软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(12):3183−3205 [doi: 10.13328/j.cnki.jos.005283]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

中文微博情感分析研究与实现*

李勇敢[1], 周学广[2], 孙 艳[3], 张焕国[1]

[1](武汉大学 计算机学院,湖北 武汉 430079)

[2](海军工程大学 信息安全系,湖北 武汉 430033)

[3](中国人民解放军 92941 部队,辽宁 葫芦岛 125000)

通讯作者: 周学广, E-mail:zxg196610@hotmail.com

摘 要: 中文微博的大数据、指数传播和跨媒体等特性,决定了依托人工方式监控和处理中文微博是不现实的,迫切需要依托计算机开展中文微博情感自动分析研究.该项研究可分为 3 个任务:中文微博观点句识别、情感倾向性分类和情感要素抽取.为完成上述任务,研制了一个评测系统:通过构建多级词库、制定成词规则、开展串频统计等给出一种基于规则和统计的新词识别方法,在情感词和评价对象的依存模式的基础上给出基于词语特征的观点句识别算法;以词序流表示文本的 LDA-Collocation 模型,采用吉布斯抽样法推导了算法,实现中文微博情感倾向性自动分类;针对中文微博情感要素抽取召回率较低的问题,利用依存关系分析理论,按主语类和宾语类把依存模式分为两类,建立了 6 个优先级的评价对象和情感词汇的依存模式,通过评价对象归并算法实现计算机自动抽取情感要素.实验包括两个部分:一是参加 NLP&CC2012 的公开评测,所提方法在微博观点句识别任务中的准确率为第 2,在中文微博情感要素抽取任务中的准确率和 F 值均为第 2,验证了该算法的实用性;二是在分析公开评测结果的基础上,分别比较了参加公开评测的各类算法在处理中文微博情感分析时的效率,给出了相关结论.

关键词: 中文微博;情感分析;依存分析;情感倾向性分类;情感要素抽取;无监督主题情感模型

中图法分类号: TP391

中文引用格式: 李勇敢,周学广,孙艳,张焕国.中文微博情感分析研究与实现.软件学报,2017,28(12):3183−3205. http://www.jos.org.cn/1000-9825/5283.htm

英文引用格式: Li YG, Zhou XG, Sun Y, Zhang HG. Research and implementation of chinese microblog sentiment classification. Ruan Jian Xue Bao/Journal of Software, 2017,28(12):3183−3205 (in Chinese). http://www.jos.org.cn/1000-9825/5283.htm

Research and Implementation of Chinese Microblog Sentiment Classification

LI Yong-Gan[1], ZHOU Xue-Guang[2], SUN Yan[3], ZHANG Huan-Guo[1]

[1](School of Computer Science, Wuhan University, Wuhan 430079, China)

[2](Department of Information Security, Navy University of Engineering, Wuhan 430033, China)

[3](Unit Number of 92941, PLA, Huludao 125000, China)

Abstract: This paper studies sentiment analysis in Weibo. The study focuses on three types of tasks: emotion sentence identification and classification, emotion tendency classification, and emotion expression extraction. An unsupervised topic sentiment model, UTSM, is proposed based on the LDA Collocation model to facilitate automatic hashtag labeling. A Gibbs sampling implementation is presented for deriving an algorithm that can be used to automatically categorize emotion tendency with computer. To address the issue of lower recall

* 基金项目: 国家重点基础研究发展计划(973)(2014CB340600); 国家自然科学基金(61332019, 61672531); 国家社会科学基金(14GJ003-152)

Foundation item: National Program on Key Basic Research Project (973) (2014CB340600); National Natural Science Foundation of China (61332019, 61672531); National Social Science Foundation of China (14GJ003-152)

收稿时间: 2016-05-19; 修改时间: 2016-07-04, 2017-01-24; 采用时间: 2017-03-21; jos 在线出版时间: 2017-07-12

CNKI 网络优先出版: 2017-07-12 15:33:26, http://kns.cnki.net/kcms/detail/11.2560.TP.20170712.1533.004.html

ratio for emotion expression extraction in Weibo, dependency parsing is used to divide dependency model into two categories with subject and object. Six dependency models are also constructed from evaluation objects and emotion words, and a merging algorithm is proposed to accurately extract emotion expression. Result of experiments indicates that the presented method has a strong innovative and practical value.

Key words: Chinese Microblog; sentiment analysis; dependency parsing; emotion tendency classification; emotion expression extraction; unsupervised topic sentiment model

截至 2016 年 12 月,我国网民已达 7.31 亿,手机网民达 6.95 亿,互联网普及率达到 53.2%,微博实际用户数超过 5 亿(http://www.cnnic.net.cn).网民的快速增加和微博的迅速发展,使得大量评论信息迅速传播[1].对这些微博评论信息进行情感分析,是正确引导用户实现个性化推荐[2,3]、开展具有较高商业价值的微博营销[4,5]、实施当前急需的网络安全监管[6,7]的基础和前提.依靠人工的方法应对微博海量信息的收集和处理难以胜任,因此,微博情感自动分析应需而生.

目前,微博情感自动分析主要的研究途径有基于机器学习的情感分类方法和基于语义词典的逻辑分类方法.前者用泛化作为核心问题,分类的泛化能力与两类样本之间的间隔(margin)有关,即,求解 margin 问题;在用机械的分词法割裂了语言之间的逻辑关联后,需要大规模的人工标注工作,且依托计算机求解后得到的分类结果与数据集标注敏感关联.后者能够利用人类知识减少机器学习的盲目性,最主要的优点是可以减少人工标注样本工作.但如何自动或半自动地构建极性词典,尤其是建立起可适应所有微博类型的完备的极性词词典,是一项值得持续研究的工作.

针对中文微博的情感自动分析主要存在以下特点和困难.

(1) 中文微博在观点句的使用、表达观点的语言以及评价对象的隐现等方面具有口语化、负面多、句子短、情感强、语言不规范以及评价对象在句子中不一定直接出现等特点,造成运用传统的文本情感分析技术进行中文微博情感分析时,很难获得理想的处理结果(http://tcci.ccf.org.cn/conference/2012/dldoc/NLP&CC2012papers/workshoppapers/sen/003.pdf).

(2) 在抽取中文微博情感要素时,只抽取到情感词是不够的.因为有些情感词虽然具有明显的情感,但是其情感倾向却由评价对象而定,如“大、小、轻、重”.比如在句子“房间很大、很宽敞”中,“大”的情感倾向为褒义;但在“房间外面噪音很大”中,“大”的情感倾向为贬义.因此,独立地使用传统情感分析规律抽取中文微博中的情感词汇,结果不一定理想.又比如,句子“他比猪还猪”中,第 1 个词“猪”属于主题词汇,第 2 个词“猪”则转义为情感词汇,这样的词性转换同样无法通过传统的情感要素抽取方法提取出来[6].

(3) 词汇标注效果需要人工干预,情感分析词典与待分析的语料是否相符或相近,将直接影响情感分析效果,如何构建高效的与语料相近或相关的情感分析词典,是中文微博情感自动化处理的重要研究内容之一.

上述特点和存在的问题,是本文开展研究的出发点.

1 相关工作

1.1 机器学习理论与微博情感分析

机器学习理论在处理文本信息方向不断演进.1995 年,Vapnik 建立了支持向量机(support vector machines,简称 SVM)理论[8],用 VSM 模型建立了文档→词映射表示.为解决 VSM 模型中无法识别同义词、多义词的问题,1998 年提出的 LSA 模型在 VSM 模型的基础上引入了一个语义维度,即文档→语义→词,实现文档在语义空间上的低维表示[9].LSA 模型并不是一个生成式主题模型,它的提出为主题模型奠定了基础.由于 LSA 模型没有很好地重建原始的 TF-IDF 矩阵,且利用基于线性代数的奇异值分解算法复杂度高,难以并行化.2001 年,Hofmann 从生成模型的角度提出 pLSA 主题模型[10].pLSA 模型寻找的是词→文档共现矩阵所表示的生成模型,而不是寻找文档本身的生成模型,这使得 pLSA 模型存在两个缺点:一是缺乏处理语料集之外的新文档;二是待估计参数随着文档数的增长而线性增长,容易出现过拟合情况.2003 年,Blei 等人在 pLSA 模型的基础上,加上超参数层来

建立潜在变量z的概率分布,提出了LDA模型.该模型以词袋表示文本,在文本与词之间加入主题,它是一个完全的产生式模型,主题在文档集中是隐藏变量,从给定的文档中计算词汇的产生概率[11].词袋法没有考虑词与词之间的顺序,表达的只是一个粗略的上下文语义信息,如“高”和“铁”表达不出“高铁”的含义.2007 年,Griffiths 等人在 LDA 模型的基础上增加了词序关系,提出了 LDA-Collocation 模型[12].该模型采用的是词序流(stream of words)表示方法,对“不喜欢”连在一起比分开的“不”“喜欢”两个词的语义更精确和丰富,因此较 LDA 模型表达语义信息更准确.

微博情感倾向性分类与主题分类有着显著区别,需要更丰富的语义信息,因此,有学者构建主题情感混合模型用于文本情感分析中.主题情感混合模型在语言模型上有两种表示方法:一种是将主题和情感描绘成单一的语言模型,在模型中,一个词可能同时与主题和情感都相关,如文献[13]提出的ASUM模型和文献[14]提出的JST (joint sentiment/topic model)模型;另一种是将情感与主题作为分开的语言模型,一个词要么是情感词要么是主题词,只能二选一,如文献[15]提出的TSM(Topic sentiment mixture)模型,TSM 模型将词分为主题词和情感词,认为情感词对主题发现没有作用,而事实上,情感词是表示主题的重要词汇,应该是主题词的一部分.

1.2 依存分析与情感要素抽取

依存分析方法的核心是,用联结表示词与词之间的关系.1970 年代,Robinson 提出了依存文法中关于依存关系的 4 条公理.由于汉语语言结构的特殊性,黄昌宁等人提出第 5 条公理,即:中心词左右两边的词相互不发生依存关系[16].宾州中文树库是出现较早且被人们使用较多的中文结构语法库[17],通过ctbparser开源工具包,可以自动获取词语之间的依存关系(http://code.google.com/p.ctbparser/downloads/list).依存分析代表性应用研究包括:文献[18]开展了中文依存树库构建与分析研究,考察了目前达到一定规模的中文树库,包括宾州中文树库、哈尔滨工业大学中文依存树库[19]和清华短语结构树库等;文献[20]通过构建文本相似矩阵和依存相似矩阵得到更准确的微博文本相似矩阵,在此基础上运行聚类算法挖掘热点主题;文献[21]综合利用不同树库对应的基线分析器解析的依存骨架,提取交叉信息,在基本框架上构建了综合句法分析器;文献[22]利用依存句法分析的结果优化上下文的选择,给出了一种基于依存关系上下文的词表示方法;文献[23]利用依存和共指关系对语料进行分析,获得基于非线性全局上下文的词表示向量.

情感要素抽取是情感分析任务的重点.文献[24]利用深度学习理论,采用递归神经网络来发现与任务相关的特征,根据句子词语间前后关联性引入情感极性,避免了人工特征设计和手工标注情感特征.文献[25]以句子作为超边、以词为节点构建超图,在超图模型下利用句子与词之间的高阶信息生成摘要和关键词.该方法的优点在于统一考虑了摘要和关键词的关联性.文献[26]结合使用SVM和RNN构造两个分类器,然后对句子进行情感二元分类,同时还加入深度学习领域中词向量的统计信息.该方法参加COAE 2014会议评测任务1表现突出.文献[27]使用微博情感的多种文本特征,选用不同的特征组合对模型 SVM 和条件随机场 CRF(conditional random fields)[28]进行了比较研究,得出在不同的情况下选择合适模型的方法.

1.3 本文研究工作安排

为完成中国计算机学会中文信息技术专业委员会主办的公开评测中文微博情感分析任务(http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html),第 2 节综合考虑中文微博特点、运用中文信息柔性匹配技术进行微博预过滤;通过构建多级词库、制定成词规则、开展串频统计等给出一种基于规则和统计的新词识别方法,在扩展情感词的基础上给出基于词语特征的观点句识别算法.本文综合运用词序流表示文本的 LDA-Collocation 模型,在表达语义信息更准确的前提下加入情感模型,采用吉布斯抽样法实现算法的推导,提出无监督的主题情感混合算法,实现中文微博情感倾向性自动分类.具体工作参见第 3 节.本文旨在计算机海量微博处理必须避免传统情感分析算法需要手工标注情感特征这一环节,按主语类和宾语类把依存模式分为两类,建立 6 个级别的评价对象和情感词汇的依存模式,通过评价对象归并算法实现计算机自动抽取情感要素任务.具体工作见第 4 节.本文第 5 节开展实验验证研究与分析,主要有微博情感发现实验、微博公开评测实验和公开评测实验算法比较这3项内容.本文中用到的符号定义说明见表1,用到的依存关系类型定义见表2,词性标注采用

宾州中文树词性标注体系.中文微博情感分析流程如图 1 所示.

Table 1 Symbol description

表 1 符号定义说明

Parameter	Definition
θ_d	文档d的主题分布
φ_d	文档d的情感分布
$\phi_{z,m}$	主题情感~词汇分布
m_s	句子s的情感
$(z,m)_n$	词汇的主题和情感
w_n	句子中的词汇
α	文档-主题分布的Dir参数
χ	文档-情感分布的Dir参数
β	主题情感-词汇分布的Dir参数
M	文档数
S	文档的句子数
N	句子的词汇数
L	情感数
K	主题数

Table 2 Dependency type

表 2 依存关系类型

依存类型	说明	例句
SUB	主谓关系	他的汽车很漂亮(汽车,漂亮).
OBJ	动宾关系	我有一个苹果(苹果,有).
PRD	系表关系	又是车祸(车祸,是).
VC	连谓关系	他背起东西走出去了(背起,走).
AMOD	数量关系	我有一个苹果(一,个).
NMOD	定中关系	客车与大货车相撞(大,货车).
PMOD	介宾关系	他在美国(美国,在).
VMOD	状中关系	汽车开的很快(很,快).
DEP	支配“的、得、地”	我的苹果很好吃(我,的).
SBAR	支配助词	你有那么多钱啊(有,啊).
P	支配标点符号	汽车开的很快(开,.).

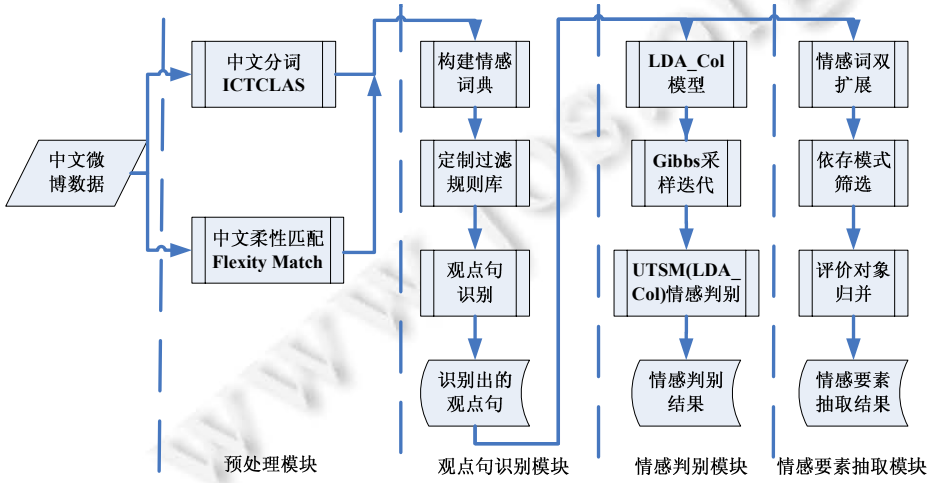


Fig.1 Flow chart of the chinese microblog sentiment classification

图 1 中文微博情感分析流程图

2 基于词语特征的观点句识别

综合考虑中文微博特点、运用中文信息柔性匹配技术进行微博预过滤;通过构建多级词库、制定成词规则、

开展串频统计等给出一种基于规则和统计的新词识别方法,在此基础上给出基于词语特征的观点句识别算法.

中文柔性匹配算法[29]采用基本的蛮力算法思想,匹配过程可以形象地看成用一个包含中文模式 p 的模板沿文本 t 滑动,同时,对文本 t 的每个字符位移注意模板上的字符是否与文本中相应顺序的字符相匹配.这里,相匹配的概念包括以下内容:如果文本关键词中夹杂了非汉字符号,直接跳过;如果文本中存在夹杂同音字、繁体字或拼音的方式进行中文主动干扰处理的关键字,采用拼音字典模式进行匹配;如果有采用英文夹杂干扰处理的关键字,采用英文字典进行匹配.最后统计模式匹配成功的次数,包括正常关键字个数和异常关键字个数.

2.1 观点句判定标准

观点句是主观性表达的一种,主观性表达包括评价(evaluation)和推测(speculation)两类,其中:评价包括个人情感、评论、判断与意见等,推测是指对非实际发生事件或非实际持有心理状态的表达.NLP&CC 2012 公开评测对观点句的判定标准为:观点句只限定于对特定事物或对象的评价,不包括内心自我情感、意愿或心情,并给出了以下例子.

(1) “我真心喜欢 iphone 的屏幕效果.”是对 iphone 的个人评价,并具有一定的感情色彩,是观点句;

(2) “我今天心情有点郁闷!”是对个人内心情感的表达,属于情感句,但由于缺乏特定的评价对象,在本次评测中被认为是非观点句;

(3) “调查表明,当前高校学生思想主流继续保持积极健康向上的良好态势.”是作者对客观现象进行的基于事实的描述,不带有个人好恶与观点,是非观点句.

在话题型微博中,评价对象一般是和话题相关的人和事.本文按照是否有特定评价对象来区分观点句和非观点句,那么所有的客观性表达均将被认为是非观点句,具有特定评价对象的主观表达被认为是观点句.

2.2 观点句识别思路

观点句识别是文本情感分析的基础,其任务是将文本中表达观点的句子识别出来.本节以微博为研究对象,从规则和统计两方面入手,对观点句识别问题开展研究.通过构建一个 3 分类情感词典,定制基于情感词典的观点句过滤规则;结合微博语言特点,采用基于规则的观点句识别算法对微博观点句分类特征进行提取,实现对观点句的识别.

2.3 五级情感词典构建

观点句识别和情感极性分析在一定程度上依赖于所构建的情感词典.词语的情感判别是句子情感分析的基础,词语按照情感极性通常可以分为褒义词、贬义词和中性词这 3 类.褒义词和贬义词通称为情感词,说话人的情感以及观点倾向主要是通过情感词来表达的,情感词在文本情感分析的研究中具有举足轻重的地位.

情感知识库结构词库包括情感词库、极性词库和否定词库.为了开展实验研究和参加公开评测,本文分别构建了情感词库、极性词库和否定词库.情感词库以知网公布的情感词表、《常用褒贬义词语详解词典》、《学生褒贬义词典》、《褒义词词典》和《贬义词词典》为基础,删除其中使用频率很低的情感词,增加网络用语和口语情感词,构建了一个情感词库,其中含褒义词 5 554 个,贬义词 6 321 个.将构建的情感词库按照词汇使用频率从最简版词库到完整版分为 5 级,分别记为 KW1~KW5,参见表 3.

Table 3 5-level emotion thesaurus description

表 3 5 级情感词库说明

	KW1	KW2	KW3	KW4	KW5
褒义词数	656	1 399	2 603	3 211	5 554
贬义词数	893	1 673	3 172	4 299	6 321
词库规模	1 549	3 072	5 775	7 510	1 1875

在情感词库中,我们还标出了一个极性词库.有部分情感词极性非常强烈,特别是一些骂人的贬义词,在观点句的识别时,只要出现这些词,就把观点句极性判定为该词的极性(否定句式中取反).为区别大的情感词库,将

这类词库称为极性词库,其中包含褒义词 16 个,贬义词 262 个.情感词库与极性词库是包含关系,即,情感词库包含极性词库中的所有词.此外,还专门构建了一个否定词库,包含“不、未、没有、欠”等否定词 20 个.

2.4 基于规则的观点句识别

对于观点句识别的研究,目前最常采用的是基于机器学习的方法,但由于人类语言有高度的灵活性和复杂性,因此研究人员依靠一些基于概率统计的机器学习方法,通过提取观点特征训练分类模型来识别观点句.这种基于概率统计的分类方法在理论上会造成一定的错判,为了减少这种错判,本文通过对大规模数据集的统计观察,定义了一些置信度较高的观点判别规则,分为非观点句判别规则和观点句判别规则这两类.

2.4.1 非观点句判别规则

微博中有些内容不属于观点句,在机器学习分类处理之前,采用规则将它们滤除,将有助于提高分类系统的处理效率和准确率.本文制定了以下规则对非观点句进行过滤.

Rule1:句子中非法词语、乱码超过 6 个的判定为非观点句;

Rule2:仅含有链接,但无实际文本信息的判定为非观点句;

Rule3:仅含有标签、标点符号的句子判定为非观点句;

Rule4:含有“数据显示”“调查表明”等客观标识词的判定为非观点句;

Rule5:含有特殊标点的句子判定为非观点句.特殊标点如“【”,“】”,这类微博通常是介绍性质的,“【”,“】”中的内容多为微博标题,这类微博中句子是观点句的概率极低,直接将其过滤.

2.4.2 观点句判别规则

汉语是一门博大精深的语言,观点句的表达方式也是多种多样的.本文通过对大量微博观点句的统计和分析,发现了几种经常在微博中使用的观点表达句式,并制定了相应的规则如下.

Rule1:包含网络热点词、情感短语的短句判为观点句.例如:#皮鞋果冻#太坑爹了吧!

Rule2:(代词|人名|地名|专有名词)+...+是+情感名词+....例如:#奖状植入广告#校长是个人才呀!

Rule3:(代词|人名|地名|专有名词)+...+程度副词+评价形容词+....例如:#洗碗工留剩菜被开除#中国人太奢侈了!

Rule4:是+...+的问题/责任.例如:#洗碗工留剩菜被开除#这位母亲很伟大,酒店不应该.

其中,规则 1 用于中文分词处理之前,目的是为了避免因中文分词不当造成的网络新词和情感短语无法识别的问题;规则 2 和规则 3 依赖于已建立的情感词典;规则 4 对话题型微博具有较强的针对性.此外,针对中文干扰微博,我们增加了柔性匹配模块,提前进行判别,将结果赋值为非观点句.

算法 1. 基于规则的观点句识别算法.

输入:句子编号集合 *SentenceID*[·];

输出:对句子编号集合的分类结果 *ValueSentenceID*[·]赋值 True/False.

步骤:

```
1)  FOR (i=0; i<32000; i++) {   //句子总数不大于公开测试集 32 000 条
2)      GET (Sentence[i]);   //获取当前句子 i
3)      IF (句子含有网络热点词或情感短语) THEN (ValueSentenceID[i]=True);   //Rule1
4)          ELSE {M=ICTCLAS(SentenceID[i]);
            //对句子 i 用 ICTCLAS 进行中文分词,M 为分词后的单词个数
5)              IF (Length(SentenceID[i])==M)   //如果分词个数与句子长度一致,
6)                  ValueSentenceID[i]=FlexibleMatch(SentenceID[i]);
                    进行中文柔性匹配,结果赋值给 ValueSentenceID[i]
7)              FOR (j=0; j<M; j++) {   //这里 M<141,因为 1 个微博句子的分词总数不超过 140 个词
8)                  IF ((word 为情感词)&&(含(是)和(程度副词))) ValueSentenceID[i]=True;   //Rule2+Rule3
9)                  IF ((word 是“问题”或“责任”)&&(含(的)和(是))) ValueSentenceID[i]=True;   //Rule4
```

10) } //end FOR (*j*)
11) } end ELSE
12) } end FOR (*i*)

3 基于主题模型的无监督主题情感分类

3.1 基于LDA-Collocation的无主题情感模型

本文选用表达语义信息更准确的 LDA-Collocation 模型,在其上添加情感模型,通过对每个句子进行情感标签采样,建立“文档-情感-句子”关系;沿用 LDA-Collocation 模型对每个词进行主题标签采样,建立“文档-主题-词序”关系,这种采样方式既符合语言的情感表达,又不会缩小词之间的主题联系.由此构建起无监督主题情感模型(unsupervised topic sentiment model,简称 UTSM),记为 UTSM(LDA-Col).模型如图 2 所示,模型中所用符号参见表 1.假设语料库中有 M 个文档,记为 $D=\{d_1,...,d_m\}$;共有 K 个主题,记为 $T=\{z_1,...,z_K\}$;共有 L 种情感,记为 $S_e=\{m_1,...,m_L\}$.文档 d 由 S 个句子组成,每个句子由 N 个词汇组成,记语料库中去重后的词汇表中词汇个数为 V.

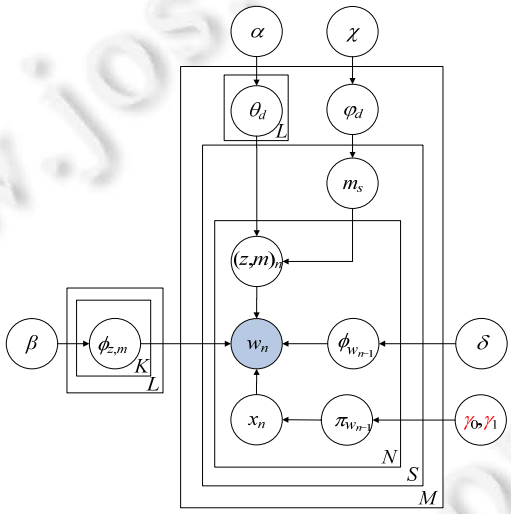


Fig.2 UTSM(LDA-Col) model
图 2 UTSM(LDA-Col)模型图

UTSM(LDA-Col)模型认为:每一篇文档 d 与 L 种情感的一个 Multinomial 分布相对应,将该 Multinomial 分布记为 φ_d,它是一个 L 维变量,服从带有超参数 χ 的 Dirichlet 先验分布,即 $\varphi_d \sim Dir(\chi)$;每一篇文档 d 与 L 种情感下的 K 个主题的一个 Multinomial 分布相对应,将该 Multinomial 分布记为 θ_d,它是一个 K 维变量,服从带有超参数 α 的 Dirichlet 先验分布,即 $\theta_d \sim Dir(\alpha)$;每个主题~情感对$(z,m)$又与词汇表中的 V 个单词的一个 Multinomial 分布相对应,将这个 Multinomial 分布记为 $\phi_{z,m}$,它是一个 V 维变量,服从带有超参数 β 的 Dirichlet 先验分布,即 $\phi_{z,m} \sim Dir(\beta)$.此外,还有连词分布 $\phi_{w_{n-1}}$ 和词搭配分布 $\pi_{w_{n-1}}$,其中:$\phi_{w_{n-1}}$ 服从带有超参数 δ 的 Dirichlet 先验分布,即 $\phi_{w_{n-1}} \sim Dir(\delta)$; $\pi_{w_{n-1}}$ 服从参数为 γ_0 和 γ_1 的 Beta 分布, $\pi_{w_{n-1}} \sim Beta(\gamma_0,\gamma_1)$,$\phi_{w_{n-1}}$ 是给定 w_{n-1} 情况下关于 w_n 的概率,即 $P(w_n|w_{n-1},x_n=1)$; $\pi_{w_{n-1}}$ 是给定 w_{n-1} 情况下关于 x_n 的概率,即 $P(x_n|w_{n-1})$.x_n 表示词搭配分配,x_n 的取值如下:

$$x_n=\begin{cases}1, & \text{if } w_n \text{ follows } w_{n-1}\\ 0, & \text{otherwise}\end{cases},$$

上式表示:若词 w_n 与 w_{n-1} 相连,则 x_n 等于 1;否则为 0.

3.2 无监督主题情感算法

3.2.1 UTSM(LDA-Col)产生算法

算法 2. UTSM(LDA-Col)产生算法.

输入:语料库中 M 个文档;

输出:文档-主题分布,文档-情感分布,主题情感-词序分布.

步骤:

1) 对于每对主题 z 和情感 m:

2) 从参数为β的 Dirichlet 分布中,抽取 Multinomial 分布$\phi_{z,m}$,即,采样$\phi_{z,m}$~$Dir(\beta)$

3) 对于每篇文档 d:

3.1) 从参数为χ的 Dirichlet 分布中,抽取 Multinomial 分布φ_d,即,采样φ_d~$Dir(\chi)$

3.2) 对于每种情感 j:

从参数为α的 Dirichlet 分布中,抽取 Multinomial 分布θ_{dj},即,采样θ_{dj}~$Dir(\alpha)$

3.3) 对于文档 d 中的每个句子 s:

3.3.1) 从φ_d中抽取一个 Multinomial 变量情感 m_s,即,采样 m_s~$Multi(\varphi_d)$

3.3.2) 对于句子 s 中的每个单词 $w_{d,n}$:

3.3.2.1) 从参数为γ_0和γ_1的 Beta 分布中,抽取二项分布 $\pi_{w_{n-1}}$,即,采样 $\pi_{w_{n-1}}$ ~$Beta(\gamma_0,\gamma_1)$

3.3.2.2) 从 $\pi_{w_{n-1}}$ 中抽取变量 x_n,若 x_n=0:

从θ_{dm}中抽取一个 Multinomial 变量主题 $Z_{s,n}$,即,采样 $Z_{s,n}$~$Multi(\theta_{dm})$;

从$\phi_{z,m}$中抽取一个 Multinomial 变量单词 w_n,即,采样 w_n~$Multi(\phi_{z,m})$;

3.3.2.3) 若 x_n=1:

从参数为δ的 Dirichlet 分布中抽取 $\phi_{w_{n-1}}$,即,采样 $\varphi_{w_{n-1}}$ ~$Dir(\delta)$;

从 $\phi_{w_{n-1}}$ 中抽取一个 Multinomial 变量单词 w_{n-1},即,采样 w_{n-1}~$Multi(\phi_{w_{n-1}})$.

3.2.2 UTSM(LDA-Col)求解结果

用 i 来表示词汇记号的索引号,i=(d,s,n),词汇记号 w_i=$w_{d,s,n}$ 表示与文档位置、句子位置相关的词汇,s_i 表示词汇记号 w_i 所在的句子. m_{s_i} 表示词汇记号 w_i 所在句子的情感分配,m_{-si} 表示除当前词汇记号所在句子外其他词汇记号所在句子的情感分配;z_i 表示词汇记号 w_i 的主题分配,z_{-i} 表示除当前词汇记号外的其他所有词汇记号的主题分配.有 w={w_i=t,w_{-i}},z={z_i=k,z_{-i}},m={m_{si}=j,m_{-si}}.利用 Gibbs 采样算法进行采样,文档-主题分布θ、文档-情感分布φ和主题情感-词分布ϕ的估计如公式(1)~公式(3)所示:

$$\hat{\theta}_{d,j,k}=\frac{n_{d,j}^{(k)}+\alpha}{\sum_{k=1}^{K}(n_{d,j}^{(k)}+\alpha)} \tag{1}$$

$$\hat{\varphi}_{d,j}=\frac{n_{d}^{(j)}+\chi}{\sum_{j=1}^{L}(n_{d}^{(j)}+\chi)} \tag{1}$$

$$\hat{\phi}_{k,j,w}=\frac{n_{k,j}^{(w)}+\beta}{\sum_{w=1}^{V}(n_{k,j}^{(w)}+\beta)} \tag{3}$$

其中,$\hat{\theta}_{d,j,k}$表示主题 k 在文档 d 的主题分布中情感为 j 的概率估计,$\hat{\varphi}_{d,j}$表示情感 j 在文档 d 的情感分布中的概率估计,$\hat{\phi}_{k,j,w}$表示词汇 w 分配的主题和情感分别为 k 和 j 的概率估计,$n_d^{(j)}$表示文档 d 中分配在情感 j 上的句子数,$n_{d,j}^{(k)}$表示文档 d 中分配在主题为 k 情感为 j 上的词数,$n_{k,j}^{(w)}$表示 w 分配在主题为 k 情感为 j 上的次数.

UTSM 算法中的 $p(w_2|w_1)$计算如下:

$$p(w_2|w_1)=\pi^{(w_1)}\phi_{w_2}^{(w_1)}+(1-\pi^{(w_1)})\sum_{(z,m)}\phi_{w_2}^{(z,m)}\frac{\phi_{w_1}^{(z,m)}}{\sum_{(z,m)}\phi_{w_1}^{(z,m)}} \tag{4}$$

3.3 微博情感倾向性分类及评价方法

利用 UTSM 算法的 $\hat{\phi}_{d,j}$ 可以得到情感 j 在文档 d 的情感分布中的概率估计,取每种情感倾向在文档 d 的情感分布中的概率估计的最大值可得到文档 d 的情感,即:

$$m_d=\arg\max_j\{\hat{\varphi}_{d,j}\mid j\in[1,...,L]\} \tag{5}$$

微博情感倾向性分类属于文本分类范畴,评价指标通常采用准确率(precision)、召回率(recall)和 F 值(F-measure):

$$F\_measure=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{6}$$

4 基于依存模式的微博情感要素抽取

微博情感要素抽取模块通过情感词双扩展、依存模式筛选和评价对象归并这 3 个步骤完成.

4.1 情感词双扩展算法

情感词双扩展算法的基本思想是:首先建立种子情感词典,通过种子情感与情感词的依存关系进行情感词扩展,再通过已知情感词和评价对象的依存关系进行评价对象获取;然后,通过已知评价对象和情感词的依存关系进行情感词扩展,这样循环进行直至无新的情感词和评价对象出现.

4.2 评价单元依存模式筛选

为了克服仅仅依靠依存关系提取情感词和评价对象可能带来的噪音,在进行依存分析时,增加对词语词性的限制,建立起评价单元依存模式优先级,从高到低依次为 1~6.在评价单元依存模式列中,粗体表示情感词和评价对象:

(1) $\{\boldsymbol{NP}\}\xleftarrow{SUB}\{\boldsymbol{AV}\}$

其中,$\{NP\}=\{NR,NT,NN,PN\}$,表示固有名词、时间名词、其他名词和代名词;$\{AV\}=\{VA,VV\}$,表示谓语形容词和其他动词.

模式(1)表示主语为评价对象,主语依存的动词为情感词,评价对象与情感词之间是直接依存关系.如例句“渭南城管真变态!”,依存分析如下:

评价单元为(变态,城管),将“渭南”与“城管”作归并操作后,得到评价单元为(变态,渭南城管).

(2) $\{NP\}\xleftarrow{SUB}\{VC\}\xrightarrow{PRD}\{NP\}\xrightarrow{NMOD}\{JV\}$

其中,$\{VC\}$表示系动词;$\{JV\}=\{JJ,VA\}$,表示形容词和谓语形容词.

此模式表示谓语为系动词,主语为评价对象,表语的定语为情感词,主语与表语定语之间是扩展的依存关联关系.如例句“他是一个优秀的学生.”,依存分析如下:

“优秀$\xleftarrow{SBAR}$的$\xleftarrow{NMOD}$学生”做剪枝操作得到“优秀 优秀$\xleftarrow{SBAR}$学生”,再用模式(2)得到评价单元为(优秀,他).

(3) $\{\boldsymbol{AV}\}\xrightarrow{OBJ}\{\boldsymbol{NP}\}$

模式(3)表示谓语不为系动词,宾语为评价对象,谓语为情感词,宾语与谓语之间是直接依存关系.如例句“我喜欢 ipad3 的屏幕.”,依存分析如下:

评价单元为(喜欢,屏幕),将“ipad3$\xleftarrow{DEP}$的$\xleftarrow{NMOD}$屏幕”做归并操作得到“ipad3 的屏幕”,最后得到评价单元为(喜欢,ipad3 的屏幕).

(4) $\{\boldsymbol{NP}\}\xleftarrow{SUB}\{VV\}\xrightarrow{VMOD}\{\boldsymbol{AV}\}$

模式(4)表示主语为评价对象,动词的补语为情感词,主语与动词的补语是依存关联或扩展的依存关联关系.如例句“他开车很溜.”,依存分析如下:

评价单元为(溜,他).

(5) $\{\boldsymbol{JV}\}\xleftarrow{NMOD}\{\boldsymbol{NP}\}\xleftarrow{SUB}\{VV\}$

模式(5)表示主语为评价对象,修饰主语的定语为情感词,主语与主语定语是直接或间接依存关系.如例句“令人敬重的方阵过来了.”,依存分析如下:

评价单元为(敬重,方阵).

(6) $\{VV\}\xrightarrow{OBJ}\{\boldsymbol{NP}\}\xrightarrow{NMOD}\{\boldsymbol{JV}\}$

模式(6)表示宾语为评价对象,修饰宾语的定语为情感词,宾语与宾语定语是直接或间接依存关系.如例句“他有部非常漂亮的手机.”,依存分析如下:

对“漂亮$\xleftarrow{NMOD}$的$\xleftarrow{SBAR}$手机”做剪枝操作得到“漂亮$\xleftarrow{SBAR}$手机”,再用模式(6)得到评价单元为(漂亮,手机).

除上述 6 种模式外,对于句子只含有偏正结构短语和标点时按模式(5)进行处理.如例句“闹心的临时工!”、“疯狂的大葱!”等.

注意,评价单元依存模式从 1~6,优先级别逐渐降低.即:先匹配评价单元依存模式 1,若匹配成功,对候选评价单元进行筛选;若不成功,则匹配评价单元依存模式 2,依此类推.

4.3 评价对象归并算法

根据对具有依存关系的 2 个成分进行分析发现,有些依存词对是无意义的.与词语“的”有关的依存词对(服务态度,的)和(的,旅店)对于倾向性分析没有意义,影响文本倾向性分析精度,因此采用以下策略对依存树进行修剪.

(1) 剪枝:对依存树的剪枝操作是指将当前节点进行删除,其作用域仅仅局限于当前节点中,其子节点并不发生作用.假设有一棵依存树 AHB,若进行剪枝操作,即,直接将其子节点 B 依靠在父节点 A 上,当前节点不加入依存词对的生成分析,剪枝操作通常应用于助词和介词标签等,可将具有间接依存关系的 2 个词变成具有直接依存关系;

(2) 归并:依存树的归并是将子节点和父节点归并为一个节点,新的节点对应的词是子节点和父节点的结合.假设进行依存树的归并操作,即,将节点 B 和 H 归并为 B'节点.归并操作用于否定词和情感词中,将否定词和情感词归并为一个新的情感词,新的情感词的极性是原情感词的极性取反.通过归并可将具有间接依存关系的 2 个词变成直接依存关系.

在情感词和评价对象的 6 种依存模式中,模式(1)、模式(2)、模式(4)和模式(5)中的评价对象都为主语,而模式(3)和模式(6)中的评价对象为宾语.为抽取出尽可能完整和明确的评价对象,需要对评价对象进行归并操作,得到完整的主语成分或宾语成分.如例句“我喜欢 ipad3 的屏幕.”中,应该抽取出“ipad3 的屏幕”,而不仅是“屏幕”.

评价对象归并算法:为区分原始的评价对象和归并后得到的评价对象,将原始的评价对象称为评价对象基准词.评价对象归并时,从评价对象基准词的左邻词开始,从右往左依次判断:若其父亲节点为基准词或为其右邻词,将其归并到评价对象中;若为句首或其父亲节点不为基准词或不为其右邻词时,停止归并.

4.4 微博情感要素抽取算法

算法 3. 微博情感要素抽取算法.

输入:已判别的观点句集合 *ValueSentenceID*[·];

输出:新抽取的完整的评价对象.

步骤:

1) 采用情感词双扩展算法对观点句进行情感词扩展;获取候选情感词和候选评价对象;

2) 采用评价单元依存模式按照优先级对候选评价单元进行筛选;

3) 通过评价对象归并算法得到完整的评价对象.

将待分析句子中所有在情感词库中出现的情感词和扩展出的情感词列为候选情感词,所有名词或代名词列为候选评价对象.

5 实验与分析

5.1 实验环境与实验系统架构

(1) 实验环境:实验所用依存分析采用ctbparser开源工具包,分词和词性标注采用宾州中文树库标准,实验是在 Visual studio 2010 上用 C++语言实现;

(2) 实验系统架构:中文微博情感评测软件包含 4 个模块,具体如图 1 所示.

5.2 基于UTSM(LDA-Col)模型的中文微博情感模块功能实验

本项实验的数据集有 2 个:我们采集了 2015 年的“东方之星沉船”微博话题作为话题 1,从中抽取 500 条评论语句;我们还选取了2012年公开评测活动中提供的“菲军舰撞船事故”微博话题作为话题2,同样从中抽取500条评论语句,在每个话题的语句中可筛选出表达主观情感的关键词汇.话题 1 和话题 2 分别代表“哀”和“怒”2 种情感.

5.2.1 主题-情感词汇发现实验

利用 Gibbs 采样算法对 LDA-Col 模型进行采样,参数设置如下:α=1,β=0.01,γ_0=0.1,γ_1=0.1,δ=0.1,迭代次数 N=1000.为方便统计主题分类准确率和召回率,取主题数 K=4,在产生词汇时,设置词的最大搭配长度 $MAXC$=4,情感分类 L=7(高兴、喜好、愤怒、厌恶、恐惧、悲哀、惊讶这 7 类).

通过 LDA-Col 模型对文档集进行主题-情感词汇发现,实验结果如下.

话题 1 有 25 类关键情感词汇,话题 2 有 30 类关键情感词汇,我们按照 7 类情感分类词库(高兴、喜好、愤怒、厌恶、恐惧、悲哀、惊讶这 7 类情绪词典),相应的词汇按照该词汇在语句集中出现的频数大小进行升序排列,获得如图 3 和图 4 所示的情感词汇散点分布图.

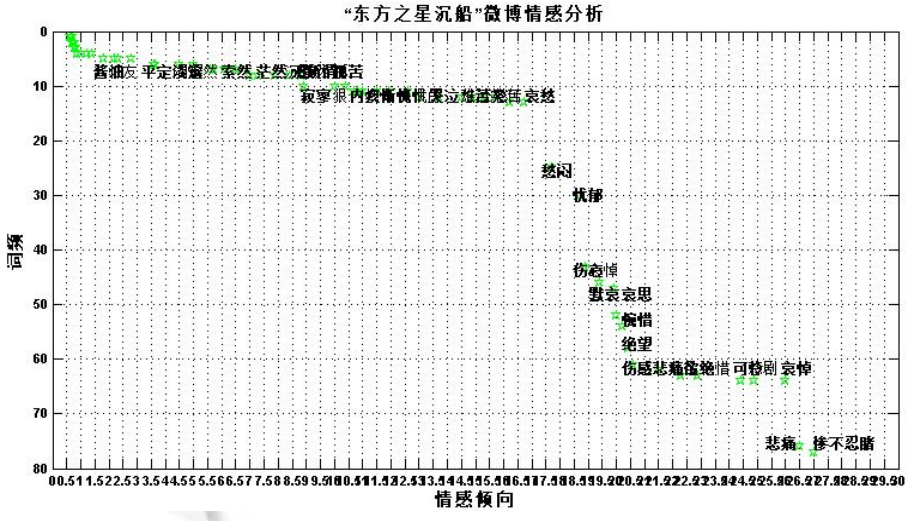


Fig.3 Scatter diagram for microblog sentiment classification of shipwreck of East Star

图 3 “东方之星沉船”微博情感分析散点图

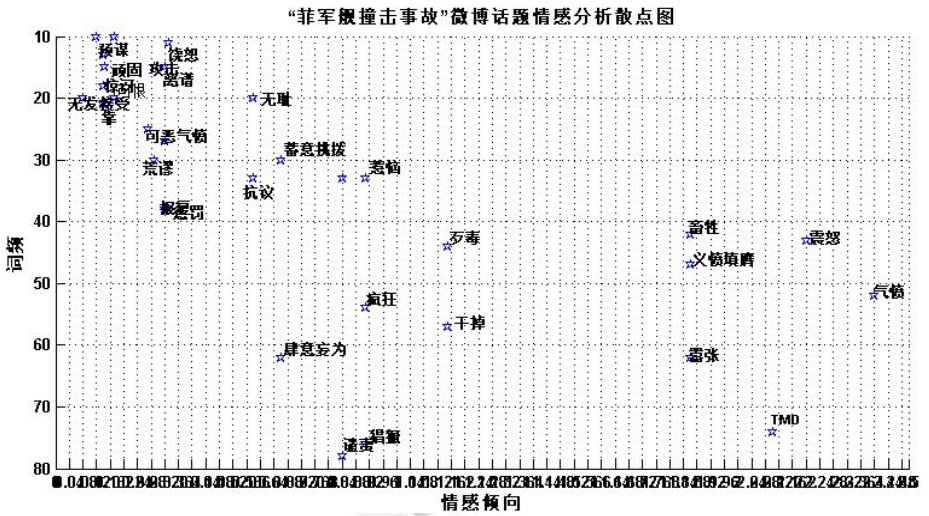


Fig.4 Scatter diagram for microblog sentiment classification of impact trouble of Philipine warship

图 4 “菲军舰撞击事故”微博情感分析散点图

相应的词汇类型有与其对应的情感强度;相应的词频也有与其对应的情感强度权重系数.词频和附加权重的情感强度共同决定了该词汇的情感倾向.图中词频出现越高,词汇对应的情感强度越强,则分布越靠近散点图右下方.图 3 中,“惨不忍睹”一词最能够体现“东方之星沉船”话题的微博情感倾向;而图 4 中,“气愤”一词最能够代表“菲军舰撞船事故”话题的微博情感倾向.

5.2.2 微博细粒度情感倾向性分析实验

当抽取的词汇量较少且与情感词典匹配率较低时,情感倾向较难判断,三维峰值分布较为散乱,峰值分布呈非线性.

图 5、图 6 是两个话题的细粒度情感分析结果,两个图的三维分布规律性较强,峰值分布更具线性,情感取向判断较直观.虽然话题 2 所表现的峰值比较离散,这是由于该话题的情感词汇种类较话题 1 更为杂乱,情感倾向性更趋多元化.但是仍旧可以将话题情感明显锁定在(31,23,90),(33,17,75),(32,15,70)这 3 个峰值坐标所表示的情感词汇“气愤”“TMD”“嚣张”上面.

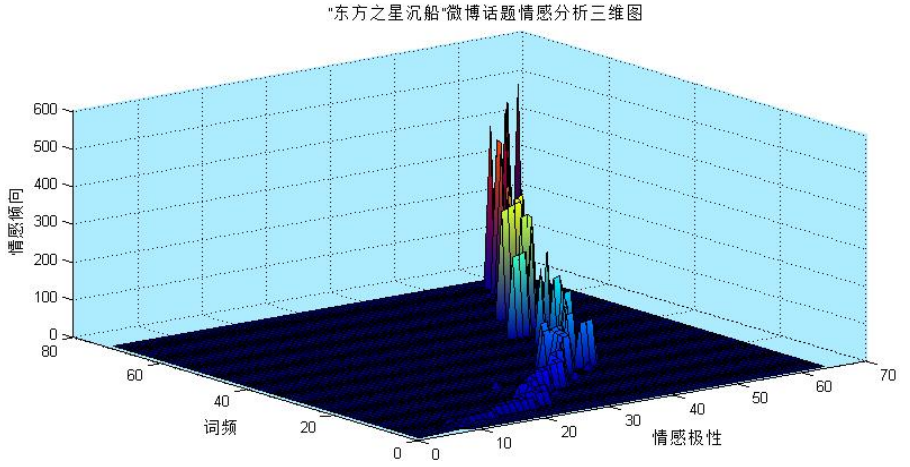


Fig.5 Diagram for microblog sentiment fine grain classification of shipwreck of East Star

图 5 “东方之星沉船”微博细粒度情感分析

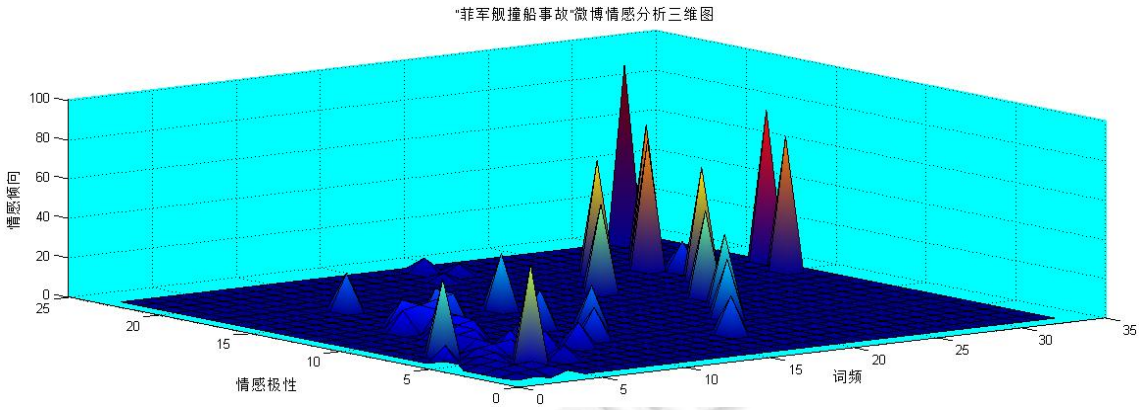


Fig.6 Diagram for microblog sentiment fine grain classification of impact trouble of Philipine warship

图 6 “菲军舰撞击事故”微博细粒度情感分析

5.2.3 微博情感随机评测

依据上述方法,我们还对“喜好恶惊惧”其他 5 类情感进行分析.随机抽取了 100 个微博评论句子,每个句子彼此相互独立.程序对该 100 个句子逐条进行评测,通过判断分析句子中的关键词汇,进行相应的量化处理,得到的评测结果涵盖了每个句子在 7 个情感维度上的分布,如图 7 所示.

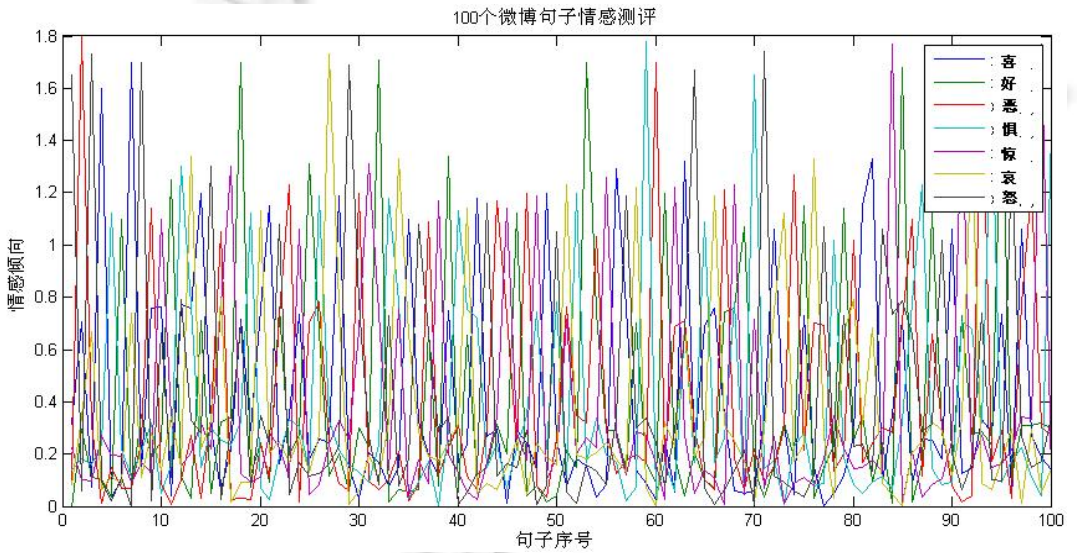


Fig.7 Diagram for microblog sentiment classification of 100 stochastic sentences

图 7 100 个随机句子微博情感分析图

图 7 中,不同颜色的线条代表不同的情感类型.可见:同一句子的不同纬度的情感分布中,总有一类情感的峰值脱颖而出,盖过其他的情感,那么这个纬度的情感倾向就可以近似为该句子的情感色彩.

为更直观地表现不同句子的情感分布特点,图 8 将分布置于三维空间中,X轴对应句子序号,Y轴对应 7 种情感类型,Z轴对应情感倾向.三维图中可见:峰值的分布是稀疏的,但每一个句子都有一个最为突兀的峰值,这意味每一句评论语句都在表达着一类情感,而该类情感倾向的强度可以掩盖其余杂乱而微弱的情感倾向.

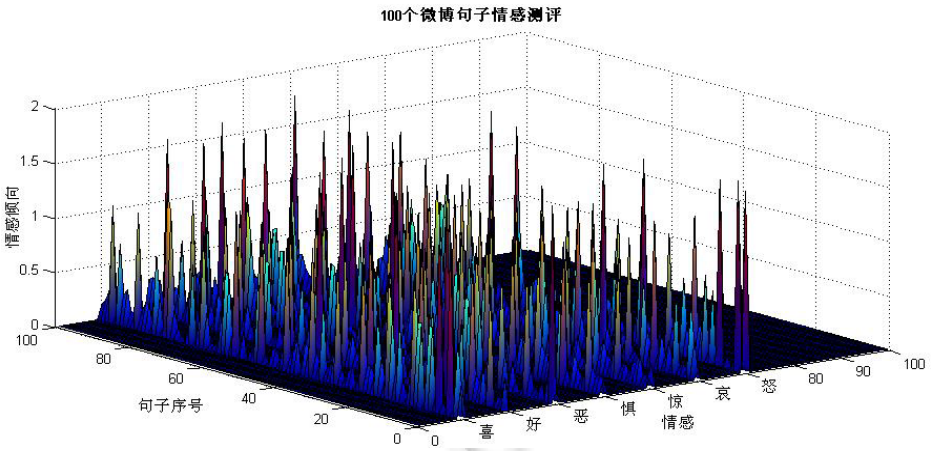


Fig.8 3-Dimensional diagram for microblog sentiment classification of 100 stochastic sentences

图 8 100 个随机句子微博情感分析三维图

5.2.4 模型对比实验结果与分析

将本文提出的 UTSM(LDA-Col)模型与 ASUM 模型、JST 模型和 Pang 方法[30]进行情感倾向性分类效果对比,其中,UTSM(LDA-Col)模型、ASUM 模型和 JST 模型都是主题情感混合模型,Pang 方法是有监督的机器学习方法.ASUM 模型和 JST 模型的原文中都用到了种子情感词作为先验知识,由于种子情感词的不同对结果影响较大,本实验统一采用无先验知识;Pang 方法中使用信息增益方法进行特征选择,选取了 2 000 个特征,分类器采用 SVM,选用 10 倍交叉验证来估计 SVM 分类器的准确率和召回率.10 重交叉验证法是将数据集分成 10 份,轮流将其中 9 份用作训练,1 份用作测试.

实验数据集建立如下:从大众点评网(http://www.dianping.com,2015-04-08)下载关于××快递、××烧烤两种类型网页和中国科学院谭松波博士公布的关于酒店和计算机的情感倾向性分类数据集(Tan Songbo.ChnSentiCorp [EB/OL]. http://www.searchforum.org.cn/tansongbo/senti_corpus.jsp,2015-9-12),整理得到 9 180 个文本,每种数据集的正负文本数见表 4.

Table 4 Data sets description

表 4 数据集说明

	快递(Corp1)	烧烤(Corp2)	酒店(Corp3)	计算机(Corp4)
Pos	1 150	910	1 130	1 270
Neg	1 140	1 230	1 200	1 150
Sum	2 290	2 140	2 330	2 420

数据集预处理在 VC 环境下进行,数据处理在 Matlab 环境下进行,需要下载主题模型工具包(Topictoolbox [DB/OL].http://psiexp,ss.uci.edu/research/programsdata/topictoolbox.zip,[2015-9-10]).

4 种方法在 4 个数据集上的情感倾向性分类准确率、召回率和 F 值见表 5~表 7.

Table 5 Sentiment classification precisions comparison (%)

表 5 情感倾向性分类准确率对比(%)

Precision	Corp1	Corp2	Corp3	Corp4
Pang	86	81.8	78.7	80.9
UTSM(LDA-Col)	78.94	77.92	81.2	76.4
ASUM	74.5	71.2	78.3	73.2
JST	58.7	51.7	53.4	50.1

Table 6 Sentiment classification recalls comparison (%)

表 6 情感倾向性分类召回率对比(%)

Recall	Corp1	Corp2	Corp3	Corp4
Pang	85.9	78.8	79.8	81.8
UTSM(LDA-Col)	83.82	77.79	82.27	73.05
ASUM	83.4	75.6	77.3	73
JST	73.9	67.5	69.7	70.3

Table 7 Sentiment classification *F*-measures comparison (%)

表 7 情感倾向性分类 *F* 值对比(%)

F-measure	Corp1	Corp2	Corp3	Corp4
Pang	85.95	80.3	79.25	81.35
UTSM(LDA-Col)	81.38	77.86	81.74	74.95
ASUM	78.95	73.4	77.8	73.1
JST	66.3	59.6	61.55	60.2

从表中可以看出:

- 综合考虑准确率和召回率,4 种方法中,效果最好的是 Pang 方法,在 4 个数据集上的 *F* 值平均值为 81.71%,但由于 Pang 方法是基于向量空间模型的有监督学习方法,需要先对标注好的样本进行训练才能测试;
- 3 种主题情感混合模型中,效果最好的是 UTSM(LDA-Col)模型,*F* 值平均值为 78.98%;接下来为 ASUM 模型,*F* 值为 75.81%;效果最差是 JST 模型,*F* 值平均值为 61.91%.UTSM(LDA-Col)模型比 ASUM 模型的 *F* 值平均值提高了 3.2%,证明了本文提出的对每个句子采样情感标签,对每个词采样主题标签的主题情感混合模型在情感倾向性分类上的有效性和对词语进行搭配合并能够提高情感倾向性分类的效率.由于 JST 模型每次采样情感标签时,对每个词进行采样,不符合自然语言的情感表达,导致 JST 模型效果最差.

总体来说,本文构建的 UTSM(LDA-Col)模型情感倾向性分类的性能比有监督情感倾向性分类方法稍差,低 2.7%;但在无监督的情感倾向性分类方法中效果最好,比 ASUM 模型提高了 3.2%,比 JST 模型提高了 17%.

5.2.5 不同特征组合对比实验结果与分析

(1) 5 级情感词库效率对比实验

为了验证本文提出的 5 级情感词库(见表 3)的大小与准确率和召回率之间的关系,采用本文第 4.1 节提出的情感词扩展方法的扩展效率对比结果如图 9 所示,结果只统计扩展识别的情感词.

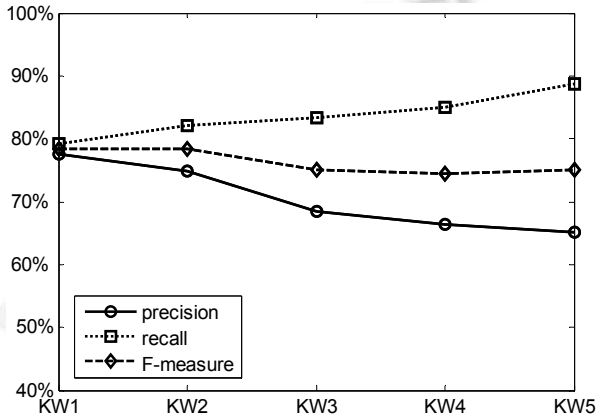


Fig.9 5-Level emotion thesaurus efficiency comparison

图 9 5 级情感词库效率对比图

从图中可以看出:准确率随着情感词库的增大而降低,召回率随着情感词库的增大而增大,最简版词库KW1的召回率最低但准确率最高,完整版词库KW5的召回率最高但准确率最低.由于KW1已包含基本常用的情感词,故其的召回率效果较理想,达到 79.2%,能够扩展出其他不常用情感词或在该句中是情感词的词.由于ctbparser的句法分析精度有限(在CTB6标准测试集上句法分析精度为81%),导致扩展准确率不高.随着情感词库的增大,分析错误导致扩展出更多错误的情感词,使得准确率下降.总体来说,本文提出的情感词扩展识别方法是有效的,在各级情感词库上平均准确率达到70.48%,平均召回率达到83.7%,平均 F 值约76.3%.

(2) 不同特征组合对比实验

结合文献[27]针对模型 SVM,CRF 等实验结果,与 UTSM(LDA-Col)算法处理公开测评数据结果进行对比,分析3个模型针对不同特征组合的准确率及其原因.

为了训练出一个较好的SVM模型,文献[27]选取了5类文本特征,包括词性、情感词、否定词、程度副词及特殊符号.将正向表情符号作为正向情感词处理,将负向表情符号作为负向情感词处理.如果情感词之前出现否定词,则情感词的情感极性反向.实验为了找出最优特征组合,评估每种特征对 SVM 模型作用的大小,首先将词性和情感词作为特征进行两组实验,然后将词性和情感词作为特征组合进行实验,最后在词性和情感词的特征组合中分别加入否定词、程度副词和特殊符号特征.通过多组特征组合的实验,评估出不同特征对SVM模型的作用,并找出最优特征组合.实验使用准确率作为评估指标,表 8 中,编号 1~编号 3 号列出了文献[27]的SVM不同特征组合的3项最佳实验结果.该实验仅开展了准确率实验,未提供召回率和 F-measure 的实验结果.

CRF 一般用于序列标注任务中,而文本情感分析是要判断整个句子的情感倾向,并不是一个典型的标注任务.为了情感分析转换成标注问题,文献[27]将文本的极性对应到文本中每个词语的极性,通过标注每个词语的极性来判断文本的极性.换句话说:如果一条微博的极性为正向,则将微博中每个词语的极性都标注为正向;如果一条微博的极性为负向,则将微博中每个词语的极性都标注为负向.表 8 中,编号 4~编号 6 列出了文献[27]的CRF不同特征组合的3项最佳实验结果.该实验仅开展了准确率实验,未提供召回率和 F-measure 的实验结果.

分析上述SVM和CRF实验结果精确的重要原因是人工预处理文本情感和极性,这样的工作方式如同人与计算机结合共同完成文本情感分析工作.而面对海量网络微博实时处理现实,没有人提供海量微博的人工标注和情感极性判断预处理结果,因此,这两种方法无法适用于计算机自动微博处理.

Table 8 Experiment result for different feature combination

表 8 不同特征组合的实验结果

编号	模型	特征组合	正面语料准确率(%)	负面语料准确率(%)	整体准确率(%)	整体召回率(%)	F-measure(%)
1	SVM[27]	词性+情感词+否定词	87.95	**89.59**	**88.72**		
2		词性+情感词+否定词+程度副词	87.84	88.87	88.32		
3		词性+情感词+否定词+特殊符号	**88.79**	88.25	88.54		
4	CRF[27]	情感极性+否定词	90.47	89.97	90.24		
5		情感极性+否定词+程度副词	90.40	**90.36**	90.38		
6		情感极性+否定词+程度副词+特殊符号	90.51	**90.36**	**90.44**		
7	UTSM^A	情感词+否定词	79.2	78.2	78.6	80.8	79.7
8	UTSM^B	情感词+极性词+否定词	84.2	81.4	82.8	53.7	65.1

为验证由本文提出的UTSM(LDA-Col)方法的有效性,分别做了以下实验.

- UTSM^A:以词库KW1为基础,采用“情感词+否定词”作为特征组合,对表4中的数据集进行处理,得到的结果,参见表 8 中的编号 7.实验结果自动处理生成,无人工干预.补充说明一点:该结果的召回率为80.8%,F 值为79.7%(如图9所示);
- UTSM^B:以词库KW5为基础,采用“情感词+极性词”作为特征组合,对NLP&CC2012的观点句任务进行处理,得到的结果,参见表8中的编号8.补充说明一点:该结果的召回率为53.7%,F 值为65.1%(如图10所示).在该公开评测活动中,UTSM^{B}获得准确率第2名,与第1名仅差0.007.其他51组结果的准确率均低于本文方法.后文图10的结果数据是由公开评测主办方提供给参评单位后画的.

通过对两组实验结果的分析得知,实验直接对原始数据进行处理、提供结果,工作速率秒级响应.与SVM,CRF 实验的最大区别是无需人工干预,人工标注和人工确定文本极性.即,我们的方法适合自动处理海量微博.

5.3 中文微博情感分析公开评测

公开评测数据集采用 NLP&CC2012 发布的来自腾讯微博的真实数据集.评测数据包括 20 个话题,17 500条微博,32 000 个句子,数据集大小为 5.6MB,数据采用 Unicode(utf-16)编码的 xml 格式,已经预先切分好句子.

5.3.1 观点句识别公开评测结果与分析

观点句识别公开评测共有 34 个单位提交了 53 组结果,所有结果的微平均评测结果如图 10 所示,其中,编号17 是采用本文提出的观点句识别模块参加观点句识别公开评测的结果.

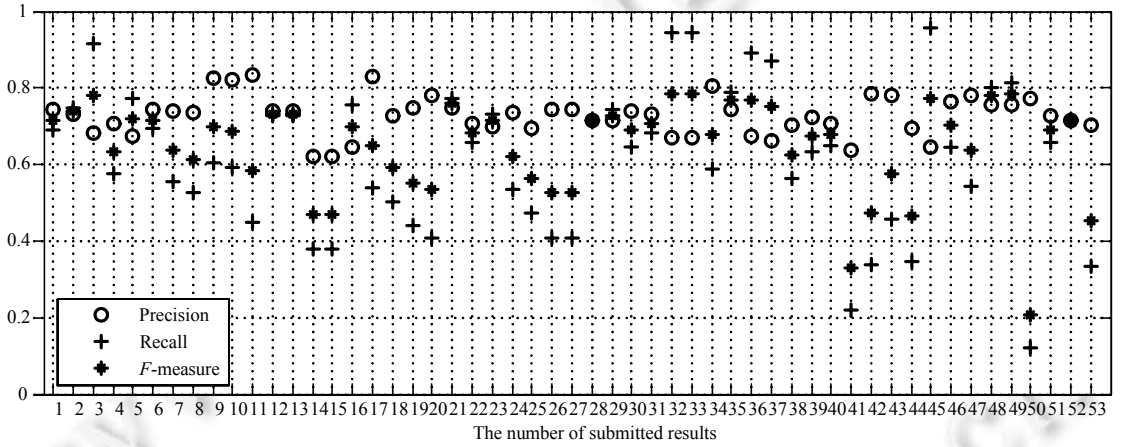


Fig.10 Open evaluating results of opinion sentence identification

图 10 观点句识别公开评测结果

从图 10 中可以看出:本文方法在微平均评测结果中准确率排名第 2,与排名第 1 的仅差 0.007;召回率结果在所有结果中处于中下,在微平均评测结果中比平均值低 0.078.虽然准确率较高,但综合考虑准确率和召回率,本文方法的 F 值仅与平均值约持平,在微平均评测结果中比平均值略高约 0.004.

通过观点句识别评测结果,分析本文方法进行观点句识别的步骤,发现有两种情况影响召回率:一是情感词典不够全面,二是对于不含情感词典的观点句不能识别出来.通过图 10 我们还发现:从所有结果看,召回率的浮动空间比准确率大多了,微平均评测结果的准确率基本都在 0.60~0.84 之间浮动,但是相应的召回率却在0.10~0.96 之间浮动,其中,召回率最差的低至 0.112,召回率最好的高达 0.96.相比于召回率,说明准确率是一个比较难提高的评价指标.

5.3.2 情感要素抽取公开评测结果与分析

微博情感要素抽取评测采用严格评价和宽松评价两种方式,在严格评价中,要求提交的情感对象在整条微博中的起始和终止位置和答案完全相同,且情感对象极性也相同时才算评价正确.根据评测大纲,在抽取情感对象时,要求抽取出完整和明确的对象.如在例句“ipad3 的屏幕很棒!”中,按照严格评价指标,要求抽取出的情感对象是“ipad3 的屏幕”而不是“屏幕”.宽松评价指标评测时,只要抽取出“ipad3 的屏幕”中的任意词语或短语都算正确.所以,宽松评价不能作为情感要素抽取任务的准确评测依据.

情感要素抽取评测要求找出微博中每条观点句作者的评价对象,即情感对象,同时判断针对情感对象的观点极性,提交格式包括微博编号、句子编号、情感对象、起始位置、终止位置、观点倾向.图 11 是 NLP&CC2012公开评测微博情感要素抽取(微平均严格评测)的全部结果.

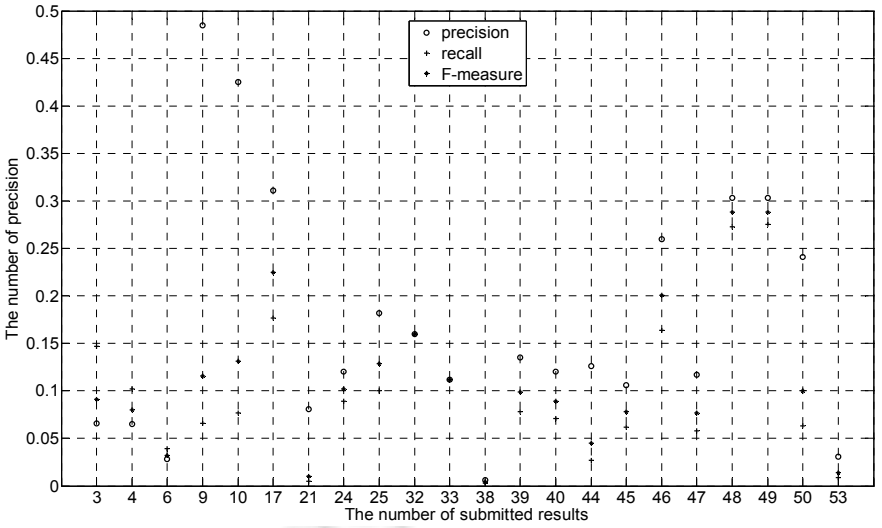


Fig.11 Strict open evaluating results of emotion expression extraction

图 11 微博情感要素抽取微平均严格评测结果

从图 11 中(http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html)可以看到:编号 9 和编号 10 的两个单位所用方法通过牺牲召回率换取高准确率,结果导致 *F* 值不理想;编号 48 和编号 49(是同一个单位)所用方法准确率和召回率结果相差较小,*F* 值最好.相比之下,本文提出的情感要素抽取方法(对应编号 17)的微平均严格评测结果的准确率、召回率和 *F* 值在所有参赛单位排名第 2.这个结果表明:通过定义 6 种评价单元依存模式对情感对象进行归并,然后进行情感要素抽取的算法是有效的.

本文方法的准确率和召回率还不够理想,原因分析如下.

(1) 实现中没有结合上下文开展指示代词消解,如在例句“小明就读于北京大学.他是一名优秀的学生.”中,抽取到的评价单元为(他,优秀),而正确答案应该为(小明,优秀);

(2) 上下文中无依存关系的情感词若不在情感词典中,无法通过依存关联扩展得到,影响了召回率;

(3) 针对中文微博分词的精度低下以及依存分析结果不足影响了准确率.由于情感要素判断需要把评价短语抽取出来,所以网页的情感要素抽取精度(<70%)仍然远低于现有的文本中文分词精度(>90%).中文微博由于自身不够遵循中文语言特点和规律,在微博信息计算机自动处理方面仍是中文信息处理的一项难题.

5.4 基于公开评测的中文微博情感分析算法比较

本节以公开评测结果及会议研讨论文为基础,比较各种中文微博情感分析算法在处理过程中的有效性.

5.4.1 观点句识别算法比较

图 12 是采用基于规则的观点句识别算法比较,图 13 是采用基于句法分析的观点句识别算法比较(图中横坐标为提交的结果编号,纵坐标为实验结果).

从以上两个图中可以得出以下结论.

(1) 观点句识别算法准确度都未突破 84%,这一点也暗合了 Ctbparser 工具句法分析精度在 CTB6 标准测试集上的句法分析最好结果为 81%;

(2) 准确率与召回率存在相关性,根据我们的实验分析,这种相关性与情感词典大小有关,随着情感词典的增大,通过扩展得到的情感词汇增多,从而导致可能出现了部分错误的情感词,使得准确率下降,但召回率得以提升.根据本文使用的情感词典大小推测,编号 29 提交的结果使用的情感词典应该是较为全面的大词典;

(3) 为完成观点句识别工作,可以采用多种方法的混合形式,如图中的编号 29 以及我们的结果(OURs),同时采用了基于规则的方法和句法分析相结合的方法.

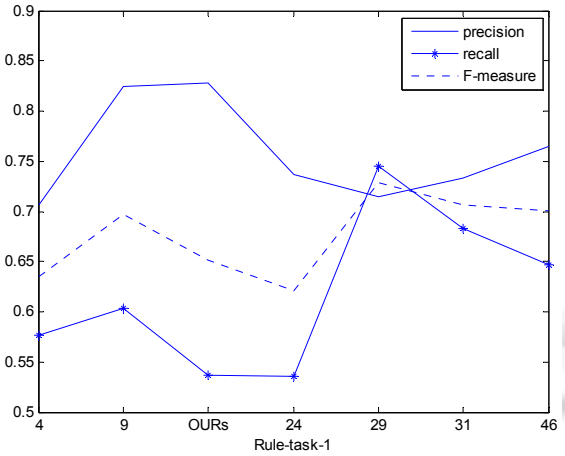


Fig.12 Comparison of opinion sentence identifications based on rule

图 12 基于规则的观点句识别算法效果比较

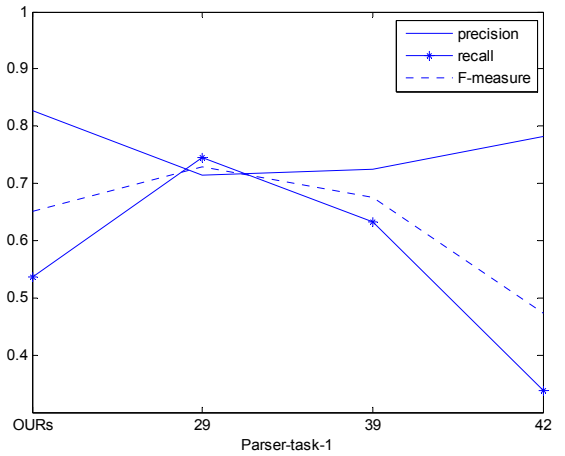


Fig.13 Comparison of opinion sentence identifications based on parser

图 13 基于句法分析的观点句识别算法效果比较

5.4.2 情感倾向性分析算法比较

图 14 给出了中文微博情感倾向性分类公开评测用到的各种算法和评测结果.

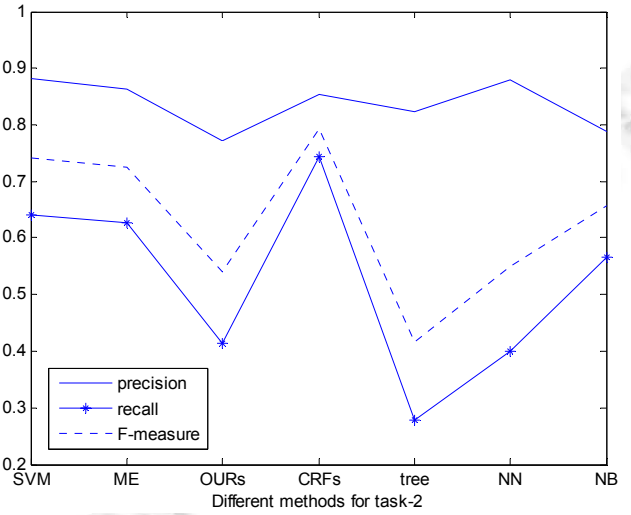


Fig.14 Open evaluating results and algorithms of emotion tendency classification

图 14 中文微博情感倾向性公开评测用到的算法和评测结果

下面给出简要分析.

(1) 经过比较可以看出:条件随机场[22]算法提交的结果成绩优异,在准确率、召回率以及 *F* 值 3 个方面都获得了十分满意的结果.其中:在准确率上,SVM,CRF,Tree,NN 以及 OURs 的准确率接近,SVM 模型的准确率最高,比 CRF 高 0.034;但是在召回率上比 CRF 少了 0.113,导致 *F* 值比 CRF 差;

(2) 最大熵模型(maximum entropy,简称 ME)的情感倾向性判断的准确率和 CRF 接近,但是召回率远远少

于 CRF,导致它的 *F* 值也低于 CRF,其他模型的分析结果都远远落后于 CRF.从中可以看出,CRF 算法在中文微博情感分析评测中效果显著;

(3) 因为 CRF 在上述公开评测中表现出优良的准确率,我们又以 CRF 为主要研究对象,查找其中的内在因素.通过查阅资料我们认为,常用的情感分析都是以句子为单位进行微博信息处理和分析,它过于单一,若是以每一条微博为单位,结合微博中的句子的上下文关系,找出观点微博,则更具有应用价值.这样,CRF 与 ME 模型就表现出相对优势了;

(4) CRF 属于无向图模型,具有很强的推理能力,并且能够使用复杂、有重叠性和非独立的特征进行训练和推理,能够充分地利用上下文信息作为特征,还可以任意地添加其他外部特征,使得模型能够获取的信息非常丰富.同时,CRF 解决了最大熵模型中的 Label Bias 问题,从理论上讲,CRF 非常适用于中文的词性标注;

(5) CRF 模型也有不足:在使用 CRF 方法的过程中,特征的选择和优化是影响结果的关键因素,特征选择问题的好与坏,直接决定了系统性能的高低.另外,训练模型的时间长达数 10 小时,且获得的模型很大,在一般的 PC 机上无法运行.

5.4.3 情感要素抽取算法比较

参加微博情感要素抽取任务的单位有 4 家到会进行了交流汇报,我们又增加了 2013 年参加同一类公开评测前 3 名队伍的结果,分别是 CUC2013,NJST2013 和 Tsinghua2013(https://www.softconf.com/e/nlpcc2013/),具体结果如图 15 所示.

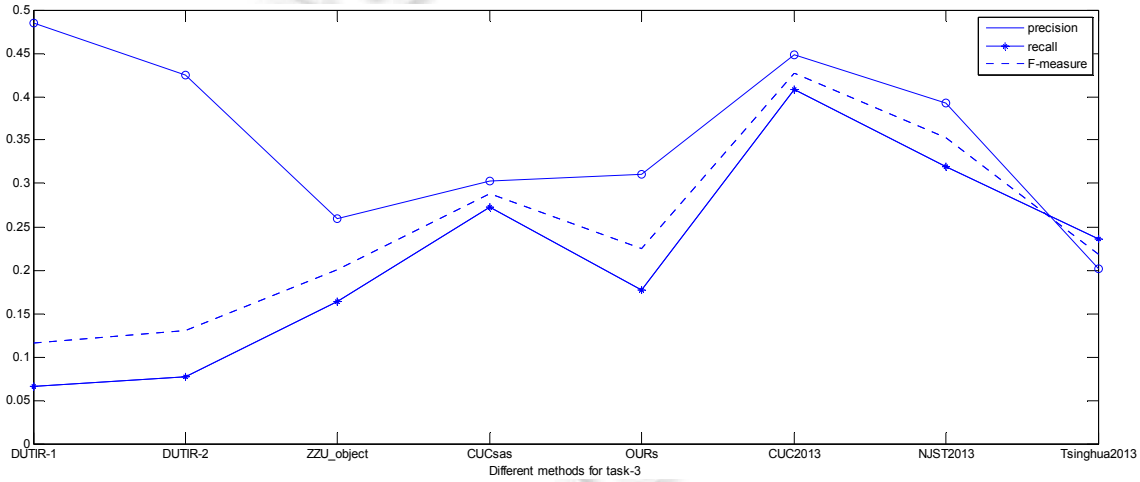


Fig.15 Comparison results of different methods for emotion expression extraction

图 15 不同的情感要素抽取算法效果比较

分析如下:

(1) 编号 DUTIR-1 以及 DUTIR-2 是同一个单位提交的两个结果,均采用基于规则与 CRF 的中文信息处理技术,其特点为准确率较高,但召回率很低,从而导致其 *F*-值也很低;

(2) 编号 ZZU_object 采用了基于哈工大汉语句法分析器的依存句法分析方法,从结果上来看,该算法的效果在 5 种算法中居中;结合该单位提交的观点句识别以及情感倾向性分类两个任务结果,可以推测得出:该单位开展中文信息处理主要注重基于规则的方法;

(3) CUCsas 方法(http://tcci.ccf.org.cn/conference/2012/dldoc/NLPCC2012papers/workshoppapers/sen/003.pdf,2016-4-10)是:在归纳微博语言特点的基础上构建基于短语的情感词典,通过短语规则确定句子极性,重点研究否定形式;并建立基于话题的 OBJ 表单等策略,完成微博情感分析.该方法采用了人工预

测经验值法,预判80%的微博是消极的,但在交流中,未给出如何依托计算机开展批处理微博情感分析研究;

(4) 本文提出的情感要素抽取方法(OURs)在公开评测中表现优良,取得准确率第 3、召回率和 *F*-值均排名第 2 的好成绩.在与 2013 年的公开评测前 3 名的结果比较后,成绩也能保持在前 3 名.这也充分说明本文提出的方法在依托计算机的情况下能够胜任开展中文微博自动处理工作,为今后实施基于计算机的实时中文微博信息处理提供了有力的软件方案.

6 结束语

本文构建了一个可自动批处理中文微博信息的情感分析系统,给出了系统 3 个主要模块的实现算法,运用该系统开展了针对私有数据的实验,并参加了 NLP&CC2012 公开评测,在观点句识别以及情感要素抽取 2 个任务中均取得第 2 名,验证了本文方法具有准确率高、自动化程度高、系统效率高的优点.

2016 年 5 月 13 日,Google Research 宣布其拥有的世界准确度最高的自然语言解析器 SyntaxNet 开源,据介绍,Google 在该平台上训练的模型的语言理解准确率超过 90%,而其核心程序 Paesey McParseface 句法分析模型只对处理英语文本有效!该项科研成果拥有者说,按照从 Google WebTreebank(谷歌网络树库,2011 年)中所学到的,从互联网上获得的句子要更难分析!语言理解是“人工智能的终极任务”,要解决这一难题,前提是要能够解决全部人类水平人工智能的问题.

中文微博情感分析若要使用 SyntaxNet 开源成果,必须进行针对中文的二次开发.而要研发出类 SyntaxNet 水平的中文自然语言解析软件,还需要我们继续做艰苦、深入、持久的研究.

致谢 衷心感谢匿名审稿专家为完善本文提出的意见和建设性建议.

References:

[1] Li Y, Chen YH, Liu T. Survey on predicting information propagation in microblogs. Ruan Jian Xue Bao/Journal of Software, 2016, 27(2):247−263 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4944.htm [doi: 10.13328/j.cnki.jos.004944]

[2] Luo LJ. Research on personalized recommendation mechanism based on online lifestyle and integrated value [Ph. D. Thesis]. Beijing: Beijing University of Posts and Telecommunications, 2015 (in Chinese with English abstract).

[3] Zhang ZJ. Research on personalized recommendation models and algorithms based on social network [Ph. D. Thesis]. Jinan: Shandong Normal University, 2015 (in Chinese with English abstract).

[4] Cao YZ. Research on user follow and information retweet prediction in enterprise micro-blogging [Ph. D. Thesis]. Chengdu: University of Electronic Science and Technology of China, 2015 (in Chinese with English abstract).

[5] Ma T. Research on co-petition relations within the enterprise microblogging platform and illegal users [Ph. D. Thesis]. Dalian: Dalian University of Technology, 2015 (in Chinese with English abstract).

[6] Sun Y. Research on problems for text sentiment analysis oriented to content security [Ph. D. Thesis]. Wuhan: Naval University of Engineering, 2012 (in Chinese with English abstract).

[7] Xie QH. Local government microblogging in Crisis: Effects of media and social properties-comparison between China and the USA [Ph.D. Thesis]. Hefei: University of Science and Technology of China, 2015 (in Chinese with English abstract).

[8] Vapnik VN. The Nature of Statistical Learning Theory. 2nd ed., New York: Springer-Verlag, 1995. 244−255.

[9] Landauer TK, Foltz PW, Laham D. An Introduction to latent semantic analysis. Discourse Processes, 1998,25(2-3):259−284. [doi: 10.1080/01638539809545028]

[10] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 2001,42(1-2):177−196. [doi: 10.1023/A:1007617005950]

[11] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3(4-5):993−1022.

[12] Griffiths TL, Steyvers M, Tenenbaum JB. Topics in semantic representation. Psychological Review, 2007,114(2):211−244. [doi: 10.1037/0033-295X.114.2.211]

[13] Jo Y, Oh A. Aspect and sentiment unification model for online review analysis. In: Proc. of the 4th ACM Int'l Conf. on Web Search and Data Mining (WSDM 2011). New York: ACM Press, 2011. 815−824. [doi: 10.1145/1935826.1935932]

[14] Lin CH, He YL. Joint sentiment/topic model for sentiment analysis. In: Proc. of the 18th ACM Conf. on Information and Knowledge Management (CIKM 2009). New York: ACM Press, 2009. 375−384. [doi: 10.1145/1645953.1646003]

[15] Mei QZ, Ling X, Wondra M, Su H, Zhai CX. Topic sentiment mixture: Modeling facets and opinions in Weblogs. In: Proc. of the 16th Int'l Conf. on World Wide Web (WWW 2007/Track: Data Mining). New York: ACM Press, 2007. 171−180. [doi: 10.1145/1242572.1242596]

[16] Huang CN, Yuan CF, Pan SM. Corpus, knowledge acquisition and parsing. Journal of Chinese Information Processing, 1992,6(3): 3−8 (in Chinese with English abstract).

[17] Xue NW, Xia F, Zhou FD, Palmer M. The Penn Chinese treebank: Phrase structure annotation of a large corpus. Natural Language Engineering, 2005,11(2):207−238. [doi: 10.1017/S135132490400364X]

[18] Qiu LK, Shi LL, Wang HF. Construction of multi-domain Chinese dependency treebanks and a study on factors influencing the statistical parsing. Journal of Chinese Information Processing, 2015,29(5):69−76 (in Chinese with English abstract).

[19] Che WX, Li ZH, Liu T. Chinese dependency treebank 1.0 LDC2012T05 [DB]. Web Download. Philadelphia: Linguistic Data Consortium, 2012. https://catalog.ldc.upenn.edu/LDC2012T05

[20] Tang XB, Xiao L. Research on micro-blog topics mining model on dependency parsing. Information Science, 2015,33(9):61−66 (in Chinese with English abstract). [doi: 10.13833/j.cnki.is.2015.09.011]

[21] Wu FX, Zhou FG. Unified framework for hybrid dependency parsing. Journal of University of Electronic Science and Technology of China, 2016,45(1):102−107 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-0548.2016.01.017]

[22] Levy O, Goldberg Y. Dependency-Based word embeddings. In: Proc. of the 52th Annual Meeting of the Association for Computational Linguistics. 2014. 302−308.

[23] Liu YB, Ouyang CP, Zhong DL, Li JZ, Yuan BZ, Li Q. Word embedding based on nonlinear global context. Science China: Information Science, 2015,45(12):1588−1599 (in Chinese with English abstract). [doi: 10.1360/N112015-00238]

[24] Liang J, Chai YM, Yuan HB, Zan HY, Liu M. Deep learning for Chinese micro-blog sentiment analysis. Journal of Chinese Information Processing, 2014,28(5):155−161 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2014.05.019]

[25] Mo P, Hu P, Huang XJ, He TT. A hypergraph based approach to collaborative text summarization and keyword extraction. Journal of Chinese Information Processing, 2015,29(6):135−140 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2015.06.018]

[26] Liu M, Zan HY, Yuan HB. Key sentiment sentence prediction using SVM and RNN. Journal of Shandong University (Natural Science), 2014,49(11):68−73 (in Chinese with English abstract). [doi: 10.6040/j.issn.1671-9352.3.2014.025]

[27] Li TT, Ji DH. Sentiment analysis of micro-blog based on SVM and CRF using various combinations of features. Application Research of Computers, 2015,32(4):978−981 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-3695.2015.04.004]

[28] Lafferty J. McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the 18th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001. 282−289.

[29] Zhou XG, Ren YZ, Sun Y. Zhang LQ. Information Content Security. Wuhan: Wuhan University Press, 2012. 168−170 (in Chinese).

[30] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 2002. 79−86.

附中文参考文献:

[1] 李洋,陈毅恒,刘挺.微博信息传播预测研究综述.软件学报,2016,27(2):247−263. http://www.jos.org.cn/1000-9825/4944.htm [doi: 10.13328/j.cnki.jos.004944]

[2] 罗莉娟.基于网络生活方式的综合价值个性化推荐机制研究[博士学位论文].北京:北京邮电大学,2015.

[3] 张志军.社交网络中个性化推荐模型及算法研究[博士学位论文].济南:山东师范大学,2015.

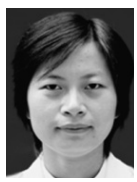
[4] 曹云忠.企业微博用户关注与信息转发预测研究[博士学位论文].成都:电子科技大学,2015.

[5] 马特.微博平台企业与利益相关者的竞合关系研究[博士学位论文].大连:大连理工大学,2015.

[6] 孙艳.面向内容安全的文本情感分析若干问题研究[博士学位论文].武汉:海军工程大学,2012.

[7] 谢起慧.危机中的地方政务微博:媒体属性、社交属性与传播效果——中美比较的视角[博士学位论文].合肥:中国科学技术大学,2015.
[16] 黄昌宁,苑春法,潘诗梅.语料库、知识获取和句法分析.中文信息学报,1992,6(3):3−8.
[18] 邱立坤,史林林,王厚峰.多领域中文依存树库构建与影响统计句法分析因素之分析.中文信息学报,2015,29(5):69−76.
[20] 唐晓波,肖璐.基于依存句法分析的微博主题挖掘模型研究.情报科学,2015,33(9):61−66. [doi: 10.13833/j.cnki.is.2015.09.011]
[21] 吴福祥,周付根.统一框架的混合依存句法分析.电子科技大学学报,2016,45(1):102−107. [doi: 10.3969/j.issn.1001-0548.2016.01.017]
[23] 刘永彬,欧阳纯萍,钟东来,李涓子,袁博志,李奇.基于非线性全局上下文的词嵌入.中国科学:信息科学,2015,45(12):1588−1599. [doi: 10.1360/ N112015-00238]
[24] 梁军,柴玉梅,原慧斌,昝红英,刘铭.基于深度学习的微博情感分析.中文信息学报,2014,28(5):155−161. [doi: 10.3969/j.issn.1003-0077.2014.05.019]
[25] 莫鹏,胡珀,黄湘冀,何婷婷.基于超图的文本摘要与关键词协同抽取研究.中文信息学报,2015,29(6):135−140. [doi: 10.3969/j.issn.1003-0077.2015.06.018]
[26] 刘铭,昝红英,原慧斌.基于 SVM 与 RNN 的文本情感关键句判定与抽取.山东大学学报(理学版),2014,49(11):68−73. [doi: 10.6040/j.issn.1671-9352.3.2014.025]
[27] 李婷婷,姬东鸿.基于 SVM 和 CRF 多特征组合的微博情感分析.计算机应用研究,2015,32(4):978−981. [doi: 10.3969/j.issn.1001-3695.2015.04.004]
[29] 周学广,任延珍,孙艳,张立强.信息内容安全.武汉:武汉大学出版社,2012.168−170.

李勇敢(1973－),男,河南平顶山人,博士生,副教授,主要研究领域为信息内容安全,网络安全.



孙艳(1983－),女,博士,工程师,主要研究领域为信息内容安全,网络安全.

周学广(1966－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为信息内容安全,网络安全,密码学.

张焕国(1945－),男,教授,博士生导师,CCF 高级会员,主要研究领域为信息安全,密码学.