DOI:CNKI: 11-2746/G2.20130325.2301.001

基于网络用户情感分析的预测方法研究*

徐 健

摘 要 网络用户情感分析领域的研究为特定领域社会行为的预测提供了新的方法和工具。本文分析了基于情感分析进行预测的逻辑基础、典型预测方法、关键技术以及当前存在的问题和发展趋势。研究发现:研究基于网络用户情感分析预测社会活动趋势的方法在政治、财经等多个领域具备应用条件;典型预测方法可归纳为以情感分析结果作为辅助依据的预测方法和以情感分析结果作为主要依据的预测方法;预测过程涉及情感分析源的选择、预测时间提前量的确定以及情感词统计处理三个关键环节;当前研究还存在网络用户情感的代表性,待分析语料的全面和正确获取,以及网络用户情感的正确分析和统计等问题,有待深入研究。图2。参考文献47。

关键词 社会化媒体 网络用户 情感分析 预测方法

分类号 G250

Research on Predicting Methods Based on Network User Sentiment Analysis

Xu Jian

ABSTRACT The research on field of network user sentiment analysis provides new methods and tools for predicting social activities in special domain. The paper analyses the logical basis, typical methods, key technologies, problems and development tendency of predicting methods based on sentiment analysis. The paper concludes that the predicting methods can be used in many domains, such as politics and finance; the typical methods include predicting methods that take sentiment analysis as auxiliary and predicting methods that take sentiment analysis as primary; the predicting process involves three key points which are the choice of sentiment analysis resources, the determination of lead time and the compute of sentiment words; current researches also have some problems that need to be researched further, which include the representativeness of network users sentiment, the comprehensive and correct acquisition of corpus, and the correct analysis and statistics of network users sentiment. 2 figs. 47 refs.

KEY WORDS Social media. Network user. Sentiment analysis. Predicting method.

1 引言

随着社会化媒体(Social Media)的快速发展,越来越多的网络用户可以直接通过各种渠道(例如在线购物网站、博客、社交网站、论坛等)表达对产品、服务、社会事件、公众人物等的意见和情感倾向。随时间不断增长的社会化媒体内容及其内在情感表达在一定程度上折射出社会集体的智慧和情感状态,而这种情感状态往往会影响到个人在社会生产生活中各种行为决策的选择结果。因此,网络用户情感状态对于许多社会活动具有预测能力,

* 本文系国家社会科学基金项目资助课题"用户评论情感分析及其在竞争情报服务中的应用研究"(项目编号:11CTQ022)的研究成果之一。

通讯作者:徐健,Email: issxj@mail.sysu.edu.cn

而这一点正在被不同领域的相关研究所证实。

近年来有不少学者利用社会化媒体用户情感分析进行特定领域社会行为的预测方法研究。由于不同的社会活动对于情感因素的敏感程度是有区别的，因此情感分析因素在不同预测任务中所起到的作用也不尽相同。在某个领域的预测任务中具有独立预测能力的情感分析方法，在另一个领域中可能被证明不具有独立预测能力，而仅能作为重要补充因素来利用。此外，不同学者所使用的情感分析方法也存在差异，这客观上影响了研究者对情感分析重要度的判断结果。根据情感因素在预测过程中的利用方式，可将当前相关研究大致分为以下两类：将情感分析作为预测的主要信息源和依据对现实世界的社会活动进行预测，以及将情感分析作为其他主要预测依据的重要补充因素以实现更好的预测效果。

研究者普遍认为，如果分析样本足够大，预测方法设计合理，那么基于社会化媒体的情感分析在预测方面能够提供与问卷调查、民意调查等传统方法相当、甚至更好的预测结果。此外，基于网络用户情感的预测方法在采样范围、预测时效性、预测成本等方面具有传统调查预测方法无法比拟的优越性。随着人们参与社会化媒体的积极性不断提高，以及情感分析方法的逐步成熟，相信这类预测方法势必会成为企业竞争情报重要、及时、廉价的来源和市场参与者制定决策的重要依据。

本文对网络用户情感分析进行阐述，对应用情感分析进行预测的典型方法及其关键技术进行分析，并对当前相关研究存在的问题及未来发展进行总结和展望。

2 网络用户情感分析

情感分析（Sentiment Analysis）又称情感分类（Sentiment Classification）、评论挖掘（Review Mining）或意见挖掘（Opinion Mining），是指通过自动分析商品、服务、人物等研究对象的相关评论文本内容，发现评论者对该研究对象的褒贬态度和意见[1-2]。而网络用户情感分析则主要针对社会化媒体产生的评论信息进行自动情感分析。目前在网络用户情感分析领域进行的研究主要以三个层面中的一种来实现，即：文档层面、语句层面和属性层面。文档层面的情感分析根据整篇评论的情感倾向将评论分成三个极性类：正面、负面或者中性，主要采用各种机器学习技术进行评论的情感分析。典型的有 Abbasi A. 等人[3]提出的用来分类多种语言 Web 论坛意见的方法。句子层面的情感分析主要集中在识别主观句子，并判断它们的情感极性。大多数这类研究采用机器学习方法。例如，Liu H. 等人[4]提出了一个基于大规模通用知识库来探测句子层面情感的方法，通过情感表达的自动学习实现对句子情感进行分类。文档层和句子层的情感分析对于精确确定用户情感都过于粗糙。为了解决这个问题，从评论中抽取商品特定属性意见的属性层面情感分析被提出来。在 Hu M. 等人[5]的相关研究中，词性标注标签序列规则被用来抽取产品的属性，而属性的情感描述短语的极性根据上下文信息来判断获得。

在情感倾向性方面，较多相关研究使用正面、负面这两类情感来区分文本中的情感倾向[6-7]。随着研究的逐步深入，也有研究认为这样的简单情感划分也许会忽略许多丰富、多维的人类情感信息。因此一些相关研究尝试对情感倾向性进行进一步细分。例如，Bollen J. 等人[8]在其研究中创建了一个情感分析工具 GPOMS，能够在“平静的”、“警惕的”、“确信的”、“至关重要的”、“宽容的”、“高兴的”这六个不同的维度上测度网络用户情感。也有学者进一步认为，网络评论中所表达的情感不仅有多维度的划分，每种情感的强弱程度对于情感分析也同样重要。Thelwall M. 等人[9]提出了 SentiStrength 算法，基于网络英文文本的语法和拼写风格来计算情感强度。

情感分析技术主要包括两类方法：机器学习方法和语义倾向性方法。基于机器学习的情感分类方法在使用时需要预先对大量的训练样本进行训练，以建立分类模型。而语义倾向性方法则不需要预先进行训练，它仅对语词倾向于正面或负面的程度进行计算。Chaovalit P. 等人[10]在电影评论数据

集上比较了语义倾向性方法和机器学习方法,发现机器学习方法更加可靠。Xu K. 等人[11]对简单比较句中的情感因素进行分析和统计,以获取厂商、零售商所需的竞争情报。Ye Q. 等人[12]在旅游目的地评论数据集上比较了常用三种机器学习方法(Naïve bayes,SVM 和 N-gram 方法)的情感分类效果,证明 SVM 和 N-gram 方法比 Naïve Bayes 方法更好。情感分析技术已经在电影、旅游资源、汽车、银行等产品或服务的网络评论方面得到了较多的应用探索。

国内相关研究起步较晚,但是近年来相关研究在多个方面取得了进展,可归纳为以下几个方面:①对情感分析的概念、类型、方法、应用等进行综述。典型的有赵妍妍等人[1]对文本情感分析的评测和资源建设情况、应用情况以及主流方法和前沿进展进行了概括和分析;陆文星等人[13]对信息抽取和情感识别这两类情感分类任务进行介绍,总结了情感分析的应用现状及存在的问题。②中文情感词识别及情感词库构建相关研究。朱嫣岚等人[14]基于知网提出了基于语义相似度和基于语义相关场两种词汇语义倾向性计算方法。张清亮等人[15]提出了一种在知网情感词集基础上利用 PMI-IR 算法进行领域情感词自动识别和词库构建的方法。③中文句子情感分析研究。李纲等人[16]对句子情感分析中的主客观句分类方法以及词汇上下文极性判断、评价主题识别、意见持有者识别等关键问题进行了总结。杨经等人[17]在词特征、词性特征、语义特征析取的基础上,使用支持向量机分类方法对句子进行情感识别和分类。④中文篇章级情感分析相关研究。李本阳等人[18]基于句型和句子位置等特征,提出了利用支持向量机模型进行篇章级情感分类的方法。夏云庆等人[19]采用基于情感单元的情感向量空间模型进行歌词情感分析,较好地解决了基于词汇的向量空间模型在文本表示效率、情感功能以及数据稀缺性等方面的不足。⑤情感分析应用研究。曹树金等人[20]将情感倾向性分析方法应用于对网络论坛中主题帖的舆情监控任务中,取得了较好的效果。郑文英[21]构建了基于逐点语义分析法、基于支持向量机、基于朴素贝叶斯以及 N 元文法的情感分类器模型,对中文旅行目的地评论进行分类和效果比较。

综上所述,作为一个新兴的研究领域,情感分析正逐渐受到计算机科学、经济学、管理学、情报学等相关学科研究者的关注。相关研究已经在词语级、句子集、篇章级情感分析方面进行了有益探索,并尝试将其应用于市场预测、舆情监测、竞争情报获取等多种任务中。

3 基于情感分析进行预测的逻辑基础

心理学相关研究认为,除了信息之外,情感在人类决策过程中扮演了重要角色[22-23]。当决策面临所需信息过于庞杂或不确定性时,决策者的情绪状态对于决策制定具有强烈影响。外部社会环境的情绪状态也会影响决策选择行为。通过社会交互行为,普遍的积极/消极社会情绪得到传递,并进而影响到包括消费者、个人投资者、机构投资者、公司、厂商等在内的各种类型的决策制定者[24]。当今社会化网络已成为网络用户表达意见和情感的平台,而对这些情感进行自动抽取和统计的情感分析技术也已日趋成熟,于是基于网络用户情感分析来预测社会活动趋势的方法在很多领域具备了应用基础。在这些不同领域相关研究中,对基于情感分析进行预测的逻辑基础进行了更为具体的阐述。

在政治活动领域,Tumasjan A. 等人[25]使用微博进行德国选举的预测研究时指出,微博最初的作用在于更新个人状态。而随着微博用户群的发展壮大和使用频率的提升,微博文覆盖了从政治新闻到产品信息几乎所有的社会生活领域。特别是在德国选举前几周,政治话题在许多微博用户中得到热议。另外,政治家也与选民通过微博进行积极交流,并尝试获得更多选民的支持。因此,微博文信息能够作为表达现实中人们政治意见的一个指示器。Hong S. 等人[26]使用推特(Twitter)上提及美国总统候选人的微博文数量进行选举预测研究,发现政治家的微博行为可以提升政治家被传统媒体(广播、电视、报纸等)提及的次数。而在传统媒体中被提及的次数与他们在微博上被提及的次数之间存

在强相关性。平均而言，在传统媒体上被提及的次数每增加 10%，那么在推特被提及次数将会相应上涨 4% 至 6%。因此政治家在微博上的发文行为会导致其在微博中被提及数的增加。对这一现象的一种解释是记者们在撰写新闻稿时可以从政治家微博中获取信息素材，而由此带来更多的传统媒体对政治家的相关报道，进一步提升了该政治家在微博中的热议度和在选民中的影响力。

在产品和服务销售领域，Gruhl D. 等人[27]认为使用博客情感分析能够预测图书销量排名峰值，其主要原因有两个方面：首先，博客对某书的提及信息量增加可能是由于较早购买者的推荐引起的，这极有可能影响到其他网络用户购买该书；其次，撰写关于一本书相关博文的门槛比购买该书的门槛要低，因此从图书在博客中得到热议到图书销售排行峰值到来之间存在一定的延迟，而这个时间延迟使得利用博文情感分析预测图书销售排行峰值成为可能。Asur S. 等人[28]在利用微博情感分析进行票房收入预测时指出，微博文中对电影正面的评论和意见可能会引起更多潜在电影观众的兴趣，并进一步推动电影票房收入的增长。Liu Y. 等人[29]在利用博客情感分析来预测电影票房收入时也持有相同的看法。他们认为由于积极的市场营销活动、存在的争议等因素，使得在电影上映前的博客提及数指标可能无法准确反映产品的销售业绩。当电影上映后，网络用户在博客中所反映出的情绪会更加趋向于客观，因此能够更好地预测票房的收入走势。

在财经领域，Joseph K. 等人[30]使用股票代码在商业搜索引擎的检索量作为投资者情感状态的有效替代物。他们认为那些正在认真考虑“买入”决策的投资者会通过股票代码检索来收集特定股票相关信息，而那些正在考虑“卖出”决策的投资者对于其所拥有的股票的历史和表现已经有较长时间的了解，因此很少对待卖股票进行检索。出于上述原因，股票代码的检索量应当可以作为投资者对于特定股票“乐观”情绪表达的替代物，并能够作为预测股票走势的依据。Zhang X. 等人[31]借助推特预测证券市场表现时认为，当人们对于未来持悲观或不确定态度时，他们将会在投资和交易时更加谨慎，并且开始使用更多的诸如“希望”(hope)、“害怕”(fear)、“担忧”(worry)这样的情感词。此外，推特中的被关注量指标通常被看做微博的流行测度。当持悲观态度的微博博主的被关注量越大，那么更多的人可能会受其微博文情绪感染，并同样感觉悲观。这些情绪会进一步影响到投资者在证券市场中的决策行为，并间接影响股市近期走势。

4 基于情感分析的典型预测方法

情感分析在很多领域的预测活动中得到了应用，并被证明在多数情况下能够获得较好的预测效果。但是由于应用领域存在差异性，情感分析与预测目标并非总是表现出密切相关的预测关系。此外，情感分析技术本身也存在适用领域或适用文本处理对象的限制性，因此在不同领域进行的预测方法研究中，情感分析所发挥的作用有所不同。按照情感因素在预测过程中被应用的方式不同，可将预测方法归纳为以下两类：以情感分析结果作为辅助依据的预测方法，以及以情感分析结果作为主要依据的预测方法。

4.1 以情感分析结果作为辅助依据的预测方法

在以情感分析结果作为辅助依据的预测方法中，采用多种指标相结合的方式来进行预测。这些指标通常都被证明与被预测指标具有相关关系，而多种指标的集成则能够使预测效果达到最优。与情感分析结果进行集成的其他预测指标包括社会化媒体中待预测对象的提及数、提及数的变形(例如微博文率，即每小时提及特定研究对象的微博文数量)以及与研究对象相关的权威机构指标(例如美国道琼斯工业指数)历史数据等。相关研究认为，在预测过程中考虑到社会化媒体所表达的情感因素，会有效提升已有指标的预测性能。

以情感分析结果和其他预测指标相结合作为预测依据的典型预测方法流程如图 1 所示。

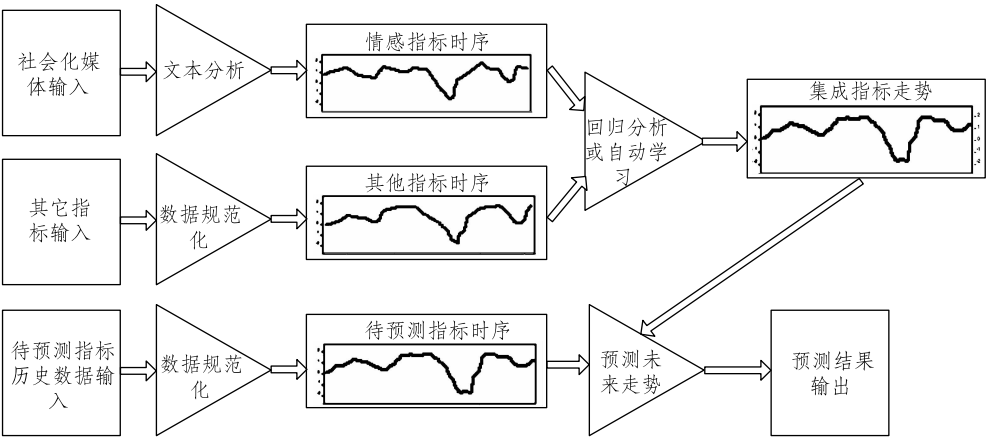

图1 以情感分析结果作为辅助依据的预测方法流程

在图1中,以情感分析结果作为辅助依据的预测方法流程主要包括四个关键步骤。

(1)对社会化媒体输入进行文本分析。文本分析可具体分为文本预处理和文本情感分析两个环节。在文本预处理环节,需要对以自然文本形式输入的社会化媒体内容进行规范化处理,以方便后续文本情感分析的开展。文本预处理环节具体任务包括:去除停用词、去除感叹号和问号以外的特殊字符、去除链接地址和用户ID、使用特定名称替换被研究对象名称(例如被预测票房收入的电影名、被预测销售量的书名)等。在文本情感分析环节,可以借助已有的情感词表对文本中出现的情感词进行标识和统计,以确定文本情感取向。也可以应用机器学习方法,通过训练建立分类模型,对文本进行情感分类。直接使用较为成熟的第三方情感分析工具(例如OpinionFinder[32-33],LIWC[34]等)进行文本情感分析的方法也常被研究者所采用。但是由于第三方情感分析工具通常针对通用领域开发,在应用到特定领域时很难进行有针对性的调整,因此情感分析的效果并不总是令人满意的。

(2)对其他指标输入和待预测指标历史数据输入进行数据规范化处理。将要与情感分析结果进行集成的其他指标输入和待预测指标历史数据输入可能存在数据表达不规范的问题,这使得后续指标集成或对比操作难以进行。应设定数据规模范围,将其他指标输入和待预测指标历史数据输入规范化到统一的数据规模范围内。

(3)对情感指标时序和其他指标时序进行集成。采用回归分析或自动学习的方法,对规范化后的情感指标时序和其他指标时序进行集成,获得集成指标时序。与单独的情感指标时序和其他指标时序相比,在减去时间提前量的情况下,集成指标时序与待预测的指标时序更加一致。

(4)利用集成指标走势对待预测指标的未来走势进行预测。当前的集成指标走势反映了待预测指标未来走势变化,而两者相差的时间提前量n可以通过训练文本集合和历史待预测指标数据计算获得,因此第t天附近的集成指标走势可以用来预测第t+n天的待预测指标走势。

以情感分析结果作为辅助依据的预测方法被应用到多个领域的预测活动中,典型代表有:Asur S.等人[28]使用电影上映日之前提及电影的推特率(每小时提及特定电影的推特数量)和情感极性指标(具有积极情感的微博与具有负面情感的微博的比值)相结合,能够在预测电影票房收入时获得最佳的效果。Tumasjan A.等人[25]对推特上提及德国政党或政治家的微博内容进行分析,发现仅仅使用提及政党的消息数量就能够预测选举结果。在结合对上述提及消息的情感分析因素后,预测方法能够更深入地描述政治家之间和政党之间的差异,对

大选后可能发生的政治结盟做出预测。Liu Y. 等人[29]认为博客提及数量可能受到市场营销、争议等因素的影响，因此很难独立用作准确预测销售业绩的信息源。他们将电影过去若干天的票房收入和提及电影的博客情感因素相结合来预测票房收入，取得了良好效果。Bollen J. 等人[8]将推特上用户情绪状态的时序变化情况与美国道琼斯工业指数(Dow Jones Industrial Average)历史数据相结合，来预测道琼斯工业指数的未来发展趋势。

4.2 以情感分析结果作为主要依据的预测方法

有研究者认为，社会化媒体的情感分析结果本身就足以预测一些领域中的现实指标。在这类预测方法中，成功的关键在于如何识别出与待预测指标最相关的情感。以情感分析结果作为主要预测依据的典型预测方法流程如图 2 所示。

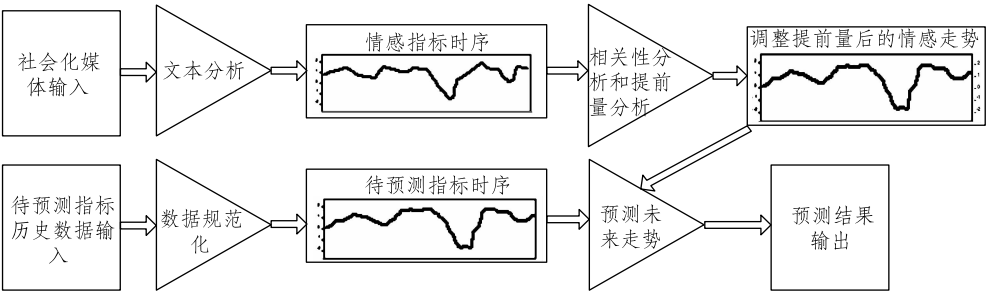

图 2 以情感分析结果作为主要依据的预测方法流程

以情感分析结果作为主要依据的预测方法流程与 4.1 小节中所介绍的方法流程基本类似，最大的不同之处在于本方法产生的情感分析结果并不需要与其他指标进行集成来实现预测功能，因此省去了其他指标输入、规范化以及指标集成的相关步骤。这类方法的典型代表有：Eric G. 等人[35]采用网络博客来预测美国标普 500 指数(S&P 500 Index)的走势，他们发现随着具有“焦虑”(anxiety)情绪的博客文章增多，则美国标普 500 指数将会面临下降的压力。这一发现进一步被 Zhang X. 等人[31]的相关研究所证实，他们发现推特上表达“希望”(hope)、“害怕”(fear)、“担忧”(worry)情感的微博数量变化与道琼斯工业指数负相关，当表达上述情感的微博数量快速增长时，往往预示着道琼斯指数将要下降。

值得注意的是，由于网络用户的情感状态在获取时可能受到情感分析技术条件限制或社会化媒体语料获取条件限制，一些学者也尝试利用网络用户情感分析的替代物作为主要预测依据来进行预测。这类方法通常选取社会化媒体中对待预测对象的提及数或网络用户对待预测对象的检索量等指标作为网络用户的情感分析替代物，认为这些指标本身就反映了网络用户情感表达的强烈程度。这类方法的典型代表有：Gruhl D. 等人[27]通过统计网络博客中与畅销书相关的评论数，来预测畅销书的销售排行峰值。Vasileios L. 等人[36]通过统计每天在推特上产生的与流行病症状相关的微博信息数，能够及时预测和定位流行病的爆发。Joseph K. 等人[30]研究了股票缩写的检索次数对股价走势的预测效应。他们认为股票缩写的网络检索次数增加，意味着潜在购买者对该股票持有乐观态度，并通过主动网络检索来收集相关信息，为下一步买入行为做准备。因此，股票缩写的检索次数可以看做投资者乐观情感的指示器。

5 关键技术分析

5.1 情感分析源的选择

在使用情感分析方法对社会化媒体资源进行分析时，情感分析源的选择和获取对分析结果有直接影响。更新及时、内容丰富、样本量大、质量较高的情感分析源往往为研究者所青睐，但有时由于预

测涉及领域不同,很难找到兼顾各方面的情感分析源,此时就需要根据具体预测任务进行有针对性的取舍。情感分析信息源的选择应充分考虑以下原则。

5.1.1 信息源应易于获取

通常基于情感分析的预测方法需要连续时间段内的社会化媒体信息单元来进行预测方法训练和验证。这些社会化媒体信息单元可以通过网络抓取程序自动累积抓取获得,通过社会化媒体检索接口调用周期性检索获得,也可以直接从其他网络资源保存系统间接获得。例如,Bollen J. 等人[8]将所抓取的 2008 年 2 月 28 日到 12 月 19 日期间的 985 万条推特微博文作为实验数据集进行情感分析和道琼斯指数预测。Tumasjan A. 等人[25]通过微博搜索引擎获取提及德国政治家或政党的微博文进行情感分析和选举预测。Gruhl D. 等人[27]在进行图书销售峰值预测时使用了 IBM 的 Fountain 项目[37]维护的 30 万博客评论中与待分析图书相关的评论信息。无论使用上述何种方式,应能保证社会化媒体资源长期可获得,以便于开展持续的情感分析和预测研究。

5.1.2 信息源更新的及时性

不同的社会化媒体信息源具有不同的更新频率,而基于这些信息源的预测方法性能也明显受到其更新频率的影响。研究者普遍认为,由于微博有 140 个字符的长度限制,因此发布微博文通常比发布博客文章更加容易和频繁,微博反映网络用户情绪变化的时效性比起博客要表现的更好。Vasileios L. 等人[36]在其流行病预测研究中指出,虽然使用微博进行流行病预测的时间提前量并不明显,但是更短的监测时间间隔、及时的流行病状态信息、较低的预测成本是其存在的根本。也有研究者[30]尝试使用网络用户情感分析的替代物——股票代码检索量来预测股票收益和交易量,但该研究受到商业搜索引擎提供检索量信息的延迟影响,无法及时做出预测行为,因此仍停留在理论方法探索阶段。

5.1.3 信息源内容的丰富性

信息源中相关研究对象的陈述数量应具有一定的规模,并且能够较为全面、客观地反映现实世界中人们对于研究对象的情感取向类别、比例和强度,否则稀疏数据或失实数据的输入必然会引起预测失效。信息源内容的丰富性原则与信息源更新的及时性原则在特定情况下可能需要作出取舍。例如在使用情感分析预测汽车产品可能存在的缺陷的任务中,微博信息虽然具有更新及时的优势,但是其内容长度限制使网络用户无法更加深入地描述产品使用体验和表达情感,因此汽车相关论坛帖子要比汽车相关微博更适合作为预测信息源。Abrahams A. 等人[38]正是使用了汽车论坛帖子来预测汽车产品缺陷。Dellarocas C. 等人[39]选择使用雅虎电影频道论坛的帖子来预测动画电影收益,也是出于上述考虑。

5.2 预测时间提前量的确定

不同领域对网络用户情感表达的反应时间和敏感程度都存在差别。此外,预测对象的促销活动、新闻报道、网上争议等社会事件也同样会对网络用户情感(进而对预测时间提前量)产生影响。因此借助实时情感分析获得的指标对于不同领域预测活动的预测时间提前量存在差别。预测时间量通常采用时间序列关联分析[40]、自回归模型[41]、格兰杰因果关系分析(Granger causality analysis)[42]等方法来实现。例如,在使用博客情感分析预测电影票房收入时,Liu Y 等人[29]提出了自回归情感感知(Auto Regressive Sentiment Aware model)模型,发现从提前一天的博客内容中捕获的情绪信息对于第二天电影票房收入预测能够达到最好的效果。在使用微博情感分析进行股票市场预测时,Bollen J. 等人[8]使用格兰杰因果关系分析对微博中各种情感时间序列与道琼斯指数时间序列进行分析,并认为"平静"(Calm)这种情绪对于道琼斯指数的预测效果最好,预测时间最佳提前量为 2 天。在使用博客提及数进行图书销售峰值预测时,Gruhl D.[27]利用时间序列关联分析方法来确定预测时间提前量。他们发现样本中不同的图书有着不同的预测时间提前量,但基本可以确定是在几天到几个星期范围内。这是由于图书销量峰值的到

来可能是受到各种社会事件的刺激而产生，这些社会事件可能包括图书的促销行为、图书获奖、图书改编电影上映等。不同的社会事件对于博客提及量的时效性影响存在一定差异，这间接影响到了预测时间的提前量。Joseph K. 等人[30]使用股票代码检索量来预测股票非正常收益时，通过分析 2005 年至 2008 年相关股票代码在 Google 中的检索量与股票历史收益情况，发现那些检索量迅速增加的股票在一周内股价会迅速增长，在持股 5 周时获得最大收益，在第 8 周后股价明显下滑。

还有一些学者并未对精确的预测时间提前量进行直接研究，而只关注采用历史相关数据来预测明天或某个具有特定意义时间点的特定指标量。例如，Asur S. 等人[28]发现在电影上映时间前后，会在微博平台上受到热议。其后相关的微博文数量会逐渐减少。票房收入也经历了类似的过程，上映日那一周周末的电影票房收入往往是最高的。因此他们通过构造基于线性回归的预测模型来预测电影上映那一周周末的票房收入。Zhang X. 等人[31]使用第 t 天、t－1 天和 t－2 天含有“希望”(hope)、“害怕”(fear)和“担心”(worry)的微博数来预测第 t＋1 天的证券市场指数，并认为这个情感指数与道琼斯、纳斯达克和标普 500 指数明显存在负相关关系。

5.3 情感词统计处理方法

在基于网络用户情感分析进行预测的方法中，通常需要构造特定情感时序图，以便与现实世界中的被预测时序进行相关分析和确定预测提前量，建立预测模型。由于情感往往是多方面的，它们以多种方式存在并且拥有极性、程度等属性，而且有时相同的情感词在不同的领域中表达的情感也有所不同，因此想要对所有情感表达进行彻底标注和统计存在较大困难，而仅将语料中表达的情感归类到正面或负面似乎又走了另一个极端。Liu Y. 等人[29]在概率潜在语义分析(Probabilistic Latent Semantic Analysis，PLSA)[43]基础上提出了情感概率潜在语义分析模型(Sentiment PLSA)，认为借助该模型可对博客中复杂的情感表达进行降维，形成低维度的隐藏情绪因素向量，从而使具有同义、近义的情感表达对应到相同或相似的隐藏情绪因素，去除部分噪音，提高情感特征的鲁棒性。

还有研究表明，并非每一种情感都具有相同的预测效力。有效甄别那些最具有预测能力的情感类别，并有针对性地对其进行识别、统计和规范化，对于提高预测准确性至关重要。Zhang X. 等人[31]在研究中对正面情感词“希望的”(hope)、“高兴的”(happy)和负面情感词“害怕的”(fear)、“担忧的”(worry)、“紧张的”(nervous)、“焦虑的”(anxious)、“心烦的”(upset)的预测能力进行了研究，认为人们在经济形式不确定时开始更多地使用“hope”，“fear”，“worry”情感词，而与经济形式将要向好转变还是向坏转变相独立。Bollen J. 等人[8]的研究中分别对 OpinionFinder 测度的语句级别的正面/负面情感和 GPOMS[44－45]测度的六种情感(“平静的”(Calm)，“警惕的”(Alert)，“确信的”(Sure)，“至关重要的”(Vital)，“宽容的”(Kind)，“高兴的”(Happy))进行了考察，认为在预测道琼斯指数走势时，GPOMS 测度的“Calm”情感具有最佳的预测效果。

深入细致的情感分析往往需要丰富完备的情感词库、精心设计的标注算法以及质量较高的训练语料。一些研究者认为可以寻找其他行之有效而且易于获得的指标来替代情感分析指标，从而使预测方法更具可行性。Kissan Joseph 等人[30]认为股票代码的在线检索强度可以作为一个有效的投资者情感表达替代品，因为投资者对股票的正向情感往往意味着即将进行的股票购买决策，而理性投资者在股票购买前需要进行股票信息检索以收集相关信息。如果投资者对某只股票持乐观情感，那么该股票的检索量会上升。利用在线检索强度作为情感表达替代品的优势在于这种指标的获得比起对大量网络评论进行情感分析要容易得多。类似的方法也被 Vasileios L. 等人[36]应用于流行病传播预测研究中，通过监测含有流行病症状描述的微博文数量变化，以获得及时、廉价的流行病状态预测信息。

6 问题及研究展望

6.1 网络用户情感能否正确代表大多数人的情感

在现有基于网络用户情感分析的预测方法中,都以网络用户的情感作为现实中特定对象的预测依据。这些预测方法所隐含的一个前提是网络用户情感能够正确代表大多数现实生活中的人们的真实情感。尽管大多数相关研究最终用相关性分析证明了网络用户情感与待预测对象之间存在相关关系,如果不对此前提进行强调,那么很可能类似的预测方法在应用到其他一些领域时无法得到预想的结果。例如,Tumasjan A. 等人[25]在对推特上涉及德国政治人物或政党的微博文进行分析后发现,经常发布微博文的用户仅占所有发布微博文用户数的3.9%,而他们所发布的政治微博文却占到总数的44.3%,微博中的政治讨论往往受到极少数人的影响。此外,预测方法也很难从社会化媒体用户的所属国别、所属年龄段、所在地域等角度对语料进行选择,并且可能因此造成预测失效。Bollen J. 等人[8]在研究中也承认可能存在这方面的问题,并认为未来需要将地点和语言因素考虑进来,排除地理和文化抽样错误。

6.2 如何全面、正确获取待分析语料

很多相关研究在获取待分析语料时采用检索方法,使用待预测对象相关语词在社会化媒体搜索引擎中检索,以获得最为相关的情感分析语料。例如,在 Asur S. 等人[28]的研究中使用类似"Twilight: New Moon"这样的电影名来检索该电影的相关微博文。但是对于类似"2012"这样名称的电影,仅采用电影名检索的方式很可能获得大量与该电影无关的微博文。一种补救措施是使用导演名、主演等信息进行联合检索,但这无疑增加了获取语料的处理复杂度。

6.3 网络用户情感如何正确分析和统计

在不同的预测领域,网络用户情感表达途径很可能存在差异[46]。在情感分析过程中需要联系具体领域的情感表达特征,对用户情感做出正确判断和统计。例如,Abrahams A. 等人[38]在汽车领域缺陷预测研究中使用 OpinionFinder 来确定负面评论的出现概率时,发现 Harvard General Inquirer[47]中的正面词汇"light"(轻的,光亮的)在汽车行业中却不一定是正面词汇,它有可能表示汽车出故障的一个部件"light"(车灯)。他们还指出,在进行汽车缺陷预测时仅仅依靠情感分析是不够的,还需要将其与问题部件相关联才能进行缺陷危害预测。例如,同样的抱怨程度并不能反映出空调故障和油门踏板故障哪一个是更加严重的缺陷。

7 结语

基于网络用户情感分析进行预测是近几年情感分析应用研究的一个重要领域。目前相关研究已经在证券市场指数预测、电影票房预测、图书销售预测、政治选举预测、传染病传播预测等多个领域进行了有益尝试和探索。本文认为,当前基于情感分析的典型预测方法可归纳为以情感分析结果作为辅助依据的预测方法和以情感分析结果作为主要依据的预测方法。在预测过程中,情感分析源的选择、预测时间提前量的确定以及情感词统计处理方法决定着预测的真实效果。不同领域中网络用户情感所具有的代表性,待分析语料的及时、正确、全面获取,以及网络用户情感的正确分析和统计等问题是后续有待研究的内容。

参考文献

[1] 赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8): 1834 - 1848. (Zhao Yanyan, Qin Bing, Liu Ting. Sentiment analysis[J]. Journal of Software, 2010, 21(8): 1834 - 1848.)

[2] 张紫琼, 叶强, 李一军. 互联网商品评论情感分析研究综述[J]. 管理科学学报, 2010, 13(6): 84 - 96. (Zhang

Ziqiong, Ye Qiang, Li Yijun. Literature review on sentiment analysis of online product reviews[J]. Journal of Management Sciences in China, 2010, 13(6): 84 -96.)

[3] Abbasi A, Chen H, Salem A. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums[J]. ACM Transactions on Information Systems, 2008, 26(3): 1 -35.

[4] Liu H, Lieberman H, Selker T. A model of textual affect sensing using real-world knowledge[C] // Proceedings of the 2003 International Conference on Intelligent User Interfaces, 2003: 125 -132.

[5] Hu M, Liu B. Mining opinion features in customer reviews[C]// Proceedings of 19th National Conference on Artificial Intellgience, 2004: 755 -760.

[6] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval, 2008, 2(1 -2): 1 -135.

[7] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis[C] //Proceedings of Human Language Technologies Conference, 2005: 34 -35.

[8] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1): 1 -8.

[9] Thelwall M, Buckley K, Paltoglou Georgios, et al. Sentiment strength detection in short informal text[J]. Journal of the American Society for Information Science and Technology, 2010, 61(12): 2544 -2558.

[10] Chaovalit P, Zhou L. Movie review mining: A comparison between supervised and unsupervised classification approaches [C]// Proceedings of the 38th Hawaii International Conference on System Sciences, 2005: 1 -9.

[11] Xu K, Liao S, Li J, et al. Mining comparative opinions from customer reviews for competitive intelligence[J]. Decision Support Systems, 2011, 50(4): 743 -754.

[12] Ye Q, Zhang Z, Law R. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches[J]. Expert Systems with Applications, 2009, 36(3): 6527 -6535.

[13] 陆文星，王燕飞. 中文文本情感分析研究综述[J]. 计算机应用研究，2012，29(6)：2014 -2017.（Lu Wenxing, Wang Yanfei. Review of Chinese text sentiment analysis[J]. Application Research of Computers, 2012, 29(6): 2014 -2017.）

[14] 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)：14 -20.（Zhu Yanlan, Min Jin, Zhou Yaqian, et al. Semantic orientation computing based on HowNet[J]. Journal of Chinese Information Processing, 2006, 20(1): 14 -20.）

[15] 张清亮，徐健. 网络情感词自动识别方法研究[J]. 现代图书情报技术，2011(10)：24 -28.（Zhang Qingliang, Xu Jian. Research on automatic extraction of Web sentiment words[J]. New Technology of Library and Information Service, 2011(10): 24 -28.）

[16] 李纲，程洋洋，寇广增. 句子情感分析及其关键问题[J]. 图书情报工作，2010，54(11)：104 -107.（Li Gang, Cheng Yangyang, Kou Guangzeng. Key problems of sentence level sentiment analysis[J]. Library and Information Service, 2010, 54(11): 104 -107.）

[17] 杨经，林世平. 基于SVM的文本词句情感分析[J]. 计算机应用与软件，2011，28(9)：225 -228.（Yang Jing, Lin Shiping. Emotion analysis on text words and sentences based on SVM[J]. Computer Applications and Software, 2011, 28(9): 225 -228.）

[18] 李本阳，关毅，董喜双，等. 基于单层标注级联模型的篇章情感倾向分析[J]. 中文信息学报，2012，26(4)：3 -8.（Li Benyang, Guan Yi, Dong Xishuang, et al. Single-label cascaded model for document sentiment analysis[J]. Journal of Chinese Information Processing, 2012, 26(4): 3 -8.）

[19] 夏云庆，杨莹，张鹏洲，等. 基于情感向量空间模型的歌词情感分析[J]. 中文信息学报，2010，24(1)：99 -103.（Xia Yunqing, Yang Ying, Zhang Pengzhou, et al. Lyric-based song sentiment analysis by sentiment vector space model

[J]. Journal of Chinese Information Processing, 2010, 24(1): 99 - 103.)

[20] 曹树金, 周小又, 陈桂鸿. 网络舆情监控系统中的主题帖自动标引及情感倾向分析研究[J]. 图书情报知识, 2012(1): 66 - 73. (Cao Shujin, Zhou Xiaoyou, Chen Guihong. A study on emotional tendency analysis of topic posts of online public opinion monitoring system[J]. Document, Information & Knowledge, 2012(1): 66 - 73.)

[21] 郑文英. 旅行目的地中文评论的情感分析研究[D]. 哈尔滨: 哈尔滨工业大学, 2010. (Zheng Wenying. Sentiment analysis of travel destination reviews in Chinese[D]. Harbin: Harbin Institute of Technology, 2010.)

[22] Dolan R. Emotion, cognition, and behavior[J]. Science, 2002, 298(5596): 1191 - 1194.

[23] Damasio A. Descarte's error: Emotion, reason, and the human brain[M]. New York: Gosset/Putnam, 1994.

[24] Nofsinger J. Social mood and financial economics[J]. The Journal of Behavioral Finance, 2005, 6(3): 144 - 160.

[25] Tumasjan A, Sprenger T, Sandner P, et al. Predicting elections with Twitter: What 140 characters reveal about political sentiment[C]// Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010: 178 - 185.

[26] Hong S, Nadler D. Which candidates do the public discuss online in an election campaign? The use of social media by 2012 presidential candidates and its impact on candidate salience[J]. Government Information Quarterly, 2012, 29: 455 - 461.

[27] Gruhl D, Guha R, Kumar R. The predictive power of online chatter[C]// Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 2005: 78 - 87.

[28] Asur S, Huberman B. Predicting the future with social media[EB/OL]. [2012 - 10 - 22]. http://arxiv.org/abs/1003.5699.

[29] Liu Y, Huang X, An A, et al. ARSA:A sentiment-aware model for predicting sales performance using blogs[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007: 607 - 614.

[30] Joseph K,Wintoki B, Zhang Z. Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search[J]. International Journal of Forecasting, 2011, 27(4): 1116 - 1127.

[31] Zhang X,Fuehres H, Gloor P. Predicting stock market indicators through Twitter "I hope it is not as bad as I fear"[J]. Procedia-Social and Behavioral Sciences, 2011, 26: 55 - 62.

[32] Wilson T, Hoffmann P,Somasundaran S. OpinionFinder: A system for subjectivity analysis[C]// Proceedings of HLT/EMNLP on Interactive Demonstrations, 2005: 34 - 35.

[33] Wiebe J, Riloff E. Creating subjective and objective sentence classifiers from unannotated texts[C]// Proceedings of Sixth International Conference on Intelligent Text Processing and Computational Linguistics, 2005: 486 - 497.

[34] Pennebaker, James W, Chung C, et al. The development and psychometric properties of LIWC2007[EB/OL]. [2012 - 10 - 22]. http://homepage.psy.utexas.edu/homepage/faculty/pennebaker/Reprints/LIWC2007_LanguageManual.pdf.

[35] Eric G,Karahalios K. Widespread worry and the stock market[C]// Proceedings of the International Conference AAAI on Weblogs and Social Media, 2010, 2(1): 229 - 247.

[36] Vasileios L, Cristianini N. Tracking the flu pandemic by monitoring the social web[C]// Proceedings of Cognitive Information Processing (CIP), 2010 2nd International Workshop on, 2010: 411 - 416.

[37] Gruhl D, Chavet L, Gibson D, et al. How to build a webfountain: An architecture for very large-scale text analytics[J]. IBM Systems Journal, 2004, 43(1): 64 - 77.

[38] Abrahams A, Jiao J, Wang G, et al. Vehicle defect discovery from social media[EB/OL]. [2012 - 10 - 22]. http://www.sciencedirect.com/science/article/pii/S0167923612001017.

[39] Dellarocas C, Zhang X, Awad N. Exploring the value of online product reviews in forecasting sales: The case of motion pictures[J]. Journal of Interactive Marketing, 2007, 21(4): 23 - 45.

[40] Chatfield C. The analysis of time series[M]. Chapman and Hall, 1984.
[41] Walter E. Applied econometric time series[M]. New York: Wiley, 2004.
[42] Freeman, John R. Granger causality and the times series analysis of political relationships[J]. American Journal of Political Science, 1983: 327 - 358.
[43] Hofmann T. Probabilistic latent semantic indexing[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999: 50 - 57.
[44] Norcross C,Guadagnoli E, Prochaska O. Factor structure of the profile of mood states (POMS): Two partial replications [J]. Journal of Clinical Psychology, 2006, 40(5): 1270 - 1277.
[45] Bergsma S,Dekang L, Goebel R. Web-scale N-gram models for lexical disambiguation[C]// Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence (IJCAI - 09), 2009: 1507 - 1512.
[46] Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10 - Ks[J]. Journal of Finance, 2011, 661(1): 35 - 65.
[47] Kelly E, Stone P. Computer recognition of English word senses[M]. North-Holland Linguistic Series, 1975.

徐 健 中山大学资讯管理学院讲师,情报学博士。
通讯地址:广州市广州大学城外环东路132号。邮编:510006。

(收稿日期:2012 - 10 - 28)