

自然语言处理在文本情感分析领域应用综述

王颖洁¹, 朱久祺¹, 汪祖民¹, 白凤波^{2,3*}, 弓箭³

(1. 大连大学 信息工程学院, 辽宁 大连 116622; 2. 中国政法大学 证据科学研究院, 北京 100088;

3. 中科金审(北京)科技有限公司 自然语言处理部, 北京 100088)

(* 通信作者电子邮箱 baifb@cupl.edu.cn)

摘要: 文本情感分析已经逐渐成为自然语言处理(NLP)的重要内容,并在系统推荐、用户情感信息获取,为政府、企业提供舆情参考等领域越来越占据重要地位。通过文献调研的方式,对情感分析领域的方法进行对比和综述。首先,从时间、方法等维度对情感分析的方法进行文献调研;然后,对情感分析的主要方法、应用场景进行归纳总结和对比;最后,在此基础上分析每种方法的优缺点。根据分析结果可以知道,在面对不同的任务场景,主要有三种情感分析的方法:基于情感字典的情感分析法、基于机器学习的情感分析法和基于深度学习的情感分析法,基于多策略混合的方法成为改进的趋势。文献调研表明,文本情感分析的技术方法还有改进的空间,在电子商务、心理治疗、舆情监控方面有较大市场和发展前景。

关键词: 自然语言处理;情感分析;情感字典;机器学习;深度学习

中图分类号: TP183 **文献标志码:** A

Review of applications of natural language processing in text sentiment analysis

WANG Yingjie¹, ZHU Jiuqi¹, WANG Zumin¹, BAI Fengbo^{2,3*}, GONG Jian³

(1. College of Information Engineering, Dalian University, Dalian Liaoning 116622, China;

2. Institute of Evidence Law and Forensic Science, China University of Political Science and Law, Beijing 100088, China;

3. Natural Language Processing Technology Department, Sino FAS (Beijing) Technology Company Limited, Beijing 100088, China)

Abstract: Text sentiment analysis has gradually become an important part of Natural Language Processing (NLP) in the fields of systematic recommendation and acquisition of user sentiment information, as well as public opinion reference for the government and enterprises. The methods in the field of sentiment analysis were compared and summarized by literature research. Firstly, literature investigation was carried out on the methods of sentiment analysis from the dimensions of time and method. Then, the main methods and application scenarios of sentiment analysis were summarized and compared. Finally, the advantages and disadvantages of each method were analyzed. According to the analysis results, in the face of different task scenarios, there are mainly three sentiment analysis methods: sentiment analysis based on emotion dictionary, sentiment analysis based on machine learning and sentiment analysis based on deep learning. The method based on multi-strategy mixture has become the trend of improvement. Literature investigation shows that there is still room for improvement in the techniques and methods of text sentiment analysis, and it has a large market and development prospects in e-commerce, psychotherapy and public opinion monitoring.

Key words: Natural Language Processing (NLP); emotional analysis; emotional dictionary; machine learning; deep learning

0 引言

随着计算机技术的发展,互联网的人数高速增长,截至2021年1月,互联网用户数量为46.6亿,而社交媒体用户数量为42亿。全球有52.2亿人使用手机,相当于世界总人口的66.6%。自2020年1月以来,手机用户数量增长了1.8%(9300万),而移动连接总数(一人拥有多部设备)增长了

1.9%(7200万),截止2021年1月达到80.2亿。90%的互联网用户通过智能手机上网,但2/3的人同时表示,他们使用笔记本电脑或台式电脑上网。由此,互联网每天都会产出难以想象的数据量。据Facebook统计,Facebook每天产生4PB的数据,包含100亿条消息,以及3.5亿张照片和1亿小时的视频浏览。可见在这样数据量的背景下,给了人工智能高速发展的机会,而人工智能的发展也会深深改变人们的生活方式。

收稿日期:2021-07-16;修回日期:2021-08-22;录用日期:2021-08-25。 基金项目:大连大学博士专项基金资助项目(2021QL09)。

作者简介: 王颖洁(1977—),女,黑龙江齐齐哈尔人,副教授,博士,CCF会员,主要研究方向:人工智能、软件工程; 朱久祺(1997—),男(彝族),云南楚雄人,硕士研究生,CCF会员,主要研究方向:深度学习、自然语言处理; 汪祖民(1975—),男,河南信阳人,教授,博士,CCF会员,主要研究方向:人工智能、物联网; 白凤波(1978—),男,吉林九台人,博士,CCF会员,主要研究方向:人工智能、数据科学; 弓箭(1994—),男,山西临汾人,主要研究方向:机器学习、自然语言处理。

自然语言处理(Natural Language Processing, NLP)是人工智能领域的一个重要方向,是一门研究如何让计算机听懂人类语言的学科。随着使用网络的人数增加,越来越多的网民可以在微博等平台上发表自己的观点,传播自己的想法和意见。因此在各大平台分析用户的语言中的情感色彩和具有倾向性的言论:对商家和企业而言,知道客户的喜好情感,有助于商家推出新的产品,改进自身的服务;在情感宣泄方面,如果能够从文字中知道用户的喜怒哀乐,那么计算机也可以充当一个情绪陪伴者和宣泄的窗口,或许能给心理疾病患者提供帮助。

因此在现阶段,通过自然语言处理,得到情感分析的结果是有重大意义的。

1 NLP 文本预处理

情感分析前的第一步,需要对文本进行预处理,一个好的文本处理可以事半功倍。

1.1 分词、停用词、词典

分词是自然语言处理的第一步,也是比较重要的一步, NLP 常用的分词算法可以分为三类:1)基于词典的分词,基于字典、词库匹配的分词方法;2)基于统计的分词,基于词频度统计的分词方法;3)基于规则的分词,基于知识理解的分词方法^[1]。

jieba 中文分词工具是一款广泛使用、分词效果较好的分词器,在 jieba 分词中有三种分词模式:1)精确模式,试图将句子最精确地切分开,适合文本分析;2)全模式,把句子中所有可以成词的词都扫描出来,速度非常快,但是不能解决歧义问题;3)搜索引擎模式,在精确模式的基础上,对长词再次进行切分,以提高召回率,适合搜索引擎分词。分词时,在面对一些专有名词的时候,分词效果不太理想,此时可以引入用户的自定义词典,人工划分出词语,字典一般为 txt 格式,导入较为简单。实际应用中,很多语气助词或者人称代词都不是需要关心的,在最终的结果中希望能够将其过滤掉,这时就需要建立停用词词典,可以从网络上下载停用词词典,也可以自己定义,在分词时,会把停用词字典中的词过滤掉,改善分词的效果。除此之外,在不同的领域会有不同的词典,有依据术语学理论制定的术语词典^[2]、态度词典、扩展情感词典、基于超短评论的图书领域情感词典^[3],等。

1.2 命名体识别

命名实体识别是一种信息抽取技术,用来解决互联网如今有大量信息资源的同时也造成信息过载的问题。第六届 MUC 会议(the sixth Message Understanding Conferences, MUC-6 第一次提出命名实体识别^[4],在自然语言处理中已经被广泛使用,信息抽取就是从非结构化的文本中(如新闻)抽取结构化的数据和特定的关系。识别文本中的人名、地名、时间等实体的名称,就叫做命名实体识别。最初 MUC-6 定义的识别的名词是人名、地名、机构名,随着使用的普及,对于不同的领域,划分的名词有了更详细的划分,在划分地名领域里加入了国家名、街道名等;在划分社交媒体领域里加入了电影名、邮箱名等。

命名实体识别的方法分为三类:

1)基于词典和规则的方法,这是提出命名实体概念后,最早提出的方法。这个方法需要构造大量的词典和规则集,多采用语言学家的规则模板,选取标点符号、关键字、中心词、指示词等特征,以模式和字符串匹配为手段,将识别对象放入对应的字典中进行匹配,需要经过多次修正和匹配,这种方法非常依赖规则和具体语言,领域和风格难以覆盖所有情况,导致只适合小规模的数据,而且系统移植性不强,遇到新的问题就需要构造新的词典,设定新的规则,系统建设周期长,代价高昂。随着机器学习的发展,进行命名实体识别的任务也考虑使用机器学习的方法。

2)基于机器学习的方法,将命名体识别看作一个序列标注问题。常使用的序列标注模型有:隐马尔可夫模型(Hidden Markov Model, HMM)、最大熵马尔可夫模型(Maximum Entropy Markov Model, MEMM)、条件随机场(Conditional Random Field, CRF)、支持向量机(Support Vector Machines, SVM)等。冯静等^[5]基于 HMM 的桥梁检测文本的命名实体识别,以 HMM 结合桥梁检测领域的词汇规则,在标注实验中取得了较好的效果;邵诗韵等^[6]基于 CRF 的电力工程标书文本实体识别方法,结合电力领域的规则,解决在工程中地名等实体的识别效果并不理想的问题;扈应等^[7]提出了一种结合 CRF 的边界组合命名实体识别方法,能够有效地识别生物医学文献中的嵌套和不连续性实体。

3)基于深度学习的方法,是在神经网络逐渐发展成熟后提出的,词向量的出现,可以解决高维空间的数据稀疏问题,也可以加入更多的特征。冯艳红等^[8]提出一种基于双向长短期记忆(Bi-directional Long Short-Term Memory, BiLSTM)神经网络的命名实体识别方法,准确率比起传统方法有所提升。此外,最近也出现很多利用 BERT(Bidirectional Encoder Representations from Transformers)、长短期记忆(Long Short-Term Memory, LSTM)网络、Transformer 等预训练模型,结合自注意力机制、迁移学习等来提高命名实体识别准确率的方法。罗熹等^[9]在采用 BiLSTM-CRF 模型的基础上,引入自注意力机制提升了中文临床命名实体的识别能力;郑洪浩等^[10]改进了 Transformer 的编码器,命名为 Transformer-P,结合 BERT 的升级模型 XLNET(transformer-XL fixed length context NET)模型和 CRF 提出了 XLNET-Transformer-P-CRF(XLNET Transformer Position CRF),模型改进了 Transformer 模型不能获得相对位置信息的缺点;董哲等^[11]提出了一种基于 BERT 和对抗训练的命名实体识别方法,在食品领域准确度有所提升;王子牛等^[12]提出了基于 BERT 和 BERT-BiLSTM 的 BERT-BiLSTM-CRF 命名实体识别方法,并和 CRF、卷积神经网络(Convolutional Neural Network, CNN)、LSTM 等方法作了对比,得到了更高的准确率。

也有很多公司已经开发了以 Python 语言为基础的语义处理库,例如 Aipnlp、BosonNLP。Aipnlp 是百度云自然语言处理的软件开发工具包(Software Development Kit, SDK),调用这个库可以获得百度云的相关服务,用来对文本进行词性标注统计和情感分析;BosonNLP 对文本进行语义聚类,同时也提供了情感分析、分词、词性标注、提取关键词、语义联想

等功能,其中的情感分析接口,用微博、新闻、汽车、餐饮等不同行业语料进行标注和机器学习,调用时请通过URL参数选择特定的模型,以获得最佳的情感判断准确率,把情绪分为正面和负面两类。使用开源的应用程序编程接口(Application Programming Interface, API)也可以帮助开发者提高开发的效率。

2 基于情感词典的情感分析法

基于情感词典的情感分析法利用情感词典获取文档中情感词的情感值,再通过加权计算来确定文档的整体情感倾向。这个方法可以对词语进行情感界定,易于分析和读者理解,如果这个词典内容足够丰富,就可以获得较好的情感分析效果;同时,情感词典不止可以考虑到文本内容,包括颜文字、文字表情等都可以进行情感分析,在性能上有较好的结果。

情感词典的一般执行过程如图1所示。首先是将文本输入,通过对数据的预处理(包含去噪、去除无效字符等),接着进行分词操作,然后将情感词典中不同类型和程度的词语放入模型中进行训练,最后根据情感判断输出情感类型。

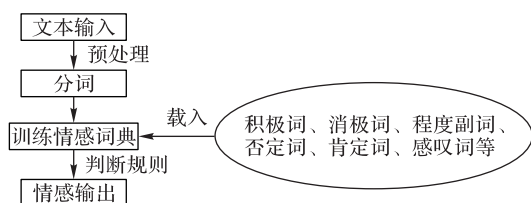


图1 基于情感词典的情感分析法一般流程

Fig. 1 General flow of sentiment analysis based on sentiment dictionary

英文的情感词典起步较早,相对来说也比较成熟。Cynthia^[13]对早期的英文情感词典进行了完善,增加其对自然语言样本的适用性。国外最早出现的英文情感词典是SentiWordNet,除此之外常用的有General Inquirer、Opinion Lexicon和MPQA (Question and Answer from Multiple Perspectives)等;中文情感词典应用比较广泛的有知网词典HowNet和大连理工大学的中文情感词汇本体库等^[14]。

为了解决在日常表达中英文混合的情况,栗雨晴等^[15]构建双语多类情感词典,提出基于双语词典的情感分析方法和基于K近邻分类算法对微博的文字分析,在多类情感分析和分类结果上都取得了不错的效果。微博的数据量对情感分析来说是巨大的资源,面对微博大量的数据,赵妍妍等^[16]基于文本统计算法,用微博大量的数据构建了一个10万词语的超大情感词典,性能超过其他中文词典,在针对微博的情感分类中,效果有了很大提升。万岩等^[17]提出在情感词典基础上,加入表情符号、网络用语等得到一个多类情感词典,利用微博的数据建立新的情感分析模型,有效提升了细粒度情感分类的准确率。邱全磊等^[18]在情感词典里加入了颜文字和语气词,然后提出了一种新的计算情感值的方法,这个方法在计算时考虑了颜文字标签和语气对情感的影响。高华玲等^[19]在通用词典的基础上,针对用户评论的情感分析,加入短语词典、否定词词典和副词词典等特殊词典。

Asghar等^[20]整合了表情符号、修饰语和领域特定术语来分析在线社区发布的评论,克服了以往方法的局限性,和基准方法相比,在考虑表情符号、修饰词、否定词和领域特定术语后,情感分析性能得到了提高。Cai等^[21]提出在面对不同的领域时使用不同的情感词典,将这个领域独有的一些新词或专业词包含在词典中,将SVM和梯度下降树(Gradient Boosting Decision Tree, GBDT)结合在一起,可以获得更好的性能。吴杰胜^[22]提出了可以同时使用多部情感词典,因此专门开发了六部情感词典并结合语义规则,提高了中文微博情感分析的准确性。

如今的大部分情感词典都是中文或者英文的,小众语言的词典非常少,文献[23]提出了用数据挖掘的方法搭建了一部韩语情感词典DecoSelex。Thien等^[24]提出一种适用于越南语的情感词典,涵盖了超10万个越南语的情感词汇。Wu等^[25]利用网络资源建立了一部俚语词典SlangSD,来衡量用户的情绪。

新词语的出现速度是很快的,相同的词也可能被赋予新的含义,为了将这些新词扩充到情感词典中,季鹏飞等^[26]提出了基于共现概率训练的情感词典扩充方法,先文本预处理,使用共现概率以及新词情感概率判断新词的情感倾向,最后把新词加入词典。李永帅^[27]提出了基于BiLSTM的动态情感词典构建方法,改进CBOW (Continuous Bag of Words)模型为ECBOW (Emotional CBOW)模型实现一个动态情感词典。Zhang等^[28]基于已构建好的程度副词词典、网络词词典、否定词词典等相关词典,用微博文本进行训练,得到更新后的情感值。郑赛乾^[29]基于Word2vec的情感词典构建子系统,通过提取用户聊天中的有效信息,帮助用户构建自己的情感词典,提取情感更加准确。Bravo-Marques等^[30]提出了一种从时变分布词向量中训练增量词的情感分类器,自动地从Twitter流中提取不断更新的情感词汇,得到基于增量词向量的时变情感词典。

但是基于情感词典的方法过度依赖于情感词典的创建,始终有作为字典的局限,并且对于成语、歇后语等的识别效果并不理想;在今天这样的信息时代,每天都会有新的词产生,有旧的词淘汰,现实中做不到每天都对词典进行更新;同时,这个方法也没有考虑词语之间的关系,没有联系上下文,导致词语的情感值不会根据文章或语句作动态的变化;情感词典的移植性也较差,一个词在不同时间、不同的语言、不同的文化中,所蕴含的意思并不相同,所以有更多的研究者开始着手基于机器学习的研究。

3 基于机器学习的情感分析法

机器学习算法(Machine Learning algorithm, ML)是一类算法的总称,可以看成机器学习就是寻找一个函数,输入海量的数据,输出是的结果,可以是分类、预测等,并追求把这个函数运用到新的样本数据上,也能得到想要的结果。基于机器学习的情感分析法,就是构造一个模型(这个模型就是一个函数的集合,只是太过复杂),输入大量有标签或者没有标签的语言信息,使用机器学习的算法,提取特征,找出代价

函数最小的结果作为输出。比起使用情感词典的方法,机器学习不依赖人工搭建,节省了大量人力,通过数据库,以及对词库进行更新。

在机器学习中, K 最近邻(K -Nearest Neighbor, KNN)算法、朴素贝叶斯(Naive Bayesian, NB)和SVM是常用的学习算法。在ML中,SVM和NB对于文本数据的分类效果较好。

3.1 支持向量机

机器学习中大部分的情感分类方法都离不开SVM和NB。SVM是一种二分类模型,SVM的学习策略就是间隔最大化,它的基本模型是定义在特征空间上的间隔最大的线性分类器,这是它和感知机的区别,而为了解决非线性的分类问题,SVM使用核函数解决了SVM无法处理非线性可分的问题。

SVM学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面,如图2所示, $w \cdot x + b = 0$ 即为分离超平面,对于线性可分的数据集来说,这样的超平面有无穷多个(即感知机),但是几何间隔最大的分离超平面却是唯一的。对于基于SVM的分类学习而言,问题已经从找到一个最好的划分超平面转换为了找到样本空间里的最大化间隔。

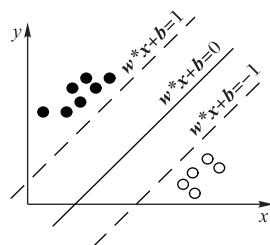


图2 支持向量机

Fig. 2 Support vector machine

对于输入空间中的非线性分类问题如图3所示,无论怎样一个线性超平面都无法很好地将样本点分类,所以在此引入核函数,核函数的主要作用是将样本从原始空间映射到一个更高维的特征空间,使得样本在这个特征空间内线性可分。具体的数学推导这里不做赘述,可以参考文献[31]或者其他相关书籍。

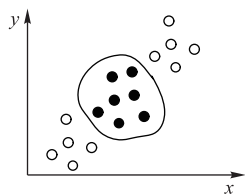


图3 非线性可分

Fig. 3 Non-linear separable

3.2 朴素贝叶斯

贝叶斯分类是一类分类算法的总称,这类算法均以贝叶斯定理为基础,故统称为贝叶斯分类。核心公式如式(1)所示:

$$P(\text{类别}|\text{特征}) = \frac{P(\text{特征}|\text{类别})P(\text{类别})}{P(\text{特征})} \quad (1)$$

朴素贝叶斯(Naive Bayes, NB)是在贝叶斯分类的基础上进行了相应的简化,即假定给定目标值时属性之间相互条件独立。虽然这个简化方式在一定程度上降低了贝叶斯分类的分类效果,但是在实际的应用场景中,极大地简化了贝叶

斯分类的复杂性。肖铮^[32]详细介绍了NB的原理和应用,并和KNN和决策树(Decision Tree)作了比较;李静梅等^[33]在特征独立性假设的基础上,加入期望值最大算法,扩展了NB分类器的应用,提高了NB分类器的分类精度。彭子豪等^[34]提出了一种基于TFIDFCF(Term Frequency Inverse Document Frequency Category Frequency)特征加权的并行化NB文本分类算法,提高了训练速度和精度。

3.3 机器学习情感分析

杨爽等^[35]提出基于SVM的情感分类方法,这个方法以词性特征、情感特征、句式特征、语义特征等作为依据,进行5级情感分类,得到了82.40%的准确率。杨妥等^[36]将SVM和LSTM方法相结合建立SVM-LSTM模型,用这个模型和传统的SVM模型、BP(Back Propagation)模型、SVM-BP模型作比较,发现这个方法实验结果的均方差(Mean Square Error, MSE)达到了0.172 2,比传统模型的均方差减小了约0.083。孙翌博等^[37]提出一种基于SD-LS-SVM(Simple optimization of Dynamic confidence interval for Least Squares SVM)算法的评论情感分析模型,先构建词网,再用上下文分析技术计算待检测评论中分词的评分,提取评论数据的特征向量。对最小二乘支持向量机(Least Squares-SVM, LS-SVM)进行基于置信区间简单动态优化的向量修剪,由改进的SD-LS-SVM算法对评论数据进行情感分类,准确率可以到85%。张成博^[38]在SA-SVM(Simulated Annealing-SVM)基础上提出FV-SA-SVM(Frequency Variance-Simulated Annealing-SVM)模型结合K-Means算法来判断文本是否消极,首先对影评进行预处理,然后使用FV-SA-SVM把文本划分为积极或者消极,分类的准确率达到94.7%。Hasan等^[39]利用Twitter的数据,结合SVM和朴素贝叶斯构建了一个情感分析器,并和仅适用SVM或者NB的情感分析器作了比较。Baid等^[40]对比了伯努利朴素贝叶斯(Bernoulli NB, BNB)、决策树、SVM、最大熵(Maximum Entropy, ME),以及多项式朴素贝叶斯(Multinomial NB, MNB)在情感分类中的效果,MNB得到了最好的88.5%。Ahmad等^[41]结合SVM提出了优化的情感分析框架OSAF(Optimized Sentiment Analysis Framework),该框架使用了SVM网格搜索技术和交叉验证。Mathapati等^[42]提出一种基于表情符号的情绪分析法,讨论了符号表情在情绪分析中的作用。Birjali等^[43]提出了一种基于WordNet语言学英语词典训练集的语义分析计算算法,使用机器学习算法SVM和NB相结合自动检测强关联消极推文。

基于机器学习的情感分类法比起构建情感词典有一定的进步,但是还是需要人工对文本特征进行标记,人为的主观因素会影响的最后的结果;其次,机器学习需要依赖大量的数据,很容易产生无效的工作,执行的速度会很慢,如果模型的效率不高,难以适应如今信息量爆炸的时代,这类方法在进行情感分析时常常不能充分利用上下文文本的语境信息,对准确性会造成影响。

基于机器学习的情感分析方法在Twitter数据集上获得的性能对比如表1所示。

因此,为了解决以上的这些问题,开始基于深度学习研

究新的方法,并取得了很多成果。

表1 基于机器学习的情感分析的实验结果 单位: %

方法	准确率	召回率	F1值
SVM多特征融合	82.40	81.91	82.10
SVM	88.20	89.00	88.00
FV-SA-SVM	92.90	93.00	89.00
SA-SVM	89.40	90.00	89.00
NB	86.60	87.00	87.00
Logistic 回归	86.90	85.00	86.00
KNN	87.00	87.00	85.00
MNB	88.50	83.33	87.87
BNB	87.50	86.33	87.35
ME	60.67	84.67	68.28
决策树	80.17	82.33	80.58

4 基于深度学习的情感分析法

深度学习其实是机器学习的一个子集,是多层神经网络在学习中的应用,覆盖了多个领域,涉及到较多知识,解决了以往的机器学习难以解决的大量问题,在图像、语音处理和情感分析领域取得了较好的效果,所以被人单独拿出来作为一门学科。目前深度学习模型包括CNN、循环神经网络(Recurrent Neural Network, RNN)、LSTM、Transformer、BiLSTM、门控循环单元(Gate Recurrent Unit, GRU)和注意力机制等。

神经网络的基本结构是输入层、隐藏层、输出层,如图4所示,输入层的每个神经元都可以作为一个对象的特征,隐藏层可以有一层或者多层,它就像一个黑盒子,不直接接收外界信号,也不向外界发送信号,它把输入转化为输出层可以使用的东西,输出层把隐藏层的结果转化为需要的结果。

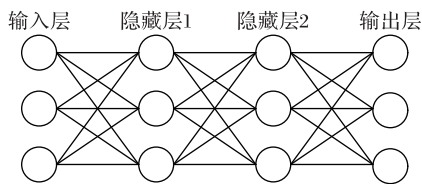


图4 神经网络结构

Fig. 4 Neural network structure

4.1 深度学习常用模型

情感分析大部分都是建立在CNN、RNN、LSTM等基本神经网络的基础上的,在这里先做一个简单介绍。

CNN如图5所示,与普通神经网络相似,它们都由具有可学习的权重和偏置常量的神经元组成。每个神经元都接收一些输入,并做一些点积计算,输出是每个分类的分数,普通神经网络里的一些计算技巧到这里依旧适用。CNN的基本结构为卷积层、池化层、全连接层。

CNN的输出都是只考虑前一个输入的影响而不考虑其他时刻输入的影响,前后的信息是不相关的,在一些需要之前信息作为参考的场景,效果不佳,所以又提出了RNN模型,如图6所示,上一时刻的 S_{t-1} 作为 S_t 的输入,保留了之前的信息。

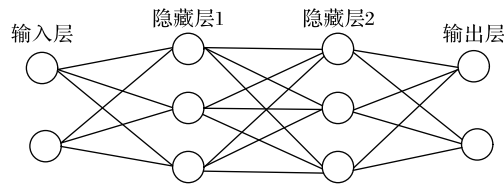


图5 CNN结构

Fig. 5 CNN structure

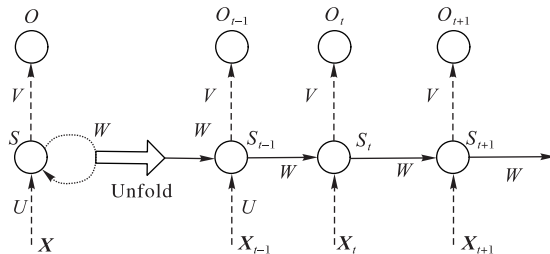


图6 RNN结构

Fig. 6 RNN structure

在有些情况下,为了解决RNN可能会出现梯度消失或者梯度爆炸问题,又提出了LSTM模型如图7所示。和RNN相比,LSTM只是运算的逻辑变了,也就是神经元的内部运算公式变了,但是结构并没有变,因此RNN的各种结构都能通过LSTM来替换。

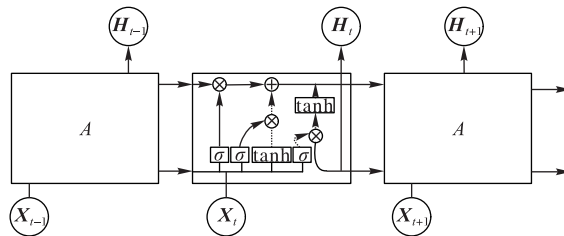


图7 LSTM结构

Fig. 7 LSTM structure

4.2 基于深度学习的分词及情感词典构建

分词、停用词、情感词典作为情感分析最基本的方法,如果能够在特定的领域构造一个专用的词典,假如这个词典的足够庞大,那么也可以得到较好的情感分析结果,出于这个想法,以深度学习为基础构造词典的想法就产生了。胡家珩等^[44]为解决训练神经网络分类器中训练语料不足和词向量的语义相关性无法区分情感信息的问题,利用词向量方法将文本信息映射到向量空间,借助已有的通用情感词典,自动标引训练语料,按照9:1的比例构建训练集和预测集,构建深度神经网络分类器,判断情感极性后,构建情感词典;赵天锐等^[45]提出了Word2vec和深度学习相结合的针对韩语的情感词典构建方法,这个方法先将已有的韩语情感词典合并,之后构建词向量模型,将词向量输入神经网络进行训练,构建情感分类器,利用神经网络对影评中单独出现的词汇进行情感倾向判别,最后整理得到所需的情感词典。

4.3 单一神经网络

在深度学习中,CNN取得了较好的效果,但是CNN没有考虑到文本的潜在主题,Dwivedi等^[46]提出一种基于受限玻尔兹曼机(Restricted Boltzmann Machine, RBM)的规则模型,用于分析句子的情感分析。曹宇等^[47]提出一种基于BGRU

(Bidirectional Gated Recurrent Unit)的中文文本情感分析方法,首先将文本转换为词向量序列,然后利用BGRU获得文本的上下文情感特征,最后由分类器给出文本的情感倾向,F1值达到了90.61%,在准确率和训练速度上都大于CNN方法。RNN结构中,输出不止取决于当前输入,之前的输入也会考虑到,就实现了联系上下文的效果。Can等^[48]提出了一种基于RNN框架的受限数据框架,使用最大数据集的语言来构建单个模型,并将其重用于资源有限的语言,对小语种的情感分析有更好的效果。潘红丽^[49]基于RNN,结合LSTM网络,对于英文文本中的情感信息,准确率达到94.5%。LSTM是RNN的一种特殊结构,为了提高训练速度,减少计算成本和时间,Wang等^[50]提出了一个基于注意力的LSTM面向方面层次情感的记忆网络分类,在LSTM的基础上进行感情分类;Gopalakrishnan等^[51]提出了6种不同参数的精简LSTM,使用Twitter数据集,比较了这几种LSTM的性能区别,为LSTM建立最佳的参数集合;Chen等^[52]基于推特和微博的数据提出了一种新的情感分析方案,关注表情对于情感的影响,通过参加这些双叉表情训练一个情感分类器,嵌入基于注意力的长短期记忆网络,对情感分析有较好的指导作用。

4.4 单一神经网络与注意力机制

在以上单一神经网络的基础上,引入注意力机制,注意力模型最初被用于机器翻译,现在已成为神经网络结构的重要组成部分,并在自然语言处理领域有着大量的应用。注意力模型借鉴了人类的注意力机制,能从众多信息中选择出对当前任务目标更关键的信息,在情绪分析中可以更加关注代表情绪的词汇。Wei等^[53]提出BiLSTM网络与多极化正交注意力机制相结合,用于隐式情感分析。Yadav等^[54]提出双向注意力机制和GRU相结合进行情绪预测,在restaurant 16数据集中取得了89.3%的准确率。Baziotis等^[55]基于LSTM,从Twitter上下载数据,引入注意力机制,放大最重要的词语的贡献来得到更好的效果。Liu等^[56]提出了一种新的基于内容注意的基于方面的情感分类模型,该模型具有两种注意增强机制:句子层面的内容注意机制能够从全局的角度捕捉特定方面的重要信息,而语境注意机制则负责同时考虑词语的顺序及其相关性。Chen等^[57]采用多注意机制捕获远距离分离的情感特征,对不相关信息具有更强的鲁棒性,通过对记忆的多重关注来提取重要信息,并结合不同关注点的特征来预测最终情感。Liang等^[58]针对目前研究中注意机制更多地与RNN或LSTM网络相结合,这种基于神经网络的结构通常依赖于复杂的结构,不能对句子中的单词进行平行化的问题,提出了一种多注意神经网络MATT-LSTM(Multi ATTention LSTM)进行基于方面的情感分类。该方法无需使用任何外部解析结果,即可通过多注意机制捕获更深层次的情感信息,明确区分不同方面的情感极性并和ATT-LSTM(ATTention LSTM)网络作了比较。Sangeetha等^[59]提出采用多层注意力机制并行处理句子输入序列,并使用不同的剪枝比例来提高准确率,随后把多层的信息融合并把结果作为输入馈送到LSTM层,多层融合的LSTM比一般的自然语言处理方法的的结果更好。

4.5 混合神经网络

除了对单一神经网络和加入自注意力机制的研究外,有不少学者在考虑了不同神经网络的优点后将这些结合起来,取长补短,并用于情感分析。Basiri等^[60]提出一种基于注意力的双向CNN-RNN深度模型(Attention based Bidirectional CNN-RNN Deep Model, ABCDM)。ABCDM利用两个独立的双向LSTM和GRU层,通过考虑两个方向上的时间信息流来提取过去和未来上下文;同时,在ABCDM双向层的输出中应用注意机制,对不同的词语进行或多或少的强调。Rehman等^[61]提出了一种利用LSTM和深度CNN模型的混合模型,首先使用Word2vec方法训练初始词嵌入,然后将卷积提取的特征集与具有长期依赖性的全局最大池层相结合;该模型还采用了dropout技术、归一化和校正线性单元来提高精度。杜永萍等^[62]提出一种基于CNN-LSTM模型的短文本情感分类方法,利用CNN,引入LSTM,并和用3D卷积代替2D卷积的E-LSTM(Embedding LSTM)方法作了比较,在此基础上再引出BiE-LSTM(Bi-directional Embedding LSTM)。Minaee等^[63]提出了基于CNN和BiLSTM模型的情感分析方法,一个用于捕获数据的时间信息,另一个用于提取数据的局部结构,效果要优于单纯使用CNN-LSTM模型;Li等^[64]提出了一种基于认知脑边缘系统(Hierarchical Attention BiLSTM based on Cognitive Brain, ALCB)的层次注意BiLSTM模型,结合随机森林算法的多模态情感分析方法。黄山成等^[65]提出一种基于EBA(ERNIE2.0-BiLSTM-Attention)的隐式情感分析方法,能够较好捕捉隐式情感句的语义及上下文信息,有效提升隐式情感的识别能力,最后在SMP2019公开数据集上取得较好分类效果,分类模型准确率达到82.3%。徐超等^[66]提出了一种基于MAML(Model Agnostic Meta Learning)与BiLSTM的情感分类方法,使用梯度下降更新参数,相较于现在流行的模型,在情感数据集上,准确率、召回率和F1值分别提升了1.68百分点、2.86百分点和2.27百分点。程艳等^[67]提出一种基于注意力机制的多通道CNN和双向门控循环单元,这个方法解决了传统RNN模型存在的梯度消失或梯度爆炸问题,而且还能很好地弥补CNN不能有效提取上下文语义信息的问题,引入注意力机制,更好地提取句子的局部特征。李卫疆等^[68]提出了一种基于多通道双向长短期记忆网络的情感分析模型(Multi-BiLSTM),取得了比普通BiLSTM更好的效果,并和SATT-BiLSTM(Self Attention BiLSTM)作了比较。Wang等^[69]提出维度情感分析,该模型由区域CNN和LSTM两部分组成,是一种树状结构的区域CNN-LSTM模型,并和T-LSTM(Tree LSTM)模型作了比较。

4.6 预训练模型

预训练模型是指用数据集已经训练好的模型,研究者希望花了很多时间训练的模型可以保留下来,在遇到类似情况的时候,可以调整参数后直接使用,就节约了再训练的时间,也能得到较好的结果。最新的预训练模型有:ELMo(Embeddings from Language Models)、BERT(Bidirectional Encoder Representation from Transformers)、XLNET、ALBERT(A Lite BERT)、Transformer等。

赵亚欧等^[70]提出基于ELMo和Transformer的混合模型,并引入LSTM和多头注意力机制用于情感分类,解决了双向语义和一词多义的问题,在NLPCC2014 Task2数据集上分类正确率提高了3.52个百分点。李铮等^[71]针对Word2vec、Glove(Global vector)、CNN等方法无法关注上下文,提取特征不充分的缺陷,引入自注意力机制,提出了双向自注意力网络Bi-SAN(Bidirectional Self-Attention Network)的情感分析模型,在NLPCC2014 task2中文数据集取得了更高的F1值,并和多尺度卷积神经网络(Multi-Scale CNN, MSCNN)作了比较,取得了较好的效果,又在此基础上提出了EBi-SAN(ELMo Bi-SAN)。胡任远等^[72]提出一种多层协同卷积神经网络模型MCNN(Multi-level CNN),并和BERT模型结合,提出BERT-MCNN(BERT Multi-level CNN)模型,并作了和Word2vec-MCNN, Glove-MCNN, ELMO-MCNN三个模型的对比实验,基于BERT的表面情感分类的能力有明显提升。Devlin等^[73]提出基于BERT的新方法,评分提高到80.5%(提高7.7个百分点)。Xu等^[74]把ELMo和BERT结合起来提出

了DomBERT(Domain oriented language model based on BERT)模型,在基于方面的情感分析中,显示了良好的效果并和BERT+Linear(BERT-Linear model)、BERT-DK(BERT on Domains Knowledge)模型作了性能比较。Sun等^[75]提出基于方面的情感分析,对BERT的预训练模型进行了微调,并在SentiHood和SemEval-2014 Task 4数据集上获得了更好的结果。Yin等^[76]提出了SentiBERT(Sentiment-BERT),是BERT的一个变种,与基线方法相比,SentiBERT方法在捕获否定关系和对比关系以及构建组合模型方面更有优势。

预训练模型作为一种迁移学习的应用,它可以将从开放领域学到的知识迁移到下游任务,以改善低资源任务,对低资源语言处理也非常有利,在几乎所有NLP任务中都取得了目前最佳的成果。同时预训练模型+微调机制具备很好的可扩展性,出现一个新任务时,不需要重复使用大量的时间和数据训练一个新的模型,只需要根据需求调整参数即可。

基于深度学习的情感分析方法在不同数据集上的性能表现如表2所示。

表 2 基于深度学习的情感分析实验结果

Tab. 2 Experimental results of sentiment analysis based on deep learning

单位:%
unit:%

模型	数据集	准确率	召回率	F1 值	模型	数据集	准确率	召回率	F1 值
CNN	NLPCC2014	72.15	74.24	73.18	BiE-LSTM	Twitter	82.20	88.10	85.00
LSTM	NLPCC2014	75.84	75.02	75.43	MATT-LSTM	Twitter	83.10	87.20	85.10
MSCNN	NLPCC2014	73.98	72.74	73.35	SentiBERT	Twitter	71.50	78.69	74.10
BiLSTM	NLPCC2014	76.12	74.52	75.31	BERT	Twitter	68.86	73.00	70.78
SATT-BiLSTM	NLPCC2014	76.23	74.84	75.53	BERT+Linear	Twitter	71.42	75.25	73.22
Bi-SAN	NLPCC2014	75.66	79.23	77.40	BERT-DK	Twitter	71.88	74.07	72.88
EBi-SAN	NLPCC2014	77.89	79.08	78.48	DomBERT	Twitter	72.17	74.96	73.45
RBM	NLPCC2014	71.30	87.01	78.41	Word2vec-MCNN	CN-o2o	88.10	90.97	90.03
T-LSTM	Twitter	70.80	82.51	76.20	Glove-MCNN	CN-o2o	89.47	91.66	90.55
E-LSTM	Twitter	81.60	82.50	82.00	ELMO-MCNN	CN-o2o	93.04	95.49	94.25
ATT-LSTM	Twitter	82.51	86.50	84.62	BERT-MCNN	CN-o2o	95.30	98.33	96.52

近几年提出的方法不再仅仅使用单一的神经网络,大多基于预训练模型、注意力机制等多种方法相结合,相互取长补短,通过对预训练模型进行微调来取得更好的效果。

5 结语

情感词典的方法易于理解,使用大量的数据,也可以获得很好的效果,但是始终有一个词典的“边界”,移植性、及时性不强,且会耗费大量的人力物力。机器学习的方法不依赖人工搭建,节省了人力,可以使用数据库及时更新词汇,但是机器学习依赖人工序列标注,不能充分利用上下文的语境,影响准确性。深度学习能够充分利用上下文的信息,保留语句的前后顺序,来实现情感分类和一词多义,使用多层神经网络,提取数据特征,学习性能更好,通过使用预训练模型,节省了之后的开发时间,降低了使用难度,在最近的NLP领域中效用很好;但是需要大量的数据支撑,算法训练时间很长,且深度越深,时间越长,开发难度较大。

随着大数据和人工智能领域的发展,各种新兴媒体的出现,有了大量的数据可以支持深度学习使用,同时也会有越来越多的新词出现,能准确判断不同语境下的词语含义对于分析和决策有至关重要的作用。

在NLP情感分析的研究中未来会面对的重点问题如下:

在语境中,目前还没有找到较好的方法处理反语;大部分的分类仍使用的二分类情感分析,对于多分类的情感分析还没有好的效果;多模态融合语料的情感分类也是近年的热点,不同模态中情感信息的权重如何分配,考虑外部语义信息对情感分类的准确性是否有帮助,也需要研究。

参考文献 (References)

[1] 敖盛,徐岚,敖清文. NLP中文分词技术在桥梁报告数据处理中的应用[J]. 交通世界, 2020(17): 3-5. (AO S, XU L, AO Q W. Application of NLP Chinese word segmentation technology in bridge report data processing[J]. Traffic World, 2020(17): 3-5.)

[2] 郑贞,王华丹,郝丽华. 术语学视角下的美军军语词典研究[J]. 中国科技术语, 2021, 23(3): 33-41. (ZHENG Z, WANG H D, HAO L H. Research on US military dictionary from the perspective of terminology [J]. Chinese Science and Technology Terminology: 2021, 23(3): 33-41.)

[3] 周知,王春迎,朱佳丽. 基于超短评论的图书领域情感词典构建研究[J]. 情报理论与实践, 2021, 44(9): 183-189, 197. (ZHOU Z, WANG C Y, ZHU J L. Research on the construction of emotional dictionaries in the book domain based on ultra-short reviews [J]. Information Theory and Practice, 2021, 44(9): 183-189, 197.)

[4] 陈基. 命名实体识别综述[J]. 现代计算机(专业版), 2016(3): 24-26. (CHEN J. Overview of named entity recognition[J]. Modern

- Computer (Professional Edition), 2016(3): 24-26.)
- [5] 冯静,李正武,张登云,等. 基于隐马尔可夫模型的桥梁检测文本命名实体识别[J]. 交通世界, 2020(8): 32-33. (FENG J, LI Z W, ZHANG D Y, et al. Named entity recognition of bridge detection text based on hidden Markov model[J]. Traffic World, 2020(8): 32-33.)
 - [6] 邵诗韵,周宇,杨蕾,等. 基于条件随机场的电力工程标书文本实体识别方法[J]. 计算机与现代化, 2020(12): 72-77. (SHAO S Y, ZHOU Y, YANG L, et al. Power engineering bid text entity recognition method based on conditional random field[J]. Computer and Modernization, 2020(12): 72-77.)
 - [7] 扈应,陈艳平,黄瑞章,等. 结合CRF的边界组合生物医学命名实体识别[J]. 计算机应用研究, 2021, 38(7): 2025-2031. (HU Y, CHEN Y P, HUANG R Z, et al. Boundary combinatorial biomedical named entity recognition combined with CRF [J]. Application Research of Computers, 2021, 38(7): 2025-2031.)
 - [8] 冯艳红,于红,孙庚,等. 基于BLSTM的命名实体识别方法[J]. 计算机科学, 2018, 45(2): 261-268. (FENG Y H, YU H, SUN G, et al. Named entity recognition method based on BLSTM [J]. Computer Science, 2018, 45(2): 261-268.)
 - [9] 罗熹,夏先运,安莹,等. 结合多头自注意力机制与BiLSTM-CRF的中文临床实体识别[J]. 湖南大学学报(自然科学版), 2021, 48(4): 45-55. (LUO X, XIA X Y, AN Y, et al. Chinese clinical entity recognition combining multi-head self-attention mechanism and BiLSTM-CRF [J]. Journal of Hunan University (Natural Science Edition), 2021, 48(4): 45-55.)
 - [10] 郑洪浩,于洪涛,李邵梅. 基于改进的Transformer编码器的中文命名实体识别[J]. 网络与信息安全学报, 2021, 7(5): 105-112. (ZHENG H H, YU H T, LI S M. Chinese named entity recognition based on improved Transformer encoder[J]. Journal of Network and Information Security, 2021, 7(5): 105-112.)
 - [11] 董哲,邵若琦,陈玉梁,等. 基于BERT和对抗训练的食品领域命名实体识别[J]. 计算机科学, 2021, 48(5): 247-253. (DONG Z, SHAO R Q, CHEN Y L, et al. Named entity recognition in the food field based on BERT and adversarial training [J]. Computer Science, 2021, 48(5): 247-253.)
 - [12] 王子牛,姜猛,高建瓴,等. 基于BERT的中文命名实体识别方法[J]. 计算机科学, 2019, 46(S2): 138-142. (WANG Z N, JIANG M, GAO J L, et al. Chinese named entity recognition method based on BERT [J]. Computer Science, 2019, 46(S2): 138-142.)
 - [13] CYNTHIA W. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language [J]. Psychological Reports, 2009, 105(2): 509-521.
 - [14] 钟佳娃,刘巍,王思丽,等. 文本情感分析方法及应用综述[J]. 数据分析与知识发现, 2021, 5(6): 1-13. (ZHONG J W, LIU W, WANG S L, et al. Summary of text sentiment analysis methods and applications [J]. Data Analysis and Knowledge Discovery, 2021, 5(6): 1-13.)
 - [15] 栗雨晴,礼欣,韩煦,等. 基于双语词典的微博多类情感分析方法[J]. 电子学报, 2016, 44(9): 2068-2073. (LI Y Q, LI X, HAN X, et al. Multi-sentiment analysis method of Weibo based on bilingual dictionary [J]. Acta Electronica Sinica, 2016, 44(9): 2068-2073.)
 - [16] 赵妍妍,秦兵,石秋慧,等. 大规模情感词典的构建及其在情感分类中的应用[J]. 中文信息学报, 2017, 31(2): 187-193. (ZHAO Y Y, QIN B, SHI Q H, et al. Construction of a large-scale sentiment dictionary and its application in sentiment classification [J]. Journal of Chinese Information Processing, 2017, 31(2): 187-193.)
 - [17] 万岩,杜振中. 融合情感词典和语义规则的微博评论细粒度情感分析[J]. 情报探索, 2020(11): 34-41. (WAN Y, DU Z Z. Fine-grained sentiment analysis of Weibo comments combined with sentiment dictionary and semantic rules [J]. Information Research, 2020(11): 34-41.)
 - [18] 邱全磊,崔宗敏,喻静. 基于表情和语气的情感词典用于弹幕情感分析[J]. 计算机技术与发展, 2020, 30(8): 178-182. (QIU Q L, CUI Z M, YU J. A sentiment dictionary based on facial expression and tone is used for bullet screen sentiment analysis [J]. Computer Technology and Development, 2020, 30(8): 178-182)
 - [19] 高华玲,张晶. 基于情感词典的酒店评论情感分析与可视化[J]. 软件, 2021, 42(1): 45-47, 66. (GAO H L, ZHANG J. Sentiment analysis and visualization of hotel reviews based on sentiment dictionary [J]. Computer Engineering & Software, 2021, 42(1): 45-47, 66.)
 - [20] ASGHAR M Z, KHAN A, AHMAD S, et al. Lexicon-enhanced sentiment analysis framework using rule-based classification scheme [J]. PLoS One, 2017, 12(2): e0171649.
 - [21] CAI Y, YANG K, HUANG D P, et al. A hybrid model for opinion mining based on domain sentiment dictionary [J]. International Journal of Machine Learning and Cybernetics, 2019, 10: 2131-2142.
 - [22] 吴杰胜. 基于多部情感词典和深度学习的中文微博情感分析研究[D]. 淮南: 安徽理工大学, 2020: 22-32. (WU J S. Research on sentiment analysis of Chinese Weibo based on multiple sentiment dictionaries and deep learning [D]. Huainan: Anhui University of Science and Technology, 2020: 22-32.)
 - [23] 신동혁, 남지순, 남지순. 정서 분석을 위한 한국 정서 사전 DecoSelex 구축 [J]. 인문학 일반, 2016, 11(28): 75-111. (SHEN D H, ZHAO D X, NAN Z S. Building the Korean sentiment lexicon DecoSelex for sentiment analysis [J]. Journal of Korealex, 2016, 11(28): 75-111.
 - [24] THIEN K T, THI P T. A hybrid approach for building a Vietnamese sentiment dictionary [J]. Journal of Intelligent & Fuzzy Systems, 2018, 35(1): 1-12.
 - [25] WU L, MORESTATTER F, LIU H, et al. SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification [J]. Language Resources and Evaluation, 2018, 52: 839-852.
 - [26] 季鹏飞,王先超,张顺香. 基于共现概率训练的情感词典的扩充[J]. 阜阳师范学院学报(自然科学版), 2019, 36(4): 49-53. (JI P F, WANG X C, ZHANG S X. Expansion of the emotional dictionary based on co-occurrence probability training [J]. Journal of Fuyang Teachers College (Natural Science Edition), 2019, 36(4): 49-53.)
 - [27] 李永帅. 基于双向LSTM的动态情感词典构建方法研究及文本情感分析[D]. 郑州: 郑州大学, 2019: 20-32. (LI Y S. Research on construction method of dynamic sentiment dictionary based on two-way LSTM and text sentiment analysis [D]. Zhengzhou: Zhengzhou University, 2019: 20-32.)
 - [28] ZHANG S X, WEI Z L, WANG Y, et al. Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary [J]. Future Generation Computer Systems, 2018, 81: 395-403.
 - [29] 郑赛乾. 基于Word2vec的客服情感词典自动构建子系统的设计与实现[D]. 北京: 北京邮电大学, 2020: 7-18. (ZHENG S Q. Design and implementation of the automatic construction subsystem of customer service emotional dictionary based on

- Word2Vec [D]. Beijing: Beijing University of Posts and Telecommunications, 2020:7-18.)
- [30] BRAVO-MARQUES F, KHANCHANDANI A, PFAHRINGER B. Incremental word vectors for time-evolving sentiment lexicon induction[J]. Cognitive Computation, 2021,14: 425-441.
- [31] 汪海燕,黎建辉,杨风雷. 支持向量机理论及算法研究综述[J]. 计算机应用研究,2014,31(5):1281-1286. (WANG H Y, LI J H, YANG F L. Support vector machine theory and algorithm research review [J]. Application Research of Computers, 2014, 31(5): 1281-1286.)
- [32] 肖铮. 常用的三种分类算法及其比较分析[J]. 重庆科技学院学报(自然科学版),2020,22(5):101-106. (XIAO Z. Three commonly used classification algorithms and their comparative analysis [J]. Journal of Chongqing University of Science and Technology (Natural Science Edition), 2020, 22(5): 101-106.)
- [33] 李静梅,孙丽华,张巧荣,等. 一种文本处理中的朴素贝叶斯分类器[J]. 哈尔滨工程大学学报,2003(1):71-74. (LI J M, SUN L H, ZHANG Q R, et al. A naive Bayes classifier in text processing [J]. Journal of Harbin Engineering University, 2003(1): 71-74.)
- [34] 彭子豪,谭欣. 并行化改进的朴素贝叶斯算法在中文文本分类上的应用[J]. 科学技术创新,2020,(26):176-178. (PENG Z H, TAN X. Application of parallelized improved naive bayes algorithm in Chinese text classification [J]. Science and Technology Innovation, 2020, (26): 176-178.)
- [35] 杨爽,陈芬. 基于SVM多特征融合的微博情感多级分类研究[J]. 数据分析与知识发现,2017,1(2):73-79. (YANG S, CHEN F. Multi-level classification of Weibo sentiment based on SVM multi-feature fusion [J]. Data Analysis and Knowledge Discovery, 2017, 1(2): 73-79.)
- [36] 杨妥,李万龙,郑山红. 融合情感分析与SVM_LSTM模型的股票指数预测[J]. 软件导刊,2020,19(8):14-18. (YANG T, LI W L, ZHENG S H. Stock index prediction based on sentiment analysis and SVM_LSTM model[J]. Software Guide, 2020, 19(8): 14-18.)
- [37] 孙翌博,濮泽堃,徐玉华,等. 应用SD-LS-SVM算法的评论情感分析模型[J]. 软件导刊,2021,20(4):43-48. (SUN Y B, PU Z K, XU Y H, et al. Comment sentiment analysis model using SD-LS-SVM algorithm [J]. Software Guide, 2021, 20(4): 43-48.)
- [38] 张成博. 基于FV-SA-SVM的电影评论情感分析[D]. 上海:上海师范大学,2020:35-39. (ZHANG C B. Sentiment analysis of movie reviews based on FV-SA-SVM [D]. Shanghai: Shanghai Normal University, 2020:35-39.)
- [39] HASAN A, MOIN S, KARIM A, et al. Machine learning-based sentiment analysis for twitter accounts [J]. Mathematical and Computational Applications, 2018, 23(1): 11.
- [40] BAID P, GUPTA A, CHAPLOT N. Sentiment analysis of movie reviews using machine learning techniques [J]. International Journal of Computer Applications, 2017, 179(7): 45-49.
- [41] AHMAD M, AFTAB S, BASHIR S, et al. SVM optimization for sentiment analysis [J]. International Journal Advanced Computer Science and Applications, 2018, 9(4): 393-398.
- [42] MATHAPATI S, NAFEESA A, MANJULA S H, et al. OTAWA-Optimized topic-adaptive word expansion for cross domain sentiment classification on tweets[M]//Advances in Machine Learning and Data Science. Singapore: Springer, 2018,705: 213-224.
- [43] BIRJALI M, BENI-HSSANE A, ERRITALI M. Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks [J]. Procedia Computer Science, 2017, 113: 65-72.
- [44] 胡家珩,岑咏华,吴承尧. 基于深度学习的领域情感词典自动构建——以金融领域为例[J]. 数据分析与知识发现,2018,2(10):95-102. (HU J H, CEN Y H, WU C Y. Automatic construction of domain sentiment dictionary based on deep learning—take the financial field as an example [J]. Data Analysis and Knowledge Discovery, 2018, 2(10): 95-102.)
- [45] 赵天锐,刘晨阳. 基于深度学习的韩国语影评情感词典构建[J]. 信息技术与信息化,2021(1):250-253. (ZHAO T R, LIU C Y. Construction of Korean film critics sentiment dictionary based on deep learning [J]. Information Technology and Informatization, 2021(1): 250-253.)
- [46] DWIVEDI R K, AGGARWAL M, KESHARI S K, et al. Sentiment analysis and feature extraction using rule-based model (RBM)[C]//Proceedings of the 2019 International Conference on Innovative Computing and Communications. Cham: Springer, 2019: 57-63.
- [47] 曹宇,李天瑞,贾真,等. BGRU: 中文文本情感分析的新方法[J]. 计算机科学与探索,2019,13(6):973-981. (CAO Y, LI T R, JIA Z, et al. BGRU: a new method of Chinese text sentiment analysis [J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(6): 973-981.)
- [48] CAN E F, EZEN-CAN A, CAN F. Multilingual sentiment analysis: an RNN-based framework for limited data [EB/OL]. [2018-06-08]. <https://arxiv.org/pdf/1806.04511.pdf>.
- [49] 潘红丽. 基于RNN弱监督网络的英语语义分析技术研究[J]. 电子设计工程,2021,29(15):97-101. (PAN H L. Research on English semantic analysis technology based on RNN weakly supervised network[J]. Electronic Design Engineering, 2021, 29(15): 97-101.)
- [50] WANG Y Q, HUANG M L, ZHU X Y, et al. Attention-based LSTM for aspect-level sentiment classification [C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2016: 606-615.
- [51] GOPALAKRISHNAN K, SALEM F M. Sentiment analysis using simplified long short-term memory recurrent neural networks [EB/OL]. [2020-05-08]. <https://arxiv.org/abs/2005.03993v1>.
- [52] CHEN Y X, YUAN J B, YOU Q Z, et al. Twitter sentiment analysis via bi-sense emoji embedding and attention-based LSTM [C]// Proceedings of the 26th ACM international conference on Multimedia. New York: ACM, 2018: 117-125.
- [53] WEI J Y, LIAO J, YANG Z F, et al. BiLSTM with multi-polarity orthogonal attention for implicit sentiment analysis [J]. Neurocomputing, 2020, 383: 165-173.
- [54] YADAV R K, JIAO L, GOODWIN M, et al. Positionless aspect based sentiment analysis using attention mechanism [J]. Knowledge-Based Systems, 2021, 226: 107136.
- [55] BAZIOTIS C, PELEKIS N, DOULKERIDIS C, et al. DataStories at SemEval-2017 task 4: deep LSTM with attention for message-level and topic-based sentiment analysis [C]// Proceedings of the 11th International Workshop on Semantic Evaluation. Stroudsburg, PA: Association for Computational Linguistics 2017: 747-754.
- [56] LIU Q, ZHANG H B, ZENG Y F, et al. Content attention model for aspect based sentiment analysis [C]// Proceedings of the 2018 World Wide Web Conference. 2018: 1023-1032.
- [57] CHEN P, SUN Z Q, BING L D, et al. Recurrent attention network on memory for aspect sentiment analysis [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language

- Processing. Stroudsburg, PA: Association for Computational Linguistics 2017: 452-461.
- [58] 梁斌, 刘全, 徐进, 等. 基于多注意力卷积神经网络的特定目标情感分析[J]. 计算机研究与发展, 2017, 54(8): 1724-1735. (LIANG B, LIU Q, XU J, et al. Aspect-based sentiment analysis based on multi-attention CNN[J]. Journal of Computer Research and Development, 2017, 54(8): 1724-1735.)
- [59] SANGEETHA K, PRABHA D. Sentiment analysis of student feedback using multi-head attention fusion model of word and context embedding for LSTM[J]. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(6): 4117-4126.
- [60] BASIRI M E, NEMATI S, ABDAR M, et al. ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis[J]. Future Generation Computer Systems, 2021, 115: 279-294.
- [61] REHMAN A U, MALIK A K, RAZA B, et al. A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis[J]. Multimedia Tools and Applications, 2019, 78(18): 26597-26613.
- [62] 杜永萍, 赵晓铮, 裴兵兵. 基于 CNN-LSTM 模型的短文本情感分类[J]. 北京工业大学学报, 2019, 45(7): 662-670. (DU Y P, ZHAO X Z, PEI B B, Short text sentiment classification based on CNN-LSTM model [J]. Journal of Beijing University of Technology, 2019, 45(7): 662-670.)
- [63] MINAE S, AZIMI E, ABDOLRASHIDI A A. Deep-sentiment: sentiment analysis using ensemble of CNN and Bi-LSTM models [EB/OL]. [2019-04-08]. <https://arxiv.org/pdf/1904.04206v1.pdf>.
- [64] LI Y Q, ZHANG K, WANG J Y, et al. A cognitive brain model for multimodal sentiment analysis based on attention neural networks[J]. Neurocomputing, 2021, 430: 159-173.
- [65] 黄山成, 韩东红, 乔百友, 等. 基于 ERNIE2.0-BiLSTM-Attention 的隐式情感分析方法[J]. 小型微型计算机系统, 2021, 42(12): 2485-2489. (HUANG S C, HAN D H, QIAO B Y, et al. Implicit sentiment analysis method based on ERNIE2.0-BiLSTM-Attention [J]. Journal of Chinese Computer Systems, 2021, 42(12): 2485-2489.)
- [66] 徐超, 叶宁, 徐康, 等. 融合 MAML 与 BiLSTM 的微博负面情感多分类方法[J/OL]. 计算机工程与应用: 1-11 [2021-01-26]. <http://kns.cnki.net/kcms/detail/11.2127.tp.20210416.1420.017.html>. (XU C, (YE N, XU K, et al. Weibo negative sentiment multi-classification method integrating MAML and BiLSTM [J/OL]. Computer Engineering and Applications: 1-11 [2021-01-26]. <http://kns.cnki.net/kcms/detail/11.2127.tp.20210416.1420.017.html>.)
- [67] 程艳, 尧磊波, 张光河, 等. 基于注意力机制的多通道 CNN 和 BiGRU 的文本情感倾向性分析[J]. 计算机研究与发展, 2020, 57(12): 2583-2595. (CHENG Y, YAO L B, ZHANG G H, et al. Multi-channel CNN and BiGRU text sentiment orientation analysis based on attention mechanism [J]. Computer Research and Development, 2020, 57(12): 2583-2595.)
- [68] 李卫疆, 漆芳. 基于多通道双向长短期记忆网络的情感分析[J]. 中文信息学报, 2019, 33(12): 119-128. (LI W J, QI F. Sentiment analysis based on multi-channel two-way long and short term memory network [J]. Journal of Chinese Information Processing, 2019, 33(12): 119-128.)
- [69] WANG J, YU L C, LAI K R, et al. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 581-591.
- [70] 赵亚欧, 张家重, 李贻斌, 等. 基于 ELMo 和 Transformer 混合模型的情感分析[J]. 中文信息学报, 2021, 35(3): 115-124. (ZHAO Y O, ZHANG J C, LI Y B, et al. Sentiment analysis based on ELMo and Transformer hybrid model [J]. Journal of Chinese Information Processing, 2021, 35(3): 115-124.)
- [71] 李铮, 陈莉, 张爽. 基于 ELMo 和 Bi-SAN 的中文文本情感分析[J]. 计算机应用研究, 2021, 38(8): 2303-2307. (LI Z, CHEN L, ZHANG S. Chinese text sentiment analysis based on ELMo and Bi-SAN [J/OL]. Application Research of Computers, 2021, 38(8): 2303-2307.)
- [72] 胡任远, 刘建华, 卜冠南, 等. 融合 BERT 的多层次语义协同模型情感分析研究[J]. 计算机工程与应用, 2021, 57(13): 176-184. (HU R Y, LIU J H, BU G N, et al. Research on sentiment analysis of multi-level semantic collaboration model fusion BERT [J]. Computer Engineering and Applications, 2021, 57(13): 176-184.)
- [73] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA: Association for Computational Linguistics, 2019: 4171-4186.
- [74] XU H, LIU B, SHU L, et al. DomBERT: domain-oriented language model for aspect-based sentiment analysis [C]// Proceedings of the 2020 Findings of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 1725-1731.
- [75] SUN C, HUANG L, QIU X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA: Association for Computational Linguistics, 2019: 380-385.
- [76] YIN D, MENG T, CHANG K W. SentiBERT: a transferable transformer-based architecture for compositional sentiment semantics [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 3695-3706.

This work is partially supported by Doctoral Special Fund of Dalian University (2021QL09).

WANG Yingjie, born in 1977, Ph.D., associate professor. Her research interests include artificial intelligence, software engineering.

ZHU Jiuqi, born in 1997, M. S. candidate. His research interests include deep learning, natural language processing.

WANG Zumin, born in 1975, Ph. D., professor. His research interests include artificial intelligence, Internet of things.

BAI Fengbo, born in 1978, Ph. D. His research interests include artificial intelligence, data science.

GONG Jian, born in 1994. His research interests include machine learning, natural language processing.