

Users' Sentiment Analysis and Application of
Chinese Online Reviews

中文在线评论的 用户情感分析及应用

郑丽娟 王洪伟◎著

中国财经出版传媒集团



经济科学出版社
Economic Science Press

——国家自然科学基金资助项目“基于在线评论文本挖掘的线上线下服务

补救：以网络零售为例”(71701085)

教育部人文社会科学研究青年基金项目“基于网络评论的‘产品质量’

网络舆情研究”(16YJCZH159)

山东省高等学校人文社会科学研究项目“物流服务的质量评估与服务

补偿：基于情感分析的方法”(J16YF25)

聊城大学学术著作出版基金资助

中文在线评论的用户情感分析及应用

郑丽娟 王洪伟 著

中国财经出版传媒集团



经济科学出版社
Economic Science Press

图书在版编目 (CIP) 数据

中文在线评论的用户情感分析及应用 / 郑丽娟,
王洪伟著. —北京: 经济科学出版社, 2018. 11
ISBN 978 - 7 - 5141 - 9942 - 0

I. ①中… II. ①郑… ②王… III. ①互联网络 -
用户 - 情感 - 分析 IV. ①TP393.094

中国版本图书馆 CIP 数据核字 (2018) 第 251563 号

责任编辑: 凌 敏

责任校对: 郑淑艳

责任印制: 李 鹏

中文在线评论的用户情感分析及应用

郑丽娟 王洪伟 著

经济科学出版社出版、发行 新华书店经销

社址: 北京市海淀区阜成路甲 28 号 邮编: 100142

教材分社电话: 010 - 88191343 发行部电话: 010 - 88191522

网址: [www. esp. com. cn](http://www.esp.com.cn)

电子邮件: [lingmin@ esp. com. cn](mailto:lingmin@esp.com.cn)

天猫网店: 经济科学出版社旗舰店

网址: [http: //jjkxcbs. tmall. com](http://jjkxcbs.tmall.com)

北京密兴印刷有限公司印装

710 × 1000 16 开 12.25 印张 170000 字

2018 年 11 月第 1 版 2018 年 11 月第 1 次印刷

ISBN 978 - 7 - 5141 - 9942 - 0 定价: 46.00 元

(图书出现印装问题, 本社负责调换。电话: 010 - 88191510)

(版权所有 侵权必究 举报电话: 010 - 88191586)

电子邮箱: [dbts@ esp. com. cn](mailto:dbts@esp.com.cn))

前言

随着互联网的飞速发展，特别是 Web 2.0 技术的逐渐普及，广大网络用户已经从过去单纯的信息获取者变为网络内容的主要制造者。人们通过互联网主动地获取、发布、共享、传播各种观点性信息。这些观点性内容——“在线评论”对于商品推荐、舆情监控、风险控制等都具有重要的意义和价值。在线评论不同于传统的结构化的数据，表现形式大多是非结构化或半结构化的评论文本。面对海量的富含感情信息的文本，情感分析应运而生。在线评论的“情感分析”，就是利用文本挖掘技术，对在线评论进行自动分析，旨在识别用户的情感趋向是“高兴”还是“伤悲”，或判断用户的观点是“赞同”还是“反对”。引入情感分析，可以在海量的在线评论中挖掘出用户的情感特征，提升在线评论的情感价值管理。

情感分析成为国内外研究的一个热点，但现有研究还存在较多的缺失点和不足，如已标注的中文情感语料库缺乏，情感分类方法的分类准确率不高，情感词强度量化方法缺乏，情感分析技术的应用领域局限等问题。为解决这些问题，本书针对中文评论的特点，围绕“提升在线评论的分类准确率，探索用户情感及应用”这一科学问题，从情感分析技术、情感分析应用两个角度，沿着“粗粒度情感分类→细粒度情感分析→情感分析应用”主线，由粗到细、由基础研究到实际应用，对“中文在线评论的用户情感分析及应用”进行了系统的研究。其中，粗粒度情感极性分类的目的是获得针对某个产品/服务/事件的整体（粗）

观点；细粒度情感强度分析的目的是获得针对产品/服务/事件的某个属性的具体（细）观点；基于情感分析视角，采用机器学习和计量经济两种范式相结合的方法，探讨情感分析技术在金融领域的应用，可以挖掘在线评论在金融领域的价值。

本书主要有三个创新点：第一，基于句子情感的段落情感极性分类方法中，通过句子的情感极性和句子的情感贡献度来对段落进行情感分类。本书提出等权重、相关度、情感条件假设3种方法，能够根据训练语料的统计数据动态地确定段落中不同位置“句子的情感贡献度”。第二，情感本体的构建过程中，根据已有在线声誉系统特点，将用户的情感强度划分若干级别。考虑到观点词情感强度的模糊性，为每个情感强度设置隶属度函数。在此基础上，通过模糊统计确定观点词情感强度。第三，基于情感分析视角，结合机器学习——支持向量机（SVM）和计量经济——广义自回归条件异方差（GARCH）两种范式，提出 GARCH-SVM 模型。并将提取出的在线股票评论中的情感信息作为 GARCH-SVM 模型的一个输入项来预测股票价格的波动。

本书的完成，要特别感谢国家自然科学基金资助项目“基于在线评论文本挖掘的线上线下服务补救：以网络零售为例”（编号：71701085）、教育部人文社会科学研究青年基金项目“基于网络评论的‘产品质量’网络舆情研究”（编号：16YJCZH159）和山东省高等学校人文社会科学研究项目“物流服务的质量评估与服务补偿：基于情感分析的方法”（编号：J16YF25）对我们研究工作的资助。感谢聊城大学商学院（质量学院）和同济大学经济与管理学院在研究和写作过程中的大力支持，感谢一起努力过的合作者们辛勤的付出，感谢经济科学出版社编辑的辛苦工作。由于作者的水平有限，书中的不足之处在所难免，殷切希望读者不吝赐教！

郑丽娟 王洪伟
2018年6月

目 录

第 1 章 绪论	(1)
1.1 本书的写作背景与意义	(1)
1.2 本书的逻辑思路与结构体系	(6)
1.3 本书的主要工作与创新点	(10)
第 2 章 在线评论情感分析的文献综述	(20)
2.1 主客观分类	(21)
2.2 情感语料库的构建	(23)
2.3 情感极性分类与情感强度分析	(25)
2.4 情感分析的应用	(42)
2.5 本章小结	(46)
第 3 章 粗粒度情感极性分类的影响因素分析	(50)
3.1 情感极性分类流程	(51)
3.2 情感极性分类准确率影响因素	(65)
3.3 情感极性分类的准确率分析	(76)
3.4 本章小结	(95)

第4章 粗粒度情感极性分类的改进方法	(97)
4.1 基于句子情感判断段落情感极性的基本假设	(98)
4.2 基于句子情感判断段落情感极性的计算方法	(99)
4.3 段落情感极性分类的准确率分析	(103)
4.4 本章小结	(115)
第5章 细粒度情感强度分析的改进方法	(117)
5.1 细粒度情感强度分析的影响因素	(118)
5.2 “属性观点对”的提取	(123)
5.3 “观点词情感”的确定	(128)
5.4 情感本体的自动构建与细粒度情感分析	(133)
5.5 细粒度情感强度分析方法的有效性分析	(136)
5.6 本章小结	(147)
第6章 情感分析技术的应用：以金融领域为例	(149)
6.1 在线股评相关文献	(151)
6.2 在线股评数据描述	(154)
6.3 在线股评情感提取	(157)
6.4 基于股评情感的 GARCH-SVM 模型设计	(160)
6.5 基于股评情感的 GARCH-SVM 模型的预测力分析	(165)
6.6 本章小结	(171)
参考文献	(172)

第1章 绪 论

1.1 本书的写作背景与意义

1.1.1 在线评论的情感价值挖掘

伴随着互联网的飞速发展，人们在线表达自己的观点越来越容易，在线参与的意愿也越发的强烈，尤其是，随着 Web 2.0 技术的逐渐普及，网民发表在线评论的主动性和频繁性得到了进一步提高，话题范围和评论水准也比以往都要广泛和深入，广大网络用户已经从过去单纯的“信息获取者”变成了网络内容的“主要制造者”。可以说，互联网已不仅仅是网民获取信息的仓库，更是情感交流的园地，情感信息已经广泛分布于互联网的各种媒体：人们可以通过购物网站对产品/服务发表意见，还可以通过微博、论坛、社交站点，甚至是新闻评论对某事件发布自己的观点或表达自己的情感，这些反映出作者对事物、人物、事件等的个人（或群体、组织等）的见解或看法，包含作者个人情感信息的观点性内容，我们称为“在线评论”。

在线评论不同于传统的结构化的数据，表现形式大多是“非结构化”或“半结构化”的评论文本形式。面对海量的富含感情信息的文

本，人们迫切地需要找到一种快速对大规模文本进行观点分析的方法。在线评论的情感分析应运而生。所谓在线评论的“情感分析”，就是利用文本挖掘技术，对在线评论进行自动分析，旨在识别用户的情感趋向是“高兴”还是“伤悲”，或判断用户的观点是“赞同”还是“反对”。引入情感分析，可以在海量的在线评论中挖掘出用户的情感特征，从而提升在线评论的情感价值管理。

1.1.2 情感分析的理论研究不足

作为一个跨学科领域，情感分析得到学术界和企业界的广泛关注，相关的专题研讨会也在定期举办，如情感分析和智能交互会议、设计与情感国际研讨会等。这些会议聚集了管理学、信息科学、心理学、语言学等领域专家，有效推动了情感分析学科的发展。目前，已经产生一些情感分析应用系统。例如，戴夫（Dave，2003）研发世界上第一个情感分析工具，也是第一个针对产品评论区别其褒贬性的系统。伽蒙等（Gamon，2005）开发的系统可自动挖掘网上用户对汽车评价中的贬褒信息和强弱程度。刘等（Liu，2005）开发的观点挖掘系统可以处理网上顾客对产品的评价，对涉及产品各种特征的优缺点进行统计，并采用可视化方式对产品特征的综合质量进行比较。情感分析虽然成了研究的热点，并取得了一定的发展，但仍存在较多的缺失点和不足，主要有：

1.1.2.1 缺乏已标注的中文情感语料库

人工标注的情感语料库是进行情感分析技术研究的前提和基础，然而已有的已标注情感分析语料库主要集中在英文领域，包括新闻评论、餐饮评论、影视评论等方面，而中文情感分析语料库相对较少。因为缺乏公共语料库，方法的有效性难以得到验证，也难以进行研究结论之间的比较。因此，研究中文情感语料库的构建是有必要的。

1.1.2.2 中文在线评论的用户情感分析相对薄弱

现有的情感分析研究成果多是面向英文文本,由于中英文语言结构及中西方文化的差异,中文评论的情感流露方式具有特殊性和复杂性,已有的研究成果不能直接应用于中文情感分类。在中文文本领域,中文在线评论的用户情感分析相对薄弱,研究还处于起步阶段,具有相当大的研究价值和研究空间,因此针对中文在线评论进行研究是有必要的。

1.1.2.3 情感分类方法的分类准确率不高

这是情感分类存在的主要问题之一。分类准确率是情感分类效果的主要衡量因素,因此,研究如何提高准确率,准确率的影响因素有哪些等问题是有必要的。

1.1.2.4 针对不同属性进行细粒度情感分析存在不足

用户情感的表露错综复杂,常给出混合观点评论,肯定某方面,同时批评其他方面。如果将评论视为原子对象,就无法挖掘包含在评论中的细节信息。因此,对在线评论进行细粒度分析,识别在线评论的属性,并针对不同的属性进行情感分析是有必要的。

1.1.3 情感分析的应用研究不足

在线评论的情感分析对于客户管理、决策制定、舆情监控、风险控制等都具有重要的意义和价值。

1.1.3.1 客户管理

针对产品(服务)的在线评论是消费者根据自己对产品的使用情

况,从使用者的角度来描述产品属性和评价产品性能。所以与商家的促销相比,在线评论具有易存性、广泛性和非商业性,深得消费者信赖。普通消费者在购买某项产品或服务的时候,一般会参考之前使用者的评论信息,来获得对这项产品或服务的直接认识。德勤咨询调查显示,80%的消费者购物前会参考在线评论。因此,商家可以通过在线评论挖掘情感信息,以便向客户推荐商品,从而提高客户满意度。例如,根据产品的好评率向消费者推荐产品。

1.1.3.2 决策制定

商业组织同样需要在线评论。通过收集互联网上的在线评论并进行分析,能够更好地理解用户的消费习惯,可以进行更全面的客户体验管理和公司反馈管理,并且能够评价公司生产和销售策略的合理性,为公司更好地完善自己的产品,制定更符合用户的销售和生產策略提供帮助。深入挖掘在线评论,商家将更了解用户对产品的评价和需求,发现与竞争对手的差异,为产品改进、价格优化等提供决策依据。因此,在线评论已被商家视为品牌管理和网络营销的工具,其所引发的意识效应和说服效应不容小觑。

1.1.3.3 舆情监控

舆情是指在一定社会空间内,围绕中介性社会事件的发生、发展和变化,民众对社会管理者产生和持有的社会政治态度,它是人们关于社会中各种现象、问题所表现的信念、态度、意见和情绪等总和。网络因其开放性和虚拟性,已经成为民意表达的重要通道和空间。挖掘在线评论中民众的情感信息,可以更加及时地了解网络民意,获得民众的反应,为人民提供更好的服务。此外,针对突发事件,国家政府可以收集在线评论来找到最关键,民众呼声最高的问题,合理地疏导和解决,对社会的稳定发展有重要作用。

1.1.3.4 风险控制

论坛作为广泛使用的媒体,影响着人们对热门话题的观点,同时,人们参与话题的讨论,也引导着话题的发展方向。如在股票市场中,大众对股票的市场前景是否看好的讨论,会影响着更多关注者的决策行为,股评中的倾向性看法,会导致受众出现“羊群效应”,进一步导致股票的升值或贬值。因此,通过分析大众对股票是否推荐的判别,能够预测普通股民是否会决定购买该股票,从而预测到股票的走势。这种对未来趋势的预测方法除需要准确的情感判断以外,还需要更多的背景知识。

通过情感分析技术,可以挖掘出在线评论中的情感特征,并将情感特征应用到商品推荐、商业调查、风险控制、舆情监控等的相关研究中。然而在已有研究中,大都只局限于分析在线评论在电子商务(客户管理和决策制定)中的商业价值。例如,李(Li, 2010)、任(Ren, 2012)研究电子商务环境下在线评论对商家、用户的交易行为的影响。而对其他应用领域研究较少。此外,还存在研究方法单一问题。已有研究大都通过相关性分析,证明了在线评论情感和销售/定价之间是否相关,而如何采用已有情感给出未来市场的预测,则几乎是空白。因此有必要结合情感分析技术,进一步挖掘在线评论在其他领域的价值。

因此本书针对中文评论的特点,提出“提升在线评论的分类准确率,探索用户情感及其应用”这一科学问题,并设计“中文在线评论的用户情感分析及应用”这个题目。本书将对缺失的中文在线评论进行情感分析研究,并对研究过程中出现的相关问题进行探讨和创新,希望通过对上述问题进行卓有成效的研究,获取有意义的方法论和研究成果,为今后的进一步研究和实际应用提供有意义的方法和结论。

1.1.4 在线评论情感分析的研究意义

1.1.4.1 丰富了中文在线评论情感计算的理论研究体系

语言结构以及文化背景的特殊性,增加了中文在线评论情感分析的复杂性。中文在线评论的情感分析是情感计算领域的重要分支。针对中文评论的特点,提出“提升在线评论的分类准确率,探索用户情感及其应用”这一科学问题,是对情感计算这一新的研究领域的有益补充。

1.1.4.2 丰富了中文在线评论价值发现的应用研究体系

伴随着情感分析技术的逐步完善,已有学者从情感分析的角度,进行在线评论价值发现的研究。研究证明了在线评论中的确存在有价值的信息,然而该类研究刚刚起步,还存在许多不足,针对这些不足,采用情感分析技术和多种研究方法,探讨在线评论情感分析的应用价值,从而丰富在线评论价值发现的研究体系。

1.2 本书的逻辑思路与结构体系

1.2.1 本书的逻辑思路

本书从技术和应用两个角度,沿着“粗粒度情感分类→细粒度情感分析→情感分析应用”主线,由粗到细、由基础研究到实际应用,对“中文在线评论的用户情感分析及应用”进行了系统的研究,从而加强对在线评论的挖掘及其商业价值的研究。具体而言,包括以下三部分。

1.2.1.1 情感语料库构建

人工标注的情感语料库是进行情感分析技术研究的前提和基础, 因为通过“实验结果(机器标注结果)”与“实际结果(人工标注结果)”的对比, 从而分析实验所采用方法的优劣, 是情感分析领域的通用处理方式。我国的情感语料库构建处在起步阶段, 缺乏有针对性的大规模情感资源的支撑, 影响着我国情感分析技术研究的发展。因此, 我们根据国外语料库建设在收集语料、制定标注规范和质量监控等方面的经验, 首先进行了情感语料库的构建。

1.2.1.2 情感分析技术

针对情感分析目的的不同, 情感分析研究分为粗粒度(coarse-grained)与细粒度(fine-grained)两个层面。粗粒度情感分析的目的是获得针对某个产品/服务/事件的整体(粗)观点, 如某一款手机是优还是差; 细粒度情感分析的目的是获得针对产品/服务/事件的某个属性的具体(细)观点, 如某一款手机的“性能”是优还是差, “外观”是美还是丑。本书粗粒度情感极性分类主要研究如何提高情感极性分类的准确率, 细粒度情感强度分析主要研究提取产品“属性观点对”的方法, 并根据所修饰的属性词“确定观点词的情感”。通过技术角度的研究, 可以进一步完善情感分析技术。

1.2.1.3 情感分析应用

针对应用研究中的研究方法的单调性、研究领域的局限性, 采用情感分析技术和多种研究方法, 探讨情感分析技术在金融领域的应用, 挖掘在线评论在金融领域的价值。通过应用角度的研究, 可以丰富在线评论价值发现的研究体系。

具体工作如图 1.1 所示。

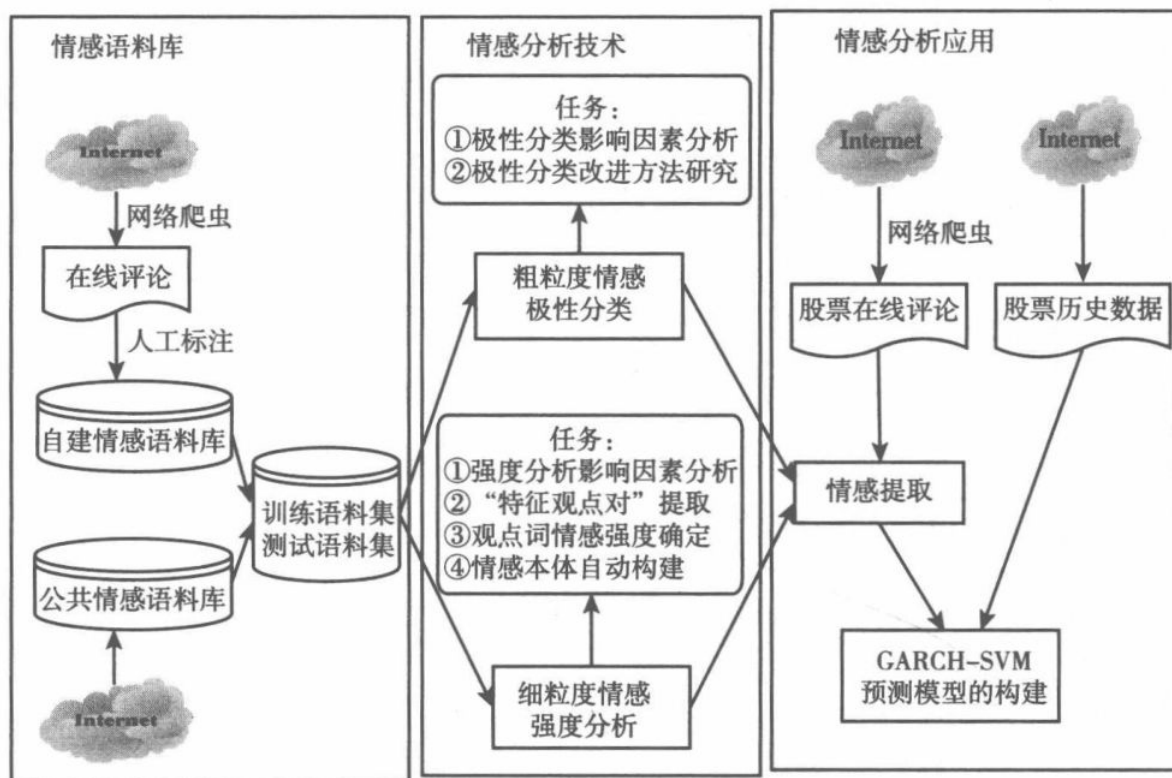


图 1.1 研究框架

1.2.2 本书的结构体系

全书共六章，第1章为绪论，第2章到第6章为本书的主要部分。第2章到第6章相互联系，构成一个较为完整的研究体系，如图1.2所示。本书的研究内容主要包括以下六部分。

第1章，绪论。介绍本书的写作背景，据此提出本书着手需要解决的问题，明确本书的研究范围；然后介绍了本书的研究思路和方法以及整体结构；最后系统地展示了本书的主要工作及研究成果。

第2章，在线评论情感分析的文献综述。本章对主客观文本分类、情感语料库构建、情感极性分类、情感强度分析、情感分析的商业价值几方面的研究现状与进展进行了总结，指出国内外现有研究的缺失点，从而为本书研究的必要性和重要性提供理论支持。

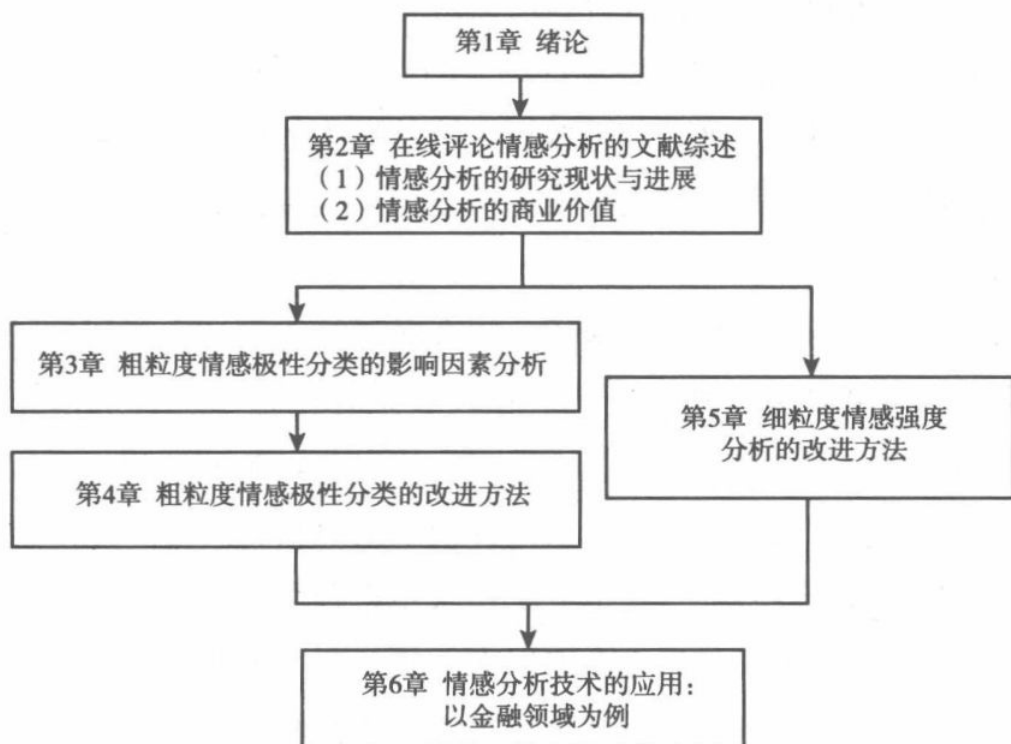


图 1.2 本书的结构体系

第3章，粗粒度情感极性分类的影响因素分析。本章首先针对本书的研究任务，设计了情感极性分类的研究流程，并提出了可能影响分类准确率的若干因素。在此基础上，结合中文在线评论的特点，选取跨度较大的两个领域“手机——产品型语料”和“酒店——服务型语料”作为实验对象，采用机器统计学习方法，通过实验，分析和比较了特征数、特征选择、特征降维、语料级别等因素对情感极性分类效果的影响。

第4章，粗粒度情感极性分类的改进方法。第3章的实验结果显示，语料级别是影响分类准确率的主要因素，即段落级语料的分类准确率远低于句子级语料。因此，第4章基于句子级评论的特点和人们表达情感的习惯，创新地提出一种基于句子情感的段落情感值计算方法，该方法结合了句子级语料的高分类准确率的特点和不同位置句子情感贡献度不同的因素，显著地提高了段落级语料的分类准确率。

第5章,细粒度情感强度分析的改进方法。本章在考虑中文网络词汇特点的基础上,通过“属性观点对”的提取和观点词情感的确定构建了情感本体。并进一步通过实验验证了该情感本体的应用,可以提高针对具体属性的分析效果,从而有效地给出用户对属性细节的满意或不满意的态度。

第6章,情感分析技术的应用:以金融领域为例。本章研究情感分析技术在金融领域的应用可行性。在线股票评论是一类特殊的在线评论,本书首先验证已有情感分析方法对在线股票评论的适用性,然后通过采用情感分析方法,提取在线股票评论中情感信息,并将提取出的情感信息作为 GARCH-SVM 模型的一个输入项来预测股票价格的波动。研究结果显示,股评情感信息在个股层面上的预测力大于在行业层面上的预测力,且预测力受到股票特征和情感特征的影响。

1.3 本书的主要工作与创新点

本书从技术和应用两个角度,沿着“粗粒度情感分类→细粒度情感分析→情感分析应用”主线,由粗到细、由基础研究到实际应用,对“中文在线评论的用户情感分析及应用”进行了系统的研究,从而加强对在线评论的挖掘及其商业价值的研究。具体研究内容如下。

1.3.1 在线评论的情感分析研究综述

通过文献分析,首先对情感分析的基础性工作——情感语料库的构建进行总结,并将情感分析归纳为三类任务:主客观分类、情感极性分类和情感强度分析,最后对在线评论情感分析的商业价值进行了总结。