基于预训练语言模型的语言学习APP评价研究

郑明鉴,徐 娟

(北京语言大学 信息科学学院,北京 100083)

摘 要: 在教育数字化转型的背景下,移动学习成为数字学习的新常态。语言学习移动应用程序(APP)已经成为外语学习者学习语言的重要工具。秉持“数据循证”的评价观,提出预训练语言模型和大数据挖掘相结合的教育APP评价技术方案,采集主流应用市场上下载量最高的20款语言学习APP的用户评论数据,并利用预训练语言模型计算出评论文本的情感分值。在文本信息的基础上,通过主题建模等技术分析学习者在使用APP辅助语言学习时的需求和偏好,并尝试从中提取出针对APP评价的多项指标。最后,综合各项分析结果建立一套准确、客观的语言学习APP评价体系,并对已采集评论信息的APP进行实例可视化分析,旨在发挥预训练语言模型和数据要素的价值,助力数字化语言教学资源的科学治理。

关键词: APP;语言学习;数据挖掘;预训练语言模型;评价可视化;情感分析

DOI: 10.11907/rjdk.241652 **开放科学(资源服务)标识码(OSID):**

中图分类号: G434 **文献标识码:** A **文章编号:** 1672-7800(2025)001-0169-07

Research on Evaluation of Language Learning APP Based on Pre-trained Language Model

ZHENG Mingjian, XU Juan

(*College of Information Science, Beijing Language and Culture University, Beijing* 100083, *China*)

Abstact: In the context of the digital transformation of education, mobile learning has become the new norm in digital learning. Language learning mobile applications (APP) have emerged as crucial tools for second language learners. This paper adheres to an "evidence-based data" evaluation perspective and attempts to propose an educational APP evaluation technique combining pre-trained language model and big data mining. The study collects user review data from the top 20 language learning apps on mainstream APP markets, and employs pre-trained language model to calculate sentiment scores for the review texts. Building upon text information, the paper utilizes techniques such as topic modeling to analyze learners′ needs and preferences when using language learning apps, aiming to extract multiple indicators for evaluating the apps. Finally, the study integrates various analytical results to establish an accurate and objective language learning APP evaluation system. Visualized analysis is conducted on collected review information for selected apps, leveraging the value of pre-trained language model and data elements to contribute to the scientific governance of digital language teaching resources.

Key Words: APP; language learning; data mining; pre-trained language model; evaluation visualization; sentiment analysis

0 引言

在教育数字化转型的背景下,移动学习以其便携性、泛在化的优势,逐渐成为数字化语言学习的新形态[1]。APP(Application)作为移动语言学习的重要工具,截至2020年,约有7 000万中国人在使用APP进行外语学习[2]。近年来,语言学习APP的功能日益强大,基本涵盖了词汇识记、考试准备、口语练习、作文批改等语言学习的方方面面[3-5]。然而,当语言学习者在面对各式各样的APP时,如何选择符合需求的APP成为关键。为解决这一问题,急需构建语言学习APP的评价体系,以支持学习者快速找到适

收稿日期: 2024-08-01

扫描二维码阅读全文:

基金项目: 中央高校基本科研业务费专项资金项目(24YCX193);北京语言大学重大专项课题(23ZDY02);世界汉语教学学会2024年全球中文教育主题学术活动计划项目(SH24Y24)

作者简介: 郑明鉴(2000-),男,CCF学生会员,北京语言大学信息科学学院硕士研究生,研究方向为人工智能、语言教育技术;徐娟(1970-),女,北京语言大学信息科学学院研究员、博士生导师,研究方向为数字化汉语教学。本文通讯作者:徐娟。

合自己的数字学习资源。

传统的资源评价体系构建多立足于专家视角，尚缺少从用户视角出发，运用人工智能和大数据相结合的方法构建评价指标体系。因此，本文尝试引入预训练语言模型，并通过情感分析、聚类分析等方法，自下而上地构建语言学习 APP 评价指标体系。

1 相关工作

语言学习 APP 兼有学习工具与软件产品的属性，前者关注 APP 的教育价值，后者则聚焦 APP 的商品特点。因此，对语言学习 APP 的评价可以从教学与产品两个视角切入。现有研究大多集中在教学视角，注重评估 APP 的助学效果，评价方法包括语言测试、调查问卷等[6]。但由于学习者只能在给定的测试或问卷范围内作出反馈，使得此类方法获取的信息维度也较为单一。此外，在产品视角下，用户使用语言学习 APP 时倾向于将其作为软件产品在应用商店中进行评价[7]。应用商店中积累的大量评论文本包含了学习者多维度的使用体验信息，可以支持其他学习者对照自身学习需求，个性化地选择适合的语言学习 APP。但这些评论文本数据大多是非结构化的，学习者难以有效处理此类数据。因此，对评论文本的数据挖掘和可视化处理有助于更好地利用过程性学习数据，从而支持个性化语言学习，也为语言学习 APP 评价提供更全面的视角。

大数据作为“互联网+”时代教育发展的新引擎，是智慧教育与个性化学习研究与实践的重要基础，也是发展智慧教育的基石[8,9]。近年来，BERT[10]、GPT[11]等预训练语言模型（Large Language Model，LLM）迅速发展，为教育大数据挖掘中的建模与分析提供了综合效果更佳的技术方案。通过预训练语言模型量化分析评论文本数据中的模式与规律，可以实现评论大数据与教学评价方法的有机结合，使教育 APP 评价更具多维性与科学性。本文以语言学习 APP 为例，尝试使用预训练语言模型对 APP 的评论文本进行数据挖掘与主题建模，自下而上地建立一套语言学习 APP 评价指标体系，并实现评价结果的可视化，从而纾解移动语言学习供需矛盾的现实困境。

2 实验设计与数据集构建

2.1 实验设计

本文以建立语言学习 APP 评价体系为最终目标，按以下步骤开展实验：首先，通过数据采集、清洗、预处理获取有效的评论数据，并依托数据特性对数据集进行标注；其次，基于标注完成的数据集微调预训练语言模型，用模型计算每条评论的情感得分，同时把所有评论分为正负面两类；再次，对两类评论分别进行主题建模与关键词提取，尝试从多角度出发，将评论中学习者关注的信息聚类为评价体系中衡量 APP 优劣的各项指标；最后，整合评价指标构建评价体系，并结合之前获取的评论得分，对示例 APP 进行可视化评价分析。研究技术路线见图 1。

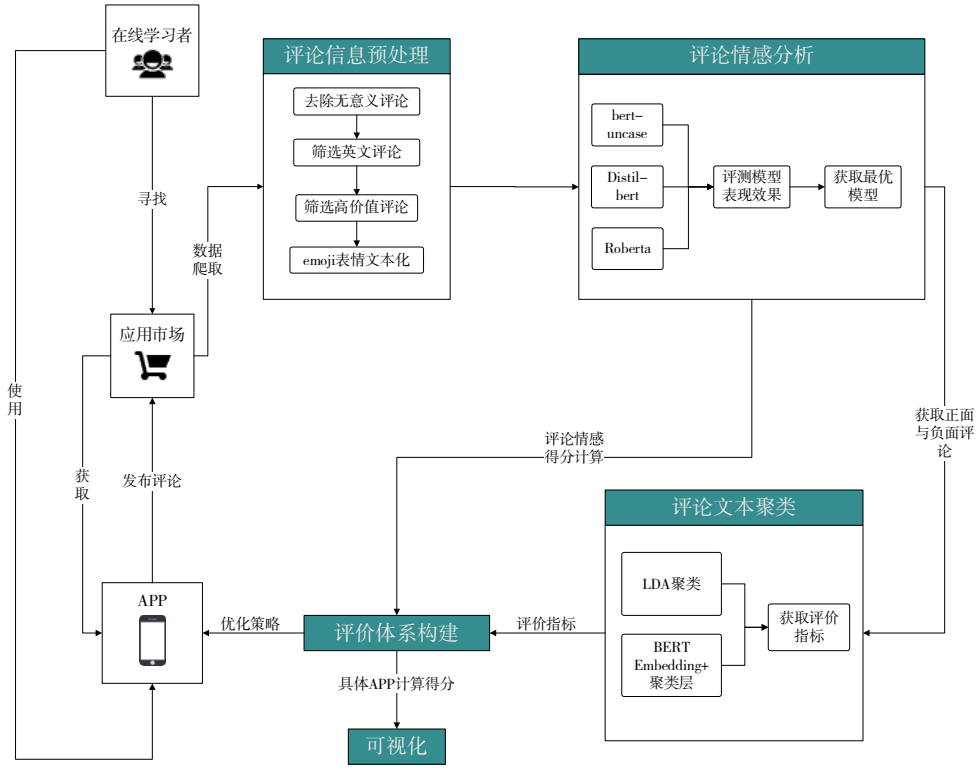


Fig. 1 Technical roadmap of the overall experiment

图 1 实验整体技术路线

2.2　数据集构建

2.2.1　数据采集

本文选取移动应用市场谷歌商城中 20 个下载量最高的语言学习 APP 作为研究对象，对于每一个 APP 的评论区，抓取每条评论的评论 ID、评论星数、评论内容、评论点赞数。抓取的时间范围限定在 2023 年 1 月 1 日—2023 年 9 月 1 日，以获取用户对最新版 APP 的真实学习体验，最终获取到评论数据 41 590 条。本文在数据采集过程中遵循谷歌商城的服务条款与使用协议，并通过限制数据访问范围与匿名化处理，确保收集数据的合法性和相关用户的隐私性。

2.2.2　数据清洗及入库

本文抓取到的用户评论涉及 10 余种不同语言，为保持后续处理的统一，选择了其中占比最大的英文评论作为处理对象。同时，本文还删除了文本过短或内容重复等没有分析价值的评论，如“good good good…”等。之后，本文又将清洗过的评论按点赞数降序排序，这一指标代表了该评论在 APP 评论区中被其他用户表示赞同的次数。由于从各个 APP 抓取到的评论数有较大差异，删除评论量大的 APP 部分点赞数为 0 的评论。该方式既可以剔除低价值的评论，又可以实现数据均衡。需要说明的是，部分评论中包含 emoji 表情符号，此类符号往往直观表达了用户情感，若简单地将其从文本中去除将会损失评论的情感信息[12]。本文选择将其转换为官方的释义文本，如将表情符号“悲伤”替换为文本“pensive_face”，以便后续进行文本处理。经过清洗并预处理后的有效评论共有 34 781 条，评论数据已开源（https://huggingface.co/datasets/Moonveiler/Google-Play_Reviews_of_Language_Learning_Apps），本文将其作为研究样本作进一步分析。

3　评论文本情感分析

3.1　情感分析模型微调

本文对清洗后的评论文本进行情感分析，试图将评论分类为正负两面，以便在之后对二类评论分别建模，从不同角度探求学习者需求。在近年的研究中，预训练语言模型 BERT 展示了双向上下文模型的优势，在多个权威的情感分析任务上取得了良好效果[11]。相较于传统方法，BERT 通过预训练习得了丰富的语言表示，使其在在线评论分析等短文本任务中能够更好地捕捉语义信息和上下文关系，同时对于用户评论中部分不严谨的表述，BERT 可以处理其中的多义性，从而减少对于此类文本的误解[13]。在自然语言处理任务中，BERT 还有一些综合表现较好的变种模型，如 DistilBERT 与 RoBERTa[14]。其中，DistilBERT 是 BERT 的轻量级版本，通过去除部分层和减少参数数量来提高模型推理效率[15]。RoBERTa 则在预训练过程中采用更长的训练时间与更大的批次，通过这些修改来提高模型性能。但与此同时，其可能需要更多计算资源[16]。此外，本文还选择了 InstructABSA 这一基于大语言模型的情感分析模型进行测试。该模型以最新的大模型架构为基础，能提供更丰富的情感细节分析功能，在本任务上可能会有相较预训练模型更佳的表现[17]。

综上，为选择出本任务最适配的模型，本文采用 BERT 模型及其变体 DistilBERT、RoBERTa，以及 InstructABSA 模型对评论文本进行情感分析，并观察其表现。

在微调阶段，模型参数需要通过与任务相关的标记数据集上的有监督学习进行调整，以适应特定的任务。为获取标记数据集，本文将每条评论数据中的评价星数视作该评论文本对应的情感标签。由于本文将评论数据进行正面与负面情感的二分类处理，因此把取值区间 1~5 的 5 种评价星级归类为正负两类。其中，将 5 星与 4 星看作正面，3 星及以下看作负面。3 星虽作为中性的分数，但其关联的评论文本中涉及更多的是不满与问题，故将其加入负面行列。

3.2　情感分析模型对比分析

经过对评论数据训练测试集的划分，在相同的评论数据集上训练并微调 BERT、DistilBERT、RoBERTa 和 InstructABSA 模型，并比较 4 种模型的表现。实验结果见表 1。

Table 1　Performance comparison of models

表 1　模型表现比较

主题	准确率	精确率	召回率	F1
BERT	0.754	0.757	0.993	0.859
DistilBERT	0.868	0.872	0.967	0.917
RoBERTa	0.776	0.775	1.000	0.873
InstructABSA	0.769	0.817	0.932	0.871

经过对比实验，可以看出 DistilBERT 在各项模型性能衡量指标中皆有明显优势，而 RoBERTa 的召回率异常高，可能是因为相较于 BERT 和规模更大的 RoBERTa 与 InstructABSA，DistilBERT 在资源受限环境下的泛化性能更强[18]。此外，DistilBERT 相对 BERT 的改进见图 2。

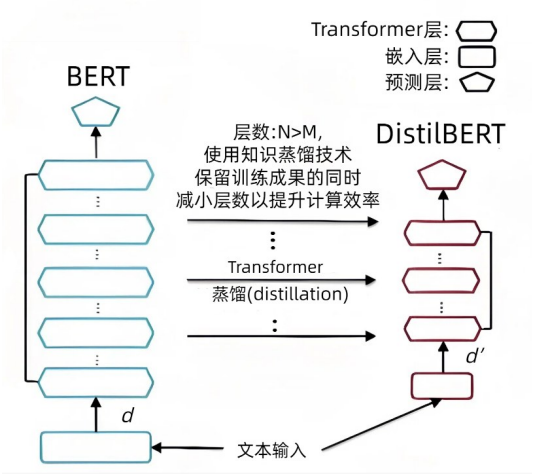


Fig. 2　Improvements of DistilBERT over BERT

图 2　DistilBERT 相对 BERT 的改进

基于此，本文选取 DistilBERT 作为之后实验的情感分析模型，利用 DistilBERT 计算每条评论的情感得分，并将评价星数归一化处理后（如 2 星转换为 0.4 分）与情感得分取平均值得出评论的总得分。总得分将反映一条评论对该 APP 的整体评价与情感态度。本文以总得分 0.6 分为界限，将高于界限的评论划分为正面评论，反之则划分成负面评论，以便后续分别进行评论文本的聚类分析。

根据模型分类结果显示，评论中的正面评论与负面评论占比相差不大，这与评价星级中显示的情况不同。在评论数据的评价星级中，高分评论占比远高于低分评论，表示有许多高分评论被模型分入了负面评论中。这是因为大多数学习者即使给出了较高评分，但在评论内容中提及更多的是 APP 尚存的缺点。为了之后的对比分析，这些评论需要被分类为负面评论，以挖掘学习者对 APP 的不满之处，这也显示了研究中采用深度学习模型的必要性。经过进一步的人工清洗与模型分类后，共得到正面评论 23 979 条，负面评论 10 802 条。各类评论数量与占比见表 2。

Table 2 Sentiment distribution of comments

表 2 评论情感分布情况

情感类型	评论量	占比/%
正面	23 979	57.6
负面	10 802	42.4

3.3 评论聚类分析

3.3.1 LDA 主题聚类

首先，本文对正负面评论文本进行分词、词形还原等处理，并通过控制词频范围剔除无意义且高频的停用词或出现次数过少的词。之后，选择 LDA（Latent Dirichlet Allocation，隐含狄利克雷分配）模型进行主题建模，其通过在大量文档中推断出一组主题，使得每个主题对应一组词汇分布[19]。对于 LDA 模型的参数 α 和 β，常用设置为 α = 50/K，β =0.01，其中 K 是文本主题的数量，通常需要取一个适中的值。若主题数量过少，让单个主题包含太多聚类，会导致主题不明确；若主题数量过多，各个主题可能界限不清晰，难以进行特定解释。经过多轮实验观察，在最终的 LDA 模型中，选择 K = 5，迭代次数 iter = 500。

在实验中，自评论中聚类出的 5 个主题由在许多文档中出现词汇的概率分布表示。这意味着一个主题可能由一组经常共同出现的词汇组成，这些词汇共同描述了某一概念或主题。这些词汇与相应主题的关联概率 $P(w|z)$ 计算公式为：

$$P(w|z) = \frac{n_{w,z} + \beta}{n_z + V \cdot \beta} \quad (1)$$

其中，$n_{w,z}$ 是在主题 z 中词 w 出现的次数；n_z 是在主题 z 中所有词出现的总次数；β 是前文中提及的超参数，用于词 Dirichlet 分布的平滑；V 是词汇表大小。受限于篇幅，表 3、表 4 仅展示了部分主题与相关词汇数据。

从表 3、表 4 可以看出，用户的正面和负面评论有几个相似的聚类：①正面的主题一和负面的主题一，高频词有“单词”“汉字”“短语”等。正面评论多是夸赞应用的教学资源丰富，而负面评论则是反馈知识内容不足。②正面的主题三和负面的主题三，涉及“简单的”“友好的”“有帮助的”和“失望的”“困惑的”等词汇，以及一些表达情绪的表情符号如 frowning_face。这些主题词表现出用户使用 APP 过程中不同的学习体验。③正面的主题四和负面的主题二、主题四，高频词有“版本”“更新”“服务”“修复”等。这些多体现了用户使用 APP 时遇到的技术方面的问题，可见随着 APP 版本的迭代，会有不同的问题得以解决或产生。正面评论多表达学习者对问题得到解决的欣慰，负面评论则多是催促开发者及时修复软件漏洞，或是投诉客服的消极服务态度。

Table 3 Keywords and associated probabilities of positive comments under the LDA model

表 3 LDA 模型下的正面评论关键词与关联概率

主题	关键词	关联概率
主题一	Word，Phrase，Grammar，App，Sentence…	0.039，0.023，0.022，0.022，0.020…
主题二	Translator，Method，Practice，Clip，Movie…	0.115，0.094，0.051，0.019，0.017…
主题三	Simple，Wow，Thank，Help，thumbs_up…	0.126，0.103，0.051，0.049，0.048…
主题四	Free，Version，Add，User，Update…	0.066，0.037，0.028，0.014，0.013…
主题五	App，Than，Other，Better，Duolingo…	0.075，0.063，0.048，0.040，0.028…

Table 4 Keywords and associated probabilities of negative comments under the LDA model

表 4 LDA 模型下的负面评论关键词与关联概率

主题	关键词	关联概率
主题一	English，Language，Native，Speaker，Word…	0.051，0.047，0.038，0.026，0.020…
主题二	App，My，Work，Problem，Fix…	0.044，0.033，0.031，0.022，0.019…
主题三	Disappoint，Confuse，Bad，Patience，frowning_face…	0.026，0.018，0.017，0.014，0.012…
主题四	Customer，Service，Bad，Support，Cancel…	0.048，0.042，0.021，0.015，0.014…
主题五	App，Subscription，Pay，Premium，Money…	0.033，0.032，0.027，0.026，0.017…

正面和负面评论也有各自独特的聚类结果，具体分析如下：①正面的主题二，高频词有“翻译”“方法”“练习”“视频”“教学卡片”等，多是对 APP 创新或有效的学习策略表示肯定。②正面的主题五，高频词有“比”“其他”“更好”“最好”等表示比较的主题词，还涉及“Duolingo（多邻国）”等其他同类竞品。此类评论基本是用户基于个人使用多种语言学习 APP 的经历并作出比较的评论。③负面的主题五，高频词有“订阅”“付款”“高级版”“广告”“价格”等付

费相关的词。当下应用市场下载量最高的APP大多含有付费内容，用户可以选择通过订阅会员或一次性付费来获取更佳的学习体验，包括但不限于无广告的应用界面、更丰富的学习内容等。但这类评论多是负面评论，说明用户针对付费内容的看法多为“价不符实”。

3.3.2　BERTopic主题聚类

除LDA方法外，本文也运用BERTopic对评论进行主题聚类。BERTopic是一种基于BERT词向量进行主题建模的方法，其结合了BERT强大的自然语言处理能力以及主题模型的优点，可快速实现对大规模文本数据的解析，并对其进行无监督的主题提取[20]。但针对本文较小的数据集，BERTopic的表现一般，聚类结果许多集中在特定的APP上，即模型偏向将特定APP的评论归为一类，而这种效果是实验不需要的。因此，对于Bertopic的工作，本文只选取有价值的部分结果作为之后量规构建的参考。

4　评价体系构建与可视化案例

4.1　APP评价体系构建

基于LDA主题聚类综合BERT嵌入+聚类层分析得到的结果，本文首先合并相似的聚类结果（如正面主题一与负面主题一），建立起“知识内容”“学习结果”“技术环境”3个一级维度。然后根据各聚类下涉关键词的词频，将部分提及频率较高的聚类“学习策略”与“竞品比较”选取为一级维度。最后，剩余的一个聚类与APP的付费内容相关，本文将其和一些高频但与其他聚类关联概率较低的主题词进行整合，确立为最后一个一级维度“支持服务”。在这6个一级维度的基础上，本文又结合关键词与具体评论内容，人工划分了共12个二级维度，并整理出了评论中从属于各维度的关键词，最终形成语言学习APP的评价指标体系，见表5。

根据表5，具体评价指标说明如下：

（1）知识内容。该一级维度代表了学习者能在APP中获取到的学习资源。其中，“语言知识”包含了各类语言要素知识，如词汇、语法；“语种需求”可以评估该款APP是否可以为学习者提供特定语种的学习内容，以满足差异化和个性化的学习需求。从主题词分析的排名可以看出，学习者在使用语言学习APP时仍然最重视自己将获得的知识内容，这也反映出APP评价应坚持“内容为本”的根本原则，确保APP所呈现的学科知识兼具科学性与丰富性。

（2）学习方式。该一级维度代表了学习者使用APP过程中习得知识的方式。其中，“工具属性”表示该款APP辅助语言学习时的实用性和易用性；“学习策略”表示APP是否为学习者提供适需的学习方法和活动，如单词卡片、闯关答题、真人视频等协作式、探究式、游戏式的语言学习活动。时下热门的语言学习APP大多具有自己代表性的学习方式，这些创新性的学习方式可以激发学习者的学习兴趣，促进语言知识的习得。根据笔者调查，语言学习APP与教育APP的用户相同，多为自我导向学习者，因此APP需要用科学的方式协助学习者选择学习策略，帮助学习者规划语言学习过程。

（3）学习结果。该一级维度代表了学习者使用APP学习后获得的成果，含有3个子维度。“知识掌握”表示学习者语言知识的积累和认知领域的发展；“能力提升”表示学习者听、说、读、写、译等语言技能的提高；“情感体验”表示学习者对APP辅助学习的情绪感受（包括学习的积极性和主动性）。当学习者使用了一段时间的APP后，大多数APP会邀请用户到应用商店进行反馈，这些反馈信息大多集中在此分类中。

（4）技术环境。该一级维度可用于评价APP的技术属性（如界面美观、系统运行、功能实现等）。其中，“界面设计”表示学习者对交互界面的主观感受；“程序系统”表示APP运行的稳定性。技术环境是APP相对于其他学习工具较为独特的属性，APP作为软件产品，需要通过用户界面设计和底层技术支持来完善软件，以提供更好的用户体验。通过评价APP的技术环境，可以判别APP能否为用户的语言学习提供稳定、高效的数字学习环境。

Table 5　Dimensions of the evaluation index system for data-driven language learning APP

表5　数据驱动的语言学习APP评价指标体系维度

一级维度	二级维度	标签词示例
知识内容	语言知识	word，phrase，grammar，sentence，vocabulary，voice，speak等
	语种需求	English，Chinese，Japanese，Spanish等
学习方式	工具属性	translate，community，search，dictionary等
	学习策略	method，practice，movie，clip，flashcard，course，lesson，listening，reading，audio等
学习结果	知识掌握	simple，hard，help，learn，confuse，quick等
	能力提升	improving，useful，insufficient等
	情感体验	love，disappointed，sloppy，angry，enjoy，patience等
技术环境	界面设计	interface，UI，friendly等
	程序系统	version，add，user，fix，update，phone，Android等
支持服务	用户反馈	support，service，customer，mail，membership等
	付费情况	pay，subscription，premium，Ad，worth，worthy，price等
竞品比较	竞品比较	APP，than，better，other，best，ever等

(5)支持服务。该一级维度代表了学习者与APP开发者的部分互动。其中,“用户反馈”表示学习者通过评论或邮件等形式请求开发者对APP作出改动或修复以及学习者对开发者回复作出的评价;“付费情况”则围绕付费内容探讨一个APP是否要求付费或其付费策略的合理性。与技术环境相似,支持服务也体现着APP作为软件产品的属性。学习者同时作为产品的消费者,可以对APP提出更多要求,并对产品的价格作出评价。通过评价APP的技术服务,可以判断APP是否积极响应学习者需求,不断迭代更新,更好地为学习者提供技术支持。

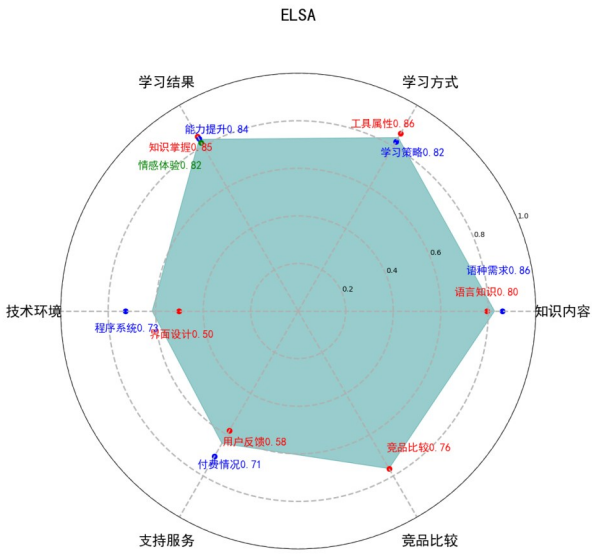
(6)竞品比较。该一级维度反映学习者将以往使用过的语言学习APP与当下评论的APP的学习体验进行比较。该指标并未进一步划分二级维度,这是因为用户比较多款APP时所用指标与以上一级指标有所重合。学习者在面对大量同类型产品时,可能会更加重视产品特性的比对。评价APP相对其竞品的优劣,可以帮助学习者进行区分与选择。

本文在每个二级维度下都设有与主题相关的数十个关键词。当进行自动评价时,首先,机器会针对每条评论文本靶向地检索和匹配这些关键词,并将评论归类到相应的一个或多个维度下;其次,根据前文所述的情感得分计算方式测算出评论情感分值;再次,对每个维度下的所有评论求加权均值,其中每条评论的权重取决于该评论在应用市场评论区的获赞数,该指标直观表达了其他用户对该条评论的认可度;最后,针对某个APP下的所有评论进行总得分计算、评论分类与加权计算后,便可得到该APP在本文评价指标下的综合得分。

4.2 APP评价可视化案例

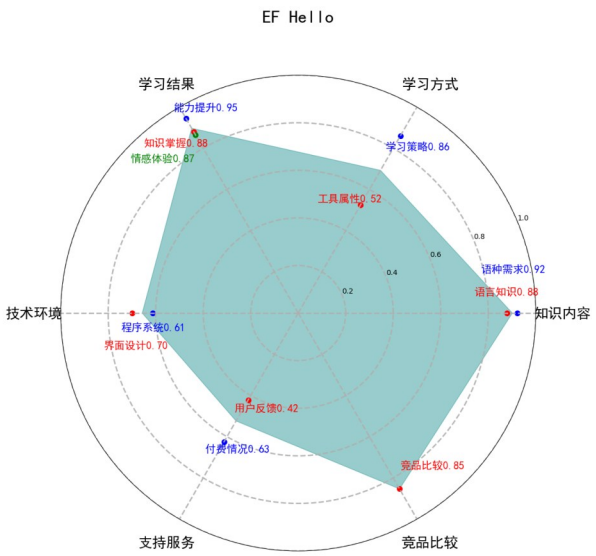
在前文构建的评价体系基础上,本文又对各款APP及其对应的评论得分通过Python第三方库(matplotlib)进行评级数据的可视化。受限于文章篇幅,本文以ELSA(一款语音助手式的外语学习APP)与EF Hello(一款面向成人、提供付费外语课程的APP)两款不同类型的APP为例,展示其评价数据的可视化结果,见图3。

由图3可知,雷达图的6个坐标代表了评价体系中的一级维度,坐标轴上显示了被评价APP在每个二级维度的对应得分,图形的半径取值为对应二级维度的平均值。学习者在选择APP时可以通过图形的涂色形状与面积直观了解每款APP在各方面的评分以及综合表现。以展示的两款APP为例,ELSA得分整体较高,相对得分较低的维度是技术环境与支持服务,表现为界面设计不合理以及开发者对用户反馈不及时;EF Hello则可以提供令用户满意的知识内容与学习方法,有助于提升学习者的语言学习效果,但作为提供付费课程的APP,其工具属性较弱,且在支持服务维度上的低分也体现出其付费策略的不合理,以及开发者对用户意见的重视不足,这可能会制约该软件的持续健康发展。



(a) Evaluation of ELSA

(a) ELSA的评价



(b) Evaluation of EF Hello

(b) EF Hello的评价

Fig. 3 Radar chart of evaluation data for two APPs

图3 两款APP评价数据可视化雷达图

5 结语

教育APP作为一种“以学习者为中心”的学习资源,在当下移动学习环境愈加普及的背景下,已存在的量化评价体系却只局限于单一的评分。本文秉持“数据循证”的评价观,采用“自下而上”的研究路径,依托在应用市场获取的在线评论数据,通过学习数据采集、情感分析模型微调、评价量规构建、数据可视化等系列流程步骤,建立起一套

数据驱动的语言学习APP评价体系，将预训练语言模型与大数据挖掘相结合应用于教育APP评价研究中。未来仍需采集语言学习APP的全息数据（如使用行为数据、人机交互数据、学习结果数据等），不断优化评价体系，以期实现教育APP评价的综合化、多元化和科学化。

参考文献：

[1] MA R L, LIANG Y. The triple logic of digital transformation in international Chinese language education: starting from ChatGPT [J]. Journal of Henan University (Social Science Edition), 2023(5):112-118.
马瑞祯，梁宇. 国际中文教育数字化转型的三重逻辑——从ChatGPT谈起[J]. 河南大学学报(社会科学版)，2023(5):112-118.

[2] iResearch. China mobile application trend insights white paper—online education 2020 [R]. 艾瑞咨询系列研究报告，2020.
艾瑞咨询集团. 2020年中国移动应用趋势洞察白皮书——在线教育篇[R]. 艾瑞咨询系列研究报告，2020.

[3] LIANG Y, XU J. Innovative research on Chinese online learning platforms based on connectivism [C]//Ho Chi Minh City: The 12th International Symposium on Modernization of Chinese Language Teaching, 2021.
梁毅，徐娟. 基于联通主义的汉语网络学习平台创新研究[C]//胡志明市：第十二届中文教学现代化国际研讨会，2021.

[4] MA R L, XU J. Empowering international Chinese smart education with language intelligence: current situation and future directions [J]. International Chinese Language Education, 2023(2): 43-52.
马瑞祯，徐娟. 语言智能赋能国际中文智慧教育：现实境况与未来路向[J]. 国际中文教育(中英文)，2023(2):43-52.

[5] LYU S. An empirical study on expanding college English speaking learning with the "English fluency" mobile APP [J]. Journal of Fujian Radio and Television University, 2017(1):18-21.
吕双. "英语流利说"手机APP拓展大学英语口语学习的实证研究[J]. 福建广播电视大学学报，2017(1):18-21.

[6] ZHOU X Q, JIAO J L, ZHAN C Q. A systematic literature review of mobile learning pedagogy for primary and secondary education [J]. Modern Educational Technology, 2023, 33(8):58-66.
周晓清，焦建利，詹春青. 中小学移动学习教学法研究的系统性文献综述[J]. 现代教育技术，2023，33(8):58-66.

[7] iResearch. Mobile application operation growth insights white paper [R]. 艾瑞咨询系列研究报告，2022.
艾瑞咨询集团. 移动应用运营增长洞察白皮书[R]. 艾瑞咨询系列研究报告，2022.

[8] YANG L N, WEI Y H, XIAO K X, et al. Research on personalized learning service mechanisms driven by educational big data [J]. Research in Educational Electronics, 2020, 41(9):68-74.
杨丽娜，魏永红，肖克曦，等. 教育大数据驱动的个性化学习服务机制研究[J]. 电化教育研究，2020，41(9):68-74.

[9] YANG X M, TANG S S, LI J H. Developing educational big data: connotations, value, and challenges [J]. Modern Distance Education Research, 2016(1):50-61.
杨现民，唐斯斯，李冀红. 发展教育大数据：内涵、价值和挑战[J]. 现代远程教育研究，2016(1):50-61.

[10] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019:4171-4186.

[11] CHENG Q, ASTON Z, ZHANG Z S, et al. Is ChatGPT a general-purpose natural language processing task solver? [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023:1339-1384.

[12] YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding [DB/OL]. https://arxiv. org/abs/1906. 08237.

[13] JIN Y C, DENG C L, WU P, et al. The social functions and applications of emoji symbols [J]. Advances in Psychological Science, 2022, 30(5): 1062-1077.
靳宇倡，邓成龙，吴平，等. Emoji图像符号的社交功能及应用[J]. 心理科学进展，2022，30(5):1062-1077.

[14] IEVA S, IGNACIO I. Compositional and lexical semantics in RoBERTa, BERT and DistilBERT: a case study on CoQA [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020:7046-7056.

[15] SANH V, BERTDEBUT L. Distil, a distilled version of BERT: smaller, faster, cheaper and lighter [DB/OL]. https://arxiv. org/pdf/1910. 01108.

[16] LIU Y, OTT M. Roberta: a robustly optimized BERT pretraining approach [DB/OL]. https://arxiv. org/pdf/1907. 11692.

[17] SCARIA K, GUPTA H, GOYAL S, et al. Instructabsa: instruction learning for aspect based sentiment analysis [DB/OL]. https://arxiv. org/pdf/2302. 08624.

[18] VARIA S, WANG S, HALDER K, et al. Instruction tuning for few-shot aspect-based sentiment analysis [DB/OL]. https://arxiv. org/pdf/2210. 06629.

[19] EKLUND A, FORSMAN M. Topic modeling by clustering language model embeddings: human validation on an industry dataset [C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track, 2022:635-643.

[20] TAEHUN C, DONGHUN L. SentenceLDA: discriminative and robust document representation with sentence level topic model [C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, 2024:521-538.

（责任编辑：黄　健）