Data in Brief
Open access
Article template

文章信息

文章标题

A 用于用户反馈分析的亚马逊应用商店低评分应用数据集

作者

Nek Dil Khan[1], Javed Ali Khan[2], Darvesh Khan[3], 李建强[1]*、Mumrez Khan[4], Shah Fahad Khan[5]

所属机构

[1]北京工业大学信息技术学院，北京，北京，100124，中国；
nekdilkhan@emails.bjut.edu.cn, lijianqiang@bjut.edu.cn

[2]赫特福德大学物理、工程与计算机科学学院计算机科学系，哈特菲尔德，AL10 9AB，英国；
j.a.khan@herts.ac.uk

[3]CECOS信息技术与新兴科学大学软件工程系，白沙瓦，开伯尔-普赫图赫瓦省，巴基斯坦
darwesh.khan.bsse-2021b@cecosian.edu.pk

[4]西安理工大学计算机科学与工程学院，中国西安，710121；mumrez.khan@stu.xaut.edu.cn

[5]大连理工大学计算机软件工程学院，中国大连116000sfahad@mail.dlut.edu.cn

通讯作者的电子邮箱地址和 Twitter 账号

李建强： lijianqiang@bjut.edu.cn ; Javed Ali Khan： j.a.khan@herts.ac.uk

关键词

用户反馈数据集、软件质量、低评分应用、问题分类

摘要

在当今的数字化环境中，终端用户反馈在软件应用的演进过程中发挥着至关重要的作用，尤其有助于解决阻碍用户体验的问题。尽管许多研究聚焦于高评分应用，但低评分应用往往仍未得到充分探索，尽管它们有可能揭示宝贵的见解。本研究介绍了一个新颖的数据集，该数据集整理自亚马逊软件应用商店（ASA）中的 64 款低评分应用，包含 79,821 条用户评论。该数据集旨在捕捉用户指出的最常见问题，这些问题对于提升软件质量至关重要。为进一步提升数据集的实用性，我们对其中 6000 条评论进行了人工标注，并将其分为六个不同的问题类别：用户界面（UI）和用户体验（UX）、功能与特性、兼容性和设备特异性、性能与稳定性、客户支持与响应能力，以及安全性与隐私问题。这个标注数据集

是开发基于机器学习的方法的宝贵资源，可用于自动将用户反馈归类为各种问题类型。公开提供标注数据集和原始数据集，为研究人员和开发人员提供了重要工具，帮助他们了解低评分应用中的常见问题，并为软件改进提供参考。对该数据集的全面分析及其开放获取，为基于数据的解决方案奠定了基础，使人们能够依据用户反馈提升软件质量。此外，该数据集还为软件供应商和研究人员提供了探索各种与软件演进相关活动的机会，包括经常缺失的功能、讽刺及相关情绪，从而有助于更好地理解应用评分相对较低的原因。

数据规格

表 1 列出了数据规格，以便更清晰地呈现。

表 1： 规格

规格康_北京pk	详情
主题	计算机科学、软件工程
具体研究领域	软件演进、众包需求工程、自然语言处理（NLP）、用户体验（UX）
数据类型	表格型
数据收集	使用谷歌浏览器扩展程序 Instant Data 自动抓取网页从网页中提取数据并将其导出为 Excel 或 CSV 文件的抓取工具。 Instant Data Scraper：
数据来源位置	北京市朝阳区平乐园100号，北京工业大学， 中华人民共和国北京市，邮编100124。
数据可访问性	原始数据集和已标注数据集已发布，详情如下 。 存储库名称：Mendeley 数据标识号： 10.17632/zztwrn36n8.2 数据直接链接： https://data.mendeley.com/datasets/zztwrn36n8/1 GitHub 链接：https://github.com/nekdil566/Dataset-of-User-Reviews-from-Low-Rated-亚马逊-应用商店-应用4ons/blob/main/README.md
数据量	原始数据集（12.1 MB），标注数据集（）
数据质量保障	该数据集经过严格的质量保证流程，以确保一致性和准确性。此外，还会对其进行清理，以删除重复项和不一致项，

	或无关信息，确保其全面且准确。经过标注的评论由多名编码者进行核验，以实现较高的编码者间一致性，进一步提高数据集对机器学习应用的可靠性。该数据集提供了高质量的人工标注数据，适用于问题检测、观点分析和情感分析等高级分析任务。
潜在用途	该数据集可以通过精准定位常见的用户问题来提升软件质量，从而便于有针对性地更新和增强功能。这些数据很有价值，可用于训练机器学习（ML）模型，以对用户反馈进行分类，并自动识别反复出现的问题。此外，该数据集还支持情感分析、用户体验研究和市场分析，为了解用户提供洞见用户不满，并为以用户为中心的软件开发提供参考。
相关研究文章	挖掘软件洞见：揭示低评分软件应用中频繁出现的问题 [1]。

数据价值

- **数据集用例：**该数据集包含丰富的终端用户评论，共计 79821 条，收集自 Amazon Play Store，覆盖 14 个类别中的 64 款应用。与以往方法不同，该数据集聚焦于评分相对较低的应用，使软件供应商、开发人员和研究人员能够专注于导致低评分的常见问题。为此，研究人员利用该数据集，通过微调机器学习和深度学习算法分析终端用户反馈，以识别应用商店中频繁出现的问题 [1]。此外，他们还使用 ChatGPT 和经过微调的深度学习算法，从定制数据集中识别评论中蕴含的终端用户情绪 [3]。通过公开这一数据集，研究界可以探究负面反馈的根本原因，为更有效的软件改进提供依据，并最终提升整体用户体验。此外，研究首先利用定制数据集，通过词性分析识别频繁出现的问题。随后，研究人员使用经过微调的机器学习分类器识别相关情绪，从而加深对终端用户不满之处的理解 [2]。此外，有报告指出，单条评论可能涉及多个问题，例如性能和安全性。因此，我们认为有必要重新标注该数据集，将其用于多标签分类问题；据我们所知，这一问题迄今尚未在软件工程文献中得到探讨。这些模型可以简化从终端用户反馈中提取有用信息的流程，为开发人员提供一种可扩展的工具，以便快速识别需要关注的方面。

- **推动用户满意度根因分析：**该研究数据集使研究人员和软件供应商能够采用主题建模技术（如潜在狄利克雷分配（LDA）），从负面反馈中挖掘潜在主题，从而找出导致应用评分偏低的系统性问题。它还将通过使用 LDA 算法识别频繁出现的问题，帮助验证研究发现 [1]。这项分析有助于免去数据标注流程，该流程既耗时又具有挑战性。此外，研究人员可以将主题聚类与功能组件关联起来，支持有关软件质量和用户感知缺陷的实证研究，并提出解决方案，以提升现有低评分软件应用的质量。

此外，还可以在句子级别重新标注该数据集，并将其与评论级别的性能进行比较，以识别问题。

- **促进机器学习模型的开发与验证：**该数据集包含79,821 条用户评论，其中 6,000 条经过人工标注，归为低评分软件应用中六种最常见的问题类型。该数据集可作为训练和评估监督式与半监督式分类模型的标准。通过进一步了解软件应用低评分的可能原因，还可以重新标注该数据集。例如，可以开发自动化流程来识别讽刺及其类型（正面和负面），以提升现有情感分析方法的性能 [4]。同样，也可以利用低评分数据集探索更细粒度的新特性，因为正如我们所提出的，用户可能报告了某些缺失功能，而这些缺失功能可能导致了低评分。此外，该数据集为验证和比较众包需求工程中使用的不同分类框架提供了可靠依据。为此，可以利用带标签数据集开展消融研究，以确定研究方法的泛化能力。研究人员可以重复使用这些数据，对 Transformer 或大语言模型（LLM）进行微调，用于特定领域的应用评论分析，从而提升模型在低资源场景下的泛化能力。

- **软件演进中的问题或特性优先级：**在现有文献中，研究者提出了许多方法，将评论归类为频繁出现的问题类型或特性。然而，有些问题或特性比其他问题或特性更为关键，需要开发人员立即响应。为此，可以先利用数据集，通过自然语言处理流程、词性标注和主题建模，识别终端用户提及的频繁出现的问题或特性。然后，通过统计共现次数，量化有多少条终端用户评论提及了各个问题或特性。可以提供一种工具，支持软件开发人员对频繁提及的问题或特性进行排序，以便立即采取行动。此外，还可以加入情感分析，以便频繁识别问题或特性，从而加深对已报告问题和特性相关的用户背景及观点的理解，最终改进排序。

- **为软件工程和用户体验课程体系提供参考：**近年来，由于人工智能（AI）和机器学习应用广泛且成效显著，它们已被纳入软件工程和计算机科学课程体系。同样，需求收集并不局限于内部用户；对于面向市场的应用，需要通过分析社交媒体平台上的终端用户反馈，自动收集需求。该数据集为学术人员提供了机会，可用于演示众包需求工程的各种概念，例如构建基础流程，将评论分类为各种与需求相关的信息，包括特性、问题和非功能性需求；还可以重新标注数据集，以更好地展示机器学习周期。它有助于在需求工程、可用性工程和软件演进课程模块中开展体验式学习。学生可以借助真实反馈获得实践经验，了解糟糕的用户体验、性能问题或缺乏支持如何导致应用失败 [1]。

- **提升客户服务：**文献中已有证据表明，应用商店及其他社交媒体平台上针对软件应用的终端用户反馈日益增多。为了提升客户服务，在这些平台上逐一手动回复每条终端用户反馈，既困难又费时。不过，通过使用自然语言处理分析终端用户反馈并提取问题类型，该数据集可作为真实值，用于开发自动化且更完善的客户服务系统。随后，可采用分类方法将问题归入预定义类别，例如

漏洞、性能、安全性等。识别问题类型后，即可从与该类别关联的词典中提取预先设定的消息。此外，还可运用先进的自然语言处理技术，提取用户姓名或确认已识别的问题，使消息更具个性化。除此之外，还可以使用 GPT、BERT 或 T5 等先进的 Transformer模型，通过识别问题类型、相关情绪、语气和意图，对终端用户评论进行语境理解。系统应基于这些信息自动添加定制回复，并以更完善、更及时的方式作出回应，从而帮助留住客户并提升质量。

背景

在 ASA 等数字市场竞争激烈的领域，如何维持终端用户的高满意度，对开发人员而言是一项重大挑战。近期研究一再表明，分析用户反馈是获取宝贵见解的重要替代途径，有助于提升现有内部利益相关者所关注的软件质量和功能。尽管所有终端用户反馈都很有用，但低评分应用的评论尤其重要，因为它们往往会详细描述软件故障、可用性问题、未满足的用户需求以及缺失功能。通过专门关注这些用户不满意的情况，该研究数据集为深入分析导致应用负面评分的因素奠定了丰富基础。本研究在基础研究的基础上，运用自然语言处理和机器学习技术，从用户反馈中提取可付诸行动的见解 [1] [2]。通过将低评分应用数据集补充到现有数据集中，该研究为开发人员创建了一个至关重要的反馈循环，帮助他们了解软件应用评分相对较低的原因。它也是一项宝贵资源，可用于训练和验证新一代分析模型，推动以用户为中心的软件开发。

数据描述

该数据集以一张内容详尽的表格呈现，涵盖亚马逊应用商店中 14 个类别的 64 款软件应用。数据集中的每条记录都提供了应用的具体信息，包括其类别、名称和终端用户评论数量，如表2所示。数据集的结构如下。

第 1 列（序号（S/no.））：此列按顺序为各条记录编号，编号范围为 1 至 14，每个编号代表一个独特的应用类别。

第 2 列（应用类别）：此列列出了应用类别。总共有 14 个类别，包括商务应用、通信应用、教育应用、食品和饮料应用、游戏应用、影视应用、音乐和音频应用、新奇应用、照片与视频应用、生产力应用、体育和健身应用、实用工具应用、生活方式应用和旅行应用。选择这些类别是为了使数据集更具通用性，能够代表主要的应用类别。

第 3 列（软件应用）：各类别下列出了具体应用。例如，Hammer Print、Sketch Guru、Sketchbook、Logo Maker 和 Logo Creator 等应用都列在商务应用类别下。各类别包含不同的应用，所有类别共计 64 个应用。

第4列（终端用户评价）：此列提供各应用类别的终端用户评论数量。这些数据概述了各类别的用户参与度和反馈情况。

表2。数据集 [1] 概述。

S/No	应用类别	软件应用	终端用户评价
1	商务应用	1：Hammer print，2：Sketch guru，3：Sketch book，4：Logo maker & logo creator	11034
2	Communica@on Apps	1：Just talk，2：Texteme，3：Skype 4：Add Me Snaps，5：免费 text	5922
3	教育应用	1：TED TV，2：Casper's company，3：World now，4：亚马逊 silk：duolingo.com	6968
4	食物和饮料 Apps	1：Food Network 2：食物规划器，3：意大利食谱，4：厨师 tap：食谱剪藏工具，规划器	5521
5	游戏应用	1：Mobile Strike，2：汽车驾驶蜘蛛模拟器，3：章节：互动故事，4：疯狂动物自拍镜头，5：帝国摇篮：配对	4748
6	电影和电视 Apps	1：Anthem TV–免费观看有线电视，2：Hauppauge my TV，3：Stream TV，4：低俗电影频道，5：TV Player观看电视直播和点播	5982
7	音乐和音频应用程序	1：red karaoke卡拉OK派对，2：MP3音乐下载，3：Red karaoke唱歌和录音，4：变声器，5：Kara Fun 卡拉 OK 与唱歌	1889
8	新奇应用应用程序	1：幽灵雷达：经典版，2：X光扫描仪，3：Cast manager，4：年龄扫描器，5：动态时钟	3104
9	照片和视频应用	1：Air Beam 电视屏幕镜像接收器，2：Air Screen，3：Snappy 照片滤镜和贴纸，4：屏幕投放，5：Rec Me 免费屏幕录制器	5089
10	生产力 Apps	1：PDF max pro–read，2：平面图创建器，3：Office Suite free	5421
11	体育和健身应用程序	1: FOX Sports：直播 NASCAR、拳击，2: fubo TV-观看直播体育、电视节目、电影和新闻，3: NBC Sports，4: CBS 体育直播并观看	10630
12	实用应用	1: Print Bot，2: 平面图创建器，3: @nyCam 监控器免费版，4: 电视屏幕镜像，5: 优化器和垃圾清理工具，适用于 Kindle Fire	3689
13	生活方式应用	1: 屏幕镜像，2: DOGTV，3: Fanmio 拳击，4: 3D 家居设计布局，5：Home Design 3D-免费	4303
14	旅行应用	1：全球网络摄像头，2：指南针 3：Tizi World My Pretend Play Town for Kids 4：My Route Planner Travel 5：公共交通离线传输映射	5521
应用总数		64	79821

数据集以表格形式呈现，便于理解和分析。表3展示了数据集中的若干实例，以说明其结构。该数据集分为四个主要列：

评分：此列记录用户给出的量化星级评分（例如“5 星中得 1.0 星”），是衡量用户对每条评论满意度的主要指标。

评论标题： 此列包含用户提供的评论标题或摘要，通常概括了用户反馈的要点。

完整评论：这是信息最详尽的一列，包含用户反馈的完整文本，也是定性和定量分析的主要来源。

问题类型： 此列用于标识用户在评论中提出的问题类型。每条评论都会根据预定义的问题类型进行分类，包括性能与稳定性、用户界面与用户体验、功能与特性、兼容性与设备特定问题、客户支持与响应能力，以及安全与隐私问题。这些问题类型依据最终用户评论中出现的频率确定。

表3. 数据集属性

评分	评论标题	基础评论	问题类型
1.0	说句公道话，也许如果我有一台无线打印机，我本来会有更好的运气	无法让这个功能ti正常运行，花费了太多精力却还是没能成功，公平地说也许如果我有一台无线打印机，我本来会有更好的运气	功能和特性
1.0	Junk	完全无法使用——即使购买了升级后也是如此。纸张穿过打印机，却没有打印出任何东西！没有可拨打的电话号码申请退款或寻求支持。简直是被骗了。千万别买！	客户支持和响应能力
1.0	不支持安卓?	无法通过我的任何安卓设备打印设备。很难将打印机与应用同步，而且直到.....才会工作。	兼容性和设备
1.0	频繁崩溃和卡死导致应用无法使用	每次尝试加载时，应用都会卡死文档。即使更新后，它仍然存在崩溃和卡死的问题，让它几乎毫无用处。	性能和稳定性
1.0	隐私问题太多问题	隐私问题太多了！这个应用要求不必要的权限，甚至未经明确同意就追踪我的位置。我不会把我的数据交给它。	安全性和隐私
1.0	非常令人困惑	导航起来极其困难，尤其是对于初学者。按钮的标签说明不清晰，我很难找到保存文件等基本功能或更改设置。	用户界面和 UX

实验设计、材料和方法

该数据集的数据采集过程采用系统化方法，收集并分类亚马逊应用商店中评分较低的软件应用信息。具体方法如下：

1. 应用的筛选标准

评分阈值：设定了特定的评分阈值，筛选用户评分为三星（满分五星）或更低的应用。这样可以确保所选应用的评分显著低于各自类别的平均水平。

最低评论数：选择评论数量众多的应用，以确保较低评分反映的是更广泛的用户意见，而非少数孤立看法。

2. 数据收集流程：

自动化抓取工具：使用了一款公开提供的谷歌浏览器扩展程序 Instant Data Scraperhttps://chromewebstore.google.com/detail/instant-data-scraper/ofaokhiedipichpaobibbnahnkdoiiah?hl=en-US&utm_source=ext_sidebar，从亚马逊应用商店自动提取数据。该工具可从网页提取数据，并将其导出为结构化文件格式。

收集的数据点：对于每款应用，均将抓取工具配置为收集评分、评论标题和完整评论，以及其他相关元数据。

3. 数据分类：

人工审核：初步自动收集完成后，对每款应用进行人工审核，以确保其符合筛选标准。随后，根据主要功能将这些应用归入 14 个预定义类别之一。

4. 数据验证与清理：

准确性检查： 对收集的数据进行了验证，以确保其准确性；验证过程包括将抓取的数据与亚马逊应用商店上的实时信息进行交叉核对。

数据清理： 对数据集进行了清理，以删除重复项、不一致项和无关信息。这是维护数据质量和可靠性的关键步骤。

5. 数据汇总：

制表： 最终清理后的数据被整理成结构化表格，使用 Microsoft Excel 编制，以便于使用和访问。

格式tti化： 数据按列整理，包含用户评分（星级）、评论标题和完整评论文本。数据已准备好以 CSV 格式进行最终分发，以确保广泛兼容。

局限性

本研究深入探讨了 ASA 商店中低评分应用的用户反馈；但仍需考虑若干局限。一个重要限制是，由于亚马逊近期调整了评论政策，我们获取终端用户评论的权限受到限制。最初，自动化网页抓取工具能够全面访问大量评论，但如今该工具每个应用最多只能抓取 10 页。这一限制影响了数据集的覆盖范围，可能遗漏 10 页阈值之后评论中提及的用户关注点。因此，数据可能无法充分反映所有问题，尤其是那些较少被提及的问题。可用数据的减少可能导致我们无法完整了解各类应用中常见的用户投诉，

更多评论。数据于2023年4月收集，亚马逊评论政策的更多详情请参阅其官方开发者政策 https://developer.amazon.com/。

此外，数据集的泛化能力仅限于 ASA 商店，只代表全球应用市场的一小部分。不同平台上的用户行为和评论模式可能存在显著差异，因此，基于该数据集得出的结论可能不太适用于其他商店中的应用。该数据集还只是特定时间点的静态快照，这意味着应用更新和变化的动态特性可能未能反映在所收集的数据中，从而可能影响用户评分和反馈。此外，本研究聚焦于至少有 350 条评论的低评分应用，因此未能涵盖较新或评分略高的应用所提供的见解，而这些应用仍可能带来有价值的信息。依赖自动化抓取工具也带来了另一项挑战：应用商店网站布局的任何变化都可能干扰数据收集并导致错误。最后，尽管问题分类是系统地进行的，但手动分类仍包含主观解读的成分，可能影响最终分类结果，尽管研究者已努力保持客观。

伦理声明

该数据集的编制和处理严格遵循伦理考量。主要数据来源亚马逊应用商店是一个公共领域，用户可以在其中获取应用信息。数据收集过程中未提取或使用任何私人或敏感用户数据。本研究仅关注可公开获取的数据，具体包括应用名称、类别和用户评论数量，这些数据不涉及任何个人或机密用户信息。

此外，本研究未涉及人类或动物受试者，也未与个人或群体进行互动。因此，本研究无需获得伦理审批。数据收集和分析均以维护最高诚信标准并尊重数据来源为宗旨。该数据集及其编制方法仅用于学术和研究目的，旨在帮助理解数字市场中的用户偏好和趋势，尤其是在低评分软件应用方面。该数据集的任何潜在用途都应遵守这些伦理标准，确保在后续研究或分析中负责任且恰当地使用数据。

CRediT 作者声明

Nek Dil Khan: 方法论、概念化、撰写——初稿、撰写——审阅与编辑。**Javed Ali Khan:** 方法论、撰写——初稿。 **Javed Ali Khan:** 方法论、撰写——初稿 **Darvesh Khan:** 软件、数据整理、调查、撰写——审阅与编辑。 **Mumrez Khan:**软件、验证、撰写——审阅与编辑。 **Shahfahad Khan:** 软件、验证、撰写——审阅与编辑。

利益冲突声明

作者声明，他们没有可能影响本研究结果的经济利益或个人关系。

数据可用性

数据标识号：*10.17632/zztwrn36n8.1*

数据的直接网址： *https://data.mendeley.com/datasets/zztwrn36n8/1*

GitHub 链接： *https://github.com/nekdil566/Dataset-of-User-Reviews-from-Low-Rated-Amazon-Appstore-Applications/blob/main/README.md*

参考文献

[1] N. D. Khan、J. A. Khan、J. Li、T. Ullah 和 Q. Zhao，“挖掘软件洞察：揭示低评分软件应用中频繁出现的问题”，*PeerJ Comput Sci*，第 10 卷，文章编号 e2115，2024 年 7 月，doi: 10.7717/PEERJ-CS.2115。[2] N. Khan、J. Khan、J. Li、T. Ullah、……A. A.-I.，以及未署名作者，2024 年，“社交媒体用户如何表达对软件应用的看法？一种细粒度分类方法”，*ieeexplore.ieee.orgND Khan, JA Khan, J Li, T Ullah, A Alwadain, A Yasin, Q ZhaoIEEE Access*，*2024•ieeexplore.ieee.org*，访问日期：2024 年 9 月 18 日。[在线]。可访问：hhps://ieeexplore.ieee.org/abstract/document/10591990/[3] N. Khan、J. Khan、J. Li、……T. U.-E. S. with，以及未署名作者，2025 年，“利用大语言模型 ChatGPT 增强对社交媒体反馈中终端用户情绪的理解”，*bdbanaly)cs.irND Khan, JA Khan, J Li, T Ullah, Q ZhaoExpert Systems with Applica)ons*，*2025•bdbanalytics.ir*，访问日期：2025 年 6 月 26 日。[在线]。可访问：
hhps://bdbanaly4cs.ir/wp-content/uploads/2025/02/Leveraging-Large-Language-Model-C
h
atGPT-for-enhanced-understanding-of-end-user-emo4ons-in-social-media-feedbacks.pdf
[4] E. Fa4ma、H. Kanwal、J. A. Khan 和 N. D. Khan，“使用微调深度学习分类器对应用商店中的讽刺检测与分类进行探索性自动化研究”，*Automated Software Engineering*，第 31 卷第 2 期，第 1–53 页，2024 年 11 月，doi：10.1007/S10515- 024-00468-3/METRICS。