计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2017.20160807
54(6):1150-1170,2017

情感分类研究进展

陈 龙 管子玉 何金红 彭进业
(西北大学信息科学与技术学院 西安 710127)
(longchen@stumail.nwu.edu.cn)

A Survey on Sentiment Classification

Chen Long, Guan Ziyu, He Jinhong, and Peng Jinye
(*School of Information Science and Technology, Northwest University, Xi'an* 710127)

Abstract Sentiment analysis in text is an important research field for intelligent multimedia understanding. The aim of sentiment classification is to predict the sentiment polarity of opinionated text, which is the core of sentiment analysis. With rapid growth of online opinionated content, the traditional approaches such as lexicon-based methods and classic machine learning methods cannot well handle large-scale sentiment classification problems. In recent years, deep learning has achieved good performance on the intelligent understanding of large-scale text data and has attracted a lot of attention. More and more researchers start to address text classification problems with deep learning. The content of this survey is organized as two parts. We firstly summarize the traditional approaches including lexicon-based methods, machine learning based methods, hybrid methods, methods based on weakly labeled data and deep learning based methods. Secondly, we introduce our proposed weakly-supervised deep learning framework to deal with the defects of the previous approaches. Moreover, we briefly summarize the research work on the extraction of opinion aspects. Finally, we discuss the challenges and future work on sentiment classification.

Key words sentiment analysis; sentiment classification; deep learning; weak-supervision; aspects extraction

摘 要 文本情感分析是多媒体智能理解的重要问题之一.情感分类是情感分析领域的核心问题,旨在解决评论情感极性的自动判断问题.由于互联网评论数据规模与日俱增,传统基于词典的方法和基于机器学习的方法已经不能很好地处理海量评论的情感分类问题.随着近年来深度学习技术的快速发展,其在大规模文本数据的智能理解上表现出了独特的优势,越来越多的研究人员青睐于使用深度学习技术来解决文本分类问题.主要分为2个部分:1)归纳总结传统情感分类技术,包括基于字典的方法、基于机器学习的方法、两者混合方法、基于弱标注信息的方法以及基于深度学习的方法;2)针对前人情感分类方法的不足,详细介绍所提出的面向情感分类问题的弱监督深度学习框架.此外,还介绍了评论主题提取相关的经典工作.最后,总结了情感分类问题的难点和挑战,并对未来的研究工作进行了展望.

关键词 情感分析;情感分类;深度学习;弱监督;主题提取

中图法分类号 TP181

收稿日期:2016-11-10;**修回日期**:2017-03-14

基金项目:国家自然科学基金优秀青年科学基金项目(61522206)
This work was supported by the National Natural Science Foundation of China for Excellent Young Scientists (61522206).

通信作者:管子玉(ziyuguan@nwu.edu.cn)

进入 Web2.0 时代已历十余载,互联网的迅猛发展和移动终端的快速普及为用户提供了发表和分享个人言论的广阔平台.日常生活中,人们经常登陆不同类型网站(如社交网站、电商网站等)发表和分享个人观点:在社交网站上评论新闻时事、在电商网站上快速浏览商品评论、在影评网站上发表影片观后感等.这些评论中包含个人情感取向,通过分析评论中的情感取向可以有效把握舆情趋势,进而惠及政府和民众.政府可以通过分析社交媒体数据来体察民意,从而合理制定或调整相关政策;商家能够从商品评论摘要中得到消费者的反馈信息,进而优化营销策略;消费者则可以通过阅读其他用户发表的商品评论来决定是否购买.图 1 所示为一条商品评论摘要,其中红色文字为商品的正面评论摘要,绿色文字为商品的负面评论摘要,深红色方框中文字表示勾选的正面评论摘要示例.

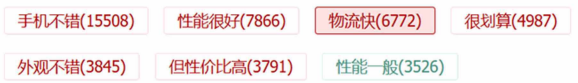


Fig. 1　Summarization of product reviews

图 1　商品评论摘要

分析上述不同类型评论数据中所包含的个人主观情感取向需要使用情感分析技术.情感分析(sentiment analysis),又称评论挖掘(opinion mining),它利用自然语言处理(natural language processing,NLP)、文本分析、机器学习、计算语言学(computational linguistics)等方法对带有情感色彩的文本进行分析、处理、推理和归纳.其标准定义为:情感分析是对文本中关于某个实体的观点、情感、情绪及态度的计算研究[1].通俗地讲,情感分析的目标就是明确评论者对所评论对象的态度.而情感分析最基本任务是在文档(document)、句子(sentences)或主题(topic,也称为 feature 或 aspect,下文统称 aspect)等不同层次上,将给定的评论文本分为积极(positive)、消极(negative)、中立(neutral)三个类别.在此基础之上,还可以根据实际问题设定多极情感分类目标,如将新闻评论分为"悲伤"、"乐观"、"愤怒".

目前,情感分析技术已经在政治、金融等领域崭露头角.文献[2]通过情感分析技术分析社交网站 Twitter 上用户的情感变化,结果显示通过情感分析技术得到的用户情感变化趋势与传统问卷调查方法的结果惊人地一致.如图 2 所示,研究人员对比了 2008 年 5 月至 2010 年 5 月期间美国民意调查结果(黑色实线)与同时期 Twitter 用户情感指数分析结果(蓝色实线),参数 *window* 表示天数,参数 *r* 表示图 2 中 2 个结果的相关度.图 2 中两者的相关性竟高达 80%.文献[3]将情感分析技术用于股票行情预测,如图 3 所示.图 3 中蓝线表示"冷静"情绪指数(CALM),该指数通过情感分析技术获得;红线表示道琼斯工业平均指数(DIJA).实验结果表明,"冷静"情绪指数沿时间轴向后推移 3 d 和道琼斯工业平均指数具有很高的一致性.因此,可以根据"冷静"情绪指数来预测股票行情.

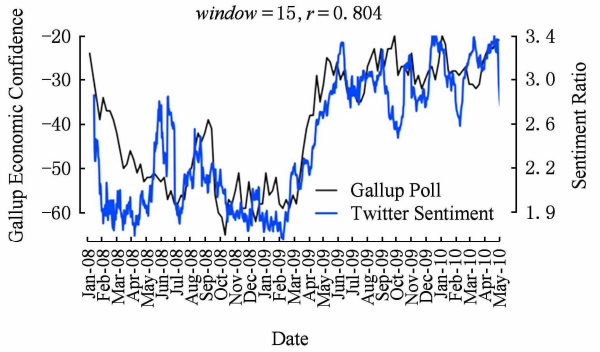


Fig. 2　Comparison between the result of sentiment analysis and polls[2]

图 2　情感分析结果与民意调查结果对比[2]

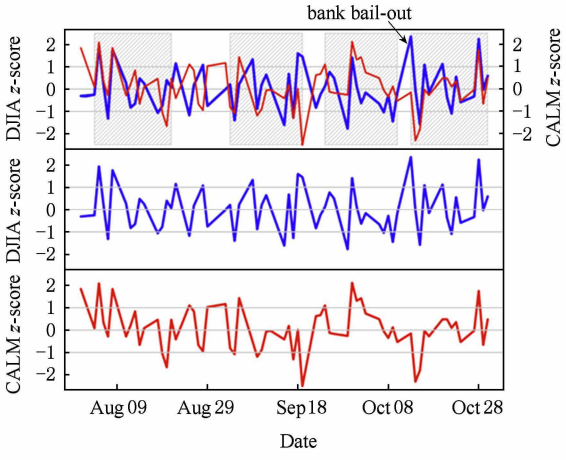


Fig. 3　Comparison between CALM and Dow Jones industrial average (DIJA)[3]

图 3　CALM 指数与道琼斯工业平均指数(DIJA)对比[3]

当前,互联网用户规模与日俱增.据《2016 年互联网趋势报告》统计显示,目前全球互联网用户数已超 30 亿,互联网全球渗透率达到 42%.其中,中国互联网用户数量达 6.68 亿,位居世界首位.日益膨胀的

① https://about.twitter.com/company

互联网用户群产生了规模庞大的评论文本数据. 据《2015 年度微博用户发展报告》统计,微博的日活跃用户量达 1 亿,日均产生数据量达 1 037 GB. 另据统计①,世界知名社交网站 Twitter 日均发布的推文(tweet)数量达 5 亿条. 如何有效分析和处理如此规模庞大的评论数据成为了研究者们面临的新挑战. 为了应对该挑战,研究者们提出了多种情感分析技术,如评论摘要技术[4-5]、对比句分析方法[6]、评论投票技术[7]等. 这些情感分析技术的关键问题有 2 个:1)提取评论主题;2)对评论内容进行情感分类.

1 评论主题提取方法

评论主题(topic),又称方面(aspect)、特征(feature),表示用户评论的对象. 评论主题的提取目标是提取或推测出评论对象的文字表达,如词汇、短语等. 评论中的主题一般分为 2 类:1)显式(explicit)主题;2)隐式(implicit)主题. 显式主题是指主题的文字表达直接存在于评论中,如手机评论"The apps are amazing."中的主题词"apps";而隐式表达中不包含主题的文字表达,但可以从语义上推测出评论的主题,如从数码相机评论"While light, it will not easily fit in pockets."中可以推测出 2 个主题词"weight"和"size". 以下内容将分为 2 个部分来梳理这 2 类主题的提取方法.

1.1 显式主题的提取方法

显式主题的提取方法主要分为 2 类:基于语言规则的方法和基于概率模型的方法.

基于语言规则的方法中,文献[8]所提出的方法最为经典. 该方法首先利用自然语言处理工具 NLProcessor 2000 对评论数据进行分词和词性标注(part-of-speech tag, POS tag);再使用 Apriori 算法挖掘评论数据中频繁出现的词汇和短语(即集合大小小于 3 的频繁项集),用这些频繁项集构建候选主题集合;之后,对候选主题集合进行过滤. 该文提出了 2 种过滤方法.

1) 紧密度过滤. 该方法用于判断候选主题集合中的短语是否紧密,若不紧密则被过滤掉. 判断标准有 2 个:①在一条评论语句 S 中,若存在候选集中的一个短语,则计算组成该短语的 2 个词汇在语句 S 中的距离,若距离小于 3 个词则称该短语在语句 S 中紧密;②在整个数据集中,若满足标准 1 的语句至少有 2 条,则称该词组是紧密的. 因此,不满足标准 2 的短语会被过滤掉. 例如 3 条评论语句:"The *battery life* is long.","The phone has long *battery life*.","The *battery* is good enough, but I spent whole *life* to get used to the huge screen."其中,候选主题词"*battery life*"在第 1,2 句是紧密的. 第 3 句话中,"*battery*"和"*life*"的距离大于 3,不满足标准 1. 如果在整个评论数据集中同时出现上述 3 句话,则"*battery life*"是一个紧密词组,因为满足"*battery life*"紧密条件的句子在整个数据集中出现了 2 次. 该方法目的是过滤掉那些频繁共现但无法构成词组的词集合.

2) 冗余过滤. 该方法定义了一个判定值 p-support. p-support 指满足下列 2 个条件的评论语句数量:①该语句中出现的主题词或短语是名词或者名词词组;②该语句中不能出现任何词组是该主题词的超集(superset). 我们通过举例说明 p-support 如何取值. 例如,候选集中词汇"manual"出现在 10 个句子中,它的超集"manual mode"和"manual setting"也出现在评论数据中,2 个词组出现在不同评论语句里的次数分别为 4 次和 3 次,且 2 个词组没有出现在同一句话里. 那么,"manual"的 p-support 值为 10−4−3=3. 论文中将 p-support 的阈值设为 3,候选集中 p-support 值小于 3 的词会被过滤掉. 该方法主要目的是过滤掉非名词词汇和词组.

在过滤步骤之后,文献[8]作者还提出了一种非频繁主题的提取方法. 该文作者通过分析数据发现了如下规律:评论者评价频繁主题所用到的情感词与其评价非频繁主题所用到的情感词相同. 例如,"*Red eye* is very *easy* to correct."和"The camera comes with an excellent *easy* to install *software*"这 2 句话都用到了情感词"*easy*",分别评价 2 个不同的主题"*Red eye*"和"*software*". 其中,"*software*"为评论中频繁出现的主题,"*Red eye*"则是非频繁主题,情感词"*easy*"将两者联系了起来. 通过"频繁主题"→"情感词"→"非频繁主题"的挖掘模式可以获得更多非频繁主题. Zhuang 等人[9]提出利用主题和情感词之间的关系来提取主题. 该方法首先利用语法依赖关系解析工具(如 MINIPAR[10])得到如图 4 所示的语法依赖关系图,图 4 中例句为"This *movie* is not a *masterpiece*."其中,"*movie*" 和"*masterpiece*"分别被标注为主题和情感词. 图 4 中的依赖关系为"NN-nsubj-VB-dobj-NN". 其中,"NN"和"VB"是词性标签,"nsubj"和"dobj"是依赖关系标签. 文献[9]作者通过大量训练数据来捕捉这种依赖关系,再利

用这种依赖关系提取“主题-情感词”对儿,从而得到评论语句的主题.

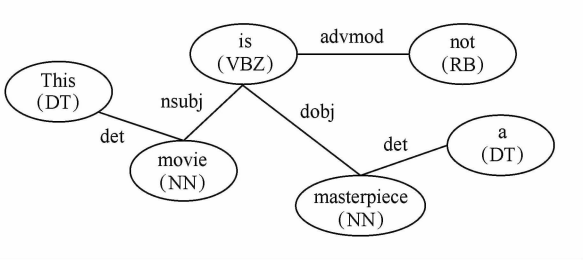


Fig. 4　Grammatical dependency graph on a review sentence[9]

图 4　评论语法依赖关系图

也有研究工作提出使用基于概率模型的方法来进行主题提取. Jin 等人[11]提出了一种词汇化隐马尔可夫模型(lexicalized HMM)来提取评论主题. 不同于传统隐马尔可夫模型,该文作者将词性标注、词典等特征融入了 HMM. Lu 等人[12]则提出了一种基于概率的隐含情感分析方法(probabilistic latent semantic analysis, PLSA)来提取短评论中的主题. 该文作者认为短评论的组成要素有 2 个:1)修饰词 modifier;2)被修饰对象 head term. 因此,一条评论可以表达为〈head term,modifier〉的形式,如〈quality, good〉,〈ship,fast〉等. 一般来说,被修饰词为主题,修饰词为情感词. 文献[12]作者利用 head term 与 modifier 之间的共现信息将这种表达形式融入到 PLSA 模型中. 该文中提出的基于 PLSA 的方法将 k-unigram 语言模型定义为 k 个主题模型(topic model),每个模型都是 head term 的多项式分布,用来捕捉对应的主题;再利用 EM 算法估计模型参数. 其他研究工作还使用到了条件随机场(conditional random fields)[13]、LDA 模型(latent Dirichlet allocation)[14-16].

1.2　隐式主题的提取方法

隐式主题的表达形式呈现多样化特点. 其中,形容词表达是最常见的一种形式[17]. 在评论数据中,一个形容词常用来评价某一特定的主题. 例如,“heavy”常用于评价“weight”,“beautiful”常用来评价“look”或“appearance”. 但是,隐式主题的提取工作的难点在于:对于不同领域的评论,相同的文字表达形式(如形容词)可能会指代不同的主题. 例如,“heavy”在数码相机评论“the camera is too heavy”指代“weight”,而在一条微博“Alas! The heavy day!”中则指代“weather”. 因此,如何捕捉文字表达与隐式主题之间的对应关系成为隐式主题提取方法的关键. 前人研究工作中,多数研究者都尝试捕捉这种隐含关系. Hai 等人[18]提出一种两步骤的方法来挖掘评论中的隐式主题:步骤 1 利用关联规则挖掘方法挖掘评论集中频繁共现的情感词和主题词,以情感词作为条件、主题词作为结论生成关联规则[情感词,主题词];步骤 2 对步骤 1 生成的关联规则[情感词,主题词]中的主题词进行聚类,形成多个主题词簇. 将主题词簇与情感词再次组合形成新的关联规则[情感词,主题词簇]. 对于给定的情感词,该方法能够找到对应的主题词簇,并将该簇中最有代表性的主题词作为所要提取的隐式主题. Su 等人[19]则提出一种聚类方法,如图 5 所示. 图 5 中,实线左侧为主题词或短语,右边为情感词. 该方法先分别对实线两侧词汇进行相似度聚类,再利用互增强关系(mutual reinforcement principle)来挖掘主题词或

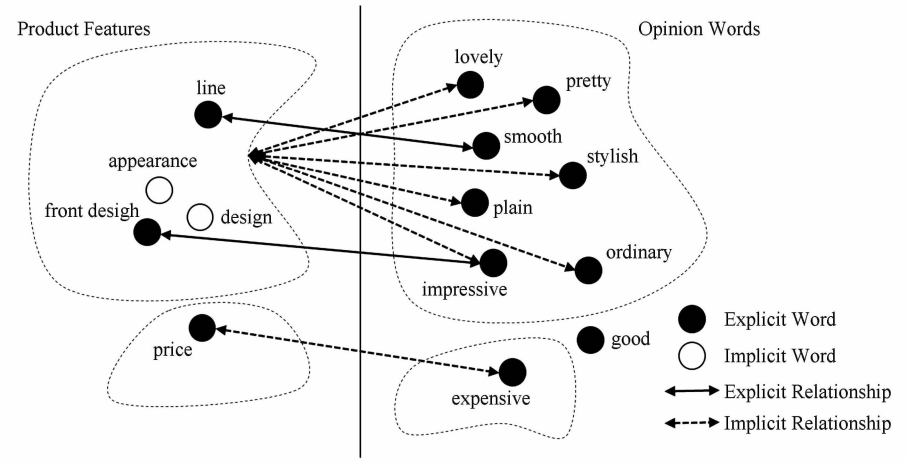


Fig. 5　The cluster-based method for apect words' extraction[18]

图 5　基于聚类的主题词提取方法

短语与情感词的对应关系. 当句子只有情感词时,通过挖掘出的关系来找到最可能的主题词.

总结上述主题提取方法的研究工作. 针对显式主题提取的研究工作中,基于语言规则的方法在同一领域内具有较强的适用性,但是推广能力有限,同一规则不一定适用于其他领域. 此外,语言规则设计需要大量的数据分析和严谨的规则设定. 基于概率模型的方法具有一定的推广能力,但在大规模评论数据上的效率较低. 隐式主题的提取难度较大,关键挑战在于如何准确捕捉文字表达与隐含主题之间的映射关系.

2 传统情感分类方法

情感分类(sentiment classification)是情感分析技术的核心问题,其目标是判断评论中的情感取向. 按区分情感的粒度可分成 2 种分类问题:1)正/负(positive/negative)二分类或者正面/负面/中立(positive/negative/neutral)三分类;2)多元分类,如对新闻评论进行"乐观"、"悲伤"、"愤怒"、"惊讶"四元情感分类[20],对商品评论进行 1 星~5 星五元情感分类[21]等. 第 1 类分类问题因为更具一般性而受到广泛关注. 以下将着重介绍针对第 1 类分类问题的方法. 主流情感分类方法按技术路线主要分为 5 类:1)基于词典的方法;2)基于机器学习的方法;3)词典与机器学习混合的方法;4)基于弱标注信息的方法;5)基于深度学习的方法. 以下,我们将介绍这 5 类方法中的经典工作.

2.1 基于词典(Lexicon-based)的情感分类方法

基于词典方法的核心模式是"词典+规则",即以情感词典作为判断评论情感极性的主要依据[22],同时兼顾评论数据中的句法结构,设计相应的判断规则(如 but 从句与主句情感极性相反). 文献[4-5,23]是基于词典的情感分类方法中最具代表性的工作. 文献[23]中,Turney 认为评论中包含形容词或副词的词组是判定整条评论情感极性的依据. 该文提出分别计算待判定词汇与"excellent"以及待判定词与"poor"之间的互信息,然后对两者求差得出词组的情感分值:

$$SO(phrase)=PMI(phrase,\text{"excellent"})-PMI(phrase,\text{"poor"}),\quad(1)$$

其中,$SO(phrase)$ 为词组(phrase)的情感分值;PMI 为互信息,利用词之间的共现关系计算得到. 计算整条评论中所有词组互信息差值的均值,将该均值作为整条评论的情感分值. 情感分值的正负和大小分别表示评论的情感极性和强弱. 对一条评论的计算结果如表 1 所示,该条评论的情感分值大于零,因此被判断为正面极性,分值越大情感极性越强. 论文实验数据共计 410 条评论,横跨手机评论、电影评论等不同领域. 实验结果显示,该方法实现了最低 65.83%、最高 84.0%的分类准确率.

Table 1 A Processed Review Judged to be Positive by Forlum (1) [23]

表 1 根据式(1)判断为正面情感的一条评论处理后的结果[23]

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
online experience	JJ NN	2.253
low fees	JJ NNS	0.333
local branch	JJ NN	0.421
⋮	⋮	⋮
Average Semantic Orientation		0.322

文献[4]中,该文作者认为评论中形容词的极性是判定评论情感极性的主要指标,提出将形容词(如"good","bad"等)作为情感词建立情感词典,再根据词典中情感词的极性来判断评论的情感极性. 该文提出通过语义词网络 WordNet 中形容词的近义词集和反义词集来判定评论中的形容词极性. 如图 6 所示,实线箭头表示近义词关系,虚线箭头表示反义词关系. 假设已知 WordNet 网络中任何一个词的情感极性,便可以利用网络中的近义词/反义词关系获取更多词汇的情感极性,进而建立起相应的情感词典. 情感极性关系为:互为近义词关系的词汇具有相同情感极性,互为反义词关系的词汇具有相反情感极性. 该方法具体步骤如下:1)从评论中统计出最频繁出现的 n 个形容词(该文中实验取值 $n=30$)建立种子集,人工标注种子集中所有词汇的极性;2)在 WordNet 中以种子集中的词汇为源头,根据上述近义词、反义词的情感极性关系,迭代地自动标注其他形容词的情感极性,从而得到 WordNet 形容词情感词典;3)根据该词典和简单规则判别评论的极性. 该文实验数据来自亚马逊购物网站,包含数码相机、DVD 播放器、MP3 播放器及手机 4 类商品评论. 该方法在测试数据集上实现了平均 84.2%的准确率. 文献[5]在文献[4]研究工作的基础上,进一步考虑提出情感词与评论主题词之间的距离对整条评论情感极性的影响. 如式(2)所示,其中,$Score(f)$ 指评论主题词 f 的情感分值;w_i 是该评论语句中除主题词外的所有词汇;$SO(w_i)$ 是词汇 w_i 的情感极性值,可

查询情感词典获得,若为正面极性则$SO(w_i)=1$,若为负面极性则$SO(w_i)=-1$;$dis(w_i,f)$指词w_i与主题词f之间的词数目.

$$Score(f)=\sum_{i=1}^{t}\frac{SO(w_i)}{dis(w_i,f)}. \tag{2}$$

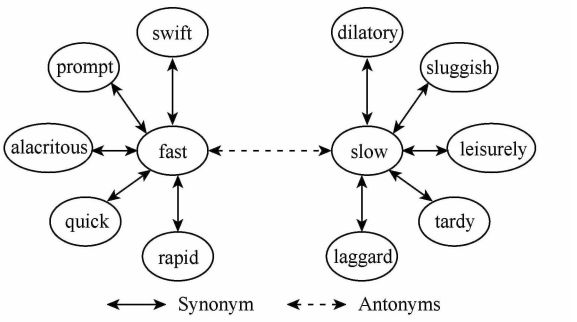


Fig. 6　Bipolar adjective structure[4]

图 6　双极性形容词结构

由式(2)可知,离主题词越远的词对情感极性影响程度越弱;相反,离主题词越近的词对情感极性影响程度越强.此外,文献[5]作者还提出了一些句法规则来调整评论语句的情感极性.该文实验结果表明,该方法的分类性能显著高于同类方法,在抓取的亚马逊商品评论数据[2]上实现了92%的精确度和91%的召回率.

简要总结2.1节工作.基于词典的情感分类方法本质上依赖于情感词典和判断规则的质量,而两者都需要人工设计,如建立词典所使用的初始种子词列表需要人工给定,判断规则的设计则需要人工分析数据集中评论语句的句法结构.因此,这类方法的优劣很大程度上取决于人工设计和先验知识,推广能力较差.如今,新事物借助于社交媒体平台得以快速传播,网络新词、旧词新义的层出不穷使得语言的更新周期变短,词汇的语义不断衍变,从而导致原先情感词典中的部分词汇不再适用于新的语言环境.此外,基于词典的方法无法解决隐含观点(implicit opinions)的挖掘问题,如客观评论语句"I bought the mattress a week ago, and a valley appeared today"指出床垫出现了质量问题,但采用了一种客观性的文字表达形式.这种客观语句是评论的重要表达形式之一[24],包含比主观表达更多的有价值信息,对用户帮助更大.但是,由于句中没有出现任何情感词导致基于词典的方法无法判断情感极性.正如文献[25]中所述,基于词典的方法只能通过专案(ad-hoc)的方式提取显式观点.此外,情感词的情感极性还依赖于上下文内容,同一情感词的情感极性会随着所评价主题的不同发生变化[26].例如在手机评论中,"large"在评价"battery"时表达负面情感,而在评价"screen"时表达正面情感.

以下,我们罗列了较为成熟的开源情感词典:

1) GI(the general inquirer)[27].该情感词典给出了较全面的词条属性.对每一个情感词都给出了对应的情感极性(negative/positive)、词性(如NN,NNs,ADJ等)、客观性指数等属性.

2) LIWC(linguistic inquiry and word count)[28].该情感词典组织结构如表2所示,表格第1列Category表示情感词类别(如第1行为否定词),第2列Example则给出了每个情感类别对应的正则表达式.

Table 2　The Structure of Sentiment Lexicon LIWC

表 2　情感词典 LIWC 的组织结构

Category	Example
Negate	aint, ain't, arent, aren't, cannot…
Swear	arse, arsehole*, arses, ass, asses…
Social	Acquainta*, admit, admits, admitted…
Affect	Abandon*, abuse, abusi*, accept…
⋮	⋮

Notes: * means regular expression.

3) MPQA(multi-perspective question answering).

由Wiebe等人[29-30]建立,词典包含2 718个正面情感词和4 912个负面情感词.每个词条具有5个属性:情感极性(Polarity:positive/negative)、情感强度(Strength:weaksubj/strongsubj)、词个数、词性(Part-of-speech:adj/noun/verb/anypos…)以及是否为过去式(Stemmed:y/n).

4) Opinion Lexicon[4].该词典包含2 006个正面情感词和4 783个负面情感词.其独特之处在于同时包含情感词的标准形式和其他形式,如俚语、拼写错误、语法变形以及社交媒体标记等.

5) SentiWordNet[31].该词典对语义词网络WordNet中所有词汇进行情感极性分类并给出情感极性的量化分数(PosScore/NegScore).

对基于词典的情感分类方法而言,选择最优情感词典也是需要注意的问题.对比同一词汇在不同词典中的一致性得到如表3所示的对比结果.表3中计算的分数表示2个词典的不一致程度.其中,分数的分母代表2个不同词典的交集中的词汇总数,分数的分子代表情感极性不一致的词汇数目.对于不一致的词条则需要根据实际数据进行人工修正,这也是基于词典方法的缺点之一.

Table 3 The Degree of Inconsistency Between Different Sentiment Lexicons

表 3 不同情感词典的不一致程度

Sentiment Lexicons	MPQA	Opinion Lexicon	GI	Senti-WordNet	LIWC
MPQA		33/5 402	49/2 687	1 127/4 214	12/363
Opinion Lexicon	33/5 402		32/2 411	1 004/3 994	9/403
GI	49/2 687	32/2 411		12/308	1/204
Senti-WordNet	1 127/4 214	1 004/3 994	12/308		174/694
LIWC	12/363	9/403	1/204	174/694	

2.2 基于机器学习的方法

Pang 等人[32]于 2002 年首次提出使用标准的机器学习方法解决情感分类问题. 该文针对文档层次(document-level)的二元情感分类问题,即判断整条评论的极性. 该工作实验对比了不同特征组合与不同机器学习方法在电影评论情感分类问题上的效果. 实验结果如表 4 所示,实验结果表明,相比于朴素贝叶斯分类(naive Bayes, NB)和最大熵模型(maximum entropy, ME),支持向量机(support vector machine, SVM)的分类效果更好,它与 Unigrams 特征结合进行情感分类的准确率达到了 82.9%.

Table 4 Performance Comparison of SVM, NB and ME[32]

表 4 SVM, NB and ME 方法情感分类准确率对比[32]

Features	NB	ME	SVM
unigrams	**0.787**		0.728
unigrams	0.810	0.804	**0.829**
unigrams+bigrams	0.806	0.808	**0.827**
bigrams	0.773	**0.774**	0.771
bigrams+POS	0.815	0.804	**0.819**
adjectives	0.770	0.777	0.751
top 2633 unigrams	0.803	0.810	**0.814**
unigrams+position	0.810	0.801	**0.816**

Notes: The bold values mean the best accuracy value among the three classifiers.

此后,多数机器学习方法的研究工作将重点放在如何设计更多有效的分类特征上. 研究者尝试了不同类特征组合在情感分类上的效果,Dave 等人[33]对比了 Unigrams 特征和 Bigrams 特征在情感分类问题上的效果,实验结果如表 5 所示. 该实验证明了相同条件下使用 Bigrams 特征比使用 Unigrams 特征的分类准确率要高.

Table 5 Performance Comparison on Unigrams and Bigrams[33]

表 5 使用 Unigrams 特征和 Bigrams 特征的分类结果比较[33]

Method	Test 1		Test 2	
	Unigrams	Bigrams	Unigrams	Bigrams
Baseline	85.0	88.3	82.2	84.6
SVM	81.1	87.2	84.4	85.8

Mullen 和 Collier[34]在前人研究工作基础上,设计出更优秀的特征组合,并利用支持向量机分类器进行情感极性分类. 该方法所提出特征组合中包含特征:词汇的互信息特征(pointwise mutual information, PMI)[23]、Osgood 语义区分度(Osgood semantic differentiation with WordNet)[35]、主题相似度特征(topic proximity)和句法关系特征(syntactic-relation features)[22]. 其中,词汇的 Osgood 语义区分度包括 3 个指标:强度(强或弱)、活跃度(积极或消极)、评估值(好或坏),这 3 个指标来源于查理斯·奥斯古德语义区分度理论(Charles Osgood's theory of semantic differentiation)[36]. 为了提取词汇的 Osgood 语义区分度特征,作者利用 WordNet 来传播这 3 个指标,其思想与文献[4]中生成情感词典的思想相似:利用 WordNet 中少量已知词汇的 Osgood 语义区分度指标来推断更多词汇的 Osgood 语义区分度指标. 该工作结合上述多种特征对评论进行情感分类,实验数据采用文献[23]中的数据集. 实验结果如表 6 所示. 从实验结果上看,该文中提出的基于

Table 6 Performance Comparison of Different Features[34] (SVM with Linear Kernel)

表 6 使用不同特征组合的分类准确率对比[34] (SVM 使用线性核)

Model	3-folds	10-folds
Pang et al. 2012	0.829	
Turney Values only	0.684	0.683
Osgood only	0.562	0.564
Turney Values and Osgood	0.690	0.687
Unigrams	0.828	0.835
Unigrams and Osgood	0.828	0.835
Unigrams and Turney	0.832	0.851
Unigrams, Turney and Osgood	0.828	0.851
Lemmas	0.841	0.857
Lemmas and Osgood	0.831	0.847
Lemmas and Turney	0.842	0.849
Lemmas Turney Osgood	0.838	0.845
Hybrid SVM(Turney and Lemmas)	0.844	**0.860**
Hybrid SVM(Turney/Osgood and Lemmma)	**0.846**	**0.860**

Notes: The bold values mean the best accuracy value among the different methods.

混合特征的分类方法 Hybrid SVM(Turney/Osgood and Lemmas)在分类准确率上明显优于使用其他特征组合的分类方法.

Saleh 等人[37]在 3 个不同数据集上进行了 27 组实验,分别测试了不同特征选择方法对情感分类效果的影响.实验选择支持向量机作为分类模型,数据集有 3 个:1)Pang 和 Lee 在文献[38]中的数据集;2)Taboada 和 Grieve 在文献[39]中的数据集;3)SINAI 数据集中的数码相机子集,实验采用 10 折交叉验证(10-FCV)方法来测试分类器的性能.3 组情感分类实验结果如表 7～9 所示:

Table 7　Performance Comparison of Different Features on Pang Dataset[37]

表 7　Pang 数据集上使用不同特征组合的分类结果对比

Metrics	TF-IDF+Unigram	TF-IDF+Bigram	TF-IDF+Trigram
Precision	0.825 4	0.837 2	0.840 1
Recall	0.843 0	0.857 0	0.848 0
*F*1	0.833 6	0.846 1	0.847 9
Accuracy	0.832 0	0.844 5	0.846 5
Kappa	0.664 0	0.689 0	0.693 0

Table 8　Performance Comparison of Different Features on Taboada Dataset[37]

表 8　Taboada 数据集上使用不同特征组合的分类结果对比[37]

Metrics	TF-IDF+Unigram	TF-IDF+Bigram	TF-IDF+Trigram
Precision	0.717 9	0.729 3	0.723 5
Recall	0.700 0	0.735 0	0.750 0
*F*1	0.705 8	0.729 5	0.733 7
Accuracy	0.710 0	0.730 0	0.732 5
Kappa	0.420 0	0.460 0	0.465 0

Table 9　Performance Comparison of Different Features on SINAI Dataset[37]

表 9　SINAI 数据集上使用不同特征组合的分类结果对比[37]

Metrics	TF-IDF+Unigram	TF-IDF+Bigram	TF-IDF+Trigram
Precision	0.920 6	0.921 7	0.920 2
Recall	0.989 5	0.987 1	0.987 7
*F*1	0.952 0	0.953 2	0.952 7
Accuracy	0.913 0	0.915 1	0.914 1
Kappa	0.488 0	0.500 0	0.489 0

实验结果表明,使用 TF-IDF 和 Trigrams 的特征组合在 Pang 数据集上实现了最高 84.65%的分类准确率;使用 TF-IDF 和 Trigrams 特征组合在 Taboada 数据集上实现了最高 73.25%的分类准确率;使用 TF-IDF 和 Bigrams 特征组合在 SINAI 数据集上实现了最高 91.51%的分类准确率.

Zhang 等人[40]使用朴素贝叶斯(NB)和支持向量机(SVM)分类器对酒店评论进行情感分类.作者对评论数据分别提取 Unigrams,Bigrams 和 Trigrams 特征,如表 10 和表 11 所示.上述特征采用 2 种表达方式:二值(binary)和频率(frequency).二值表达用 0 或 1 表示一个特征是否出现在评论文档中;频率表达则是统计特征在评论文档中的出现次数.实验测试了使用不同数量 *n*-gram 特征进行情感分类的准确率,结果如表 12 所示,其中,表格第 1 列为不同类别特征,*n*-gram 和 *n*-gram_freq 分别表示基于二值表达的 *n*-gram 特征和基于频率表达的 *n*-gram 特征,NB 和 SVM 对应 2 种分类器,表格中分类结果由 2 部分组成:括号外数字为情感分类准确率,括号内数字为特征数目.从结果可以看出,使用 NB 和

Table 10　*n*-gram Feature Selected from Binary-Based Documents[40]

表 10　基于二值表达方式的 *n*-gram 特征[40]

Top-*k*	Unigrams	Bigrams	Trigrams
1	滑	好味	唔會再
2	味	好香	好好味
3	差	好好	DISH 同 DISH
4	香	唔錯	唔新鮮
5	濃	DISH 好	都好好
6	正	幾好	DISH 好香
7	脆	唔掂	DISH 好好
8	甜	點知	知所謂
9	错	难食	不知所

Table 11　*n*-gram Feature Selected from Frequency-Based Documents[40]

表 11　基于频率表达方式的 *n*-gram 特征[40]

Top-*k*	Unigrams	Bigrams	Trigrams
1	好	好味	唔知點
2	味	好香	DISH 係有
3	差	好好	話蘇絲
4	滑	唔錯	之極之
5	香	DISH 好	佢個同
6	濃	幾好	野只是
7	正	唔掂	！跟住
8	脆	態度	得幾件

Table 12 The Best Performance on Different Number of Features[40]

表 12 不同特征数目对应的最高情感分类准确率[40]

Features	NB(Number of n-gram Features)	SVM(Number of n-gram Features)
Unigram	0.9317(400)	0.9076(600)
Unigram_freq	0.8883(150)	0.8633(1550)
Bigram	**0.9567(900−1100)**	0.9067(1900−2100)
Bigram_freq	**0.9483(1300)**	**0.9483(1950)**
Trigram	0.9533(1700,1800)	0.8250(2250)
Trigram_freq	0.9417(1200)	0.9017(2550)

Notes: The bold values mean best accuracy, and the values in the parentheses means the number of n-gram features.

基于二值的 Bigram 特征在特征数目落入 900 至 1 100 区间时,能够达到最高 95.67%的分类准确率(accuracy). 使用 SVM 和基于频率的 Bigram 特征在特征数目为 1 950 时,能够达到最高分类准确率 94.83%.

简要总结上述基于机器学习技术的情感分类研究工作:

1) 特征工程(feature engineering)是此类研究工作的核心. 情感分类任务中常用到的特征有 n-gram 特征(unigrams, bigrams, trigrams)、Part-of-Speech(POS)特征、句法特征[41]、TF-IDF 特征等. 然而,这类方法仍旧依赖于人工设计,研究过程中容易受到人为因素的影响. 此外,人工设计的特征在不同领域的推广能力较差,在某一领域表现优秀的特征集不一定在其他领域也同样优秀[42].

2) 基于机器学习的情感分类方法多使用经典分类模型如支持向量机、朴素贝叶斯、最大熵模型等. 其中,多数分类模型的性能依赖于标注数据集的质量[43],而获取高质量的标注数据则需要耗费大量的人工成本.

2.3 词典与机器学习混合的方法

部分情感分类的研究工作将基于词典的方法和基于机器学习的方法相融合. 这类混合方法的思路主要分为 2 种:1)将“词典+规则”视作简单的分类器,然后融合多种不同分类器进行情感分类;2)将词典信息作为一种特征与现有特征(如句法特征、POS 特征等)进行组合,然后选择最优的特征组合进行情感分类. 以下,我们对这类方法中的代表性工作进行简要介绍.

Prabowo 等人[44]提出了一种基于规则的分类器(rule-based classifier, RBC)和支持向量机分类器(SVM)[32]混合的方法,解决文档级别的情感分类问题. 其中,RBC 设定了 3 种规则:

1) 基于情感词的判定规则[情感词]→[+/−]. 其中,“+/−”表示“正面情感/负面情感”. 该规则根据出现在评论中的情感词的极性来判断整条评论的情感极性,情感词的极性通过查询 GI 词典[27]获得. 具体实例如[excellent]→[+],[absurd]→[−].

2) 基于主题词的判定规则,如[# more expensive than?]→[−]. 其中,“#”表示主题词,“?”表示被比较的对象. 该规则主要针对包含多主题词的对比句的情感分类问题. 例如“A is more expensive than B”,若主题词为 A,则该评论的情感极性为负,即[# more expensive than?]→[−];若 B 为主题词,则评论的情感极性为正,即[?more expensive than #]→[+].

3) 基于互信息的判断规则[PMI of review]→[+/−]. 该规则基于 Turney 的研究工作[23],计算整条评论中所有词组互信息差值的均值,根据均值的正负来判断评论的情感极性.

上述 3 种判定规则中,基于情感词的判定规则和基于互信息的判定规则属于基于词典的情感分类方法. SVM 采用文献中[23]的方法,该方法属于机器学习方法. 该文作者将上述 2 种分类器混合进行情感分类:先使用 RBC 进行分类,若得到分类结果则返回该结果;若没得到分类结果,则使用 SVM 分类器进行情感分类. 实验数据集来自文献[38],该数据集包含电影、商品和社交网站 3 个不同领域的评论数据. 该混合方法在实验数据集上达到了 90.45%的准确率.

Fang Ji 等人[45]提出将词典信息融入到支持向量机分类器中,解决语句级别的情感分类问题. 该方法中,作者将评论语句中的名词、动词、形容词和副词作为该语句的 Unigrams 特征词. 例如,一条评论语句“The case is rigid so it gives the camera extra nice protection.”通过判断词性可以抽取句中的 Unigram 特征词序列:〈case, rigid, give, camera, extra, nice, protection〉. 若 Unigrams 特征词序列中出现了包含于 MPQA[29]中的情感词,则将该情感词的极性词(positive 或 negative)插入到特征词序列中. 例如,上述词序列中“nice”的情感极性为“positive”,则将“postive”插入到语句的词序列中得到〈case, rigid, give, camera, extra, nice, protection, positive〉. 若词序列中出现多个情感词,仍按上述方法在 Unigrams 特征词序列中插入相应的极性词.

然后,利用 Bag-of-Words 模型将特征词序列转化成对应的特征向量.特征向量中的元素代表词序列中词汇出现的次数.例如,词序列中出现了 2 个“positive”和 2 个“negative”,则对应的特征向量中“positive”和“negative”位置都为 2.通过这种方法将词典信息融入到语句的特征向量中,再使用支持向量机分类器进行情感分类.不同于上述 Fang Ji 等人的工作,Abbasi 等人[46]将研究重点放在特征工程上,提出了一种新的特征选择技术,称为特征关系网络(feature relation network, FRN).该技术融合了规则特征、n-grams 特征、句法特征等多种特征,达到了较高的分类性能.

综上所述,尽管混合方法改进了基于词典和基于机器学习方法的性能,但本质上并没有从特征设计和词典构建中解放出来.

2.4　基于弱标注信息的方法

由于人工标注训练数据费时费力,近年来情感分析领域的研究者开始考虑从用户产生的数据中挖掘有助于训练情感分类器的信息,如评论的评分(ratings)、微博中的表情符号等.由于互联网用户的“标注”行为没有统一标准,具有较大随意性,所产生的标注信息存在噪声(如高评分的负面评论),因此我们将这种标注信息称为弱标注信息.弱标注信息能够在一定程度上反映评论的情感语义,因此很多研究者尝试在情感分类研究工作中引入弱标注信息.

Qu 等人[47]提出使用包含评分信息的评论数据作为弱标注数据训练概率模型来解决语句的情感分类问题.Täckström 等人[48]提出利用条件随机场(conditional random fields, CRF)模型结合文档标签(即评论评分)和语句标签来解决情感分类问题.但是,上述 2 种方法都还依赖于人工设计的特征.

Maas 等人[49]提出在概率模型中引入评论评分信息来学习反应情感属性的词向量,然后用一篇文档中所有词的词向量平均值作为特征学习情感分类器.Tang 等人[50]提出利用推文中的表情符号(如“:)”表示开心)作为情感标签来训练一种 C&W 模型[51]的变种,从而学习出反映情感属性的词向量.对于给定的一篇推文,对其词的词向量进行最大、最小和平均池化(pooling)操作,进而获得该推文的特征表达向量.最后,利用该特征表达向量进行情感分类.上述 2 种方法都没有考虑如何减轻弱标注信息中的噪声影响.此外,尽管这 2 种方法能够自动生成用于情感分类的特征表达,但只是简单的池化操作,并不能很好地捕捉文本到高层语义的复杂映射函数.而捕捉这种复杂映射函数正是深层神经网络的专长.接下来我们将介绍基于深度学习的情感分类方法.

2.5　基于深度学习的方法

自 2006 年无监督逐层学习技术(greedy layer-wise training)[52]的提出,深度学习逐渐成为机器学习领域的热门研究方向.深度神经网络模仿人脑的分层组织结构,具有指数倍于浅层计算模型的表达能力,理论上能够更好地捕捉从数据本身到高层语义的复杂映射函数.目前,深度学习模型在不同应用问题上的推广能力得到了一定验证[53],如图像识别[54-57]、语音识别[58-60]、药物分子活性预测[61-62]等.更令人惊喜的是,深度学习还在很多自然语言理解任务上得到了令人满意的效果,如智能问答系统[63]、自然语言翻译[64-65]、情感分析[50,66-72]等.其中,情感分析作为自然语言理解的重要应用之一,也受到了越来越多研究者的广泛关注.

正如第 2 节第 1 段所述,情感分析的核心在于解决情感分类问题.因此,很多研究工作尝试使用深度学习技术来解决情感分类问题.现有研究工作中,针对情感分类问题的深度学习方法有 2 个主要步骤:1)从海量评论语料中学习出语义词向量(word embedding);2)通过不同的语义合成(semantic composition)方法用词向量得到所对应句子或文档的特征表达[73].现有合成方法主要基于语义合成性原理(principle of compositionality)[74],该原理指出:长文本(如一个句子、一篇文档)的语义由它的子成分(如词汇、短语)的语义按不同规则组合而成.本质上讲,语义合成就是利用原始词向量合成更高层次的文本特征向量.

Bespalov 等人[66]提出通过潜在语义分析(latent semantic analysis)初始化词向量,再用带权重的 n-gram 特征进行线性组合从而得到整篇文档的情感特征向量.Glorot 等人[67]提出利用除噪堆叠自编码器(stacked denoising autoencoder, SDA)来解决海量评论数据情感分类中的领域适应性问题(domain adaptation)[75].自编码器是一种通过重建自身输入进行模型优化的特征学习器.除噪堆叠自编码器是 Bengio 等人提出的堆叠自编码器(stacked autoencoder)[76]的一种扩展算法.作者用无监督方法训练该深度模型去捕捉不同领域数据之间的共性表达,在 22 个不同类别的商品评论数据上进行模型的推广能力测试.实验结果显示,与同类方法相比,

SDA 方法达到了较低的平均传输推广误差(averaged transfer generalization errors)10.9%. 该文献表明,基于除噪堆叠自编码器的深度学习系统可以通过无监督方法提取不同领域评论文本的潜在共性特征,从而有效地解决跨领域情感分类问题. Socher 等人在 2011—2013 年间的研究工作中[68-70]提出了一系列基于递归神经网络(recursive neural network, RecNN)的分类模型来解决情感分类问题. RecNN 模型通过递归计算来学习变长语句的特征向量. Kim[71]则使用卷积神经网络(convolutional neural network, CNN)来解决情感分类问题. 实验结果表明,卷积神经网络的分类性能明显优于递归神经网络. 对于卷积神经网络模型的研究,Kalchbrenner 等人[72]提出了一种新颖的卷积神经网络模型,该模型特点在于采用了动态 k-max 池化(dynamic k-max pooling)操作和多层卷积神经网络层相结合的结构. 不同于上述工作,有研究者提出使用序列模型如循环神经网络(recurrent neural network, RNN)来解决情感分类问题,例如文献[77]中,作者提出使用长短期记忆网络(long short term memory, LSTM),将评论语句建模成词序列来解决情感分类问题. 与 CNN 相比,LSTM 可以捕捉到评论语句中的长依赖关系(long-term dependencies),可以从整体上"理解"评论的情感语义.

相比于传统机器学习方法,深层神经网络的表达能力有了质的飞跃,并摆脱了特征工程的束缚. 利用语义合成性原理通过不同深度模型将低层词向量合成高层文本情感语义特征向量,从而得到文本的高层次情感语义表达,有效提升了模型的推广能力. 但是,上述针对文本情感分类问题的深度学习方法仍然在较大程度上依赖于有标注训练数据,即依赖于有监督学习方法来训练深层神经网络[50-51,78]. 大规模的训练数据是深度学习成功的关键. 然而,要获得有标注训练数据,便要耗费大量的人力成本. 通过人工标注方式获得大规模有标注训练数据的成本十分高昂. 尽管传统的无监督预训练技术能够利用无标注数据训练神经网络,但是该方法只有在数据分布与要预测的语义之间具有较强相关性时才能很好地发挥作用[79]. 但是,文本中的词共现信息通常与所要预测的情感语义没有很强的相关性[49]. 因此,缺乏大规模的训练数据已成为深度学习在情感分类问题上的瓶颈.

3 基于弱监督深度学习的情感分类

传统的情感分类方法中,基于词典的方法依赖词典设计,基于机器学习的方法则倚重特征设计,两者要求相关人员具有较高的领域知识和研究经验,且方法的推广能力较差. 近年来深度学习在情感分类问题上表现优秀. 但是,缺乏标注的训练数据是深度学习的瓶颈问题. 互联网用户产生的弱标注信息给我提供了突破瓶颈的新思路. 由于弱标注信息与评论情感语义具有一定的相关性,因此可以用于训练深度模型来解决情感分类问题.

为此,我们提出了一种利用深层神经网络和弱标注信息解决情感分类问题的新思路:利用互联网上产生的海量弱标注评论数据作为训练集训练深度模型进行情感分类任务. 但是,使用弱标注数据的挑战在于如何尽量减轻数据中噪声对模型训练过程的影响. 针对该挑战,我们设计了一种弱监督深度学习框架(weakly-supervised deep learning, WDE)来解决文本情感分类问题. 其总体框架如图 7 所示. 该框架以评论语句 s 作为输入,抽取低层次定长的特征向量表达,并在隐含层引入了上下文信息. 训练方法采用"弱监督预训练+有监督学习微调"的思路来训练深层网络模型. 框架的核心是弱监督预训练方法,该方法利用弱标注数据预训练出一个能够捕捉文本语句情感语义分布的嵌入空间(embedding layer),如图 7 所示,使得具有相同情感极性的语句互相接近,而具有不同情感极性的语句互相远离. 得到较好的嵌入空间之后,再增加分类层(classification

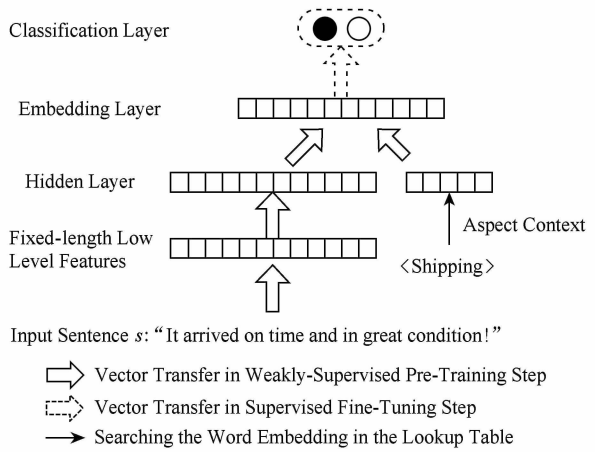


Fig. 7 Network architecture in general for sentence sentiment classification

图 7 语句情感分类的通用网络结构

layer),如图 7 所示,用少量有标注数据训练最终分类模型.

我们将该框架应用在商品评论情感分类问题上.通过分析大量商品评论数据,我们发现:一条商品评论的评分在一定程度上能够反映这条评论的情感取向.因此,我们提出利用一条评论的评分信息作为该评论中所有语句的情感标签来训练深度模型.但是,商品评论的评分是一种弱标注标签,评论中可能存在实际情感语义与评分不一致的情况,如一条 5 星级的评论中仍然存在负面评论语句,具体实例如图 8 所示,图 8 中框内语句为负面评价语句.

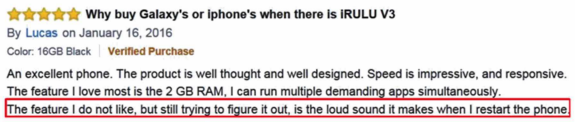


Fig. 8　A negative sentence in a 5-stars review

图 8　一条 5 星评论中的负面语句

我们将这种情感语义与评分不相符的标注数据称为噪声.为了确定噪声的影响,我们人工标注了一些语句(人工标注集在实验部分有详细说明),并统计噪声数据在人工标注语句中的占比情况,即:人工标注数据集中,属于 4,5 星评论的语句中人工标注为负面评价的语句所占百分比,以及属于 1,2 星评论的语句中人工标注为正面评价的语句所占百分比.统计结果如图 9 所示.在人工标注数据中,上述 2 个占比值都超过了 10%,总体上看,噪声占总量的 13.4%.这表明弱标注数据中存在一定噪声,直接作为有标注信息会影响模型的训练效果,因此无法直接用于深度模型的有监督训练.

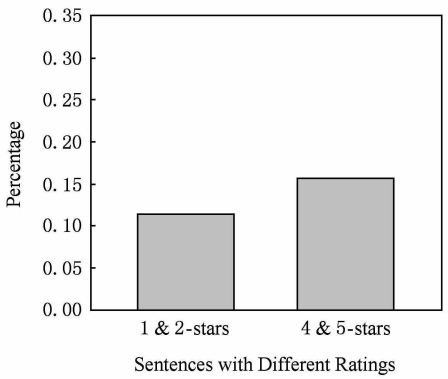


Fig. 9　Percentages of wrong-labeled sentences by ratings in our labeled review dataset

图 9　标注数据中以评分错标语句的比例

为解决该问题,我们设计了一种抗噪声的预训练方法,即前文所述的"弱监督预训练"方法.该方法的目标是通过预训练得到一个能够捕捉语句情感语义分布的嵌入空间,从弱标注数据中汲取有用信息的同时,避免其对最终分类模型的直接影响.以下内容,我们将按如下顺序组织:1)详细介绍 2 种实现 WDE 框架的深度模型,包括基于 CNN 的深度模型和基于 LSTM 的深度模型;2)具体给出弱监督预训练方法并简要描述有监督微调步骤;3)我们通过实验验证了 2 种深度模型在商品评论情感分类问题上的效果.

3.1　基于 CNN 的深度模型介绍

由于卷积神经网络在语句级别的情感分类问题上表现优秀[71],因此我们选择卷积神经网络作为 WDE 的一种模型实现.该模型称为 WDE-CNN,是文献[51,71]中 CNN 模型的一种变体结构,其结构如图 10 所示.图 10 中,将一条评论语句 s 输入到模型中,$w_1,w_2,\cdots,w_t$ 表示句子中的词语,对每个词语查询词向量列表 $\boldsymbol{X}$ 得到对应的词向量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t$.从而将语句 $s=\langle w_1,w_2,\cdots,w_t\rangle$ 转化为 $\langle \boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t\rangle$.我们使用 Word2Vec 在谷歌新闻语料库上的训练结果[80]来初始化词向量列表,对于不在谷歌新闻训练结果中的词汇则随机初始化.

卷积层(convolutional layer)包含多个卷积滤波器,每个滤波器通过滑动能容纳 n 个词汇的窗口进行卷积计算,进而生成局部特征值,计算公式如下:

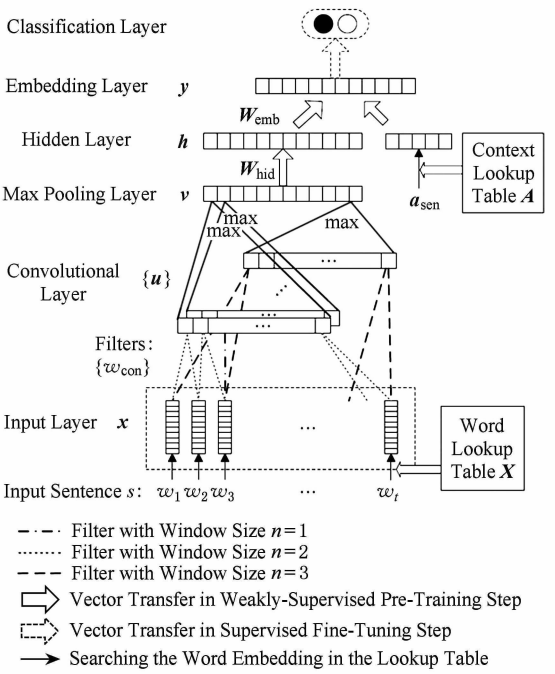


Fig. 10　The CNN network architecture for sentence sentiment classification WDE-CNN

图 10　基于 CNN 的深度模型 WDE-CNN

$$u(i)=f(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_{i:(i+h-1)}+b), \tag{3}$$

其中,$\boldsymbol{x}_{i:(i+h-1)}$表示输入语句位置 i 窗口中词向量首尾相接的向量$(\boldsymbol{x}_i^{\mathrm{T}},\boldsymbol{x}_{i+1}^{\mathrm{T}},\cdots,\boldsymbol{x}_{i+h-1}^{\mathrm{T}})^{\mathrm{T}}$,$u(i)$是某个卷积滤波器在位置 i 的卷积输出,b 是该卷积滤波器的偏置,$f(\cdot)$是非线性激活函数,我们选择的是双曲正切函数(hyperbolic tangent). 在整条输入语句 s 上滑动窗口计算得到一个 $t-h+1$ 维的特征向量 $\boldsymbol{u}=(u(1),u(2),\cdots,u(t-h+1))$(即卷积滤波器的特征图(feature map)).

最大池化层(max pooling layer)对向量 $\boldsymbol{u}$ 中所有 $u(i)$求最大值以获取其中最显著的特征值[51]:

$$v(j)=\max_i\{u_j(i)\}, \tag{4}$$

其中,j 表示第 j 个卷积滤波器. 在最大池化层中,最大池化操作提取特征图显著特征的同时还生成了定长的特征向量 $\boldsymbol{v}$.

需要说明的是,词容量为 g 的卷积滤波器本质上是一个特征采集器,用来提取句子的 n-gram 特征. 对输入的 n-gram 匹配其相应的 w 从而得到高层次特征值. 该方法虽然与传统情感分类工作中的特征选择方法[42]相像,但其优势在于特征是自动学习的而非人工设计. 考虑到机器学习方法常使用 unigrams,bigrams 和 trigrams 特征[42],因此我们使用不同词容量的卷积滤波器,设置参数 $n=1,2,3$.

隐含层(Hidden layer)与传统神经网络结构相同,所有神经元之间全连接. 该层以 $v(j)$作为输入,通过非线性激活函数得到更高层次的特征向量 $\boldsymbol{h}$. 计算是标准的前向传播(forward)计算:

$$\boldsymbol{h}=f(\boldsymbol{W}_{\mathrm{hid}}\boldsymbol{v}+\boldsymbol{b}_{\mathrm{hid}}), \tag{5}$$

其中,$\boldsymbol{W}_{\mathrm{hid}}$为权重矩阵,$\boldsymbol{b}_{\mathrm{hid}}$为偏置(bias)向量.

隐含层之后为嵌入层(embedding layer),该层输入由 2 个部分拼接而成:隐含层的输出向量 $\boldsymbol{h}$ 和句子 s 的上下文向量 $\boldsymbol{a}_{\mathrm{sen}}$. 在商品评论中,上下文向量是对商品某一具体主题(aspect)的语义表达,如“battery life”是“cell phone”的一个主题. 引入上下文向量的原因在于,相似或相同文字表达对于不同的上下文可能会表现出完全相反的情感极性,例如“the screen is big”和“the size is big”. 与输入层获取词向量方法类似,所有上下文向量都可以从上下文向量列表 $\boldsymbol{A}$ 中查询获得,该列表的初始值由随机初始化获得. 嵌入层的计算为

$$\boldsymbol{y}=f\left(\boldsymbol{W}_{\mathrm{emb}}\begin{bmatrix}\boldsymbol{h}\\ \boldsymbol{a}_{\mathrm{sen}}\end{bmatrix}+\boldsymbol{b}_{\mathrm{emb}}\right) \tag{6}$$

3.2 基于 LSTM 的深度模型

基于 CNN 的深度模型中,卷积滤波器的词容量有限,导致其无法捕捉到句子中的长距离依赖关系. 因此,我们提出一种基于 LSTM 的深度模型实现,称为 WDE-LSTM. LSTM 是 RNN 的一种特殊类型. RNN 网络根据前一次迭代过程的隐层输出和当前的数据输入来更新隐层的状态,使神经元具有了“记忆”功能,可以更自然地处理文本数据. LSTM 则在 RNN 基础上引入了门机制,利用 3 种不同的门函数,即输入门、遗忘门和输出门,来控制记忆的长短. 一个 LSTM 记忆单元在时刻 t 的前向计算过程为[81]

$$\boldsymbol{d}_t=f(\boldsymbol{W}_{\mathrm{iu}}\boldsymbol{x}_t+\boldsymbol{U}_{\mathrm{iu}}\boldsymbol{z}_{t-1}+b_{\mathrm{iu}}). \tag{7}$$

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{\mathrm{ig}}\boldsymbol{x}_t+\boldsymbol{U}_{\mathrm{ig}}\boldsymbol{z}_{t-1}+b_{\mathrm{ig}}). \tag{8}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{\mathrm{fg}}\boldsymbol{x}_t+\boldsymbol{U}_{\mathrm{fg}}\boldsymbol{z}_{t-1}+b_{\mathrm{fg}}). \tag{9}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{\mathrm{og}}\boldsymbol{x}_t+\boldsymbol{U}_{\mathrm{og}}\boldsymbol{z}_{t-1}+b_{\mathrm{og}}). \tag{10}$$

$$\boldsymbol{c}_t=\boldsymbol{i}_t\times\boldsymbol{d}_t+\boldsymbol{f}_t\times\boldsymbol{c}_{t-1}. \tag{11}$$

$$\boldsymbol{z}_t=o_t\times f(c_{t-1}). \tag{12}$$

式(7)~(11)中,$\{\boldsymbol{W}_*,\boldsymbol{U}_*,\boldsymbol{b}_*\}_{*\in\{\mathrm{ig,iu,og,fg}\}}$是模型的参数集合,$\times$表示 2 个向量的元素乘积;$\boldsymbol{d}_t$,$\boldsymbol{i}_t$,$\boldsymbol{f}_t$,$\boldsymbol{o}_t$ 分别表示时刻 t 记忆单元的输入单元、输入门、遗忘门和输出门的输出值;$\boldsymbol{c}_t$ 表示时刻 t 记忆单元的内部状态,$\boldsymbol{z}_t$ 表示时刻 t 记忆单元的输出;$\sigma(\cdot)$是 sigmoid 激活函数,$f(\cdot)$是双曲正切激活函数. 以上述结构 LSTM 作为基本构件,我们设计了基于 LSTM 的深度模型,如图 11 所示:

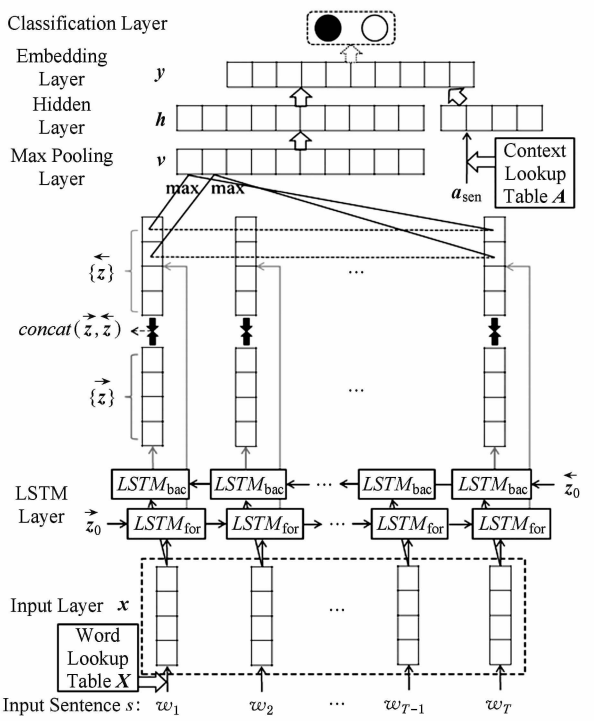


Fig. 11 The LSTM network architecture for sentence sentiment classification (WDE-LSTM)

图 11 基于 LSTM 的语句情感分类网络结构(WDE-LSTM)

模型的输入与 WDE-CNN 模型相同. LSTM 层包含 2 组不同方向的 LSTM 形成了一个双向 RNN 结构[82],该层的操作为

$$\vec{z}_t = LSTM_{\text{for}}(\boldsymbol{x}_t, \vec{z}_{t-1}),$$
$$\overleftarrow{z}_t = LSTM_{\text{bac}}(\boldsymbol{x}_t, \overleftarrow{z}_{t+1}),$$
$$\boldsymbol{z}_t = concat(\vec{z}_t, \overleftarrow{z}_t).$$

其中,$LSTM_{\text{for}}$ 和 $LSTM_{\text{bac}}$ 分别表示前向 LSTM 和反向 LSTM;$\vec{z}_t$,$\overleftarrow{z}_t$ 是它们在时刻 t 的输出向量. $\boldsymbol{z}_t$ 将 $\vec{z}_t$,$\overleftarrow{z}_t$ 拼接起来作为 LSTM 层的输出. 这种双向结构融合了每个词在整个语句中的上下文信息,生成了更为完整的上下文相关的词向量表达. 另外,LSTM 中的记忆单元也可以视为特征采集器,用来识别句子中的有用序列模式.

与 WDE-CNN 相似,在 LSTM 层之后,我们进行了最大池化操作以提取显著特征值,从而得到定长的特征向量 $\boldsymbol{v}$,计算方法如式(13)所示. 最大池化层之上各层结构操作与 WDE-CNN 相同:

$$v(j) = \max_t \{z_t(j)\}. \tag{13}$$

3.3　基于弱标注数据的预训练方法

3.1 节和 3.2 节 2 种模型都采用嵌入层的弱监督预训练方法. 其思想是:通过预训练得到一个可以较好捕捉数据情感语义分布的嵌入空间,之后再使用少量有标注数据学习目标预测函数. 相较而言,前人基于弱标注信息的训练方法通常直接学习目标预测函数. 这种做法会使弱标注信息中的噪声直接影响预测函数. 而我们提出的训练方法避免了直接使用弱标注信息训练目标预测函数,能在一定程度上削弱噪声对预测函数学习的影响.

弱监督预训练方法首先将 5 级评分离散化为二值标签,标签分配的规则是:将评分高于 3 星的评论中的句子标为正面标签 $l(s)=\text{pos}$,将评分低于 3 星的评论中的句子标为负面标签 $l(s)=\text{neg}$,其中 $l(s)$ 表示句子 s 的弱标注标签. 通过标签分配,我们将评论语句划分到 2 个集合里. 弱监督学习的训练目标是让 P 集合和 N 集合各自内部的语句更接近,同时使分别属于 2 个集合的语句相互远离.

一种直观的训练方法是,采样句子对,利用随机梯度下降法(stochastic gradient descent, SGD)对句子对进行操作[83]:如果 2 句话的弱标签相同,则减小它们在嵌入空间中的距离;反之,则增大它们在嵌入空间中的距离. 但是,当采样到噪声时会导致语句向错误类别移动. 为了减弱噪声影响,我们提出一种三元训练准则,每次采样弱标注数据中的 3 条评论语句组成三元组,再利用 Ranking Loss[51] 目标函数对嵌入空间中语句的相对距离进行惩罚,目标函数为

$$\mathcal{L}_{\text{weak}} = \sum_{\langle s_1, s_2, s_3 \rangle} \max(0, \lambda - dst(s_1, s_3) + dst(s_1, s_2)), \tag{14}$$

$$dis(s_i, s_j) = \| y_i - y_j \|_2. \tag{15}$$

式(14)中,λ 表示间隔;$\langle s_1, s_2, s_3 \rangle$ 表示训练集中一组三元采样,其中句子标签 $l(s_1)=l(s_2)\neq l(s_3)$;$dst(\cdot)$ 表示语句在神经网络嵌入层所表示的空间中的欧式距离,该距离的计算方法如式(15)所示. 式(14)目标函数的含义是:让具有相同弱标签的语句 s_1 与 s_2 之间的距离至少比具有相反弱标签的语句 s_1 与 s_3 之间的距离小 λ. 预训练过程中,三元采样方法的具体步骤是:先从 P 或者 N 中随机选其中之一,然后随机抽取该集合中的 2 个语句,再从另一个集合中随机抽取 1 个语句.

图 12 为二元训练准则与三元训练准则的对比图. 图 12 中,圆圈表示弱标签为 *pos* 的样本;三角表示弱标签为 *neg* 的样本;黑色为错标语句(即噪声,实际语义与标签不符的句子);白色为正确标注语句;①,②,③为 3 种包含错标语句的采样实例. 对于二元训练准则(图 12(a)所示),例①、例②中错标语句在训练过程中会向错误类别的语句靠近;例③中,错标语句则远离了其正确类别的语句. 对比来看,对于三元训练准则而言,目标函数确保 s_1 与 s_2 之间的距离至少比 s_1 与 s_3 之间的距离小 λ. 例①中由于同时采样到 2 个错标语句,因此仍然会导致 s_2 和 s_3 向错误类方向移动. 例②和例③中则混合了 2 种情况:一个语句向正确方向移动,而另一个向错误方向移动. 因此,在三元训练准则中,例②和例③

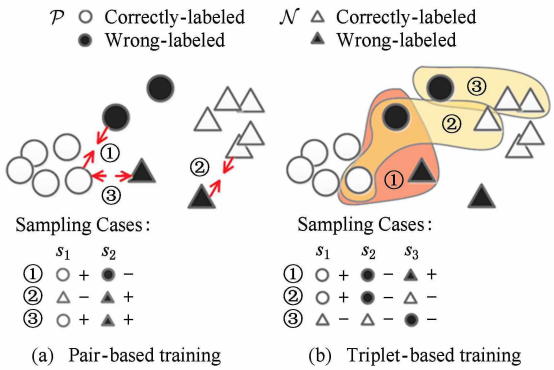


Fig. 12　Comparison between pair-based training and triplet-based training

图 12　二元训练准则与三元训练准则对比

中的噪声对训练过程的影响弱于二元训练准则中的噪声对训练过程的影响. 此外,在三元准则训练过程中,如果 2 对句子的距离之差大于间隔参数 λ,那么预训练目标函数 $\mathcal{L}_{\text{weak}}$ 的导数为零. 这种情况下,在训练过程中句子不发生任何移动. 例如,在图 12(b)的例②中,s_2 的真实类别是 neg,因此 s_2 不应该向 s_1 移动. 由于 s_3 与 s_1 之间的距离较大,使得预训练目标函数中 2 对句子的距离之差大于 λ. 因此,训练过程中句子不发生任何移动,s_1 与 s_2 不会互相靠近. 对比来看,在图 12(a)的例①和例②中,s_1 与 s_2 会朝对方移动直至两者之间的距离变为 0. 此时,噪声对训练过程的影响较大. 上述分析表明,与二元训练准则相比,三元训练准则在一定程度上减弱了噪声对训练过程的影响.

3.4 有监督学习微调模型

通过弱监督预训练步骤,我们得到了能够较好捕捉情感语义分布的模型. 有监督微调步骤用该模型的参数作为有监督训练的初始参数,同时在嵌入层上添加分类层,学习目标分类函数. 分类层采用 Softmax 输出函数,用少量人工标注的数据对整个模型进行有监督训练,从而得到最终的分类模型.

3.5 实验验证

我们在亚马逊评论数据集上验证 2 种实现 WDE 模型的性能. 我们从公开的亚马逊评论数据集[84]上收集了 3 类商品的评论数据:数码相机、手机和笔记本电脑,所有评论数据都包括相应评分信息. 我们利用文献[5]中的方法提取了 107 个商品主题关键词(aspect keywords). 之后,我们对所有评论数据进行了分句并过滤掉无 aspect keywords 和多 aspect keywords 的语句. 经过以上预处理操作我们得到了 1 143 721 条弱标注语句. 另外,我们人工标注了 11 754 条语句用于有监督训练. 标注工作包含 2 个部分:主客观标注和正负情感标注. 其中,主客观标注的目的是为了进一步分析情感分类方法分别在主观和客观语句上的性能. 标注数据被随机划分为训练集(50%)、验证集(20%)和测试集(30%). 标注数据详细情况如表 13 所示:

Table 13 Statistics of the Labeled Dataset

表 13 人工标注数据集

Label	Positive	Negative	Total
Subject	3750	2024	5774
Object	1860	4120	5980
Total	5610	6144	11754

3.5.1 WDE-CNN 和 WDE-LSTM 与其他分类方法对比实验

实验中采用的对照组方法描述如下:

1) Lexicon. 基于词典的方法[5].

2) SVM. "支持向量机+n-gram 特征"是情感分类中最常见的一种方法[32],实验中我们使用 trigrams 特征,并使用 Liblinear 分类器[85].

3) NBSVM. 文献[86]中将 NB 分类器和 SVM 分类器融合在情感分类上取得了较好的效果.

4) SSWE. SSWE 通过在弱标注信息上训练神经网络以得到词向量. 给定一条语句,对语句中所包含词的词向量求最大、最小和均值,从而得到语句的特征向量表达进行情感分类[50].

5) SentiWV. 该方法使用评分信息训练词向量,再使用线性分类器进行情感分类[49]. 用词向量生成语句特征表达的过程与 SSWE 相同.

6) CNN-rand. 在有标注数据集上训练基于 CNN 的网络模型(如图 10 所示),随机初始化网络参数.

7) LSTM-rand. 在有标注数据集上训练基于 LSTM 的网络模型(如图 11 所示),随机初始化网络参数.

8) CNN-weak. 直接将弱标注数据当作有标注数据训练基于 CNN 的网络模型(使用基于 LSTM 的网络模型效果相似,因此只展示基于 CNN 的网络模型的结果).

表 14 展示了实验结果. 通过对比可以看出,WDE-CNN 和 WDE-LSTM 的准确率和 Macro-$F1$

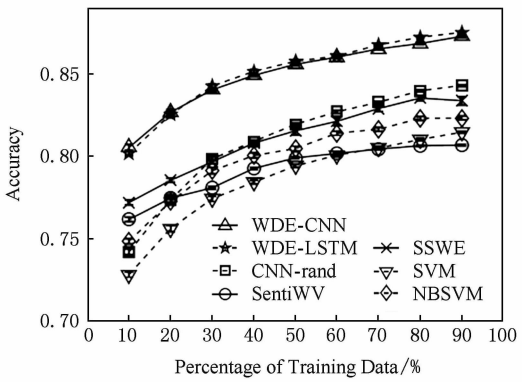
Table 14 Performance Comparison

表 14 性能比较

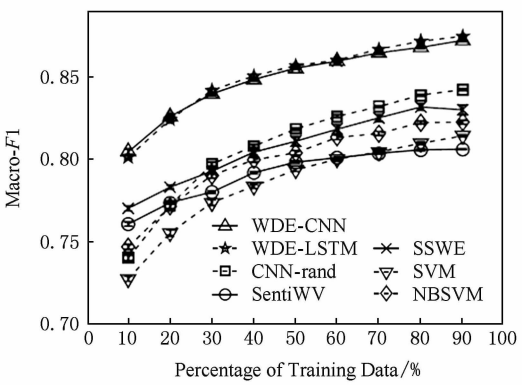
Method	Accuracy			Macro-$F1$		
	All	Subj	Obj	All	Subj	Obj
Lexicon	0.722	0.827	0.621	0.721	0.812	0.613
SVM	0.818	0.838	0.800	0.818	0.821	0.765
NBSVM	0.826	0.844	0.808	0.825	0.831	0.773
SSWE	0.835	0.857	0.815	0.834	0.826	0.804
SentiWV	0.808	0.806	0.809	0.807	0.786	0.771
CNN-rand	0.847	0.861	0.835	0.847	0.848	0.802
CNN-weak	0.771	0.773	0.770	0.771	0.755	0.741
LSTM-rand	0.845	0.863	0.829	0.845	0.852	0.794
WDE-CNN	**0.877**	**0.886**	**0.868**	**0.876**	**0.875**	**0.843**
WDE-LSTM	**0.879**	**0.889**	**0.870**	**0.879**	**0.878**	**0.844**

Notes: The bold values mean the best accuracy and Marco-$F1$ values among the different methods.

值都超越了其他方法;另外,WDE-CNN 和 WDE-LSTM 在小规模有标注训练数据上的性能也明显优于其他方法.如图 13 所示,在只采用 10%训练数据的情况下,2 种方法都达到了 80%左右的分类准确率,显著高于其他模型.原因在于 WDE 中引入了商品评分这种与句子情感语义有一定关联性的弱标注信息,并通过三元训练准则和"弱监督预训练+有监督微调"训练框架减弱了噪声对模型训练过程的影响,较好地将弱标注数据中的大量有用信息"记忆"到深度模型中.从表 14 中还可以看出,CNN-weak 效果较差,说明弱标注数据并不能直接用于有监督学习.



(a) Performance of accuracy



(b) Performance of Macro-$F1$

Fig. 13　Impact of labeled training data size on each method's performance

图 13　不同规模训练集对模型分类性能的影响

3.5.2　WDE-CNN 与 WDE-LSTM 对比

与 WDE-CNN 相比,WDE-LSTM 模型中的 LSTM 层可以捕捉评论语句中的长距离依赖关系,因此 WDE-LSTM 更善于从整体上"理解"句子的语义.我们对 2 个模型在测试数据上的分类结果进行了详细分析,发现:1)WDE-CNN 更善于对句法结构简单的语句进行分类,例如"Sound is not that good";2)对于句法结构复杂的语句,WDE-LSTM 则更加适用.表 15 展示的是 WDE-LSTM 分类正确而 WDE-CNN 分类错误的例句,表格第 2 列是评论语句的真实情感标签.可以看到,前两句话都是转折句,转折词前后子句的情感极性发生了反转.由于 WDE-CNN 提取的是局部特征,情感含义冲突的局部文字表达容易导致错误分类,如第 1 句中的"not the greatest"和"is ok".最后 1 句话中,否定词"None"和表达用户观点的内容之间的距离超出了滑动窗口的最大容量,因此 WDE-CNN 很难捕捉到两者之间的依赖关系.对于 WDE-LSTM 而言,它可以从整体上捕捉语句内的长距离依赖关系,从而能够正确分类句法结构复杂的语句.

Table 15　Example Sentences on Which WDE-LSTM Makes Correct Prediction While WDE-CNN Fails

表 15　WDE-LSTM 分类正确而 WDE-CNN 分类错误的例句

Sentence	Label
Battery capacity is not the greatest, but it is ok.	Positive
The internet drops randomly while the yoga in the same room is absolutely fine.	Negative
None of these cameras has an articulating front view screen.	Negative

3.5.3　预训练间隔参数 λ 对模型分类性能的影响

预训练目标函数式(14)中的间隔参数 λ 本质上是控制我们要将弱标注正类和弱标注负类分开的程度.若 λ 参数设定过小会导致无法有效捕捉情感分布,而 λ 参数设定过大会导致噪声影响被放大.在实验中,我们测试了不同 λ 取值对分类结果的影响.首先需要设定 λ 的测试范围.由于嵌入层特征是 300 维的向量且神经元的输出值范围为[−1,1].这就形成了一个超立方体,立方体内任意 2 点间的最大距离约为 35.因此,我们将 λ 的测试范围设为 1～30

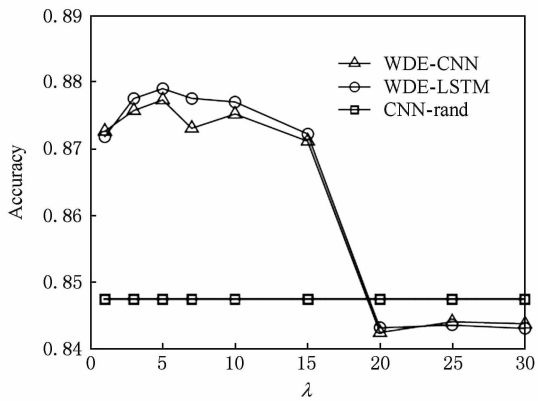


Fig. 14　Impact of λ on classification performance

图 14　不同 λ 取值对情感分类性能的影响

之间. 图 14 展示了不同 λ 值对最终情感分类性能的影响. 从图 14 中可以看出,当 $\lambda>15$ 时,分类性能严重下降;在 $\lambda<15$ 时,WDE-CNN 和 WDE-LSTM 都达到了较高的分类准确率. 此外,当 λ 取值较大时(即大于嵌入空间最大距离的一半),模型经过长时间训练更容易陷入饱和区域[87]. 综合上述分析和实验结果,我们将设置优化目标间隔参数 $\lambda=5$.

4 总结与展望

本文对情感分类技术进行了系统性归纳,并着重介绍了弱监督深度学习在情感分类问题上的最新研究进展. 本节我们简要梳理传统情感分类方法的不足,并总结弱监督深度学习的要点和挑战.

传统情感分类方法中,基于词典的方法过于依赖情感词典的构建,而机器学习方法的关键在于特征设计. 无论是生成情感词典还是设计分类特征,都要求相关人员具有丰富的领域知识. 此外,传统机器学习方法中的分类特征一般只能针对特定问题,推广能力有限. 相比而言,深度模型拥有更强大的表达能力,能够更好地学习从数据到情感语义的复杂映射函数. 但是,深度模型的训练是关键挑战. 一方面,由于文本数据分布与所要预测的情感语义之间没有很强的相关性,所以无监督预训练方法在情感分类问题上效果欠佳;另一方面,有监督训练方法需要大量有标注数据来训练深度模型,而获得大规模有标注评论语句需要耗费大量人力进行数据标注工作.

基于弱监督的深度学习方法则提供了一种解决情感分类问题的新思路:先使用互联网用户产生的大量评分信息对深度模型进行弱监督预训练得到一个能够捕捉情感语义分布的语句高层特征表达,再利用少量有标注数据进行监督学习预测情感极性. 该方法引入了互联网用户产生的弱标注数据作为深度模型的训练集,能够较好地利用弱标注数据中的有用信息. 对于采用其他互联网用户产生的弱标注数据(如 tagging 数据、表情符号等)来训练深度模型也有借鉴意义. 相比于其他深度学习方法,基于弱监督的深度学习方法的优势有 3 点:1)该方法更容易获取训练数据且很大程度上减少了人工标注成本;2)该方法中深度模型的预训练方法具有抗噪能力,能有效减弱训练数据中噪声对模型训练过程的影响;3)该方法可以推广应用到很多文本智能理解应用问题上. 互联网中存在大量的用户产生的文本弱标注信息,如百度知道问答社区的最佳答案①、美味书签网站的用户标签②等.

因此,可以将基于弱监督的深度学习方法推广到相应的文本智能理解应用问题上,如智能问答系统、推荐系统等等. 另一方面,基于弱监督的深度学习方法性能的好坏一定程度上取决于弱标注数据中噪声的影响. 因此,如何有效过滤弱标注数据中的噪声是未来研究工作中亟待解决的问题.

由于情感分类在不同现实场景中有着广泛应用,如电影票房预测、股指预测、政府政策调控等. 因此,探索更好的情感分类方法仍然会是情感分类领域的热点问题. 另外,如何将 WDE 有效地应用在其他包含弱标注信息的问题上也是未来的重要挑战之一.

① https://zhidao.baidu.com/.

② https://del.icio.us/.

参 考 文 献

[1] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey [J]. Ain Shams Engineering Journal, 2014, 5(4): 1093-1113

[2] O'connor B, Balasubramanyan R, Routledge B R, et al. From tweets to polls: Linking text sentiment to public opinion time series [C] // Proc of the 4th Int AAAI Conf on Weblogs and Social Media. Menlo Park, CA: AAAI, 2010: 122-129

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

[4] Hu Mingqi, Liu Bing. Mining and summarizing customer reviews [C] //Proc of the 10th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2004: 168-177

[5] Ding Xiaowen, Liu Bing, Yu P S. A holistic lexicon-based approach to opinion mining [C] //Proc of Int Conf on Web Search and Web Data Mining. New York: ACM, 2008: 231-240

[6] Liu Bing, Hu Miaowen, Cheng Junsheng. Opinion observer: Analyzing and comparing opinions on the Web [C] //Proc of Int Conf on World Wide Web. New York: ACM, 2005: 342-351

[7] Zhu Jingbo, Wang Huizhen, Zhu Muhua, et al. Aspect-based opinion polling from customer reviews [J]. IEEE Trans on Affective Computing, 2011, 2(1): 37-49

[8] Hu Mingqi, Liu Bing. Mining opinion features in customer reviews [C] //Proc of the 19th National Conf on Artifical Intelligence. Menlo Park, CA: AAAI, 2004: 755-760

[9] Zhuang Li, Jing Feng, Zhu Xiaoyan. Movie review mining and summarization [C] //Proc of Conf on Information and Knowledge Management. New York: ACM, 2006: 43-50

[10] Lin D. Dependency-Based Evaluation of Minipar [G] // Treebanks. Berlin: Springer, 2003: 317-329

[11] Jin Weijin, Ho H H, Srihari R K. OpinionMiner: A novel machine learning system for Web opinion mining and extraction [C] //Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2009: 1195-1204

[12] Lu Yue, Zhai Chengxiang, Sundaresan N. Rated aspect summarization of short comments [C] //Proc of Int World Wide Web Conf. New York: ACM, 2009: 131-140

[13] Jakob N, Gurevych I. Extracting opinion targets in a single- and cross-domain setting with conditional random fields [C] //Proc of Empirical Methods in Natural Language Processing. Cambridge, MA: MIT Press, 2010: 1035-1045

[14] Branavan S R K, Chen H, Eisenstein J, et al. Learning document-level semantic properties from free-text annotations [J]. Journal of Artificial Intelligence Research, 2014, 34 (1): 569-603

[15] Zhao W X, Jiang Jing, Yan Hongfei, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid [C] //Proc of Empirical Methods in Natural Language Processing. Cambridge, MA: MIT Press, 2010: 56-65

[16] Brody S, Elhadad N. An unsupervised aspect-sentiment model for online reviews [C] //Proc of North American Chapter of the Association of Computational Linguistics. New York: ACM, 2010: 804-812

[17] Zhang Lei, Liu Bing. Aspect and Entity Extraction for Opinion Mining [M]. Berlin: Springer, 2014

[18] Hai Zhen, Chang Kuiyu, Kim J. Implicit feature identification via co-occurrence association rule mining [C] // Proc of Computational Linguistics and Intelligent Text Processing. Berlin: Springer, 2011: 493-514

[19] Su Qi, Xu Xinying, Guo Honglei, et al. Hidden sentiment association in Chinese Web opinion mining [C] //Proc of Int Conf on World Wide Web. New York: ACM, 2008: 959-968

[20] Duan Xiuting, He Tingting, Song Le. Research on sentiment classification of blog based on PMI-IR [C] //Proc of 2010 Int Conf on Natural Language Processing and Knowledge Engineering (NLP-KE). Piscataway, NJ: IEEE, 2010: 1-6

[21] Lu Yao, Kong Xiangfei, Quan Xiaojun, et al. Exploring the sentiment strength of user reviews [C] //Proc of Int Conf on Web-Age Information Management. Berlin: Springer, 2010: 471-482

[22] Nasukawa T, Yi J. Sentiment analysis: Capturing favorability using natural language processing [C] //Proc of Int Conf on Knowledge Capture. New York: ACM, 2003: 70-77

[23] Turney P D. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews [C] //Proc of the 40th Annual Meeting of the Association for Computational Linguistics. Cambridge, MA: MIT Press, 2002: 417-424

[24] Feldman R. Techniques and applications for sentiment analysis [J]. Communications of the ACM, 2013, 56(4): 82-89

[25] Zhang Lei, Liu Bing. Identifying noun product features that imply opinions [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics. Cambridge, MA: MIT Press, 2011: 575-580

[26] Lu Yue, Castellanos M, Dayal U, et al. Automatic construction of a context-aware sentiment lexicon: An optimization approach [C] //Proc of Int World Wide Web Conf. New York: ACM, 2011: 347-356

[27] Stone P J, Dunphy D C, Smith M S. The general inquirer: A computer approach to content analysis [J]. American Journal of Sociology, 1968, 73(5): 375-376

[28] Pennebaker J W, Francis M E, Booth R J. Linguistic inquiry and word count 2001 [J]. Lawrence Erlbaum Associates Mahwah Nj, 2001, 10(2): 22-32

[29] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis [C] //Proc of the Conf on Human Language Technology and Empirical Methods in Natural Language Processing. Cambridge, MA: MIT Press, 2005: 347-354

[30] Riloff E, Wiebe J. Learning extraction patterns for subjective expressions [C] //Proc of Empirical Methods in Natural Language Processing. Cambridge, MA: MIT Press, 2003: 105-112

[31] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining [C] //Proc of Int Conf on Language Resources and Evaluation. Piscataway, NJ: IEEE, 2010: 2200-2204

[32] Pang B, Lee L, Vaithyanathan S. Thumbs up?: Sentiment classification using machine learning techniques [C] //Proc of Empirical Methods in Natural Language Processing. Cambridge, MA: MIT Press, 2002: 79-86

[33] Dave K, Lawrence S, Pennock D M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews [C] //Proc of Int World Wide Web Conf. New York: ACM, 2003: 519-528

[34] Mullen T, Collier N. Sentiment analysis using support vector machines with diverse information sources [C] //Proc of Empirical Methods in Natural Language Processing. Cambridge, MA: MIT Press, 2004: 412-418

[35] Kamps J, Marx M. Words with attitude [C] //Proc of the 14th Belgian-Netherlands Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2002: 332-341

[36] Osgood C E. The nature and measurement of meaning [J]. Psychological Bulletin, 1952, 49(3): 197-237

[37] Saleh M R, Mart N-Valdivia M T, Montejo-R Ez A, et al. Experiments with SVM to classify opinions in different domains [J]. Expert Systems with Applications, 2011, 38 (12): 14799-14804

[38] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts [C] //Proc of Meeting on Association for Computational Linguistics. Cambridge, MA: MIT Press, 2004: 271-278

[39] Taboada M, Grieve J. Analyzing appraisal automatically [C] //Proc of AAAI Spring Symp. Menlo Park, CA: AAAI, 2004: 158-161

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

[41] Feng Shi, Fu Yongchen, Yang Feng, et al. Blog sentiment orientation analysis on dependency parsing [J]. Journal of Computer Research and Development, 2012, 49(11): 2395-2406 (in Chinese)
(冯时, 付永陈, 阳锋, 等. 基于依存句法的博文情感倾向分析研究[J]. 计算机研究与发展, 2012, 49(11): 2395-2406)

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

[43] Sindhwani V, Melville P. Document-word co-regularization for semi-supervised sentiment analysis [C] //Proc of the 8th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2008: 1025-1030

[44] Prabowo R, Thelwall M. Sentiment analysis: A combined approach [J]. Journal of Informetrics, 2009, 3(2): 143-157

[45] Fang Ji, Chen B. Incorporating lexicon knowledge into SVM learning to improve sentiment classification [C] //Proc of the Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP). New York: ACM, 2011: 94-100

[46] Abbasi A, Chen H, Salem A. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums [J]. ACM Trans on Information Systems, 2008, 26 (3): 12-47

[47] Qu Lizhen, Gemulla R, Weikum G. A weakly supervised model for sentence-level semantic orientation analysis with multiple experts [C] //Proc of the 2012 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Cambridge, MA: MIT Press, 2012: 149-159

[48] Täckstoröm O, Mcdonald R. Semi-supervised latent variable models for sentence-level sentiment analysis [C] //Proc of the Meeting of the 49th Annual Meeting of Association for Computational Linguistics. Cambridge, MA: MIT Press, 2011: 569-574

[49] Maas A L, Daly R E, Pham P T, et al. Learning word vectors for sentiment analysis [C] //Proc of the Meeting of the Association for Computational Linguistics. Cambridge, MA: MIT Press, 2011: 142-150

[50] Tang Duyu, Qin Bing, Liu Ting. Deep learning for sentiment analysis: Successful approaches and future challenges [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2015, 5(6): 292-303

[51] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12(Aug): 2493-2537

[52] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural computation, 2006, 18(7): 1527-1554

[53] Yu Kai, Jia Lei, Chen Yuqiang, et al. Deep learning: Yesterday, today and tomorrow [J]. Journal of Computer Research and Development, 2013, 50(9): 1799-1804 (in Chinese)
(余凯, 贾磊, 陈雨强, 等. 深度学习的昨天、今天和明天[J]. 计算机研究与发展, 2013, 50(9): 1799-1804)

[54] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012: 1097-1105

[55] Farabet C, Couprie C, Najman L, et al. Learning hierarchical features for scene labeling [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2013, 35(8): 1915-1929

[56] Tompson J, Jain A, Lecun Y, et al. Joint training of a convolutional network and a graphical model for human pose estimation [J]. Eprint Arxiv, 2014: 1799-1807

[57] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1-9

[58] Mikolov T, Deoras A, Povey D, et al. Strategies for training large scale neural network language models [C] //Proc of Automatic Speech Recognition and Understanding. Piscataway, NJ: IEEE, 2011: 196-201

[59] Hinton G, Deng Li, Yu Dong, et al. Deep neural networks for acoustic modeling in speech recognition [J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97

[60] Sainath T N, Mohamed A R, Kingsbury B, et al. Deep convolutional neural networks for LVCSR [C] //Proc of Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2013: 8614-8618

[61] Leung M K, Xiong H Y, Lee L J, et al. Deep learning of the tissue-regulated splicing code [J]. Bioinformatics, 2014, 30 (12): 121-129

[62] Xiong Huiyuan, Alipanahi B, Lee L J, et al. The human splicing code reveals new insights into the genetic determinants of disease [J]. Science, 2015, 347 (6218): 1254806-1254806

[63] Bordes A, Chopra S, Weston J. Question answering with subgraph embeddings [J]. Computer Science, 2014, 8(4): 23-33

[64] Jean S, Cho K, Memisevic R, et al. On using very large target vocabulary for neural machine translation [J]. Computer Science, 2015(10): 35-45

[65] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 3104-3112

[66] Bespalov D, Bai Bing, Qi Yanyu, et al. Sentiment classification based on supervised latent n-gram analysis [C] //Proc of ACM Conf on Information and Knowledge Management. New York: ACM, 2011: 375-382

[67] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: A deep learning approach [C] //Proc of Int Conf on Machine Learning. New York: ACM, 2011: 513-520

[68] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces [C] //Proc of Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Cambridge, MA: MIT Press, 2012: 1201-1211

[69] Socher R, Pennington J, Huang E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions [C] //Proc of Empirical Methods in Natural Language Processing. Cambridge, MA: MIT Press, 2011: 151-161

[70] Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank [C] //Proc of Empirical Methods in Natural Language Processing. Cambridge, MA: MIT Press, 2013: 1631-1642

[71] Kim Y. Convolutional neural networks for sentence classification [J]. arXiv preprint, arXiv: 1408. 5882, 2014

[72] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences [J]. arXiv preprint, arXiv: 1404. 2188, 2014

[73] Mitchell J, Lapata M. Composition in distributional models of semantics [J]. Cognitive Science, 2010, 34(8): 1388-1429

[74] Frege G. On sense and nominatum [J]. Philosophy of Science, 1949, 59(16): 35-39

[75] Wu Qiong, Liu Yue, Shen Huawei, et al. A unified framework for cross-domain sentiment classification, [J]. Journal of Computer Research and Development, 2013, 50 (8): 1683-1689 (in Chinese)
(吴琼, 刘悦, 沈华伟, 等. 面向跨领域情感分类的统一框架 [J]. 计算机研究与发展, 2013, 50(8): 1683-1689)

[76] Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2007: 153-160

[77] Zhu Xiaodan, Sobihani P, Guo Hongyu. Long short-term memory over recursive structures [C] //Proc of Int Conf on Machine Learning. New York: ACM, 2015: 1604-1612

[78] Hu Baoting, Lu Zhengdong, Li Hang, et al. Convolutional neural network architectures for matching natural language sentences [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015: 2042-2050

[79] Bengio Y. Learning deep architectures for AI [J]. Foundations & Trends in Machine Learning, 2009, 2(1): 1-127

[80] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representations of words and phrases and their compositionality [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 3111-3119

[81] Greff K, Srivastava R K, Koutnik J, et al. LSTM: A search space odyssey [J]. IEEE Trans on Neural Networks & Learning Systems, 2016(7): 10-18

[82] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks, 2005, 18(5/6): 602-610

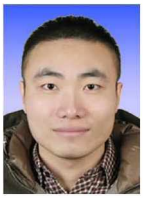
[83] Weston J, Ratle F, Mobahi H, et al. Deep Learning via Semi-Supervised Embedding [G] //Neural Networks: Tricks of the Trade. Berlin: Springer, 2012: 639-655

[84] Mcauley J, Pandey R, Leskovec J. Inferring networks of substitutable and complementary products [C] //Proc of the 21st ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2015: 785-794

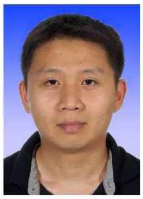
[85] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification [J]. Journal of Machine Learning Research, 2008, 9(Aug): 1871-1874

[86] Wang S, Manning C D. Baselines and bigrams: Simple, good sentiment and topic classification [C] //Proc of the 50th Annual Meeting of the Association for Computational Linguistics. Cambridge, MA: MIT Press, 2012: 90-94

[87] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828



Chen Long, born in 1989. Received his BSc degree in electronic information engineering from City College, Xi'an Jiaotong University in 2012 and received his MSc degree in electronics and communications engineering from Northwest University, Xi'an, China, in 2015. PhD candidate at the School of Information Science and Technology, Northwest University, Xi'an, China. His main research interests include deep learning, sentiment analysis, text mining and natural language processing.



Guan Ziyu, born in 1982. Received his BSc and PhD degrees in computer science from Zhejiang University, in 2004 and 2010, respectively. Full professor in the School of Information Science and Technology of Northwest University. His main research interests include attributed graph mining and search, machine learning, expertise modeling and retrieval, and recommender systems.

He Jinhong, born in 1983. Received his BSc degree in management engineering from People's Liberation Army Guilin Air Force Academy in 2009. After serving in the army for 3 years, he joined Northwest University. His main research interests include image processing, machine learning and information security.

Peng Jinye, born in 1964. Received his MSc degree in radio electronics from Northwest University in 1996 and received his PhD degree in signal and information processing from Northwestern Polytechnical University in 2002. Full professor in Northwest University in 2003. He was awarded as "New Century Excellent Talent" by the Ministry of Education of China in 2007. His main research interests include machine learning, image/video analysis and retrieval, and face recognition.