http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0404

基于情感对象识别和情感规则的微博倾向性分析

王泽辰[1], 王树鹏[1,*], 孙立远[2], 张磊[2], 王勇[1], 郝冰川[1]
(1. 中国科学院信息工程研究所, 北京 100193; 2. 国家计算机网络应急技术处理协调中心, 北京 100085)

摘 要: 微博平台数据中含有大量反映用户情感喜恶的信息,对于涉及博文倾向性分析的应用尤为重要。现有的分析方法往往聚焦在博文情感的简单分类上,无法分析特定类型实体的微博倾向性。为解决微博倾向性分析问题,实现博文立场判定,采用半监督学习的方法,通过协同训练和主动学习,训练实体识别模型,并构建基于主成分分析的情感规则,提取句子的主成分,将口语化的文本规范化为指定格式。再利用指向性实体的正负面性、情感词的褒贬义及情感词充当的句子成分,实现情感分类的更深层次分析——立场判定。针对实际问题进行立场判定实验,在不同规模数据集上的自对比实验和他比实验显示,随着标注实体的博文数量增加,模型对博文立场判断的正确率持续提升,而且所提方法判断博文立场的正确率显著高于对比方法,相较已有研究方法分别提高了2.79%和10.00%。

关 键 词: 情感分析; 立场判定; 半监督学习; 倾向性; 情感规则; 协同训练; 主动学习

中图分类号: P391

文献标志码: A **文章编号**: 1001-5965(2022)02-0301-10

微博作为中文社交平台,具有强大的影响力和渗透力。微博用户不断增加,2020 年第 1 季度财报显示,微博日活跃用户达 2.41 亿,月活跃用户达 5.5 亿。这些用户在平台上建立关系、获取信息及生产大量内容。在微博文本信息中,大量博文包含带有明显情感色彩的内容。挖掘这些文本背后用户的观点与倾向,可以判断流行趋势和热点,有助于企业分析消费者购买倾向,进行精准化营销,也使得政府可以对网民的舆论立场变化做出及时反应。因此,对微博文本进行倾向性分析,进而完成对博文立场的判定,是情感分析中一项重要的研究任务。

情感分析以对特定文本进行处理及挖掘其中情感色彩为目的,是自然语言处理的重要分支,近年来受到了广泛关注。提取情感词是分析微博所表达情感最直接的方法。微博文本具有长度较短、内容形式多样、观点倾向性强、表达方式口语化、普遍缺少上下文信息等特点。针对微博数据有许多情感分析的研究,结合情感词典对微博进行分析的方法较为直观。通过对微博中包含的情感词进行分类,可以简单地概括其中蕴含的情绪。但是 1 条微博所表达的信息不仅与其包含的情感词相关,即使 2 条微博包含完全相同的情感词,当情感词所指向的对象不同时,微博的观点与倾向性会存在明显差异。例如,电子产品制造商可能只希望关注用户与手机、电脑等相关的微博情感信息,而不关心对食品、日用品等的购买倾向;在舆情方面,要分析网民在讨论社会热点问题时的

收稿日期: 2020-08-09; **录用日期**: 2020-09-25; **网络出版时间**: 2021-03-09 20:12
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210309.0928.002.html
基金项目: 国家自然科学基金 (61931019)
* **通信作者**. E-mail: wangshupeng@iie.ac.cn

引用格式: *王泽辰, 王树鹏, 孙立远, 等. 基于情感对象识别和情感规则的微博倾向性分析[J]. 北京航空航天大学学报, 2022, 48(2): 301-310. WANG Z C, WANG S P, SUN L Y, et al. Weibo tendency analysis based on sentimental object recognition and sentimental rules[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 301-310 (in Chinese).*

http://bhxb.buaa.edu.cn

立场,不仅需要辨别微博表达的态度是积极还是消极,还需要确定微博支持或反对的实体是什么。另外,中文词汇的多义性使得同一词语在不同语境下指代的实体含义可能不同。随着网络语言的发展,用于实体表示的新词也层出不穷,用户常常用品牌缩写、谐音、流行语等进行指代,这些依赖人工标注的实体往往需要很大的工作量。因此,如何针对海量微博数据进行倾向性分析,从而实现立场判定,是目前亟须解决的问题。

为解决微博倾向性分析的问题,本文将深度学习的方法与情感规则相结合,基于句法规则进行博文情感分析,指定相应的情感规则,结合协同学习和主动学习的方法,可以在仅依赖少量标注数据的前提下准确判断博文对指定类型实体表达的情感。

本文主要有以下 3 点创新:

1)利用半监督学习的方式,通过协同训练加主动学习的方式,同时利用半监督学习结合主动学习的模式,确定指向性实体集。

2)不同于传统的情感规则,本文提出了基于主成分分析的情感规则,通过识别指定方向实体,结合情感词对博文进行倾向性判断。

3)生成指定类别实体集,结合情感规则,判断博文的立场,实现对博文更深层次的分析。

1 相关工作

网络舆情中的情感识别问题得到了学界广泛关注。网络社交平台数据是情感识别任务的重要数据来源。例如,基于 Twitter 平台数据,Giachanou 等[1]建立了情感分析方法,训练、分析推文内容的情感倾向性。类似地,微博作为受众广、数据规模大的中文平台,吸引了许多学者对其数据开展研究。

微博情感分析方法主要有 3 种:

1)基于语义词典的方法。该方法需要先构建微博情感分析数据库,一般包含多个词典和句法规则库,再利用知识库进行聚合计算。例如,王志涛等[2]提出了一种基于词典和规则集的中文微博情感分词方法,将词典和语义规则结合进行情感分析。王灿伟[3]提出了结合情感词典和感情符号来计算微博情感值的方法,实现主题归类和情感分类。语义词典通过构造情感词典,筛选出微博文本中的情感词,计算文本片段的情感权值,从而判断微博的情感倾向。Ebrahimi 等[4]提出了将情感极性融入到情感对象和情感立场中的方法,通过对数线性联合建模实现立场判定。

2)传统机器学习方法。该方法普遍应用于情感分类,其核心思想是构建特征向量,找到特征与分类结果之间的关联。朴素贝叶斯(naive Bayes)、最大熵(maximum entropy,ME)、支持向量机(support vector machine,SVM)等方法常用于情感分类[5]。Pang 等[6]使用了电影评论作为数据,采用以上 3 种机器学习方法进行情感分类。奠雨洁等[7]使用 SVM、随机森林和梯度提升决策树对文本特征进行立场检测,并利用特征分类器进行立场融合。这些方法具有训练时间短及特征维度高的特点,但不能充分利用上下文信息和语法信息,且需要大量手工标注。

3)基于深度学习的方法。该方法对微博博文进行分词,将其表示为词向量,通过深度神经网络提取语义信息,构建情感表征向量,并利用微博情感表征完成情感分类任务。深度学习模型在情感分析上取得了很好的效果[8],特别是基于长短期记忆网络(long and short term memory,LSTM)方法,在长语句分析上取得了很好的结果[9]。除此之外,还有基于双重注意力模型[10]、监督学习[11]、集成学习[12]的方法。

通过对上述已有工作进行总结和分析,可以将基于语义词典、传统机器学习和深度学习的微博情感分析方法的特点进行对比,结果如表 1 所示。由表 1 可知,深度学习方法能够充分利用文本信息,往往具有更好的效果。基于深度学习的情感分析又可细分为无监督学习、有监督学习和半监督学习 3 种方法。

在无监督学习方法中,所有数据都缺少情感标注。Turney[13]提出了情感倾向性分类的方法,通过选取指定词汇,分别计算每个词与情感词典中积极、消极情感词的交叉熵,以 ME 值作为情感词分类判断标准。除此之外,还有基于语法分析[14]、句法模式[15]等方法的研究。上述方法可以发掘文本数据中内在的词汇情感规律,节约人工标注的成本。

表 1 微博情感分析方法的特点

Table 1 Features of Weibo sentiment analysis methods

情感分析方法	特点
语义词典	包含多个词典和句法规则库;利用知识库进行聚合计算;需要构建微博情感分析数据库
传统机器学习	构建特征向量;找到特征与分类结果间的关联;需要大量手工标注
深度学习	分词并将词语表示为词向量;深度神经网络提取语义信息;构建情感表征向量;需要较长训练时间

http://bhxb.buaa.edu.cn

有监督学习情感分类问题主要关注特征选择和分类器设计。例如,文献[16]中采用了词性特征,Cambria 等[17]引入了符号特征和词嵌入特征。基于商品评论中产品特征信息的情感分类采用有监督学习方法[18],具有较高的运行效率。但是,如果在有监督学习中出现分类错误,那么之后的学习都会受到这个错误的影响。为了解决这个问题,有学者提出了半监督学习方法避免准确率的下降。

Sindhwani[19]和 Liu[20]等分别提出了采用半监督学习方法进行情感分析。半监督学习包括自训练、协同训练等方法。自训练[21]方法是较早提出的一种半监督学习方法,先训练初始分类器,再使用该分类器对未标注的数据进行标注,选取分类准确度较高的样本确认标注,直到经过多轮迭代后所有样本都完成标注为止。半监督学习方法仅需要标记少量数据,适用于数据量大的任务。

对基于深度学习的方法进行总结和分析,将基于无监督学习、有监督学习和半监督学习的深度学习情感分析方法的特点进行归纳,结果如表 2所示。

由表 2 可知,采用半监督学习方法,在情感分析任务上可以取得较好的效果。基于情感分析研究现状,本文提出了基于情感对象识别与情感规则的倾向性分析(orientation analysis based on sentiment object recognition and sentiment rules, OASOSR)算法,不仅能够确定博文中的情感词,还可以分析情感词指向的实体是否为目标实体。

表 2　基于深度学习的情感分析方法特点对比

Table 2　Comparison of features of sentiment analysis methods based on deep learning

情感分析方法	特点
无监督学习	发掘文本数据中内在的词汇情感规律;无需人工标注
有监督学习	运行效率较高;出现分类错误时后续训练会受到影响
半监督学习	仅需要标记少量数据,适用于数据量大的任务

2　OASOSR 算法

OASOSR 算法框架如图 1 所示,主要分为 2 个部分:①基于协同训练和主动学习的实体识别模型;②基于情感词典和句法分析的微博倾向性分析。

基于协同训练和主动学习的实体识别模型部分采用半监督学习方法,将少量已标注的微博文本数据集作为初始输入,协同训练 2 个不同的实体识别模型。为了在协同训练中比较模型训练效果,基于这 2 个实体识别模型构建 2 个分类器,分别包括实体集、情感词典和情感规则 3 个部分。在协同训练的过程中,首先,从未进行实体标注和立场判定的数据集中抽取一定数量的数据,并利用这 2 个训练过的实体识别模型对未标注语料进行实体识别。同时,2 个实体识别模型通过判断微博文本中的实体类型,提取指定方向的实体,各形成 1 个实体集。然后,在这 2 个实体集的基础上,构造 2 个分类器,利用分类器分别结合情感词典和情感规则来判断微博立场,并对同一条博文进行倾向性分析。比较 2 个分类器得到的倾向性分析结果,判断样本置信度,从中选择高置信度(即 2 个分类器输出结果完全相同)的样本,并为此类样本添加指定类别的实体标记和倾向性标签,合并到已标记博文数据中。同时,对于低置信度的样本,采用主动学习模式,将分类器挑选出的分歧大(2 个分类器标注结果不同)的样本添加到已标注的数据集中。将更新扩充后的有标注数据集重新输入到协同训练模型,再次训练上述 2 个实体识别深度学习模型,不断迭代直到已标记的博文数据集达到足够的规模,进而获取最大的指定方向实体集。2.1 节将详细介绍微博实体识别模型及指定方向实体集的训练过程。

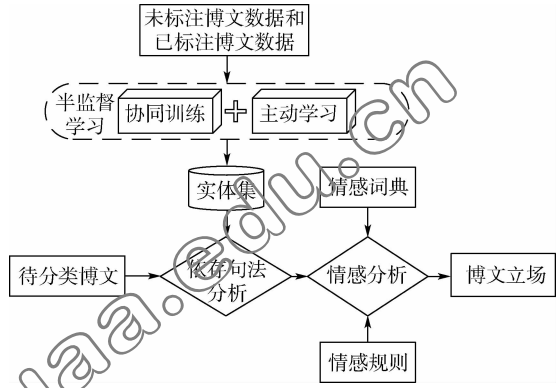

图 1　OASOSR 算法总体架构

Fig. 1　Algorithm architecture of OASOSR

在基于情感词典和句法分析的微博倾向性分析部分,首先,模型基于学习得到指定方向实体集。然后,针对需要立场判断的微博数据,判断微博是否与指定方向的实体相关。如果微博文本中包含指定方向实体集内的任意实体,则对此类微博根据预先编写的情感规则和情感词典进行倾向性分析,实现对微博文本的立场判定。2.2 节将重点讨论 OASOSR 算法中基于情感词典和情感规则的倾向性分析。

http://bhxb.buaa.edu.cn

2.1 基于协同训练和主动学习的实体识别模型

本节详细介绍基于协同训练和主动学习模型的构造细节、训练过程及微博情感对象实体集的提取流程。

协同训练假设数据拥有 2 个充分且条件独立的视图,即每个视图所包含的信息都可以支持生成最优学习器,且在给定类别标记的条件下 2 个视图相互独立。利用未标记数据在每个视图上都训练生成 1 个分类器,每个分类器选择置信度最高的样本生成伪标记,再将这些伪标记作为标记样本训练另 1 个分类器。本文对层叠隐马尔可夫模型(cascaded hidden Markov model,CHMM)和条件随机场(conditional random field,CRF)实体识别学习模型进行协同训练。

隐马尔可夫模型(hidden Markov model,HMM)是一种在自然语言处理领域中被广泛应用的统计模型。中文命名实体识别中的人名识别、地名识别、译名识别及机构名识别等都可以用 HMM 来解决。本文利用 CHMM 在统一的 HMM 中识别各类实体命名,自底向上分为人名识别 HMM、地名识别 HMM 和机构名识别 HMM 3 层。每一层 HMM 将产生的最好的若干个结果送到词图中供高层模型使用,最终实现人名、地名和机构名的实体识别。

CRF 模型是一种基于统计的判别式模型,结合了 ME 模型和 HMM 模型的特点。CRF 模型根据样本的特征生成预测,对所有的特征权重进行最优化,进而得到最优解,解决了 HMM 模型因独立性假设而导致其不能考虑上下文特征的问题。在本文模型中,从经过分词的数据中挑选出包含在特定微博主题标签内的数据实体,形成候选词集。再利用 CRF 模型对候选词集中的候选实体进行实体抽取。

本文采用协同训练和主动学习的方法训练以上 2 个模型。主动学习方法具有在训练集较少的情况下能够获得较高分类准确率的优势,是对有监督学习方法的一种改进。有监督学习方法在学习过程中被动接受人工标注的样本集,通过将标注样本映射到目标函数空间来学习知识。而主动学习方法则试图从未标注的池中选择数量尽可能少的、机器无法正确标注的样本(低置信度的样本),并由领域专家进行标注后添加到下一次迭代过程中作为训练集,以提高分类的效率。主动学习通过选择质量高的少数样本进行学习以保证分类器的分类性能,同时,也可以减轻样本的标注复杂度。

情感对象实体集提取流程如图 2 所示,采用了半监督学习方法,通过协同训练和主动学习的方式对 2 个实体识别学习模型进行训练。在实体及情感类别已标注数据集中,有少量已标注微博文本分别标注了博文中包含的实体(实体 1、实体 2 等)和博文情感类别(正向、负向)。将有标注的微博文本进行切词处理,分别输入到 CHMM 和 CRF 实体识别学习模型,初步训练这 2 个模型。

在对 CHMM 和 CRF 实体识别学习模型进行协同训练的过程中,先从未标注实体及情感类别

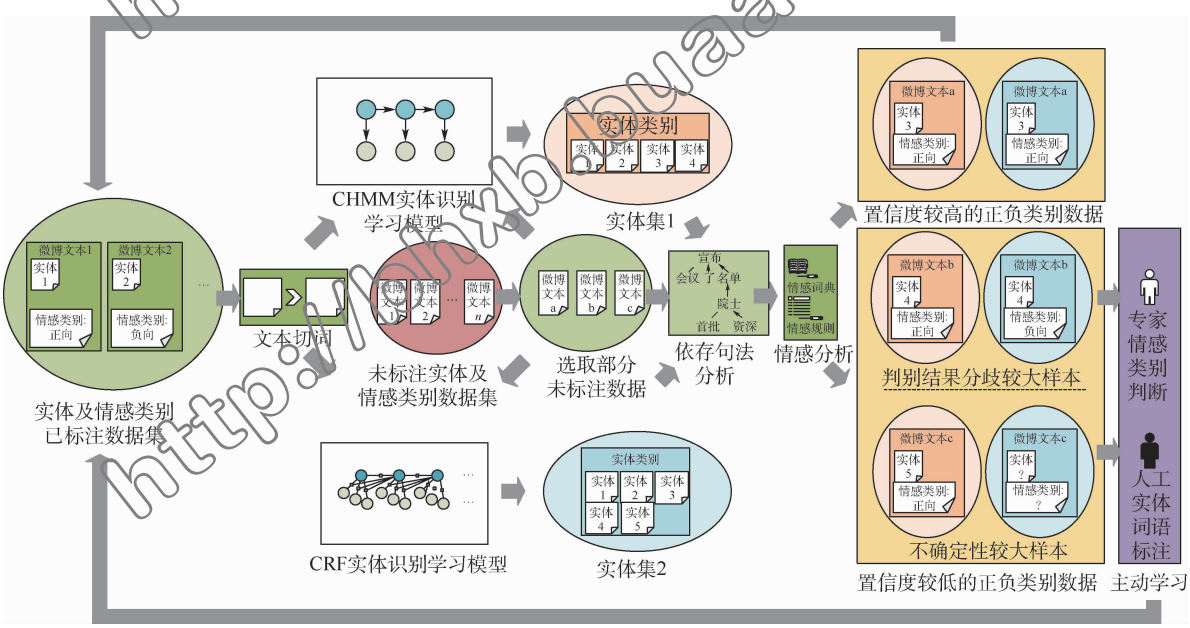

图 2 情感对象实体集提取流程

Fig. 2 Extraction flowchart of sentimental object entity sets

数据集中选取部分数据(微博文本 a、b、c),输入到经过初步训练的 CHMM 和 CRF 实体识别学习模型,提取博文中指定方向的实体,分别得到实体集 1 和实体集 2。分别基于 2 个实体集对相同的未标注数据进行依存句法分析,再基于给定的情感词典和情感规则进行情感分析,判断微博文本的情感类别。对于每条微博,都有基于实体集 1 和实体集 2 的 2 个标记结果,每个标记结果都包括微博所含有的实体(实体 n)和倾向(正向、负向、无法判断)。比较基于实体集 1 和实体集 2 得到的标记结果,针对同一微博文本,若 2 个分类器得到的结果中所标记实体和情感类别都完全相同,则判断该结果为置信度较高的正负类别数据。直接将这条微博文本和对应实体、情感类型标注添加到实体及情感类别已标注数据集;若 2 个结果中所标记实体或情感类别不完全相同,则判断该结果为置信度较低的正负类别数据,需要采用主动学习的方法处理。对于这类数据,针对同一微博文本,若 2 个分类器提取的指定方向实体完全相同,但情感类别相反,则认为该样本是分歧较大样本,需要将结果交由领域专家进行情感类别判断。若 2 个分类器提取的指定方向实体不同,或有 1 个分类器没有提取出指定方向实体,则认为该样本是不确定性较大样本,需要进行人工实体词语标注。同样地,将经由主动学习得到的微博文本和对应实体、情感类型标注添加到实体及情感类别已标注数据集。这样就完成了协同学习的第一轮循环。

对这一过程进行迭代,不断扩充实体及情感类别已标注数据集,直到已标记数据的数据量达到设定的停止阈值。获取该数据集中的所有标注实体所组成的实体集,作为 2.2 节中微博倾向性分析的基础。

2.2　基于情感词典和情感规则的微博倾向性分析

本节介绍基于情感词典和情感规则进行句法分析的微博立场判断算法。该算法通过情感规则识别微博中情感词所充当的句子成分,分析情感词与目标实体之间的修饰关系,从而判断微博的倾向性。然而,情感规则的制定依赖于依存句法分析。汉语语句以词语为基本单位,在词语之间拥有依存和支配关系。依存句法分析理论能够借用各级语言单位的依存关系,清晰地提炼出成分之间的修饰和搭配信息,从而达到解析语句的目的,进而总结句子成分之间的关联,制定情感规则。

基于情感词典和情感规则的微博倾向性分析算法包含文档预处理、条件判断、实体信息判断、句法分析及规则判断 5 个步骤,输出结果为对应输入博文的博文倾向性类别。算法流程如图 3 所示。

首先,对原始博文进行数据清洗,经过简化繁体字、删除无效评论等操作后,得到有效数据,再使用结巴分词(数据来源:https://github.com/fxsjy/jieba),并去除分词结果中的停用词。然后,基

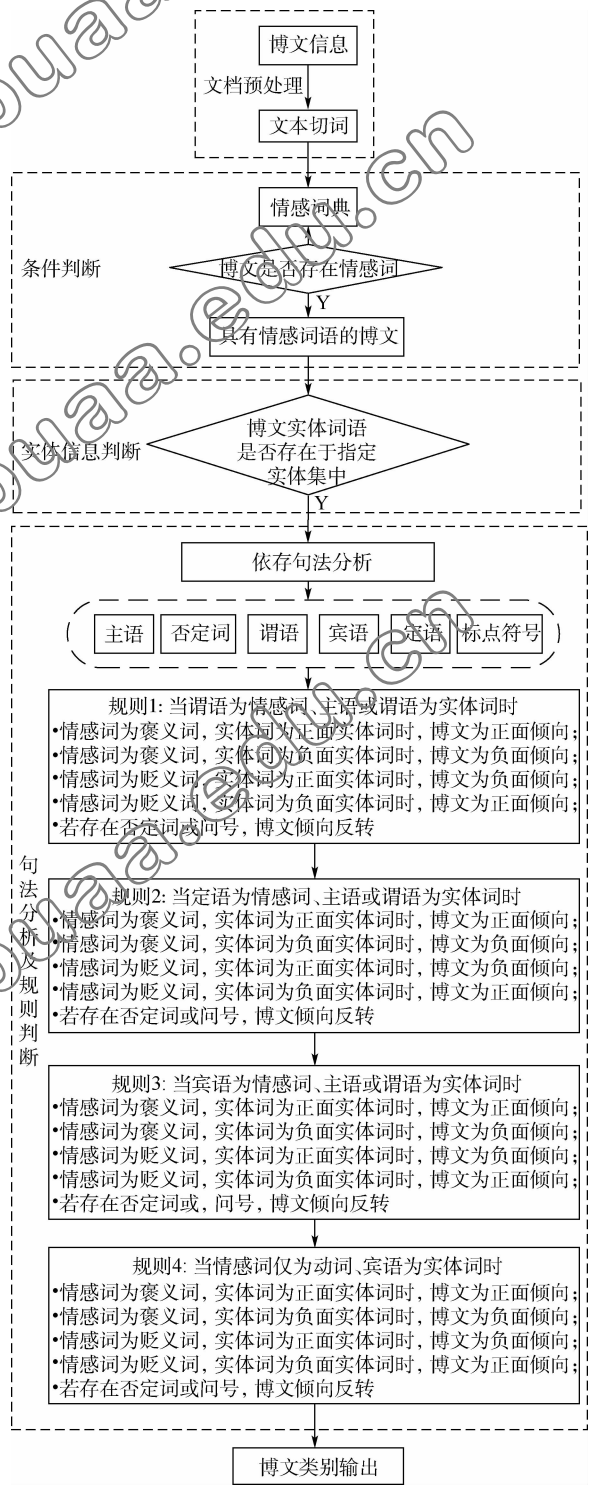

图 3　OASOSR 算法流程

Fig. 3　OASOSR algorithm flowchart

http://bhxb.buaa.edu.cn

于情感词典,对经过处理的微博文本数据进行情感词判断。显然,只有包含情感词的微博文本才能表达倾向性,不包含情感信息的微博不具有立场分析的价值。因此,本文采用大连理工大学情感词汇本体库 DUTIR 情感词典(数据来源:http://ir.dlut.edu.cn)判断博文中是否存在情感词,筛选出含有情感词的微博文本,进行下一步处理。在含有情感词的微博文本基础上,进行实体信息判断,筛选出包含指定方向实体的微博。实体信息判断基于2.1节中通过协同训练和主动学习得到的实体集完成。最后,提取微博文本中实体词语,若实体存在于该指定方向的实体集之中,则筛选出这条微博,对其中的句子进行依存句法分析。通过识别句子的主语、否定词、谓语、宾语、定语、标点符号等成分,基于句法分析、情感规则将句子分为4种不同的类型:①当谓语为情感词、主语或谓语为实体词时;②当定语为情感词、主语或谓语为实体词时;③当宾语为情感词、主语或谓语为实体词时;④当情感词仅为动词、宾语为实体词时。

情感词在句子中充当不同成分时,具体分类规则如 OASOSR 算法所示(见图3)。经过对应的情感规则判断,最终输出该微博文本的立场。

算法1 OASOSR 算法。

输入:微博文本。

输出:博文类别。

步骤1 文档预处理。文本切词。

步骤2 条件判断。依据情感词典判断博文是否存在情感词,若存在情感词,则对句子进行下一步处理。

步骤3 实体信息判断。提取博文中实体词语,若实体存在于指定实体集之中,则进行下一步处理。

步骤4 句法分析及规则判断。

步骤4.1 对句子进行依存句法分析,识别句子的主语、否定词、谓语、宾语、定语、标点符号等成分。

步骤4.2 设定情感判断规则。

1)规则1:当谓语为情感词、主语或谓语为实体词时:

① 情感词为褒义词,实体词为正面实体词时,博文为正面倾向。

② 情感词为褒义词,实体词为负面实体词时,博文为负面倾向。

③ 情感词为贬义词,实体词为正面实体词时,博文为负面倾向。

④ 情感词为贬义词,实体词为负面实体词时,博文为正面倾向。

⑤ 若存在否定词或问号,博文倾向反转。

2)规则2:当定语为情感词、主语或谓语为实体词时:

① 情感词为褒义词,实体词为正面实体词时,博文为正面倾向。

② 情感词为褒义词,实体词为负面实体词时,博文为负面倾向。

③ 情感词为贬义词,实体词为正面实体词时,博文为负面倾向。

④ 情感词为贬义词,实体词为负面实体词时,博文为正面倾向。

⑤ 若存在否定词或问号,博文倾向反转。

3)规则3:当宾语为情感词、主语或谓语为实体词时:

① 情感词为褒义词,实体词为正面实体词时,博文为正面倾向。

② 情感词为褒义词,实体词为负面实体词时,博文为负面倾向。

③ 情感词为贬义词,实体词为正面实体词时,博文为负面倾向。

④ 情感词为贬义词,实体词为负面实体词时,博文为正面倾向。

⑤ 若存在否定词或问号,博文倾向反转。

4)规则4:当情感词仅为动词、宾语为实体词时:

① 情感词为褒义词,实体词为正面实体词时,博文为正面倾向。

② 情感词为褒义词,实体词为负面实体词时,博文为负面倾向。

③ 情感词为贬义词,实体词为正面实体词时,博文为负面倾向。

④ 情感词为贬义词,实体词为负面实体词时,博文为正面倾向。

⑤ 若存在否定词或问号,博文倾向反转。

步骤4.3 输出博文的立场。

在 OASOSR 算法中,文本切词的时间复杂度为 $O(n^2)$,条件判断、实体信息判断的时间复杂度为 $O(n)$,而句法分析的时间复杂度为 $O(Kn^3)$,可以得知 OASOSR 算法时间复杂度为 $O(Kn^3)$。

3 实 验

3.1 实验设置

为证明 OASOSR 算法的实用性,本文实现了 OASOSR 算法的自对比实验和他比实验。本文通过网络爬虫获取了38 175条新冠肺炎疫情舆论

http://bhxb.buaa.edu.cn

微博数据(数据来源:https://github.com/ZuQin131/Weibo-public-opinion-analysis),并对新冠肺炎疫情舆情言论微博数据进行人工实体标注,分为正面和负面情感博文,用于自对比实验;并选用 weibo_senti_100k(数据来源:https://github.com/SophonPlus/ChineseNlpCorpus/tree/master/datasets)的开源微博情感数据集,从中选取部分数据进行实体标注和人工正面、负面情感标注,应用于自对比实验和他比实验。上述数据集的特点总结如表 3 所示。

表 3　对比实验中选取的数据集

Table 3　Datasets selected for comparative experiment

数据类型	数据集	大小/条	数据格式
爬取数据	#新冠肺炎疫情#话题	38 175	标签(正向、负向),文本
公开数据	weibo_senti_100k 开源微博情感	119 989	标签(正向、负向),文本

实验中构造的微博情感词典是在大连理工大学情感词汇本体库 DUTIR 的基础上,添加与新冠肺炎疫情舆论相关的情感词得到的。最终构建的情感词典共有 23 651 个中文情感词,包括 9 833 个正面情感词和 13 818 个负面情感词。

为了测试在数据量变化情况下 OASOSR 算法的稳定性、鲁棒性与准确率变化趋势,本节设计了基于 OASOSR 算法的自对比实验。如表 4 所示,从新冠肺炎疫情话题数据集中随机选取 485 条新冠肺炎疫情舆情言论微博数据,进行人工实体标注,并将博文分为正面情感博文和负面情感博文,其中已标注正面情感博文 331 条,负面情感博文 154 条,作为小规模数据集。

同时,也在 weibo_senti_100k 的开源微博情感数据集上使用该方法进行实验,这是一个开源的大规模数据集,共有 119 989 条微博情感数据,从中随机选取 4 000 条微博数据进行数据标注,并作为大规模数据集,其中标注了正面情感博文 2 200 条,负面情感博文 1 800 条。

为证明 OASOSR 算法在不同规模数据集上的有效性,设计了 4 组实验,在不同规模下对 OASOSR 算法的鲁棒性进行了测试。首先,在小规模数据集(新冠肺炎疫情话题数据集)上进行实验,从新冠肺炎疫情话题数据集中随机选取 200 条作为训练集,剩余的 285 条博文作为测试集。然后,在大规模数据集上进行实验,从 weibo_senti_100k 开源微博情感数据集中选取了 2 000 条作为训练集,剩余 2 000 条作为测试集进行实验。在 2 个数据集上分别进行 4 组实验,每组实验分别从训练集中选取 25%、50%、75% 和 100% 的数据用于训练,生成训练结果,并对比观察 2 个模型在测试集上情感分类的正确率,正确率对比如图 4 所示。

表 4　对比实验中选取的数据集大小

Table 4　Size of datasets selected for comparative experiment

数据集	实验数据条数
#新冠肺炎疫情#话题	485
weibo_senti_100k 开源微博情感	4 000

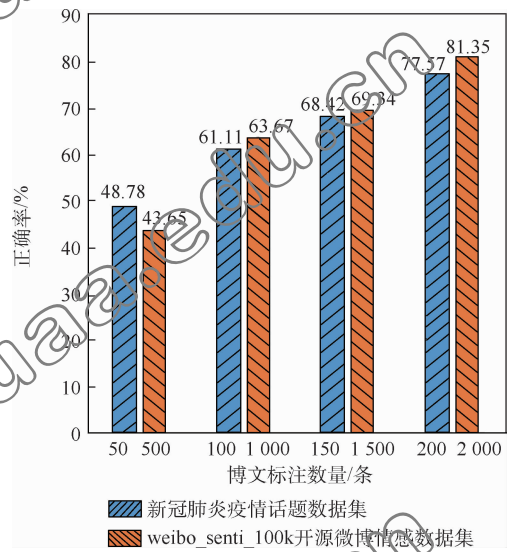

图 4　不同规模数据集上 OASOSR 算法立场判断正确率

Fig. 4　Accuracy of standpoint judgement by OASOSR algorithm on different datasets

在小规模数据集上,当实体标注博文的数量为 50 条时,标注数据过少,识别出的情感对象实体较少,因此博文情感识别率较低。当实体标注博文的数量为 100,150,200 条时,正面和负面情感博文的判断正确率持续提高。相似地,在大规模数据集上,当实体标注博文的数量为 500 条时,正面和负面情感博文的判断正确率都比较低,而当情感标注博文的数量为1 000,1 500,2 000 条时,正面和负面情感博文的判断正确率持续提高。

可以看出,随着训练数据中标注实体的博文数量增加,在不同规模的数据集上,算法对微博立场的判断正确率都在逐渐提高。OASOSR 算法在数据集规模变化差别较大的情况下,仍具有很好的鲁棒性。

3.2　他比实验

他比实验选取 weibo_senti_100k 开源微博情感数据集作为实验语料。为了验证 OASOSR 算法的有效性,本文选取了基于 SVM 的有监督情感文本分类方法(sentiment classification of texts

http://bhxb.buaa.edu.cn

based on SVM, SCSVM)[22]和基于极性词典情感分析的无监督情感分类方法(sentiment analysis method based on a polarity lexicon, SAMPL)[23],分别与本文算法进行对比。在 SCSVM 方法中,先构建情感词典,并进行情感特征选取及情感特征加权,再使用 SVM 分类的方法对文本进行情感识别及分类。SAMPL 方法则综合了基础词典、领域词典、网络词典及修饰词词典,将极性词和修饰词组合成极性短语,以极性短语为极性计算的基本单元,进行情感分析。

表 5 展示了实验中基于上述 2 个模型的微博立场判断正确率和基于 OASOSR 算法的微博立场判断正确率。对比表 5 中正确率可知,在同等数量的训练集和测试集中,OASOSR 算法要优于 SCSVM 方法和 SAMPL 方法。这是因为:本文提出的 OASOSR 算法充分考虑了对情感对象的识别,微博情感识别更有针对性,在立场判断任务上的正确率相对较高。由对比实验结果可以证明,相较于有监督学习方法和无监督学习方法,半监督学习方法更适用于指定实体的博文立场判断。

表 5 基于不同模型的微博立场判断正确率

Table 5 Accuracy of Weibo standpoint judgement based on different models

方法	正确率/%
SCSVM	78.56
SAMPL	71.35
OASOSR	81.35

3.3 规则分析实验

为证明 OASOSR 算法中不同情感规则在微博文本倾向性分析中会起到作用,本文设计了 2 组规则分析实验。在第 1 组实验中,设置变量为仅保留情感规则 1 ~ 4 条中的 1 条,去除其他 3 条规则;在第 2 组实验中,设置变量为去除情感判断规则 1 ~ 4 条中的 1 条,按原有顺序保留其余 3 条规则。在新冠肺炎疫情话题数据集和 weibo_senti_100k 开源微博情感数据集上进行这 2 组实验,生成实验结果,并将结果与去除判断规则前的实验结果进行对比。图 5 中,“单一”代表仅保留某1 条规则,去除其他 3 条情感规则得到的实验正确率;“缺失”表示缺失某 1 条规则,但保留其他3 条规则的得到的实验正确率。从实验结果对比可以看出,在去除某 1 条情感规则后,立场判定的正确率均低于原有算法模型。特别是去除规则 1后,正确率下降了 34.36%。因此,OASOSR 算法中 4 条规则互相支撑且都具有必要性,缺一不可。

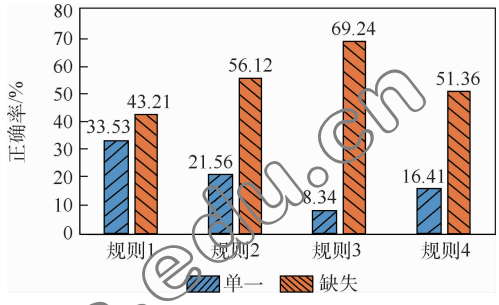

图 5 基于不同模型筛选条件的微博立场判断正确率

Fig. 5 Accuracy of Weibo standpoint judgement based on filtering conditions of different models

4 结 论

微博情感的分类较为简单,无法判断微博对特定方向实体的立场问题。针对微博文本数据信息量大、情感倾向性强的特点,本文提出了使用 OASOSR 算法解决微博立场判定问题。该算法结合深度学习与情感规则,实现微博立场判定。首先,本文采用半监督学习的方法,分别基于 CHMM 和 CRF 模型,对微博文本进行协同训练,再通过主动学习的方法,提取微博文本中的指向性实体集,并构造了 2 个分类器。然后,通过分析大量微博文本中的句法特征,构建情感规则,提取句子的主成分,去除与立场判断不相关的信息,将表达方式较为随意的微博文本规范化。基于经过处理的博文数据,综合判断指向性实体的正面和负面性,结合博文句子中情感词的褒贬义、分析情感词充当的句子成分及情感词修饰的实体类型,进而分析博文的倾向性,确定微博对指定主题所表达的立场。为证明 OASOSR 算法的实用性,本文实现了自对比实验和他比实验。自对比实验显示,随着标注实体的博文数量增加,模型对博文立场判断的正确率持续提升,证明了该算法可以有效增强立场判断效果。将 OASOSR 算法与 SCSVM 方法、SAMPL 方法进行对比实验,在不同标注数量博文上进行对比的实验结果显示,OASOSR 算法判断博文立场的正确率显著高于对比方法。

参考文献(References)

[1] GIACHANOU A, MELE I, CRESTANI F. Explaining sentiment spikes in twitter[C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York: ACM, 2016: 2263-2268.

[2] 王志涛,於志文,郭斌,等. 基于词典和规则集的中文微博情感分析[J]. 计算机工程与应用,2015,51(8):218-225.
WANG Z T, YU Z W, GUO B, et al. Sentiment analysis of Chinese micro blog based on lexicon and rule set[J]. Computer Engineering and Applications, 2015, 51(8): 218-225 (in Chi-

http://bhxb.buaa.edu.cn

nese).

[3] 王灿伟. 基于主题提取的海量微博情感分析[J]. 南京大学学报(自然科学),2017,53(3):549-556.
WANG C W. Sentimental analysis of massive micro-blog based on topic extraction[J]. Journal of Nanjing University (Natural Sciences),2017,53(3):549-556(in Chinese).

[4] EBRAHIMI J, DOU D J, LOWD D. A joint sentiment-target-stance model for stance classification in tweets[C]// Proceedings of the 26th International Conference on Computational Linguistics,2016:2656-2665.

[5] PAK A, PAROUBEK P. Twitter as a corpus for sentiment analysis and opinion mining[C]// Proceedings of International Conference on Language Resource and Evaluation,2010:13-20.

[6] PANG B, LEE L, VAITHYANATHAN S, et al. Thumbs up?: Sentiment classification using machine learning techniques[C]// Proceedings of the ACL-02 Conference on Empirical Methods on Natural Language Processing. New York: ACM,2002:79-86.

[7] 奠雨洁,金琴,吴慧敏. 基于多文本特征融合的中文微博的立场检测[J]. 计算机工程与应用,2017,53(21):77-84.
DIAN Y J, JIN Q, WU H M. Stance detection in Chinese microblogs via fusing multiple text features[J]. Computer Engineering and Applications,2017,53(21):77-84(in Chinese).

[8] 李俭兵,刘梁材. 基于改进型神经网络的影评文本情感分析算法[J]. 计算机工程与科学,2019,41(12):2261-2269.
LI J B, LIU S C. A film criticism sentiment analysis algorithm based on improved neural network[J]. Computer Engineering and Science,2019,41(12):2261-2269(in Chinese).

[9] LI D, QIAN J. Text sentiment analysis based on long and short term memory[C]// 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI). Piscataway: IEEE Press,2016:471-475.

[10] 张仰森,郑佳,黄改娟,等. 基于双重注意力模型的微博情感分析方法[J]. 清华大学学报(自然科学版),2018,58(2):122-130.
ZHANG Y S, ZHENG J, HUANG G J, et al. Microblog sentiment analysis method based on a double attention model[J]. Journal of Tsinghua University (Science and Technology),2018,58(2):122-130(in Chinese).

[11] 朱晓光,聂培尧,林培光. 基于监督学习的微博情感分类方法[J]. 计算机应用与软件,2015,32(8):238-242.
ZHU X G, NIE P Y, LIN P G. Supervised learning based on microblogging sentiment classification method[J]. Computer Applications and Software,2015,32(8):238-242(in Chinese).

[12] 段吉东,刘双荣,马坤,等. 基于集成学习的文本情感分类方法[J]. 济南大学学报(自然科学版),2019,33(6):483-488.
DUAN J D, LIU S R, MA K, et al. Text sentiment classification method based on ensemble learning[J]. Journal of University of Jinan(Science and Technology),2019,33(6):483-488(in Chinese).

[13] TURNEY P D. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,2002:417-424.

[14] BLOOM K, ARGAMON S. Automated learning of appraisal extraction patterns[J]. Language and Computers,2010,71(2):249-260.

[15] GUO J L, PENG J E, WANG H C. An opinion feature extraction approach based on a multidimensional sentence analysis model[J]. Cybernetics and Systems,2013,44(5):379-401.

[16] AGRAWAL A, XIE B, VOVSHA I, et al. Sentiment analysis of Twitter data[J]. International Journal of Computer Applications,2013,139(11):880-887.

[17] CAMBRIA E, POIRIA S, HAZARIKA D, et al. Senticnet5: Discovering conceptual primitives for sentiment analysis by means of context embeddings[C]// 32nd AAAI Conference on Artificial Intelligence,2018:1795-1802.

[18] BANDAPAT S. Handbook of natural language processing(second edition)[J]. Machine Translation,2011,25(4):377-381.

[19] SINDHWANI V, MELVILLE P. Document-word co-regularization for semi-supervised sentiment analysis[C]// 18th IEEE International Conference on Data Mining. Piscataway: IEEE Press,2008:1025-1030.

[20] LIU Z, DONG X, GUAN Y, et al. Reserved self-training: A semi-supervised sentiment classification method for Chinese micro-blogs[C]// Proceedings of IJCNLP,2013:455-462.

[21] SCUDDER H. Probability of error of some adaptive pattern-recognition machines[J]. IEEE Transactions on Information Theory,1965,11(3):363-371.

[22] 陈培文,傅秀芬. 采用 SVM 方法的文本情感极性分类研究[J]. 广东工业大学学报,2014,31(3):95-101.
CHEN P W, FU X F. Research on sentiment classification of texts based on SVM[J]. Journal of Guangdong University of Technology,2014,31(3):95-101(in Chinese).

[23] 张成功,刘培玉,朱振方,等. 一种基于极性词典的情感分析方法[J]. 山东大学学报,2012,47(3):47-50.
ZHANG C G, LIU P Y, ZHU Z F, et al. A sentiment analysis method based on a polarity lexicon[J]. Journal of Shandong University,2012,47(3):47-50(in Chinese).

http://bhxb.buaa.edu.cn

Weibo tendency analysis based on sentimental object recognition and sentimental rules

WANG Zechen[1], WANG Shupeng[1,*], SUN Liyuan[2], ZHANG Lei[1], WANG Yong[1], HAO Bingchuan[1]

(1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100193, China;
2. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100085, China)

Abstract: Weibo contains a large number of information reflecting users' likes and dislikes, which is important for popular trend judgment, precision marketing, public opinion monitoring, etc. However, the existing methods tend to focus on the classification of Weibo sentiment. In order to solve the problem of Weibo tendentiousness analysis and position detection, we employ semisupervised learning method, through collaborative training and active learning. We train entity recognition models and combine deep learning with emotional rules. Moreover, the sentiment rules based on principal component analysis are constructed to extract the main components of sentences, normalize the spoken text into the specified format. Then we use the positive and negative aspects of directional entities, the positive and negative meanings of emotional words, and the sentence components of emotional words to judge the tendency of blog posts, and conduct deeper analysis on position classification. Finally, the self comparison experiment and other comparison experiment on different scale data sets show that with the increase of the number of blog posts of labeled entities, the accuracy of the model continues to improve, and the accuracy of this method is significantly higher than the comparison method, which is 2.79% and 10.00% higher than the existing research methods.

Keywords: sentiment analysis; position detection; semi-supervised learning; tendency; sentimental rule; co-training; active learning

http://bhxb.buaa.edu.cn

Received: 2020-08-09; **Accepted**: 2020-09-25; **Published online**: 2021-03-09 20:12
URL: kns.cnki.net/kcms/detail/11.2625.V.20210309.0928.002.html
Foundation item: National Natural Science Foundation of China (61931019)
* **Corresponding author**. E-mail: wangshupeng@iie.ac.cn