强策略。对不同情感词划分两级,并对不同级别的情感词作了不同程度的增强,这样处理可以区分出不同情感词对情感分类的不同作用;另外因为情感词的增强,使得短文本中出现次数很少的情感词对整句情感贡献的权重人为得到提升,这些都对短文本情感分类起到很好的促进效果。
- 使用 N-Gram 语言模型来找特征。传统情感分类中,使用分词并查找情感词典方法找到情感词,然后使用这些情感词充当特征,在这种情况下,情感分类结果将受限于分词准确性,也容易遗漏情感新词。本文使用 N-Gram 语言模型,将不会遗漏任何其他情感信息,而这些信息在短文本情感倾向性分析中相当重要,比如未登录情感词和否定倾向词等。

另外从表 4 可以看出:本文方法在训练数据较少的情况下表现更好。在小数据集下,本文方法能够更好地抓住仅有的情感词判断出微博情感倾向性。而在大数据集下,对效果的提高没有小规模情况下明显,其原因仍然是因为本文方法中的情感词分级增强策略不同程度地提升了核心情感词和普通情感词在句中的权重,导致小规模语料集中,情感信息更容易凸显并促进情感分类。但在大规模语料集下,情感信息充沛,增强策略效果已经不明显,意义不大。所以本文方法在小规模语料集下对情感信息更为敏感,分类效果更好。

通过表 5 的评测结果可以看出本文方法在负类样本识别上效果优于正类样本。分析其原因可能有以下几点:

- 情感词典规模问题。负类情感词典规模大,收集齐全,而正类情感词典规模较小。
- 微博训练语料库。微博语料因其媒体属性特点,吐槽、约架、批评等负面博文较多,而且博主为了表达负面情绪往往使用很多情感词,虽然训练集是平衡语料,但出现了负类样本中情感信息充沛,正类样本中情感信息不明显的现象,这对正类样本识别有很大的影响。
- 否定词处理效果不佳。中文微博存在很多否定现象:否定、反讽、双重否定等,而否定现象的存在对语句情感倾向性将起到颠覆作用。本文方法在这一点上效果不佳,导致正负类样本识别出现很大差别。

因为在负类样本上的出色表现,本文方法在 COAE2014 收到的 50 个提交结果中排名第 7 位。

4　结语与展望

可以预见在不久的将来,网络舆情疏导、产品评价反馈、消费者购物指导等方面都会使用短文本情感倾向性分析技术。而本文在情感分类的过程中提出了一些改进,提高了分类效果,并在实际应用评测中也取得了良好的成绩。但该方法存在着正负样本分类效果差别明显的问题,影响了总体效果。原因可能是正向语料情感倾向性不充分,没有为分类器提供足够的分类指导信息。但也正因为此特性,可以将此方法用于舆情监控和评价跟踪等特别关注负面信息的场合。接下来的工作将主要集中在提高正面样本的识别率上,争取将总体效果优化提升,并尝试一些其他特征以及多特征融合的方法,以进一步提高系统性能。

参考文献:

[1] DAVIDOV D, TSUR O, RAPPOPORT A. Enhanced sentiment learning using Twitter hashtags and smileys[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing: Tsinghua University Press, 2010:241-249.

(下转第 13 页)

Annual Meeting of the Association for Computational Linguistics. Somerset: ACL, 2011: 151-160.

[12] DAVIDOV D, TSUR O, RAPPOPORT A. Enhanced sentiment learning using Twitter hashtags and smileys[C]// Proceedings of the 23rd International Conference on Computational Linguistics. Beijing: Tsinghua University Press, 2010: 241-249.

[13] KOULOUMPIS E, WILSON T, MOORE J. Twitter sentiment analysis: the good the bad and the omg! [C]// Proceedigns of ICWSM. [S. l.]: AAAI Press, 2011, 11: 538-541.

[14] 张珊,于留宝,胡长军. 基于表情图片与情感词的中文微博情感分析[J]. 计算机科学,2012,39(3):146-148.
ZHANG Shan, YU Liubao, HU Changjun. Sentiment analysis of Chinese micro-blogs based on emoticons and emotional words[J]. Computer Science, 2012, 39(3):146-148.

[15] 吴迪. 汉语微博文本特征研究[D]. 长春:吉林大学,2012.
WU Di. Research on text features of micro-blogging in Chinese[D]. Changchun: Jilin University, 2012.

[16] 丁建立,慈祥,黄剑雄. 网络评论倾向性分析[J]. 计算机应用,2010(11):2937-2940.
DING Jianli, CI Xiang, HUANG Jianxiong. Orientation analysis of Web review[J]. Journal of Computer Applications, 2010(11):2937-2940.

[17] 唐慧丰,谭松波,程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报,2007,21(6):88-94.
TANG Huifeng, TAN Songbo, CHENG Xueqi. Research on sentiment classification of chinese reviews based on supervised machine learning techniques[J]. Journal of Chinese Information Processing, 2007, 21(6):88-94.

(编辑:许力琴)

(上接第 7 页)

[2] BARBOSA L, FENG J. Robust sentiment detection on Twitter from biased and noisy data[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing:Tsinghua University Press, 2010:36-44.

[3] HASSAN A, QAZVINIAN V, RADEV D. What's with the attitude? Identifying sentences with attitude in online discussions [C]//Proceedings of 2010 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010:1245-1255.

[4] MEENA A, PRABHAKAR T, AMATI G, et al. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis[C]//Advances in Information Retrieval. Heidelberg: Springer Berlin, 2007:573-580.

[5] TURNEY P D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of 40th Annual Meeting of the Association for Computational Linguistics. Somerset: ACL, 2002:417-424.

[6] SOCHER R, PENNINGTON J, HUANG E H, et al. Semi-supervised recursive auto-encoders for predicting sentiment distributions[C]//Proceedings of 2011 Conference on Empirical Methods in Natural Language Processing. [S. l.]:[s. n.], 2011: 151-161.

[7] 王根,赵军. 基于多重标记 CRF 句子情感分析的研究[C]//内容计算的研究与应用前沿:第九届全国计算语言学学术会议论文集. 北京:清华大学出版社, 2007:609-614.
WANG Gen, ZHAO Jun. Sentence sentiment analysis based on multi-redundant-labeled CRFs[C]//Advanced Research and Application of Frontier Content Computing: the 9th China National Conference on Computational Linguistics. Beijing: Tsinghua University Press, 2007:609-614.

[8] TAN Chenhao, LEE Lilian, TANG Jie, et al. User-level sentiment analysis incorporating social networks[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2011:1397-1405.

[9] LI Shoushan, WANG Zhongqing, ZHOU Guodong, et al. Semi-supervised learning for imbalanced sentiment classification [C]//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. [s. l.]: AAAI Press, 2011:1826-1831.

[10] Iadh Ounis, Craig Macdonald, Ian Soboroff. Overview of the TREC 2010 BlogTrack[C]//Proceedings of the 19th Text REtrieval Conference Proceedings (TREC 2010). NIST, 2010.

[11] SEKI Y, KU L-W, SUN L, et al. Overview of multilingual opinion analysis task at NTCIR-8: a step toward cross lingual opinion analysis[C]//Proceedings of the 8th NTCIR Workshop. [S. l.]:[s. n.], 2010:209-220.

(编辑:许力琴)