中/青/文/库

文本情感分析关键技术研究

朱　俭◎著

中国社会科学出版社

中/青/文/库

本书得到中国青年

文本情感分析关键技术研究

朱　俭◎著

中国社会科学出版社

图书在版编目(CIP)数据

文本情感分析关键技术研究/朱俭著. —北京：中国社会科学出版社，2015.11
ISBN 978-7-5161-5996-5

Ⅰ.①文… Ⅱ.①朱… Ⅲ.①文本编辑—研究 Ⅳ.①TP311.11

中国版本图书馆 CIP 数据核字(2015)第 081336 号

出 版 人 赵剑英
责任编辑 李炳青
责任校对 朱妍洁
责任印制 李寡寡

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 http://www.csspw.cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印 刷 北京君升印刷有限公司
装 订 廊坊市广阳区广增装订厂
版 次 2015 年 11 月第 1 版
印 次 2015 年 11 月第 1 次印刷

开 本 710×1000 1/16
印 张 18.75
插 页 2
字 数 320 千字
定 价 65.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社营销中心联系调换
电话：010-84083683

《中青文库》编辑说明

中国青年政治学院是在中央团校基础上于1985年12月成立的，是共青团中央直属的唯一一所普通高等学校，由教育部和共青团中央共建。中国青年政治学院成立以来，坚持“质量立校、特色兴校”的办学思想，艰苦奋斗、开拓创新，教育质量和办学水平不断提高。学校是教育部批准的国家大学生文化素质教育基地，中华全国青年联合会和国际劳工组织命名的大学生KAB创业教育基地。学校与中央编译局共建青年政治人才培养研究基地，与北京市共建社会工作人才发展研究院和青少年生命教育基地。

目前，学校已建立起包括本科教育、研究生教育、留学生教育、继续教育和团干部培训等在内的多形式、多层次的教育格局。设有中国马克思主义学院、青少年工作系、社会工作学院、法律系、经济系、新闻与传播系、公共管理系、中国语言文学系、外国语言文学系等9个教学院系，文化基础部、外语教学研究中心、计算机教学与应用中心、体育教学中心等4个教学中心（部），轮训部、继续教育学院、国际教育交流学院等3个教学培训机构。

学校现有专业以人文社会科学为主，涵盖哲学、经济学、法学、文学、管理学5个学科门类。学校设有思想政治教育、法学、社会工作、劳动与社会保障、社会学、经济学、财务管理、国际经济与贸易、新闻学、广播电视学、政治学与行政学、汉语言文学和英语等13个学士学位专业，其中社会工作、思想政治教育、法学、政治学与行政学为教育部特色专业。目前，学校拥有哲学、马克思主义理论、法学、社会学、新闻传播学和应用经济学等6个一级学科硕士授权点和1个专业硕士学位点，同时设有青少年研究院、中国马克思主义研究中心、中国志愿服

务信息资料研究中心、大学生发展研究中心、大学生素质拓展研究中心等科研机构。

在学校的跨越式发展中，科研工作一直作为体现学校质量和特色的重要内容而被予以高度重视。2002 年，学校制定了教师学术著作出版基金资助条例，旨在鼓励教师的个性化研究与著述，更期之以兼具人文精神与思想智慧的精品的涌现。出版基金创设之初，有学术丛书和学术译丛两个系列，意在开掘本校资源与移译域外菁华。随着年轻教师的剧增和学校科研支持力度的加大，2007 年又增设了博士论文文库系列，用以鼓励新人，成就学术。三个系列共同构成了对教师学术研究成果的多层次支持体系。

十几年来，学校共资助教师出版学术著作百余部，内容涉及哲学、政治学、法学、社会学、经济学、文学艺术、历史学、管理学、新闻与传播等学科。学校资助出版的初具规模，激励了教师的科研热情，活跃了校内的学术气氛，也获得了很好的社会影响。在特色化办学愈益成为当下各高校发展之路的共识中，2010 年，校学术委员会将遴选出的一批学术著作，辑为《中青文库》，予以资助出版。《中青文库》第一批（15 本）、第二批（6 本）、第三批（6 本）出版后，有效展示了学校的科研水平和实力，在学术界和社会上产生了很好的反响。本辑作为第四批共推出 12 本著作，并希冀通过这项工作的陆续展开而更加突出学校特色，形成自身的学术风格与学术品牌。

在《中青文库》的编辑、审校过程中，中国社会科学出版社的编辑人员认真负责，用力颇勤，在此一并予以感谢！

前　言

在 Web 2.0 时代里，互联网上存在着大量可作为情感语料数据原型的评论。如何高效精确地获取基于这些语料的情感信息，并依此进行相关研究成为当前信息科学与技术领域面临的重大挑战。情感分析，又称意见挖掘（opinion mining），是用于分析人对特定对象及其相关属性的观点、态度以及其他主观感情的技术。

为了给广大研究者提供科研参考，本书在深入研究文本情感问题及现状的基础上，充分结合计算语言学、统计学、机器学等相关理论及方法，利用语义块、句子、文本等不同语言粒度进行文本情感倾向性的建模、分析与研究，从而提出高效、精确的文本情感分类技术与方法。

本书阐述的技术可广泛应用于推荐系统、社会舆情分析、产品在线跟踪和质量评价、影视评价、Blogger 声誉评价、新闻报道评述、事件分析、股票评论、图书推荐、敌对信息检测、企业情报系统等方面。

本书可分为九个模块，从情感分析的具体流程，包括语料、文本预处理、特征选择与情感分类等为文本的情感分析提供相关的技术指导，并着重介绍了数据获取、自然语言处理、算法运用与文本情感比较等多种技巧，内容全面丰富，语言详细。

本书不但可以满足广大科研工作者、博士、硕士研究生的科学技术研究和情感语言分析的实际操作的需求，也可以作为高校研究生的参考教材，让读者在进行技术研究时得到更多的技术指导。

在本书编写过程中，感谢中国青年政治学院计算机搜索团队成员马敬贤、纪彬伟、陈思、吴佳妮等不辞辛劳地审校，感谢张洪喆、王函石、李志晓、闫晓宇等挚友对本书稿提出的宝贵建议和指导，感谢中国社会科学出版社的支持和帮助。

有关文本情感分析的知识、理论和技能还有很多，并且随着计算机

应用技术的扩展以及新兴产品的普及，情感分析所涉及的领域会变得更为宽广。本书对于目前关键性的技术内容进行研究、总结和撰写，书中难免有不妥与欠缺之处，敬请读者批评指正或提出修改建议。

朱　俭

2014 年 4 月于北京

目　录

第一章　绪论

本章从研究背景和意义入手，充分阐释了情感分析的研究现状和背景技术，以帮助读者从宏观角度理解自然语言处理的分析方法。

第一节　研究背景和研究意义

一　自然语言处理

自然语言处理（Natural Language Processing，NLP[①]）也有人称为自然语言理解（Natural Language Understanding，NLU[②]），可见对所提供语言信息进行“理解”的重要性。自然语言处理是一项非常庞大的工程，是自然科学和社会科学相交叉的学科，其所涉及的领域包括：计算机科学、语言学、逻辑学以至心理学等。自然语言处理的目的是实现计算机对语言信息的自动分析和理解，它以实验、理论和计算为三大支柱，通过对人脑及语言认知的实现途径进行模拟研究，建立起多层次网络处理模型来阐明人脑语言信息处理系统，以期待取得突破性的进展。它的研究具有很强的生命力，是当代科学新的生长点，这不仅对信息科学，而

---

① NLP：自然语言处理是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的学科。因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，所以它与语言学的研究有着密切的联系，但又有重要的区别。自然语言处理并不是一般地研究自然语言，而在于研制能有效地实现自然语言通信的计算机系统，特别是其中的软件系统。因而它是计算机科学的一部分。

② NLU：自然语言理解的关键是要让计算机“理解”自然语言，所以自然语言处理也叫作自然语言理解（Natural Language Understanding，NLU），也称为计算语言学（Computational Linguistics）。一方面这是语言信息处理的一个分支，另一方面它是人工智能（Artificial Intelligence，AI）的核心课题之一。

且对认知语言学、心理学以及对国民经济和社会的发展都会起到推动作用。

自然语言处理研究方向分为：自动阅卷（Automatic marking）、自动问答系统（Question Answering）、自动文摘（Automatic Abstract）、自动翻译（Machine Translation）、情感分析（Sentiment Analysis）。

随着计算机技术的高速发展和计算机的日益普及，为了提高阅卷效率而提出了自动阅卷的需求。自动阅卷系统的优势既体现在人力上也体现在物力上，自动阅卷系统能够自动评阅、计分、成绩存档等。可以有效地避免资源浪费，有利于环保，还大大减少人力物力，提高了工作效率。相比传统的人工阅卷，自动阅卷可以动态地管理试卷，可以当场给出成绩，较好地保证了考试的公平。如果一个学校或公司使用了该系统，老师或领导即使在外地出差也可以审批试卷，非常方便。但是现在的自动阅卷系统还不够完善，特别是对主观题的评阅还不够成熟。我们可以用自然语言处理技术对主观题的答案进行判断，与标准答案进行分析比较以提高自动阅卷的准确率。

自动问答系统是目前人工智能和自然语言处理领域中一个备受关注并具有广泛发展前景的研究对象。人们希望能够快速、准确地获取信息以满足需要。但由于用户提问的形式复杂多变，让机器理解用户问话的意思就显得非常困难。通过自然语言处理技术可以推断用户的真正询问意图，这样就提高了问答系统的准确率。如问题“中国的领土有多大?”，我们可以推断出用户是想询问中国领土的面积，这样经过筛选后的回答就不全是跟关键词“中国”“领土”相关的信息，如“中国领土争端”相关的信息，甚至是只跟“中国”或“领土”相关的信息，而这些信息不是用户想要的。因此，自动问答系统与基于关键词检索并返回有关网页、文档集的传统搜索引擎有一个重要的区别，即自动问答系统能够为用户提供真正有用的和准确的信息，这将是新一代的信息获取的理想手段。

自动文摘技术大体分为机械文摘与理解文摘两种。近 40 年以来，自然语言理解技术逐渐朝着真实语料并且实用化的发展方向前进，鉴于机械文摘非常适合在非受限领域内使用，因此得到了蓬勃的发展。但是目前这种技术受限于仅仅分析文本表层结构，因此在技术发展上常常遇到瓶颈，文摘的提取质量无法继续令研究者满意。而理解文摘应用于受

限领域，虽然领域宽度很窄，但是理解深度较高，这种方法作为理论探索有着较高的价值，但在现实生活中的实用性较低，因此，理解文摘的发展前景较为黯淡，无法应用于未来互联网上纷杂的海量数据分析。一种基于篇章结构的自动文摘算法不但可以应用于非受限领域，而且由于篇章结构远远优于语言表层结构[①]，并且这种结构能更加确切地反映文章的核心内容，因此，基于篇章结构的算法能够适应未来纷杂的海量数据分析和非受限领域，排除了机械文摘的缺点，提高了文摘的质量。如果能让机器基于篇章结构的算法在推断出短文的意思之后把短文的主旨提取，那么可想而知，自动文摘的提取质量将会得到根本性的提高。

二　文本情感分析

所谓文本情感分析（Sentiment Analysis），就是对说话人的观点、态度和情感倾向性进行分析，即分析文本中表达的主观性信息。根据立场、出发点、个人态度和喜好的不同，人们对各种对象和事件表达的信念、态度、意见和情感的倾向性不可避免地存在各种差异。在论坛、博客（blog）等反映人们观点的网络媒体上，尤其表现出了这种差异。

文本情感分析在实际生活中有着广泛的应用：

推荐系统：对产品用户的在线反馈进行自动分类和整理，分析和挑选出值得推荐的产品和服务，推荐给其他用户。

过滤系统：自动过滤一些对政府和商业机构不利的文字信息，并且鉴别出撰稿者的情感倾向、政治倾向及态度、观点和看法。如根据对文

① 所谓语言的深层结构，也就是一般所谓语言的思维形态结构。语言是附着在思维上的结构体，语言是受思维支配的，它是处在交际中的思维的载体。人作为社会的成员必然具有这个社会的思维特征、思维方式和思维风格，我们统称为思维的形态。思维形态是一种历史的产物，又是一种共时的产物，它无时无刻不在支配语言表现并模式化为语言的深层机制，这种人类所共有的、内在的、心理的东西称为语言的“深层结构”。对比语言学这种分析语言的原理首先是布龙菲尔德创立的结构主义学说，是在结构主义语言理论以及外国语教育的双重刺激下崛起的，它运用同一种原理对两种或两种以上的语言进行描述分析。我们强调对比语言的深度发展，并不是说语言的表层结构无关紧要。恰恰相反，语言表层结构的对比研究是必不可少的，因为语言的形式结构，正是其异质性的表现现象。语言的形式结构表现为其基本形式手段和句法形式手段两方面内容。基本形式手段包含语言系统、文字系统、词语系统；句法形式手段包含句法成分系统、句型结构系统、语序分布系统。而对比语言学又是一门经验学科，因此，在进行对比研究之前，我们不可能先去主观地规定出某些语言的条条框框来，而总是先去描述其中的一门语言，然后再把它跟其他语言进行比较分析，从而找出其固有的规律。

本中反映出的作者情感进行分类，对攻击政府及个人的 E - mail 可以实现自动加入黑名单的功能。

问答系统：对询问者问题中透露的情感色彩进行分析和文本分类，尽量用适合的语气回复，防止答案情感色彩出现错误而适得其反。

此外，文本情感分析还可应用于有害信息过滤、社会舆情分析、产品在线跟踪和质量评价、电影书籍评论、博客声誉、新闻报道评论、事件分析、股票评论、推荐书籍、敌对信息检测、企业情报分析等方面。

文本情感分析属于计算语言学的研究领域。研究者们以前普遍关注的是客观性信息的分析和提取，对主观信息的分析和提取研究仍处于起步阶段，还有很多问题需要进行全面的探索。这项研究涉及计算语言学、人工智能、数据挖掘、机器学习等多方面的内容。因此，文本情感分析具有重要的研究价值。

本书研究的情感分析方向是指定的语句、段落、文本等文字信息，判断文字信息所反映出来的情感倾向。在自然语言处理领域的研究中，此类问题也可以被描述为 opinion classification（意见分类）、genre classification（流派分类）、sentiment polarity（情感极性）、sentiment classification（情感分类）、semantic orientation（语义倾向）、opinion mining（观点挖掘）、opinion extractive（观点抽取）、sentiment analysis（情感分析）等，本书为了保持术语的表述一致，将此类研究问题统一描述为情感分析。

文本情感分析通常包含四个子问题：一是确定文本情感的类别有多少；二是文本的主客观的区分，即区分出文本内容是主观的评论还是客观的陈述；三是文本情感的极性分类（polarity classification），又称为正负面倾向性分类，即判别文本内容是肯定赞赏的，还是否定批判的；四是文本情感强度分类，即判定文本情感倾向性的强弱程度，如强烈贬义、一般贬义、客观、一般褒扬、强烈褒扬五个类别，这一问题通常又被称为等级推理（rating inference）。本书主要关注其中第三方面，即文本情感的极性分类。

第二节　文本情感分析整体研究现状

近年来，文本情感分析成为一个非常新颖的研究方向。一开始并没

有一个文本情感分析的评测规范对该领域的研究任务进行清晰的定义，同时也没有一个普遍接受的文本情感倾向性分析的标准语料库来支撑关键技术的研究、评测和应用系统的开发。

目前公认的关于文本情感分析的研究工作开始于 Pang 在 2002 年提出的基于文本的 N 元语法（n-gram）和词类（POS）等特征分别使用朴素贝叶斯（naive bayes）、最大熵（maximum entropy）和支持向量机[①]（Support Vector Machine，SVM）将电影评论文本的倾向性分为正向和负向两类。此外还有 Turney 在 2002 年提出的基于无监督学习[②]（Unsupervised Learning）对文本情感倾向性分类的研究。同时，他们在实验中使用的电影评论数据集目前已成为广泛使用的情感分析的测试集。

如今，国内外都已经掀起了文本情感的研究热潮，很多研究团体、科研院校、公司已经对文本情感展开了研究。根据本文涉及的研究内容，我们把这些相关的研究分为四个阶段：1. 语料阶段；2. 文本的预处理阶段；3. 特征标注与特征选择阶段；4. 情感分类阶段。

一　语料阶段

目前绝大部分语料都来自博客、专业的评论站点、新闻站点、电子商务站点。而其中影评资料、产品的用户评论、Web 2.0 博客文章是研究者的首选。

（一）评测语料

电影评论数据集[③]以及 Theresa Wilson 等建立的 MPQA 库[④]是目前研

① 支持向量机 SVM 作为一种可训练的机器学习方法基本情况，Vapnik 等人在多年研究统计学习理论基础上对线性分类器提出了另一种设计最佳准则。其原理也从线性可分说起，然后扩展到线性不可分的情况，甚至扩展到使用非线性函数中去，这种分类器被称为支持向量机。

② 无监督学习：设计分类器时，用于处理未被分类标记的样本集。目标是我们不告诉计算机怎么做，而是让它（计算机）自己去学习怎样做一些事情。无监督学习一般有两种思路。第一种思路是在指导 Agent 时不为其指定明确的分类，而是在成功时采用某种形式的激励制度。需要注意的是，这类训练通常会置于决策问题的框架里，因为它的目标不是产生一个分类系统，而是做出最大回报的决定。这种思路很好地概括了现实世界，Agent 可以对那些正确的行为做出激励，并对其他的行为进行处罚。

③ 康奈尔大学电影评论数据集下载地址是：http：//www. cs. cornell. edu/people/pabo/movie-review-data/。

④ Theresa Wilson 等建立的 MPQA 库下载地址是：http：//www. cs. pitt. edu/mpqa/。