软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(12):2790−2807 [doi: 10.13328/j.cnki.jos.004728]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

轻型评论的情感分析研究*

张 林[1,2], 钱冠群[3], 樊卫国[4], 华 琨[5], 张 莉[1]

[1](北京航空航天大学 计算机学院,北京 100191)

[2](浙江财经大学 信息学院,浙江 杭州 310018)

[3](百度公司,北京 100085)

[4](Department of Information Systems, Pamplin College of Business, Virginia Technological University, USA)

[5](Electrical and Computer Engineering Department, Lawrence Technological University, USA)

通讯作者: 张林, E-mail: zhanglin_hz@163.com

摘 要: 以在智能移动设备上发表的用户评论作为研究对象,并将该类评论称为轻型评论.指出了轻型评论与早期互联网评论及短文本研究的异同点,并通过实验总结轻型评论的独有特性:字数少、跨度大,短小评论数量众多,评论长度与数量满足幂率分布.同时,针对轻型评论的情感分类研究展开了一系列的实验研究,发现:(1) 情感分类效果随着评论长度的增加而下降;(2) 传统的特征筛选方法以及特征加权方法对于轻型评论效果都不够理想;(3) 极性词在短评论中比例高于长评论;(4) 长、短评论在用词上存在较高的重叠度.在此基础上,提出了一种基于短评论特征共现的特征筛选方法,将短小评论中的优势信息和传统的特征筛选方法相结合,在筛选掉无用噪音的同时增补有利于分类的有效特征.实验结果表明,该方法可以有效地提高轻型评论中较长评论的分类效果.

关键词: 情感分析;用户评论;短文本;意见挖掘

中图法分类号: TP181

中文引用格式: 张林,钱冠群,樊卫国,华琨,张莉.轻型评论的情感分析研究.软件学报,2014,25(12):2790−2807. http://www.jos.org.cn/1000-9825/4728.htm

英文引用格式: Zhang L, Qian GQ, Fan WG, Hua K, Zhang L. Sentiment analysis based on light reviews. Ruan Jian Xue Bao/ Journal of Software, 2014,25(12):2790−2807 (in Chinese). http://www.jos.org.cn/1000-9825/4728.htm

Sentiment Analysis Based on Light Reviews

ZHANG Lin[1,2], QIAN Guan-Qun[3], FAN Wei-Guo[4], HUA Kun[5], ZHANG Li[1]

[1](School of Computer Science and Engineering, BeiHang University, Beijing 100191, China)

[2](Department of Information Technology, Zhejiang University of Finance & Economic, Hangzhou 310018, China)

[3](Baidu Inc., Beijing 100085, China)

[4](Department of Information Systems, Pamplin College of Business, Virginia Technological University, USA)

[5](Electrical and Computer Engineering Department, Lawrence Technological University, USA)

Corresponding author: ZHANG Lin, E-mail: zhanglin_hz@163.com

Abstract: This paper researches the newly emerging user reviews (referred here as “light reviews”) generated from smart mobile devices. The similarities and differences between this research and the early studies are pointed out. The unique characteristics of the light review can be summarized as having shorter texts, bigger span, and in most cases fewer words per review. The review length and scale also meet the power-law distribution. A series of experiments are studies based on light reviews, resulting in some interesting findings: (1) There is an inverse relationship between classification accuracy and review length; (2) The traditional classical feature selection and feature weight method do not perform well enough on light reviews; (3) The polar word ratio in short reviews, which is the most important feature in sentiment analysis, is higher than in long reviews; (4) There is a higher shared feature term proportion between short

* 收稿时间: 2014-05-05; 修改时间: 2014-08-21; 定稿时间: 2014-09-30

review and long review. Based on above studies, the paper puts forward a feature selection method based on short text co-occurrence feature. By combining the information advantages in short reviews with the traditional feature selection methods, the presented method preserves useful information and details as much as possible while removing noise. The results of experiment show that the method is effective and the classification rate is higher.

Key words: sentiment analysis; user review; short-text; opinion mining

基于文本的意见挖掘是近 10 年来备受关注的研究领域,其目的是研究如何从文本中获得用户对各种事物的意见倾向,从而了解互联网上用户的普遍观点和意见[1−3].其中,基于评论的情感分析是意见挖掘的一个重要分支和研究方向,主要研究如何采用自动化的方法,从用户的评论中获取用户对事物的褒贬倾向.从本质上讲,判断褒贬倾向是一个分类问题,因此也称为情感分类、极性分类.前期的研究中,有较多针对互联网用户评论的情感分类研究,如 Turney[3],Dave[2],Fan[4]等人的研究.

随着智能手机、平板电脑、无线传感器网络等各种普适系统的快速发展,人类已经生活在由通信网、互联网、传感网相互融合的混合网络环境.同时,各类实用的 App 软件已经日益盛行.有数据显示:谷歌的 Andriod market 在 2012 年就拥有 40 万款 App 应用[5];而苹果的 Appstore 在 2013 年 6 月达到了 90 万款 App 应用,下载次数更是达到了 150 亿次以上[6].与此同时,人们已经习惯在日常生活中随时随地用手机或平板电脑发表意见和观点.在 Appstore 上各类 App 页面下,成千上万的用户在吐槽、发表评论.这些评论不同于发表在互联网上的评论,它更加短小精干,很多时候仅用一个字、词甚至一个标点符号来表明自己的态度.相比早期在互联网上那些文字较多、具有一定组织结构、较为正式的评论,这种评论可以称为轻型评论.

随着更多移动智能应用的出现,轻型评论将成为一种新的评论形态.因此,情感分析研究应该将研究的广度扩展到这类轻型评论上,重新审视传统的方法和流程,讨论该类评论在情感分析上独有的特性.目前,针对这类短小简洁而数量众多的轻型评论的研究还很少.根据我们的统计分析,这类评论具有以下几个较为鲜明的特点:

(1) 平均字数少,不仅远远小于早期在 PC 用户评论的平均字数,也小于大多短文本的平均字数;

(2) 字符数量跨度大.最短的评论只有一个字或一个标点符号,最长的却有上千字,这种长短严重不齐的字符跨度为后期情感分析带来一定的难度;

(3) 最为重要的是,评论的文本长度与数量之间呈现出幂律关系,短小评论占了整体评论中的大多数,长评论数量相对较少.

这些特点都与前期文献研究涉及到的互联网用户评论的特点有很大的不同.互联网用户评论不仅平均字数较多,且短小文本的数量相对较少.此外,很多文献显示:这些过于短小且数量不多的评论在文本预处理阶段就会被清洗及过滤掉[5,8,9],而轻型评论中的主体正是极其短小且数量众多的短小评论,这种评论在早期研究中正是被忽略和过滤掉的那部分.因此,本文所针对的轻型评论不同于早期情感分析的研究对象,呈现出不同于其他评论的特性.

本文研究的轻型评论来自苹果 Appstore 发布的基于各类 App 的用户评论.选择这类评论为研究对象,是由于 App 软件用户评论数量巨大,符合社会计算研究中对海量、真实数据的处理要求.同时,此类用户评论具有一定的代表性,它非常类似社会网络中的随笔、吐槽,具有典型的字数少、数量大等特点,可以作为社会化计算、扩展情感分析广度的一个典型样本.此外,对这类用户评论进行情感分析,可以帮助我们了解 App 用户的情感和意见,辅助指导 App 的进一步改进及演化.这对于软件供应商、App 开发商以及用户都很重要,同时也可作为软件评测的一个重要指标.

本文对轻型评论在情感分析上进行了一系列的实验研究,发现轻型评论中的一些特性:

(1) 评论长度对情感分类效果有较大的影响;评论长度与分类效果成反比的关系;短小评论分类效果更好;这与传统的短文本由于特征向量稀疏,而分类效果不好[7]的结论有所不同;

(2) 对于文本分类有显著作用的特征筛选方法并不完全适用于轻型评论,对轻型评论中的短评论而言,特征筛选方法并没有提高分类效果,反而有所降低;

(3) 传统的特征加权方法在各评论分组中表现各异,不够稳定;

(4) 短评论中极性词的比例明显高于长评论中极性词的分布比例,而长评论中的极性词含量会逐步趋于稳定;

(5) 长评论和短评论中的特征项重叠率较高,即长短评论中有较多的特征词是重复出现的.

在上述实验结果的基础上,我们提出了基于短评论共现的特征筛选法(简称 SCO),以提高轻型评论中长评论的分类效果.该方法利用短评论富含极性词以及与长评论特征重叠率高的特性,在传统特征筛选的基础上,将短评论中的特征引入到长评论中,既降低了噪音又补充了优势信息.实验证明,这种方法取得了较好的效果.

本文第 1 节综述相关的工作.第 2 节给出轻型评论的定义及统计特性.第 3 节是轻型评论在情感分析方面的一些讨论.第 4 节是基于短评论共现的特征筛选方法的介绍,包括方法思路、定义.第 5 节展示该方法的验证.第 6 节是总结和将来的工作.

1 相关研究

1.1 情感分析的相关研究

情感分析是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的一种技术,当前研究使用的方法主要分为两种:基于机器学习的方法和基于语义词典的方法:基于机器学习的方法是利用机器学习的各种分类方法来识别情感;而基于语义词典的方法则先构造情感词词典或是列表,借助该字典判断情感的倾向.两种方法各有优势:

- 基于机器学习的情感分类方法可以跨领域应用且稳定性较好,对于数据量较小、标注完整的数据集具有较好的效果[8].如:Pang 和 Li 等人采用不同的特征选择方法,应用了 NB,ME,SVM 等不同的分类方法对电影评论进行分类,取得了较好效果[9];Ni 等人利用 CHI 方法和信息增益方法(IG)进行特征选择,并采用 NB,SVM 和 Rocchio 算法对情感分类[10];Mullen 等人和 Whitelaw 等人用 SVM 做分类算法,但他们在特征的选择和处理上不同[11,12];Cui 等人比较了 PA(passive-aggressive),LM(language modeling)和 Winnow 分类器的性能,发现 PA>Winnow>LM[13].这些方法需要大量繁琐的人工标注工作,同时,分类方法由于与数据集较相关,不同类型的数据需要采用不同的分类方法才能有较好的效果,很多研究通过实验对比找到针对某类数据较好的分类方法.
- 基于语义词典方法的优点是能够更好地利用人类的知识,减少机器学习的盲目性,同时也能大幅度地减少人工标注样本的工作,但如何自动或半自动地构造极性词典或极性词列表是一个关键问题.这方面大部分的研究是通过人为观察语言中的一些现象,诸如同义词或者词语之间搭配,挖掘出众多极性词及极性词的倾向.例如:Wiebe 等人利用一种相似度分布的词聚类方法,标记了形容词极性词及极性[14];Riloff 等人利用手工制定的模板并选取种子评价词语,使用迭代的方法获取名词极性词及极性[15];Turney 和 Littma 提出了点互信息(point mutual information)的方法判别某个词语是否是极性词语[16].此外,不少研究人员利用现有的通用词典 WordNet 或 HowNet,将手工采集的种子评价词语进行扩展来获取大量的极性词语及极性[17,18].事实上,由于情感的表达与领域息息相关,不同领域的极性表达方式各不相同,甚至截然相反,需要依据更多的上下文表达才能了解整个句子或者段落的极性倾向,这就需要词法、句法等更深层次的分析.如文献[19]使用依存句法结构(如 ADV,ATT 以及 DE 结构),即,利用极性词之间的句法关系来获取句子的极性.由于自然语言的复杂性,建立起一个完备的极性词词典几乎是不可能的任务[20].当然,也有一些研究是将这两种方法相结合,但整体上分析效果并没有明显的提升[21,22].

中文的基于情感分析研究还不是很多,大多数处于借用英文的相关技术.叶强等人在 PMI-IR 基础上探讨了中文的情感分类方法[23,24];朱嫣岚基于 HowNet 提出了两种词语语义倾向性计算的方法[18];上海交通大学[25]则开发了一个用于汉语汽车论坛的意见挖掘系统,并给出了统计及可视化结果.

1.2 用户评论的相关研究

情感分析的研究语料很大一部分都来自于用户评论,早期针对用户评论的研究语料大多来自于互联网[26],其文本长度较长,与本文研究的轻型评论有较为明显的差异.我们列举一些目前能够搜集到的、具有明确语料长度的研究案例,以说明在用户评论研究中,类似轻型评论语料的研究尚不多见.

Pang 等人[9]在 2002 年~2004 年的研究中,采用的是发表在 IMDB[27]上的电影评论,据我们统计,该语料集合中,每个评论平均长度为 3500+字符,平均包含 700+的单词.

Fan 等人[4]对用户发表在汽车论坛上的反馈意见进行了研究,该研究语料来自 2010 年美国交通部、国家高速安全部门等数据库,在这些研究语料中,他们主要研究超过 50 个单词以上的用户评论,他们的研究语料平均包含 502 个单词,最小 50 个单词,最大 8 586 个单词.

Cui 和 Mittal 的研究中[13]收集了很多针对电子产品的评论,如照相机、笔记本电脑、PDA、MP3 等,整个语料库大约 0.4G,包括了 32W 的用户评论,而评论的平均长度是 875 字节.

Kasthuriarachchy 的研究[28]中比较了各种不同领域的用户评论,诸如电影、DVD、手机甚至 tweets,在他的研究语料中,每个评论至少包含 1 个句子,每个句子的平均单词数是 17 个~22.2 个.根据论文中的统计,其中电影评论语料平均包含 35.8 个句子.

在中文情感分析的语料研究中,ChnSentiCorp 语料库[29]由谭松波博士提供.该语料包含酒店、电脑与书籍这 3 个领域的相关评价,每种类型的语料来源单一,但不同类型间的来源差异性较大.根据杨震的研究[33]发现:在文本清洗、分词和预处理之后,得到的文本有效属性为 29 550 个,单一文本平均由 76.35 个词组成.在去除无效字符后,平均文本长度为 70.4114 个词.此外,根据刘鲁的研究[43]可知,中文微博的平均长度为 45 个中文字符.

由于书写的局限,来自手机的评论与大部分研究文献中的评论有非常明显的不同.手机评论的长度要远远小于 PC 评论的长度.在上述研究中的语料大部分经过了筛选,如 Fan[4]的研究中只保留超过 50 个单词的评论进行研究.而手机评论中平均只有 17 个字,70%的评论都不超过 30 个字(详细信息参见第 2 节),因此,我们认为非常有必要对这类轻型评论进行专门的研究,重新审视早期的研究方法和研究流程.

1.3 短文本的相关研究

短文本是指那些文本长度小于 160 个字符的文本[30],一般以微博、手机短信、网页评论以及聊天等形式存在(短文本的概念是相对于长文本而言).研究学者普遍认为,短文本由于特征向量的维度过少,在整体特征矩阵中不可避免地出现极度稀疏的问题,即:每个短文本样本中,只有极少数的维数上是有取值的.由此,给短文本的处理带来了极大的不确定性和困难.解决这种稀疏性主要从两个方面入手:一是通过降低整体的特征向量维度来避免稀疏性问题;二是通过各种方法扩展短文本的信息,从而提高短文本自身的向量维度.

一方面,研究者通过特征选择方法来降低维度,对高维数据进行有效降维的典型方法有信息增益方法(information gain)[31]、CHI 方法(统计量)[32]以及互信息(MI)[33]方法,其他的方法有潜在语义索引(latent semantic indexing,简称 LSI)方法[34,35]、基于聚类重心数据降维(centroidmethod,简称 CM)[36]的方法等.这些降维的方法或者需要计算大矩阵的特征值和特征向量,或者需要对数据进行频繁的聚类迭代分析,其计算复杂度和计算时间都比较大.

另一方面,研究者希望通过属性联想或组合来扩充短文本特征矩阵,这些研究大多集中在对主题的检测与分类上.如:Wang 等人[37]利用 WR-kmeans 聚类方法综合相关手机短消息解决相似短文本发现问题;Fan 等人[38]利用特征扩展和控制模型,有效提高短文本的分类精度;Adams 等人[39]利用 WordNet 解决即时聊天信息话题检测与抽取问题.

文献[7]研究了中文短文本评论的情感分类,该文献认为:短文本存在着数据稀疏性及上下文缺失的情况,需要用某种方法来补充和扩展信息.但该文献中情感分类不仅包括了对情感的极性分类,还包括了对领域主题的划分,文献中对短文本的缺失补充主要体现在对于领域主题的补充和扩展上,而不是情感的极性分类.

我们研究针对的是轻型评论,它与短文本的概念之间有很大的相似之处,都是较为短小的文本信息.但是,

我们的研究重点与短文本的研究重点是不一样的,短文本的相关研究更注重补充缺失的信息,根据我们的实验研究,短文本缺失的信息并不体现在情感极性分类上,而是体现在评论主题的识别和抽取.由于不同研究学者对情感分析概念划分的不同,因此,尽管一些文献也在研究短文本的情感分类问题[33],但该文献里的情感分类还包括了主题的识别和主题的抽取,并不是单纯的情感极性分类问题.

根据本文研究,短评论尽管信息含量少,但其情感信息并不少.如“很好!”、“坑爹啊”等语句,其中就包含了强烈倾向性的情感信息(垃圾及无用的评论不计),短评论缺失的信息仅仅是产品的主题或者是产品的特征信息.实验结果显示(详见第 3.1 节):越是文本短小的评论,其情感分类的效果越好;而文本长度越长,其携带的信息中,不利于情感分类的噪音信息也会加大,分类效果远远不如短小文本评论.早期对于短文本的研究大多集中在主题的分类上,而对于情感极性分析,几乎没有人专门针对短小评论进行研究.而手机评论不仅短小,而且短小评论的数量极其众多,但同时包含丰富的极性词,在情感分类上表现更好.轻型评论与短文本前期研究是有非常大区别的.

综上所述,本文的研究对象是文本长度较短,但数量极其众多的轻型评论,它不同于早期互联网上的用户评论,因为它的文本长度远远小于互联网用户评论;同时,它也不同于短文本的主题分类研究,不存在情感信息缺失的问题.目前,这类轻型评论的情感分析研究尚未见到,本文扩展了情感分析的研究范围.

2 轻型评论的定义及特点

2.1 轻型评论的定义

由于书写方式的局限,人们发表在手机或者其他移动设备上的用户评论通常呈现出更加短小精干的特点,其平均字数只有不到 20 个汉字;此外,手机评论的长度跨度大,短到只有 1 个字甚至一个标点,长的可以达到数千甚至上万;其长度和规模符合幂律分布,即非常多的短评论和非常少的长评论交织在一起.我们将符合这几个特点的评论称为轻型评论.

本文中的轻型评论与短文本概念有点类似,但却有明显的不同:

- 短文本是针对长文本提出的概念,虽然也指那些长度较小的文本,但更主要是指人们谈论的主体或者某些领域信息过于简短、有所省略,因此,短文本的主要研究大部分集中在如何补充这些重要信息,提高主题分类的准确性.在短文本研究中,“短”是不足,是需要通过各种方式进行弥补的.此外,短文本只是泛指那些较为短小的文本,一般不超过 160 个字符;
- 而轻型评论中,不仅有更加短小的文本也有上千字符的长文本;且长短文本的规模分布极端不均衡;对于轻型评论的情感分类,“短”是优点,因为短评论中包含了更多的情感极性词,噪音少指向更加明确,充分利用规模巨大的“短”评论更有利于整体分类效果的提升.

2.2 轻型评论的特点

本文通过真实数据的一些统计实验来论证和说明轻型评论的具体特点.本文采集了 2011-01-21~2013-06-06 期间 Appstore 上各类 App 用户评论,共计 145 263 条.每条 App 评论具体包含标题、正文、评分等信息.其中,评分为 1 分~5 分,1 分最低,5 分最高.我们对原始评论进行了过滤和筛选,其中被过滤掉的评论有:① 标题和正文不含任何文字的评论;② 用户打分为 3 的评论,因人为都很难判断其正负情感倾向.剩下共计 133 575 条评论.用户打分在 1 分~2 分的评论作为负例,打分在 4 分~5 分的评论作为正例.

通过一些统计及实验,我们发现,轻型评论具有以下自身独有的特点:

2.2.1 字数少,字符分布跨度大

样本数据的统计结果见表 1,可以看出,样本评论包括标点符号的平均字数只有 17 个中文字符.这不仅远远小于文献中有记载的研究语料长度,也远远小于微博 40 个字的平均长度.同时也可以看出:评论的字符最小的只有 1 个字(空评论已经过滤掉),最大到 6 336 个字,可见其字符分布跨度大.

Table 1 Sample data statistics result

表 1 样本数据统计结果

		正例	反例	总体
字符数	最小~最大	1~1891	1~6336	**1~6336**
	平均	15.75	34.11	**17.08**
	标准差	21.850 47	65.960 21	34.396

2.2.2 评论长度与数量之间满足幂律分布

通过观察发现:在轻型评论中,评论长度与数量存在着极端不均衡的现象.即:有非常多的评论字数很少,而字数多的评论其数量又很少,平均长度已经不能代表轻型评论的分布特性.其文本长度与数量之间满足幂率分布,如图 1 所示.其中,x 表示文本长度,y 表示长度为 x 的评论数量.从图中可以看出:在双对数坐标系下,这是一个典型的幂律分布关系,满足 $y=kx^{\alpha}$,其中,幂律指数 $\alpha=-4.177$.

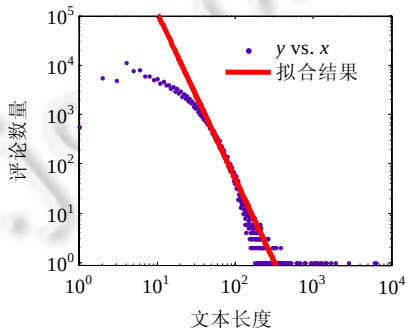

Fig.1 Power-Law relation of the length of reviews and the scale of reviews

图 1 评论长度与评论数量幂率关系图

3 轻型评论在情感分类的特性

为了考察文本长度、特征筛选等对轻型评论情感分类的影响,我们进行了一系列实验.实验基于 WEKA 工具[40],采用 10-fold 交叉进行测试和验证.中文分词采用 mmseg4j[41]工具.本论文中的分类算法均采用 Naive Bayes Multinomial 方法(据前期实验研究,采用 SVM 方法实验效果类似).

3.1 评论长度与情感分类效果之间的关系

为了考察评论长度是否影响情感的分类效果,我们采用了两种方式进行了实验:

(1) 按照字符长度进行分组,考察其对分类的影响.

这里,特征向量的赋值采用了 TF 方法.实验结果见表 2,可以看到,*F*-Value 随着评论字符的增长而下降.它说明评论长度越长,情感分类越不够准确.字数在 300 以上的评论,分类的准确性达到最低,只有 0.4 左右.

(2) 按照样本数量及字符长度进行分组,考察对分类的影响

为避免样本数量影响分类效果,我们在设计实验时,按照实验样本的数量重新进行了分组,使得每一组参与的样本数量保持在同一个级别上.表 3 为分组情况及实验结果,可以看出:在样本数量都较为均衡的条件下,不同长度的评论其分类依然存在评论越长,情感分类准确性越低的结论.

Table 2 Classification result comparison based on length group

表 2 基于不同字符长度的分组分类效果

分组(评论长度)	Precision	Recall	*F*-Value
1+	0.923	0.911	0.915
5+	0.909	0.896	0.9
10+	0.881	0.868	0.872
15+	0.857	0.846	0.849
20+	0.844	0.836	0.837
25+	0.825	0.821	0.822
30+	0.827	0.824	0.814
35+	0.808	0.807	0.799
40+	0.809	0.809	0.794
50+	0.796	0.795	0.785
60+	0.791	0.794	0.773
70+	0.795	0.795	0.765
90+	0.753	0.729	0.755
100+	0.757	0.735	0.735
150+	0.639	0.738	0.678
200+	0.649	0.637	0.642
250+	0.539	0.543	0.543
300+	0.333	0.5	0.4

Table 3 Classification result comparison based on balanced sample set

表 3 基于均衡样本数的分组分类效果

分组(评论长度)	评论数量	Precision	Recall	*F*-Value
1# (1~5)	31 978	0.981	0.981	0.975
2# (6~10)	32 086	0.955	0.955	0.955
3# (11~20)	33 316	0.900	0.893	0.893
4# (21+)	36 125	0.815	0.811	0.812

从上述两组实验看出,较短评论的情感分类效果要好于较长评论的分类效果.

3.2 特征筛选方法对轻型评论的作用

在文本分类的流程中,特征筛选方法是一个非常重要的步骤,它可以有效地减少特征维度,降低噪音,提高分类的效果.那么,特征筛选方法对于轻型评论是否还存在类似的作用呢?

这里,我们选用经典特征筛选方法——信息增益方法(简称 IG)以及 CHI 方法(即χ^2统计法)进行了实验,两种特征筛选方法得到的实验结果非常相似,见表 4.

Table 4 Feature selection classification effect contrast in different length group

表 4 特征筛选在不同长度分组中分类效果

分组 (评论长度)	*F*-Value(不用特征筛选)	*F*-Value(采用特征筛选(IG))	*F*-Value(采用特征筛选方法(CHI))
1+	0.915	0.900↓	0.899↓
10+	0.872	0.872	0.872
20+	0.837	0.821↓	0.821↓
30+	0.814	0.810↓	0.811↓
50+	0.785	0.781↓	0.780↓
70+	0.765	0.761↓	0.762↓
90+	0.755	0.755	0.754↓
100+	0.735	0.739↑	0.738↑
150+	0.678	0.688↑	0.687↑
200+	0.642	0.640↓	0.643↑
250+	0.543	0.546↑	0.545↑
300+	0.4	0.560↑	0.562↑

实验结果显示:采用两种特征筛选方法之后,分类效果并没有显著的提高,相反,在大多数短评论分组内还有所下降;只有评论字数超过 100 时,即在评论字数较长时,特征筛选方法才使得分类效果有提高,但幅度有限.

同时我们注意到:在特征筛选后,特征向量大幅度减少,特征矩阵变得越来越稀疏.图 2 显示了在使用 IG 特征筛选前后特征数量的对比.非常多样本的特征被筛选为空,这违背了特征筛选的初衷,给机器学习带来了更多的不确定性,分类效果很难预测.因此,我们建议对于较短评论可以不进行特征筛选的工作.因为筛选后的分类效果不够理想,筛选后特征数量减少过多.通过对筛选词的分析发现:筛选掉的一些看似噪音的特征词,很可能是网络中最常见的错字、别字或者网络新词,像“神马”、“滚粗去”、“肿么办”,这些特征词的保留,有利于提高整体的分类效果.

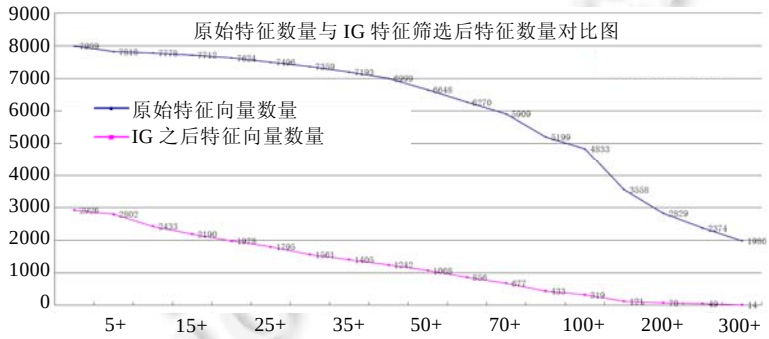

Fig.2 Feature decline tendency after feature selection (IG method)

图 2 原始特征数量与 IG 筛选后特征数量对比图

3.3 特征加权方法对轻型评论的作用

在机器学习过程中,通常要给每个特征项赋予一定的权重,以衡量某个特征项在文档表示中的重要程度或者区分能力的强弱.权重计算方法有很多,如:TF 方法直接根据词语的频率赋予权重,出现频率较高的词汇会得到更高的权重;IDF 方法则给予区分度较高但频率出现较低的特征项更高的权重,以降低那些出现频率高但又不重要的常用词权重;TF-IDF 方法则充分考虑了两者的结合,给予了那些出现频率高并且对分类具有良好区分度的特征更高的权重;TFC 方法是对 TF-IDF 方法进行了归一化处理;ITC 方法则是将频率数值改成了频率的对数值;TF-IWF 则是用特征频率的倒数的对数值来代替 IDF 方法.具体公式见文献[42].

这里,我们选择上述几种经典的权重赋值方法,按照评论长度不同,分组对比不同加权算法对分类效果的影响,结果见表 5.

Table 5 Classification effort comparison in different feature weight methods

表 5 不同特征加权方法的分类效果对比

分组 (评论长度)	*F*-Value					
	TF	IDF	TF-IDF	TFC	ITC	TF_IWF
1+	0.915	0.917	0.900	0.911	0.909	0.902
10+	0.872	0.861	0.862	0.865	0.854	0.876
20+	0.837	0.820	0.825	0.805	0.806	0.806
30+	0.814	0.785	0.815	0.765	0.789	0.781
40+	0.794	0.745	0.765	0.749	0.774	0.743
50+	0.785	0.743	0.761	0.763	0.760	0.752
60+	0.773	0.754	0.750	0.761	0.765	0.748
70+	0.765	0.749	0.755	0.752	0.768	0.744
90+	0.755	0.706	0.723	0.745	0.752	0.731
100+	0.735	0.714	0.705	0.733	0.750	0.727
150+	0.678	0.693	0.678	0.654	0.684	0.686
200+	0.642	0.52	0.642	0.518	0.552	0.588
250+	0.543	0.612	0.543	0.479	0.52	0.516
300+	0.4	0.567	0.4	0.433	0.433	0.524

表中数据表明:这些特征加权方法在长短不同的样本分组里表现各异,并没有一个特征加权方法能够整体

上优于其他方法.为简便起见,本文中大部分实验在特征形式化时均采用TF加权方法.此外,本实验与第3.1节的实验结论一致:不论对特征向量如何形式化,情感分类效果依然随着文本长度而呈现较为明显的下降趋势.

3.4 情感极性词在长短评论中的分布对比

情感极性词是能够明显表达用户情感倾向的词,如“太好了!”、“垃圾”等.不论是构造情感字典的方式,还是采用机器学习的方式,情感极性词对于情感分类都是至关重要的[47].本实验考察极性词在长短评论中是否均匀分布,是否会随着评论长度的增长而线性递增.对此,我们进行了如下的实验过程:

(1) 人工标注所有样本中出现的极性特征词.我们挑选了5名学生分别将评论中的极性词标识出来,并进行交叉验证以保证标注的一致性和正确性;

(2) 按评论的文字长度对样本进行分组,划分成长度不同、跨度相互重叠的评论组;

(3) 统计不同分组内,特征数量、极性词数量以及在特征中的占比情况.

表6为实验结果,可以看出:随着文本长度的增加,极性特征词的数量在攀升,但这种攀升是逐步收敛的,会逐步稳定在某个数值上;极性词在整体特征中比例逐步下降.在长评论中,极性词比率变小,说明长评论的表达大多采用了间接或更复杂的表述方式,不像短评论那样更直接地使用极性词.本实验可以看出:随着极性词比例的相对减少,具有噪音特性的其他特征词越来越多,因此也越来越难以识别长评论的情感倾向.由此可见,极性词比例的逐步下降很可能是长评论分类效果不好的原因之一.

Table 6 Count and proportion of polar word in different length group

表6 极性词在不同长度分组中的数量及比例关系

评论长度	特征总数	情感词个数	比例(情感词个数/特征总数)(%)
1~5	1 907	200	11.01
1~10	4 234	370	9.08
1~20	6 309	466	7.45
1~30	6 911	490	7.18
1~50	7 453	512	6.87
1~70	7 603	520	6.84
1~100	7 693	520	6.76
1~150	7 739	524	6.77
1~300	7 748	525	6.78
1~∞	7 899	526	6.66

3.5 长短评论特征的共现程度度量

用户在写评论表达自己观点时,表达的方式或用词存在一定的共性,也可以称为共现.这里,我们不做语义上的分析,仅从特征词上的重叠比例来考察长短评论之间的共性程度.

为了后续讨论的方便,我们先给出共现程度的相关定义:

设 $R_{i,j}$ 表示字数从 i 到 j 的评论集合,$F_{i,j}$ 代表 $R_{i,j}$ 中所有特征的集合,$F_{IG(i,j)}$ 表示经过IG特征筛选之后的特征集合;用 $|X|$ 表示集合 X 的元素个数.如果有两组评论集合分别是 $R_{i,j}$ 和 $R_{k,l}$,那么两组评论之间共现的特征集合为 $F_{i,j}\cap F_{k,l}$;$|F_{i,j}\cap F_{k,l}|$ 则是共现特征集合包含的元素个数.因此,两组评论之间的共现特征比为 $\frac{|F_{i,j}\cap F_{k,l}|}{|F_{i,j}\cup F_{k,l}|}$,共现特征占某评论特征 $F_{i,j}$ 的比例可表示为 $\frac{|F_{i,j}\cap F_{k,l}|}{|F_{i,j}|}$.这里,我们用3个指标来度量两组评论之间的共现程度:$\frac{|F_{i,j}\cap F_{k,l}|}{|F_{i,j}\cup F_{k,l}|},\frac{|F_{i,j}\cap F_{k,l}|}{|F_{i,j}|},\frac{|F_{i,j}\cap F_{k,l}|}{|F_{k,l}|}$.实验过程如下:

(1) 将评论按照字符长度大致分成4组:1~10,11~20,21~30,31~+∞;

(2) 分别统计各组的特征向量个数$|F_{1,10}|,|F_{11,20}|,|F_{21,30}|,|F_{31,+\infty}|$;

(3) 统计两组间的共现特征个数.如:$|F_{1,10}\cap F_{11,20}|$为1~10与11~20两组评论中共现的特征数;$|F_{1,10}\cap F_{21,30}|$为1~10与21~30两组评论中共现的特征数;

(4) 计算两组间的共现特征比.如,$\frac{|F_{1,10}\cap F_{11,20}|}{|F_{1,10}\cup F_{11,20}|}$表示 1~10 与 1~20 两组评论中共现的特征占两组总体特征的比例;

(5) 计算共现特征占某一评论分组中的比例.如$\frac{|F_{1,10}\cap F_{11,20}|}{|F_{1,10}|}$,$\frac{|F_{1,10}\cap F_{11,20}|}{|F_{11,20}|}$表示 1~10 与 11~20 两组评论中共现的特征分别占 1~10 组、11~20 组中特征的比例.

表 7 为各评论分组统计及计算结果.我们以 1~10 组为例,该组与 11~20 组的共现特征占其自身特征数量的 92%,与 21~30 组的共现特征占其自身特征数量的 90%,与 31~+∞组的共现特征占其自身特征数量的 95%;该共现特征又分别占两组总体特征数量的 62%,59%以及 54%.如此高的共现率说明,各组之间在表达方式上有很大的共性.

表 8 为各评论分组经过 IG 特征筛选前后的统计结果.以 1~10 组为例,筛选之前,该组内特征为 4 234 个,IG 筛选之后该组内特征为 948 个,共现特征个数占筛选前特征个数的 22%.未经筛选的 1~10 组与经过筛选的 11~20 组之间,共现特征个数是 1 103 个,共现特征个数仅占 1~10 组中特征个数的 26%,占两组所有特征的 25%.可以看出:经过筛选之后,各组间共现的特征项变少,共现率下降严重,信息损失严重.这可能是特征筛选后分类效果不够理想的原因之一.

Table 7 Feature co-occurrence degree measurement in different length group

表 7 不同长度评论间特征向量共现程度度量

	$\lvert F_{i,j}\cup F_{k,l}\rvert$				$\lvert F_{i,j}\cap F_{k,l}\rvert$				$\frac{\lvert F_{i,j}\cap F_{k,l}\rvert}{\lvert F_{i,j}\rvert}$, $\frac{\lvert F_{i,j}\cap F_{k,l}\rvert}{\lvert F_{k,l}\rvert}$, $\frac{\lvert F_{i,j}\cap F_{k,l}\rvert}{\lvert F_{i,j}\cup F_{k,l}\rvert}$			
	k=1, l=10	k=11, l=20	k=21, l=30	k=31, l=+∞	k=1, l=10	k=11, l=20	k=21, l=30	k=31, l=+∞	k=1, l=10	k=11, l=20	k=21, l=30	k=31, l=+∞
i=1, j=10	4234	6 305	6 486	7 527	4 234	3 894	3 796	4 032	1	0.92, 0.65, 0.62	0.90, 0.63, 0.59	0.95, 0.55, 0.54
i=11, j=20		5 965	6 797	7 598		5 965	5 216	5 692		1	0.87, 0.86, 0.77	0.95, 0.78, 0.75
i=21, j=30			6 048	7 588			6 048	5 785			1	0.96, 0.79, 0.76
i=31, j=+∞				7 325				7 325				1

Table 8 Feature co-occurrence degree measurement used feature selection method in different length group

表 8 特征筛选后不同字数评论之间特征向量共现程度度量

		经过 IG 特征筛选											
		$\lvert F_{i,j}\cup F_{k,l}\rvert$				$\lvert F_{i,j}\cap F_{k,l}\rvert$				$\frac{\lvert F_{i,j}\cap F_{IG(k,l)}\rvert}{\lvert F_{i,j}\rvert}$, $\frac{\lvert F_{i,j}\cap F_{IG(k,l)}\rvert}{\lvert F_{IG(k,l)}\rvert}$, $\frac{\lvert F_{i,j}\cap F_{IG(k,l)}\rvert}{\lvert F_{i,j}\cup F_{IG(k,l)}\rvert}$			
		k=1, l=10	k=11, l=20	k=21, l=30	k=31, l=+∞	k=1, l=10	k=11, l=20	k=21, l=30	k=31, l=+∞	k=1, l=10	k=11, l=20	k=21, l=30	k=31, l=+∞
未经过IG特征筛选	i=1, j=10	4 234	4 396	4 455	4 533	948	1 103	1 102	1 224	0.22, 1, 0.22	0.26, 0.87, 0.25	0.26, 0.83, 0.25	0.29, 0.8, 0.27
	i=11, j=20	5 999	5 965	6 008	6 024	914	1 265	1 280	1 464	0.15, 0.96, 0.15	0.21, 1, 0.21	0.21, 0.97, 0.21	0.25, 0.96, 0.24
	i=21, j=30	6 102	6 065	6 048	6 088	894	1 248	1 323	1 483	0.15, 0.94, 0.15	0.21, 0.99, 0.21	0.22, 1, 0.22	0.25, 0.97, 0.24
	i=31, j=+∞	7 343	7 326	7 329	7 325	930	1 264	1 319	1 523	0.13, 0.98, 0.13	0.17, 1, 0.17	0.18, 1, 0.18	0.21, 1, 0.21

从前面的各个实验看出,对于轻型评论的情感分类,具有以下几个特点:

(1) 短评论分类效果更好,长评论分类效果会随着文本长度的增加呈现下降趋势;

(2) 传统的特征筛选方法对于大部分是短评论的手机评论来讲并不是特别适用,对于短评论部分,它没有起到提高分类效果的作用;同时,筛选后特征数量急剧减少,甚至导致部分样本所有特征权重为空,有价值的特征词并没有被保留下来,也使得分类难以判断;

(3) 各种特征加权方法对于轻型评论来讲表现并不稳定;

(4) 短评论中极性词含量较高,而长评论中极性词并没有因为文本长度增加而显著增加;相反,其极性词比例在下降.这有可能是长评论中分类准确性降低的原因之一;

(5) 短评论中的特征词与长评论中的特征词有较大的重叠率,说明在长评论中存在与短评论类似的基本表达方法,但长评论中存在着比短评论更多的特征,而这些特征很可能是噪音,是造成长评论分类准确性较低的原因之一.

4 基于短评论共现的特征筛选法

4.1 方法思想

轻型评论中,短评论呈现出分类效果好、极性词含量高、与长评论的共现特征多等优点;而长评论呈现出分类效果不够理想、极性词含量低、噪音特征较多等不足.本文希望借助轻型评论中短评论的优点,提高长评论的分类效果.主要考虑以下几个因素:

(1) 短评论的数量极大.

从统计数据上看,轻型评论中短评论占了整体评论的绝大多数,因此,短评论所包含的信息应该被充分地挖掘和利用,而不是被忽略或被过滤掉.

(2) 短评论中极性词和情感表达方式多,这使得短评论具有其他评论所不具备的优势信息.

一些需要借助领域词典或上下文才能判断出情感倾向的词,诸如“快”、“慢”、“多”、“少”,在短评论中已经给出了明显的标示,如:“功能好少”、“下载速度快”等.这样的词在短评论中大批量重复地出现,并且短评论的句子简单、同现的其他特征词少,无关的噪音也少.这些都给训练提供了良好的学习样本.此外,短评论中不断有新生的网络词汇出现,如“坑爹啊”、“有木有”等,这也为有效识别网络用语、提高分类效果提供了很好的条件.

(3) 短评论中的表达方式与长评论中的表达方式有很高的重叠性.

如,短评论中有90%多的特征词都会出现在长评论中,而长评论中大概有50~60%的特征词与短评论中的特征词相重叠,这说明长评论中有和短评论类似的表达,可以认为长评论特征的一部分是由短评论中的特征构成的,一部分是由其他的特征构成的.而短评论中的大部分特征有利于极性的判断,长评论中的另一部分非常有可能是极性特征不够明显的噪音.

基于上述的分析,本文提出了一种利用短评论优势以提高长评论情感分类的方法——基于短评论共现的特征筛选法简称 SCO(base on short-review co-occurrence)特征选择方法.

4.2 方法定义

如前面实验所示:传统的特征选择方法对于长评论来讲,分类的准确率稍微有所提高,但提高的幅度很小;同时,由于特征筛选方法使得特征更加稀疏,很多样本变成了空实例,分类效果没有预期的好.

方法的前提假设是:短评论特征词富含极性词和极性表达方式.基于这个假设,我们扩展了传统的特征选择方法,在现有特征筛选方法基础上增加长评论中与短评论共现的特征,即增加那些在长评论和短评论中共同出现的特征.增加的这部分特征由于在短评论中曾经出现过,因此富含极性词或极性表达方式,能够辅助长评论进行情感分类.

这里给出该方法的定义:设评论集合为 $R_{i,j}$,其中,$i,j(i<j)$表示评论的字符长度,其特征集合为 $F_{i,j}$,经过 IG 筛选之后,特征集合为 $F_{IG(i,j)}$.

假设有两组评论 $R_{i,j}$ 和 $R_{k,l}$,其中:$R_{i,j}$ 为长评论,其原始特征集合为 $F_{i,j}$,传统特征筛选方法,筛选出来的特征集合为 $F_{IG(i,j)}$,该集合通常为参与情感分类的特征集合;为了提高分类的准确性,SCO 方法将分类效果较好的短小评论 $R_{k,l}$ 的特征集合 $F_{k,l}$ 引入进来,并获取其共现的特征集合 $F_{i,j}\cap F_{k,l}$.最终,评论 $R_{i,j}$ 的最终特征集合为 $F_{IG(i,j)}\cup(F_{i,j}\cap F_{k,l})$.该集合作为新的情感分类的特征集合,通过重新计算各特征值,再进行情感分类.

该方法与其他方法的对比如图 3 所示.如果不采用特征筛选方法,则长评论 $R_{i,j}$ 的最终特征集合为其原始的特征集合 $F_{i,j}$;如果采用特征筛选方法以降低噪音,则长评论 $R_{i,j}$ 的最终特征集合为 $F_{IG(i,j)}$;如果采用 SOC 方法,在降低部分噪音的同时引入短评论 $R_{k,l}$ 的特征 $F_{k,l}$,将 $F_{k,l}$ 与 $F_{i,j}$ 的共现集合 $F_{i,j}\cap F_{k,l}$ 作为辅助集合增加到传统的特征筛选集合 $F_{IG(i,j)}$ 中,即,长评论 $R_{i,j}$ 的最终特征集合为 $F_{IG(i,j)}\cup(F_{i,j}\cap F_{k,l})$.

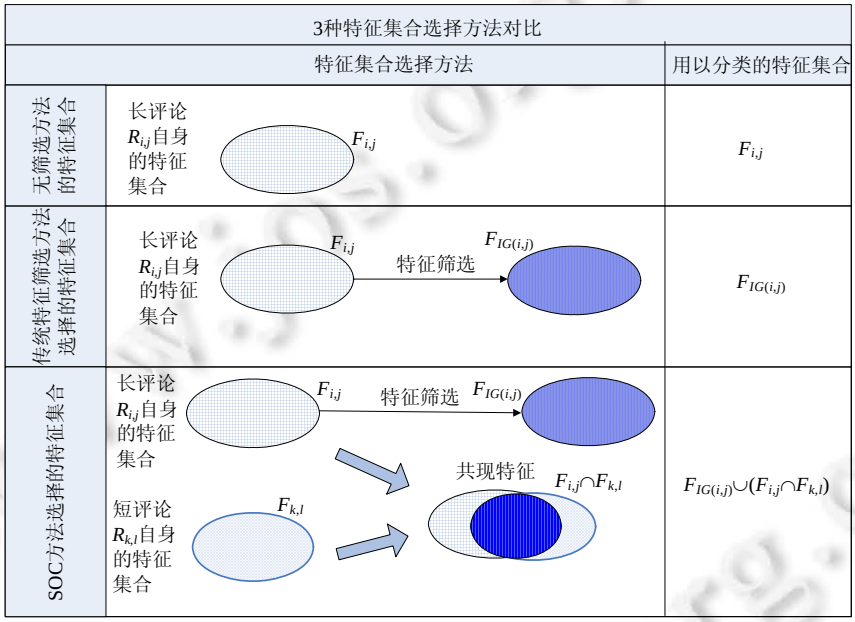

Fig.3 Three feature selection methods comparison

图 3 3 种特征集合选择方法对比

5 方法验证

5.1 实验流程

为了验证本文提出的 SCO 方法的有效性,我们进行了情感分类的对比实验.分别选择了:① 不采用特征筛选方法;② 采用特征筛选方法(IG 和 CHI);③ 采用本文提出的 SCO 方法等 3 种不同形式的分类效果进行比较.这里,我们重点阐述 SCO 方法的实验流程.

第 1 步:我们将评论分为两部分:短评论,记为 $R_{k,l}$;长评论,记为 $R_{i,j}$.

这里,我们选择长度≥100 字符的评论作为长评论.从前面的实验结果看,长度在 100 以上的评论其分类效果开始急剧下降,从 75%降到了 40%,因此,我们选择字符在 100 以上的评论作为实验用的长评论,作为需要提高和改善的主要对象.

在 SCO 方法中,短评论中的特征是用来补充长评论特征的.选择多少长度区间的短评论,即 k,l 的选择是一个值得研究的问题.如果选择的区间过小,短评论的特征过少,与长评论中的共现特征也会变少,不能起到补充信息的作用;如果选择的区间过大,短评论与长评论的共现特征会增大,但同时也会引入噪音,影响分类效果.本文拟通过实验的方法,比较不同短文本长度区间对分类效果的影响.

由于长度≤50 的评论分类效果较好且样本实例众多,因此选择该部分作为短评论,并将其分成 5 个长度不同的实验区间:1~10,1~20,1~30,1~40 和 1~50,以考察不同区间对分类效果的影响.这里,我们的 k 值都从 1 选取,主要考虑到越是短小的评论,其极性词比例越高,因此,我们将最短评论作为短评论选择区间的起点.

第 2 步:对长评论进行特征筛选处理.

如果筛选前的特征集合记为 $F_{i,j}$,则 IG 筛选之后的特征集合记为 $F_{IG(i,j)}$.

第 3 步:获取不同区间的短评论特征集合 $F_{k,l}$.根据实验分组,分别为 $F_{1,10}$,$F_{1,20}$,$F_{1,30}$,$F_{1,40}$ 和 $F_{1,50}$.

第 4 步:采用 $F_{IG(i,j)}\cup(F_{i,j}\cap F_{k,l})$公式,进行长、短评论特征集合的交、并操作.

例如,长度在 100+的长评论将分别与 5 个不同区间的短评论特征集合相交,找出其共现特征,并将该共现特征集合与其自身筛选后的特征集合进行合并.

第 5 步:将第 4 步得到的集合作为新的特征集合,重新计算各个特征的权重值,这里采用 TF 方法.

第 6 步:采用分类算法(这里选用 NB 分类算法)进行分类,对比分类效果.

5.2 实验效果对比

表 9 显示了不采用特征筛选方法、采用 IG 特征筛选方法以及采用 SCO 方法后的效果比较.从实验结果看:SCO 方法分类效果要普遍好于 IG 特征筛选方法;特别是当选择 1~30 及 1~40 的短评论区间时,分类效果最优.为了进一步验证 SCO 方法,我们按照同样的思路,将特征筛选方法换成了 CHI 方法,表 10 显示 CHI 特征筛选方法与 SCO 方法的效果比较,SCO 方法同样也普遍好于单纯采用 CHI 方法.

Table 9 Compare SCO method classification with IG-methods

表 9 SCO 方法与 IG 方法的分类性能对比

NB 分类算法效果 F-VALUE		不采用特征筛选	采用 IG 特征筛选	基于短评论共现的特征筛选方法(SCO) 采用 $F_{IG(i,j)}\cup(F_{i,j}\cap F_{k,l})$作为分类特征				
				不同选择区间的短评论 $R_{k,l}$				
				k=1 l=10	k=1 l=20	k=1 l=30	k=1 l=40	k=1 l=50
不同长度的长评论 $R_{i,j}$	i=100, j=+∞	0.735	0.739	0.721	0.723	0.773↑	0.771↑	0.736
	i=150, j=+∞	0.678	0.688	0.645	0.71↑	0.718↑	0.721↑	0.677
	i=200, j=+∞	0.642	0.640	0.643↑	0.67↑	0.726↑	0.724↑	0.652↑
	i=250, j=+∞	0.543	0.546	0.6↑	0.611↑	0.667↑	0.644↑	0.64↑
	i=300, j=+∞	0.4	0.560	0.587↑	0.649↑	0.692↑	0.674↑	0.673↑

Table 10 Compare SCO method classification with CHI-methods

表 10 SCO 方法与 CHI 方法的分类性能对比

NB 分类算法效果 F-VALUE		不采用特征筛选	采用 CHI 特征筛选	基于短评论共现的特征筛选方法(SCO) 采用 $F_{CHI(i,j)}\cup(F_{i,j}\cap F_{k,l})$作为分类特征				
				不同选择区间的短评论 $R_{k,l}$				
				k=1 l=10	k=1 l=20	k=1 l=30	k=1 l=40	k=1 l=50
不同长度的长评论 $R_{i,j}$	i=100, j=+∞	0.735	0.738	0.730	0.730	0.772↑	0.762↑	0.740↑
	i=150, j=+∞	0.678	0.687	0.652	0.677	0.717↑	0.724↑	0.673
	i=200, j=+∞	0.642	0.643	0.655↑	0.668↑	0.718↑	0.712↑	0.656↑
	i=250, j=+∞	0.543	0.545	0.620↑	0.619↑	0.662↑	0.651↑	0.658↑
	i=300, j=+∞	0.4	0.562	0.585↑	0.657↑	0.682↑	0.691↑	0.683↑

在 SCO 方法中,如何选择最理想的短评论长度区间,也是一个较为关键的问题.选择不同长度区间的短评论,情感分类效果有所不同,其原因有两种:(1) 如果选择的长度区间过小,则该区间的短评论特征与长评论特征的共现程度太低,特别是针对每条长评论来讲,其共现的特征更少,并没有起到补充极性特征或极性表达方式的作用;(2) 如果选择的长度区间过大,长短评论之间的共现程度有所增加,但同时也带来一定的噪音,从而影响其分类效果.

以表 9 中数据为例,图 4 显示了 SCO 方法中不同区间的分类效果对比.当我们采用 1~10 区间的短评论特征

与长评论 IG 特征合并时,效果并不很理想;而采用 1~20 区间时,性能略有所上升;选择 1~30 及 1~40 区间时,效果较为理想,各项分组均高于 IG 特征筛选方法时的效果;但当选择 1~50 区间时,部分长评论分组的效果开始有所下降.从实验效果来看:选择 1~30 或 1~40 特征区间,各个评论长度的分类效果最为理想.这里,我们仅通过实验对比的方法,发现长度在 1~30 区间分类效果最好.如何提出一套判定方法或判定指标,能够判断出最佳的短评论长度区间,还有待于后期进一步研究.

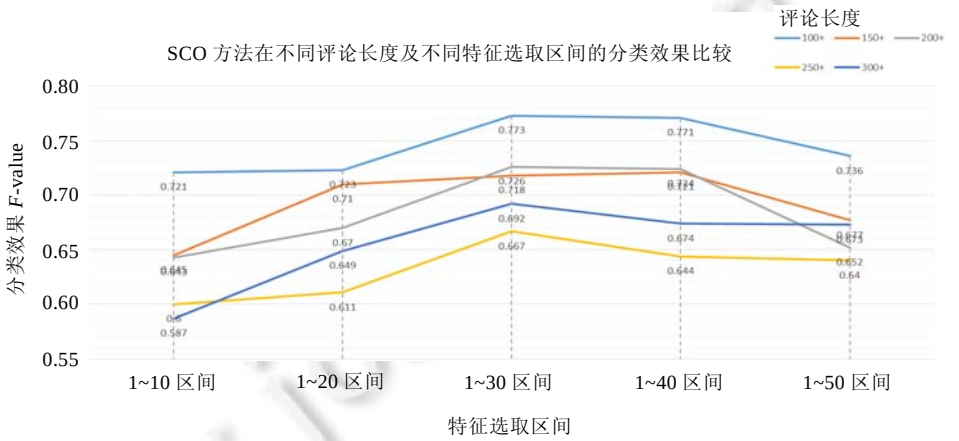

Fig.4 SCO method classification comparison in different reviews and different interval

图 4 SCO 方法在不同评论长度及不同特征选取区间的分类效果比较

5.3 方法总结

传统的特征筛选方法可以降低文本中的噪音,减少特征的维度.但在长评论中,其特征词重复率较低,相互间共同点较少,在特征筛选过程中,一些有用的信息因为出现在文档中的次数少而容易被过滤掉.因此,特征筛选方法对于轻型评论来讲更容易造成信息缺失,将短评论与长评论的共现特征补充进来,可以有效地增加有利于极性判断的特征词,从而提高分类效果.

这里,我们给出样本实例:

样本评论 1:“好友圈——把微信好友圈,弄个有我的好友圈,自己发过什么,点击我的好友圈,就可以看见了,就不要找半死,每天好友圈发的那么东西,我要找我发字和我分享的音乐都得找半死,还有照片,怎么那么烂啊,为什么不可以设置回答答案就可以看见照片的???!!!!!.”(负向)

该样本在经过 IG 特征筛选了之后,只保留了“每天”、“可以”、“别”这样的特征词,机器判别其为正向.显然,筛选后没有保留该评论中的关键特征词,这有悖于特征筛选的初衷.经过短评论特征的补充,该句子中的大部分有用的特征词,如“好友圈”、“分享”、“半死”、“烂”、“!”等,由于在短评论中曾经大量出现过而被保留下来,并重新赋予了权重,机器判别为负向,从而提高了分类效果.

样本评论 2:“我好像是沙发啊,哈哈!希望能出 ipad 版,不要总是 iphone 班(版)啊,.....然后就是要提高下流畅度,在左划删除会话的时候,总是会有黑条在闪动,需要改进啊.”(正向)

该样本在经过 IG 特征筛选了之后,保留的特征词有“删除”、“沙发”,判断出的极性是负向.但是经过短评论的特征补充之后,更多的特征词被保留下来,如“提高”、“流畅度”、“闪动”、“改进”、“iphone”等,机器判断也改为正向.

由上述两个样本实例我们可以看出:短评论中的特征有效地弥补了特征筛选所造成的信息过少的不足,补充的信息也有效地提高了分类的效果.

本论文提出的方法不仅能够辅助提高轻型评论中较长评论的分类效果,还可以扩展应用到其他的场景中.例如,游戏厂商可以通过采集手机游戏类的轻型评论,建立游戏类的短评论特征模型,利用该特征模型,辅助判

断互联网游戏用户发表在论坛或微博上的情感,提高同领域长评论的分类效果.此外,利用轻型评论中的短评论特征模型中的优势信息,也可以建立该领域内的极性词词典或典型极性表达方式列表,这个领域字典或列表对于分析用户的情感也具有较好的辅助作用.

6 结论和未来的工作

随着智能移动设备的日渐普及,文本短小、数量众多的轻型评论将成为移动设备上评论的主流.本文将这类轻型评论作为研究对象,探讨该类评论的特点;在相关实验基础上,发现该类评论在情感分析上具有的特性;根据这些特性,本文提出了基于短评论共现的特征筛选法,以提高轻型评论中长评论的分类效果.

本文贡献主要有以下 3 点:

1) 提出了轻型评论的概念、特点及定义.

智能移动设备上日益丰富的用户评论是一种新兴的评论形态,是文本挖掘以及意见挖掘值得关注的研究对象.本文将这类评论定义为轻型评论,并根据统计及实验分析总结了该类评论的特点,即:平均字数少、跨度大,且短小评论占了评论中的主流,评论长度与评论数量满足幂率分布规律.

同时,本文也指出了轻型评论与互联网评论、短文本研究的不同点.在长度上,轻型评论比互联网评论更加短小;在数量上,短小评论成了评论的主流;这使得传统的分析方法值得重新审视;短文本的概念尽管与轻型评论有相似之处,但在分类的特性上,短文本研究主要是针对主题的分类研究,解决的是短文本信息缺失的问题.事实上,轻型评论中短文本中的情感极性信息并不缺失,相反,其信息含量更高,因此其情感分类效果更好.这也是轻型评论研究与短文本研究的不同点.

2) 发现轻型评论在情感分类方面具有的特性

本文按照传统的情感分析方法进行了一系列实验对比研究,发现轻型评论中:

(1) 文本长度越短,其情感分类效果越好;而随着文本长度的增加,其分类效果逐步降低;

(2) 传统的特征筛选方法,比如 IG,CHI 方法对轻型评论并不完全适用,甚至会降低短评论的分类效果;

(3) 传统的特征加权方法在不同的评论长度中,分类表现很不稳定;

(4) 短评论中的极性词比例要明显高于长评论中的含量;长评论中的极性词数量不是线性递增的,而是逐步收敛趋向稳定;

(5) 短评论与长评论中的特征有较高的共现度,说明长短评论在用词上存在着很高的相似性.

这些特性的发现,不仅使得我们重新审视传统的情感分类方法和结论,也有利于我们找到适合轻型评论自身特点的情感分类方法.

3) 提出了基于短评论共现的特征筛选方法.

针对轻型评论的特点,本文提出了基于短评论共现的特征筛选方法(SCO).该方法充分利用了短评论中富含极性词及极性表达方式的特点,在传统特征筛选方法的基础上,将长评论与短评论共现的特征保留下来;在筛选掉噪音的基础上,扩大了特征的维度,将优势信息补充进来.实验证明,该方法在长评论的情感分类上取得了良好的效果.

在未来的工作中,我们将会从以下几个方面深入研究:

(1) 继续探讨针对轻型评论的情感分类方法,例如,如何定义一套指标或方法,帮助选择最佳的短评论长度区间;

(2) 扩大研究样本,将论文中提出的方法应用到其他领域的评论中;

(3) 尝试利用轻型评论中的短评论信息,建立某一领域内的极性词词典,辅助语义角度的情感分析.

致谢 本文是作者在美国访学期间完成的.在此,向给予本文的工作支持和建议的北京航空航天大学计算机学院软件研究所、美国弗吉尼亚理工大学、美国劳伦斯理工大学的老师及同学表示感谢.

References:

[1] Hu MQ, Liu B. Mining and summarizing customer reviews. In: Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Seattle: ACM Press, 2004. 168−177. [doi: 10.1145/1014052.1014073]

[2] Dave K, Lawrence S, Pennock DM. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proc. of the 12th Int'l Conf. on World Wide Web. Budapest: ACM Press, 2003. 519−528. [doi: 10.1145/775152.775226]

[3] Turney PD. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proc. of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002. 417−424.

[4] Abrahams AS, Jiao J, Wang GA, Fan WG. Vehicle defect discovery from social media. Decision Support Systems, 2012,54(1): 87−97. [doi: 10.1016/j.dss.2012.04.005]

[5] Tencent. Google android market. 2014 (in Chinese). http://tech.qq.com/a/20120104/000474.htm

[6] Encyclopedia B. Appstore. 2014 (in Chinese). http://baike.baidu.com/view/2771827.htm

[7] Yang Z, Lai YX, Duan LJ, Li YJ. Short text sentiment classification based on context reconstruction. Acta Automatica Sinica, 2012, 38(1):55−67 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2012.00055]

[8] Su JS, Zhang BF, Xu X. Advances in machine learning based text categorization. Ruan Jian Xue Bao/Journal of Software, 2006, 17(9):1848−1859 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/1848.htm

[9] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Philadelphia: Association for Computational Linguistics, 2002. 79−86. [doi: 10.3115/1118693.1118704]

[10] Ni XC, Xue GR, Ling X, Yu Y, Yang Q. Exploring in the weblog space by detecting informative and affective articles. In: Proc. of the 16th Int'l Conf. on World Wide Web. Banff: ACM Press, 2007. 281−290. [doi: 10.1145/1242572.1242611]

[11] Mullen T, Collier N. Sentiment analysis using support vector machines with diverse information sources. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Barcelona: Association for Computational Linguistics, 2004. 412−418.

[12] Whitelaw C, Garg N, Argamon S. Using appraisal groups for sentiment analysis. In: Proc. of the 14th ACM Int'l Conf. on Information and Knowledge Management. Bremen: ACM Press, 2005. 625−631. [doi: 10.1145/1099554.1099714]

[13] Cui H, Mittal V, Datar M. Comparative experiments on sentiment classification for online product reviews. In: Proc. of the 21st National Conf. on Artificial Intelligence. Boston: AAAI Press, 2006. 1265−1270.

[14] Wiebe JM. Learning subjective adjectives from corpora. In: Proc. of the 17th National Conf. on Artificial Intelligence and 12th Conf. on Innovative Applications of Artificial Intelligence. Austin: IEEE, 2000. 735−740.

[15] Riloff E, Wiebe J. Learning extraction patterns for subjective expressions. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Sapporo: Association for Computational Linguistics, 2003. 105−112. [doi: 10.3115/1119355.1119369]

[16] Turney PD, Littman ML. Measuring praise and criticism: Inference of semantic orientation from association. ACM Trans. on Information Systems, 2003,21(4):315−346.

[17] Kim SM, Hovy E. Identifying and analyzing judgment opinions. In: Proc. of the Main Conf. on Human Language Technology Conf. of the North American Chapter of the Association of Computational Linguistics. New York: Association for Computational Linguistics, 2006. 200−207. [doi: 10.3115/1220835.1220861]

[18] Zhu YL, Min J, Zhou YQ, Huang XJ, Wu LD. Semantic orientation computing based on hownet. Journal of Chinese Information Processing, 2006, 20(1):14−20 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2006.01.003]

[19] Ni MS, Lin HF. Mining product reviews based on association rule and polar analysis. In: Proc. of the NCIRCS. 2007. 628−634 (in Chinese with English abstract).

[20] Xu J, Ding YX, Wang XL. Sentiment classification for Chinese news using machine learning methods. Journal of Chinese Information Processing, 2007,21(6):95−100 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2007.06.013]

[21] Moschitti A, Basili R. Complex linguistic features for text classification: A comprehensive study. In: Proc. of the Advances in Information Retrieval. Springer-Verlag, 2004. 181−196. [doi: 10.1007/978-3-540-24752-4_14]

[22] Kehagias A, Petridis V, Kaburlasos VG, Fragkou P. A comparison of word- and sense-based text categorization using several classification algorithms. Journal of Intelligent Information Systems, 2003,21(3):227−247. [doi: 10.1023/A:1025554732352]

[23] Ye Q, Lin B, Li Y. Sentiment classification for chinese reviews: A comparison between SVM and semantic approaches. In: Proc. of the Int'l Conf. on Machine Learning and Cybernetics. Guangzhou: IEEE, 2005. 2341−2346. [doi: 10.1109/ICMLC.2005.1527335]

[24] Ye Q, Shi W, Li YJ. Sentiment classification for movie reviews in Chinese by improved semantic oriented approach. In: Proc. of the 39th Annual Hawaii Int'l Conf. on System Sciences. Kauai: IEEE, 2006. 53b. [doi: 10.1109/HICSS.2006.432]

[25] Yao TF, Nie QY, Li JC, Li LL, Lou DC, Chen K, Fu Y. An opinion mining system for Chinese automobile reviews. In: Proc. of the Chinese Society of Chinese Information 25 Anniversary Conf. Beijing: Tsinghua University Press, 2006. 260−281 (in Chinese with English abstract).

[26] Ye Q, Zhang ZQ, Luo ZX. Automatically measuring subjectivity of Chinese sentences for sentiment analysis to reviews on the Internet. China Journal of Information Systems, 2007,1(1):79−91 (in Chinese with English abstract).

[27] Corpus-Imdb. 2014. http://www.cs.cornell.edu/people/pabo/movie-review-data/

[28] Kasthuriarachchy BH, De Zoysa K, Premaratne HL. A review of domain adaptation for opinion detection and sentiment classification. In: Proc. of the Int'l Conf. on Advances in ICT for Emerging Regions. Colombo: IEEE, 2012. 209−213. [doi: 10.1109/ICTer.2012.6423023]

[29] Tan S, Wang YF. Chnsenticorp corpus. 2010 (in Chinese). http://www.searchforum.org.cn/tansongbo/corpus.htm

[30] Wang S, Fan XH, Chen XL. Chinese short text classification based on hyponymy relation. Journal of Computer Applications, 2010, 30(3):603−606 (in Chinese with English abstract). [doi: 10.3724/SP.J.1087.2010.00603]

[31] Mitchell TM. Machine learning and data mining. Communications of the ACM, 1999,42(11):30−36. [doi: 10.1145/319382.319388]

[32] Dunning T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 1993,19(1):61−74.

[33] Wiener E, Pedersen JO, Weigend AS. A neural network approach to topic spotting. In: Proc. of the 4th Annual Symp. on Document Analysis and Information Retrieval. 1995. 317−332.

[34] Chen L, Tokuda N, Nagai A. A new differential LSI space-based probabilistic document classifier. Information Processing Letters, 2003,88(5):203−212. [doi: 10.1016/j.ipl.2003.09.002]

[35] Kim H, Howland P, Park H. Dimension reduction in text classification with support vector machines. Journal of Machine Learning Research, 2005,6:37−53.

[36] Park H, Jeon M, Rosen JB. Lower dimensional representation of text data based on centroids and least squares. BIT Numerical Mathematics, 2003,43(2):427−448. [doi: 10.1023/A:1026039313770]

[37] Wang L, Jia Y, Han W. Instant message clustering based on extended vector space model. In: Proc. of the 2nd Int'l Conf. on Advances in Computation and Intelligence. Springer-Verlag, 2007. 435−443. [doi: 10.1007/978-3-540-74581-5_48]

[38] Fan XH, Hu HG. A new model for Chinese short-text classification considering feature extension. In: Proc. of the Int'l Conf. on Artificial Intelligence and Computational Intelligence. Sanya: IEEE, 2010. 7−11. [doi: 10.1109/AICI.2010.125]

[39] Adams PH, Martell CH. Topic detection and extraction in chat. In: Proc. of the IEEE Int'l Conf. on Semantic Computing. Santa Clara: IEEE, 2008. 581−588. [doi: 10.1109/ICSC.2008.61]

[40] Hall M, Frank E, Holmes G, Pfahringer B, Reatemann P, Witten IN. The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter, 2009, 11(1):10−18. [doi: 10.1145/1656274.1656278]

[41] Mmseg4j. 2014. http://code.google.com/p/mmseg4j/

[42] Zong CQ. Statistical Natural Language Processing. Tsinghua University Press, 2008 (in Chinese).

[43] Liu ZM, Liu L. Empirical study of sentiment classification microblog based on machine learning. Computer Engineering and Applications, 2012,48(1):1−4 (in Chinese with English abstract). [doi: 10.3778/j.issn.1002-8331.2012.01.001]

附中文参考文献:

[5] 腾讯科技.谷歌 Android Market. 2014. http://tech.qq.com/a/20120104/000474.htm

[6] 百度百科.AppStore. 2014. http://baike.baidu.com/view/2771827.htm

[7] 杨震,赖英旭,段立娟,李玉.基于上下文重构的短文本情感极性判别研究.自动化学报,2012,38(1):55−67. [doi: 10.3724/SP.J.1004.2012.00055]

[8] 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展.软件学报,2006,17(9):1848−1859. http://www.jos.org.cn/1000-9825/17/1848.htm

[18] 朱嫣岚,闵锦,周雅倩,黄萱菁,吴立德.基于 HowNet 的词汇语义倾向计算.中文信息学报,2006,20(1):14−20. [doi: 10.3969/j.issn.1003-0077.2006.01.003]

[19] 倪茂树,林鸿飞.基于关联规则和极性分析的商品评论挖掘.见:第 3 届全国信息检索与内容安全学术会议.苏州,2007.635−642.

[20] 徐军,丁宇新,王晓龙.使用机器学习方法进行新闻的情感自动分类.中文信息学报,2007,21(6):95−100. [doi: 10.3969/j.issn.1003-0077.2007.06.013]

[25] 姚天昉,聂青阳,李建超,李林琳,娄德成,陈珂,付宇.一个用于汉语汽车评论的意见挖掘系统.见:中国中文信息学会 25 周年学术会议.北京:清华大学出版社,2006.260−281.

[26] 叶强,张紫琼,罗振雄.面向互联网评论情感分析的中文主观性自动判别方法研究.信息系统学报,2007,1(1):79−91.

[29] 谭松波,王月粉.ChnSentiCorp 语料.2010. http://www.searchforum.org.cn/tansongbo/corpus.htm

[30] 王盛,樊兴华,陈现麟.利用上下位关系的中文短文本分类.计算机应用,2010,30(3):603−606. [doi: 10.3724/SP.J.1087.2010.00603]

[42] 宗成庆,统计自然语言处理.北京:清华大学出版社,2008.

[43] 刘志明,刘鲁.基于机器学习的中文微博情感分类实证研究.计算机工程与应用,2012,48(1):1−4. [doi: 10.3778/j.issn.1002-8331.2012.01.001]

张林(1970－),女,山西翼城人,博士生,副教授,CCF 会员,主要研究领域为意见挖掘,需求验证.

E-mail: zhanglin_hz@163.com

钱冠群(1978－),男,博士,主要研究领域为大数据处理,用户属性挖掘,复杂网络分析.

E-mail: Qianguanqun@buaa.edu.cn

樊卫国(1973－),男,博士,教授,博士生导师,主要研究领域为商业智能挖掘,文本挖掘.

E-mail: wfan@vt.edu

华琨(1977－),男,博士,副教授,主要研究领域为人工智能,多媒体挖掘.

E-mail: Khua@ltu.edu

张莉(1968－),女,博士,教授,博士生导师,CCF 会员,主要研究领域为软件过程,需求分析,软件测试.

E-mail: Lily@buaa.edu.cn