

# 情感倾向性分析及应用研究综述

李建华, 刘功申, 林 祥

上海交通大学电子信息与电气工程学院, 上海, 中国 200240

**摘要** 在情感倾向性分析领域, 关于情感的收集、分析和聚合等技术一直是近年来的关注热点。该领域的相关发展带动了各个子任务及其相关研究的大力发展。本文主要综述了面向情感的信息系统中使用的情感分析相关的需求、技术、应用以及评测方法等。在情报分析方面, 存在许多不同于传统的主题分析的新需求, 这就是对情感分析技术的强烈需求。接着, 介绍了词级、句子级、段落篇章级等不同层次的情感分析技术。然后, 还综述了采用情感分析技术的各种典型应用。最后, 为了工作开展的便利, 讨论了情感分析领域的词库资源、样本集资源、评测方法及重要会议等。

**关键词** 情感倾向性分析; 意见挖掘; 文本挖掘; 语言处理

**中图法分类号** TP311.13 **DOI号** 10.19363/j.cnki.cn10-1380/tn.2017.04.005

## Survey on Sentiment Orientation Analysis and Its Applications

LI Jianhua, LIU Gongshen, LIN Xiang

School of Electronic & Electronics Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

**Abstract** The field of sentiment analysis, in which sentiment is gathered, analyzed, and aggregated from text, has seen a lot of attention in the last few years. The corresponding growth of the field has resulted in the emergence of various subareas, each addressing a different level of analysis or research question. This survey covers requirements analysis, techniques, applications and evaluations that promise to directly enable sentiment-oriented information systems. In intelligence area, there are new challenges raised by sentiment aware applications, as compared to those that are already present in more traditional subject-based analysis. Techniques, such as word-level sentiment analysis, sentence-level sentiment analysis, paragraph-level sentiment analysis, are introduced in this paper. There also are several typical sentiment aware applications being reviewed. To facilitate future work, a discussion of available resources, benchmark datasets, and evaluation campaigns is also provided.

**Key words** sentiment orientation analysis; opinion mining; text mining; linguistic processing

### 1 情报分析需求

近年来, 恐怖组织将不同空间的恐怖活动发展为一种超越空间地域限制的新型恐怖主义活动形式——网络恐怖主义活动。1997年, 美国加州情报与安全研究所 Barry Collin 提出了“网络恐怖主义”一词。2000年2月, 英国在《反恐怖主义法案 2000》中第一次的明确提出了“网络恐怖主义”的概念。2001年“9.11”恐怖袭击事件发生后, 美国国会通过了《美国爱国者法案》(USA PATRIOT Act), 将“网络恐怖主义”正式列为新的法律术语<sup>[1,2]</sup>。

2001年“9.11”恐怖事件发生后, 据美国联邦调

查局(FBI)查实, 基地通过电子邮件进行联络, 借助互联网来组织实施恐怖活动。基地组织还曾使用雅虎提供的聊天、邮件、群组、论坛等各种功能进行网上交流。从2014年就名噪全球的“伊斯兰国”(The Islamic State, IS)组织最为突出的特点是善于利用先进技术来壮大自己, 其中尤为重视网络技术手段。IS依托最新的IT技术, 把网络作为宣传自身恐怖主义意识形态的工具。

近年来, 我国境内的恐怖势力也充分利用网络手段进行在线交流、网络串联、发展会员等活动。他们在网络上使用了大量的行话、黑话、暗语(上山、下山)、音译术语(伊吉拉特)等非正常的语言现象。

**通讯作者:** 刘功申, 博士, 副教授, Email: lgshen@sjtu.edu.cn。

本课题得到国家自然科学基金(编号: 61472248 和 61431008)资助。

收稿日期: 2016-06-20; 修改日期: 2016-08-30; 定稿日期: 2017-03-07

识别这些语言现象有助于识别涉恐怖信息,防止这些信息在网络上扩散,进而可以有效阻止恐怖组织的扩张。

科学技术的发展往往是一柄双刃剑。发展日新月异的 IT 技术在带给人们便捷生活的同时,也让恐怖主义者有了新的发展思路。恐怖组织肆意利用各种网络手段传播极端思想、组织恐怖活动,引发各方关注。为了对抗网络恐怖主义,学界在情报监测、网络推演等网络反恐技术方面有了不少积累。查克拉巴蒂(S.Chakrabarti)等提出了一个聚焦爬虫的超文本资源发现系统,可以对预先定义的热点事件相关的网络信息的快速提取和存储更新<sup>[3]</sup>。梅塔(Mehta R)<sup>[4]</sup>,西尼奥里尼(Signorini A)等<sup>[5]</sup>利用 Twitter API 获取了 Twitter 中热点事件的实时数据。陈(Hsinchun Chen)等以开源情报的获取和处理为基础构建“暗网”(Dark web)平台,对恐怖活动进行实时在线监控,可以实现站点内容、网络关系和活动水平的可视化。中科院自动化所的情报与安全信息学研究团队基于 ACP(Artificial societies、Computational experiments、Parallel execution)方法所构建的“天网工程”,能在线实时收集互联网上的信息,建立针对国际反恐、社会治安等的门户式平行管理系统,实时监测和评估网民群体动态变化,并对可能发生的社会事件进行预警<sup>[6]</sup>。棱镜项目是美国国家安全局(NSA)自 2007 年起实施的电子监听计划,通过监听用户的视频、音频聊天记录、照片、电子邮件、私人文档以及相关网络信息来预防恐怖活动。2013 年 6 月,美国国家安全局局长基思·亚历山大(Keith Alexander)在众议院特设情报委员会听证会时指出,通过“棱镜”等监视项目所获得的情报数据及相应分析,美国政府至少防止和挫败了 50 起恐怖袭击事件。FBI 副局长西恩·乔伊斯(Sean Joyce)也证实这其中包括挫败了纽约证券交易所的爆炸阴谋、纽约地铁的背包爆炸案、以及阻止了企图轰炸发表有关先知穆罕默德卡通漫画的丹麦报纸的阴谋等<sup>[7]</sup>。

网络反恐的难度在于从海量的网络信息(大数据)中精确(较精确)地识别出真正的涉恐信息。例如,同样是讨论“东突”的网络信息(微博、微信、tweet 等),大部分是持批判观点的信息,也有一部分是持支持观点的信息。如果不加区分地根据“东突”或其相关关键字进行处理,那么无论是支持还是批判都将被识别为可疑对象,这显然不是我们期望的结果。要区分“批判”和“支持”,也就是主题情感性识别,就需要采用不同于主题分类的研究方法。为了正确地区分对“涉恐”信息的支持或批判,识别“涉恐”信息情感倾向性是核心技术。

“涉恐”领域的信息识别是文本倾向性技术的最新应用方向,出此之外,文本倾向性分析技术已经在多个领域得到成功应用并取得了较好的成绩。这些传统领域包括:商业情报分析、政府情报分析、网络舆情分析、群体情感计算等。

本文接下来的部分将首先把国内外的情感倾向性技术进行归纳总结,并给出技术知识图谱。然后,总结了采用情感倾向性分析和意见挖掘技术为核心的典型应用(系统)。接着,概括了情感倾向性技术的相关语料库资源,会议资源等。最后,在总结部分,提出了该技术的潜在的发展方向。

## 2 情感倾向性分析技术

情感分析及意见挖掘涉及的技术领域宽泛,本文将沿着分析的粒度从小到大为主线,详细总结该领域的核心技术(如图 1 所示)。图 1 主要是给出了情感倾向性分析的两大类别。整体判断是指直接判断对象(词、句、段、篇等)整体的褒义或贬义。元素提取一般是指针对某对象的具体特征(例如,手机的价格、待机时间、屏幕等特征)进行详细的判断。

### 2.1 情感分析技术特点

随着互联网技术的发展,在自然语言处理领域出现了很多新型的实际应用,例如,主客观句的自动识别(识别一个句子是对事物或现象的客观描述还

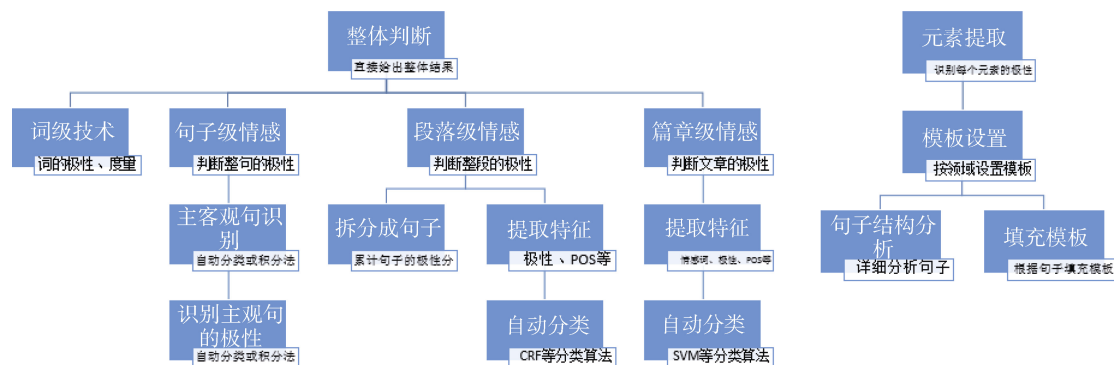


图 1 情感分析技术谱系图

表 1 传统的主题分类和新应用的区别

项目	传统分类	新的分类需求
最终目的	面向主题分类	面向情感、风格、用词等
分类依据	依据主题词(概念)	依据词法、句法、词性等
常用特征	名词、动词等特征词	形容词、助词、词性、句模等
特征表示	向量空间模型	需要更复杂的模型(张量空间模型、分域的结构表示法)
分类算法	Bayes\SVM\KNN 等传统分类算法	基于张量空间模型的分类算法、结构化模式识别算法等

是表达作者的主观情感)、文章写作风格的自动识别(判断一段文字采用何种写作风格,例如,是否采用讽刺手法)、文章作者自动判断研究(自动识别一段文字的作者是谁)、基于文本信息隐藏的自动检查研究(自动判断一段文字是否隐藏了其他信息)、面向抄袭判断的译文特征自动识别(判断一段文章是否由其他语言翻译而来,而不是母语书写)等。

传统的面向主题(subject)的分类是指把文档分到不同主题中,例如,按中图法进行分类、按路透社新闻体系分类等都是主题分类。主题分类算法需要自动判断一篇文章在描述哪个主题,而不关心其描述是支持该主题还是批判该主题。顺理成章,主题分类的特征重点考虑刻画主题的内容词,例如名词、名词短语等。因此,在主题分类领域,一篇文章通常以无序的内容词(概念)做为特征的向量,也就是所谓的词袋(bag of word)及其改进方法。

面向主题的传统分类技术和新型应用需求在多个方面有区别。经过总结,现将这两类应用的区别罗列如下(表 1 所示)。表中所提及的张量空间模型、分域的结构表示法、基于张量空间模型的分类算法、结构化模式识别算法都是较新的研究方法。

很显然这些新出现的应用需求和传统的面向主题内容的分类应用有明显的区别。这些新的实际应用不仅局限于对语义的分析,而且也需要对文本外在特征进行详细的分析。

## 2.2 情感分析类别体系

情感分类可以看作是一种两类或三类分类问题,当然,也有学者给出了更精细的分类。

三类观点分类可形式化地定义为:假设预定义三类文本集,类型为  $C=\{C_1, C_2, C_3\}$ , 其中  $C_1$  表示属于正面评价(持褒扬、支持、积极态度)的文本类型,  $C_2$  表示中立或者无关主题的文本类型,  $C_3$  表示属于负面评价(持批判、反对、消极态度)的文本类型。而要进行分类的文本集为  $D=\{d_1, d_2, \dots, d_n\}$ , 则情感分类的任务就是给文本集  $D$  中的文档  $d_i$  ( $i=1, 2, 3, \dots, n$ ) 计算一个类型标记  $C_1, C_2$  或者  $C_3$ 。

两类观点分类是对三类分类体系的简化,也就是去掉了中立或无关类别。

在心理学领域,埃克曼(Ekman)提出了六种不同的情绪:害怕、愤怒、悲伤、厌恶、愉快和轻蔑,他发现每一种情绪都与相对应的不同的跨文化的表情相联。但是,这种划分方法并没有在情感分析领域得到应用。

林鸿飞等在 Ekman 的基础上,把情感的类别体系进行了更细致的划分,最终的情感共分为 7 大类 21 小类,分别为(1)好:尊敬、赞扬、相信、喜爱、祝愿;(2)恶:烦闷、憎恶、贬责、妒忌、怀疑;(3)乐:快乐、安心;(4)哀:悲伤、失望、疚、思;(5)惧:慌、恐惧、羞;(6)怒:愤怒;(7)惊:惊奇<sup>[8,9]</sup>。

就分类体系而言,几乎所有研究论文和应用都采用了两类或三类分类体系。两类或三类分类体系不但样本构造简单,而且几乎满足现有所有的应用需求。在人工标注样本时,标注 7 个类别的样本的难度非常大。应用领域,两类或三类分类体系完全能把“涉恐”信息识别、商业情报分析提升到一个新的高度。正是基于上述原因,详细的情感分类体系没有得到广泛的应用。

## 2.3 词级的技术

词级技术的思路是用词汇(短语)的情感性来量化表示文本的情感性。这种思路的关键在于计算词汇(短语)的情感性。

Hatzivassiloglou 和 McKeown 在 1997 年首先开始词汇情感性的研究<sup>[10]</sup>。他们利用词汇之间的连词训练生成词汇间的同义或反义倾向的连接图,然后用聚类的方法将词汇聚成褒义和贬义两类。

Jaap Kamps 等国外学者提出了基于词库(例如 WordNet)的词汇相似度计算,通过分析词汇与特定基准词(如 good 和 bad)的相似度来度量其褒贬程度<sup>[11,12]</sup>。

国内方面,朱嫣岚等学者提出了基于 HowNet 的词汇语义倾向计算<sup>[13]</sup>,徐琳宏等学者也利用知网中义原标注为良和莠的词作为基准词<sup>[14]</sup>,利用刘群等人<sup>[15]</sup>的义原相似度公式计算语料中的待选词汇与基准词之间的相似度,从而得出待选词的褒贬倾向度。最后将选出的倾向度明显的词作为特征,用支持向量机进行分类,取得一定识别效果。

除了利用词典计算相似度,Turney 则提出了用词汇互信息来评价相似度<sup>[16]</sup>。也有学者在词汇情感分析的基础上,考虑了简单的句法分析(POS),从短语的角度分析情感性<sup>[17]</sup>。Turney 提出了简单的二词短语模式匹配分析,将特定的短语进行模式匹配后进行短语情感性分析,取得了不错的效果<sup>[16]</sup>。

词级技术的核心工作是为每个词赋予褒贬值, 这个褒贬值可以是二值性的, 也可以是连续的。词级技术的优点是简单易行, 缺点也非常明显: 有些词在不同语境下的褒贬性完全不同, 不能靠简单的值来表示。例如, “薪水高”和“税费高”中的“高”就表达了两种完全相反的情感。

## 2.4 句子级技术

句子级的情感倾向性分析主要有两个步骤: 首先自动识别句子的主观或客观, 然后, 再判断主观句的情感倾向性。在识别句子的情感倾向性方面, 也存在两种思路: 一是基于句子结构、词的极性等基本信息, 采用累加的方式计算出某个句子的倾向性。二是基于句子的特征, 采用自动分类的方法判断句子的情感倾向性。

### 2.4.1 主客观句判断

在文本中, 作者主要通过主观句来表达自己的情感。主观性是指作者在文本中传达了自己的立场、态度和感情, 因此主观性判断就是据此来衡量一个句子偏于主观还是客观。

识别主客观句的研究思路主要有两种: 语言领域专家制定模式和模式自动识别抽取。专家定制模式是指由语言学的专家来制定一些能够区分句子主客观性的规则, 这种方法优点是不需要大量的训练集, 但是对于不规范的文本特别是网络上的数据效果并不理想, 跨领域的合作研究也比较缺乏。模式自动识别抽取是指通过给定一些已经标注了主客观的文本训练集, 通过机器学习的方式来抽取其中的主观模式或者客观模式, 这种方法的优点是可以发掘一些人工无法识别的隐藏的模式, 并且不需要相关领域专家的参与, 但是缺点是需要先期准备大量的训练集以达到比较好的训练结果。

Hatzivassiloglou 与 Wiebe 在文献[18]中提出了从语料中学习得到形容词的语义方向, 然后将之运用于句子的主观性识别。Wiebe 随后又在文献[19]中讨论了如何选择某些词类(代词、形容词、序数词、情态动词和副词)和句子位置作为特征来实现对主观句的识别, 平均准确率达到 72.17%。Riloff 和 Wiebe 在文献[20]中提出使用 Bootstrapping 算法从主观性文本中抽取主观性名词, 在单独使用主观性名词作为特征分类的结果查准率为 77%, 查全率为 64%, 在加上预先确定的主观线索(包括来自词典和已有的研究成果)查准率将达到 81%, 查全率达到 77%。Riloff 和 Wiebe 在文献[21]中进一步提出了如何从未经过人工标注的文本中自动提取主观句的方法, 最终主观分类器的查准率达到 90.2%, 查全率达到 40.1%。而

国内在这方面也有一些初步的研究, Hu Mingqing 和 Liu Bing<sup>[22]</sup>将形容词作为主客观句的分界线, 当句子中同时包含形容词和所检测的评价主题时, 即认为该句为主观句。Hong Yu 等人研究了贝叶斯分类器算法在区分主客观类型的文本的应用<sup>[23]</sup>。该文章则以词汇、二元组、三元组、词性等作为特征, 训练朴素贝叶斯分类器, 期间为了提高训练集的准确率, 采用多重贝叶斯分类的方法, 达到较好的效果。

### 2.4.2 句子情感倾向识别

在句子情感分析领域同样存在两种方法, 基于句子结构分析的方法和基于有监督的机器学习方法。

基于句子结构和句法分析的方法简便有效, 是早期经常被采用的方法。Yu 和 Hatzivassiloglou<sup>[24]</sup>基于词典的方法, 他们首先计算句子中所有词 LLR 分数的平均值, 然后通过设定阈值判断句子的情感倾向性。薛丽敏<sup>[25]</sup>等人提出了基于句法分析的 5 种结构用于情感倾向性识别。句法分析方法的优势在于处理复杂语句时, 能够很好地定位到情感词语以及其修饰副词与评价对象的关系。宋艳雪等<sup>[26]</sup>采用语义分析的方法, 在构建的语境歧义词搭配词典基础上对句子进行情感倾向性分析。韦向峰等<sup>[27]</sup>把影响倾向性分析的词语分为 4 类: 对象词、褒贬词、逻辑词和程度词, 建立了语句倾向性分析的二元模型和三元模型, 在语句语义块分析的基础上实现对语句和篇章的倾向性获取。

由于一个句子可能包含多种观点或包含多子句, 因此, 使用简单的句子分析法困难重重。基于有监督的机器学习是研究者常用的另外一种方法。Wilson 等人提出了机器学习的自动情感分类问题<sup>[28]</sup>, 把句子情感倾向性分析转化为了分类问题来处理, 取得了很好的效果。近年来, 斯坦福大学自然语言处理小组致力于利用深度学习模型(Deep Learning)构建句子表示<sup>[29,30]</sup>, 从而便于进行句子级的情感分析。国内的王根、赵军等人使用 CRF 和冗余标签对句子序列进行情感倾向性标注<sup>[31]</sup>, 这种方法的缺陷在于实现需要给予人工标注的语料进行训练, 需要消耗大量人力。

另外, 针对特殊的句式也有学者专门开展了研究。邸鹏<sup>[32]</sup>等人提出了一种专门对转折句式进行有效情感分析的方法, 文中详细分析了转折、并列、选择、因果等 8 种句式及常用的转折词。陈豪<sup>[33]</sup>在基于句法分析的商品情感倾向性分析考虑了否定句的具体影响, 分析了否定句的具体结构。

句子即不像词那么琐碎, 又不像篇章那样由于包罗太多的内容而难于把握, 因此, 句子是识别情

感倾向性的较理想单位。基于句子结构分析的方法和基于有监督的机器学习方法是识别句子倾向性的两类常用方法, 它们各有长处。前者需要依赖对语言本身的理解, 很难跨语种使用; 后者通用性较强, 但需要给出合格的成熟样本。

## 2.5 篇章(段)级技术

在识别篇章(段)级的情感倾向性时, 国内外学者采取了各种不同的研究思路和方法, 也取得了较大的进展。按照研究的层次来分, 这些方法可分为 3 大类: 基于词汇的情感量化进一步量化文本的情感; 通过分析句子的语法规则, 从语义的角度量化文本的情感性; 通过数据挖掘的思路来分析文本的情感性<sup>[34]</sup>。

### 2.5.1 文本表示模型研究

原始文本必须转换为计算机能够识别的格式, 才能进行处理分析, 常见的表示元素为字、词、短语、N-Gram 等, 并将这些元素构造为特征向量结构, 然后提交经典的基于机器学习的自动分类方法。下面主要介绍常用的一些文本表示模型。

布尔模型(Boolean Model)是基于布尔代数与集合论的一种常见的数学模型。布尔模型既可以采用 0 或 1 为每个元素赋值, 也可以是介于 0 和 1 之间(包括 0 与 1)的数值。比如 Lee 提出的 fuzzy set, Waller-Kraft, P-Norm 与 Infinite-One<sup>[35]</sup>, 以及 Salton 等提出的 p-norm 模型<sup>[36,37]</sup>。

向量空间模型(VSM - Vector Space Model)将文本表示为一个向量, 其中向量中的每个元素分别表示一个特征项。在文本分类中, 特征常常由字、词、短语、N-Gram 等组成。TF\*IDF 法是最经典的一种特征权重计算方法<sup>[38]</sup>。

概率模型(Probability Model)是另一个比较成熟的模型, 它已经在很多系统中应用并且取得了不错的成果<sup>[39]</sup>。概率模型有自己的假设前提和比较完整的推理过程, 用概率的高低来计算文本间的相似度大小<sup>[40]</sup>。

语言模型是一种比较新型的模型, 这种模型也是基于概率和统计的一种数学模型。语言模型的思路是, 将文本视作由不同字符组成的一个概率分布, 要判定某一句子、段落或者文章是否符合该类别, 可以由概率分布来计算得出。也即是说, 特征集合在该文本中形成了一个概率分布, 这个概率分布就被称之为一个语言模型。当建立了所有待计算的文本的语言模型之后, 一种常用于计算距离的方法是计算 Kullback-Leibler 距离, 也被称之为 K-L 距离<sup>[41]</sup>。

$$KL(A, B) = \sum_{w_i \in A} P_A(w_i) \log \frac{P_A(w_i)}{P_B(w_i)}$$

在上述各个模型中, 布尔模型由于其简单且易用的特性较为实用, 因此在商业上比较常用。但是从学术角度上而言, 向量空间模型以及概率模型因其形式化备受推崇。语言模型虽然较新且发展时间最短, 尚待进一步的研究, 但是它能很合理地表现文本信息, 有比较好的发展前景。四种文本模型的区别和比较如表 2 所示。

表 2 四种文本表示模型对比表

模型	布尔模型	概率模型	向量空间模型	语言模型
提出时间	20 世纪 50 年代	20 世纪 80 年代	20 世纪 60 年代	20 世纪 90 年代
理论基础	集合论	概率论	代数理论	概率论、随机过程
实现难度	简单	较难	简单	简单
表示方法	词向量	词向量	词向量	N-Gram
商用情况	采用	采用	常采用	未采用

对于中文文本而言, 词并不是特征表示的唯一单位, 在词与词之间本身并没有分割界限, 这也说明中文文本的特征可以表示成多种粒度。最常见的中文文本的特征表示粒度包括字、词、词组、概念、N-Gram 和 N-POS 等几种<sup>[42]</sup>。由于前几种非常容易理解, 接下来简单概括一下后两种粒度。

N-Gram 以连续出现的 N 个单位作为一个特征<sup>[43]</sup>, 这种表示方法容易在各种领域应用。在文本处理领域没有字、词和词组的语义上的限制。中文词汇大多是由两个单字组成, 因此 2-Gram 是一种最常见的表示方法<sup>[44]</sup>。

当处理传统的主题分类问题时, N-Gram 特征表示法的缺点也很明显: 它会将毫无意义的虚词、无语义的字的搭配作为特征。这些只是单纯的数据组合, 并不能着重体现出文本原有的含义, 因此引入了很多噪声, 且效率会随着 N 的上升快速下降。

N-POS 的概念类似于 N-Gram 的概念<sup>[45]</sup>, 它是指连续 N 个相连的词性。N-POS 可以通过分词并标注词性后得到, 比如句子“今天天气不错。”, 经过分词后得到“今天/t 天气/n 不错/a . /wj”。

1-POS 表示为: (t), (n), (a), (wj)

2-POS 表示为: (t, n), (n, a), (a, wj)

3-POS 表示为: (t, n, a), (n, a, wj)

相比于 N-Gram, N-POS 明显有着更低的特征空间, 在训练阶段所需要的训练集数量也相应大大减少, 不过这两种特征表示方法有不同的应用场景。

N-Gram 主要用于分词及新词识别, N-POS 主要用于文本主观性判断中。在观点句和非观点句中, 遭

词造句方式有所不同, 比如(副词, 形容词), (形容词, 助词)这样的组合在观点句中更多被使用到。

## 2.5.2 自动分类研究

文本自动分类就是将目标空间中未知类别的文本归于指定的类别空间, 这一过程需要构造一个分类函数或者分类模型, 这也是文本分类算法的核心<sup>[46]</sup>。现在常用的基于向量空间的分类算法有以下几种: 简单向量距离(Vector Distance)法<sup>[47]</sup>、K 最近邻算(K-Nearest Neighbor)法<sup>[48]</sup>、贝叶斯(Bayesian)算法<sup>[49]</sup>、决策树(DecisionTree)法、支持向量机(Support Vector Machine, SVM)<sup>[50]</sup>、条件随机场(CRF - Conditional Random Field)和一类支持向量机(One-Class SVM)等。下面将介绍几种新近出现的分类方法:

一类支持向量机算法<sup>[51]</sup>, 即 one-class SVM。该算法作为支持向量机的一个重要分支算法, 主要针对训练样本比例严重失衡的情况下设计的。一类支持向量机在解决小样本聚类学习、非线性问题、异常值检测等无指导领域取得了广泛的应用。

条件随机场(CRF - Conditional Random Field)在数据分段、序列标注、命名实体识别、中文分词、短文本分类<sup>[52]</sup>等自然语言处理任务中都有很好的表现。CRF 使用条件特征<sup>[53]</sup>, 可以对特征进行全局归一化。它不是在给定当前状态的条件定义下一个状态的分布, 而是在给定需要标记的观察序列的条件下, 计算整个标记序列的联合概率, 从而避免了HMM 的对数据进行不必要的独立性假设。

深度学习(Deep Learning)是机器学习领域的算法集合, 它使用多层非线性结构对数据进行建模<sup>[54]</sup>。深度学习常被当作是实现深度人工智能的一个方法, 许多公司在解决实际问题时经常采用这种方法<sup>[55]</sup>。Facebook 最近宣布在加利福尼亚、伦敦和纽约成立了 AI 实验室。这些实验室的主要目的就是采用深度学习方法对用户上传的照片进行标注<sup>[56]</sup>。

基于高阶张量的分类算法。在数学领域, 张量(Tensor)是对向量的高阶扩展。张量在数据分析中的应用也非常广泛, 早期主要用于心理、化学领域的建模。近年来, 在机器学习、数据挖掘等领域的研究和应用中, 越来越多的大规模数据和高阶数据对研究者提出了新的挑战, 张量模型逐渐被作为新的工具引入到图像、视频等领域<sup>[57]</sup>。例如, 在分析视频数据时, 二维的图像数据, 再加上时间维度一起组成了三阶张量<sup>[58]</sup>。

采用基于机器学习的方法对实践中的问题进行分类已经取得了很大成绩, 因此, 不能简单地说那个分类算法不好, 关键是如何针对实际需求选择特

征, 以及针对实践问题选择合适的算法。单一的分类器往往不能在有效性和时间空间复杂度间找到一个好的平衡点, 准确度和高效率兼得的分类器尚待研究发现。

## 2.5.3 基于机器学习的篇章情感分析

把有监督的机器学习用于篇章情感分析时, 研究者的工作主要体现在特征选择和分类算法选择两个方面。

有部分学者在针对倾向性分析进行特征选择时, 参考了句法结构信息, 例如, Gamon<sup>[59]</sup>在 unigram + bigram + trigram 项的基础上加入句法结构相关特征后, SVM 的分类精度有所提高。Ng 等<sup>[60]</sup>发现将根据依存关系抽取的特征加入 bigram + trigram 项不能明显改善分类效果, 但如果仅有 unigram 项, 使用依存关系特征有助于提高分类精度, 并认为原因是依存关系特征与 bigram /trigram 可以获得相近的信息。

许歆艺等<sup>[61]</sup>在识别篇章情感时, 提出了“文本纹理”特征的新概念。其特征主要包括句式、修辞、词语间依赖关系等在内的文本纹理, 以情感词汇、评价词汇、语气词以及部分文本纹理为基础特征, 并根据所识别的文本纹理调整基础特征的权重。在特征的基础上, 采用了 SVM 作为分类器, 取得了非常好的效果。

近年来, 由于深度学习算法和计算框架的崛起, 该方法也被很多学者引入到了情感倾向性识别领域。Glorot 等<sup>[119]</sup>首次将深度学习应用到情感倾向性识别领域, 该文采用早期的堆叠式 Auto-encoder 自动获取文本的特征, 在获取特征后采用经典的 SVM 监督学习算法自动识别倾向性类别。Zhou 等提出了基于动态深度学习方法半监督情感倾向性分类算法, 该算法主要的目的是解决情感分类领域样本不足的问题<sup>[120]</sup>。此外, 针对段文本、Twitter 帖子等特殊对象, 也有学者采用深度学习算法进行研究并获得了较好的效果<sup>[121-123]</sup>。

## 2.6 对象特征项的情感挖掘

细粒度的对象抽取一般是用在意见挖掘, 特别是针对商品或服务多个组成特征分别进行分析, 目的是获取评论者评价对象的所有特征的逐项情感。很显然, 对同一对象的多个特征, 不一定全持否定态度, 也不一定全持肯定态度。

Kim 等<sup>[62]</sup>认为意见由四个元素组成, 即特征、持有者、陈述、情感。其中, 意见的持有者针对某主题发表了具有情感的意见陈述。意见挖掘的工作就是自动确定这些元素以及它们之间的关系, 其主要子任务包括:



a) 产品特征抽取: 识别特征术语和指派领域相关的概念。

b) 识别评论持有者: 确定评论表述的作者。

c) 确定陈述部分: 确定意见表述的范围和过滤客观性表述。

d) 情感分析: 在 c) 的基础上, 判断其语义倾向。

对于上述 4 个子任务, 需要分别处理。其中, d) 采用的技术和句子、段落级情感分析是相同的。接下来, 将总结其他几个部分的研究现状及成果。

### 2.6.1 产品特征抽取

在抽取产品特征时, 研究者分别采用了专家总结和自动学习抽取两大类思路。

姚天昉等<sup>[63]</sup>通过人工定义的汽车本体抽取中文汽车评论的产品特征。Carenini 等<sup>[64]</sup>利用领域知识建立了一个产品特征的层次模型, 然后利用关联规则挖掘产品特征, 并将挖掘到的产品特征通过相似性映射到该产品特征层次模型上。

黄永文<sup>[65]</sup>则采用了自动抽取的方法。他首先定义了一些常见的产品特征和观点词, 然后利用这些产品特征和观点词的语法联系抽取特征和观点的表达模式, 并采取 Bootstrapping 方法迭代抽取新的产品特征和观点词以及新的表达模式。Somprasertsri<sup>[66]</sup>等标注了评论训练集中的产品特征, 利用句子中词周围特定窗口中的词、词的词性等语法特征构造了一个极大熵的分类器, 然后抽取所有的名词和形容词作为候选产品特征, 并用分类器来判断该候选产品特征是否是产品特征。

### 2.6.2 评论持有者

由于人名或机构名都是可能的意见持有者, 因此, 识别评论持有者可采用命名实体方法<sup>[62]</sup>。当然, 也能够通过模式识别和机器学习技术解决<sup>[67]</sup>。文献[68]把意见持有者与真实世界的实体关联起来, 根据选择提到相同实体的意见持有者来解决互指问题。

### 2.6.3 确定陈述部分

研究者主要利用产品特征和观点之间的关系来抽取陈述部分, 也有些是利用人工定义的字典去抽取陈述部分。

Popescu 等<sup>[69]</sup>根据产品特征和观点词往往在句子中同现的规律, 人工定义了观点词和产品特征以及相关元素之间的语法规则, 利用已挖掘出的产品特征结合语法规则抽取观点。Hu 等<sup>[70]</sup>认为观点主要由形容词描述并且应该和产品特征邻近, 因此要抽取产品特征邻近的形容词作为观点。文献[68]采用了细颗粒度陈述范围确定, 从而可以抽取主题和陈述之间紧密的关系, 如定义包含形容词或副词的名词

短语作为陈述。

## 3 典型应用

意见挖掘和情感分析技术已经被应用到很多实践活动中。例如, 针对评论的意见挖掘和情感分析、商业或政府情报分析以及网络空间的情感计算等。

### 3.1 商业情报分析系统

早期的情感分析技术非常适合用于各种类型的情报分析应用中。事实上, 商业情报是推动情感分析和意见挖掘技术进步的主要推动力。常见的做法, 网络商家会要求他们的客户在线评价所购买的产品。随着越来越多的人利用网络来表达意见, 一种产品被评价的数量快速增长。利用情感分析技术, 商家能从这些评论信息中自动分析出那些是“推荐”或“不推荐”的评价<sup>[71]</sup>。

考虑如下应用场景<sup>[72]</sup>。某智能手机制造商的手机销量比预期低很多, 因此, 就会设法知道原因。途径之一就是通过对用户的个人评论信息中挖掘。影响销售的客观因素可能包括手机太重、带机时间短、价格太贵等。主观评价包括: “手机设计太俗气了”、“客服非常差劲”等。

文献[73]提出一个新的通过消费者评价分析和比较竞争对手的产品的框架。该文实现了一种称为意见观察员(Opinion Observer)的原型系统。系统主要包括两个任务: 一是确定产品特征。基于语言模型, 把消费者的意见和产品特征对应起来。二是, 针对产品的每个特征, 确定该特征是正面还是负面的。

文献[74]根据用户使用产品的时间长短和使用经验等对不同的评论赋予不同的权值, 针对用户的查询, 基于用户评论给出了一个产品推荐系统。文献[75]计算评论中产品特征出现的次数作为权值, 认为一个评论中某特征出现的次数越多则相应权值越高, 并在此基础上给出了一个产品推荐系统。

商业情报分析系统能够通过情感分析与意见挖掘自动统计汇总网络用户的意见, 形成强有力的反馈, 为产品的改进提供支持。典型的商业情报系统有 Epinion([www.epinion.com](http://www.epinion.com))等。

### 3.2 政府情报分析系统

政府情报分析系统是情感倾向性分析的另一个典型应用。例如, 可以通过这类系统监控不断上升的敌意或负面通信<sup>[76]</sup>。

文献[77]尝试通过意见挖掘技术, 在美国国会的辩论文稿中, 自动分析发言者对所提的法律条文是支持或是反对。文献[78]针对政治讨论信息提出一个基于统计的情感分析系统, 并通过该系统分析是否

存在着对主贴的反面评论。

话题检测与追踪(TDT)是一项面对新闻媒体信息流进行未知话题识别和已知话题跟踪的技术,主要应用于从海量信息中发掘热点敏感问题,并追踪其趋势变化<sup>[79]</sup>。话题检测和跟踪技术的研究始于1996年,James Allan<sup>[80]</sup>等人通过跟踪信息的时空进程来从海量信息中检索感兴趣的信息。

对TDT的研究主要分为数据表示模型和话题聚类两方面:数据表示模型方面,James Allan<sup>[80]</sup>和Schultz<sup>[81]</sup>采用向量空间模型描述特征空间,根据特征在文本中的概率分布估计权重,利用余弦夹角衡量报道间的相似性。Yang, Steinberger等人<sup>[82-84]</sup>使用命名实体来表示话题和新闻,对不同的命名实体种类分配不同的权重,其权重计算方法各不相同。在话题聚类方面,应用最广泛的当属single-pass聚类算法<sup>[85]</sup>,基于短语的聚类算法主要有文档索引图算法<sup>[86, 87]</sup>和后缀树算法<sup>[88]</sup>,上述算法的共同之处在于匹配任意两个文档之间的共享短语。

### 3.3 群体情绪或情感计算

在情感分析和意见挖掘技术的支持下,结合互联网大数据分析技术能够计算某些群体甚至整个社会的情绪或情感。经过量化处理后的情绪和情感就是情绪指数或情感指数。

在网络空间的社会情感方面,目前的研究不是特别多,见诸于报端的有哈尔滨工业大学的微博情感指数系统和刘翠娟等提出社会情感可视化计算方法<sup>[89]</sup>。该文给出微博文本的情感计算模型,提出一种基于依存句法和人工标注相结合的情感分析方法,并以三维立体的人脸表情呈现群体情感。

经济的推动力是巨大的,因此,有大量的研究集中在投资者情绪(investor sentiment)计算方面。投资情绪是反映投资者心理的重要因素,它是反映投资者的投资意愿或预期的市场人气指标,对证券市场的运行和发展有很大的影响。文献[90]首先通过知网文库、微博舆情、百度推荐确定关联性关键词,在此基础上,通过一定的筛选方法获得领先于我国股市行情的关键词,最终基于这些领先性关键词数据构建可量化的投资者情绪指数,以此建立可跟踪的投资者情绪的动态监测机制。

### 3.4 网络舆情系统

观点或主张是人们对某一问题所持有的看法,而舆论,即大众观点,意味着人群中的全部或绝大部分人所达成的主张或立场。由于存在众多的因素影响大众观点的形成过程,舆论最终可能会形成,也可能无法形成<sup>[91]</sup>。研究舆论形成过程的目的,重点是

研究群体中的观点形成过程、观点两极分化或观点分裂<sup>[92]</sup>。

舆情系统是综合监测网络社会观点形成的重要工具,它综合运用搜索引擎技术、文本处理技术、知识管理方法等,通过对互联网海量信息自动获取、提取、分类、聚类、主题监测、专题聚焦,以满足用户对网络舆情监测和热点事件专题追踪等需求。

目前,在我国市场上有较多的舆情系统,影响力较大的包括TRS舆情系统、百度舆情监控、中科天玑、军犬舆情系统、上海交大-鹏越惊鸿舆情系统以及来之国外的Autonomy等。为了准确地把握网上的观点,这些系统无一例外都采用了情感倾向性分析或者观点挖掘技术。

### 3.5 其他应用

除了上述几个典型的系统之外,情感分析及观点挖掘技术还是以下几种应用的重要模块。它们包括如下几个典型应用:

#### 1) 推荐系统

在推荐系统中,通过情感倾向性技术可以识别并过滤掉那些受到大量负面评价的物品,避免将这类物品推荐给用户<sup>[93,94]</sup>。

#### 2) 信息抽取系统

在信息抽取系统中,通过情感倾向性技术可以帮助避免抽取主观句中的信息,而主要从客观句中抽取信息<sup>[95]</sup>。

#### 3) 问答系统

在问答系统中,用户倾向于得到系统对某个问题的观点<sup>[96]</sup>。通过情感倾向性技术可以帮助过滤掉非观点问题。

#### 4) 自动摘要系统

自动摘要系统需要自动综合来自不同数据源的观点,并加以综合<sup>[97]</sup>。通过情感倾向性技术可以帮助过滤非观点问题。

#### 5) 学术声誉系统

学术声誉系统需要明确,当作者引用某个文献时的意图。是用来支持某个论点?还是仅仅完善其文献完整性<sup>[98,99]</sup>?

#### 6) 信息过滤系统

对于讨论同样的一个主题的文本,可以选择性地过滤支持该主题的部分或者反对该主题的部分<sup>[100]</sup>。这正是情感倾向性技术的作用所在。

## 4 数据资源及评测

一般来说,不同实验之间的可比性需要满足如下几个条件:1)采用同样的数据集合;2)用同样的比



例切分训练数据和测试数据; 3) 用同样的评价方法。由此可见, 公开的测试数据, 测试方法对提高该领域的研究是至关重要的。

与自然语言处理、信息检索和数据挖掘领域的其它任务类似, 基准语料库和公共评测对于推动文本倾向性分析技术的发展有着至关重要的作用。本小节将分别介绍一些较有影响力的基准语料库和公共评测。

#### 4.1 测试方法

如果把情感分析看成一个自动分类问题, 则可以用评价自动分类的方法来评价情感分析的准确性。文本分类的有效性标准包括查准率(Precision)、查全率(Recall)和 F 值(F-Measure)三部分。这三个评价指标都是针对某一类别的分类指标。

##### 1) 查准率

查准率  $P$ , 又称准确率, 衡量的是所有被分类器分为类别  $C$  的文本中正确文本的比率。公式为:

$$P = \frac{\text{正确判为类别 } C \text{ 的文本数}}{\text{所有被判为类别 } C \text{ 的文本数}} \times 100\%$$

##### 2) 查全率

查全率  $R$ , 又称召回率, 衡量的是所有实际属于类别  $C$  的文本被分类器分到该类别中的比率。公式为:

$$R = \frac{\text{正确判为类别 } C \text{ 的文本数}}{\text{所有属于类别 } C \text{ 的文本数}} \times 100\%$$

##### 3) F 值

查准率和查全率是对分类结果的两个不同的侧面作出的评价, 在通常情况下, 这两个指标是互补的, 是两个互相矛盾的衡量指标: 查全率会随着查准率地升高而降低, 两者不可兼得。所以产生了将这两者综合考虑的指标: F 值, 公式为:

$$F_{\beta}(P, R) = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

其中  $\beta$  是一个调整参数, 引入  $\beta$  的目的是以不同权重综合查准率和查全率。通常取  $\beta = 1$ 。此时 F 值的公式为:

$$F_1(P, R) = \frac{2PR}{P + R}$$

为了准确地评价某一分类器整体上的有效性, 需要将各类别的有效性指标综合起来, 可以用分类精度(Accuracy)来衡量。

精度计算公式为:

$$\text{Accuracy} = \frac{\text{分类正确的文本数}}{\text{测试文本总数}}$$

上述的这几个评价指标的值越高, 说明文本分类系统的有效性越好。

在情感分析领域, 研究者们基本都采用这几个指标来评价其算法的有效性。例如, Leshed<sup>[101]</sup>等就用准确率和召回率评估其提出的博主情绪识别算法, Ye 等<sup>[102]</sup>也用这两个指标评价论坛帖子的正反识别算法。除了准确率和召回率指标外, 也有研究者提出了另外的评价方法, 但这些方法使用的范围非常窄, 有兴趣的读者可以阅读 Mishne 等<sup>[103]</sup>提出的关联因素和相对错误评价方法。

#### 4.2 数据资源

国内外研究者们已经针对文本情感倾向分析的不同任务、不同语种构建了多个公开的语料库。

##### 4.2.1 情感词资源库

在词语倾向性分析研究方面, 主要是公布了较多的用于情感分析的词库。它们包括如下这些。

Hatzivassiloglou 和 McKeown 与 1997 年发布的英文 HM 情感词库包含 1336 个形容词(657 褒义, 679 贬义)<sup>[104]</sup>。

Stone 等人在 1966 年给出 GI 英文词表是从 General Inquirer Lexicon<sup>[105]</sup>抽取而来, 它包含 3596 个形容词、副词、名词和动词, 其中, 1614 个褒义词, 1982 个贬义词。

在非中文词库方面还有 SentiWordNet<sup>[106]</sup>, MPQA<sup>[107]</sup>, LIWC<sup>[108]</sup>等, 这些词典的特点各不相同, 为研究文本情感分析问题提供了巨大的帮助。

在中文词语倾向性方面, 上海交通大学姚天昉等在 2007 年构建了包含强度信息的汉语情感词词典, 包含 3120 个褒义情感词和 3485 个贬义情感词<sup>[109]</sup>, 为中文词语级倾向性分类的研究提供了帮助。

中文情感词汇本体库是大连理工大学信息检索研究室在林鸿飞教授的指导下经过全体教研室成员的努力整理和标注的一个中文本体资源<sup>[110]</sup>。该资源从不同角度描述一个中文词汇或者短语, 包括词语词性种类、情感类别、情感强度及极性等信息。词汇本体加入情感类别“好”对褒义情感进行了更细致的划分。最终词汇本体中的情感共分为 7 大类 21 小类。该词汇库含有情感词共计 27466 个, 情感强度从弱到强分为 1、3、5、7、9 五档分值。

《知网情感分析词语集》包含了中英文的主张词语、正面情感词语、正面评价词语、负面情感词语、负面评价词语、程度级别词语共 12 个词语列表。其中, 中文的正负面情感词语和正负面评价词语列表, 共 8936 个词语。中文的程度副词列表, 共 219 个词语<sup>[111]</sup>。

##### 4.2.2 句子级的数据集

2004 年 Pang 等提供了共计 10000 个句子, 并包

含倾向性信息的标注语料, 该语料库广泛应用于句子和篇章级倾向性分析中。

在中文句子资源方面主要是来源是中文信息学会主办的评测会议(COAE2008 - 2015)系列会议。以及上海交通大学信息安全工程学院团队构建的中文句子主客观句句子库(共 10000 句, 主观和客观句各 5000)<sup>[112]</sup>。

#### 4.2.3 篇章级的数据集

Internet Movie Database(IMDb, <http://reviews.imdb.com/reviews/>) 数据库有 27000 个 HTML 格式的电影评论。这个库虽然未经过后期加工, 但它有较为客观的分数可供参考。分数体系有多种, 例如 5 星体系、字母 A-F 体系的分数、数字 1-10 的分数体系等。

在 Internet Movie Database 的基础上, 2002 年 Pang 和 Lee 公开了 1000 篇正面和 1000 篇负面影评成熟数据库(其中, 每类中有 312 位作者的评论, 且每个作者贡献 20 篇)。

中国科学院计算技术研究所吴琮等在 2010 年文章中, 介绍了他们所创建的包含 17122 篇正面和负面评论文章, 涵盖影视、教育和书籍等在内的 9 个主题的语料库<sup>[113]</sup>。

谭松波教授提供的中文情感挖掘语料 ChnSentiCorp<sup>[114]</sup> 是国内非常有影响数据资源。ChnSentiCorp 的内容主要来源于携程网上的酒店评论, 原始语料包含正面情感倾向性文本 7000 篇和负面情感倾向性文本 3000 篇。

#### 4.2.4 意见挖掘数据集

针对倾向性信息抽取任务, Hu 等在 2004 年提供了包含 4 个类别共 5 种产品的语料库, 该语料库对每个句子标记了评价对象、倾向性极性及其强度信息。

MPQA(Multi-Perspective Question Answering)最初是由 Wiebe 等人于 2002 年构建的带有深度标注的倾向性分析语料库, 包含 692 篇文档, 其中每句都标记了评价对象、评价表达式、极性和强度等在内的信息。在 2008 年, 中国科学院自动化研究所和复旦大学一起构建了由 478 篇评论构成的包含汽车、数码相机、笔记本、手机四个领域的语料库, 标注了其中的评价对象以及作者对该评价对象的倾向性所构成的倾向性单元<sup>[115]</sup>。

### 4.3 主要会议

除了各种类型的基准语料库之外, 近年来国内外也开展了一系列的倾向性相关评测活动, 对推动这个领域的发展起到了很好的作用。

在情感分析领域, 为了学术的交流, 近年来也陆续出台了一些测试项目, 例如, TREC(文本检索会

议)<sup>[116]</sup>, NTCIR(亚洲语言信息检索评测会议)<sup>[117]</sup>, 中文信息学会主办的评测会议(COAE2013)<sup>[118]</sup>, 中国计算机学会自然语言处理专委会主办的情感倾向性评测会议等。TREC 的测试证实观点的检索和常规的相关性检索相比, 在准确率方面尚有很大的提升空间。根据国内外相关文献报道的测试数据, 以及中文信息学会和计算机学会举办的多届中文倾向性分析评测会议的报告可以看出, 目前的主题观点情感性(倾向性)分类效果远远不如主题分类, 其准确率一般维持在 70%到 80%左右, 有些测试指标甚至很低。

## 5 展望及结论

举个简单的例子作为本文总结对象:

- (1) “这瓶洗发水, 适合头发很干的人用”
- (2) “用了这瓶洗发水, 头发变得很干”
- (3) “这瓶洗发水太好了, 我的头发都快被洗光了”

前两个句子中的主要成分都差不多, 描述的中心议题也集中在“洗发水”, “头发”, “很干”三个主题词上。如果采用传统的特征向量表示方法, 两个句子的向量几乎没有区别。但是其本质意义却不同, 第(1)句是褒义(正面), 第(2)句则是贬义(负面), 第(3)个句子则更难识别, 因为其采用了讽刺的手法, 表达了贬义。

由此可知, 尽管在情感分析和意见挖掘领域研究很多, 但仍有很大的研究空间。本文作者认为可以在如下方向进行深入研究。

#### (1) 研究适合情感分析的特征表示方法

很显然, 这种传统的把所有主题词像装袋子(bag)一样组织在一起的方法, 适合于面向主题分类, 而不适合于倾向性分析等新的应用领域。新的表示法要考虑依赖关系、词性组合、句型、句式结构、句子修辞手法, 以及词语间的相互联系等的重要因素。

#### (2) 选择适合情感分析的有监督学习算法

例如, 深度学习算法、基于高阶张量的分类算法、结构模式识别算法等可能更适合进行情感分析类别识别。

#### (3) 写作风格和表现手法的识别

为了解决自然语言处理领域的情感倾向性分析, 还可能需识别句子或段落等字面意义背后的隐含意思。因此需要解决比喻、讽刺等问题。

**致谢** 在本文撰写过程中, 参考了大量国内外研究者的论文、网站等资料, 在此表示诚挚的谢意。文

本情感倾向性分析和意见挖掘是个极具挑战的领域,国内外有很多研究人员参与其中。由于文献的数量实在太多,难免会疏漏一些好的研究成果,如果疏漏了您的成果,还请海涵。

## 参考文献

- [1] Collin B C. The future of cyberterrorism: The physical and virtual worlds converge[J]. *Crime and Justice International*, 1997, 13(2): pp.15-18.
- [2] Act U S P. Uniting & Strengthening America by Providing Appropriate Tools Required to Intercept & Obstruct Terrorism[J]. *Public Law*, 2001, 107(56).
- [3] Chakrabarti S, Van den Berg M and Dom B. Focused crawling: a new approach to topic-specific Web resource discovery[J]. *Computer Networks*, 1999, 31(11): pp.1623-1640.
- [4] Mehta R, Mehta D, Chheda D, et al. Sentiment analysis and influence tracking using twitter[J]. *International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE)*, 2012, 1(2): pp: 72-79.
- [5] Signorini A, Segre A M and Polgreen P M. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic[J]. *PloS one*, 2011, 6(5): e19467.
- [6] J.L. Fu, D.Y. Sun, J. Xiao, et al. Review of the research on the terrorist networks based on social network analysis [J]. *Systems Engineering—Theory & Practice*, 2013, 33(9): pp. 2177-2186.  
(付举磊, 孙多勇, 肖进, 等. 基于社会网络分析理论的恐怖组织网络研究综述[J]. 系统工程理论与实践, 2013, 33(9): pp.2177-2186.)
- [7] John Koetsier, “NSA Director General Keith Alexander: PRISM saved us 50 times”, <http://venturebeat.com/2013/06/18/nsa-director-general-keith-alexander-testifies-before-house-committee-on-intelligence/>, June 18, 2013.
- [8] J.M Chen, H.F. Lin and Z.H Yang. Bayesian Model based Word Sentiment disambiguation [C]. *Research and application of cutting-edge content computing, The Ninth China National Conference on Computational Linguistics*. 2007.  
(陈建美, 林鸿飞, 杨志豪. 基于贝叶斯模型的词汇情感消歧[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. 2007.)
- [9] L. H Xu, H.F. Lin, Y. Pan, et al. Constructing the Affective Lexicon Ontology [J]. *Journal of The China Society for Scientific And Technical Information*, 2008, 27(2): pp. 180-185.  
(徐琳宏, 林鸿飞, 潘宇, 等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2): pp. 180-185.)
- [10] Hatzivassiloglou V and McKeown K R. Predicting the semantic orientation of adjectives[C]. *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics. Association for Computational Linguistics*, 1997: pp.174-181.
- [11] Kamps J, Marx M J, Mokken R J, et al. Using wordnet to measure semantic orientations of adjectives[J]. *LREC 2004*, volume IV, pp. 1115-1118.
- [12] Kamps J, Marx M, Mokken R J, et al. Words with attitude[M]. *Institute for Logic, Language and Computation (ILLC)*, University of Amsterdam, 2001.
- [13] Y.L. Zhu, J. Min, Y.Q. Zhou, et al. Semantic Orientation Computing Based on HowNet [J]. *Journal of Chinese Information Processing*, 2006, 20(1): pp.16-22.  
(朱嫣岚, 闵锦, 周雅倩, 等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1): pp.16-22.)
- [14] L.H Xu, H.F. Lin and Z.H Yang. Text Orientation Identification Based on Semantic Comprehension [J]. *Journal of Chinese Information Processing*, 2007, 21(1): pp.96-100.  
(徐琳宏, 林鸿飞, 杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报, 2007, 21(1): pp.96-100.)
- [15] Q. Liu and S.J. Li, Word Similarity Computing Based on How-net [J]. *Computational Linguistics and Chinese Language Processing*, 2002, 7(2): pp.59-76.  
(刘群, 李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学, 2002, 7(2): pp. 59-76.)
- [16] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews[C]. *Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics*, 2002: pp.417-424.
- [17] Z.M. Wang, Z.F. Zhu and S.W. Yu, Research on Word Emotional Evaluation Based on the Grammar Knowledge-base of Contemporary Chinese[J]. *Computational Linguistics and Chinese Language Processing*, 2005, 10(4): pp.581-592.  
(王治敏, 朱学锋, 俞士汶. 基于现代汉语语法信息词典的词语情感评价研究[J]. 中文计算语言学, 2005, 10(4): pp.581-592.)
- [18] Hatzivassiloglou V and Wiebe J M. Effects of adjective orientation and gradability on sentence subjectivity[C]. *Proceedings of the 18th conference on Computational Linguistics-Volume 1. Association for Computational Linguistics*, 2000: pp. 299-305.
- [19] Wiebe J, Wilson T and Bell M. Identifying collocations for recognizing opinions[C]. *Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*. 2001: pp. 24-31.
- [20] Riloff E and Wiebe J. Learning extraction patterns for subjective expressions[C]. *Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics*, 2003: pp.105-112.
- [21] Wiebe J and Riloff E. Creating subjective and objective sentence classifiers from unannotated texts[M]. *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, 2005: pp. 486-497.
- [22] Hu M and Liu B. Mining and summarizing customer reviews[C]. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004: pp. 168-177.
- [23] Yu H and Hatzivassiloglou V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences[C]. *Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics*, 2003: pp.129-136.
- [24] Yu H and Hatzivassiloglou V. Towards answering opinion ques-

- tions: Separating facts from opinions and identifying the polarity of opinion sentences[C], *Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics*, 2003: pp. 129-136.
- [25] L.M Xue, D.W. Li and B. Xiao. Study on Novel Quintuple Model for Chinese Text Sentiment Orientation[J]. *Communications Technology*, 2011, 44(7): 130-132.  
(薛丽敏, 李殿伟, 肖斌. 中文文本情感倾向性五元模型研究[J]. 通信技术, 2011, 44(7): 130-132.)
- [26] Song Y, Zhang S and Lin H. Sentence Sentiment Analysis Based On Ambiguous Words[J]. *Journal of Chinese Information Processing*, 2012, 26(3): pp.38-32.
- [27] Wei X, Zhang Q, Miao J, et al. Event Sentiment Analysis Based on Semantic Chunks[J]. *Journal of Chinese Information Processing*, 2012, 3: pp.009.
- [28] Wilson T, Wiebe J and Hwa R. Just how mad are you? Finding strong and weak opinion clauses[C], *aaai*. 2004, 4: pp.761-769.
- [29] Socher R, Pennington J, Huang E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions[C], *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, 2011: pp.151-161.
- [30] Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C], *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. 2013, pp.1631-1642.
- [31] G. Wang and J. Zhao. Sentence Sentiment Analysis Based on Multi-redundant-labeled CRFs [J]. *Journal of Chinese Information Processing*, 2007, 21(5): pp. 51-55.  
(王根, 赵军. 基于多重冗余标记 CRFs 的句子情感分析研究[J]. 中文信息学报, 2007, 21(5): pp.51-55.)
- [32] P. Di, A.P. Li and L.G. Duan. Text sentiment polarity analysis based on transition sentence [J]. *Computer Engineering and Design*, 2014, 35(12): pp.4289-4295.  
(邸鹏, 李爱萍, 段利国. 基于转折句式的文本情感倾向性分析[J]. 计算机工程与设计, 2014, 35(12): pp. 4289-4295.)
- [33] Chen H, Liu G and Huang C. Sentiment Classification of Commodity Evaluation based on Parsing[J]. *Information Security and Communications Privacy*, 2013, 2: pp.040.
- [34] Gongshen Liu, Wenlei He and Huoyao Lai. Study on Feature Representation Based on Sentimental Orientation Classification[J], *China Communications*. 2011, 8(3): pp. 90-98
- [35] Lee J H. Properties of extended Boolean models in information retrieval[C], *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc., 1994: pp.182-190.
- [36] Z.F. Yang. A Stable Information Retrieval Algorithm And Its Application In Peer to Peer Network[D]. The Institute of Computing Technology of the Chinese Academy of Sciences, 2003.  
(杨志峰. 稳定的信息检索方法及其在分布式环境下的应用[D]. 博士学位论文, 中科院计算所, 2003.)
- [37] Salton G, Fox E A and Wu H. Extended Boolean information retrieval[J]. *Communications of the ACM*, 1983, 26(11): pp. 1022-1036.
- [38] Salton G, Wong A and Yang C S. A vector space model for automatic indexing[J]. *Communications of the ACM*, 1975, 18(11): pp. 613-620.
- [39] Tanaka T. Statistical-mechanics analysis of Gaussian labeled-unlabeled classification problems[C], *Journal of Physics: Conference Series*. IOP Publishing, 2013, 473(1): 012001.
- [40] R.L. Li, J.H. Wang, X.Y. Chen, et al. Using Maximum Entropy Model for Chinese Text Categorization [J]. *Journal of Computer Research And Development*, 2005, 42(1): pp. 94-101.  
(李荣陆, 王建会, 陈晓云等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展, 2005, 42(1): pp. 94-101.)
- [41] Somashekara M T, Kiran R R S and Muralidhara B L. Probabilistic Measures of Similarity/Dissimilarity Between Markov Models for Construction of Guide Tree for Multiple Sequence Alignment[J]. *IJECCE*, 2013, 4(3): pp.1091-1094.
- [42] J.K. Yu, Y.X. Wang and H.C. Chen. An Improved Text Feature Extraction Algorithm Based on N-Gram [J]. *Library and Information Service*, 2004, 48(8): pp. 48-50.  
(于津凯, 王映雪, 陈怀楚. 一种基于 N-Gram 改进的文本特征提取算法[J]. 图书情报工作, 2004, 48(8): pp.48-50.)
- [43] Tomović A, Janičić P and Kešelj V. n-Gram-based classification and unsupervised hierarchical clustering of genome sequences[J]. *Computer methods and programs in biomedicine*, 2006, 81(2): pp.137-153.
- [44] Yen S J, Lee Y S, Wu Y C, et al. Automatic chinese text classification using n-gram model[M], *Computational Science and Its Applications-ICCSA 2010*. Springer Berlin Heidelberg, 2010: pp.458-471.
- [45] Z.Q. Zhang. Patterns of Word Class Combination for Sentiment Analysis in Chinese[D]. Harbin Institute of Technology, 2007.  
(张紫琼. 面向中文情感分析的词类组合模式研究[D]. 哈尔滨工业大学, 2007.)
- [46] Y. Xu, J.T. Li, B. Wang, et al. A Study on Constraints for Feature Selection in Text Categorization [J]. *Journal of Computer Research And Development*, 2008, 45(4): pp. 596-602.  
(徐 燕, 李锦涛, 王 斌, 等. 文本分类中特征选择的约束研究[J]. 计算机研究与发展, 2008, 45(4): 596-602.)
- [47] Manning C D, Raghavan P and Schütze H. Introduction to information retrieval[M]. Cambridge: Cambridge university press, 2008.
- [48] Bremner D, Demaine E, Erickson J, et al. Output-sensitive algorithms for computing nearest-neighbour decision boundaries[J]. *Discrete & Computational Geometry*, 2005, 33(4): pp.593-604.
- [49] Chen J, Huang H, Tian S, et al. Feature selection for text classification with Naïve Bayes[J]. *Expert Systems with Applications*, 2009, 36(3): pp. 5432-5435.
- [50] Y.L Liu. A Study on Machine Learning Based Chinese Text Classification[D]. Xidian University, 2009.  
(刘依璐. 基于机器学习的中文文本分类方法研究[D]. 西安电子科技大学, 2009.)
- [51] Manevitz L M and Yousef M. One-class SVMs for document classification[J]. *the Journal of machine Learning research*, 2002, 2: pp.139-154.
- [52] J.N. Zeng, G.S. Liu, B. Su. Research on Sentiment Classification

- of Chinese Micro-blog [J]. *China Information Security*, 2013 (3): pp. 56-58.  
(曾佳妮, 刘功申, 苏波. 微博话题评论的情感分析研究[J]. 信息安全与通信保密, 2013 (3): pp. 56-58.)
- [53] Lafferty J, McCallum A and Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]. *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*, 2001. pp.282-289
- [54] Bengio Y, Courville A and Vincent P. Representation learning: A review and new perspectives[J]. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2013, 35(8): pp.1798-1828.
- [55] D. Hernandez., "The Man Behind the Google Brain: Andrew Ng and the Quest for the New AI," <http://www.wired.com/wiredenterprise/2013/05/neuro-artificial-intelligence/all/>. Wired, 10 May 2013.
- [56] C. Metz., "Facebook's 'Deep Learning' Guru Reveals the Future of AI," <http://www.wired.com/wiredenterprise/2013/12/facebook-yann-lecun-qa/>. Wired, 12 December 2013.
- [57] Tao D, Li X, Hu W, et al. Supervised tensor learning[C], *Data Mining, Fifth IEEE International Conference on*. IEEE, 2005: 8
- [58] Renard N and Bourennane S. Dimensionality reduction based on tensor modeling for classification methods[J]. *Geoscience and Remote Sensing, IEEE Transactions on*, 2009, 47(4): pp.1123-1131.
- [59] Ng V, Dasgupta S, Arifin S M. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews[C], *Proceedings of the COLING/ACL on Main conference poster sessions. Association for Computational Linguistics*, 2006: pp.611-618.
- [60] Gamon M. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis[C], *Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics*, 2004: pp.841.
- [61] X.Y. Xu, G.S. Liu. Texture Based Sentiment Orientation Identification for Chinese Texts [J]. *Journal of Chinese Information Processing*, 2015, 29(3): 106-112.  
(许歆艺, 刘功申. 基于文本纹理特征的中文情感倾向性分类[J]. 中文信息学报, 2015, 29(3): 106-112.)
- [62] Kim S M, Hovy E. Determining the sentiment of opinions[C], *Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics*, 2004: pp. 1367.
- [63] T.F. Yao, Q.Y. Nie, J.C. Li, et al. A Mining System for Chinese Automobile Reviews [C], *The Twenty-fifth Anniversary of Chinese Information Processing Society Conference*, 2006: pp. 260-281.  
(姚天昉, 聂青阳, 李建超, 等. 一个用于汉语汽车评论的意见挖掘系统 [C], 中文信息处理前沿进展-中国中文信息学会二十五周年学术会议论文集, 2006: pp. 260-281.)
- [64] Carenini G, Ng R T, Zwart E. Extracting knowledge from evaluative text[C], *Proceedings of the 3rd international conference on Knowledge capture. ACM*, 2005: pp. 11-18.
- [65] Y.W. Huang. A Study on Mining Technology in Chinese Product Review[D]. Chongqing University, 2009.  
(黄永文. 中文产品评论挖掘关键技术研究[D].重庆大学, 2009.)
- [66] Somprasertsri G and Lalitrojwong P. A maximum entropy model for product feature extraction in online customer reviews[C], *Cybernetics and Intelligent Systems, 2008 IEEE Conference on. IEEE*, 2008: pp. 575-580.
- [67] Choi Y, Cardie C, Riloff E, et al. Identifying sources of opinions with conditional random fields and extraction patterns[C], *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, 2005: pp. 355-362.
- [68] Stoyanov V and Cardie C. Toward opinion summarization: Linking the sources[C], *Proceedings of the Workshop on Sentiment and Subjectivity in Text. Association for Computational Linguistics*, 2006: pp. 9-14.
- [69] Popescu A M and Etzioni O. Extracting product features and opinions from reviews[M], *Natural language processing and text mining. Springer London*, 2007: pp. 9-28.
- [70] Hu M and Liu B. Mining opinion features in customer reviews[C], *AAAI*. 2004, 4(4): pp.755-760.
- [71] Pang, Bo., Lee, and L., Vaithyanathan, S, Thumbs up? Sentiment classification using machine learning techniques[C], *In Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP)*, 2002: pp. 79-86.
- [72] Lee Y B and Myaeng S H. Text genre classification with genre-revealing and subject-revealing features[C], *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM*, 2002: pp. 145-150.
- [73] Liu B, Hu M and Cheng J. Opinion observer: analyzing and comparing opinions on the web[C], *Proceedings of the 14th international conference on World Wide Web. ACM*, 2005: pp.342-351.
- [74] Aciar S, Zhang D, Simoff S, et al. Recommender system based on consumer product reviews[C], *Proceedings of the 2006 IEEE/WIC/ACM international Conference on Web intelligence. IEEE Computer Society*, 2006: pp. 719-723.
- [75] Scaffidi C, Bierhoff K, Chang E, et al. Red Opal: product-feature scoring from reviews[C], *Proceedings of the 8th ACM conference on Electronic commerce. ACM*, 2007: pp. 182-191.
- [76] Abbasi A. Affect intensity analysis of dark web forums[C], *Intelligence and Security Informatics*, 2007 IEEE. IEEE, 2007: pp. 282-288.
- [77] Thomas M, Pang B and Lee L. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts[C], *Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics*, 2006: pp. 327-335.
- [78] Mullen T and Malouf R. A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse[C], *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. 2006: pp. 159-162.
- [79] M.Q. Yu, H.W. Luo, H.B. Xu, et al. Research on Hierarchical Topic Detection in Topic Detection and Tracking [J]. *Journal of Computer Research And Development*, 2006, 43(3): pp. 489-495.  
(于满泉, 骆卫华, 许洪波, 等. 话题识别与跟踪中的层次化话题识别技术研究[J]. 计算机研究与发展, 2006, 43(3): 489-495.)

- [80] James Allan, Jaime Carbonell and George Doddington. Topic Detection and tracking pilot study[C]. *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. 1998: pp.194-218.
- [81] Michael Schultz and Mark Liberman. Towards a universal dictionary for multi-language IR applications [R]. In *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic, Massachusetts, 2002, pp.225-241.
- [82] Yang Y, Zhang J, Carbonell J, et al. Topic-conditioned novelty detection[C]. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002: pp.688-693.
- [83] Ignat C, Steinberger R and Pouliquen B. Navigating multilingual news collections using automatically extracted information[J]. *CIT. Journal of computing and information technology*, 2005, 13(4): pp. 257-264.
- [84] Ralf Steinberger, Bruno Pouliquen and Camelia Ignat. News Explorer - Multilingual News Analysis with Cross-lingual Linking [C]. *Machine Learning for Multilingual Information AccessNIPS Workshop*, 2006, pp.5-32.
- [85] Allan J, Papka R and Lavrenko V. On-line new event detection and tracking[C]. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998: pp.37-45.
- [86] Hammouda K M and Kamel M S. Efficient phrase-based document indexing for web document clustering[J]. *Knowledge and Data Engineering, IEEE Transactions on*, 2004, 16(10): pp. 1279-1296.
- [87] Momin B F, Kulkarni P J and Chaudhari A. Web Document Clustering Using Document Index Graph[C]. *Advanced Computing and Communications International Conference on*. IEEE, 2006: pp. 32-37.
- [88] Chim H and Deng X. Efficient phrase-based document similarity for clustering[J]. *Knowledge and Data Engineering, IEEE Transactions on*, 2008, 20(9): pp. 1217-1229.
- [89] C.J. Liu, J. Liu, Y.J. Cai, et al. Research on the Visualization Method of Social Crowd Emotion Based on Microblog Text Data Analysis [J]. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 2016, 52(1): pp.178-186.  
(刘翠娟, 刘箴, 柴艳杰, 等. 基于微博文本数据分析的社会群体情感可视计算方法研究[J]. 北京大学学报 (自然科学版), 2016, 52(1): pp.178-186.)
- [90] X. J. Meng, X.L. Meng and Y. Y. Hu. A Research on Text Mining and Baidu Index based Investor Sentiment Index [J]. *Macroecomics*, 2016 (1): 144-153.  
(孟雪井, 孟祥兰, 胡杨洋. 基于文本挖掘和百度指数的投资者情绪指数研究[J]. 宏观经济研究, 2016 (1): 144-153.)
- [91] Shen B and Liu Y. An opinion formation model with two stages[J]. *International Journal of Modern Physics C*, 2007, 18(08): pp. 1231-1242.
- [92] Latane B. The psychology of social impact[J]. *American psychologist*, 1981, 36(4): pp. 343.
- [93] Tatemura J. Virtual reviewers for collaborative exploration of movie reviews[C]. *Proceedings of the 5th international conference on Intelligent user interfaces*. ACM, 2000: pp.272-275.
- [94] Terveen L, Hill W, Amento B, et al. PHOAKS: A system for sharing recommendations[J]. *Communications of the ACM*, 1997, 40(3): pp. 59-62.
- [95] Riloff E, Wiebe J and Phillips W. Exploiting subjectivity classification to improve information extraction[C]. *Proceedings of the National Conference On Artificial Intelligence*. 1999, 2005, 20(3): pp.1106.
- [96] Lita L V, Schlaikjer A H, Hong W C, et al. Qualitative dimensions in question answering: Extending the definitional QA task[C]. *Proceedings Of The National Conference On Artificial Intelligence*, 2005, 20(4): pp.1616.
- [97] Seki Y, Eguchi K, Kando N, et al. Multi-document summarization with subjectivity analysis at DUC 2005[C]. *Proceedings of the Document Understanding Conference (DUC)*. 2005.
- [98] Piao S, Ananiadou S, Tsuruoka Y, et al. Mining opinion polarity relations of citations[C]. *International Workshop on Computational Semantics (IWCS)*. 2007: pp. 366-371.
- [99] Taboada M, Gillies M A and McFetridge P. Sentiment classification techniques for tracking literary reputation[C]. *LREC workshop: towards computational models of literary analysis*. 2006: pp.36-43.
- [100] Gongshen Liu, Wenlei He. Study on Feature Representation Based on Sentimental Orientation Classification[J]. *China Communications*. 2011, 8(3): pp. 90-98
- [101] Leshed G and Kaye J J. Understanding how bloggers feel: recognizing affect in blog posts[C]. *CHI'06 extended abstracts on Human factors in computing systems*. ACM, 2006: pp. 1019-1024.
- [102] Ye Q, Shi W and Li Y. Sentiment classification for movie reviews in Chinese by improved semantic oriented approach[C]. *Proceedings of the 39th Annual Hawaii International Conference on*. IEEE, 2006, 3: pp. 53b-53b.
- [103] Mishne G and De Rijke M. Capturing Global Mood Levels using Blog Posts[C]. *AAAI spring symposium: computational approaches to analyzing weblogs*. 2006: pp. 145-152.
- [104] Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives[C]. *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the European chapter of the association for computational linguistics*. Association for Computational Linguistics, 1997: pp. 174-181.
- [105] Stone P J, Dunphy D C and Smith M S. The General Inquirer: A Computer Approach to Content Analysis [M]. MIT Press, Cambridge, US. 1966.
- [106] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining[C]. *LREC*. 2010, 10: pp. 2200-2204.
- [107] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis[C]. *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 2005: pp. 347-354.
- [108] Pennebaker J W, Francis M E, Booth R J. Linguistic inquiry and word count: LIWC 2001[J]. *Mahway: Lawrence Erlbaum Association*



- ates, 2001, pp. 71: 2001.
- [109] T. F. Yao and D.C. Lou. Study of discrimination of Chinese emotional words meaning tendency [C], *International Conference On Chinese Computing*, 2007.  
(姚天昉, 娄德成. 汉语情感词语义倾向判别的研究[C], 中国计算技术与语言问题研究—第七届中文信息处理国际会议论文集. 2007 年, 2007.)
- [110] J. M. Chen. A Construction and Application of Chinese Emotion Word Ontology[D]. Dalian University of Technology, 2009.  
(陈建美. 中文情感词汇本体的构建及其应用[D]. 大连理工大学, 2009.)
- [111] Z.D. Dong and Q. Dong. “HowNet Knowledge database”. <http://www.keenage.com>, 2007.  
(董振东, 董强. “知网情感分析词语集”, <http://www.keenage.com>, 2007.)
- [112] X.Y. Xu, Text Texture Model and Its Application[D], Shanghai Jiao Tong University, 2014.  
(许歆艺. 文本纹理模型及其应用研究[D]. 上海交通大学, 2014.)
- [113] Q. Wu, S.B. Tang, X.Q. Chen. Chinese Words Emotional Tendency Analysis and Related Research[J]. *Information Technology Letter*, 2010, 8(4): pp.16-38.  
(吴琼, 谭松波, 程学旗. 中文情感倾向性分析的相关研究进展[J]. 信息技术快报, 2010, 8(4): pp.16-38.)
- [114] S.B. Tang, ChnSentiCorp. <http://www.searchforum.org.cn/tansongbo/corpus-sen-ti.htm>.  
(谭松波, 中文情感挖掘语料. <http://www.searchforum.org.cn/tansongbo/corpus-sen-ti.htm>.)
- [115] J. Zhao, H.B. Xu, X.Q. Huang, et al. A Test Report of Chinese Emotional Tendency[J]. *The first Conference on Chinese Emotional Tendency*. 2008: pp.1-20.  
(赵军, 许洪波, 黄萱菁, 等. 中文倾向性分析评测技术报告[J]. 第 1 届中文倾向性分析评测研讨会论文集. 2008: pp.1-20.)
- [116] Ounis I, Macdonald C, Soboroff I. Overview of the TREC-2008 blog track[R]. GLASGOW UNIV (UNITED KINGDOM), 2008.
- [117] National Institute of Informatics Retrieval Test Dataset. <http://reasearch.nii.ac.jp/ntcir-ws6/opinion/index-en.html>.
- [118] S.B. Tang, S.G. Wang, X.W. Liao, et al. The Overall Report of the Fifth Chinese Emotional Tendency Conference [J]. *the Fifth Chinese Emotional Tendency Conference*. 2013: pp. 5-53.  
(谭松波, 王素格, 廖祥文, 等. 第五届中文倾向性分析评测总体报告[J]. 第五届中文倾向性分析评测报告论文集. 2013: pp. 5-53.)
- [119] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: A deep learning approach[C]// *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011: 513-520.
- [120] Zhou S, Chen Q, Wang X. Active deep learning method for semi-supervised sentiment classification[J]. *Neurocomputing*, 2013, 120: 536-546.
- [121] Duyu T, Bing Q, Ting L. Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5:10.1002/widm.2015.5.issue-6, 292-303.
- [122] dos Santos C N, Gatti M. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts[C]//*COLING*. 2014: 69-78.
- [123] Tang D, Wei F, Qin B, et al. Coooolll: A deep learning system for Twitter sentiment classification[C]//*Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. 2014: 208-212.



**李建华** 于 1998 年上海交通大学通信息与信息系统专业获得工学博士学位。现任上海交通大学网络空间安全学院院长、教授、博士生导师, 信息内容分析技术国家工程实验室主任。研究领域为信息安全。研究兴趣包括: 信息安全综合管理与互联网内容安全等。Email: lijh888@sjtu.edu.cn



**刘功申** 于 2003 年上海交通大学计算机专业获得工学博士学位。现任上海交通大学网络空间安全学院副教授、硕士生导师。研究领域为内容安全。研究兴趣包括: 网络舆情分析与监控、不良信息过滤技术、恶意代码及其防范技术等。Email: lgshen@sjtu.edu.cn



**林祥** 于 2005 年上海交通大学通信息与信息系统专业获得工学硕士学位。现任上海交通大学信息内容分析技术国家工程实验室副总工程师。研究领域为内容安全。研究兴趣包括: 互联网公开大数据感知采集、融合分析与表达呈现等。Email: lionel@sjtu.edu.cn