计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(11):296−308 [doi: 10.15888/j.cnki.csa.008779]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于预训练模型和基础词典扩展的酒店评论情感分析①

丁美荣, 冯伟森, 黄荣翔, 罗嘉俊

(华南师范大学 软件学院, 佛山 528225)
通信作者: 冯伟森, E-mail: 3150950776@qq.com

摘　要: 本文主要针对酒店领域的评论信息进行情感分析, 研究用户对于酒店配置、服务等方面的态度, 以便为酒店提高个性化服务质量提供一定的帮助. 本文基于 BiLSTM 神经网络构建预训练模型进行实验, 同时与传统的机器学习算法进行比较, 实验结果显示, 相较于朴素贝叶斯, 支持向量机的分析准确率更为稳定, 而利用预训练模型进行预测的精确率相比前两者有小幅度的提高; 同时以基础词典为主体, 构建适用于酒店评论的扩展情感词典, 对否定词的权重进行了弱化处理, 减小对带有相反含义语句的分类效果的影响, 将基础词典与扩展词典对获取的同一语料进行情感分类, 比较二者的结果表明采用扩展词典进行正向分类的准确率为 86%, 负向分类的准确率为 84%, 结果显示扩展词典分类比基础词典的分类效果更好.

关键词: 情感分析; 预训练模型; 情感词典; 支持向量机; 自然语言处理

引用格式: 丁美荣,冯伟森,黄荣翔,罗嘉俊.基于预训练模型和基础词典扩展的酒店评论情感分析.计算机系统应用,2022,31(11):296–308. http://www.c-s-a.org.cn/1003-3254/8779.html

Hotel Review Sentiment Analysis Based on Pretraining Model and Basic Dictionary Extension

DING Mei-Rong, FENG Wei-Sen, HUANG Rong-Xiang, LUO Jia-Jun

(School of Software, South China Normal University, Foshan 528225, China)

Abstract: This study mainly analyzes the sentiment of user reviews on hotels by investigating the attitudes of users toward hotel configuration and service to help hotels improve the quality of personalized service. Specifically, a pretraining model based on the BiLSTM neural network is built and compared with traditional machine learning algorithms. The experimental results reveal that the analysis accuracy of support vector machines (SVMs) is more stable compared with that of naive Bayes, while the prediction accuracy using the pretraining model is slightly improved compared with that of the previous two. Moreover, an extended dictionary of sentiment, with the basic dictionary as the main part, is constructed for reviews on hotels, and the weights of negatives are weakened to reduce the impact on the classification of sentences with opposite meanings. The basic dictionary and the extended dictionary are used to classify the sentiment of the same corpus obtained, and the comparison of the results indicates that with the extended dictionary, the accuracy of the positive classification and negative classification is 86% and 84%, respectively. This indicates that the classification effect of the extended dictionary is better than that of the basic dictionary.

Key words: sentiment analysis; pretraning model; sentiment dictionaries; support vector machine (SVM); natural language processing (NLP)

① 基金项目: 2021 年度广东省基础与应用基础研究基金 (2021A151501117)
收稿时间: 2022-01-30; 修改时间: 2022-03-10, 2022-03-25; 采用时间: 2022-04-02; csa 在线出版时间: 2022-07-07

1 引言

自然语言处理 (natural language processing, NLP) 是一种基于计算机的技术, 用于处理和加工人类社会特有的书面形式和口头形式的自然语言信息, 在挖掘海量文字信息的过程中逐渐产生了自然语言处理领域的各种技术与研究方向, 其中情感分析是一个重要的研究方向, 无论是对于企业领域还是社会生活均具有重要的研究价值和实际的应用价值. 如: 对商品评论信息中蕴含的情感进行分析, 不仅可以为消费者购置商品提供指导, 还能让商家客观地认识到自身商品的优缺点, 从而改善商品的功能或提高商品的质量[1]; 可以通过挖掘出隐藏在酒店评论中的用户情感倾向[2], 对用户提供消费引导的同时为酒店管理者提供改善建议, 达到双赢局面; 通过社交媒介观点信息的精确情感量化, 构建相关的股市预测模型, 有助于提升市场预测水平, 增加投资者决策依据[3], 从而间接带动社会经济的发展等. 本文主要针对酒店领域的用户评论进行情感分析, 研究用户对于酒店配置、服务等方面的态度, 为酒店提供更加精准高质量个性化服务提供一定的帮助, 同时根据用户对不同酒店评价的褒贬比例构建了一个酒店推荐网站, 为用户的出行旅居提供更优质的选择.

在现代社会, 随着电子商务与社交网络的快速发展, 人们在作为获取信息的客体的同时, 逐渐发展为创造信息的主体. 据第 47 次中国互联网络发展状况统计报告统计, 截至 2020 年 12 月, 我国使用互联网的人口数达 9.89 亿, 相较于 2020 年 3 月的数据实现了 8 540 万的增长, 互联网普及率提升至 70.4%[4]. 而互联网中海量的信息数据很大一部分以文字的形式呈现: 人们在通过网络购物平台进行购物的同时, 会留下对商品的评论; 人们在订购外卖时, 会针对某一饭店的菜品留下自己的评论; 最普遍的是人们在各种网络社交平台发表自己对某些问题的评论等. 这些评论中蕴含着人们的各种情感, 这些情感数据存在潜在的商业财富. 但由于信息量的规模庞大且具有实时性, 若以人工的方式来挖掘这些数据信息, 效率过于低下, 因此需要利用计算机对海量信息进行清洗、处理与分析, 以达到对有效的科学应用, 让数据为我们的社会和生活提供有效的服务, 让数据能真正赋能各行业的个性化高效精准的服务, 彰显数据的应用价值.

情感分析方面的研究在国外起步较早, 诸多学者对于英文文本的情感分析已开展过大量研究, 构建出如 SentiWordNet 等的情感词典. 中文因其词语的多义性与灵活的表达方式, 挖掘出文本中蕴藏的情感具有不小的难度. 近年来, 国内相关领域的学者逐渐对中文文本情感分析开展深入研究, 有两大类主流方法. 一种是基于机器学习的算法, 该方法利用词向量模型对获取的语料信息进行向量化类型转换, 然后利用最大熵[5]、贝叶斯方法[6]、支持向量机[7] 等分类技术, 对文本的情感倾向进行判断. 如: 徐勇等人[8] 建立了电子商务产品评价指标的结构模型, 通过机器学习算法, 提取并标注文本中的情感标签, 实现了对淘宝网站商品的模糊综合评价. 王祖辉等人[9] 使用粗糙集方法挖掘在线评论中的固定搭配特征, 并将其融合到支持向量机和朴素贝叶斯等情感分析模型中. 这种方法需要以大规模的评论样本作为训练支撑, 结合特征加权方法实现对评论情感值的计算. 而且朴素贝叶斯、支持向量机等方法效果主要依赖人工标注的数据数量和质量, 因此其效果受人的主观意识影响较大. 另一种是基于情感词典的算法: Hu 等人[10] 通过对评论中蕴含的产品属性进行评估来判断评论的情感倾向. 史伟等人[11] 从语义的角度出发构建模糊情感本体, 实现对在线评论的情感分析. 魏慧玲[12] 通过构建情感词典, 采用语义相似度算法对手机行业的在线评论进行分类, 实现对产品关键特征的情感判断. 这种方法需要构造出与研究领域适用的情感词典, 赋予词语不同的权重, 然后根据相关的语义规则计算文本情感值, 从而实现对文本的褒贬倾向分析. 基于情感词典的方法相较于机器学习算法更具通用性, 因为情感词典的构建与所分析的领域密切相关且有良好的成长性.

杨飞等人[13] 通过对大连理工情感词典进行高频词汇的扩展构建扩展情感词典, 对爬取的酒店评论语料进行了情感分析, 选用正确率 (precision)、召回率 (recall)、*F*1-score (*F*1) 作为指标来评价分类效果. 最终在对褒义评论、贬义评论进行情感分析的精准率上分别实现较大幅度的提升.

目前酒店评论情感分析仍然是一个持续关注的问题, 首先该研究是基于对庞杂的评论语料数据的读取, 评论数据具有更新快, 词性词义多变、流动强的特点, 所以需不断探索研究, 以期找到一种更合适的算法, 实现对数据的分析, 提炼出能针对具体问题的新方法. 随着机器学习的快速发展, 预训练技术在自然语言处理领域引起广泛的重视, 各研究领域利用其实现了迁移

学习[14]的概念, 因此, 预训练技术在解决大规模文本分类的问题上取得了显著成效. 为避免人工筛选语料特征产生主观性的错误, 以及使用多种自然语言处理工具造成的误差累计, 本文在传统方法的基础上也引入预训练模型进行情感分析实验探究.

2 研究原理与实现方法

在解决酒店评论情感分析的问题上, 诸多团队采用了构造扩展情感词典与结合机器学习算法的传统方法, 本文则以酒店评论作为研究对象, 结合多个实验角度, 在传统方法基础上, 融入了基于 BiLSTM 的预训练模型来计算文本情感值并进行褒贬倾向分析. BiLSTM 是 2015 年 Zhang 等人[15]在 LSTM 的基础上提出的双向长短期记忆网络模型 (bidirectional long short-term memory), 该模型通过设计使用前后两个方向的 LSTM, 可以分别获得当前词与上文和下文的关系, 学习上下文相关信息的情感特征[16]. 所以本文将 BiLSTM 作为预训练模型的一部分, 并将其应用于酒店评论文本情感分析, 以分析判别其中蕴含的酒店情感倾向. 主要研究内容如下.

(1) 搭建 Python 的 Scrapy 框架, 针对酒店领域的相关内容利用 Xpath 方法爬取携程、美团等平台上用户的评论信息. 利用 Python 的 Pandas 模块去除重复文本, 删除对模型建立无影响的特殊字符, 通过采集到的数据集了解用户对酒店服务的关注重点.

(2) 对数据进行标注分类, 其中分为预训练数据与待测试数据. 在自然语言处理的任务中, LSTM 通常被用于时序数据类预训练模型的建模, 而由双向 LSTM 组合而成的 BiLSTM 能够较好地捕捉到上下文语句的依赖, 非常适用于处理文本数据, 所以本文使用深度学习框架 TensorFlow 实现 BiLSTM 神经网络作为预训练基础模型, 为了获取待测试数据所有词的词向量, 通过嵌入中科大的开源词向量模型 Word2Vec 工具对预训练数据进行处理. Word2Vec 采用的负采样能够有效提高数据的处理效率, 在利用训练集对基础模型进行训练后最终得到能够判断酒店评论情感倾向的模型.通过抽取相关词构建关联字典, 对评论数据进行分词处理后形成判断向量, 最后利用朴素贝叶斯和支持向量机两种分类器以及基于 BiLSTM 的预训练模型对评论进行褒贬倾向分析比较. 过程如图 1 所示.

(3) 在基础词典的前提下, 通过构建程度副词词典、否定词词典和停用词词典, 将评论高频词赋予权重从而实现词典的扩展; 考虑到不同语境下同一词的使用可能代表完全相反的情感意义, 而否定词的出现往往会使文本的情感向相反的方向扭转, 于是我们弱化了否定词的权重, 使其权重的绝对值位于 0.5–0.75 区间, 结合加权扩展情感词典和相关语义规则对文本情感值进行计算, 即判断语句中每个情感词的前方是否存在否定词或程度副词, 并将位于它之前的否定词和程度副词列为一个组别, 若存在否定词, 则将情感词的权值乘上否定词的权重, 出现若干个则进行连乘, 若存在程度副词, 则将情感词的权值乘上程度副词的程度值, 最后将各组别的值相加, 情感值大于 0 判定为正向评论, 小于 0 判定为负向评论, 从而实现对酒店评论的情感分析. 我们利用经中文文本分词和去停用词处理的同一语料, 根据精确率指标对基础词典和扩展词典的分类效果进行评估对比, 研究思路如图 2 所示.

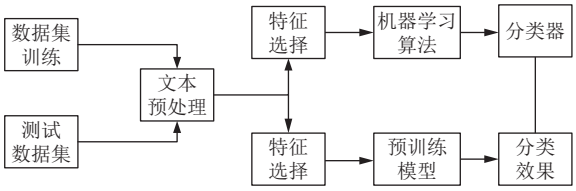

图 1 基于预训练模型与机器学习的褒贬倾向分析

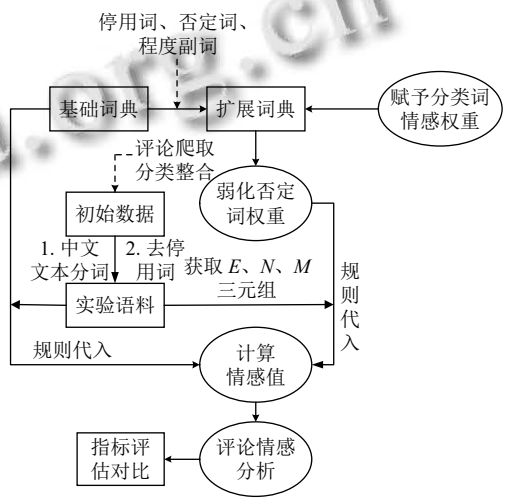

图 2 基于基础词典扩展的酒店评论分析方法

2.1 机器学习算法

本文采用的机器学习算法主要围绕两种展开: 朴素贝叶斯以及支持向量机.

朴素贝叶斯算法 (naive Bayesian algorithm)[17]是

机器学习中应用最广的分类算法之一, 它在贝叶斯算法的基础上进行了简化并以特征条件独立假设为前提. 其基本原理是: 对于给定的训练集, 以特征词相互独立为前提假设, 通过训练使模型学习输入输出的联合概率分布, 从而获得先验概率、条件概率, 在此基础上依据贝叶斯定理求解出使得后验概率最大的输出, 最终得到文本的分类结果.

大量实验研究显示, 朴素贝叶斯算法在情感分析领域具有良好的分类效果. 其特有的条件独立性和健壮性使得在对样本集进行一次扫描后就可以估计出其中所有的概率, 在数据量较小的情况下具备较好的学习效果. 朴素贝叶斯算法相对简单, 依靠设置少量的参数, 适用于多类型的分析, 对残缺的数据也不敏感, 然而, 该算法对输入数据的格式要求较严格[18,19].

支持向量机 (support vector machine, SVM)[20] 是一种基于监督学习方式的二元分类器, 具有稀疏性和稳健性, 通过在有限的样本空间中寻找最大边距超平面进行求解.

作为一个线性分类器它包含两种情况, 一是基于线性可分, 即根据输入数据的样本特征构造特征空间, 从中寻找能够区分不同类别且使得类别之间的间距为最大值的超平面. 二是基于分类问题不具有线性可分性时, 需要通过引入核函数[19] 求解.

(1) 线性可分

线性可分的情况是指样本集所在的特征空间存在能够将数据划分为正类和负类的超平面, 在样本空间中, 任意样本的点到该平面的距离大于等于 1, 如式 (1) 所示:

$$y_i(\omega^{\mathrm{T}}x_i+b)\geqslant 1 \tag{1}$$

其中, y_i 代表二元变量, 值域为{−1, 1}, $\omega=(\omega_1, \omega_2, \cdots, \omega_d)$ 代表法向量, 表示超平面的方向; x_i 代表样本特征, b 则代表截距, 即为超平面到原点之间的间距, 由法向量 ω 和位移 b 可以确定一个超平面, 记作 (ω, b).

在样本空间中, 存在若干支持向量, 任意两个支持向量到超平面的间距之和称为间隔[21], 如图 3 所示.

(2) 非线性可分

面对诸如用户评论等复杂的文本数据集, 由于无法找到能够将样本类别完全区分开的超平面, 需采用非线性方式进行处理. 通过非线性函数将样本数据映射至更高维空间, 而后采用线性支持向量机, 可使得样本数据在新的空间里是线性可分的, 这就回到了最优超平面求解问题上.

解决非线性问题的关键在于核函数的选取, 常见的核函数有线性核 (linear kernel)、多项式核 (polynomial kernel)、指数核函数 (exponential kernel) 等. 支持向量机算法具有优秀的泛化能力, 简化了分类问题, 适用于小样本学习. 但其缺点在于时间与空间的开销较大, 在面对大规模的训练样本时将耗费大量运算时间, 且支持向量机的难以解决多分类问题, 对于参数与核函数的选取比较敏感[20].

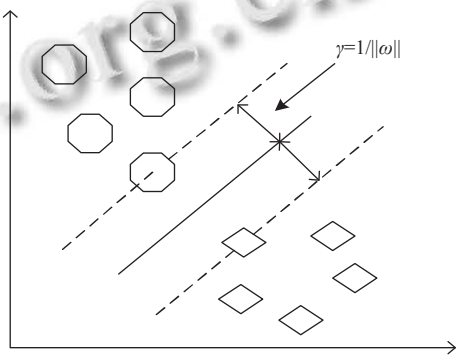

图 3 支持向量与间隔

2.2 预训练模型

由于文本的情感分类受到当前语义环境的影响, 因此, 探索一种能够适用于大部分文本分类的分类模型对自然语言处理具有重大的意义. 预训练模型 (pre-training models, PTMs)[22,23] 就是利用在生活中大量出现过的文本对模型进行训练, 使模型学习到相关的词或字出现的概率, 最终构建出满足这些文本分布特征的模型. 预训练模型是一种能够适用于大部分文本分类的模型, 在它的基础上, 进行相关语义环境的扩充和调整, 可以使该模型契合当前环境下的文本语义情感分析. 一个语言模型的语料库标签等同于它的上下文, 这就决定了人们几乎可以无限制地使用大型语料库进行语言模型的训练, 这些语料使预训练模型获得了强大的能力, 进一步在其他具体的相关任务上展现出优秀的分类效果.

PTMs 拥有灵活的适应性和优秀的性能, 采用基于预训练模型的方法对酒店评论进行文本分类的模型构建, 通过丰富原训练集的文本内容重新训练预训练模型, 可以得到适合于酒店评论环境下的情感分类模型.

相较于用 Word2Vec、SVM、PCA 等传统方法来构建的情感分类模型, 预训练模型不管是在数据集的

规模, 亦或是在分类的准确率上, 都要高于前者.

经典预训练模型包括 ELMo[24]、GPT[25]、BERT[26] 等. 本文使用深度学习框架 TensorFlow 中的 keras 接口实现基于 BiLSTM 的预训练模型, 从而对酒店评论情感分析的预测准确率进行比较, 该模型框架如图 4 所示. BiLSTM 由前向 LSTM 与后向 LSTM[27] 组合而成. LSTM[28] 是 RNN (recurrent neural network) 的一种, 由于其设计的特点, 非常适用于对文本数据进行建模, 但若采用单向的 LSTM 进行建模会存在一个问题, 它只能编码到从前往后传递的信息, 而无法捕捉到从后往前的信息依赖, 易对高细粒度的分类任务产生影响. 所以本文采用 BiLSTM 进行实验. 模型中包含输入门、输出门、记忆门、遗忘门、细胞状态及隐藏层, 而隐藏层贯穿于整个神经网络的运算, 所谓隐藏层, 是通过将输入数据的特征抽象化, 使其在另一个维度空间能够进行更好的线性划分, 诸如记忆门、遗忘门等的输入信息都与隐层状态相关联. 模型整体的输入是语料中每个词对应的词向量, 输出该词对应类别的值, 为了便于判断分类的准确性, 在输出层后构建一个线性层, 以此将输出结果映射至带有正负向标签的特征区间, 最终根据数据计算出分类的准确率. BiLSTM 的运作原理如下, 本文以“酒店设施齐全”一句为例进行分析, 编码模型如图 5 所示.

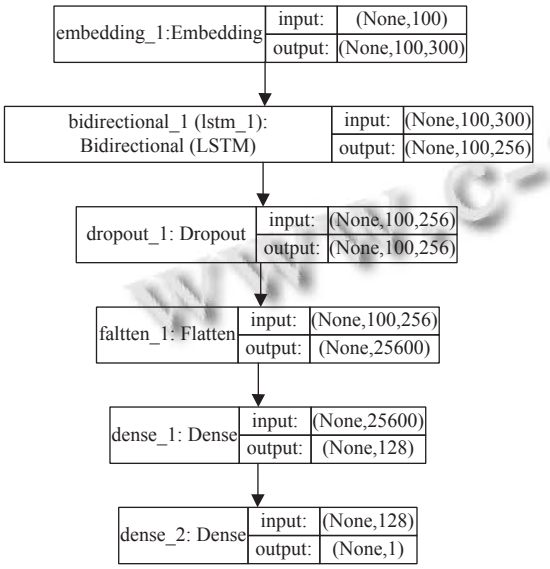

图 4 BiLSTM 模型架构

前向的 LSTM 依次输入“酒店”“设施”“齐全”, 后向的 LSTM 则依次输入“齐全”“设施”“酒店”, 共得到两组向量, 每组为 3 个, 分别是{h_{L0}, h_{L1}, h_{L2}}和{h_{R0}, h_{R1}, h_{R2}}, 将前后所得向量进行拼接, 得到结果向量{h_0, h_1, h_2}. 其中 h_0 由 h_{L0} 和 h_{R2} 拼接而成, h_1 由 h_{L1} 和 h_{R1} 拼接而成, h_2 由 h_{L2} 和 h_{R0} 拼接而成.

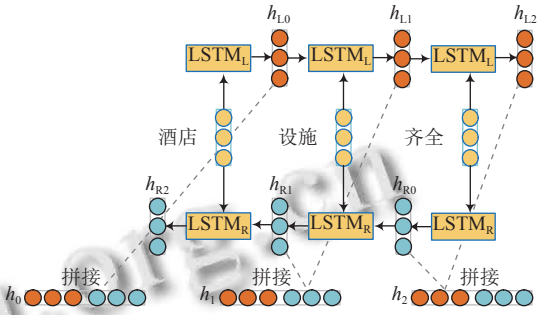

图 5 BiLSTM 编码模型

但经过这种处理得到的向量并未有效结合前后向信息, 仅仅包含了一部分语句的信息, 将拼接方式调整为对应序列拼接: 即 h_{L0}-h_{R0}, h_{L1}-h_{R1}, h_{L2}-h_{R2}, 能够包含前后向的所有信息. 如图 6 所示.

在 Softmax 层对拼接得到的数据进行归一化处理, 将数据缩放至 0–1 的区间内, 从而进行多分类.

2.3 扩展情感词典

情感词典是由情感词组成的集合, 其内包括词语及对应的情感值. 由于情感值的不同, 词典又可划分为正向 (褒性) 词典和负向 (贬性) 词典. 利用情感词典对语料进行情感分类的关键在于情感词语的匹配, 因此构建情感词典是进行情感分析的一项关键步骤. 伴随着在情感分析领域的深入研究, 学者们总结了诸多情感词典, 这些词典中收集了常见的情感词语, 同时标注了对应的词性, 具有很好的参考价值.

国外较为主流的英文情感词典, 有如 WordNet 词典、General Inquirer (GI) 词典[29] 及 Sentiword Net 词典等. 国内中文方面的情感词典资源相对缺乏且需不断完善, 知名度较高的有知网 HowNet 词典、大连理工情感词典等. 本文参考文献 [13] 的实验方法, 也在大连理工情感词典的基础上进行扩展, 借助 GitHub 发布的程度级别词语 (中文) 中的程度副词、否定词 (中文) 分别构建了程度副词词典、否定词词典, 同时又在由中科院计算所中文自然语言处理开放平台发布的中文停用词表上加以调整, 构建了用于酒店评论分析的停用词表, 与文献 [13] 不同的是, 我们对否定词的权重进行了弱化处理, 以减小对带有相反含义语句的分类

效果的影响.

在利用爬虫获取数据并分类归纳后, 针对酒店某个服务类别的用户评论开展定量分析. 结合构建的扩展情感词典与相关的语义规则, 计算每条语料的情感极性. 以大连理工情感词典为基础, 将分类词及其情感权重扩展进情感词典, 然后通过酒店评论的情感值计算方法计算情感倾向. 具体实现如下: 1) 读取待测试的酒店评论文本, 提取特征词 (情感词 E, 否定词 N, 程度副词 M); 2) 利用构建的程度副词、否定词、停用词词典实现情感词典的扩展; 3) 获取特征词所包含的情感词、否定词等对应的权重, 最后将评论中所有特征词情感强度数值进行累加, 得到的结果即为该条评论的情感值[13].

$$V=\sum V_e\times V_m\times V_n \tag{2}$$

其中, V_e 代表情感词在扩展情感词典中的对应权重; V_m 代表程度副词在程度副词词典中的对应权重; V_n 作为否定词标识, 若语料中存在否定词则 V_n 取值为–1, 若不存在, V_n 取值为 1. 如果评论的情感值大于 0, 则判定为正面评论; 情感值小于 0, 判定为负面评论.

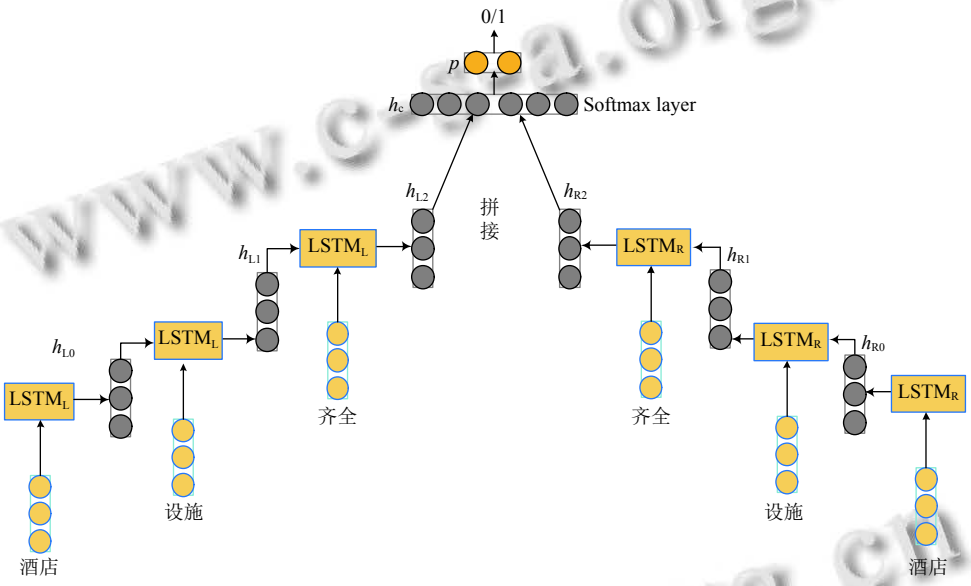

图 6 BiLSTM 数据拼接

在进行情感倾向分析后, 采用精确率和召回率两个指标对分析的准确性进行评估, 对于精确率和召回率的计算分别如式 (3) 和式 (4) 所示:

$$P=\frac{TP}{TP+FP} \tag{3}$$

$$R=\frac{TP}{TP+FN} \tag{4}$$

其中, TP 代表模型对评论的判断结果为真的正确率; FP 代表本应被判断为负面评论的测试数据却被模型判断为正面评论的这一类数据, 在所有负面评论中占有的比重, 作为模型的误报率; FN 代表本应被判断为正面评论的测试数据却被模型判断为负面评论的这一类数据, 在所有正面评论中占有的比重, 作为模型的漏报率.

在实际问题中, 一个变量往往受到多个变量的影响, 因此本文还采用多元线性回归模型[30] 对影响分类的因素进行估计. 表达式如式 (5) 所示:

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki}+u_i,\ i=1,2,\cdots,n \tag{5}$$

其中, Y 作为因变量代表酒店的销量; k 代表解释变量的数目; β_0 为常数项; β_j (j=1, 2, ⋯, k) 代表回归系数, 反映了解释变量对被解释变量的影响程度, 也被称为偏回归系数. 自变量 X_{1i}, X_{2i}, ⋯, X_{ki} 分别对应酒店星级、酒店评分、酒店评论情感倾向等与因变量线性相关的因素[31].

扩展情感词典与其面向的领域密切相关, 相当于一个可移动的语料库, 具有便捷性和易用性, 本文构建的是酒店领域的扩展情感词典, 以某酒店的评论为例, 进行情感评论的褒贬分析, 主要步骤如下.

(1) 首先爬取某酒店相关的情感评论, 将得到的评

论文本进行数据预处理, 进而得到所需的具有一定情感参考价值的文本数据集.

(2) 将文本数据集中的情感词与扩展情感词典中的情感词进行逐一检索, 若在词典中检索到情感词, 则将该情感词对应的情感权值提取出来. 当检索完文本数据集中所有的情感词后, 得到相应的带有情感权值的评论文本.

(3) 通过基于语义规则的情感值计算方法, 计算得到该文本中每条评论的情感极性.

(4) 设定阈值对结果进行分类, 超过该阈值的文本评论判定为正向评论, 否则判定为负向评论.

通过扩展情感词典对评论进行褒贬分析时所需注意的是, 在当前研究领域中对某一事物的情感评论可能在其他领域具有不同的情感倾向, 所以不可否认扩展情感词典存在一定的局限性, 但将其应用在同一领域的不同评论中均有较高的情感分类准确度.

3 实验与分析

本文利用爬取的酒店评论语料集, 选择其中的 2000 条作为实验的语料, 包含 1 000 条正向评论以及 1 000 条负向评论.

3.1 预处理

3.1.1 数据获取与分类

本文在中科院计算所中文自然语言处理开放平台发布的中文停用词表上进行更新调整, 共得到 1 200 个停用词. 通过建立 Scrapy 框架[32], 使用 Xpath 方法[33]以酒店相关内容为关键字爬取各平台上用户的评论信息. 对文本进行预处理操作, 剔除数据中的无用内容, 降低数据的噪声, 进而获取较好较完整的文本语料.

本文从爬取的数据中抽取 2 000 条评论组成数据集进行分析处理, 其中正向评论和负向评论各占一半. 将二者分别划分到两个 txt 文件中, 且均是一条评论占据一行. 将所有评论整合成文档, 作为接下来的正负向集合文档. 部分实现结果如图 7、图 8 所示.

3.1.2 中文文本分词

中文分词[34] 是自然语言处理领域的一个重要环节, 是情感分析实验中不可或缺的部分. 中英文本语料的主要差别是英文单词可以由空格隔开, 由此可以使用空格符的分割方式对英文文本进行分割处理. 而中文文本语料里有逗号等标点符号作为分割标准, 但是这样分割后的单位是一个个的句子. 因此还需要进行额外的操作, 将句子继续分割为有意义的字词, 这种操作称为“分词”. 中文分词主要有 3 种方法, 即基于词典分词法、基于统计分词法以及基于理解分词法[35].

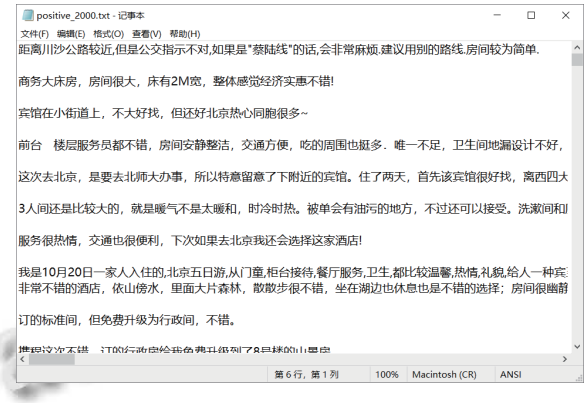

图 7 正向语料集合文档

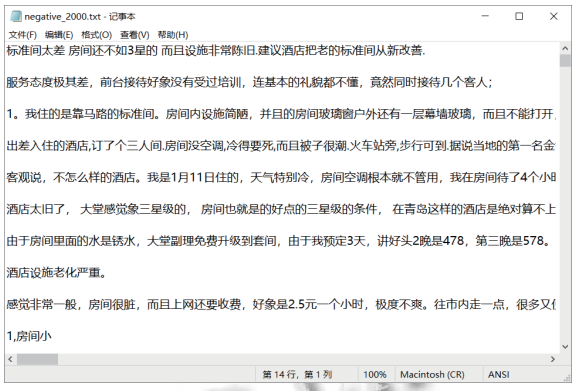

图 8 负向评论集合文档

建立在词典之上的分词方式就是把文本语料中的词和词典中已存在的词进行匹配操作, 进而得出已分类的分词文本. 本文采用结巴分词组件 (Jieba) 对正负向评论进行分词操作.

Jieba 工具是 Python 的一个集成模块, 它具有易操作和便捷性的特点, 且能提供多种方法进行分词操作. 在进行分词前, 对文本进行过滤或去噪处理, 删去特殊符号等以减少该类符号对分类效果的影响. 利用 Python 本身已有的模块可以实现对这些特殊符号的处理, 如字符串模块和正则表达式模块等.

处理完成后, 会得到正负语料的分词结果. 部分实现代码及结果如图 9、图 10 所示. 分词后的所有词汇共同组成文本语料的初始特征, 在接下来的情感分析操作中只用抽取出对判断正负向情感有效果的特征.

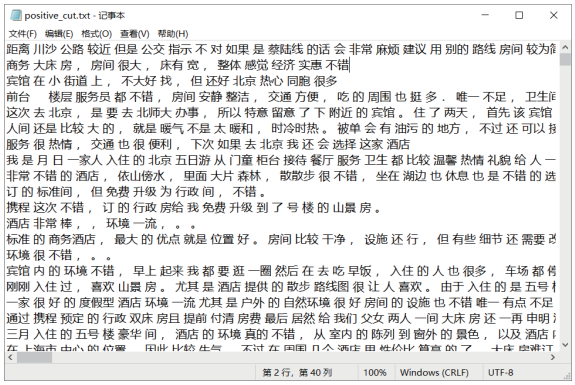

图 9 正向语料分词文档

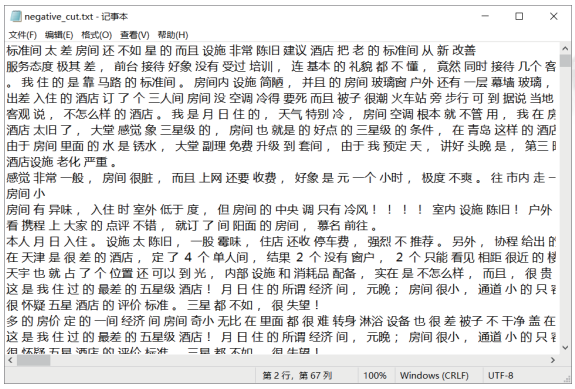

图 10 负向语料分词文档

3.1.3 去停用词

在中文语料的处理操作中, 名词、动词等在文本语料中一般起到十分关键的情感表达作用, 而其他词性例如介词、连词等在文本中对情感分析没有太大的影响, 这类词则为停用词. 按照停用词的定义, 抽取符合条件的词组成的词典叫停用词词典, 包括“呢”“的”“了”以及一些符号“. ”“%”“》”等. 在预处理过程中过滤这些没有太大意义的词, 可以降低无关信息的干扰, 进而为接下来的特征处理等过程提供一个清晰且干净的文本内容[36].

在对文本进行分词处理后, 将正向文本和负向文本与停用词表进行词与词的配对, 从而去除文本中的停用词. 去除的主要步骤如下.

1) 读取停用词表.

2) 分别遍历正负向语料文档, 将读取的词归档到该表中进行配对, 若当前词在表中出现, 则在正负向语料文档中过滤该词.

由于抽取出的停用词可能还包括换行符和制表符等不同的词, 若不对其进行处理可能会影响匹配结果, 所以还需要对停用词执行一次去符号操作. 将分词操作与去除停用词操作放置在相同的代码程序段中执行, 方便直接调用去停用词的函数, 在得到去停用词的语料文本后, 将其写入结果文件中.

3.2 传统模型的实现

本文构建传统机器学习算法模型使用到了多个著名的第三方模块, 主要有 Jieba、Gensim[37]、Pandas、Numpy、Scikit-learn、Matplotlib 以及 TensorFlow[38]. 实验中的中文分词版块由 Jieba 模块实现; 词向量模型的训练由 Gensim 中的 Word2Vec 模块实现; Pandas 是一种建立在 Numpy 模块基础之上的工具, 主要是为了解决数据分析; Scikit-learn 是一个工具包, 用来处理机器学习相关内容; Matplotlib 用于绘制二维图形; TensorFlow 通过采用数据流图用于数值计算.

由于模型的输入需要数值型的数据, 所以在经过以上处理后得到正负向语料文本还需进行类型转换操作. 诸如 bag of words (BOW)、Word2Vec 等是自然语言处理领域常见的转化算法. 本文采用通用性较强且维度较低的 Word2Vec[36] 词向量模型将文本内容处理成词向量.

由于语料中有较多的繁体字, 因此本文使用了 OpenCC 工具进行转换, 字体转换完成后采用结巴分词对处理后的语料文本完成分词操作, 由于采集的数据已经进行预处理, 所以在操作中不必进行数据冲洗即可使用.

在模型建立的基础上, 将已经处理的测试集与模型进行匹配, 获得特征向量, 第 1 列包含两种数值, 即 1 (正向) 和 0 (负向), 表示对应类别. 第 2 列及之后为数值向量, 每一行代表处理的一条评论.

利用 Word2Vec 模块获取特征向量如图 11 所示.

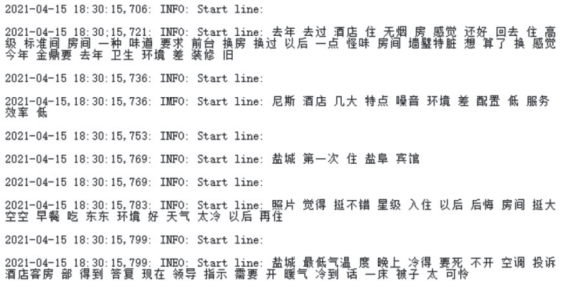

图 11 通过模型获取特征向量

3.3 传统机器学习测试结果

本文在利用机器学习算法进行测试时, 经过数据

分析和判断, 选取 400 作为临界维度进行实验, 训练得到相应维数的词向量. 由于特征空间的维度会影响向量的疏密程度, 导致分类结果有较大偏差, 因此采用了 PCA 算法[39] 对结果进行降维. 结果如图 12 所示.

根据维度图可以发现, 在 100 维以后曲线趋于平缓, 大部分原始信息在前面的维度就能够被包括在内, 所以最后选择 100 维为分界点, 前面维度的数据作为模型的输入部分.

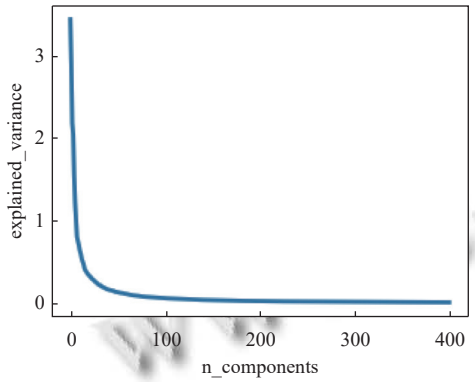

图 12 维度图

经过前期的运算比较, 团队发现基于机器学习的算法中, 支持向量机[40] 更为稳定, 因而在后续实验中使用支持向量机作为分类器算法, 先算出测试集的预测准确率, 然后使用相应的工具构建出 ROC 曲线, 进而验证分类器的效果, 大多数情况下模型的表现随 ROC 曲线面积 (AUC) 的增大而越来越好. 实验结果如图 13 所示.

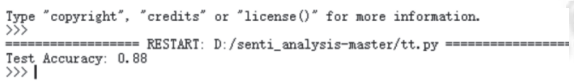

图 13 预测准确率

经过多次调整变量进行测试, 得到测试集的平均预测精度为 0.88. 相较于杨飞等人[13] 基于大连理工情感词典进行扩展的实验结果, 精确度呈现出相同的变化趋势. 本模型的 ROC 曲线如图 14 所示.

3.4 预训练模型测试结果

本文基于深度学习框架 TensorFlow 中的 Keras 接口实现 BiLSTM 神经网络, 利用爬取的预训练数据集进行训练. 训练集包含了 8 884 条中文酒店评论, 其中正、负向评论各有 4 442 条. 模型训练完成后, 采用上述实验中已经过分类的 2 000 条正负向评论进行测试 (正向评论 1 000 条, 负向评论 1 000 条; 评论均已打好标签, 正向评论标记为 1, 负向评论标记为 0). 实验的部分代码及结果如图 15–图 17 所示.

算法 1. BiLSTM 对实验集分类

```
model equals Sequential()
model.add
Embedding( //获取词向量
    words, embedding_dim,
    //嵌入维度矩阵
    weights equals [embedding_matrix])
    trainable equals False))
model.add( //载入 BiLSTM 模型
Bidirectional(LSTM(
units equals 64, return_sequences equals True,
dropout equals 0.2)))
model.add(
Dense( //处理数据过拟合
    1, activation equals "sigmoid"))
metrics equals [keras.metrics.BinaryAccuracy()])
```

算法 2. 预测

```
result equals model.predict(x) //输出预测结果
    x → tokens_pad
coef equals result[0][0]
if coef >= 0.5 then //情感值达到阈值激活函数
    print('正面评价', 'output=%.2f' % coef)
else
    print('负面评价', 'output=%.2f' % coef)
```

在利用预训练模型对已分类评论进行测试后, 我们根据输出结果计算预测的准确率, 系统显示准确率高达 92%

但由于本文在测试时采用的已分类数据集存在误差, 以及预训练模型参数设置的不贴切, 导致实验在准确率结果上还有待提升.

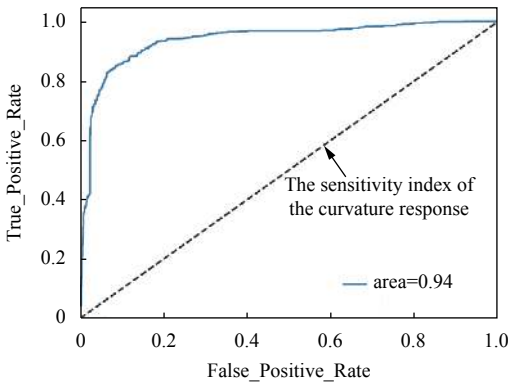

图 14 ROC 曲线

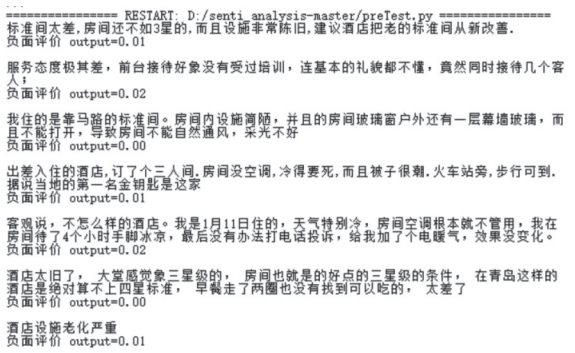

图 15 负面评论情感预测结果

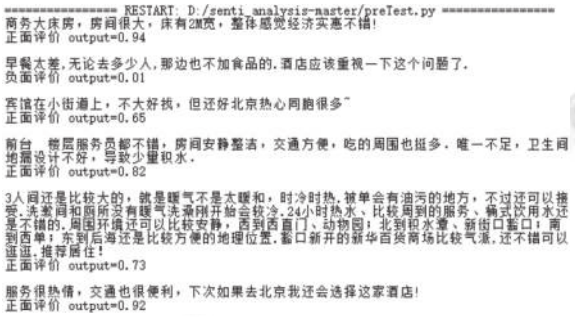

图 16 正面评论情感预测结果

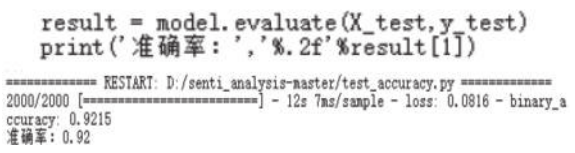

图 17 评论情感分析预测准确率

3.5 基于扩展词典的情感值计算

团队在对实验评论语料完成预处理操作后, 使用扩展情感词典进行匹配, 将匹配结果依照算法计算文本的情感值. 根据最终计算结果的值对评论语句进行分类, 若值为正数, 则归为正向语料, 若值为负数, 则归为负向语料, 词语匹配及情感值计算的部分实现代码如算法 3 和算法 4, 结果如图 18.

算法 3. 词典匹配

```
1. //找出文本中的情感词、程度副词和否定词
2. def classify_words(word_list):
3. //读取情感词典
4.     file equals open(Expanded_dictionary.txt, 'r+')
5. //获取词典内容
6.     list equals file.readlines()
7. //创建情感字典
8. emo_dict equals defaultdict()
9. //读取词典每一行的内容, 将其转换成字典对象
10. //key 作为情感词索引, value 为其对应权重
11.     for i in list
12.         if len (i.split(' ')) equals 2
13.             emo_dict[i.split(' ')[0]] → i.split(' ')[1]
14. //读取否定词、程度副词词典
15.     neg_word_list equals neg_word_file.readlines()
16.     deg_word_list equals deg_word_file.readlines()
17. //匹配
18.     for i in range(len(word_list)):
19.         if word in emo_dict.keys() //匹配情感词
20.         and word not in neg_word_list
21.         and word not in deg_word_list:
22.             word[i] → emo_dict[word]
23.         elif word in neg_word_list //匹配否定词
24.         and word not in deg_word_list:
25.             neg_word[i] → –0.75
26.         elif word in deg_word_list: //匹配程度副词
27.             deg_word[i] → deg_dict[word]
```

算法 4. 评论情感值计算

```
1. //初始化情感值与权重
2. value → 0
3. weight → 1
4. //情感词下标初始化
5. emo_index → -1
6. //遍历分词结果
7. for i in range(0, len(seg_result)):
8.     if i in emo_word.keys(): //判断是否为情感词
9.         value → value+weight*float(emo_word[i])
10. //权重乘以情感词得分
11.         emo_index → emo_index+1
12.         if emo_index < len(emo_index_list)-1:
13. //判断两个情感词间是否有否定词、程度副词
14.         for i in range():
15.             if j in neg_word.keys():
16.                 weight → weight * –0.75 //调整权重
17.             elif j in deg_word.keys():
18.                 weight → weight * float(deg_word[j])
19.     emo_index → emo_index+1 //读取下一词
20.     return value
```

图 18 评论语料情感值运算结果

经过统计分析得出正向分类的准确率为 86%, 负向分类的准确率为 84%, 比基础词典的分类效果更好.

虽相较于预训练模型的预测结果, 基于情感词典方法的预测准确率略低, 但相比于机器学习算法, 用情感词典构建的分类模型更具有适用性. 基于扩展情感词典进行情感分析的过程如图 19 所示.

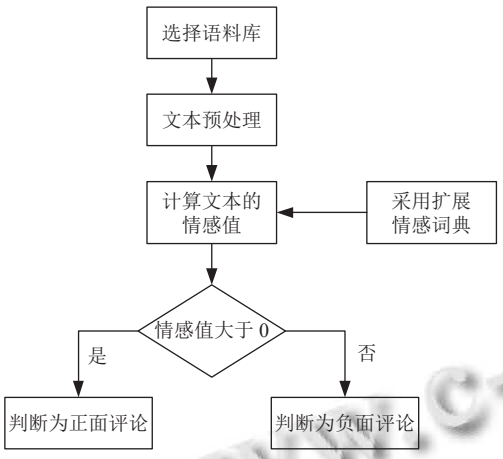

图 19 基于情感词典的情感分类过程

3.6 实验结果分析

本文采用两套研究方案进行分析比较, 一是基于扩展情感词典的方法, 二是基于预训练模型的方法. 在大连理工情感词典的基础上, 添加新的情感词以及调整在不同场景下相应词性的权值, 使得扩展的情感词典更加细化, 从而提高情感计算结果的精确性. 本文利用结巴 (Jieba) 分词组件, 分别对正向和负向的文本进行分词操作, 将读取出的每一个停用词进行去符号操作, 以提高分类模块匹配的精确度. 在得到正负语料的特征词文本后, 使用 Word2Vec 工具进行向量转换, 即让文本信息转为词向量数据. Wiki 中文语料是常用的语料库, 其中包含丰富且全面的语料. 故本文从该语料得出的特征向量中获取所需的特征词向量, 以此作为实验的输入数据. 由于 Word2Vec 模型的模型训练得到的词向量维度过大不便于分析, 故使用降维算法 (PCA) 对结果进行操作, 得到了较好的模型输入. 而在预训练模型的实验中采用基于 BiLSTM 神经网络进行模型的搭建, 该方法的优点是基于 LSTM 能够通过处理过程学习所需记住的内容和要忘记的内容, 进而更方便的获取较长距离的关联关系[41].

经过对实验结果的分析, 发现利用机器学习算法中的支持向量机、基于 BiLSTM 构建的预训练模型和基于扩展情感词典方法得到的分析结果准确率更高. 原因在于: 1) SVM 是一种有监督的学习[40,42] 方法, 是能够进行数据二分类的分类器. 它是线性的分类, 且具有鲁棒性和稳定性等特点. 2) 支持向量机在遇到复杂情况时能够把样本数据映射到高维度的空间, 这样确保了在新空间里的数据能够线性可分. 3) 预训练模型可以通过学习内容丰富的语料库上的常见表示方法来辅助完成后续的工作. 4) 预训练可以给模型提供优秀的初始化操作, 这往往有好的泛化体验, 而且可以提高选定目标的收敛速度. 5) 预训练能够被看作是一种符合规则化的过程, 它可以防止一些较小数据的过度拟合. 6) 扩展情感词典的方法通过提取特征词, 赋予相应权重的方式, 将文本数据转换为数值进行多元线性回归量化运算, 很大程度上避免了出现大的误差. 7) 基于扩展情感词典的方法使用两个标准进行评价, 即精确率和召回率, 从而提高分析结果的准确性.

3.7 实验对比

近年来在针对酒店评论情感分析这一领域, 为提高分析的精确率, 不少研究者通过额外构建其他的词典进行实验, 作出了较为显著的贡献.但这种方法存在的不足之处就是所构建的词典里一切常用的词都被打上了唯一分数, 在日常的评论交流中, 往往会涉及到部分不带情感色彩的停用词, 这类词语会对文本情感值的计算造成影响. 其次, 相同的词可能有不同的词性, 其相应的情感数值评价也会有差别, 例如: (1) 这家酒店的环境很差, 地板很脏; (2) 这家酒店的自助早餐会提供脏咖啡, 很好喝. 很显然, “脏”字在第 1 句中显示负向含义, 但在第 2 句中却显示中性含义, 只是一个简单的名词修饰. 因此, 对此类问题的分类需要考虑一词多义, 否则必然会产生偏见. 因而本文在传统实验的基础上, 结合了预训练模型与机器学习算法进行实验, 诸如 BiLSTM 神经网络、BERT 表征模型[16] 等预训练模型以及 NB、SVM 等算法在文本情感分类领域具有通用性及显著性, 有监督的机器学习方法可以达到分类器训练准确度的较高标准, 从而弥补扩展情感词典受复杂句式限制的不足之处, 同时为扩充词汇的数量控制、高频词汇赋予权重的调整提供参考.

实验分析精确率结果如表 1 所示. 从实验数据上看, 直接使用基础词典 (以大连理工情感词典为例) 进行情感分析的精确率平均为 79.26%, 杨飞等人[13] 在此基础上采用扩展词典 (扩展高频词数为200 时) 的分析精确率平均为 91.96%, 本文采用机器学习算法得到的分析精确率平均为 88%, 而采用预训练模型得到的分

析精确率为 92.15%, 实现了小幅度的提升.

王宏鹏[43] 在对基础情感词典进行词汇扩充的基础上, 结合细粒度判别方法, 实现了对情感倾向强度的计算, 同时, 利用机器学习算法进行实验, 并根据分类的评价指标对结果进行了比较分析. 与本文的不同之处在于王宏鹏[43] 采用了 K 近邻分类, 这是机器学习中最基础的分类方式. 它的优点是对噪声和异常值的容忍度高, 分类精度高, 但是根据样本数据的不同, 计算量也随之产生变化, 这对计算机的内存需求较大. 本文则利用基于 BiLSTM 的预训练模型去优化计算量的问题. BiLSTM 的设计使得其适用于大部分文本类数据模型, 它能够很好地捕捉双向的语义依赖, 很大程度上降低了语义环境的影响, 其灵活的适应性和优秀的性能使得在分类的准确率上要高于采用 PCA 等传统方法构建的情感模型. 王宏鹏[43] 的结合细粒度判别、加权词向量[44] 和 SVM 的方法最终实验结果显示准确率为 85.7%, 低于本文的预训练模型预测准确率 92.15%, 由此看出使用预训练模型有助于提高分类效果.

表 1 实验分析精确率结果 (%)

方法	平均分析准确率
基础情感词典	79.26
扩展情感词典	86.0 (pos) / 84.0 (neg)
SVM算法	88.0
基于BiLSTM的预训练模型	92.15

4 结论与展望

本文针对酒店这一领域进行具体的模型构建分析, 在寻找有利于酒店发展的方向的同时, 对 NLP 技术的实用性进行验证. 通过构建适用于酒店评论分析的扩展情感词典, 结合机器学习算法进行建模分析, 能够得到较高的情感分析准确率. 由于网络情感评论内容的丰富化, 表情符号的普遍使用致使评论分析变得更为困难, 因此在进行扩展词典的构建时可以通过加入网络用语并赋予不同的情感值来适应当下的语义环境, 弱化否定词的叠加效用以减小对带有相反含义语句的分类效果的影响, 从而提高分析的精确度. 使用扩展的情感词典判别文本的积极和消极情绪具有良好的可移植性, 可以对不同结构的语句进行分类. 另外, 在构造机器学习的分类模型时, 采用预训练模型能够加速训练, 使得模型收敛更快, 学习数据中的普遍特征, 产生更好的泛化效果.

由于这种模型受训练样本的内容影响, 会对此具有一定的依赖关系, 所以对于测试集的数据来源, 不同方式得到的数据可能会导致最终的分类模型构建出现偏差, 从而影响测试结果. 因此对于情感分析的研究仍具有很广的研究空间以及研究价值.

参考文献

1 於雯, 周武能. 基于 LSTM 的商品评论情感分析. 计算机系统应用, 2018, 27(8): 159–163. [doi: 10.15888/j.cnki.csa.006483]

2 Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 2014, 5(4): 1093–1113. [doi: 10.1016/j.asej.2014.04.011]

3 刘斌. 融合社交情感分析的股市预测探究. 计算机系统应用, 2018, 27(2): 250–256. [doi: 10.3969/j.issn.1003-3254.2018.02.043]

4 中国互联网络信息中心. 中国互联网络发展状况统计报告. 北京: 中国互联网络信息中心, 2021.

5 樊娜, 蔡皖东, 赵煜. 基于最大熵模型的观点句主观关系提取. 计算机工程, 2010, 36(2): 4–6. [doi: 10.3969/j.issn.1000-3428.2010.02.002]

6 苏莹, 张勇, 胡珀, 等. 基于朴素贝叶斯与潜在狄利克雷分布相结合的情感分析. 计算机应用, 2016, 36(6): 1613–1618. [doi: 10.11772/j.issn.1001-9081.2016.06.1613]

7 Jiang L, Yu M, Zhou M, *et al*. Target-dependent Twitter sentiment classification. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland Oregon: Association for Computational Linguistics, 2011. 151–160.

8 徐勇, 张慧, 陈亮. 一种基于情感分析的 UGC 模糊综合评价方法——以淘宝商品文本评论 UGC 为例. 情报理论与实践, 2016, 39(6): 64–69.

9 王祖辉, 姜维. 基于粗糙集的在线评论情感分析模型. 计算机工程, 2012, 38(16): 1–4.

10 Hu MQ, Liu B. Mining opinion features in customer reviews. Proceedings of the 19th National Conference on Artificial Intelligence. San Jose: AAAI, 2004. 755–760.

11 史伟, 王洪伟, 何绍义. 基于语义的中文在线评论情感分析. 情报学报, 2013, 32(8): 860–867. [doi: 10.3772/j.issn.1000-0135.2013.08.009]

12 魏慧玲. 文本情感分析在产品评论中的应用研究 [硕士学位论文]. 北京: 北京交通大学, 2014.

13 杨飞, 吴颖丹, 王鑫颖. 基于基础词典扩展的中文酒店评论情感分析. 湖北工业大学学报, 2019, 34(1): 107–110. [doi: 10.3969/j.issn.1003-4684.2019.01.024]

14 潘常玮. 迁移学习中预训练中文词向量优化方法研究 [硕士学位论文]. 北京: 北京交通大学, 2018.
15 Zhang S, Zheng D, Hu X, *et al*. Bidirectional long short-term memory networks for relation classification. Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. 2015. 73–78.
16 刘文秀, 李艳梅, 罗建, 等. 基于 BERT 与 BiLSTM 的中文短文本情感分析. 太原师范学院学报 (自然科学版), 2020, 19(4): 52–58.
17 陈翠娟. 改进的多项朴素贝叶斯分类算法和 Python 实现. 景德镇学院学报, 2021, 36(3): 92–95. [doi: 10.3969/j.issn.1008-8458.2021.03.032]
18 Yosinski J, Clune J, Bengio Y, *et al*. How transferable are features in deep neural networks? Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 3320–3328.
19 汪廷华, 陈峻婷. 核函数的选择研究综述. 计算机工程与设计, 2012, 33(3): 1181–1186. [doi: 10.3969/j.issn.1000-7024.2012.03.068]
20 秦淼. 二分类问题非平行超平面支持向量机的应用与优化 [硕士学位论文]. 长春: 吉林大学, 2020.
21 葛霓琳. 基于词典和机器学习的酒店评论情感分析 [硕士学位论文]. 镇江: 江苏科技大学, 2019.
22 余同瑞, 金冉, 韩晓臻, 等. 自然语言处理预训练模型的研究综述. 计算机工程与应用, 2020, 56(23): 12–22. [doi: 10.3778/j.issn.1002-8331.2006-0040]
23 刘睿珩, 叶霞, 岳增营. 面向自然语言处理任务的预训练模型综述. 计算机应用, 2021, 41(5): 1236–1246.
24 Peters ME, Neumann M, Iyyer M, *et al*. Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans: NAACL, 2018. 2227–2237.
25 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020.
26 Chang WC, Yu FX, Chang YW, *et al*. Pre-training tasks for embedding-based large-scale retrieval. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: ICLR, 2020.
27 谭咏梅, 刘姝雯, 吕学强. 基于 CNN 与双向 LSTM 的中文文本蕴含识别方法. 中文信息学报, 2018, 32(7): 11–19. [doi: 10.3969/j.issn.1003-0077.2018.07.002]
28 陈葛恒. 基于极性转移和双向 LSTM 的文本情感分析. 信息技术, 2018, (2): 149–152.
29 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv: 1301.3781, 2013.
30 郑强. 浅谈多元线性回归模型. 女人坊 (新时代教育), 2020, (1): 00190.
31 郭丽环, 韩越, 王伟. 在线评论对旅游者酒店选择的影响——基于细粒度文本情感分析. 泉州师范学院学报, 2019, 37(6): 93–100.
32 鲁继文. 基于 Scrapy 的论文引用爬虫的设计与实现. 现代计算机 (专业版), 2017, (9): 131–133.
33 王胜. 基于 XPath 的网页信息抽取 [硕士学位论文]. 合肥: 中国科学技术大学, 2006.
34 王佳楠, 梁永全. 中文分词研究综述. 软件导刊, 2021, 20(4): 247–252. [doi: 10.11907/rjdk.201673]
35 张启宇, 朱玲, 张雅萍. 中文分词算法研究综述. 情报探索, 2008,(11):53-56.[doi:10.3969/j.issn.1005-8095.2008.11.022]
36 Le QC, Mikolov T. Distributed representations of sentences and documents. Proceedings of the 31th International Conference on Machine Learning. Beijing: ICML, 2014. 1188–1196.
37 肖元君, 吴国文. 基于 Gensim 的摘要自动生成算法研究与实现. 计算机应用与软件, 2019, 36(12): 131–136. [doi: 10.3969/j.issn.1000-386x.2019.12.021]
38 Abadi M, Agarwal A, Barham P, *et al*. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv: 1603.04467, 2015.
39 Ji YF, Song LB, Sun J, *et al*. Application of SVM and PCA-CS algorithms for prediction of strip crown in hot strip rolling. Journal of Central South University, 2021, 28(8): 2333–2344. [doi: 10.1007/s11771-021-4773-z]
40 薛贞霞. 支持向量机及半监督学习中若干问题的研究 [博士学位论文]. 西安: 西安电子科技大学, 2009.
41 李舟军, 范宇, 吴贤杰. 面向自然语言处理的预训练技术研究综述. 计算机科学, 2020, 47(3): 162–173. [doi: 10.11896/jsjkx.191000167]
42 王倩文. 基于 SVM 的验证码识别算法研究. 黑龙江科技信息, 2013, (20): 164.
43 王宏鹏. 基于词典与机器学习的酒店评论情感分析的研究 [硕士学位论文]. 大连: 大连交通大学, 2020.
44 胡万亭, 贾真. 基于加权词向量和卷积神经网络的新闻文本分类. 计算机系统应用, 2020, 29(5): 275–279. [doi: 10.3969/j.issn.1003-3254.2020.05.041]

(校对责编: 牛欣悦)