基于 BERT-LDA 的中文学习 APP 评价指标体系构建研究 (Construction of an Evaluation Indicator System for Chinese Learning Apps Based on BERT-LDA)

张邝弋
(Zhang, Kuangyi)
北京语言大学
(Beijing Language and Culture University)
zky.staybirds@foxmail.com

侯尚余
(Hou, Shangyu)
云南大学
(Yunnan University)
houshangyu@stu.ynu.edu.cn

宋靖雯
(Song, Jingwen)
云南大学
(Yunnan University)
songjingwen@stu.ynu.edu.cn

肖锐
(Xiao, Rui)
云南大学
(Yunnan University)
ruixiao@ynu.edu.cn

摘要：本研究围绕中文学习 APP 在信息化及智能化趋势下的评价标准和质量挑战，提出了一种基于 BERT-LDA 模型的主题聚类算法，并结合 LLMs 的专家模型主题提取方法，从评价内容（内容质量）、评价过程（用户体验）、评价效果（学习成效）等核心维度构建了中文学习 APP 的多维度、动态化评价指标体系，并在情感分析任务验证其有效性，最后从智能化、动态化以及安全性等方面指明了未来国际中文教育数字资源评价指标体系构建的未来方向及风险挑战。

Abstract: This study explores the evaluation standards and quality challenges associated with Chinese learning apps in the context of increasing informatization and intelligence. It introduces a topic clustering algorithm derived from the BERT-LDA model and integrates an expert model for topic extraction utilizing Large Language Models (LLMs). A multidimensional and dynamic evaluation indicator system for Chinese learning apps is developed, focusing on core dimensions such as evaluation content (content quality), evaluation process (user experience), and evaluation outcomes (learning effectiveness). The validity of this system is confirmed through sentiment analysis tasks. Lastly, the study identifies future directions and potential risk challenges for creating evaluation indicator systems in international Chinese education digital resources, emphasizing intelligent, dynamic, and secure approaches.

关键词：中文学习 APP, BERT-LDA, 大语言模型

Keywords: Chinese learning APP, BERT-LDA, Large language model

1. 引言

以 ChatGPT 为代表的大语言模型(Large Language Model, LLM)在智能教育助手、课程定制、学习评价和语言交互等多个领域的应用，进一步突显了人工智能技术在全球中文教育普及与深化进程中的核心驱动作用，并揭示了 LLM 作为通用人工智能发展的重要里程碑，对中文教学产生了深远的影响（Wu et al., 2023）。

聚焦在移动学习的特定领域，中文学习 APP 凭借其方便高效的学习模式和出色的内容个性化功能，正逐渐成为众多中文学习者掌握知识和提高语言能力的关键工具（郭晶等，2021）。然而，对于如何精确有效地评价这些中文学习 APP 的质量和效率，以及它们在推动中文教学向数字化、智能化转型中所做的实际贡献，仍缺少一套广泛适用的评价指标体系。

传统的评价方法如层次分析法(Analytic Hierarchy Process, AHP) （Kharat et al., 2016）和德尔菲专家咨询法(Delphi Method)（Alon et al., 2025）虽在一定程度上解决了评价复杂系统的问题，但在应对快速迭代更新的学习环境，尤其是融合了先进人工智能技术的中文学习 APP 时，这些方法的局限性日益显现（王春枝等, 2011; 邓雪等, 2012）。鉴于此，本研究旨在借鉴现有评价理论及方法的基础上，提出一种基 BERT(Bidirectional Encoder Representations from Transformer)-LDA(Latent Dirichlet Allocation)型[1]的主题聚类算法，同时基于 LLMs(Large Language Models)[2]的专家模型对聚类主题进行提取，从而构建动态适应性增强的中文学习 APP 评价指标体系，并在实际案例中对指标体系进行验证，以实现对中文学习 APP 的多维动态评价，最终为中文学习 APP 的持续优化改进与健康发展提供有力支持和科学依据。

2.文献综述

评价指标作为量化评价与决策支撑的重要依据，在数据分析和业务优化过程中扮演着核心角色（虞晓芬等，2004）。数据挖掘作为一种强大的工具和技术手段，为评价指标的精准量化设定与深层次洞察力发现提供了强有力的技术支持和实质性的改进空间。在现代教育信息化背景下，中文学习 APP 作为普及语言学习及促进文化交流的数字媒介，能够通过对海量数据的挖掘提炼出有价值的信息。因此，构建一套完善的指标体系至关重要， 这不仅能有效实现对教学效果的实时监测与精确度量，也能深入剖析用户行为特征。这一举措将有力驱动中文智能教学效率的提升、数字化教育资源管理水平的进步，使得相关领域的研究和实践逐步摆脱传统上过度依赖人工操作、孤立分散的数据分析方式和相对有限的个性化服务，进而迈向自动化、规模化以及高度集成化的智慧教育新时代。

[1] https://www.kaggle.com/code/dskswu/topic-modeling-bert-lda

[2] LLMs（Large Language Models）为多个大语言模型，指基于提示引导的群体智能；LLM（Large Language Model）为单个大语言模型。

2.1 传统指标体系构建的研究现状

经验驱动的传统指标体系构建方法主要依赖专家经验和定性分析手段，具有较强的主观性和过程复杂性。例如，梁宇等（2023）综合运用德尔菲法和层次分析法，从专家经验和逻辑推理出发构建了国际中文教材评价指标体系；杨甜等（2023）基于广泛的问卷调查和用户反馈定性数据，构建了国际中文教师智能素养指标体系；方紫帆等（2023）参照《国际中文教师专业能力标准》[3]，结合理论研究与实践需求，构建了国际中文教师数字素养指标体系；程涛等（2024）利用德尔菲专家咨询法，尝试性地建构了具有中国特色的跨文化职业胜任力评价指标体系；宫雪等（2023）运用词频统计、多词序列提取、搭配分析等量化手段改进了国际中文教材评价指标基础框架的构建方式，减轻了其原有的“重定性、轻定量”问题。由此可知，以层次分析法、德尔菲方法等为代表的经验主义与半定量研究策略，在语言教学评价、教育政策制定及课程质量评估等多个领域发挥了重要作用（袁海红等, 2014；杨绪辉，2019）。然而，此类方法同样存在显著局限性：首先，它们对大规模客观数据的利用不足，过度依赖专家的专业见解和判断，可能导致评价结果的主观性强、稳定性差；其次，建立指标体系的过程往往涉及多次循环的匿名咨询、意见整合、反馈调整等环节，周期长且成本高；最后，由于专家观点的主观偏倚以及数据采集阶段可能出现的操作不一致，所得到的评价指标权重分配和预测结果，在客观性和精确性方面可能与基于大数据挖掘方法所得出的结论存在一定差距。

2.2 基于数据驱动的指标体系构建研究现状

数据驱动(Data-Driven)是指利用大规模客观数据，结合统计学和机器学习技术，以数据内在规律为基础，自下而上地构建评价指标体系的过程（杨现民等，2017）。这种方法强调通过算法模型揭示数据间的深层关联和模式，克服传统经验主义方法的主观性和不确定性，从而提高评价体系构建的客观性、准确性和普适性。随着深度学习和自然语言处理技术的发展，数据挖掘和机器学习算法已在不同领域指标体系构建中广泛应用，并已历经多个发展阶段：（1）传统模型的独立应用。早期的数据驱动指标体系构建多依赖于 LDA 等单一的模型，这些模型在处理文本数据时，能够初步揭示数据中的隐含主题或模式。（2）模型融合与技术创新。随着对更深层次数据关联需求的增长，研究者开始探索模型的融合使用，旨在通过结合不同模型的优势来提升分析的全面性和准确性。这一时期 Convolutional Neural Networks（CNN）等深度学习模型因其强大的语境理解能力而被引入，与 LDA 等传统主题模型结合使用成为趋势。例如，贾海楠等（2023）的工作展示了 LDA 与扎根分析法的融合，Lai（2023）使用 CNN 和 Bi-LSTM 模型对已有指标体系进行验证，都是这一阶段创新的体现。此外，潘小宇等（2023）提出的 HBL-LDA 方法，则是模型集成思想的实践，它通过结合多种模型特性，提高了书法价值评估指标构建的效率与准确性。（3）面向特定领域的最优模型选择与定制化融合，研究更加注重模型优化，以适应特定领域的独特需求。李天义等（2024）等从文本特征融合的视角出发，创造性地结合了 BERT-LDA 与 K-means 聚类算法，针对绘画作品的价值要素

[3] https://shihan-org.chinese.cn/index/build/detail.html?id=239

进行深度挖掘，这种融合模型不仅继承了 BERT 对复杂语境的强大理解力，还利用 LDA 捕获主题结构，同时通过 K-means 进一步细化类别，实现了对绘画领域高度定制化的价值评估指标体系构建。这标志着数据驱动方法在特定领域应用趋向成熟，不仅追求技术的先进性，更强调模型与实际应用场景的紧密结合。由此可知，基于数据驱动与主题挖掘的研究方法与指标构建研究已结合得十分紧密。

2.3 中文学习 APP 研究现状

诸如 *Duolingo*、*HelloChinese* 等中文学习 APP 因其丰富的用户交互数据、多样的学习行为记录以及实时更新的内容反馈等数字化资源特征，为教育研究和个性化学习提供了前所未有的可能性和挑战。相关研究主要呈现出以下特点：第一，中文学习 APP 评价数量较少，覆盖面不足，难以全面反映各类产品的优劣（高传智等, 2025; 李姝姝等, 2025）；第二，中文学习 APP 评价维度较为单一，往往集中在功能设计或用户体验上，无法做到对教学内容、学习效果、技术性能等方面的综合评价（刘永俊, 2021）；第三，缺乏系统的评价理论作为支撑，容易导致评价标准不一、主观性强的问题（杨倩, 2018）。由此可知，借助大数据与人工智能技术高效、科学地构建更具针对性、动态适应性的中文学习 APP 指标评估体系显得尤为迫切且必要。

2.4 国际中文教育数字化资源多维评价

人工智能、大数据、云计算、虚拟现实等技术的不断进步与广泛应用正深刻重构国际中文教育生态，其不仅促进目标受众角色从传统语言习得者向具备多元文化表征的网络用户转型，更通过技术赋能的增效机制，显著提升了该群体对数字化学习工具的探索动能、应用黏性及其对技术的接受度和融合能力。在这一演变过程中，赵学铭等（2017）基于模糊层次分析法对学习 APP 的易用性进行评价；张熠等（2019）基于 D－S 证据理论，从用户体验视角构建了针对中国大陆学习 APP 的指标，验证了用户体验与 APP 使用、内容资源之间的紧密关系；蔡燕等（2022）基于技术接受模型(Technology Acceptance Model, TAM），构建了解释和预测中文学习者在线直播课程学习意愿的理论模型；梁宇等（2023）则更进一步以技术接受扩展模型为理论框架，构建了中文数字学习资源使用意愿模型，并且特别强调了感知易用性、感知有用性、使用态度具有关键的中介作用。由此可知，中文学习 APP 作为数字教育资源的一种创新形式，显著增强了学习的便捷性和互动性，促进了个性化学习路径的发展。因此，从用户体验视角出发，系统性地评价与分析用户对该类新兴数字资源的应用效果及内容反馈对于优化产品设计、提升教学效果至关重要。

2.5 基于 LLMs 的专家模型主题提取与效果评价

LLM 在多项基准测试中展现出媲美人类专家的水平和表现（Achiam et al., 2023）。提示工程作为一种有效引导 LLMs 的方法，通过专门设计的提示词或短语，能够在零样本（Kojima et al., 2022）或少量样本（Brown et al., 2020）条件下显著提升模型在特定 NLP 任务上的表现。进一步而言，群体智能决策机制能进一步强化

LLMs 的性能，甚至在某些任务上超越人类（Wu et al., 2023; Jang et al., 2023）。例如，何多魁等（2025）提出了一种微调大语言模型驱动的短文本动态主题建模方法，通过结合指令微调、检索增强生成(Retrieval-Augmented Generation, RAG)和聚类技术，有效提升了主题识别的准确度，并揭示了主题的演化规律，为主题建模提供了新的思路和方法。翟洁等（2025）则针对计算机实验报告评阅过程中评语模板化、缺乏个性化内容等问题，提出了基于 LLM 的个性化实验报告评语自动生成框架，通过主题-评价决策-集成提示策略，实现了从实验要求和代码质量需求中抽取评价体系，自动生成具有可解释性的实验评语，提高了评阅效率和质量 。Reuter（2024）介绍了 GPTopic 软件包，利用 LLM 创建动态、互动的主题表征，通过聊天界面让用户能够探索、分析和优化主题，使主题建模更加易于访问、更加系统全面。这些研究均体现了大语言模型在不同领域的应用潜力，以及在提升数据分析和决策支持方面的显著优势。基于此，本研究着重关注在提示工程与群体智能决策双重赋能下的 LLMs 在文本聚类主题提取任务的应用潜力，旨在探索这一策略如何实现高效自动化处理并显著提升文本聚类和主题提取的准确性与鲁棒性，同时减少对大量标注数据的依赖。

2.6 已有研究的启示与本研究的具体问题

构建中文学习 APP 的评估指标体系是一项极具挑战性且意义深远的工作。它涵盖众多多维度与多层次，需要全方位、多角度的综合考量。传统构建方式往往依赖专家的见解与经验，然而在当下技术革新日新月异、用户需求瞬息万变的大环境下，这种模式逐渐显露出局限性。与之相对地，数据驱动的评价模型凭借深度学习算法的强大能力，能够更加精准且灵活地适配当前中文教学的发展态势以及用户不断变化的需求，为构建全新的评价指标体系提供了崭新的视野。近年来，以 ChatGPT 为代表的 LLM 与教育领域的深度融合发展，更是对中文学习 APP 评价体系的构建以及用户体验感的提升产生了深远且重要的影响。

基于上述情况，本研究旨在构建一套智能、客观、动态、综合的中文学习 APP 评价指标体系，以实现教学资源评价的科学化、精准化和自动化。为此，提出以下研究问题：

（1） 如何设计并实现一个覆盖用户多样性与教学场景多变性的中文学习 APP 效能评价框架，以精准监控中文学习者的学习过程并有效评价 APP 的应用效能？
（2） 如何在构建与优化中文学习 APP 评价指标体系中，整合 LLMs 促进群体智能决策，确保评价体系的高效性、准确性和对技术动态的敏捷响应能力？
（3） 如何通过情感分析技术，结合中文学习 APP 的特点，深入挖掘用户对中文学习 APP 的情感倾向和具体反馈，从而为评价指标体系的验证和优化提供更具针对性和实用性的依据，进一步提升中文学习 APP 的用户体验和教育效果？

3. 研究思路

本研究借鉴文本特征向量融合的理念，融合了 BERT 模型的语义特征向量和 LDA 主题特征向量，进而设计了一种适用于中文学习 APP 短评文本的主题识别、评估指标构建及验证的整体框架，该框架如图 1 所示，具体实施步骤如下：

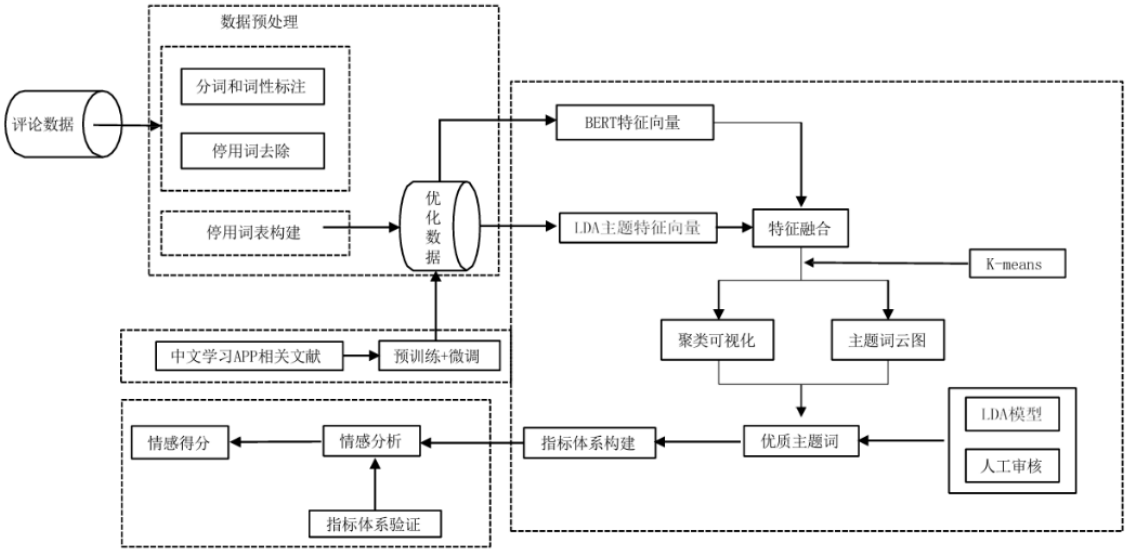


图 1 中文学习 APP 评估指标体系构建和验证的整体框架

3.1 数据采集与预处理阶段

第一，以爬虫软件“后羿”为数据采集工具[4]，从“七麦数据平台”[5]上抓取大量中文学习 APP 的用户短评文本，以此作为下游任务的训练数据集；第二，通过“中国知网”平台[6]整合中文学习相关的学术文献标题与摘要信息，从而构建预训练数据集；第三，整合百度、四川大学以及哈工大的通用停用词表[7]，并据此对原始数据进行深度筛选与结构化处理；第四，为增强模型主题提取方面的性能，本研究将 BERT 模型通过[CLS]标记符产生的综合文本向量与 LDA 模型生成的主题特征向量相结合，借助加权求和、拼接等方式进行特征融合，以构建融合深层语义信息及主题结构的复合特征向量。

3.2 K-means 聚类

首先，将语义和主题相近的关键词分成若干群组，通过此方法探究它们之间的深层联系。然后，通过计算困惑度来挑选 K-means 算法[8]的合适 K 值，以此来决定

[4] https://www.houyicaiji.com/

[5] https://www.qimai.cn/rank/featured

[6] https://www.cnki.net/

[7] https://www.csdn.net/

[8]https://baike.baidu.com/item/K%E5%9D%87%E5%80%BC%E8%81%9A%E7%B1%BB%E7%

恰当的主题数量；接着，为使聚类结果更加直观且易于理解，运用统一流形逼近与投影(Uniform Manifold Approximation and Projection，UMAP)算法[9]，对多维聚类结果进行了降维并进行了可视化处理；最后，在此基础上，构建对应主题的词云图，用以呈现各个主题的核心词汇组成及其相互之间的关系。

3.3 基于 LLMs 的专家模型主题提取方法

首先，通过文本聚类技术提炼出一系列主题，每个主题内都包含相关关键词。其次，引入群体智能体参与分析流程，以文本聚类主题下的关键词为处理对象，鉴定其作为构建中文学习 APP 评价指标体系的适用性。再次，引导智能体进一步深化执行关键词的语境分析任务，力图实现关键词内涵与既定评价理论体系的无缝对接，确保分析的深度与精度。最后，将所有智能体的分析结果集成为统一知识库，并经自一致性(Self-Consistency)投票机制进行过滤与强化，从而高信度地确立核心评价主题群集，为后续评价体系构建奠定坚实基础。

3.4 构建并验证中文学习 APP 评估指标体系

首先基于 LDA 主题模型挖掘用户评论中的核心主题特征，结合 BERT 语义特征与 LDA 主题特征进行多维度融合，构建涵盖功能体验、内容质量、用户情感等维度的评价指标框架。通过 K-means 聚类分析提炼高频主题词，筛选出与学习效果强相关的优质主题词作为核心评价维度。上述过程采用基于 SnowNLP[10]的情感分析技术对中文学习 APP 的短评文本数据进行情感得分量化处理。再将得出的情感得分与用户的实际评分进行对照，以此来检验评价指标体系的准确性与实效性。

4. 研究工具和方法

4.1 BERT 模型

（1）模型介绍：BERT 模型由 Google 公司在 2018 年 10 月推出，与传统的基于静态词嵌入的 Word2Vec 模型不同，BERT 在基于 Transformer 双向编码器架构的基础上将词在不同语境的文本特征纳入考虑（Devlin et al., 2019）。为了使模型能够进行跨任务应用，并能深入语境中捕捉文本语义联系，BERT 在其输入层融合了词向量(Token Embedding)、段落标识向量(Segment Embedding)以及位置向量(Position Embedding)这三种向量嵌入技术，同时融入独特的标记符[CLS]和[SEP]。一方面，在[CLS]的帮助下，模型可为整个序列创建统一的句子向量表征，有助于其执行分类任务。另一方面，[SEP]的作用在于划分和标识文本序列中的各个句子或片段，有助于模型在处理涉及多个句子或段落的情况下，依然能够维持文本顺序和结构信息的稳定性（如图 2 所示）。借助上述结构设计，BERT 模型有效实现了对文本内容

AE%97%E6%B3%95/15779627

[9] https://blog.csdn.net/CRUSH8496052/article/details/132926453

[10] https://developer.baidu.com/article/details/3330267

深入且全面的双向语境理解，增强了其在自然语言处理任务上的表现和精确度。

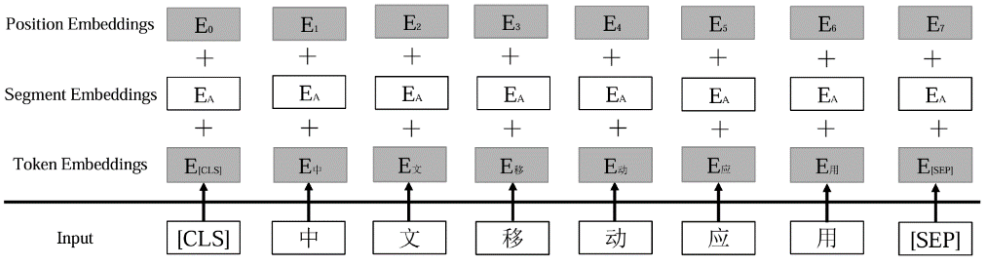


图 2 BERT 的句子级表示

（2）预训练: BERT 模型的训练起初采用大规模的无监督学习方法，这一过程涉及两个主要任务：一是遮蔽语言模型(Masked Language Model, MLM)；二是预测下一句(Next Sentence Prediction, NSP)。MLM 的目标是预测随机遮蔽情况下的词汇，模型必须依赖上下文信息来填充这些缺失，这样它就能学习到更加丰富的语言表达。而 NSP 任务则是评价两个连续的句子片段是否在逻辑上构成前后关系，其目的是辅助模型掌握文本的连贯性和整体结构。

（3）微调: BERT 模型在完成预训练之后，能够针对特定的自然语言处理任务进行微调。在微调过程中，经过有监督学习，模型会在某个特定的数据集上进行训练，这通常意味着在预先训练过的模型之上，仅需增加一个输出层便可满足任务需求，并在此基础上针对这一层进行细致的调整，可实现模型参数的精确优化。BERT 通过迁移学习的方法，在自然语言处理领域，例如文本分类、命名实体识别、情感分析等多个任务中都展现了广泛的应用价值。

4.2 LDA 模型

LDA 是一种无监督学习的概率生成模型，包含文档、主题和词语三层结构，其主要思想是：文档是由若干主题组成的，主题是由文档中一组特定词汇组成的，文档中的每个词都是以一定概率分布的，由此可将一篇文档的主题以出现频率最高的一组词汇表示。LDA 主题模型可以在文档、主题、词语三个层面进行概率建模，计算主题与文档、主题与词语之间的语义关联度，已在文本挖掘、信息检索和情感分析等领域得到了广泛应用（Blei et al., 2003），具体计算过程如图 3 所示。

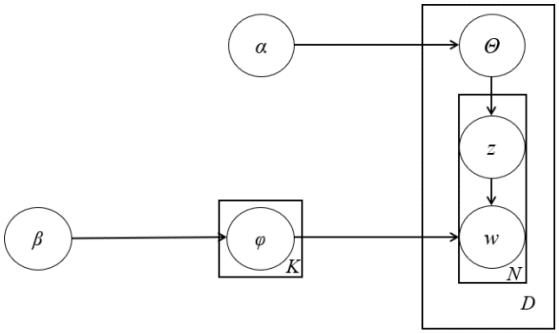


图 3 LDA 主题模型

图 3 中每个符号的含义见表 1，变量间的箭头表示条件依赖关系(Conditional Dependencies)，即文档(Documents)、主题(Topics)以及词语(Words)之间的生成概率关系。

表 1 主题模型中各参数意义

参数	描述
α	狄利克雷分布，θ 的超参数
β	狄利克雷分布，φ 的超参数
θ	评论-主题分布
φ	主题-词分布
z	评论中词语对应的主题
w	评论中的词语
K	主题数
M	文档数目
N	一篇文档的词数

4.3 BERT-LDA 模型

（1）构建 LDA 主题特征向量: 首先对原始文本数据集进行处理，运用 LDA 主题模型对其进行训练。通过无监督学习的方式，挖掘出文本潜在的主题分布。在 LDA 模型中，每个文本都被表示为一系列主题的概率分布，从而可以提取出每个文本对应的主题特征向量。这些向量记录了文本在各个主题上的权重信息。

（2）构建 BERT 语义特征向量: 使用 BERT 模型对预处理数据执行词嵌入操作，以此构建 BERT 语义特征向量。Transformer 编码器单元中，输入向量首先通过多头自注意力机制进行上下文依赖建模，随后经由残差连接与层归一化操作实现梯度稳定，继而通过前馈神经网络进行非线性空间变换并叠加二次残差连接，最终输出具有多层抽象特征的 BERT 语义向量表征（王秀红等, 2021）。

（3）BERT-LDA 特征向量融合: 借助加权求和、拼接以及深度神经网络融合等方法，融合文本特征表示，这种表示兼顾主题结构和深层语义信息，可优化自然语言处理任务的表现，如图 4 所示。

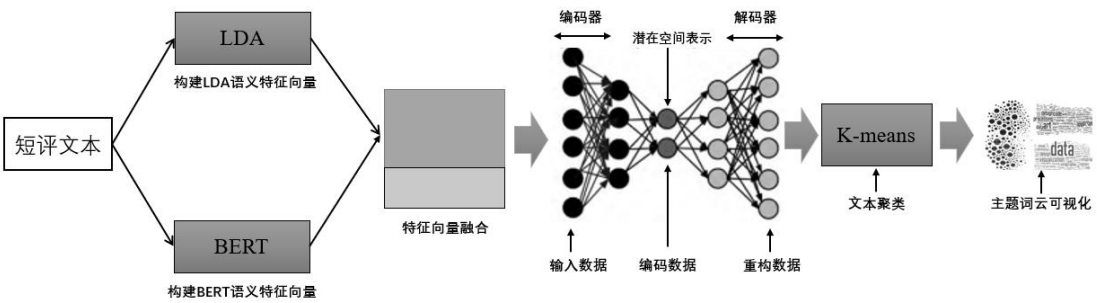


图 4 BERT-LDA 模型示意图

4.4 K-means 聚类及可视化

引入 BERT 语义特征向量对 LDA 主题特征向量进行补偿，虽然提升了文本高层次语义的保持和底层主题模式的捕捉能力，丰富了表达的多样性和深度，但向量拼接操作在信息稀少的高维空间中容易引发维度灾难[11]和过拟合[12]的问题。为此，本研究采用 K-means 算法，通过聚类实现降维，并提取关键词，以降低模型的复杂性且提升分类的效率。K-means 算法是一种无监督学习方法，擅长处理大量数据。该算法以欧氏距离为基准，将相似的数据点划分为同一类别，从而实现数据的聚类（Sinaga et al., 2020）。在进行聚类分析时，可通过评价潜在语义主题模型的困惑度找出最合适的主题数量，该数量将用作 K-means 算法中的 K 值。随着 K 值的逐步上升，模型的困惑度前期呈现降低趋势。然而，在达到一个局部最小值之后，如果继续提高 K 值，模型的表现会开始退化，出现过拟合现象，从而影响其泛化能力。本研究采用 UMAP 算法对特征空间进行非线性可视化处理，目的是为维护数据的全局与局部结构信息，直观展示高概率主题词及其对应概率，进而以此为基础，构建中文学习 APP 的评价指标体系。

4.5 基于情感计算的指标体系验证

基于情感词典的短评文本计算方法是一种利用预先建立的情感词汇库来量化分析文本情感倾向的技术，主要有以下步骤：首先对情感词权重赋值，其次对短评文本中情感词的位置进行定位，最后进行情感强度的计算。基于此，本研究调用 SnowNLP 库中的自然语言处理工具对中文学习 APP 短评文本进行情感打分，计算出每条短评文本的情感得分。接着，对大量短评文本的情感得分进行统计和分析，以获取整体的情感倾向分布。然后，根据分析结果对基于情感计算的指标体系进行验证和调整，确保其准确性和有效性。

5. 实验设计

5.1 数据集构建

本研究选择“中国知网”和“七麦数据”两个平台的中文学习 APP 相关文献以及短评文本作为数据采集对象。首先，在“中国知网”平台以“教育 APP”、“在线汉语/中文”、“汉语/中文技术”、“汉语/中文词典”以及“汉语/中文学习”为主题字段，检索得

[11] 维度灾难（Curse of Dimensionality），又称为维数灾难、维度诅咒，最早由美国数学家理查德・贝尔曼（Richard Bellman）在 20 世纪 50 年代末研究动态规划时提出。随着问题维度的增加，解决问题的难度呈指数级增长，计算量和存储需求等也急剧增加，使得问题变得难以处理。

[12] 过拟合（Overfitting）是指在机器学习和统计建模中，模型在训练数据上表现得过于完美，过度学习了训练数据中的噪声和细节，导致在新的、未见过的数据上表现不佳，泛化能力差的现象。

到 3459 篇，经人工筛选后删除了 356 无效文献，最终得到 3103 篇有效文本摘要数据，并将其作为语料用于增强 BERT 模型的语言理解能力，例如提升模型在识别特定 APP 名称、关键技术与教学模式等方面的准确性，更好地理解用户对“交互性”、“课程涉及”等维度的评价。其次，从“七麦数据”平台搜集了 17 款中文学习 APP 的用户短评，共 10866 条短评数据，将其作为特定领域语料用于下游任务的微调。这些 APP 涵盖了多种学习场景与用户群体，包括综合学习、词典查询、汉字书写、考试备考等类型，具备一定的市场代表性进而功能多样性。本次所选取的 17 款在用户基数、活跃度与功能类型上具有一定代表性，能够反映主流中文二语学习工具的使用体验和反馈特征。所选 APP 多数同时提供中文及英文名称，其开发者背景多样，既包括中国本土企业（如 HELLOCHINESE TECHNOLOGY CO., LTD.），也有 CHINEASY LTD.等国际团队(如表 2 所示)。这些应用主要面向非母语者，提供从零基础到高级水平的中文学习支持，包括词汇、语法、听力、阅读、写作等多个维度。所有 APP 均可通过 IOS APP Store 在全球多个地区下载，覆盖中国、美国、日韩、欧洲以及东南亚等广泛区域，具有较高的可获取性和使用普及度。

表 2 中文学习 APP 评论数据信息

序号	中文学习 APP 名称	开发者	评论数量
1	ChineseSkill	YIYANTECHNOLOGY CO., LTD.	3776
2	HelloChinese	HELLOCHINESE TECHNOLOGY CO., LTD.	2902
3	PlecoChinese Dictionary	PLECO INC.	1425
4	LearnChineseEasily	CHINEASY	1038
5	Scripts:LearnChinesewriting	TOUCHSCREEN LEARNING LTD	642
6	Chineasy:LearnChineseeasily	CHINEASY LTD	345
7	DuChinese–ReadMandarin	SINAMON AB	195
8	ChineseParents	LITTORAL GAMES	111
9	DailyChinese|Words&Idioms	MOJAY, LLC	90
10	MandarinChinesebyNemo	NEMO APPS LLC	83
11	LearnChinese-Mandarin	BRAINSCAPE	79
12	HSKStudyandExam-SuperTest	SHANGHAI YUXUAN INFORMATION TECHNOLOGY CO., LTD	76
13	DominoChinese	ZIMAD	40
14	HanYou-ChineseDictionary	Nomad AI OU	28
15	DotLanguages-LearnChinese	/	17
16	LearnChineseHSK1Chinesimple	KHANJI SCHOOL DIGITAL FACTORY SOCIEDAD LIMITADA	13
17	LearnChineseforBeginners	HECTOR GONZALEZ LINAN	6
	合计	/	10866

5.2 数据预处理

鉴于从“七麦数据”平台所采集的短评文本数据涵盖了英语、俄语、法语等多种语言，为确保后续对这些文本进行一致性处理，首先将所有非中文的短评翻译转写为中文版本。然而，直译过程中往往难以避免情感色彩和初始语义信息的部分损失。为此，本研究利用 LLM 并结合提示技术，针对性地设计了适用于机器翻译任务的提示策略，旨在最大程度上缓解统一翻译过程中可能产生的语义流失问题。其次，针对收集到的中文学习 APP 短评文本，进行系统化的数据预处理步骤：第一，整合了百度、四川大学及哈尔滨工业大学发布的停用词表，通过去除文本摘要中的常见停用词，有效地减少无关噪音信息的影响。第二，通过统计分析文本中的高频词汇，并基于其对主题内容的实质性贡献度，过滤了诸如空格、标点符号等出现频率较高但贡献微弱的词汇单元。

5.3 文本聚类与主题分析

（1）困惑度计算:困惑度是确定主题模型最优主题数目的重要判断指标，困惑度值越小，模型泛化能力越强，当前主题数目就越优（关鹏等, 2016），然而随着主题数增大，提取的主题噪声也会随之增多。因此，本研究将模型最终主题数选定为 6 个。如图 5 所示：

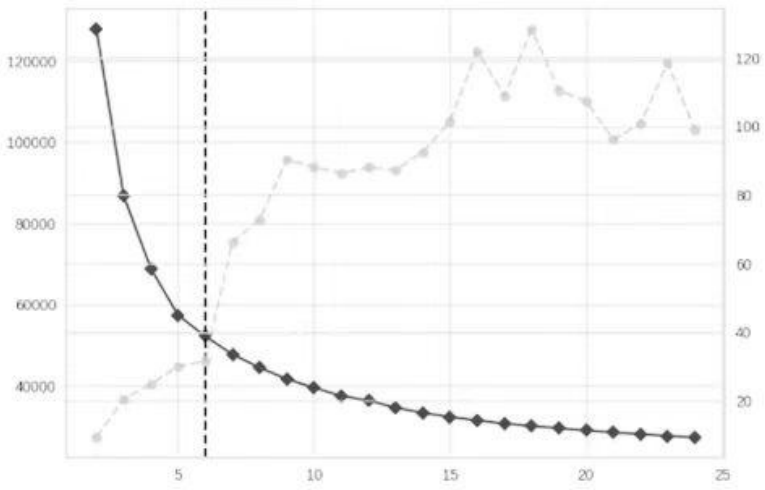


图 5 BERT-LDA 模型不同主题数的困惑度变化

（2）Umap 聚类可视化：通过 UMAP 算法将高维文本向量降维至二维空间并进行可视化。结果显示，不同主题对应的文本在低维空间中形成分布清晰的聚类簇，且各簇间边界较为明确，表明模型能够有效区分语义差异。尽管部分主题簇存在局部重叠，反映了主题间的潜在关联性，但整体聚类结构与预设的 6 个主题数较为吻合，如图 6 所示：

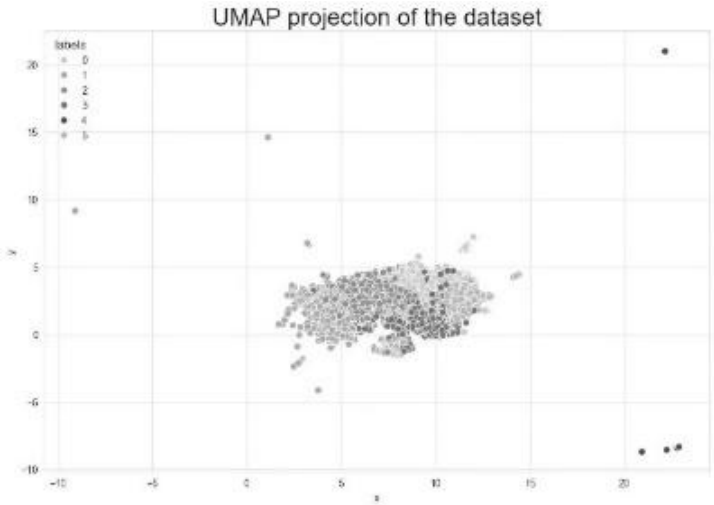
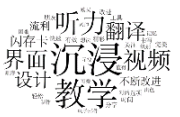


图 6 基于 UMAP 的二维聚类可视化

（3）词云图：基于 BERT-LDA 模型所识别出的 6 个相关主题，选择每个主题下的前 40 个核心词汇进行深入的可视化探索，这一过程旨在通过构建词云图，直观展示这 6 个主题的词汇分布特征(图 7)。



(a)主题 1

(b)主题 2

(c)主题 3

(d)主题 4

(e)主题 5

(f)主题 6

图 7 中文学习 APP 短评词云图

（4）主题关键词：由表 3(下页)可知，BERT-LDA 主题模型提取的主题为 6 个，6 个主题类型分别为“多媒体学习与交互设计”“学习效率与标准测试”“工具创新与社区互动”“语言技能与文化传播”“用户参与与个性化服务”“文化沉浸与深度学习”。

5.4 基于 LLMs 的专家模型主题提取与效果评价

基于 LLMs 的专家模型主题提取与效果评价的核心原理是基于提示去引导群体智能体，并利用自一致性投票机制协同自适应学习策略执行主题提取与识别任务，如图 8(下页)所示。这主要包括主题识别提取算法的定义、主题识别矩阵构建、效果评价等步骤，旨在通过智能化协同提升主题识别的精度与效率，并通过定量与定性分析确保评价的全面性与客观性。

表 3 中文学习 APP 文本聚类主题关键词

序号	主题	关键词
1	多媒体学习与交互设计	沉浸、听力、视频、翻译、闪卡、设计、界面、互动性、流利、教学
2	学习效率与标准测试	记忆 、最佳、笔画、关注、制作、激励、值得、直观、语音、HSK
3	工具创新与社区互动	工具、程序、结构、出色、困难、分享、对话、声调、测验、轻松
4	语言技能与文化传播	语言、故事、教授、开发、声调、精彩、西班牙语、说话、答案
5	用户参与与个性化服务	购买、美好、关注、反馈、电子邮件、发送、故事、抖音、口音、老师
6	文化沉浸与深度学习	深刻、繁体字、教授、改进、笔画、视频、易用性、访问、文化、经典著作

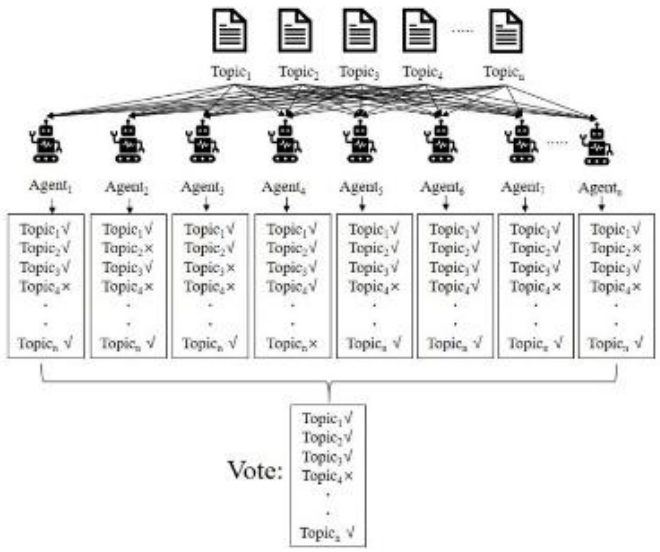


图 8 LLM 自一致性投票机制

（1）基于 LLMs 的主题识别提取算法：本研究定义了适用于 LLMs 处理的主题识别提取算法流程，如表 4 所示。具体来说，包括符号定义、智能处理、筛选准则、综合评价与排序以及集体决策与输出等 7 个流程。算法核心包括智能体执行函数，该部分在于引导智能体根据自身逻辑识别出与主题相关的关键词子集。随后，通过筛选准则去除与评价指标语义不相关的关键词，确保关键词的高关联度。综合评价与排序步骤中，每个智能体对其识别的主题根据相关关键词数量进行排序并优选，排除关键词数量不足的主题。最后，集体决策与输出阶段采用投票机制，比较各主题在不同智能体排序中的流行程度，仅保留那些出现频率超过预设阈值的主题，作为构建中文学习 APP 评价指标的最终主题集合。

表 4 基于 LLM 的主题识别提取提示构建

（1）定义符号	·$T=\{T_1, T_2 \ldots T_i\}$，其中每个 $T_j=\{K_{j1}, K_{j2}, \ldots, K_{jn}\}$，表示序列 T 包含 i 个主题，每个主题有 n 个关键词。$R=\{A_1, A_2, \ldots, A_a\}$，表示序列 R，由 a 个智能体组成。 ·$C=\{I_1, I_2 \ldots I_d\}$，表示一级评价指标的集合。
（2）智能体执行函数	·对于每个智能体 $A_r \in R$ 和每个主题 $T_j \in T$，定义识别函数 $f_r(T_j)=K'_{rj,}$ 其中 $K'_{rj} \subseteq T_j$ 为智能体 A_r 认定的相关关键词集合。
（3）筛选准则	·定义筛选函数 $g(K'_{rj})= K''_{rj}$,使得 $K''_{rj} \in K'_{rj}$，且 $k \in K''_{rj}$，存在 $c \in C$ 使得 k 与 C 在语义上相关。
（4）综合评价与排序	·对于每个智能体 A_r，定义排序和选择函数 $h(A_r)=S_r$,其中 S_r 是按与一级评价指标相关的关键词数量降序排列的主题集合，且$\lvert S_r \rvert < I$,表示除去了一些关键词数量较少的主题。
（5）集体决策与输出	·定义投票函数 $v(R,S)=F$，其中 $F \subseteq T$，基于所有智能体 R 的排序结果 $S=\{S_1, S_2, \ldots, S_a\}$,通过比较各主题在不同智能体排序中的出现频率，选择频率高于预设阈值的主题作为最终结果。 ·输出结果为F，代表经集体决策确定的，用于构建中文学习 APP 评价指标的主题集合。
（6）输入主题	$[T_1, T_2 \ldots T_i]$
（7）输出结果	$[T'_1, T'_2 \ldots T'_i]$

（2）基于 LLMs 的专家模型主题识别矩阵：表 5 展示了不同 LLM 模型对前述定义的各个主题（Topic1 至 Topic6）的支持情况。符号“✓”表示对应主题能够为 LLM 有效识别，并能够作为中文学习 APP 评价指标体系构建的基础，而“×”则表示支持度较低或不具备直接关联的主题。通过汇总各模型对聚类主题的支持情况，并在最后一行统计出每个主题的投票支持率，可直观反映各聚类主题识别提取情况。

表 5 基于 LLM 的专家模型主题识别矩阵

	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
Agent1	✓	✓	✓	✓	×	×
Agent2	✓	✓	✓	✓	×	✓
Agent3	✓	×	✓	×	×	✓
Agent4	✓	×	✓	✓	✓	×
Agent5	×	×	✓	×	✓	✓
Agen6	✓	✓	×	✓	×	✓
Agent7	✓	×	✓	×	×	✓
Agent8	✓	✓	✓	✓	×	✓
Agent9	×	✓	✓	×	✓	✓
Agent10	✓	×	✓	×	×	✓
投票支持率	80%	50%	80%	50%	30%	80%

（3）评估指标：为全面评价基于 BERT-LDA 模型的主题聚类效果，本研究采用了以下 3 个评价指标：查准率(*Precision*)、查全率(*Recall*)与 *F* 值(*F-measure*)，以下分别用 *P*，*R* 和 *F* 表示。这些指标帮助本研究从不同维度理解模型识别主题的准确性和全面性。其中 $T_{correct}$ 是指 LLM 专家模型识别的正确主题数量，$T_{extract}$ 是指基于 LLM 专家模型提取或识别出的主题数量，$T_{standard}$ 是指专家总结出的主题数量。

$$（1）\ P = \frac{T_{correct}}{T_{extract}}$$

$$（2）\ R = \frac{T_{correct}}{T_{standard}}$$

$$（3）\ F = \frac{2PR}{P+R}$$

（4）评价结果：本研究评价对比了 BERT-LDA 模型与传统 LDA 模型在主题抽取任务上的性能。从结果可以看出前者在查准率、查全率以及 *F* 值上均优于后者，分别高了 16.07%、33.3%以及 27.96%。BERT-LDA 模型识别出的主题类别更加清晰、准确，其中主题 1、主题 3 和主题 6 更有利于构建中文学习 APP 评价指标体系。

表 6 不同模型主题抽取效果对比结果

	主题数	$T_{extract}$	$T_{correct}$	$T_{standard}$	查准率	查全率	F 值
BERT-LDA	6	6	3	6	50%	50%	50%
LDA	3	3	1	6	33.3%	16.7%	22.04%

6. 中文学习 APP 评估指标体系构建及验证

6.1 中文学习 APP 评估指标体系构建

基于 BERT-LDA 模型和 K-means 聚类，同时运用基于 LLMs 的主题识别提取方法对“多媒体学习与交互设计”“学习效率与标准测试”“工具创新与社区互动”、“语言技能与文化传播”“用户参与与个性化服务”“文化沉浸与深度学习”等 6 个主题及其主题关键词进行了识别，并基于主题 1、主题 3 和主题 6，最终归纳总结得到中文学习 APP 评价指标体系。需要特别指出的是，本研究中归纳总结出的主题及最终构建的评价指标体现（如表 7 所示），均由机器基于数据挖掘和算法模型自动生成，整个过程未引入人类专家进行审核或验证。这种纯粹数据驱动的方法虽然保障了规模与效率，但也可能引入算法固有的偏见，例如不能捕捉到凭借专家经验才能洞察的深层教学逻辑与核心质量维度。

表 7 中文学习 APP 评价指标体系

一级维度	二级维度	三级维度	指标描述
内容评价（内容质量）	内容组织	语言表达	清晰准确，符合语法规范
		词汇覆盖	广泛多样，适应不同水平
	视听融合	音质协调	声音清晰，无杂音干扰
		视频指导	情景模拟，直观展示应用
过程评价（用户体验）	信息架构	页面布局	简洁高效，便于信息查找
		导航设计	逻辑清晰，快速定位功能
	交互体验	流畅体验	响应迅速，操作无卡顿
		平台兼容	多系统适配，运行稳定
	更新迭代	问题修复	及时反馈，解决用户难题
		版本优化	持续改进，提升应用性能
效果评价（学习成效）	实用工具	笔顺演示	正确书写顺序，动画展示
		字典查询	例句丰富，快速查词解义
	技能应用	听力强化	多场景练习，增强交流能力
		文化适应	跨文化融入，提升理解能力

6.2 中文学习 APP 短评情感分析

（1）中文学习 APP 短评文本评分分布：调用 SnowNLP 库对中文学习 APP 的短评进行情感计算，旨在对 17 款中文学习 APP 的情感倾向进行深入分析，即积极、消极或中性等评论数量的占比。基于此，本研究对 10866 条评论进行情感分析，并量化了每款 APP 的情感得分，进而得出中文学习 APP 总体短评情感分布情况，见表 8。

表 8 中文学习 APP 短评情感分布情况

情感类型	评论量	占比
积极情感	10132 条	93.25%
中性情感	311 条	2.87%
消极情感	422 条	3.88%

（2）基于情感词典的中文学习 APP 短评文本计算：本研究选择了适合中文语境的情感词典，结合了否定词识别、程度副词权重调整等方法，以提高情感分析的准确性。并以 *HanYou-Chinese Dictionary* APP 的部分短评情感得分为示例结果，如表 9 所示。

表 9 “HanYou - Chinese Dictionary”APP 部分短评情感得分

序号	评论内容	情感得分
1	我特别喜欢绘图词典。	92.11%
2	汉友确实帮助我增加了中文词汇量。闪存卡也很有用。	82.48%
3	超级有用且易于使用。强烈推荐给汉语学习者和来中国的游客！	95.71%
4	我建议该APP可以包含一些同义词及有关定义的更多详细信息。	24.69%
5	即使它是免费的，它仍然是垃圾。	10.48%

基于评论量的梯度分布性与情感均值的层级覆盖性双重筛选原则，选取用户评分排名前三的 3 款中文学习 APP 作为研究对象进行深入分析，如表 10 所示。通过该分析，可以深入了解用户反馈的情感倾向，为 APP 开发者提供改进建议，并帮助潜在用户做出更明智的选择。

表 10 部分中文学习 APP 短评情感分布情况

中文学习 APP 名称	APP 分数	评论量	情感均值
Du Chinese – Read Mandarin	4.7	195 条	87.09%
Domino Chinese	4.6	40 条	83.63%
Learn Chinese HSK1 Chinesimple	4.7	13 条	92.22%

Learn Chinese HSK1 Chinesimple 和 *Du Chinese–Read Mandarin* 两款 APP 不仅获得了较高的平均评分，而且用户情感均值也相对较高，表明这两款 APP 在用户满意度方面表现突出；相比之下，*Domino Chinese* APP 在内容深度及用户体验优化方面需进一步改进。（3）基于中文学习 APP 的多维评价：*Chinese Parents* APP 在所选的中文学习 APP 中评分最低，其在内容、用户体验以及学习成效等方面可能存在较多有待改进的地方，因此选择 *Chinese Parents* APP 作为验证中文学习 APP 评价框架的主要对象，期望通过对其进行详细评价，为提升此类 APP 的整体质量和用户体验提供有价值的案例借鉴。具体分析结果如表 11(下页)所示，展示了该 APP 在各个评价维度上的情感均值及用户反馈的详细情况。

7. 讨论与分析

7.1 研究价值

（1）理论贡献

第一，推动中文教育数字化与智能化转型。通过整合 BERT-LDA 模型与 K-means 聚类算法，并结合 LLMs 的主题识别与提取方法，本研究在“内容质量—用户体验—学习成效”三维框架下构建了系统的评价指标体系（见表 7）。这一过程不仅

验证了人工智能驱动评价体系构建的可行性，也为教育理论与技术融合提供了新路径。

表 11 “Chinese parents”APP 短评文本多维评价结果

一级维度	二级维度	情感均值	三级维度	情感均值
内容评价（内容质量）	内容组织	83.21%	语言表达	86.25%
			词汇覆盖	80.17%
	视听融合	47.71%	音质协调	32.74%
			视频指导	62.67%
过程评价（用户体验）	信息架构	50.03%	页面布局	52.25%
			导航设计	44.81%
	交互体验	49.31%	流畅体验	42.18%
			平台兼容	56.44%
	更新迭代	70.35%	问题修复	65.47%
			版本优化	75.22%
效果评价（学习成效）	实用工具	86.65%	笔顺演示	84.74%
			字典查询	88.51%
	技能应用	77.84%	听力强化	90.27%
			文化适应	65.41%

第二，扩展评价方法论的适用性。在对 10866 条用户短评的情感计算中，本研究验证了基于 SnowNLP 与情感词典结合的方法能够兼效率与精准性。例如，*HanYou-Chinese Dictionary* APP 在部分评论中的情感得分差异显著，显示了细颗粒度分析在揭示用户真实情感上的独特价值。这为教育技术评价提供了新的方法论支撑。

第三，推动个性化与动态化评价的形成。在 *Chinese Parents* APP 的多维评价结果中。用户对“听力强化（90.27%）”表现高度认可，但对“音质协调（32.74%）”则明显不满。这种维度差异凸显了动态指标不仅要面向整体水平，更要识别局部薄弱环节，为后续模型的个性化适配提供了理论依据。

（2）实践意义

第一，服务教师教学与资源甄选。对于外语教师而言，本研究的成果可直接应用于课堂资源筛选与教学辅助。例如，教师可参考 APP 在“词汇覆盖”或“文化适应”维度的情感均值，判断该应用是否适合于初级、中级或跨文化教学场景，从而提升教学资源配置的科学性。

第二，提升教育决策的数据驱动水平。基于情感分布结果（积极评估占比 93.25%），教师与教育管理者能够快速了解不同 APP 的整体口碑，并通过多维度细分（如“信息架构”“交互体验”）洞察学生学习中的真实痛点。这使得教学管理更具针对性和实效性。

第三，促进教育市场竞争与应用生态优化。评价框架不仅帮助开发者精准把握用户需求，还能为学生与教师提供直观的参考。例如，*Du Chinese–Read Mandarin* 和 *Learn Chinese HSK1 Chinesimple* 两款 APP 的情感均值分别达到 87.09%和 92.22%，对教师而言，这类数据有助于有限推荐更受认可的资源，提升课堂学习效果。

7.2 发展建议

优化和完善基于 BERT-LDA 的中文学习 APP 评价指标体系的构建，不仅能提升其实用性和科学性，还能为中文教学资源的优化发展提供有力的数据支持和决策依据。由此，本文从智能化、动态化、安全性三个层面对其评估指标体系构建提出建议：

（1）智能化导向。结合 LLM 与大数据分析，教师可依赖评价系统快速识别适配不同学习水平的资源。例如，在 APP 教学应用中，系统可自动生成“听力练习难度梯度”或“词汇拓展路径”，为教师布置差异化作业提供支持。同时，未来模型还应在跨语种支持与文化内容融合等方面加大投入，提升全球推广的适配度。

（2）动态化适应。面对互联网教育资源更新迅速的现实，指标体系必须保持灵活可扩展。例如，*Chinese Parents* APP 在“页面布局”和“导航设计”上评分偏低，若能及时反馈并优化，则可快速提升用户体验。教师在选择 APP 时，也能依靠这种动态评价，避免教学过程中因工具落后而导致的学习障碍。

（3）安全性保障。在教育应用的推广中，教师和学生的数据安全问题尤为关键。未来的评价框架需要纳入“隐私保护”维度，并遵循国际数据保护法规。这不仅关系到用户信任度，也直接影响教育资源的可持续发展和跨国推广。

（4）人机协同与专家介入。本研究所构建的指标体系完全基于机器算法，虽展现了自动化处理的潜力，但缺乏教育领域专家的深度干预。未来研究应积极探索“人机协同”(Human-AI Collaboration)的混合模式，将本研究的数据驱动方法与德尔菲法相结合。例如。可以在机器初步生成主题和指标以后，引入国际中文教育领域的专家和一线教师进行多轮审议、修正与验证，对机器可能存在的偏差进行校准，对指标的权重和表述进行优化。这种融合了算法广度与专家深度的模式，有望构建出既客观全面又符合教育理论与实践的评价体系。

8. 结语

本研究通过整合 BERT-LDA 模型与 LLMs，不仅科学构建了一套科学的中文学习 APP 评价指标体系，更重要的是，通过对真实用户数据的深入挖掘，验证了该体系的有效性和实用性。研究表明将 BERT-LDA 模型与基于 LLMs 引导的群体智能决策相结合的评价方法，能有效构建客观、动态的中文学习 APP 评价指标体系，

并通过情感分析验证了该体系在精准反映用户体验与学习成效方面的实用性与科学性。未来研究应进一步探索评价体系的自动化与自适应机制，融合更多维度的用户数据，并加强对数据安全与伦理问题的关注，以推动更加智能、全面的国际中文教育数字资源评价生态。

致谢：本文受教育部人文社会科学重点研究基地重大项目“国际中文教育数字资源综合评价理论与方法研究”（22JJD740016）的资助。

参考文献

Achiam, J., Adler, S., Agarwal, S., et al. (2023). Gpt-4 technical report. *arXiv Preprint arXiv*:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Alon, I., Haidar, H., Haidar, A., & Guimon, J. (2025). The future of artificial intelligence: Insights from recent Delphi studies. *Futures*, 165, 103514. https://doi.org/10.1016/j.futures.2024.103514

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993-1022. https://dl.acm.org/doi/abs/10.5555/944919.944937

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

Cai, Y., & Wang, Z. (2022). Research on the Learning Intention of Chinese Learners in Live Courses Based on the Technology Acceptance Model. *Language Teaching and Research*, (5), 35-46. [蔡燕, 汪泽. (2022). 基于技术接受模型的中文学习者直播课程学习意愿研究. *语言教学与研究*, 5, 35-46.]

Cheng, T., & Wang, Z.Q. (2024). Research on the construction of a cross-cultural professional competence model for "Chinese + Vocational Skills". *China Vocational and Technical Education*, (1), 59-70. [程涛, 王正青. (2024). “中文+职业技能”人才跨文化职业胜任力模型构建研究. *中国职业技术教育*, 1, 59-70.]

Deng, X., Li, J.M., Zeng, H.J., et al. (2012). Analysis and application research on weight calculation methods in the analytic hierarchy process. *Mathematical Practice and Knowledge,* (7), 93-100. [邓雪, 李家铭, 曾浩健, 等. (2012). *层次分析法权重计算方法分析及其应用研究. 数学的实践与认识,* 7, 93-100.]

Devlin, J., Chang, M. W., Lee, K., et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint* arXiv:1810.04805. https://aclanthology.org/N19-1423/

Fang, Z.F., & Xu, J. (2023). Construction of an indicator system for digital literacy of international Chinese teachers. *Journal of Tianjin Normal University (Social Sciences Edition)*, (6), 25-33. [方紫帆, 徐娟. (2023). 国际中文教师数字素养指标体系建构研究. *天津师范大学学报(社会科学版)*, 6, 25-33.]

Gao, C.Z., & Ge, Z.Y. (2025). A study of the supply and demand of the interactive digital products of international Chinese-language education and its supply strategies. *Journal of Yunnan Normal University (Humanities and Social Sciences*

Edition), (1), 53-60. [高传智, 戈兆一. (2025). 数字交互式国际中文教育产品供需状况与供给策略研究. *云南师范大学学报(哲学社会科学版)*, (1),53-60.]

Gong, X., & Liang, Y. (2023). Construction of a basic framework for evaluation indicators of international Chinese textbooks based on descriptive corpus. *Journal of Research on Education for Ethnic Minorities*, (3), 161-167. [宫雪, 梁宇. (2023). 基于描述语库的国际中文教材评价指标基础框架构建. *民族教育研究*, 3, 161-167.]

Guan, P., & Wang, Y.F. (2016). Study on the determination of optimal number of topics in LDA topic model in scientific and technological intelligence analysis. *Modern Library and Information Technology*, (9), 42-50. [关鹏, 王曰芬. (2016). 科技情报分析中 LDA 主题模型最优主题数确定方法研究. *现代图书情报技术*, 9, 42-50.]

Guo, J., Wu, Y.H., Gu, L., et al. (2021). Current status and prospects of digital resource construction in international Chinese education. *International Chinese Language Teaching Research*, (4), 86-96. [郭晶, 吴应辉, 谷陵, 等. (2021). 国际中文教育数字资源建设现状与展望. *国际汉语教学研究*, 4, 86-96.]

Jang, M. E., & Lukasiewicz, T. (2023). Consistency analysis of ChatGPT. *arXiv Preprint* arXiv:2303.06273. https://doi.org/10.48550/arXiv.2303.06273

Jia, H. N., & Chen, L. H. (2023). Topic mining and indicator construction of clothing brand information in online social networks. *Wool Textile Science & Technology, 51*(1), 121-129. [贾海楠, 陈李红. (2023). 在线社交网络中服装品牌信息主题挖掘及其指标构建. *毛纺科技, 51*(1), 121-129.]

Kojima, T., Gu, S. S., Reid, M., et al. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199-22213. https://doi.org/10.48550/arXiv.2205.11916

Kharat, M. G., Kamble, S. J., Raut, R. D., & Kamble, S. S. (2016). Identification and evaluation of landfill site selection criteria using a hybrid Fuzzy Delphi, Fuzzy AHP and DEMATEL based approach. *Modeling Earth Systems and Environment*, *2*(2), 98. https://link.springer.com/article/10.1007/s40808-016-0171-1

Lai, W. (2023). Deep learning network-based evaluation method of online teaching quality of international Chinese education. *3c Tecnología: glosas de innovación aplicadas a la pyme, 12*(1), 87-106. https://dialnet.unirioja.es/servlet/articulo?codigo=8881472

Li, S. S., Liu, F., Cao, H. Y. (2025). An analysis of Chinese app teaching design from the perspective of mobile microlearning theory: A case study of *HelloChinese*. *International Chinese Language Education*, (1), 112-127+142. [李姝姝, 刘芳, 曹洪豫. (2025). 移动微学习理论视角下的中文 App 教学设计分析——以 HelloChinese 为例. *国际中文教育(中英文)*, 1, 112-127+142.]

Li, T. Y., & Liu, Q. M. (2024). Construction of an evaluation indicator system for painting works based on BERT-LDA and K-Means clustering. *Software Engineering,* (1), 68-73. [李天义, 刘勤明. (2024). 基于 BERT-LDA 和 K-means 聚类的绘画作品价值评估指标体系构建. *软件工程,* 1, 68-73.]

Liang, Y., & Li, N. E. (2023). Construction of an evaluation indicator system for

international Chinese textbooks—Based on the Delphi Method and the Analytic Hierarchy Process. *Journal of Guizhou Normal University (Social Sciences Edition)*, (6), 30-40. [梁宇, 李诺恩. (2023). 国际中文教材评价指标体系构建——基于德尔菲法和层次分析法. *贵州师范大学学报(社会科学版)*, 6, 30-40.]

Liang, Y., & Li, N.E. (2023). Research on the intention to use Chinese digital learning resources and its influencing factors: Based on the extended TAM Model. *Language and Text Application*, (2), 23-35. [梁宇, 李诺恩. (2023). 中文数字学习资源使用意愿及其影响因素研究——基于 TAM 扩展模型. *语言文字应用*, 2, 23-35.]

Liu, Y. J. (2021). Research on the optimization path of lexicographical integrated publishing: A review of the Modern Chinese Dictionary (7th Edition) APP. *Journal of Beijing Union University (Humanities and Social Sciences Edition),* (2), 109-115. [刘永俊. (2021). 辞书融合出版的优化路径研究——兼评《现代汉语词典》(第 7 版)APP. *北京联合大学学报(人文社会科学版),* 2, 109-115.]

Pan, X. Y., Ni, Y., Jin, C. H., Zhang, J. (2023). Extraction of value elements and construction of an indicator system for calligraphy works based on Hyperplane-BERT-Louvain optimized LDA model. *Data Analysis and Knowledge Discovery,* (10), 109-118. [潘小宇, 倪渊, 金春华, 张健. (2023). 基于超平面-BERT-Louvain 优化 LDA 模型的书法作品价值要素提取及指标体系构建. *数据分析与知识发现,* 10, 109-118.]

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, 8, 80716-80727. https://ieeexplore.ieee.org/abstract/document/9072123

Wang, C. Z., & Si, Q. (2011). Data statistical processing methods and their application research in the Delphi Method. *Journal of Inner Mongolia University of Finance and Economics (Comprehensive Edition),* (4), 92-96. [王春枝, 斯琴. (2011). 德尔菲法中的数据统计处理方法及其应用研究. *内蒙古财经学院学报(综合版),* 4, 92-96.]

Wang, X. H., & Gao, M. (2021). Key technology identification method based on BERT-LDA and empirical research: A case study of agricultural robots. *Library and Information Service,* (22), 114-125. [王秀红, 高敏. (2021). 基于 BERT-LDA 的关键技术识别方法及其实证研究——以农业机器人为例. *图书情报工作,* 22, 114-125.]

Wu, Q, Bansal, G, Zhang, J, Wu, Y, Li, B, Zhu, E, Jiang, L, Zhang, X, Zhang, S, Liu, J, Awadallah, A, White, R, Burger, D, Wang, C. (2023). Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. *arXiv Preprint* arXiv:2308.08155. https://openreview.net/forum?id=BAakY1hNKS

Wu, T; He, S; Liu, J; Sun, S; Liu, K; Han, Q. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica, 10*(5), 1122-1136. DOI: 10.1109/JAS.2023.123618

Yang, Q. (2018). Implications for the construction of Chinese language network resources from Chinese learning APPs: A case study of iChinese APP. *Publishing Horizon*, (14), 77-79. [杨倩. (2018). 中文学习 APP 对汉语网络资

源建设的启示——以 iChinese APP 为例. *出版广角*, 14, 77-79.]

Yang, T., Xu, T., & Li, Q. (2023). Construction, empirical study, and optimization of an indicator system for intelligent competence of international Chinese teachers. *Journal of Yunnan Normal University (Teaching & Studying Chinese as a Foreign Language Edition),* (3), 41-52. [杨甜, 许桐, 李琴. (2023). 国际中文教师智能素养指标体系的构建、实证及优化. *云南师范大学学报(对外汉语教学与研究版),* 3, 41-52.]

Yang, X. H (2019). Design thinking training: A way out of the dilemma of thinking teaching. *China Educational Technology,* (7)*,* 54-59[杨绪辉. (2019) 设计思维培养: 基础教育思维教学困境的出路. *中国电化教育*, 7, 54-59.]

Yang, X. M., Luo, J. J., Liu, Y. X., Chen, S. C. (2017). Data-driven teaching: New directions of teaching paradigms in the era of big data. *Research in Electro-Education,* (12), 13-20+26. [杨现民, 骆娇娇, 刘雅馨, 陈世超. (2017). 数据驱动教学:大数据时代教学范式的新走向. *电化教育研究,* 12, 13-20+26.]

Yu, X. F., & Fu, D. (2004). Overview of multi-indicator comprehensive evaluation methods. *Statistics and Decision Making*, (11), 119-121. [虞晓芬, 傅玳. (2004). 多指标综合评价方法综述. *统计与决策*, 11, 119-121.]

Yuan, H. H., & Gao, X. L. (2014). Assessing the economic vulnerability to disasters of cities: A case study of Haidian District in Beijing. *Journal of Natural Resources*, (7)*,* 1159-1172. [袁海红, 高晓路. (2014), 城市经济脆弱性评价研究——以北京海淀区为例 [J]. *自然资源学报*, 7, 1159-1172.]

Zhang, Y., Zhu, Q., & Li, M. (2019). Construction of evaluation index system for domestic learning APPs from the perspective of user experience: Based on D-S Evidence Theory. *Journal of Intelligence,* (2), 187-194. [张熠, 朱琪, 李孟. (2019). 用户体验视角下国内学习学习 APP 评价指标体系构建——基于 D-S 证据理论. *情报杂志,* 2, 187-194.]

Zhao, X. M., Shu, J., & Zhang, Z. X. (2017). Research on the evaluation of learning APPs based on user experience. *Heilongjiang Science and Technology Information*, (2), 186-189. [赵学铭, 舒珺, 张振兴. (2017). 基于用户体验的学习学习 APP 评价研究. *黑龙江科技信息*, 2, 186-189.]

附录 中文 APP 情况简介

APP	发布日期	应用特点
ChineseSkill	2014-02-08	内置中文语音评估、汉字手写、动画技术。
HelloChinese	2015-06-18	致力于为初级中文学习者提供优质语言服务。
PlecoChineseDictionary	2009-12-17	集成词典/文档阅读器/单词卡系统，支持全屏手写输入及实时 OCR 功能。
LearnChineseEasily	2018-02-14	以“积木式”方法组织汉字学习的形式。

Scripts:LearnChinesewriting	2018-10-10	配置极简的语言插图及快节奏的语言游戏。
Chineasy:LearnChineseeasily	2018-02-12	内置中文词汇学习游戏，且具备 28 个贴合实际生活场景的汉语学习主题及 1834 个词汇。
DuChinese–ReadMandarin	2015-12-05	Du Chinese 是一款分级阅读应用程序，为各级汉语学习者提供广泛的阅读练习。
ChineseParents	2022-04-22	以真实生活为背景，沉浸式体验中文学习
DailyChineseWords&Idioms	2019-06-20	遵循艾宾浩斯记忆曲线规律，间隔复习并逐步引入新单词，确保已学单词的有效记忆。
MandarinChinesebyNemo	2011-04-12	个性化追踪进度，重点学习高频词汇，逐步构建长期记忆，轻松应对日常对话。
LearnChinese-Mandarin	2011-05-19	应用内含超过 200 条常用词汇及短语，并由母语者录音，支持离线使用。
HSKStudyandExam-SuperTest	2018-02-14	应用结合 AI 技术提供精准水平测试与定制化课程，拥有丰富题库及模拟考试资源。
DominoChinese	2022-02-16	通过视频教程和真实情境，清晰解释并演示日常普通话使用形式
HanYou-ChineseDictionary	2014-09-19	具备强大的离线 OCR 功能，能识别万余个汉字，辅助阅读各种文本。
DotLanguages-LearnChinese	2021-04-12	通过丰富多样的 HSK 级别文章提升普通话水平。每日新增六篇以上文章确保学习材料充足。
LearnChineseHSK1Chinesimple	2020-03-11	告别枯燥课本与昂贵课程，透过宾果系统分析进步状况，集中练习要点，快速高效备考 HSK。
Learn Chinese for Beginners	2022-01-11	无需注册账号，所有内容完全免费且可离线使用。课程覆盖拼音、汉字及日常生活中的各种实用词汇与表达。

