

基于深度学习的优质在线课程影响因素分析及方法研究

孙宗仁

四川文理学院, 四川 达州 635000

摘 要 : 随着信息技术和互联网资源的快速发展, 大数据已深入到生活的各个领域, 深度学习技术已成为不可或缺的战略资源。在全球化与科技双重驱动的教育新时代, 优质在线课程的需求日益增长, 深度学习成为提升教育质量的关键。尽管在线课程取得显著成就, 但是课程质量参差不齐的问题仍制约其发展, 亟需技术手段去解决。本文采用多学科交叉视角, 运用文本挖掘技术分析在线课程评论, 提出神经网络的卷积核结合 BLSTM 更好地提取局部和全局的特征信息, 揭示优质课程的关键因素, 为在线教育优化提供理论支持和实践建议。研究通过自然语言处理方法比较卷积核结合 BLSTM 模型在情感分析中表现优异, 有效解决了文本深层信息提取问题, 为在线教育发展提供科学建议。

关 键 词 : 深度学习; 在线课程; 多学科交叉; BLSTM

Analysis and Research on Influencing Factors and Methods of High-Quality Online Courses Based on Deep Learning

Sun Zongren

Sichuan University of Arts and Sciences, Dazhou, Sichuan 635000

Abstract : With the rapid development of information technology and Internet resources, big data has penetrated into all aspects of life, and deep learning technology has become an indispensable strategic resource. In the new era of education driven by globalization and technology, the demand for high-quality online courses is increasing day by day. Deep learning has become the key to improving educational quality. Although online courses have achieved remarkable accomplishments, the problem of uneven course quality still restricts their development, and technical means are urgently needed to solve it. This paper adopts a multi-disciplinary perspective and uses text mining technology to analyze online course reviews. It proposes that the convolutional kernel of neural networks combined with BLSTM can better extract local and global feature information, revealing the key factors of high-quality courses and providing theoretical support and practical suggestions for the optimization of online education. The research uses natural language processing methods to compare the performance of the deep learning BLSTM model in sentiment analysis, effectively solving the problem of extracting deep information from texts, and providing scientific suggestions for the development of online education.

Keywords : deep learning; online courses; interdisciplinary integration; BLSTM

一、研究背景和意义

随着互联网与大数据技术的深度融合, 在线学习技术正悄然改变着传统教育的传授模式, 尤其在挖掘课程内容与教学模式潜力方面展现出卓越能力。在线学习以便捷、高效和多元化的特点打破了时空的限制, 迅速兴起并成为传统教育的重要补充, 疫情期间更呈指数级增长。深度学习技术的应用为在线课程注入个性化体验, 拓展了教育可能性。然而, 课程数量激增导致质量参差不齐, 学习者难以快速筛选优质内容, 加之数据收集与整理的挑战, 在线教育的持续发展仍需克服诸多问题。

近年来, 数据科学领域在文本处理方面取得突破性进展, 中国外学者在文本情感分析等领域成果显著。Taboada 等^[1]提出

通过引入语音注释的情感词典来提升情感识别的性能。陶梦园^[2]通过构建情感词典, 显著提升了评论文本情感极性分析的分类效果。林江豪等^[3]则对收集的数据进行了文本情感倾向性分类。黄天元^[4]在文本数据挖掘方面详细介绍了利用 TF-IDF 值实现文本向量化的具体步骤。潘越、高雪芬^[5]采用 TF-IDF 方法提取关键词特征词, 进一步优化了分析过程。李晶、罗泰晔^[6]通过信息熵和组合力指标对高频关键词进行聚类, 提升了分析的精准性。赵殿国^[7]运用数据分布特性降低计算复杂度, 显著提高了聚类效率。这些研究共同推动了文本数据处理和分析技术的发展。

面对庞杂的教育数据, 深度学习技术应运而生, 为教育数据分析提供了新的解决方案。Kandhro 等^[8]通过预训练模型和长短时记忆网络 (LSTM) 分析评论文本的情感倾向, 为精准识别和打

项目信息: 达州市社科联重点项目: SCMF202302。

作者简介: 孙宗仁, 四川文理学院统计学专任教师。

造优质课程开辟了新路径。Imran等^[9]利用预训练嵌入模型,在情感分析任务中证明了其性能的优越性。Maryam^[10]基于评论数据集对多种模型进行评估,结果表明深度学习算法在情感分类任务中表现优异。杨志林等^[11]提出结合自回归和自编码的模型,验证了双向编码的可行性,进一步拓展了深度学习在教育数据分析中的应用范围。Ma D^[12]等通过长短时记忆单元生成情感特征向量和范围词来获得上下文情感特征。这些研究为教育数据的深度挖掘和情感分析提供了重要的技术支撑。尽管已有研究取得一定进展,但数据来源局限性和研究方法不普适性导致结果难以全面反映课程真实情况,结论的准确性和可靠性有待提升。因此还需研究优化方法、拓展数据来源,以增强分析的全面性与可靠性。

本文探讨利用深度学习技术重构优质课程评判体系,分析影响课程质量的关键因素,为课程开发与优化提供科学依据。传统评价方法受限于主观判断,难以全面反映教学质量,而深度学习能从海量数据中提取关键信息,揭示教学效果,提供更精准、客观的分析工具。本文通过深度分析学习者评论文本,挖掘潜在信息,精准把握核心因素,探索多因素间复杂关系,完善在线教育质量理论。研究结果为课程开发者提供优化方向,为政策制定者提供实证依据,并帮助学习者提升学习体验。通过自然语言处理技术,本研究减少了人工干预,提高分析效率,为在线教育提质增效提供理论支持与实践指导,深化对在线课程质量的理解,推动教育技术创新与应用。

二、相关理论及方法介绍

本研究以中国大学MOOC平台为研究对象,通过R语言爬虫技术获取在线课程的用户评论文本作为原始数据,并对其进行文本预处理、分词、词频统计及编码,以提取用户关注的核心内容。研究采用TF-IDF方法对文本进行向量化处理,通过最佳聚类簇数检验实现文本聚类,并结合关键词提取各类用户的关注点。基于情感词典对评论文本进行情感极性分析,计算积极与消极评论的占比,最后利用LDA主题模型提取主题词,深入分析影响在线课程质量的正负向因素。为了突破传统研究对课程指标的依赖,通过深度挖掘学习者评论文本中的潜在信息,精准识别影响在线课程质量的核心因素,并探索多因素间的复杂关系,进一步完善了在线教育质量理论。研究结果为课程开发者提供了优化课程设计的科学依据,为政策制定者推动在线教育规范化发展提供了实证支持,同时帮助学习者在海量课程中筛选优质内容,提升学习体验。

通过自然语言处理技术,本研究显著减少了人工干预,提高了分析效率,为在线教育的提质增效提供了理论支持与实践指导。这一研究不仅深化了对在线课程质量的理解,也为在线教育技术的创新与应用奠定了坚实基础。

情感分析是基于情感词典对词汇赋予权重分配的方法便于理解文本中的情感内容,因而在在线课程评价研究中具有重要意义。基于词典的情感分析方法通过预先构建的情感词典,将评论中的词汇与词典中的情感词进行匹配,从而判断评论的情感极

性。这种方法简单直接,但面临情感词典更新滞后于网络语言发展的问题,难以识别新兴表达。卷积网络则已广泛应用于从丰富内容中提取特征,为情感分析和课程推荐提供了技术支持。总体而言,各种方法各有优劣,需根据具体场景选择合适的技术方案,以提升在线课程评价的准确性和实用性。

然而,已有的情感分析对文本所在的全局与局部信息未能充分利用,为此本研究提出的神经网络的卷积核结合BLSTM更好地提取局部和全局的特征信息方法可用于提取不同方面文本特征的信息,得到综合的特征,以实现更好的情感分类效果。

神经网络依据结构提取有效的局部特征,对文本使用卷积核操作所得的特征进行降维:

$$c_i = W^{(c)} * [\text{conv}_k(x_{(i+k-1)})] + s^{(c)} \quad (1)$$

其中对应层 $W^{(c)}$ 为权重, $s^{(c)}$ 为偏置。卷积核大小为 k ,卷积的过程操作是 conv , conv_k 为 $x_{(i+k-1)}$ 词向量拼接的结果。

为了更好地理解句子就需捕捉到句子的上下文特征,借助词向量化后的文本序列:

$$o_i = [f(x_{(i)}), b(x_{(i)})] \quad (2)$$

这里 $[f(x_{(i)}), b(x_{(i)})]$ 表示前后向在 x_i 处的输出拼接结果。

通过上述分析,把局部的输出特征用 $g_{(i)} = g\{g_1, g_2, \dots, g_n\}$ 表示,全局的输出特征用 $h_{(i)} = \{h_1, h_2, \dots, h_n\}$ 表示, $n=1, 2, \dots, n$ 。推导得出的加权特征向量为:

$$w_i = w_N^T \tanh(w_{N1}g_i + w_{N2}h_i + s_N)$$

这里 $N1$ 为局部的权重, $N2$ 为全局的权重, s_N 是偏置。

令 $\hat{R}$ 为预测矩阵, A 矩阵的行为用户嵌入, B 矩阵列为项目嵌入。令 b 和 c 是分别包含针对每个用户以及每个项目的偏置向量。因此,双线性预测获得:

$$\hat{R}_{\mu,i} = b_{\mu} + c_i + \sum_j A_{\mu,j} B_{j,i} \quad (3)$$

当用户嵌入和项目嵌入首次缩小到低维度时就可以方便的可视化,或者可以将用户或项目彼此之间进行比较。获得这这些嵌入的一种方式是对实际目标的矩阵 R 进行奇异值分解。这对应于将 $\hat{R} = UDV^T$ 分解为两个因子的乘积,低秩矩阵 $A = UD$ 和 $B = V^T SVD$ 的一个问题是它以任意方式处理缺失条目,如同它们对应于目标值0。

在自然语言处理中通常模型需要考虑分类正确的正样本数占分类器判定为正的样本数及占实际正样本数比例大小,鉴于此考虑F1值指标对模型评估:

$$F1 = \frac{2(TP)^2}{(TP + FP)(TP + FN)} \quad (4)$$

其中 TP 表示将正类评估结果预测为正类评估结果的样本数, FP 表示负类评估结果预测为正类评估结果的样本数, FN 正类评估结果预测为负类评估结果的样本数。

在深度学习BLSTM中对当前时间步考虑到前后的序列信息,这增强了模型评估的稳健性。因此本文运用卷积核结合BLSTM技术,全面挖掘在线课程评论文本中的复杂信息,旨在揭示影响优质课程的关键因素,为在线教育发展提供支撑。关键词提取以无监督方法为主,如文本聚类,因其普适性更强。

三、应用研究

数据预处理包括评论清洗与结构化处理，研究对预处理后的数据进行分词处理，统计词频以提取文本关键词，并计算分词文本的 TF-IDF 值，将文本数据向量化为结构化数据。基于情感词典对评论文本进行情感极性分析，识别对应的影响因素。最后，采用 LDA 主题模型对在线课程受欢迎影响因素的优先级进行排序。由于网络课程评论数据在评分和评论信息方面缺乏最新资源，本文采用网络爬虫技术从互联网 HTML 代码中自动提取所需数据，包括用户 ID、网页链接及评论文本等。考虑到数据的有效性与时效性，评论文本及评论时间对分析至关重要。通过选取涵盖多学科、多层次且评论数量充足的课程样本，结合爬虫技术获取评论数据，并经过严格的剔除无效信息、使用停用词表和 Jieba 词库进行分词及去停用词处理预处理操作，最终生成高质量数据集，为后续分析提供可靠基础。

利用爬虫技术从中国大学 MOOC 平台 2019–2023 年部分用户课程评论信息作为研究对象。在进行用户评论爬虫时，由于采集量较大不可避免会出现一些重复、无效等冗余数据，重复评论数据可设置相应条件初步去除。然而评论数据质量参差不齐，甚至存在许多没有意义的文本数据，这就需要机械压缩法对去重后的文本做进一步处理，同时还需要删除评论字数过少的短句，有效的保留用户情感文本。中文分词由停用词和分词两部分构成。停用词有助于简化文本数据，提高处理的效率，同时减少噪音对后续分析的干扰，达到了较好的分词效果。本文研究采用混合模式对评论文本进行分词处理。

为了解在线课程受众的基本情况，通过对课程评分情况和评论分布年份的描述性分析，预测各类课程近年来的发展趋势，绘制出词云图探索评论关键词，初步分析在线课程受欢迎可能存在的影响因素。

> 图1 课程概述核心词云图



词云图 1 中词汇的大小通常反映了对应词汇在整个文本中的重要性。从图中可以看出“课程”、“学习”、“学生”等词汇相对较大，这表明它们在课程中占据较为明显比例的地位。这些词汇涵盖了教育的多个方面，还体现了在线课程的多元性和包容性。

> 图2 汇总词云图



从图2四类课程的词云图可知，词频较高的关键词侧面反映了用户对网络课程的关注重点。可为接下来课程受欢迎影响因素的寻找提供初步探索方向。并且从词中可以看出积极的评论较多，故课程总体评分较高。

将课程评论数据词结果整合之后进行词频统计，筛选词频最高的词语进行汇总，有助于理解和解读课程评论。

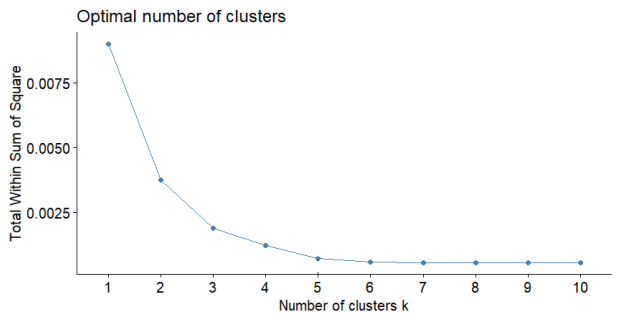
表1 汇总分词表部分展示

Trem	Frequency	Term	Frequency
老师	1175	作业	124
内容	736	素材	121
课程	689	受益匪浅	121
学习	369	详细	117
细致	363	丰富	116
讲解	280	感谢	112
知识	276	讲课	110
学到	219	了解	110
不错	211	提升	105
内容	209	讲得	99

表1中可以发现出现频率最高的词汇，从这里就可以初步看出学生对课程的关注情况，引导下一步的探索课程受欢迎影响因素的分析方向。

分析在线课程评论文本的情感倾向是一个循序渐进的过程，可通过关键词的提取对文本进行向量化，得到结构化数据；再将所得关键词进行聚类分析得到在线课程受欢迎的影响因素；最后通过构建情感词典的方法分析影响因素关键词的情感倾向。计算个关键词的 TF-IDF 值后发现分词后的文本具有明显的类别差异，可考虑利用无监督的聚类算法对分词后的文本进行聚类分析，本研究采用 K-means 法对文本进行聚类。

> 图3 聚类簇数



根据聚类簇数图结果，当聚类数据量为四类时，数据点之间的相似性和差异性得到了较好的平衡，这意味着每一类的文本数据都具有较高的内部相似。

> 图4 每类文本的关键词显著差异

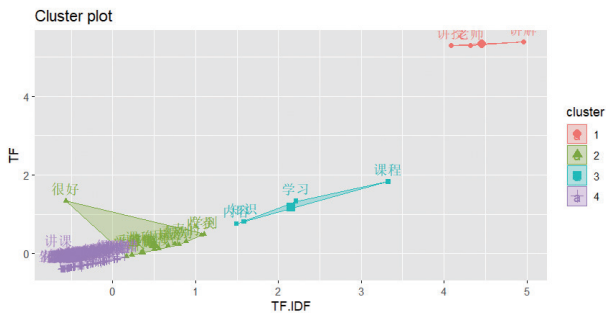


图4表明，将文本关键词分为四类后，每类文本的关键词具有显著差异。

基于情感词典对词汇进行权重分配，使我们能够深入理解文本中的情感内容。文本情感分是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程，可以通过研究这些主观色彩的评论来了解大众的看法。

在情感倾向中，-1代表消极性情感，0代表中立性情感，1代表积极性情感，情绪值越高代表情感倾向越积极。通过对每条评论文本进行分析，可以探索该用户对相应课程的情感态度，从而研究课程的受欢迎程度。

表2 情感倾向统计

情感倾向	评论数
积极	3099
中立	334
消极	155

从表2可知，对所有评论数进行统计之后得到评论的情感倾向，其中绝大部分是积极性评论，中立性或消极性评论很少，可见大多数学习用户对在线课程较为满意。

但由于中立性评论难以对其细节进行倾向性提取，对主题特征抽取没有明显情感价值，故后续只对积极性和消极性两类评论进行分析。运用 LDA 主题模型可进一步确定优质在线课程的影响因素，并得出各个影响因素的正负向关键词。图模型用于刻画复杂系统中多个变量或观测之间的关系，是高维数据结构信息提取中的重要工具，其基本思想是通过揭示多维随机向量的相依结构刻画变量之间的条件独立性。使用 LDA 分析工具综合运用多种分析，深入挖掘文本中的潜在信息，构建理解课程影响因素的基础框架；通过确定 LDA 模型的最佳主题数，分别获取积极评论和消极评论的热点主题，详细分析学习者的学习体验、期望以及遇到的问题。

本文通过评估不同主题数模型的困惑度来选择最佳的模型主题数，其计算公式为：

$$\text{per}(D_{\text{test}}) = \exp \left\{ \frac{-\sum_{d=1}^M \log(p(z_d))}{\sum_{d=1}^M N_d} \right\} \quad (5)$$

其中， M 是测试文档的数量， N_d 是第 d 篇文本大小， z 是

文档。模型生成能力越强， per 值越小。

> 图5 主题数与困惑度

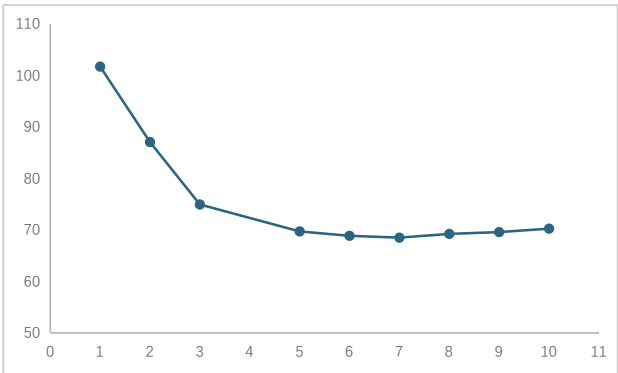


图5显示，当主题数为1至4时，困惑度下降速度较快，在此之后困惑度变化较为平缓，故将主题数确定为4比较合适。

由于主题数确定为4，与上文在进行主题词挖掘时得到的主题数相一致，故将 LDA 主题模型的四个主题分别确定为课程内容、教师讲授、学习成效、以及教学资源四个方面的主题内容结果，得到如下所示主题词得分情况：

表3 主题词得分情况

TOPIC	Total Score	Average Score
课程内容	297.13	0.65
教师讲授	243.47	0.70
学习成效	276.64	0.70
教学资源	211.20	0.74

表3显示课程内容的总得分最高，说明评论文本当中对此部分内容提及最多，但仍然表明课程内容的重要性程度最高；其次是学习成效，比较而言，评论涉及教学资源方面的较少。

为了验证卷积核结合 BLSTM 方法在文本挖掘上的有效性，选用本文在线课程数据集和 https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/weibo_senti_100k/intro.ipynb 数据集。

表4 卷积核结合 BLSTM 深度学习实验结果

模型	F1
在线课程数据集	88.46
weibo_senti_100k	89.26

从表4中观察 F1 值可知，分类正确的正样本数占分类器判定为正的样本数及占实际正样本数比例高，论证了深度学习在文本挖掘分析中的有效性。

在计算语言学中，文档以词袋格式表示，其维度可能是数万，通常主题的个数远小于词项的个数，使用主题模型进行降维。在以文本为处理对象的领域中，降维后的新坐标往往具有更显著的语义特征，得到的模型解释性更强。

四、结论与展望

在线课程因其资源广泛、学习方式灵活等特点，已成为学生自学的主要方式之一。尽管在线课程行业整体发展趋势良好，但

仍存在诸多问题。通过自然语言处理技术，本文有效挖掘了在线课程受欢迎的影响因素。利用爬虫技术采集评论数据，经整理后采用 TF-IDF 方法将文本向量化，初步提取关键词，为深层次分析奠定基础，同时运用聚类算法拓展和深化关键词研究，揭示主题的复杂性和多样性。研究比较显示卷积核结合 BLSTM 模型在

情感分析中表现优异。未来可进一步探讨研究主题的深层机制，并将本方法应用于其他领域，验证其普适性和有效性。

随着自然语言处理技术的进步，深度学习将能更精准地分析大规模文本数据，揭示语言背后的深层含义和结构，从而推动网络在线课程的实际发展，也为相关领域提供了起点和思路。

参考文献

[1]Taboada M, Brooke J, Tofiloski M. Lexicon Based Methods for Sentiment Analysis [J]. Computational Linguistics, 2011, 37(2): 267-307.

[2]陶梦园. 基于文本挖掘的民宿房客满意度分析 [D]. 云南财经大学, 2023.

[3]林江豪, 周咏梅, 阳爱民等. 基于词向量的领域情感词典构建 [J]. 山东大学学报 (工学版), 2018, 48(03): 40-47.

[4]黄天元. 文本数据挖掘 [M]. 北京: 机械工业出版社, 2021: 92-94.

[5]潘越, 高雪芬. 大学数学精品慕课课程质量影响因素研究——基于评论文本挖掘的视角 [J]. 浙江理工大学学报 (社会科学版), 2024, (3): 1-9.

[6]李晶, 罗泰晖. 基于文本挖掘的 5G 技术研究热点分析 [J]. 科技管理研究, 2020, 40(19): 153-158.

[7]赵殿国. 面向无监督学习的大数据聚类方法及其应用研究 [J]. 统计与咨询, 2023, (06): 7-11.

[8]Kandhro I A, Wasi S, Kumar K, et al. Sentiment Analysis of Students' Comment Using Longshort Term Model [J]. Indian Journal of Science and Technology, 2019, 12(8): 1-16.

[9]Imran A S, Daudpota S M, Kastrati Z, et al. Cross-cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets [J]. Ieee Access, 2020, 8: 181074-181090.

[10]Edalati M, Imran A S, Kastrati Z, et al. The Potential of Machine Learning Algorithms for Sentiment Classification of Students' Feedback on MOOC [C]. Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference Volume 3. Springer International Publishing, 2022: 11-22.

[11]Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding [J]. Advances in Neural Information Processing Systems, 2019, 32.

[12]Ma D, Li S, Zhang X, et al. Interactive Attention Networks for Aspect-Level Sentiment Classification [C]. Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017: 4068-4074.