

基于多特征融合的微博主题情感挖掘

黄发良^{1),2),4)} 冯 时^{1),3)} 王大玲^{1),3)} 于 戈^{1),3)}

¹⁾(东北大学计算机科学与工程学院 沈阳 110819)

²⁾(福建师范大学软件学院 福州 350108)

³⁾(医学影像计算教育部重点实验室(东北大学) 沈阳 110819)

⁴⁾(福建省公共服务大数据挖掘与应用工程技术研究中心 福州 350108)

摘 要 微博情感分析是社会媒体挖掘中的重要任务之一,在恐怖组织识别、个性化推荐、舆情分析等方面具有重要的理论和应用价值。但与传统文本数据不同,微博消息短小而凌乱,包含着大量诸如微博表情符号之类的特有信息,同时微博情感是与其讨论主题是密切相关的。多数现有的微博情感分析方法都没有将微博主题与微博情感进行协同分析,或者在微博主题情感分析过程中没有考虑将用户关系、用户性格情绪等特征数据,从而导致微博情感分析与主题检测的效果难尽人意。为此,提出了一个基于多特征融合的微博主题情感挖掘模型 TSMMF(Topic Sentiment Model based on Multi-feature Fusion),该模型将情感表情符号与微博用户性格情绪特征纳入到图模型 LDA 中实现微博主题与情感的同步推导。实验结果表明,与当前用于短文本情感主题挖掘的最优模型(JST,SLDA 与 DPLDA)相比较,TSMMF 具有更优的微博主题情感检测性能。

关键词 情感分析;主题检测;LDA;微博短文本;社会媒体处理;自然语言处理

中图法分类号 TP301

DOI号 10.11897/SP.J.1016.2017.00872

Mining Topic Sentiment in Microblogging Based on Multi-feature Fusion

HUANG Fa-Liang^{1),2),4)} FENG Shi^{1),3)} WANG Da-Ling^{1),3)} YU Ge^{1),3)}

¹⁾(School of Computer Science and Engineering, Northeastern University, Shenyang 110819)

²⁾(Faculty of Software, Fujian Normal University, Fuzhou 350108)

³⁾(Key Laboratory of Medical Image Computing of Ministry of Education(Northeastern University), Shenyang 110819)

⁴⁾(Fujian Engineering Research Center of Public Service Big Data Mining and Application, Fuzhou 350108)

Abstract Sentiment analysis in microblogging is an important task in mining social media, and has important theoretical and application value in the terrorist organization identification, personalized recommendation, public opinion analysis, etc. However, different from traditional texts, messages in microblogging are short and irregular, and contain multifarious features such as emoticons, update time and etc, and in microblogging sentiment of a message is closely related to its topic. Most existing sentiment mining approaches cannot achieve cooperating analysis of topic and sentiment of messages in microblogging, or do not take factors such as social relations and users emotional personality into consideration, and this may lead to unsatisfactory sentiment classification and topic detection. To address the issues, a probabilistic model, TSMMF (Topic Sentiment Model based on Multi-feature Fusion) is proposed, which introduces emoticons and microbloggers personality into LDA inference framework, models emotion and personality of microbloggers under the guidance of emotional psychology theory, uses social relations among microbloggers to initialize topics of messages, utilizes Gibbs sampling techniques to estimate parameters in the

收稿日期:2016-03-11;在线出版日期:2016-09-21。本课题得到国家自然科学基金(61370074)资助。黄发良,男,1975年生,博士,副教授,硕士生导师,主要研究方向为数据挖掘、社会媒体处理。E-mail: huangfiliang@163.com。冯 时,男,1981年生,博士,主要研究方向为情感分析、数据挖掘。王大玲,女,1962年生,博士,教授,博士生导师,主要研究领域为数据挖掘、社会媒体处理。于 戈(通信作者),男,1962年生,博士,教授,博士生导师,主要研究领域为数据科学、先进数据库。E-mail: yuge@ise.neu.edu.cn。

model, and finally achieves synchronized detection of sentiment and topic in microblogging. Extensive experiments show that TSMMF outperforms state-of-the-art unsupervised approaches JST, SLDA and DPLDA significantly in terms of sentiment classification accuracy, and compared to the typical semi-supervised sentiment analysis algorithm SSA-ST, TSMMF performs as well as SSA-ST, but unlike SSA-ST, TSMMF can work without labeled training datasets. And so it has promising performance.

Keywords sentiment analysis; topic detection; Latent Dirichlet Allocation; microblogging short text; social media processing; natural language processing

1 引言


以自由开放共享为核心精神的 Web 2.0 使得用户成为互联网的主角, 诸如社交网站、微博和 BBS 论坛之类的平台为网民发表意见和交流情感提供了经济便捷的渠道。一般来说, 用户在微博平台上发表的言论比较简短却又包含着丰富的个人情感与主观倾向性, 例如, 不同读者对于同一条新闻事件持有不同的看法, 不同用户对于某款手机有着其个性化的用户体验, 不同影视爱好者对于同一部电影会留下不同的观影评论, 等等。研究如何高效挖掘隐藏于这些鱼目混杂的微博消息中的主题与情感有助于各级政府机构、企业组织与理性个体的管理决策, 例如, 政府机构可以对网络舆论进行实时监测与导向^[1]、网上商家能够根据用户反馈意见及时调整生产服务实现利润最大化^[2]、个体网民可以敏捷获取目标信息^[3], 等等。


文本情感分析的巨大价值诱导出各式各样的微博情感分析方法, 然而, 这些方法都不同程度地存在着其自身不足, 比如, 基于词典等方法受制于词典覆盖率的问题; 像 word2vec 这种表示学习方法, 虽然可以依赖海量的数据学习出新的表征模型, 但这些模型无法体现个性化; 有监督的情感挖掘方法不同程度地利用训练语料来训练生成文本情感分类器, 一般具有较高的分类准确率, 但获取训练样本的昂贵代价极大地限制此类方法的应用性。因此, 以 JST^[4]、SLDA^[5] 和 DPLDA^[6] 等为代表的无监督情感分类方法近年来备受青睐, 此类方法能有效地避免传统无监督情感分类方法具有的情感词典依赖性缺点, 能达到较好的情感识别效果。然而, 现有的无监督情感分析方法都不同程度地忽略了几个基本事实: (1) 微博消息的情感极性是与微博主题密切相关的, 很少存在游离于主题之外的情感, 脱离主题的

情感挖掘效果难尽人意(值得说明是, “情感极性”在此实际上是一个与主题密切相关的变量, 而不是具体指某种确定的情感极性。这个变量可以取值为积极(赞成、同情)或消极(反对、厌恶), 例如, 在微博话题“厦门公交车纵火案”中, 随着网民对主题事件的了解深入, 网民在发表关于该事件的微博中表现出来的情感极性会发生改变, 由愤怒到同情); (2) 微博具有篇幅短小、不规范与多模态等特性, 许多在传统文本挖掘被视为噪声的数据在微博消息中却显得格外重要, 例如 ASCII 艺术化的表情符号; (3) 微博用户群的性格存在差异性, 而不同性格微博用户的思想感情表达方式不同, 含蓄委婉抑或坦率直接。这些基本事实的忽略导致此类无监督方法的情感分析与主题检测效果难尽人意。下面以微博评论为例加以说明。

一个因新种一批小菜苗而迫切希望天下雨的微博用户在不同时间点发布的两条微博消息 m_1 与 m_2 如下:

m_1 : “sunny day !”—2014/04/03

m_2 : “sunny day again ”—2014/04/04

在对 m_1 与 m_2 进行情感分析时, 现有的无监督情感分类方法会很大可能将 m_1 与 m_2 都归类为积极(positive)型情感, 由于这些传统分析方法仅根据文本信息进行微博情感极性分类, 而 m_1 与 m_2 都含有积极情感词 sunny。然而, 如果将表情符号  加以考虑, 可以发现从 m_2 中传递出来的微博用户情感更可能是消极(negative)型情感, 由于该微博用户在 m_2 中连续使用两个表示伤心的表情符号。另外, 由于 m_1 与 m_2 是同一个微博用户在相距较短的两个时间点上发布的消息, 根据情感心理学研究的结果“用户在相邻时间内表达的情感具有一致性”^[7], 可以推知 m_1 的情感极性也很大可能是消极型。

从此例可以看出, 在进行微博情感分析时若将表情符号视为噪声数据而抛弃, 这会使得微博情感

极性误判的产生成为可能。

另外,社会心理学中的积极心理研究表明^[8-9]:性格乐观者往往会通过不同方式与不同媒介来表达自己的积极情感与态度,而性格悲观者则相反。

基于此,本文提出了一个基于多特征融合的微博主题情感挖掘模型 TSMMF (Topic Sentiment Model based on Multi-feature Fusion),该模型以主题模型 LDA 为基本框架,将表情符与微博用户性格情绪特征因素引入微博消息的情感推理中,利用概率图模型理论与 MCMC (Markov Chain Monte Carlo) 采样技术对模型参数进行推导与估计,最后实现微博消息隐含主题与情感的同步检测。实验表明,与当前用于短文本(主要是指字符数少于 140 的诸如新浪微博、Twitter 等社交媒体)情感主题挖掘的最优模型 (JST, SLDA 与 DPLDA) 相比较, TSMMF 具有更优的微博主题情感检测性能。值得指出的是,虽然 TSMMF 是基于微博主题情感分析应用场景而提出的,但由于诸如微信、QQ、人人网、论坛、Facebook、LiveJournal 等等之类社交网络媒体都有着用户交互信息相对较短、表情符备受欢迎与用户情感倾向流露等特征,将 TSMMF 扩展到其他各种社交网络媒体并非难事,因而其有着广阔的应用前景。

本文第 2 节对微博主题情感挖掘的相关工作进行简介;第 3 节给出 TSMMF 模型的详细描述、推理过程及算法分析;第 4 节给出真实微博数据的测试与比较结果;第 5 节对全文进行总结。

2 相关工作

2.1 文本情感分析

文本情感分析,其实质是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程,目前正在吸引着来自人工智能、数据挖掘、自然语言处理等不同领域研究者的广泛关注^[10-13],涌现出的具体算法不胜枚举,其中以有监督情感分类方法与无监督情感分类方法最为引人注目。

有监督情感分类的基本思想是通过具有情感极性标记的训练样本模型学习并以此训练好的模型对未见文本进行情感分类。该类方法主要涉及训练集获取、文本向量化、分类器训练与分类器检测等步骤。具有里程碑意义的是 2002 年 Pang 等人^[14]应用 3 个代表性分类器(支持向量机 SVM、朴素贝叶斯 NB 与最大熵 ME)对文本情感分类任务进行实验研

究,得出机器学习的文本情感分类性能远高于随机猜选,可达到大致 80% 的准确率。文本向量化过程涉及的特征工程在情感分类任务中起着至关重要的作用,因而,很多研究者围绕着文本情感表示模型展开研究。Mullen 等人^[15]提出基于 WordNet 特征项的 Osgood 情感特征权重值计算机制。Ng 等人^[16]设计出基于 n -gram 的 5 类特征项用于文本情感识别。Kennedy 等人^[17]提出在使用单个单词特征的基础上,通过语境情感值转换器构造双词组形式的上下文敏感的特征项。Xia 等人^[18]提出一种词语关系特征与单个单词特征相结合的文本情感特征模型。Li 等人^[19]提出“短语-句子”的混合表示模型,在短语粒度上运用 Left-Middle-Right 模板与条件随机场 CRF 来提取情感词。近来基于深度学习的文本情感表示备受关注。Bespalov 等人^[20]提出一种基于单词、短语与文档等多级嵌入的短文本情感表示方法,然后利用深度神经网络学习情感的分布式表示向量。Tang 等人^[21]提出面向情感的单词嵌入方法 SSWE (Sentiment Specific Word Embedding) 进行微博情感分析。文献^[22-24]提出利用深度学习模型 CNN (Convolutional Neural Network) 充分挖掘单词、短语与句子信息以提升微博情感分析效果。文献^[25-26]报告了深度学习模型 DBN (Deep Belief Network) 在微博情感分析中的成功应用。

尽管各种基于不同情感表示模型的有监督情感分类方法取得长足的进步,但绝大多数情感表示模型是针对长文本的,很难直接应用于微博情感分析任务。

虽然有监督情感分类器能取得较高的分类准确率,但其训练过程需要大量带情感标签的训练数据,人工标记的高昂代价与机器标记的低劣质量使得有监督情感分类方法的实际应用大打折扣。因而,无需标签数据的无监督情感分类策略因势而动。基于主题模型的情感分类方法是使用最为广泛的无监督情感分类方法,其代表性算法有:Mei 等人^[27]提出一个主题情感模型 TSM 进行主题及其相关情感的演化分析。TSM 一方面存在着类似 pLSI 所有的学习过度问题,另一方面需要相关后处理操作才能完成文档情感的预测。Titov 等人^[28-29]应用 MG-LDA 提取评论对象中的各个被评价,然后提出 MAS 模型对情感进行总结,MAS 模型要求评论对象的每个方面至少在部分评论中被评价过,然而,这对真实评论文本数据集来说是不实际的。Dasgupta 等人^[30]提出一种基于用户反馈的谱聚类技术进行网络文本的无

监督情感分类,聚类分析过程涉及数据特征都是具有情感倾向的主题,然而,在该分析过程中需要人为指定最重要的特征维。Lin 等人^[4]提出一种基于 LDA 模型的 JST 模型,该模型将文本情感标签加入 LDA,形成一个包含词、主题、情感和文档的四层贝叶斯概率模型。后一类方法中的代表性算法有:Hu 等人^[31]提出一种基于 WordNet 的情感词典构建法,首先人工选出情感极性已知的情感词作为种子词,然后迭代进行同义或反义搜索不断将扩展情感词典。Andreevskaia 等人^[32]以不同的种子词集在 WordNet 同义关系图中进行 bootstrap,然后通过运用模糊理论对各个词在所有 bootstrap 结果中表现为积极或消极情感的频数进行规范化得到词的情感值。欧阳继红等人^[33]提出两个多粒度主题情感混合模型:文档级 MGR-JST 与局部 MG-JST。Rao 等人^[34]提出有监督的多标签主题模型 MSTM 和隐含情感主题模型 SLTM 对社交情绪分类。Li 等人^[35]提出一个基于文本主题与用户-商品潜在因子的有监督情感分析模型 SUIT(Supervised User-Item based Topic)。Yang 等人^[36]提出一个用户感知的主题情感模型 USTM,该模型把评论者的人口统计学信息纳入到主题建模过程中。Liu 等人^[37]提出一种主题自适应的半监督微博情感分析模型 TASC(Topic-Adaptive Sentiment Classification)。Tan 等人^[38]提出 FB-LDA 与 RCB-LDA 模型对公众微博情感的演化进行解释。黄发良等人^[39]提出一种新的基于 LDA 和互联网短评行为理论的主题情感混合模型 TSCM(Topic Sentiment Combining Model)。这些微博文本信息情感主题模型都是以微博文本信息为依据对单条微博的主题与情感极性进行分析,虽然一定程度上克服了有监督学习所面临的训练数据难以获取的问题,但由于其对情感表情符与微博用户特征的忽视,导致模型的微博情感分析性能欠佳。

与上述主题情感模型不同,TSMMF 不仅能实现微博情感与主题的协同分析,同时由于微博情感符与微博用户性格情绪特征的加入使得微博情感分析体现出个性化特点。

微博包含着传统文本中所没有的表情符等特有数据,受此启发,近几年越来越多的研究人员在研究如何充分利用表情符来提升微博情感识别性能的方面做了不少尝试性工作。Go 等人^[40]提出一种遥远监督(distant supervision)的方法,通过噪声标签(情感符号😊与😞)的 Twitter 查询结果集来构造情感分类器的训练集。Davidov 等人^[41]提出基于 hashtag

或 smiley 的训练集半自动构造方法。Pak 等人^[42]根据微博表情符自动构造 3 类微博数据集:positive(含 happy 情感符),negative(含 sad 情感符),neutral(主流报刊杂志的微博)。Kouloumpis 等人^[43]运用 Twitter hashtag 构建训练数据集,将 POS 标签特征与词的字典情感值相结合来训练 PNN 三分类器。Liu 等人^[44]提出一个用于微博情感分析的语言模型 ESLAM,该模型的设计者首先利用人工标注数据集训练生成一个语言模型,然后再运用微博表情符对此模型进行平滑,以实现基于训练数据的人工标识法与基于噪声数据的自动标识法无缝结合。Jiang 等人^[45]提出表情符空间模型 ESM,在该模型中微博情感分类任务可分解为两个子任务:首先是将微博映射到 ESM 中,然后根据 ESM 中的距离测度来训练微博情感分类器。Zhao 等人^[46]开发基于表情符的中文微博情感分析系统 MoodLens,根据微博所包含的表情符将微博分类为愤怒(angry)、厌恶(disgusting)、高兴(joy)与悲伤(sad)。Zhang 等人^[47]构建一个以表情符为节点集、表情符对之间的互信息关联度为边权重的加权网络,然后根据此网络对微博进行情感分析。

上述研究^[40-47]表明,充分挖掘表情符号价值能有效提升微博情感分析性能,受此启发,TSMMF 继续推进这条研究思路,然而,与上述工作不同的是,TSMMF 是一种基于概率图模型的无监督情感分析方法,而上述工作都是基于有监督学习思想进行的,比如文献^[40-43]都是为了降低有监督学习所必需的训练集的人工标记成本,而文献^[45,47]是基于表情符空间的有监督情感分析方法。

2.2 情感表达理论

正如达尔文所说“面部与身体的富于表达力的动作有助于发挥语言的力量”的那样,人类在面对面的交流中常常会借助和利用自己的面部表情、手势动作、身体姿态的变化来简便直观地表达和传递思想感情,然而,微博空间的虚拟性使得功能强大的肢体语言失去了其赖以生存的物理基础。为了弥补肢体语言的缺失,以新浪微博、Twitter 等为代表的国内外微博平台(新浪微博、Twitter)都推出其自身特有的表情符号系统。借此表情符号,微博用户可以迅速直观地表达自己的观点与情感,例如,新浪的开心😊,愤怒😡与伤心😞。另有来自社会神经系统科学的研究表明^[48],在面对表情符号时,人类大脑不是将其简单地处理为一个标点符号而是将其视为真实物理行为进行响应。

人工心理学研究表明,个体的遗传基因及生长环境的不同使得每个个体都具有不同性格,人类个体的情感反应成因于外部输入的奖惩信号超过某一阈值.具有不同性格的不同个体对相同输入信号的情感反应并不是完全相同的,而是在一个可接受的范围内变化的,各不相同.对个体而言,其性格影响着他的积极与消极情绪的变化范围和变化率,也就是说,一个时间步长内,情感变化多快以及变化到何种程度,是性格的函数.相关学者在基于性格的情感建模研究中做了不少成功的尝试. Kshirsagar 等人^[49]提出一个“性格-心情-情感”多层情感模型用于模拟生成人的面部表情. Yang 等人^[50]构建一个基于情绪心理学的情感、心情和性格相结合的情感模型以人工合成表情. 李海芳等人^[51]设计出基于性格、心情和情感多层模型的情感预测系统,提出了一种适应不同性格特征的 HMM 情感模型. 类似地有,微博用户的性格特征与其在社交媒体上情感表述方式与主题偏好有着很大相关性^[52-53].

3 主题情感挖掘模型

在这一部分我们首先对与 TSMMF 模型密切相关的两个模型(主题模型 LDA 与主题情感模型 JST)进行简介,然后对 TSMMF 模型进行详细描述,讨论 TSMMF 模型的相关参数推理机制与模型参数先验值的确定,在此基础上提出一种 TSMMF 模型的微博主题情感挖掘算法. 模型相关符号及说明如表 1 所示.

表 1 符号说明

类型	符号	备注
随机变量	m	微博消息
	u	微博用户
	t	主题
	s	情感极性
	w	词语
分布及相关参数	e	表情标签
	α	Dirichlet 分布 $A(m, s, t)$ 的先验参数
	β	Dirichlet 分布 $B(s, t, w)$ 的先验参数
	η	Dirichlet 分布 $H(m, s)$ 的先验参数
	ζ	Dirichlet 分布 $Z(s, t, e)$ 的先验参数
算法预定参数	λ	分布 $H(m, s)$ 的性格参数
	T	微博集隐含的主题数
	S	微博情感类别数
微博消息集统计量	$PROB$	词语与表情标签的比例
	W	微博消息集的词典大小
	E	微博消息集的情感标签库大小
	M	微博消息集的微博数
	V	微博消息集的词语总计数
	P	微博消息集的情感标签总计数
	U	微博消息集的博主总计数

3.1 模型描述

LDA 模型是 Blei 等人于 2003 年提出的“文档-主题-单词”三层贝叶斯模型(图 1(a)),通过运用概率推导方法来寻找数据集的语义结构,从而得到文本的主题.该模型是建立在如下假设之上的:文档是由不同主题组成的,而一个主题是单词集合的概率分布,在此假设下,文档单词的产生步骤可以分为两个阶段,首先从文档-主题分布中选择一个主题,然后根据随机选择的主题从主题-单词分布中选择一个单词.牢固的数学基础与良好的扩展性使得 LDA 在文本主题挖掘研究中广为使用,但情感层的缺失使得 LDA 无法完成文档情感的分析.

为了使得能够同时对主题与情感进行挖掘, Lin 等人提出 JST 模型对 LDA 进行嵌入情感层的改造,模型的图表示可参见图 1(b).在 JST 中,情感标签与文档(产品短评)关联,主题标签与情感标签关联,在此基础上,词语同时与主题和情感标签关联.与 LDA 相比较, JST 是一个可以同时进行文本情感与情感分析的 4 层图模型.

考虑到微博消息的固有特质与微博博主的社会性,类似 JST 对 LDA 的改造方式,本文提出基于社会学理论与人工心理学的微博主题情感挖掘模型 TSMMF(图 1(c)).对比图 1 中的 3 个图模型可以发现,在 TSMMF 中,除了嵌入情感层之外,还进行两个方面的改造:(1)引入表情符号变量 e 以更准确

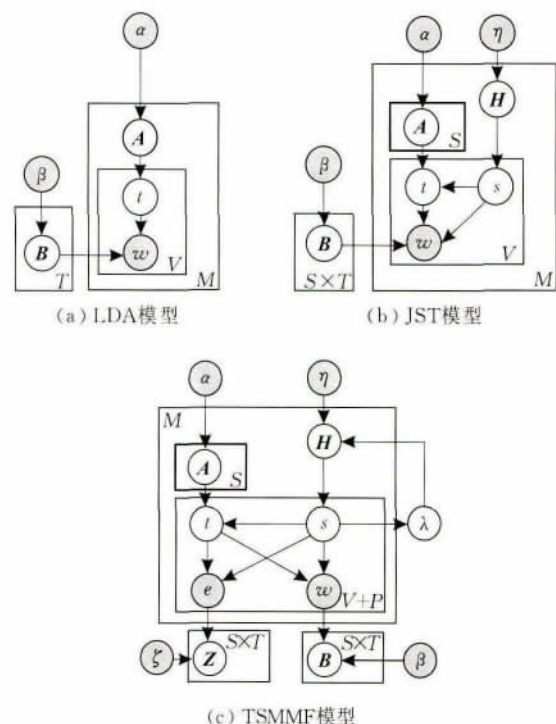


图 1 3 个图模型

地捕捉微博消息的隐含情感;(2)新增性格情绪参数 λ 以在模型迭代推导微博主题情感的过程中对微博消息的情感进行调整.与图1(c)相对应的TSMMF微博消息的生成过程可简述如下:

对于微博消息集 $C = \{m_1, m_2, \dots, m_M\}$,其中 M 为集合的基数,与微博消息集 C 对应的词典的大小为 W ,微博 m_i 是由 W_i 个单词与 E_i 表情符号组成的序列,即 $m_i = \{w_1, w_2, \dots, w_{W_i}; e_1, e_2, \dots, e_{E_i}\}$.TSMMF产生微博消息集 C 的过程可简单归结为如下两个步骤:(1)初始化TSMMF模型的先验分布参数 $\Theta = \{\alpha, \beta, \eta, \zeta\}$ 并以此参数进一步生成分布 A, B, H 和 Z ,其中 α, β, η 与 ζ 分别服从狄利克雷分布 $\text{Dir}(\alpha), \text{Dir}(\beta), \text{Dir}(\eta)$ 与 $\text{Dir}(\zeta)$, α 是指具有情感极性 s 的主题 t 在文档 m 中出现的先验次数, β 是指单词 w 在具有情感极性 s 的主题 t 中出现的先验次数, η 是指情感极性 s 在文档 m 中出现的先验次数, ζ 是指情感标签 e 在具有情感极性 s 的主题 t 中出现的先验次数;(2)概率生成微博消息集 C 的单词与情感标签,此生成过程可简单描述如下:首先利用微博-情感分布 $H(m, \cdot)$ 生成微博消息 m 的情感极性 s ,其中 s 服从 $\text{Mul}(H(m, \cdot))$ 分布;接着根据产生的情感极性 s ,应用(微博,情感)-主题分布 $A(m, s, \cdot)$ 生成具有情感极性 s 的微博消息 m 的主题 t , t 服从 $\text{Mul}(A(m, s, \cdot))$ 分布;最后根据前面生成的微博消息 m 的情感 s 与主题 t 产生 m 中的单词 w 或者表情符 e ,具体地,若随机数 rand 大于给定的情感标签单词比例 PROB 则微博消息 m 的当前位置为单词 w ,否则为表情符 e .TSMMF微博消息生成过程可形式化为算法1,为了方便理解,我们给出与其对应的流程图(图2).

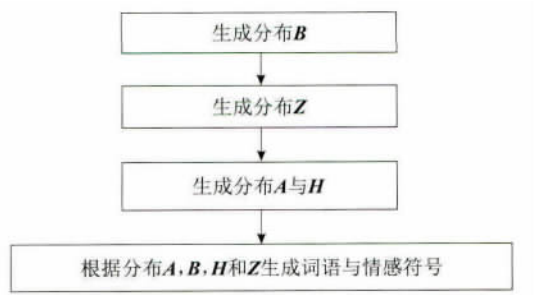


图2 微博消息生成过程流程图

算法1. 微博消息生成过程.

//第1部分负责生成分布 B ,其伪代码为:

1. for each $s \in \{1, 2, \dots, S\}$
2. for each $t \in \{1, 2, \dots, T\}$
3. for each $w \in \{1, 2, \dots, W\}$
4. choose $B_{s,t,w} \sim \text{Dir}(\beta)$

//第2部分负责生成分布 Z ,其伪代码为:

5. for each $s \in \{1, 2, \dots, S\}$
6. for each $t \in \{1, 2, \dots, T\}$
7. for each $e \in \{1, 2, \dots, E\}$
8. choose $Z_{s,t,e} \sim \text{Dir}(\zeta)$

//第3部分负责生成分布 A 与 H ,其伪代码为:

9. for each microblog $m \in \{1, 2, \dots, M\}$
10. for each $s \in \{1, 2, \dots, S\}$
11. choose $H_{m,s} \sim \text{Dir}(\eta)$
12. for each $t \in \{1, 2, \dots, T\}$
13. choose $A_{m,s,t} \sim \text{Dir}(\alpha)$

//第4部分根据分布 A, B, H 和 Z 生成词语与情感符号,其伪代码为:

14. for each word w or emoticon e in document m
15. choose $s \sim \text{Mul}(H_m)$
16. choose $t \sim \text{Mul}(A_{ms})$
17. if $\text{rand} > \text{PROB}$ then choose $w \sim \text{Mul}(B_{st})$ else choose $e \sim \text{Mul}(Z_{st})$

由TSMMF的图模型与产生式过程可以看出,TSMMF具有如下特征:(1)主题表示上,TSMMF利用单词与表情符号等不同特征来建模微博主题,而JST将微博文本视为微博主题的唯一基础数据;(2)情感捕捉上,TSMMF不仅通过单词与表情符号等不同特征来挖掘微博情感,还通过性格情绪参数来建模博主性格情绪对其微博情感的调节与影响;(3)从主题情感空间容量上看,TSMMF中空间容量要远大于JST的空间容量,即 $V+P>V$,这意味着TSMMF具有比JST更强大的微博主题情感表示能力;(4)模型扩展性,尽管当前TSMMF仅考虑了微博中的词语与表情符号,但其实很多其他含有主题与情感含义的微博特征可以很容易加入到TSMMF,因而,TSMMF具有良好的可扩展性.

3.2 性格情绪建模

从心理学上讲,性格是指表现在人对现实的态度和相应的行为方式中的比较稳定的、具有核心意义的个性心理特征,主要体现在对自己、对别人、对事物的态度和所采取的言行上,它是一种与社会相关最密切的人格特征.尽管心理学研究者提出诸多模型(三因素模型、七因素模型、Big Five模型等)对性格进行量化研究,但他们多数都是采用基于小规模调查数据的统计分析策略.而以诸如微博之类的大量社交媒体数据为数据源的研究成果少见报道.

用户群在微博上的言行表现是一种复杂系统行为,用户可以采用不同模态形式的媒介(文本、图像、音频与视频等)、不同种类的语言(姿态语言、文字语言等)与不同交互行为(点赞、收藏、评论等)来表现自己对事物的态度与情感.

作为人的个性心理特征,性格有着多种不同方面的特征,勇敢或怯懦的意志特征,乐观或悲观的情绪特征、思维敏捷或思维迟缓的理智特征.对性格进行基于所有特征的全方位建模是不现实的,受社会科学[8-9]的启发,我们选取情绪特征对性格建模.在 TSMMF 模型中,我们将用户性格情绪特征简单表征如下:对微博博主 u 而言,其性格情绪参数 λ 是在其发布的主观性微博消息中具有情感 s 的微博消息的比例,可形式化为式(1)

$$\lambda_s = \frac{US[u][s]}{\sum_{s \in SSet} US[u][s]} \quad (1)$$

其中: US 是微博博主与情感类型之间的关联矩阵, $US[u][s]$ 是指微博博主 u 发布的具有情感 s 的微博消息数.本文假设一条微博消息的情感极性是“非消极即积极”的,也就是说, $SSet = \{POS, NEG\}$.

由于在模型参数推理过程(3.3节中的吉布斯迭代采样过程)中,微博主题情感相关分布 A 、 B 、 H 与 Z 是动态变化的,因而,性格参数 λ 是依据关联矩阵 US 来自适应调节的.

虽然 TSMMF 模型仅用正负情感微博比例来量化微博用户性格情绪特性,但该模型具有很好的扩展性,可以比较容易实现用户发布频率、发布微博总数量等更多因子的纳入以更加全面刻画用户性格情绪特性.无疑,更多因子的纳入会大大增加模型的复杂度,因而,如何选择合适的因子以使模型复杂度与情感识别性能达成最优平衡是我们下一步要开展的工作.

3.3 模型推理

由 TSMMF 微博消息的生成过程可知,分布 A 、 B 、 H 与 Z 是模型推理的核心.通过对这些分布的估计,可以容易推导出每条微博消息的主题与情感极性信息.为了获得这些分布,我们需要对如下后验分布进行估计:

$$p(t_i = t, s_i = s | t_{-i}, s_{-i}, w, e, \alpha, \beta, \eta, \zeta) \quad (2)$$

其中: t_{-i} 与 s_{-i} 分别是指除了微博 m 中第 i 个位置上的元素(词语或表情符号)以外的其他元素(词语或表情符号)的主题向量与情感极性向量.

当前已经提出很多估值概率图模型参数的方法(变分法(variational inference)[54],期望传播法(expectation propagation)[55]与吉布斯(Gibbs Sampling)[56]).吉布斯采样是一种快速高效的 MCMC (Markov Chain Monte Carlo) 采样方法,它可以通过迭代的采样方式对复杂的概率分布实现推导.考虑到吉布斯采样的简单高效,TSMMF 模型参数的估值采用吉布斯采样方法.

要估计后验分布(式(2))就必须计算词语 w 、表情符 e 、主题 t 与情感 s 的联合分布 $P(w, e, t, s)$, 根据图 1(c) 与概率图模型理论可知:

$$P(w, e, t, s) = P(w, e | t, s) P(t, s) \\ = P(w | t, s) P(e, | t, s) P(t | s) P(s) \quad (3)$$

下面通过欧拉展开方法对联合分布 $P(w, e, t, s)$ 的各个因子分别进行求解,也即是通过分布 B 、分布 Z 、分布 A 与分布 H 进行积分实现因子 $P(w | t, s)$ 、 $P(e | t, s)$ 、 $P(t | s)$ 与 $P(s)$ 的推导,具体见式(4)、式(5)、式(6)与式(7).

$$P(w | t, s) = \left(\frac{\Gamma(V\beta)}{[\Gamma(\beta)]^W} \right)^{S \times T} \times \prod_{s=1}^S \prod_{t=1}^T \frac{\prod_{w=1}^W \Gamma(n_{s,t,w} + \beta)}{\Gamma(n_{s,t} + W\beta)} \quad (4)$$

其中: $n_{s,t,w}$ 表示单词 w 同时属于主题 t 和情感极性 s 的频数, $n_{s,t}$ 表示所有同时属于主题 t 和情感极性 s 的单词总的频数, $\Gamma(*)$ 表示伽马函数.

$$P(e | s, t) = \left(\frac{\Gamma(E\zeta)}{[\Gamma(\zeta)]^E} \right)^{S \times T} \times \prod_{s=1}^S \prod_{t=1}^T \frac{\prod_{e=1}^E \Gamma(p_{s,t,e} + \zeta)}{\Gamma(p_{s,t} + E\zeta)} \quad (5)$$

其中: $p_{s,t,e}$ 表示表情符 e 同时属于主题 t 和情感极性 s 的频数; $p_{s,t}$ 表示所有同时属于主题 t 和情感极性 s 的表情符总频数.

$$P(t | s) = \left(\frac{\Gamma(T\alpha)}{[\Gamma(\alpha)]^T} \right)^{M \times S} \times \prod_{m=1}^M \prod_{s=1}^S \frac{\prod_{t=1}^T \Gamma(n_{m,s,t} + \alpha)}{\Gamma(n_{m,s} + T\alpha)} \quad (6)$$

其中: $n_{m,s,t}$ 表示微博消息 m 中的元素(词语或表情符号)同时属于主题 t 和情感极性 s 的频数; $n_{m,s}$ 表示所有同时属于主题 t 和情感极性 s 的微博消息 m 中的元素(词语或表情符号)总频数.

$$P(s) = \left(\frac{\Gamma(S(\eta + \lambda))}{[\Gamma(\eta + \lambda)]^S} \right)^D \times \prod_{m=1}^M \prod_{s=1}^S \frac{\Gamma(n_{m,s} + \eta + \lambda)}{\Gamma(n_m + S(\eta + \lambda))} \quad (7)$$

其中: $n_{m,s}$ 表示在微博 m 中情感极性 s 出现的频数; n_m 表示微博 m 中情感极性标签的总频数,即微博 m 的元素总数.

根据上述联合概率可进一步得到微博主题情感的后验分布如下:

$$p(t_i = t, s_i = s | t_{-i}, s_{-i}, w, e) \propto \frac{\{n_{m,s}^t\}_{-i} + \alpha}{\{n_{m,s}\}_{-i} + T\alpha} \times \frac{\{n_{s,t}^w\}_{-i} + \beta}{\{n_{s,t}\}_{-i} + W\beta} \times \frac{\{p_{s,t}^e\}_{-i} + \zeta}{\{p_{s,t}\}_{-i} + E\zeta} \times \frac{\{n_m^s\}_{-i} + \eta + \lambda_s}{\{n_m\}_{-i} + \sum_{s=1}^S (\eta + \lambda_s)} \quad (8)$$

$$p(t_i=t, s_i=s | t_{-i}^w, s_{-i}^w, w) \propto \frac{\{n_{m,s}^t\}_{-i} + \alpha}{\{n_{m,s}\}_{-i} + T\alpha} \times \frac{\{n_{s,t}^w\}_{-i} + \beta}{\{n_{s,t}\}_{-i} + W\beta} \times \frac{\{n_m^s\}_{-i} + \eta_s + \lambda_s}{\{n_m\}_{-i} + \sum_{s=1}^S (\eta_s + \lambda_s)} \quad (9)$$

$$p(t_i=t, s_i=s | t_{-i}^e, s_{-i}^e, e) \propto \frac{\{n_{m,s}^t\}_{-i} + \alpha}{\{n_{m,s}\}_{-i} + T\alpha} \times \frac{\{p_{s,t}^e\}_{-i} + \zeta}{\{p_{s,t}\}_{-i} + E\zeta} \times \frac{\{n_m^s\}_{-i} + \eta_s + \lambda_s}{\{n_m\}_{-i} + \sum_{s=1}^S (\eta_s + \lambda_s)} \quad (10)$$

其中: t_{-i}^w 与 s_{-i}^w 分别表示除了微博 m 中的第 i 个位置的词语外, 其他所有词语的主题向量与情感向量; t_{-i}^e 与 s_{-i}^e 分别表示除了微博 m 中的第 i 个位置的表情符外, 其他所有表情符的主题向量与情感向量; $\{n_{m,s}^t\}_{-i}$ 表示在微博 m 中, 除微博 m 中第 i 个位置上的特征外, 具有主题 t 和情感极性 s 的特征总计数; $\{n_{m,s}\}_{-i}$ 表示在微博 m 中, 除微博 m 中第 i 个位置上的特征外, 具有情感极性 s 的特征总计数; $\{n_{s,t}^w\}_{-i}$ 表示在微博集合 C 中, 除微博 m 中第 i 个位置上的元素外, 单词 w 同时属于主题 t 和情感极性 s 的频数; $\{n_{s,t}\}_{-i}$ 表示在微博集合 C 中, 除微博 m 中第 i 个位置上的元素外, 所有其他具有主题 t 和情感极性 s 的词语总计数; $\{p_{s,t}^e\}_{-i}$ 表示在微博集合 C 中, 除微博 m 中第 i 个位置上的元素外, 表情符 e 同时属于主题 t 和情感极性 s 的频数; $\{p_{s,t}\}_{-i}$ 表示在微博集合 C 中, 除微博 m 中第 i 个位置上的元素外, 所有其他具有主题 t 和情感极性 s 的表情符总计数; $\{n_m^s\}_{-i}$ 均表示在微博 m 中, 除微博 m 中第 i 个位置上的特征外, 具有情感极性 s 的特征总计数; $\{n_m\}_{-i}$ 表示在微博 m 中, 除微博 m 中第 i 个位置上的特征外, 所有其他特征的总计数。

根据上述后验分布, 可以借助吉布斯采样技术从 Markov 链中获得采样样本, 从而实现主题情感分布 A, B, H 与 Z 的近似。

$$A_{m,s}^t = \frac{n_{m,s}^t + \alpha}{n_{m,s} + T\alpha} \quad (11)$$

$$B_{s,t}^w = \frac{n_{s,t}^w + \beta}{n_{s,t} + W\beta} \quad (12)$$

$$Z_{s,t}^e = \frac{p_{s,t}^e + \zeta}{p_{s,t} + E\zeta} \quad (13)$$

$$H_m^s = \frac{n_m^s + \eta_s + \lambda_s}{n_m + \sum_{s=1}^S (\eta_s + \lambda_s)} \quad (14)$$

3.4 模型先验

研究表明^[40], 合理地利用先验知识可以有效提高无监督学习算法的性能. 为了提升 TSMMF 的微博主题情感学习能力, 本文从词语情感先验、表情符情感先验与主题先验这 3 个方面来充分利用数据中蕴含的先验知识。

(1) 词语情感先验. 微博是一个论题开放的平台, 从而微博数据集具有极强的主题跨域性, 单一情感词典往往存在覆盖率问题, 这是由于现有很多情感词典都是根据语料库来产生的, 而语料库涵盖主题范围总是有限的. 另外, 微博交互是网络文化的一种重要表现形式, 微博消息中往往会出现一些新情感词. 因此, 我们选择 3 个广为使用的情感词典 (NTUSD、SentiWordNet 与 HowNet) 进行组合, 在组合过程中以 HowNet 为基准. 词语情感先验赋值的具体操作为: 对于微博消息集合的词典中的每个词语 w , 若 w 出现在新构建的情感词典中则赋以对应的情感值, 否则赋以随机的情感值。

(2) 表情符情感先验. 由于表情符能在微博交互的过程中起到很好的传情达意作用, 不同微博系统设计各自独立的表情符号系统. 尽管这些表情符号系统所包含的符号种类繁多, 但只有少部分符号为绝大多数微博用户在使用. 在此, 我们从中选择出 20 个表情符号, 分别构造积极情感表情符号集合 $POSE = \{:,), :-), :P, =), :], (:, :-), XD, D:, =D, =]\}$ 与消极情感表情符号集合 $NEGE = \{:/, :o, :-(, :(, =/, =(, :-||, :-(\}$. 要指出的是, 虽然我们只选择这 20 种表情符, 但这 20 种表情符具有很强的代表性, 例如, 在我们的实验数据集中, 这 20 种表情符的数量在所有表情符总计数中占 95% 以上. 当然, 从理论上讲, 那些低频度微博表情符的抛弃有浪费之嫌, 但是表情符种类的增加同样也会带来模型的时间复杂度, 由于算法时间复杂度是一个与表情符种类数密切相关的数量级 $O(M \times (V + P))$.

(3) 主题先验. 在微博系统中, 微博消息的主题表达有显式与非显式之分. 显式主题是指微博发布者通过井号 (#) 标签对来明确一条微博的所属主题. 类似 Hu 等人^[57] 提出的基于社交关系的微博情感分析方法, 我们将此思想应用于微博主题先验初始化。

值得说明的是, 情感词语与表情符号情感极性、主题属性等先验知识的应用仅仅是为了给生成模型 TSMMF 以更好的初始化, 这与需要大量带

情感极性标签的训练数据的有监督微博情感学习算法是不同的,当然,若从先验知识应用的角度上看,TSM MF 模型可视为一个弱监督模型.

综合上述,我们设计了一个模型先验初始化算法 PKInitializer(Prior Knowledge Initializer). 算法 PKInitializer 主要可分为两大部分:第 1 部分(步骤 1 与步骤 2)负责对微博词语与表情符赋予初始情感值;第 2 部分(步骤 3~步骤 7)主要负责基于微博用户关系的微博主题词集构造.

算法 2. PKInitializer.

输入: 微博数据集 $C = \{m_1, m_2, \dots, m_M\}$

输出: 被赋予情感先验值的情感词语与表情符、提取的微博主题词集

BEGIN

1. 利用情感词典对微博词语赋予初始情感值;
2. 根据表情符集对表情符赋予初始情感值;
3. 根据发布者身份将微博消息集合划分为 $\{M_1, M_2, \dots, M_u, \dots, M_U\}$, 其中 $M_u = \{EM_u, IM_u\}$, M_u 表示用户 u 发布的微博消息, EM_u 表示用户 u 发布的具有显式主题的微博消息, IM_u 表示用户 u 发布的具有隐式主题的微博消息;
4. 对于具有显式主题的微博 EM_u , 从井号标签对“#……#”中随机提取词语作为微博主题;
5. 重复如下操作对 IM_u 中的消息进行主题初始化: 从 EM_u 选择消息 m_1 初始化 IM_u 中的消息 m 的主题, 消息 m_1 必须满足条件: m_1 的发布时间与 m 的发布时间的距离 $dist$ 最小且 $dist$ 要小于指定的时间阈值;
6. 从所有用户中选择 10 个与用户 u 在指定窗口内具有最高交互频数的用户, 然后将这 10 个用户的微博消息对用户 u 的消息 m 进行主题初始化, 上述操作若成功则转至步骤 5 初始化 IM_u 中的下一条消息的主题;
7. 随机初始化消息 m 的主题, 并转至步骤 5 初始化 IM_u 中的下一条消息的主题

END

3.5 算法描述与分析

为了更好地理解 TSM MF 模型的微博主题情感分析过程, 本节将此过程形式化为算法 TSM MF_Miner. 该算法步骤可划分为 4 个部分: (1) 第 1 部分(步骤 1)负责微博数据的预处理, 主要包括微博特征(词语和表情符)的提取(中文分词、英文 stemming、表情符提取)、利用先验知识对微博特征项的情感与主题进行初始赋值等; (2) 第 2 部分(步骤 2)完成模型超参数的初始化, 为叙述方便, 令与词语 w 相关的变量集 $VW = \{n_{s,t,w}, n_{m,s,t}, n_{s,t}, n_{m,s}, n_m, US\}$, 与表情符号 e 相关的变量集 $VE = \{p_{s,t,e}, p_{s,t}, n_{m,s,t}, n_{m,s}, n_m, US\}$; (3) 第 3 部分(步骤 3~步骤 6)主要通过 Gibbs

采样技术实现 TSM MF 模型参数的估值; (4) 第 4 部分(步骤 7)根据微博情感分布 H 与微博主题情感分布 A 进行微博消息的主题情感判别, 具体为: 对于消息 m , 若 $H_m^{Pos} > H_m^{Neg}$ 则该消息的情感极性为 Pos , 否则为 Neg .

算法 3. TSM MF_Miner.

输入: 微博数据集 $C = \{m_1, m_2, \dots, m_M\}$

输出: 微博数据集 C 中的每条消息 m_i 情感极性

BEGIN

1. 微博数据预处理并调用 PKInitializer 初始化微博特征(词语和表情符)的情感与主题;
2. 初始化超参数 $(\alpha, \beta, \eta$ 与 $\zeta)$, 分布 $(A, B, Z$ 与 $H)$ 和矩阵 US ;
3. 对微博数据集 C 中的每个特征(词语 w 和表情符 e) 执行如下操作: 分别从变量集 VW 与 VE 中排除具有情感 s 与主题 t 的词语 w 与表情符 e , 并根据式(9)与式(10)分别为 w 和 e 构造新的情感主题对 (s_{new}, t_{new}) , 并运用情感标签 s_{new} 与主题标签 t_{new} 更新变量集 VW 与 VE , 最后根据式(8)更新 λ ;
4. 迭代次数加 1;
5. 每 100 次迭代运用新采样结果对分布 A, B, H 与 Z 更新 1 次;
6. 重复步骤 3 至步骤 5 直到指定迭代次数;
7. 主题情感判别: 对于消息 m , 若 $H_m^{Pos} > H_m^{Neg}$ 则该消息为积极情感, 否则为消极情感.

END

由于 B 是一个 $S \times T \times W$ 的矩阵, 生成 B 需要循环 $S \times T \times W$ 次, 所以产生 B 的时间复杂度是 $O(S \times T \times W)$, 由于 Z 是一个 $S \times T \times E$ 的矩阵, 生成 Z 需要循环 $S \times T \times E$ 次, 所以产生 Z 的时间复杂度是 $O(S \times T \times E)$, 由于 H 是一个 $M \times S$ 的矩阵, 所以产生 H 的时间复杂度是 $O(M \times S)$, 由于 A 是一个 $M \times S \times T$ 的矩阵, 所以产生 A 的时间复杂度是 $O(M \times S \times T)$. 对于产生语料库 C 中每个单词的过程, 重复这个过程 $M \times (V + P)$ 次, 所以这个过程的时间复杂度是 $O(M \times (V + P))$. 由于 $O(M \times S) < O(M \times S \times T) < O(S \times T \times E) < O(S \times T \times W) < O(M \times (V + P))$, 所以算法的时间复杂度为 $O(M \times (V + P))$.

4 实验与分析

为了定量地分析 TSM MF 模型的性能, 我们用 3 个不同的真实微博数据集进行实验, 分别从情感分类准确率、主题提取、主题数对准确率的影响、表情符号对准确率的影响、博主性格情绪参数对准确率的影响以及用户活跃度对情感分类正确率的影响

等多方面进行分析. 实验环境为: CPU 为 Intel(R) Core(TM) i7-2600, 内存 8GB, OS 为 Windows 7.

4.1 数据集

由于微博主题情感分析研究目前还处于萌芽状态, 再加上诸如新浪、Twitter 之类的微博平台处于隐私安全需要对其提供的微博服务加以各种不同限制条件, 从而导致在科研上很少有用于实验比较的标准数据集. 因此, 我们通过调用新浪微博 API 接口编写网络爬虫来构造实验数据集(见表 2). 对于采集到的微博数据, 我们进行如下预处理: 标识低活跃度的微博用户(在一个采集周期(10 天)内所发微博数小于 3 条的用户)与广告推广用户, 并将这些用户及其微博消息删除. 预处理后的微博数据集(First-Ten、Mid-Ten 与 Last-Ten)分别是微博用户在持续时间段(2011/12/1~2011/12/10、2012/1/1~2012/1/10 与 2012/2/1~2012/2/10)发表的微博(这 3 个数据集又统称为 Data-old), 并通过人工方式对 3 个数据集中的微博消息进行情感标记. 考虑到前面使用的实验微博数相对久远, 可能难以捕捉到微博发展现状的特性, 我们另外采集并构造新近时间(2016/3/10~2016/3/19)内的数据集 Data-new, 并进行类似处理. 在微博情感人工标注的过程中, 我们请 3 个微博情感标注者对采集微博数据进行情感极性标注, 并对标注结果的一致性进行 Kappa 检验, 检测结果见表 2. 对于情感标注不一致的微博, 我们根据 high-voting 的投票原则来确定其情感极性归属. 从表 2 可以看出, 与数据集 First-Ten 和 Last-Ten 相比较, Mid-Ten 的情感人工标注结果一致性较低, 这说明 Mid-Ten 中的微博情感模糊性相对较强, 可能会给微博情感自动分析提出更大的挑战. 在表 3 中, #Pos 与 #Neg 分别是指积极情感与消极情感消息数, EM 是指含有表情符号的消息数, EMP 是指含有表情符号的积极情感消息数, AVL 是指消息平均长度, U 是微博用户数, E 是指表情符号数, M 是指微博消息数. 从表 2 可以看出, Data-new 的标注一致性高于其他 3 个数据集, 类似地从表 3 可知, Data-new 的表情符比例与积极情感比例也要高于其他 3 个数据集.

表 2 微博情感标注一致性检测

数据集	标注者对(1-2)	标注者对(1-3)	标注者对(2-3)
First-Ten	0.829	0.831	0.828
Mid-Ten	0.783	0.802	0.794
Last-Ten	0.828	0.815	0.836
Data-new	0.835	0.847	0.842

表 3 实验数据及其统计特征

数据集	U	E	M	#Pos	#Neg	EM	EMP	AVL
First-Ten	121	243	12364	7053	5311	6264	2749	42.76
Mid-Ten	98	249	10972	5158	5814	4376	2533	44.29
Last-Ten	128	229	11875	6983	4892	5421	2645	41.12
Data-new	146	387	12208	8136	4072	8295	6192	56.68

4.2 情感分类正确率分析

为了评价 TSMMF 的微博情感识别能力, 考虑到 TSMMF 模型学习的无监督性, 我们将其与当前最具代表性的无监督情感学习模型(JST^[4], SLDA^[5]和 DPLDA^[6]), 半监督情感学习模型 SSA-ST^[57]和基于 SVM 的有监督情感学习模型在 3 个数据集(First-Ten, Mid-Ten 与 Last-Ten)进行微博情感分类正确率 ACC 比较. 要加以说明的是, 该实验中 SVM 算法具体为基于分词特征(1-gram+2-gram)的 libSVM. 通过独立同分布的随机抽样, 对数据集(First-Ten, Mid-Ten 与 Last-Ten)分别构造 8 组实验数据集. 值得指出的是, SVM 是采用 10-fold 方式进行训练的, 因而在每一次实验中其训练集是从各组数据中随机抽取的. 实验结果分别见表 4~表 7, 表中的 Avg 表示各组数据分类正确率的平均值(Average). 由上述表可以看出: (1) TSMMF 在数据集 First-Ten, Mid-Ten 与 Last-Ten 上的情感分类表现分别为最好(70.75%)、最差(66.81%)与次差(69.15%). 结合表 3 的统计结果(First-Ten, Mid-Ten 与 Last-Ten 的表情符比率分别是 50.66%, 39.88%和 45.65%), 不难发现, 数据集所含表情符的比例对 TSMMF 的情感分类正确率有着重要影响; (2) 对比 4 个表的列(TSMMF, JST, SLDA 和 DPLDA), 尤其是 Avg 行, 可以发现 TSMMF 的微博情感分类正确率远高于其他 3 种非监督情感分类算法 JST, SLDA 和 DPLDA, 这说明表情符数据的利用与用户性格情绪特征的引入能有效提升微博情感分析的有效性; (3) 与半监督情感学习模型 SSA-ST 相比较, TSMMF 在数据集 First-Ten, Mid-Ten 与 Last-Ten 上的情感识别率稍逊一筹, 而在 Data-new 上 TSMMF 的情感分类正确率要略高; (4) 检测 4 个表中的最大 ACC 可以发现, 所有最大的 ACC 都出现在 SVM 中, 这说明 SVM 具有比其他非监督情感分类算法高的情感识别能力, 而这正好与 Pang 等人^[14]的实验结论相吻合, 尽管与 SVM 相比较, TSMMF 的情感分类能力存在着差距, 但考虑到有标签微博情感训练数据的高昂代价, 其差距还是可以接受的, 尤其是在表情符比例较高

的数据集中;(5)比较表 7 与其他表(表 4~表 6)可以发现,相对于 Data-old, TSMMF 在 Data-new 上的情感分类正确率有较大幅度的提高,与 SVM 的表现几乎相当,而其他 4 种算法在 Data-old 与 Data-new 上的情感分类正确率基本一致并没有出现较大起伏.结合表 3,对此刻做出如下解释:Data-new 中更多的表情符信息给 TSMMF 提供了展示其优势的机会.

表 4 数据集 First-Ten 上的情感分类正确率比较

组号	SVM	SSA-ST	JST	SLDA	DPLDA	TSMMF
1	72.05	70.34	58.52	56.29	58.17	70.62
2	71.84	71.57	59.54	52.99	61.38	71.73
3	72.89	71.28	58.41	58.92	55.42	69.92
4	76.83	74.41	54.27	59.38	57.87	74.33
5	72.01	70.15	53.77	55.70	55.54	67.67
6	73.93	71.33	58.14	62.11	54.37	72.49
7	70.85	68.52	59.34	58.93	56.35	68.30
8	71.56	72.09	62.43	54.53	57.37	72.90
Avg	72.74	71.21	58.05	57.37	57.06	70.75

表 5 数据集 Mid-Ten 上的情感分类正确率比较

组号	SVM	SSA-ST	JST	SLDA	DPLDA	TSMMF
1	71.23	70.07	60.51	58.94	58.55	66.90
2	75.40	74.72	59.30	61.83	56.47	65.31
3	75.40	74.29	62.70	59.13	58.30	72.37
4	76.84	71.16	64.61	60.26	58.19	63.41
5	75.92	69.95	61.67	64.24	56.31	66.45
6	76.66	70.23	63.80	60.04	58.87	68.73
7	71.78	68.98	54.01	67.32	60.44	66.05
8	70.42	66.66	59.81	64.69	56.85	65.21
Avg	74.21	70.75	60.08	62.06	57.99	66.81

表 6 数据集 Last-Ten 上的情感分类正确率比较

组号	SVM	SSA-ST	JST	SLDA	DPLDA	TSMMF
1	72.26	71.95	57.25	60.70	54.72	72.15
2	70.80	68.82	55.92	59.86	56.92	65.67
3	72.41	69.06	59.08	55.84	58.74	67.76
4	71.56	71.24	58.27	50.45	52.17	71.60
5	72.89	71.86	53.87	60.29	55.56	71.23
6	69.39	68.59	60.40	53.86	58.47	67.91
7	71.03	70.55	55.57	52.99	57.57	67.93
8	70.59	69.11	58.35	54.05	58.81	68.91
Avg	71.37	70.14	57.34	56.01	56.62	69.15

表 7 数据集 Data-new 上的情感分类正确率比较

组号	SVM	SSA-ST	JST	SLDA	DPLDA	TSMMF
1	72.83	73.31	57.83	58.76	52.58	74.39
2	74.65	72.58	59.54	59.83	56.07	72.86
3	72.02	71.46	59.26	61.02	58.17	71.33
4	76.19	75.51	62.31	59.12	61.22	75.45
5	72.77	72.89	59.92	56.39	54.21	73.21
6	77.96	74.22	63.37	61.96	51.84	74.14
7	71.83	70.65	59.48	54.85	59.72	72.23
8	74.55	72.93	62.65	59.13	57.03	70.55
Avg	74.10	72.94	60.55	58.89	56.36	73.02

4.3 主题提取

TSMMF 的一个重要目标就是从微博数据集中提取话题.与 LDA 不同,TSMMF 在进行话题提取时需要考虑话题的情感极性.因此,本节从积极与消极两个情感面对抽取的话题展开分析,实验结果见表 8.由该表可以看出,来自 3 个数据集中隐含话题紧凑且具有较好解释性.例如,在积极情感话题方面,从隐含在数据集 First-Ten 中的主题词“中国,一个,问题,两,国家”可以看出,微博用户在讨论“一国两制”的话题,由隐含在数据集 Mid-Ten 中的主题词“幸福,继续,朋友,准备,漂亮”可推知微博用户正在讨论诸如婚礼等庆典活动,而从数据集 Last-Ten 中的主题“人生,电影,警察”可知微博用户可能谈论成龙主演的电影《警察故事》,这些微博用户在对待“一国两制”制度,庆典活动与《警察故事》都表现出支持、欢喜与喜爱等积极情感态度.类似地,在消极情感话题方面,数据集 First-Ten 中的主题词“北京,上海,工作”反映北京、上海等一线城市工作压力带来的不满等消极情绪,在数据集 Mid-Ten 中的主题词“分享,图片,骗子”中微博用户表现出对网络欺诈等行为的痛恨,而数据集 Last-Ten 中的主题词“饿,腐败,报道”体现网络民众对政府官员腐败的深恶痛绝.从上面分析可知,与传统话题提取模型 LDA 相比较,TSMMF 可以在微博消息集话题提取的同时进行话题情感倾向分析.

表 8 TSMMF 提取的主题示例

数据集	积极情感	消极情感
First-Ten	人,事,请,懂,钱,世界,年,版,需要,中国,一个,时,手机,问题,两,演唱会,全,月,国家	新,北京,哥,活动,工作,上海,互联网,国,人员,一直,小,线,称,报,投资,提供,间,话,刚刚,鞋
Mid-Ten	里,后,请,走,技术,岁,男人,却,挺,继续,幸福,酷,水,朋友,准备,漂亮,一下,错,元,句	分享,转,图片,说,话,只,博文,加油,送,地球,够,加入,骗子,顺便,西班牙,老婆,篇,一个,儿子,掉
Last-Ten	支持,最,女人,男人,美国,书,问题,搞,买,滴,爆,感谢,送,错,哈哈,哈哈,电影,道,人生,警察,阿根廷	猫,强,公司,今晚,只,杯,买,换,难,饿,想,电锯,日,先,衣服,非,腐败,理由,报道,黄晓明

另外,表 8 中出现了一些单字词,如“先、只、里”等,这些单字词的出现一方面冲淡了主题的集中度,另一方面也损伤了微博情感极性的直观性.那么能否直接采用预处理策略将这些单字词过滤掉?深入分析表 8 可以发现,表 8 中的单字词大致可分为 3 类:其一是情感倾向性与主题指示性都很弱的单字词,如“先、够、里、称”;其二是情感倾向性很弱但具有一定主题指示性的单字词,如“钱、书、报、买”;

其三是主题指示性很弱但具有较强情感倾向性的单字词类是“饿、挺、难”等。由此可见,直接过滤策略难以奏效的;由于能够有效帮助微博主题情感分析的第2类与第3类单字词会被误删。为了探索研究思路,我们利用简单的微博数据集全文检索方法对部分单字词进行统计分析,发现很多单字词都有其自身的出现模式:如单字词“挺”出现在“挺+人称代词”(如“挺你”)和“挺住”中的频率高达80%,单字词“懂”出现“你懂的”、“你们懂的”与“大家懂的”之类的“人称名称+懂的”模式中的概率也达到82%。从这些现象可以看出,第2类与第3类单字词的产生很大可能是由于产生于传统长文本的语料库的现有通用中文分词算法难以适应包含大量网络流行语的形式不规范的微博用语,因此,我们将在未来的工作中对微博分词展开研究,吸收采纳更好的微博分词算法来对微博消息进行预处理。当然,除了分词算法的研究外,还可研究单字词同义扩展策略以增加情感主题词的直观性,如将“挺你”中的“挺”同义扩展为“支持”,这也是一条值得尝试的思路。

为了更好地分析 TSMFM 的主题提取性能,我们进一步引入 KL 散度对 TSMFM 与 LDA 进行实验比较分析(图3)。

$$KL = \frac{1}{k} \sum_{P, Q \in DT} \sum_{w \in W} P(w) \log \frac{P(w)}{Q(w)} \quad (15)$$

其中: k 为指定主题数; DT 为算法从数据集中提取的主题集合; $P(w)$ 与 $Q(w)$ 指不同主题中单词 w 的出现概率。 KL 值越大表示所提取主题之间的相似度越低。

从图3中可以看出,与 LDA 相比较,TSMFM 提取的主题集合具有更高 KL 值,这说明由 TSMFM 从同一个数据集中提取的不同主题具有更大的区分度,从而能更好地为微博话题发现服务,当然,TSMFM 在不同数据集上表现出的这种优势存在着差异。这种优势可以从 TSMFM 的主题提取过程得到解释,即 TSMFM 的主题提取与微博情感识别是协同进行的。

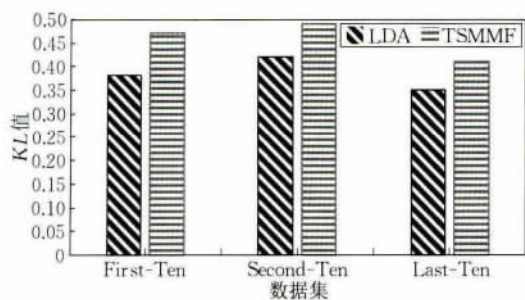
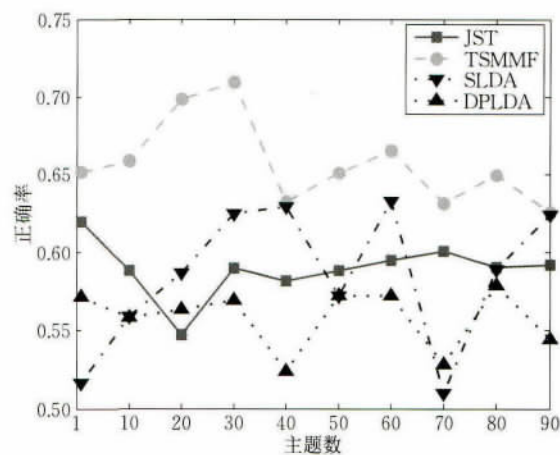


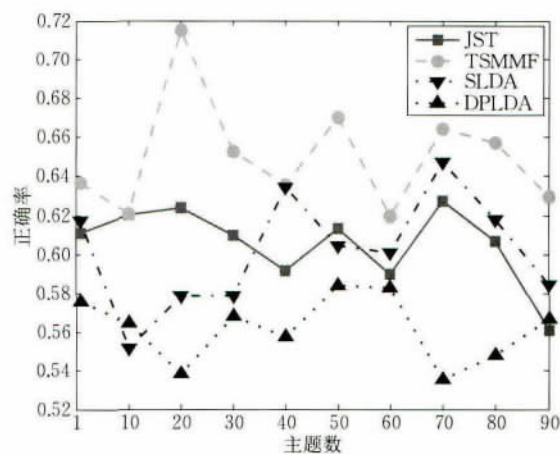
图3 提取主题的质量比较

4.4 主题数对情感分类正确率的影响

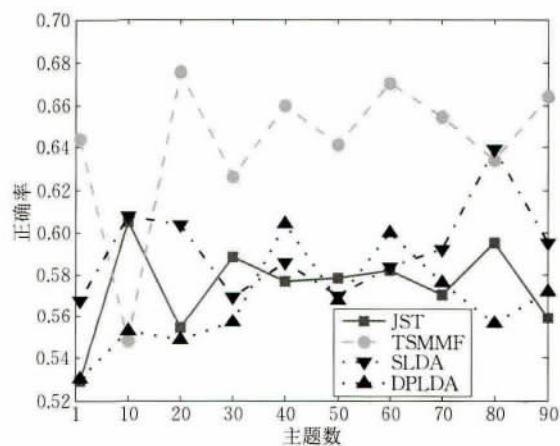
由于 TSMFM 可以同步进行微博消息集的主题检测与情感分析,因而有必要对情感分类与话题检测是如何相互作用进行探究。本节对主题数对情感分类准确率的影响展开实验分析,实验结果见图4。由图4可知,主题数在不同数据上对 TSMFM 的情感分类能力的影响不同。具体来说,TSMFM 在



(a) First-Ten



(b) Second-Ten



(c) Last-Ten

图4 主题数对情感分类正确率的影响

如下情形下获得最高的情感分类正确率:主题数为 30(First-Ten)、主题数为 20(Second-Ten)与主题数为 30(Last-Ten). 由此可见,不适当的主题粒度(主题数)会降低微博情感分类正确率. 这可以从两个方面进行解释:一方面过低的主题数会使得 TSMMF 弱化为完全忽略情感与主题相关性的 LDA,从而情感分类正确率下降;另一方面过高的主题数使得一些广泛关注的完整主题碎片化从而出现噪声主题,这也会降低模型的情感分析能力.

与无监督情感分析模型(JST, SLDA 和 DPLDA)相比较,除了在少数情形下(主题数为 10 和 80 时的 Last-Ten),无论在单一主题还是在多主题数下, TSMMF 都体现出不同程度的优势. 从这些观测可知,就微博主题情感同步建模性能而言, TSMMF 要远优于其他典型无监督主题情感分析模型.

4.5 表情符号比例对情感分类正确率的影响

在 4.2 节的实验结果中,我们观察到表情符号的利用会有效提升 TSMMF 的主题情感建模能力,那么其二者之间的量化关联关系是怎样的呢?为了测度表情符号的利用程度,我们提出表情符号比例指标(在一个实验微博消息集中,含有表情符号的消息所占比例),实验结果见图 5. 观察图 5 可知,在所有的实验数据集中,微博情感分类正确率在表情符号比例为 0 时最低,而在 100% 时最高. 从情感分类正确率的变化趋势来看,尽管在不同数据集中,情感分类正确率随表情符号比例增加而提高的程度不同,但从整体上看,而 TSMMF 的情感分类正确率与数据集表情符号比例是呈线性正相关的.

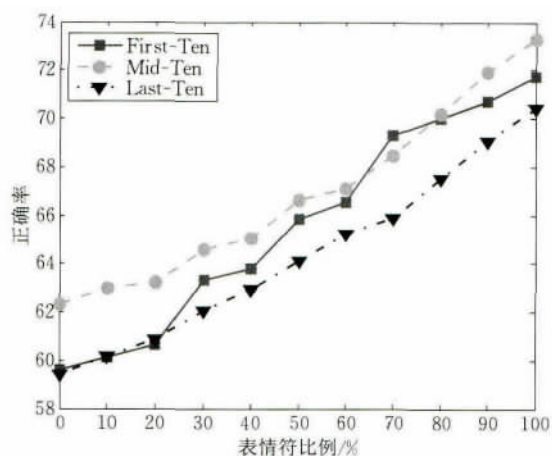


图 5 表情符比例对情感分类正确率的影响

4.6 性格情绪特征对情感分类准确率的影响

TSMMF 模型的一个重要特点就是在微博消息情感推理过程中加入了博主性格情绪特征因素.

因此,在本小节中,我们尝试分析博主性格情绪特征因素对情感分类准确率的影响.

考虑到性格情绪特征参数 λ 是一个与微博情感分布在 TSMMF 学习过程中协同更新的动态变量,而情感分类准确率是情感分析结果的度量指标,这二者并直接构成一一对应关系,因而,给出微博用户情绪特征参数值与微博情感分类准确率的依赖关系是非常困难的,当然也可以研究 λ 矩阵的演化特性与微博情感分类准确率之间的关系,以寻求更新 λ 矩阵的更优方法. 我打算将这一部分研究做为未来的兴趣点.

在此,我们没有直接度量微博用户性格参数值与微博情感分类准确率的依赖关系,而是通过定义性格情绪特征使用率 PUR 来量化分析析博主性格情绪特征因素对情感分类准确率的影响.

$$PUR = UP / U \quad (16)$$

其中: UP 是指在情感分析过程需要进行性格情绪特征建模的用户数; U 为微博消息集的博主总计数.

实验结果见图 6. 从图 6 可以看出,尽管在不同数据集中,性格情绪特征的利用度对情感分类准确率的提升存在着差异, TSMMF 的情感分类正确率在总体上是与 PUR 呈线性正相关的,即随着 PUR 的增加, TSMMF 的情感分类正确率会得到有效提升,尤其是对比 $PUR = 0$ 与 $PUR = 100\%$ 两种极端情形,可以发现,性格情绪特征信息的利用使情感分类准确率提高 3~4 个百分点. 由此可见,引入博主性格情绪特征因素对提升模型的微博情感分析能力有着积极意义.

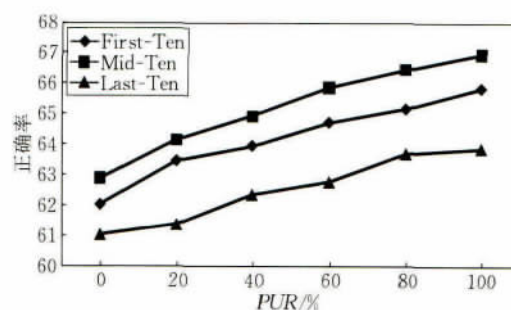


图 6 用户性格值对情感分类正确率的影响

4.7 用户活跃度对情感分类准确率的影响

在前面的实验分析中,我们采用的微博数据集是源自活跃度相对较低的微博用户. 用户活跃度在通信网络、社交网络等不同形式的网络中都得到不同程度的研究,李全刚等人^[59]根据模型的复杂度和准确度将全体用户划分成 K 个活跃度等级. 刘晶等人^[60]等建立微博用户活跃度模型实现用户的活跃

与非活跃的二分类. 本文中的用户活跃度是从用户单位时间参与微博活动行为的频率, 这与从微博粉丝数角度定义的用户活跃度是有区别的, 那么对于高活跃度用户的微博数据, TSM MF 的性能表现又如何呢? 为此, 我们根据用户发微博的日频率 f 将 3 个数据集 (First-Ten、Mid-Ten 与 Last-Ten) 中的用户分为 3 类: 日频率 ($f \leq 3$) 的用户为低活跃度用户, 日频率 ($4 < f \leq 6$) 的用户为中等活跃度用户, 其他为高活跃度用户. 然后由此形成 9 个微博数据集 (具有不同等级活跃度的用户所发微博). 表 9 是 TSM MF 在 9 个不同数据集的情感分类实验结果. 从整体上观察表 9 的“分类准确率”列, 可以得知, 在绝大多数情况下, 对于同一时间周期内的不同微博数据集, TSM MF 在用户活跃度越高的数据集上所获得情感分类准确率越低, 如在数据集 First-Ten 上, 随着用户活跃度沿着高-中-低逐步减弱, TSM MF 的情感分类准确率由 70.75 到 72.34 再到 74.57 逐步提高; 在数据集 Last-Ten 上, 随着用户活跃度沿着高-中-低逐步减弱, TSM MF 的情感分类准确率由 69.15 到 73.39 再到 74.12 逐步提高, 当然, 也有少数例外: TSM MF 在中等活跃度数据集 Mid-Ten-1 的情感分类正确率要比高活跃度数据集 Mid-Ten-2 更优. 为什么会出现这种情况? 通过进一步分析微博数据, 我们发现: 低活跃度用户参与讨论的事件往往是社会影响力大的事件, 其可能是天怒人怨的官员腐败事件, 也可能是深得人心的社会正能量事件, 而且与中高等活跃度用户相比较, 这些低活跃度用户的微博用语更加情绪化, 这意味着低活跃度用户的微博具有更强的情感倾向性, 而这恰好与 TSM MF 在低活跃度用户微博数据集上的情感分类正确率相对更高是相吻合的, 由于在情感倾向性更强的低活跃度用户微博数据集中微博极性 (消极与积极) 分界更加清晰.

表 9 用户活跃度对情感分类正确率的影响

数据集	用户活跃度	用户数	PN	分类正确率
First-Ten-1	高	121	0.52	70.75
First-Ten-2	中	457	0.46	72.34
First-Ten-3	低	832	0.39	74.57
Mid-Ten-1	高	98	0.57	66.81
Mid-Ten-2	中	494	0.52	65.28
Mid-Ten-3	低	905	0.48	69.06
Last-Ten-1	高	128	0.53	69.15
Last-Ten-2	中	513	0.49	73.39
Last-Ten-3	低	987	0.41	74.12

另外, 我们还对 9 个不同数据集的用户数与微博情感极性比 ($PN = \text{消极情感微博数} / \text{积极情感微博数}$) 进行描述性统计, 由统计量“用户数”可以看出, 在 3 个不同采集周期的数据中, 随着用户活跃度的降低, 用户数目不断增加, 这一现象是与微博复杂系统中的节点度幂律分布特性相吻合的. 还可由统计量 PN 看出, 随着活跃度的增加, 负面情绪出现的概率明显增大, 相应地, 正面情绪出现的概率明显减小.

5 结束语

随着微博服务的普及, 人们可以在微博平台上针对现实各种事件进行所见分享与所感交流, 挖掘隐藏在海量微博消息中的主题情感能有效辅助用户个体、企业组织与政府机构等的决策. 针对传统无监督的主题情感分析模型的不足, 本文提出了一个新的基于多特征融合的主题情感分析模型 TSM MF. 实验表明, TSM MF 不仅能实现微博消息的主题情感同步检测与分析, 而且具有比现有典型无监督主题情感模型 (JST, SLDA 和 DPLDA) 更优的情感分类能力.

将来的工作将在以下几个方面进行. 首先我们将微博用户的档案属性信息纳入微博消息情感极性与主题推理以提升情感分类的准确率. 其次, 微博消息具有实时特征, 我们将对微博消息的情感主题动态演化模式进行分析. 另外, 利用统计分析语料解释本文假设的合理性并结合微博的中性或客观的内容分析进一步提高情感识别正确率也是一个重要的后续工作.

参 考 文 献

- [1] Jiang H, Lin P, Qiang M. Public-opinion sentiment analysis for large hydro projects. *Journal of Construction Engineering and Management*, 2015, 142(2): 05015013
- [2] Zhang Y. Incorporating phrase-level sentiment analysis on textual reviews for personalized recommendation//*Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM'15)*. Shanghai, China, 2015: 435-440
- [3] Xie H, Li X, Wang T, et al. Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy. *Information Processing & Management*, 2016, 52(1): 61-72

- [4] Lin C, He Y, Everson R, et al. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering*, 2012, 24(6): 1134-1145
- [5] Li F, Huang M, Zhu X. Sentiment analysis with global topics and local dependency//*Proceedings of the 24th AAAI Conference on Artificial Intelligence*. Atlanta, USA, 2010: 1371-1376
- [6] Moghaddam S, Ester M. On the design of LDA models for aspect-based opinion mining//*Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. Maui, USA, 2012: 803-812
- [7] Abelson R. Whatever became of consistency theory? *Personality and Social Psychology Bulletin*, 1983, 9(1): 37-64
- [8] Guo Hui-Qing. The Research of Citizen Positive Emotions Experience in Micro-Blogging Public Opinion Field [Ph. D. dissertation]. Northeast Normal University, Changchun, 2015(in Chinese)
(郭慧清. 微博舆论场中的公民积极情绪体验研究[博士学位论文]. 东北师范大学, 长春, 2015)
- [9] Peterson C. *A Primer in Positive Psychology*. London, UK: Oxford University Press, 2006
- [10] Tang H, Tan S, Cheng X. A survey on sentiment detection of reviews. *Expert Systems with Applications*, 2009, 36(7): 10760-10773
- [11] Liu B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 2012, 5(1): 1-167
- [12] Tsytsarau M, Palpanas T. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 2012, 24(3): 478-514
- [13] Li Y, Gao H, Yang M, et al. What are Chinese talking about in hot weibos? *Physica A Statistical Mechanics & Its Applications*, 2013, 419: 546-557
- [14] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques//*Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*. Philadelphia, USA, 2002: 79-86
- [15] Mullen T, Collier N. Sentiment analysis using support vector machines with diverse information sources//*Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain, 2004: 412-418
- [16] Ng V, Dasgupta S, Arifin S M. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews//*Proceedings of the COLING/ACL on Main Conference Poster Sessions*. Sydney, Australia, 2006: 611-618
- [17] Kennedy A, Diana I. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 2006, 22(2): 110-125
- [18] Xia R, Zong C. Exploring the use of word relation features for sentiment classification//*Proceedings of the 23rd International Conference on Computational Linguistics*. Uppsala, Sweden, 2010: 1336-1344
- [19] Li S, Zhang H, Xu W, et al. Exploiting combined multi-level model for document sentiment analysis//*Proceedings of the 20th International Conference on Pattern Recognition*. Istanbul, Turkey, 2010: 4141-4144
- [20] Bespalov D, Bai B, Qi Y, et al. Sentiment classification based on supervised latent n -gram analysis//*Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. Glasgow, UK, 2011: 375-382
- [21] Tang D, Wei F, Yang N, et al. Learning sentiment-specific word embedding for Twitter sentiment classification//*Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*. Baltimore, USA, 2014: 1555-1565
- [22] Dos Santos C N, Gatti M. Deep convolutional neural networks for sentiment analysis of short texts//*Proceedings of the 25th International Conference on Computational Linguistics (COLING'14)*. Dublin, Ireland, 2014: 69-78
- [23] Liu Long-Fei, Yang Liang, Zhang Shao-Wu, et al. Convolutional neural networks for Chinese micro-blog emotional tendency identification. *Journal of Chinese Information Processing*, 2015, 29(6): 159-165(in Chinese)
(刘龙飞, 杨亮, 张绍武等. 基于卷积神经网络的微博情感倾向性分析. *中文信息学报*, 2015, 29(6): 159-165)
- [24] Chen Zhao, Xu Rui-Feng, Gui Lin, et al. Combining convolutional neural networks and word sentiment sequence features for Chinese text sentiment analysis. *Journal of Chinese Information Processing*, 2015, 29(6): 172-178(in Chinese)
(陈钊, 徐睿峰, 桂林等. 结合卷积神经网络和词语情感序列特征的中文情感分析. *中文信息学报*, 2015, 29(6): 172-178)
- [25] Sun X, Li C, Xu W, et al. Chinese microblog sentiment classification based on deep belief nets with extended multi-modality features//*Proceedings of the IEEE International Conference on Data Mining (ICDM'14)*. Shenzhen, China, 2014: 928-935
- [26] Zhou S, Chen Q, Wang X. Active semi-supervised learning method with hybrid deep belief networks. *Plos One*, 2014, 9(9): e107122-e107122
- [27] Mei Q Z, Ling X, Wondra M, et al. Topic sentiment mixture: Modeling facets and opinions in weblogs//*Proceedings of the 16th International Conference on World Wide Web*. Alberta, Canada, 2007: 171-180
- [28] Titov I, McDonald R. A joint model of text and aspect ratings for sentiment summarization//*Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technology*. Columbus, USA, 2008: 308-316
- [29] Titov I, McDonald R. Modeling online reviews with multi-grain topic models//*Proceedings of the 17th International Conference on World Wide Web*. Beijing, China, 2008: 111-120

- [30] Dasgupta S, Ng V. Topic-wise, sentimentwise, or otherwise? Identifying the hidden dimension for unsupervised text classification//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore, 2009: 580-589
- [31] Hu M, Liu B. Mining and summarizing customer reviews//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, USA, 2004: 168-177
- [32] Andreevskaia A, Bergler S. Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses//Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy, 2006: 209-216
- [33] Ouyang Ji-Hong, Liu Yan-Hui, Li Xi-Ming, et al. Multi-grain sentiment/topic model based on LDA. Acta Electronica Sinica, 2015, 43(9): 1875-1880(in Chinese)
(欧阳继红, 刘燕辉, 李熙铭等. 基于 LDA 的多粒度主题情感混合模型. 电子学报, 2015, 43(9): 1875-1880)
- [34] Rao Y, Li Q, Mao X, et al. Sentiment topic models for social emotion mining. Information Sciences, 2014, 266(5): 90-100
- [35] Li F, Wang S, Liu S, et al. SUIT: A supervised user-item based topic model for sentiment analysis//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec, Canada, 2014: 1636-1642
- [36] Yang Z, Kotov A, Mohan A, et al. Parametric and non-parametric user-aware sentiment topic models//Proceedings of the 38th International ACM Conference on Research and Development in Information Retrieval (SIGIR'15). Santiago, Chile, 2015: 413-422
- [37] Liu S, Cheng X, Li F, et al. TASC: Topic-adaptive sentiment classification on dynamic tweets. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(6): 1696-1709
- [38] Tan S, Li Y, Sun H, et al. Interpreting the public sentiment variations on Twitter. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(5): 1158-1170
- [39] Huang Fa-Liang, Li Chao-Xiong, Yuan Chang-An, et al. Mining sentiment for web short text based TSCM model. Acta Electronica Sinica, 2016, 44(8): 1887-1891(in Chinese)
(黄发良, 李超雄, 元昌安等. 基于 TSCM 模型的网络短文本情感挖掘. 电子学报, 2016, 44(8): 1887-1891)
- [40] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009: 1-12
- [41] Davidov D, Tsur O, Rappoport A. Enhanced sentiment learning using Twitter hashtags and smileys//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10). Uppsala, Sweden, 2010: 241-249
- [42] Pak A, Paroubek P. Twitter as a Corpus for sentiment analysis and opinion mining//Proceedings of the International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta, 2010: 1320-1326
- [43] Kouloumpis E, Wilson T, Moore J. Twitter sentiment analysis: The good the bad and the OMG !//Proceedings of 5th International AAAI Conference on Weblogs and Social Media. Barcelona, Spain, 2011: 538-541
- [44] Liu K L, Li W J, Guo M. Emoticon smoothed language models for Twitter sentiment analysis//Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12). Toronto, Canada, 2012: 1678-1684
- [45] Jiang F, Liu Y Q, Luan H B, et al. Microblog sentiment analysis with emoticon space model. Journal of Computer Science and Technology, 2015, 30(5): 1120-1129
- [46] Zhao J, Dong L, Wu J, et al. MoodLens: an emoticon-based sentiment analysis system for chinese tweets//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 1528-1531
- [47] Zhang L, Pei S, Deng L, et al. Microblog sentiment analysis based on emoticon networks model//Proceedings of the 5th International Conference on Internet Multimedia Computing and Service. Huangshan, China, 2013: 134-138
- [48] Churches O, Nicholls M, Thiessen M, et al. Emoticons in mind: An event-related potential study. Social Neuroscience, 2014, 9(2): 196-202
- [49] Kshirsagar S, Magnenat-Thalmann N. A multilayer personality model//Proceedings of the International Symposium on Smart Graphics. NY, USA, 2002: 107-115
- [50] Yang G, Wang Z, Wang G, et al. Affective computing model based on emotional psychology//Jiao Li-Cheng, Wang Li-Po, Gao Xin-Bo, et al, eds. Advances in Natural Computation. Springer Berlin Heidelberg, 2006: 251-260
- [51] Li Hai-Fang, He Hai-Ping, Chen Jun-Jie. A multi-layer affective model based on personality, mood and emotion. Journal of Computer-Aided Design & Computer Graphics, 2011, 23(4): 725-730(in Chinese)
(李海芳, 何海鹏, 陈俊杰. 性格、心情和情感的多层情感建模方法. 计算机辅助设计与图形学学报, 2011, 23(4): 725-730)
- [52] Seidman G. Self-presentation and belonging on Facebook: How personality influences social media use and motivations. Personality & Individual Differences, 2013, 54(3): 402-407
- [53] Stoughton J W, Thompson L F, Meade A W. Big five personality traits reflected in job applicants social media postings. Cyberpsychology Behavior & Social Networking, 2013, 16(11): 800-805
- [54] Li C, Cheung W K, Ye Y, et al. The author-topic-community model for author interest profiling and community discovery. Knowledge & Information Systems, 2014, 44(2): 359-383
- [55] Minka T P. Expectation propagation for approximate Bayesian inference//Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence. Seattle, USA, 2001: 362-369
- [56] He L, Jia Y, Han W, et al. Mining user interest in microblogs with a user-topic model. China Communications, 2014, 11(8): 131-144

- [57] Hu X, Tang L, Tang J, et al. Exploiting social relations for sentiment analysis in microblogging//Proceedings of the 6th International Conference on Web Search and Data Mining. Rome, Italy, 2013: 537-546
- [58] Lu T. Semi-supervised microblog sentiment analysis using social relation and text similarity//Proceedings of the 2nd International Conference on Big Data and Smart Computing (BigComp'15). Jeju, South Korea, 2015: 194-201
- [59] Li Quan-Gang, Liu Qiao, Qin Zhi-Guang. Modeling and simulation of communication based on topic model. Journal of Computer Research and Development, 2016, 53(1): 206-215 (in Chinese)
(李全刚, 刘桥, 秦志光. 基于主题模型的通信网络建模与仿真. 计算机研究与发展, 2016, 53(1): 206-215)
- [60] Liu Jing, Wang Feng, Hu Ya-Hui, et al. Detecting inactive users from behavior data based on Weibo. Journal of University of Electronic Science and Technology of China, 2015, 44(3): 410-414(in Chinese)
(刘晶, 王峰, 胡亚慧等. 基于微博行为数据的不活跃用户探测. 电子科技大学学报, 2015, 44(3): 410-414)



HUANG Fa-Liang, born in 1975, Ph.D., associate professor, M.S. supervisor. His research interests include data mining and social media processing.

FENG Shi, born in 1981, Ph.D. His research interests include sentiment analysis and data mining.

WANG Da-Ling, born in 1962, Ph.D., professor, Ph.D. supervisor. Her research interests include data mining and social media processing.

YU Ge, born in 1962, Ph.D., professor, Ph.D. supervisor. His research interests include data science and advanced database.

Background

With the fast development of Web 2.0 technology, microblogging, like Twitter, has become a popular platform of human expressions. The huge amount of microblogging data is a useful and timely source that carries massive information about sentiment and opinions on various topics. How to automatically detect sentiment polarities and discover hidden knowledge in microblogging data has become the major concern for both the academic and commercial communities. Different from traditional texts, microblogging data are dynamic and contain multifarious features such as emoticons, update time etc. Most existing sentiment and topic detection approaches treat the unique microblogging data as noise. However, this may lead to unsatisfactoriness in sentiment classification and topic identification.

In this paper we study the problems which belong to the field of social media processing and it is also related to

data mining. We proposed a probabilistic model, TSMMF, to discover topic and sentiment distribution hidden in microblogging based on multi-feature fusion is proposed to overcome the abovementioned drawbacks. This model extends widely-used topic model LDA to 4-layer joint topic sentiment model by adding a sentiment layer between topic layer and word layer. Secondly, TSMMF introduces emoticons and microbloggers personality into LDA inference framework and achieves synchronized detection of sentiment and topic in microblogging. Finally, the experimental results show that TSMMF outperforms state-of-the-art unsupervised approaches JST, SLDA and DPLDA significantly in terms of sentiment classification accuracy, and it has promising performance.

This work is supported by the National Natural Science Foundation of China(61370074).