基于语义约束主题模型的商品特征和情感词提取研究

彭　云　万红新◎著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

基于语义约束主题模型的商品特征和情感词提取研究

彭　云　万红新◎著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

内 容 提 要

为了更多地提取符合语义要求的细粒度特征词和情感词，改善主题模型对于中文商品评论文本语义理解能力的不足，本书提出了语义约束主题模型并进行了相关研究。首先从句法依存、词义理解和语境相关等角度获取语义关系，并将语义关系转化为 LDA 模型容易识别和方便嵌入的方式；然后在 LDA 模型中嵌入语义先验知识来影响不同层级的词语分布关系，构建语义约束 LDA 主题模型，指导 LDA 提取符合语义要求的特征词和情感词，并实现特征级别的细粒度情感分析。

图书在版编目（CIP）数据

基于语义约束主题模型的商品特征和情感词提取研究/彭云，万红新著.—北京：北京理工大学出版社，2017.11

ISBN 978-7-5682-4962-1

Ⅰ.①基…　Ⅱ.①彭…　②万…　Ⅲ.①语义学－研究　Ⅳ.①H030

中国版本图书馆CIP数据核字(2017)第273926号

出版发行 / 北京理工大学出版社有限责任公司
社　　址 / 北京市海淀区中关村南大街 5 号
邮　　编 / 100081
电　　话 / （010）68914775（总编室）
　　（010）82562903（教材售后服务热线）
　　（010）68948351（其他图书服务热线）
网　　址 / http://www.bitpress.com.cn
经　　销 / 全国各地新华书店
印　　刷 / 北京紫瑞利印刷有限公司
开　　本 / 710 毫米 ×1000 毫米　1/16
印　　张 / 8
字　　数 / 149 千字
版　　次 / 2017 年 11 月第 1 版　2017 年 11 月第 1 次印刷
定　　价 / 48.00 元

责任编辑 / 梁铜华
文案编辑 / 梁铜华
责任校对 / 周瑞红
责任印制 / 边心超

前 言
PREFACE

随着互联网和在线购物的普及，网络购物呈现出了前所未有的爆发式增长势头，购物网站上也产生了大量的商品评论文本数据。利用自然语言文本处理中的情感分析技术，可以从这些海量的文本数据中获得有用的评价知识。情感分析可以获取评价对象的情感极性分类，从粒度上包括三个层面：①文档级别的情感分析；②句子级别的情感分析；③特征级别的情感分析。文档级别和句子级别的情感分析可以获取评价对象的粗粒度情感极性，但难以满足人们进一步了解更细致的商品部件及属性评价情况的要求。要获取商品局部部件及属性的情感极性分类知识，必须对商品评论进行特征级别的情感分析，即细粒度的情感分析，其核心任务是有效提取特征词和情感词以及发现它们之间的关联性。相对于粗粒度的情感分析，细粒度的情感分析更具有挑战性。

商品评论是用自然语言表达的非结构化的文本数据，其语义关系和语法结构具有随意性，并且数据量非常庞大，给特征词和情感词的提取带来了极大的困难。综合运用自然语言理解及数据挖掘技术，在有效降低文本数据维度的基础上，才有可能实现细粒度的特征词和情感词挖掘。由于潜在狄利克雷分配（Latent Dirichlet Allocation，LDA）主题模型可以对文本数据进行降维，实现大规模文本的主题词提取，同时能利用主题聚类功能自动获取词语间的关联关系，所以LDA主题模型在特征词和情感词的提取研究中受到了极大的关注，并得到了广泛的应用。

特征级别的情感分析需要更多地发现局部结构关系中的特征词和情感词，这些词语相对于全局特征词和全局情感词来说词频更低，并且它们之间的关系隐含在句子、短语等结构中，尤其在具有复杂词语语义关系的中文商品评论中，局部特征词和局部情感词的

提取难度明显要高于全局特征词和全局情感词。现有LDA主题模型偏向于发现全局特征词和全局情感词，在主题—词语的概率分配过程中没有考虑词语间的语义关系，导致一些低频的、具有隐含语义关系的特征词和情感词提取的准确率和召回率不高。为了解决上述问题，实现细粒度的特征词和情感词提取，需要有指导地进行主题词挖掘，即利用先验知识对主题模型进行约束，形成监督效应来提取符合挖掘目标的主题词。考虑到LDA模型的语义理解能力的欠缺，首先从语义关系来探索词语间的关联性，然后利用关联性知识对主题模型形成约束机制，以更多地发现特征词和情感词之间的隐含关系。引入词语之间的语义关系约束机制可以在保留LDA主题模型的大规模文本主题词提取功能的同时，提升主题模型的语义理解能力，提高识别局部词语间关联关系的能力，以更多地提取细粒度的特征词和情感词。

本书的独创性工作主要体现在：

（1）提出了商品评论文本的词语语义关系获取方法。针对中文商品评论的特点，从句法分析、词义理解和语境相关等多角度设计了特征词和情感词之间的语义关系发现规则，并考虑语义关系作为约束先验知识加入LDA模型的方便性，获取的词语关系能够较好地反映中文商品评论中特征词—特征词、特征词—情感词和情感词—情感词之间的语义关联。

（2）设计了语义关系对LDA主题模型的约束机制。其包括两个方面：一是设计了语义关系约束下的主题—词语分配机制，实现主题下细粒度特征词和情感词的有效聚合和区分；二是设计了全局特征词主题分配约束机制，减少全局特征词对局部词语分配的干扰，尽可能多地发现局部特征词和局部情感词。语义约束可以指导LDA进行主题—词语的概率分配，影响主题下词语的聚合度和分离度，弥补LDA对于语义关系理解的不足。

（3）构建了4个带语义约束的LDA主题模型。对LDA模型进行了扩展，在语义先验知识的指导下进行细粒度主题词提取，提出了WC-LDA、AC-LDA、SRC-LDA和SWS-LDA模型。在保留LDA主题词提取特点的基础上，对LDA结构进行了改进，利用词语间的语义关系知识来指导LDA进行主题词挖掘，使得词语分配更符合商品特征和情感词提取的语义需求，提高了隐含在句式结构中的低频特征词和情感词的识别率，同时增加了词语分配的主题聚类程度，有利于发现更多的细粒度特征词和情感词以及它们之间的关联关系。

本书得到了国家自然科学基金“语义约束主题模型的细粒度商品特征和情感词提取研究”（课题编号：61662032）、江西省高校人文社科项目“语义弱监督主题模型的Web评论特征词提取”（课题编号：JC1544）和“时态化主题模型的网络舆情热点词链发现”（课题编号：TQ1505）的帮助。本书可以作为从事文本挖掘、自然语言理解专业计算机同行的参考用书。本书的撰写和出版还得益于江西师范大学、江西科技师范大学的各位专家及同事的大力支持，为此，笔者在此向其表示衷心感谢！

著　者

目 录
CONTENTS

第1章 引 言

1.1 研究背景与意义

截至2016年6月，我国网民人数达到7.10亿人，半年共计新增网民2 132万人，我国网络购物用户人数达到4.48亿人，较2015年年底增加3 448万人，增长率为8.3%[1]。随着互联网的普及和网络购物所带来的便捷性，网络购物呈现出前所未有的爆发式增长趋势，在购物网站上也产生了大量的商品评论文本数据，且日益呈现大数据化趋势。人们在进行网络购物时，通常会将以往的评论信息作为自己购物的依据。Deloitte公司的一项调查表明，82%的人在购买商品或服务时会阅读网上评论，网上评论直接影响了他们的购买决策[2]。面对如此大数量的评论文本，顾客、商家、厂家直接阅读这些评论来帮助做决定已变得十分困难，因此，对商品评论数据的自动化处理尤为重要。

要从海量的非结构化在线评论文本数据中获得有用的信息，通过人工方式进行处理的难度越来越大，人们希望通过相应的技术对这些评论文档进行自动化处理、分析，提取有用的知识。在这样的一个应用需求背景下，出现了针对文本的情感分析技术。情感分析(Sentiment Analysis)，也称观点挖掘(Opinion Mining)，主要研究和分析人们对实体对象，如商品、服务、组织、个人、问题、事件和主题及其属性所表达的观点、情感、评价和态度[3,4]。情感分析是近几年在自然语言处理和文本挖掘研究领域中兴起的研究热点，在电子商务、商业智能、信息监控和舆情分析等方面都有着重要的应用。对商品评论进行情感分析，可以获取并提供历史购物者对所购商品及其属性的情感倾向，包括正向的、中性的和负向的，从而使潜在购物者无须逐条浏览和分析历史商品评价文本，就能做出决策，大大提高了商品选择的便捷性。

在查看商品评论时，用户不再满足于商品总体性的多元情感分类判断，即商品的总体评级(分)或粗粒度的商品评价，而满足于掌握关于细粒度商品特征的情感评价情况。如购买手机，有些用户对手机的屏幕及像素有较高的要求，以满足其上下班路上观影的需要；有些用户则可能对电量、待机时间的要求较高。因此在对商品评论的分析中，需要生成基于细粒度特征的商品评论情感分析，即细粒度情感分析，以满足用户获取更细致的商品评价的需求。而细粒度情感分析的首要任务和关键技术就在于从复杂文本中有效提取商品特征及对应的情感词。在此基础上才可以进行特征级别的情感分析，从而获取各个特征上的评价信息。

商品评论是用自然语言表达的非结构化的文本数据，数据量非常庞大，需要综合运用自然语言理解及数据挖掘技术，并有效降低文本的数据表示维度，才有可能实现细粒度的特征词和情感词的挖掘。利用主题模型可以进行文本数据的降维，实现大规模文本数据的主题词提取，并通过主题聚类来获取词语间的关联关系，但其提取的主题词一般是粗粒度、全局性的特征词和情感词。细粒度特征词和情感词具有局部性特点，词频偏低且词语间关联关系不易被发现，其挖掘难度远远大于全局性特征词和情感词。

特征级别的情感分析需要更多地发现局部结构关系中的特征词和情感词，这些词语相对于全局特征词和全局情感词来说词频更低，并且它们之间的关系隐含在句子、短语等结构中，尤其在具有复杂词语语义关系的商品评论中，局部特征词和局部情感词的提取难度明显要高于全局特征词和全局情感词。现有 LDA 主题模型偏向发现全局特征词和全局情感词，在主题—词语的概率分配过程中没有考虑词语间的语义关系，导致一些低频的、具有隐含语义关系的特征词和情感词提取的准确率和召回率不高，主要表现在：

(1)难以提取低词频的特征词和情感词。商品评论的局部特征词和情感词相对于全局特征词和情感词，其词频明显偏低，包括一些次级特征词、非典型特征词及专属性情感词等。LDA 主题模型偏向发现高频的词语，导致词频相对较低的这类特征词和情感词的提取率不高，而这些局部性词语提取的准确率和召回率会对细粒度情感分析产生重要的影响。

(2)难以发现低频的局部特征词和情感词的匹配关系。特征词和情感词之间具有修饰关系，其共现频率越高，则关系越明显，但一些低频共现同样包含了特征和情感词的修饰匹配关系。由于 LDA 主题模型倾向于发现具有较高文档共现频率的词语间匹配关系，对于低频的句子级别的共现关系识别度不高，难以发现一些真实存在的但低频共现的局部特征词和情感词之间的关联性。

(3)全局特征词对局部特征词概率分配的干扰。全局特征词相对于局部特征词具有更高的词频和文档频率，容易被 LDA 主题模型以较高概率分配到不同主题，会对其他相对低频的局部特征词和局部情感词的主题分配产生影响，使这些局部词语在主题的概率分配值相对偏低。这种分配形式会造成高频全局特征词的重复提取，同时给低频局部特征词的概率分配产生干扰，从而造成局部特征词和情感词的提取率降低。

(4)难以识别特征词和情感词之间的语义关系。LDA 是词袋模型，其主题语义关系是建立在词频及文档级别词语共现的基础上，其词语间和文档间的无关性假设使得词语之间的语义关联性难以理解，包括句法关系、词义理解和上下文关联等，从而可能将文档共现频率高但无语义关联的特征词和情感词分配到同一主题，或将共现低但语义关联强的特征词和情感词分配到不同主题，造成提取的主

题词不能符合语义关系的要求。

要实现细粒度的特征词和情感词提取，需要有指导地进行主题词挖掘，即对主题模型进行约束，形成弱监督效应来提取符合挖掘目标的主题词。从语义关系的发现来探索词语间的关联性，利用关联性进一步对主题模型形成约束机制，从而发现特征词和情感词之间的隐含关系。引入词语之间的语义关系可以提升主题模型的语义理解能力，提高识别局部词语间关联关系的能力，更多地发现细粒度的特征词和情感词。

本书将从新的角度对商品评论中主题模型的作用机理进行研究，探索大数据背景下的主题模型在商品评论情感分析研究中的新途径，研究目标具体如下：

(1)提出中文商品评论文本语义知识发现和获取的规则和方法。

(2)设计语义知识对主题模型产生约束机制的思路和方案。

(3)构建商品评论文本的细粒度特征词和情感词提取主题模型。

以国内知名购物网站的实际商品交易评论文本为数据源，通过对中文商品评论的语法、语义结构进行分析，构建符合中文商品评论特点的细粒度特征词和情感词提取模型，具有以下研究意义：

(1)基于 LDA 语义提取目标，从句法分析、词义理解和语境相关等角度进行语义获取研究，拓宽了文本语义挖掘的研究视角。

(2)从提高主题模型的语义理解能力入手，对主题模型的语义化机制进行研究，拓展了用概率型主题模型进行语义提取的研究思路。

(3)针对商品评论大数据的知识挖掘要求，对语义约束主题模型进行构建研究，丰富了复杂语境下情感分析的研究方法。

同时，研究成果还可以应用于其他领域海量文本的主题知识的自动化提取、分析及处理，具有广泛的应用价值和推广前景。

1.2 国内外研究现状概述

商品评论的情感分析，主要有三个层面：①文档级别的情感分析；②句子级别的情感分析；③特征级别的情感分析。三个层面的区分依据主要来自情感极性分类对象的粒度，从文档到句子再到特征级别，其对应情感分类的细致化程度也越来越高。细粒度的情感分析更能满足用户了解商品的细节评价的需求，所以目前情感分析的研究热点主要集中在特征级别的情感分析研究上。要实现商品评论的细粒度情感分析，首先要有效提取商品特征和情感词，同时发现特征词和情感词之间的关联关系。

有很多学者已经研究了评论文本的特征词和情感词的提取方法，包括频繁名词(名词短语)及共现规则、机器学习方法和句法依存分析等。随着评论数据规模

的不断扩大，有些研究试图使用主题模型方法，在实现文本降维的同时提取主题词，并形成主题词和特征词、情感词的映射关系。由于契合了海量评论文本的特征词和情感词提取要求，基于主题模型的情感分析逐渐成为研究热点。许多研究在标准主题模型的基础上进行改进和扩展，以弱监督或半监督的形式指导主题模型进行特征词和情感词的挖掘。

1.2.1 词语频繁度及共现规则方法

在商品特征及情感词的提取中，由于商品特征通常是名词或名词短语，且特征词和情感词具有一定共现性，因此有些研究基于频繁名词、关联规则和共现规则的方法提取特征词和情感词。

Moghaddam 等(2010)[5]利用细粒度商品评论中定义的标准商品特征挖掘商品特征的出现模式，根据高频名词短语在模式中的命中率过滤高频非特征词；Hu 等(2004)[6]抽取出现频率大的名词及名词短语作为候选商品特征，通过压缩剪枝和冗余剪枝策略对提取的频繁商品特征进行筛选，抽取特征词附近的形容词作为情感词，再使用关联规则挖掘识别频繁商品特征，最后利用抽取的情感词来识别非频繁的特征词；文献[7,8]提出基于关联规则算法的产品特征挖掘算法，并结合监督型情感分析算法，实现对产品特征及其情感倾向的分析和挖掘。

Popescu 等(2005)[9]将商品特征看作商品的一部分，使用候选商品特征和领域特征之间的共现来提取商品特征，并使用点互信息(Pointwise Mutual Information，PMI)表示关联程度，最终按关联程度大小选择商品特征，该方法提高了商品特征提取的准确率，但召回率有所下降；高磊等(2015)[10]提出基于特征选择、词频和点互信息剪枝的商品属性提取方法；文献[11,12,13]利用相邻词语的共现规则来识别特征词，并提出了相应的特征筛选方法，对噪声数据进行筛除。

1.2.2 机器学习方法

一些研究采用有监督的机器学习方法来提取特征词和情感词，需要对训练数据进行标注。

吴含前等(2016)[14]提出了一种二次剪枝算法来改善序列模式挖掘算法(Generalized Sequential Patterns，GSP)在中文商品评论特征提取中准确率不高的问题；Lipenkova 等(2015)[15]提出了在汽车评价中特征级情感分析的管线模型。输入的是中文评论数据，输出的是评价和评价对象之间的关联性；Wang 等(2015)[16]提出了一个新颖的情感和特征提取模型，模型基于限制玻尔兹曼机(Restricted Boltzmann Machines，RBM)，反映了评论的生成过程，在隐藏层引

入了异质结构并加入先验知识；Madan 等(2016)[17]提出了基于贝叶斯网络的统计转化模型进行情感分类。先提取文本关键词，然后将其转换为数值形式，并用贝叶斯网络模型处理实现分类。

吴钰洁等(2015)[18]提出了一种基于概率图模型的文本情感分析方法，先通过训练语料建立先验概率图模型，用于计算词语的情感概率值，再利用信息熵将概率值归一化为情感特征值，并使用该特征值训练 SVM 对测试语料数据进行情感分类；文献[19,20,21]使用半监督的分类算法，实现特征和情感词提取，并进行正、负向情感分类；宋佳颖等(2016)[22]在模糊集框架下探索基于词语情感隶属度的情感极性分类方法。结合模糊推理的隶属度设置方法，为词语设定情感极性隶属度，获得基于词语情感隶属度的特征值表示方法；Wu 等(2015)[23]提出了利用非标注的大量微博数据的上下文信息来改善微博的情感分类。上下文知识通过监督学习算法形式化为标准词语，并利用优化算法进行模型的学习；Chang 等(2015)[24]提出了灵活的原则导向方法(Principle-based Approach，PBA)用来进行情感分类。PBA 从原始文本中获取情感模板，模板被采用来预测用户情感，并进一步辅助情感评论的生成；文献[25,26]利用聚类算法实现了特征识别和情感分类。

王荣洋等(2012)[27]基于条件随机场(Conditional Random Fields，CRFs)模型研究了多种特征及其组合在特征提取上的效果，重点引入了语义角色标注新特征，实验结果表明语义角色标注新特征对特征识别有较好的指示作用；文献[28,29,30]利用条件随机场(CRFs)模型，并结合其他方法来改善特征和情感词识别的鲁棒性；文献[31,32]将特征词和情感词的提取看作一个序列标注任务，利用隐马尔科夫模型(Lexicalized Hidden Markov Model，LHMM)进行特征词和情感词的标注；一些研究利用神经网络对特征和情感建模，并实现情感分类[33,34,35]。

1.2.3 句法依存关系分析方法

句法依存关系分析方法是从自然语言所表达的文本的句式、句法结构来分析特征词和情感词之间的匹配关系。

刘鸿宇等(2010)[36]基于句法分析获得名词和名词短语的候选特征词，然后结合 PMI 和名词剪枝算法对候选特征词进行筛选获得最终结果；文献[37,38,39,40]在使用句法分析的同时，结合 PMI 或其他共现规则方法实现产品特征的自动提取。

赵妍妍等(2011)[41]利用统计方法来获取描述评价对象及其评价词语之间修饰关系的句法路径，提出了一种基于句法路径的情感评价单元自动识别方法，并通过句法路径编辑距离的计算来提高情感评价单元抽取的性能；姚天昉等(2006)[42]基于依存句法分析总结出“上行路径”和“下行路径”的匹配规则，进而总结出主谓关系(SBV)极性传递的一些规则，用于情感评价单元的识别；一些研

究结合句法分析设置模板和规则来发现特征和情感词之间的关系[43,44,45,46]。

姚天昉等(2007)[47]利用领域本体知识来抽取语句主题以及它的属性，在句法分析的基础上识别语句的主题、主题与情感表达项之间的关系，同时利用定中关系可以查找语句中其他的主题词及情感修饰词；董丽丽等(2014)[48]首先构建面向大量商品评论的领域本体，然后利用情感词典与上下文极性计算情感词极性，并通过将本体与主谓结构(SBV)极性传递算法相结合，实现评价对象和情感词的二元组关系抽取，最后完成句子级别的情感倾向分析；文献[49,50]利用句法依存树、因子图模型实现特征发现和情感分类。一些研究利用结合分类方法、情感词典、学习框架模型和动词表达式进行句法依存分析，用来识别隐藏的特征和情感[51,52,53,54]。

1.2.4 主题模型的方法

由于商品评论是非结构的文本数据，且数据量极大，同时行文较为自由，有研究者试图利用 LDA(Latent Dirichlet Allocation)主题模型[55]的文本降维及主题聚类作用，通过提取主题词来发现特征词和情感词。LDA 是一种无监督概率生成模型，不需要进行人工数据标注，结构包括三层：文档、主题和词语。主要思想是：①文档是主题的随机混合；②主题是满足一定概率分布的词语组合。LDA 将表达文本的词向量转化为主题向量，大大降低了文本维度，同时在文本生成过程中可以提取主题词。由于 LDA 倾向于产生全局性的主题词，为了提取更多的局部主题词，一些研究对 LDA 主题模型进行了改进和扩展，再加入先验知识，形成了弱监督或半监督机制下的 LDA 主题模型。

(1)特征提取主题模型。

1)无监督主题模型。Titov 等(2008)[56]将标准 LDA 模型扩展为多粒度主题模型(Multi-grain LDA，MG-LDA)，并假设全局主题倾向于捕获商品的总体属性，而局部主题倾向于捕获用户评价的商品特征，在此基础上对全局主题和局部主题两类不同类型的主题建模，使得 MG-LDA 可以进行细粒度的主题建模；Das 等(2014)[57]主要从微博文本中发现主题短语或特征词，这些都是关注度较高的话题热点。这些主题特征词的推导可以应用于趋势发现和观点挖掘。

2)弱监督主题模型。Andrzejewski 等(2009)[58]将领域知识以 Dirichlet 森林先验的方式加入 LDA 中，提出了结合领域知识的 DF-LDA(Dirichlet Forest LDA)模型；Zhai 等(2011)[59]提出了带约束的 LDA(Constrained-LDA)模型来实现商品特征抽取及分组，设置了 must-link 和 cannot-link 两种约束类型；Bagheri 等(2014)[60]提出了基于 LDA 的特征发现模型(Aspect Detection Model Based on LDA，ADM-LDA)，关注的核心任务是如何从评价句子中提取所需的特征；马柏樟等(2014)[61]首先对网络评论文本进行分词和词性标注，得到最初的产品特

征名词集合，然后利用LDA筛选出候选产品特征词集合，进而通过同义词词林拓展和过滤规则得到最终的产品特征集；Chen等(2014)[62]针对无监督主题模型抽取的特征容易产生不一致性，在模型中加入先验知识来指导特征提取，提出了AKL(Automated Knowledge LDA)模型。先验知识的获取无须人工输入，而是自动从商品评论大数据中得到，并且来自不同的商品领域；Mukherjee等(2012)[63]在根据用户提供种子词分类的基础上，提出了实现自动对特征进行提取和分类的概率主题模型，这样可以更好地满足用户的特定需求；Chen等(2013)[64]将must-set和cannot-set引入LDA，提出了MC-LDA(LDA with M-set and C-set)模型，用于提取特征词。

(2)特征词和情感词提取主题模型。

1)无监督主题模型。Titov等(2008)[65]对MG-LDA模型进行了扩展，提出了MAS(Multi-aspect Sentiment)模型；Lin等(2009)[66]在标准LDA模型的基础上，加入了情感层并考虑每一个情感不同的特征分布，提出了JST(Joint Sentiment Topic)模型用来同时识别主题和情感；He等(2011)[67]对JST模型进行了扩展，使用词先验极性实现跨领域的特征挖掘，并通过修改主题—词语Dirichlet先验分布来加入词语的先验极性信息；Li等(2010)[68]也是在标准LDA模型中加入了情感层，提出了Sentiment-LDA模型；Jo等(2011)[69]首先提出了发现特征词的SLDA(Sentence-LDA)模型，然后将其扩展为ASUM(Aspect and Sentiment Unification Model)模型，用来发现特征词—情感词匹配关系；Moghaddam等(2011)[70]将评价文本分解为情感短语的形式，提出了ILDA(Interdependent LDA)模型，试图从情感短语中提取特征词及对应的情感词；孙艳等(2013)[71]考虑到有监督、半监督的评论文本数据集的标注工作量较大，且存在标注样本不容易获取的问题，提出了一种无监督的主题情感混合模型(Unsupervised Topic and Sentiment Unification，UTSU)，通过在标准LDA模型中融入情感来实现文档级别的情感分类。

2)弱监督主题模型。Mukherjee等(2012)[72]提出的TME(Topic and Multi-Expression)模型对评论中共现的各类情感短语和主题建模，并利用最大熵知识改善TME中的Beta先验分布的粗糙度和微弱性；文献[73,74]利用最大熵来影响LDA对于主题和情感的生成；一些研究利用贝叶斯模型[75]、变分平均场推导算法[76]、分层特征模型[77]和图学习算法[78]对LDA进行扩展，有效俘获文本中的特征和情感。

Chen等(2014)[79]提出了AMC(Automatically Generated Must-links and Cannot-links)主题模型，利用词语关系约束用来提高LDA提取特征词和情感词的效果；李丕绩等(2012)[80]对每个评价对象的所有评论文档，根据句子结构中的依存关系，构成对应实体的特征标签库，并从显式语义角度对标签库去重；

Lu 等(2011)[81]提出了 STM(Sentiment Topic Model)模型，利用极少量先验知识(种子词形式)来加强主题和特征词的直接关联性；Dermouche 等(2015)[82]提出了主题—情感(Topic-sentiment TS)主题模型，并基于 Gibbs 抽样过程进行模型参数推导；欧阳继红等(2015)[83]基于主题情感混合模型 JST 和 R-JST(Reverse Joint Sentiment Topic Model)，考虑到整体分布与局部分布的关系会影响分类效果，提出了 MG-JST(Multi Grain JST)和 MG-R-JST(Multi Grain Reverse JST)模型；有些研究利用外部线索和知识来辅助 LDA 进行特征挖掘和情感分类，将用户和产品信息加入主题模型以便更真实地捕捉用户的情感倾向[84]、加入用户行为的 AS-AC(Aspect-action)LDA 模型[85]、引入人口学知识的 USTM(User-aware Sentiment Topic Models)情感主题模型[86]和利用机器学习进行约束的 LDA 模型[87]。

1.3 研究现状评述

对上述研究方法进行梳理和分析，可以发现，基于频繁名词和关联规则的方法会造成部分低频特征词丢失，并容易产生高频的非特征词；基于数据标注的机器学习方法需要人工标注数据集，当数据量太大时，要耗费大量的人力；基于句法依存关系的分析方法，由于商品评论文本的语法结构较为随意，句法依存规则难以穷尽其句式结构关系，增加了非规范格式下的特征词和情感词关系的识别难度。

通过对 LDA 主题模型方法的研究现状进行分析，可以看到，LDA 主题模型适于提取全局特征词和全局情感词，难以满足细粒度情感分析的要求，其无监督学习方式也使提取的主题词往往难以符合预期的领域知识挖掘目标。对 LDA 主题模型进行改造，加入先验知识来提高局部主题词的发现率，是目前细粒度情感分析研究的热点和趋势。

LDA 是词袋型概率生成主题模型，提取的词语关联性主要体现在文档级别的共现，难以更深入地理解词语之间的语义关联，从而可能将共现高但无语义关联的词语分配到同一主题，或将共现低但语义关联强的词语分配到不同主题，造成提取的主题词不能真实反映特征词和情感词的关系。本书将基于大数据背景下的中文商品评论文本特点，在保留 LDA 的大容量文本主题词提取功能的基础上，从语义约束角度对主题模型进行弱监督改造，提升 LDA 对中文商品评论文本的语义理解能力，使它能够按照预定语义目标进行主题词挖掘，实现细粒度商品特征和情感词的提取。关于本书涉及的各具体问题的研究现状和相关工作，详见以下相关章节。