HARPT：用于分析消费者对移动健康应用的信任和隐私担忧的语料库

Timoteo Kelly
University of Missouri
USA

Abdulkadir Korkmaz
University of Missouri
USA

Samuel Mallet
University of Missouri
USA

Connor Souders
University of Missouri
USA

Sadra Aliakbarpour
Rockbridge High School
USA

Praveen Rao
University of Missouri
USA

摘要

我们提出了 HARPT，一个大规模的移动健康应用商店评论注释语料库，旨在推进用户隐私和信任方面的研究。该数据集包含超过 480,000 条用户评论，并被标注为七个类别，涵盖应用程序中的信任、对提供商的信任以及隐私问题等关键方面。创建 HARPT 需要解决多个复杂问题，包括定义一个细致的标签模式，从大量嘈杂的数据中分离出相关的内容，设计一种既能保证可扩展性又能确保准确性的注释策略。该策略整合了基于规则的过滤、迭代手动标注与审阅、定向数据增强以及使用基于转换器的分类器进行弱监督以加速覆盖面。同时，一个精心挑选的 7,000 条评论子集被人工标注，以支持模型开发和评估。我们对广泛的一系列分类模型进行了基准测试，证明了强大的性能是可实现的，并为未来的研究提供了基线。HARPT 作为公共资源发布，旨在支持健康信息学、网络安全以及自然语言处理方面的工作。

Keywords

移动健康应用，用户评论，信任与隐私，弱监督，文本分类

ACM Reference Format:

Timoteo Kelly, Abdulkadir Korkmaz, Samuel Mallet, Connor Souders, Sadra Aliakbarpour, and Praveen Rao. 2025. HARPT：用于分析消费者对移动健康应用的信任和隐私担忧的语料库. In *Proceedings of The 34th ACM International Conference on Information and Knowledge Management (CIKM'25).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

CIKM'25, Seoul, KR

ACM ISBN 978-1-4503-XXXX-X/2025/11
https://doi.org/XXXXXXX.XXXXXXX

1 介绍

移动健康（mHealth）应用程序在医疗保健服务中扮演着日益重要的角色，使患者能够通过智能手机安排预约、查看医疗记录、参与远程医疗服务，并直接管理慢性疾病。随着这些平台收集和处理越来越多的敏感个人健康信息，隐私、信任和数据治理等问题已经成为患者采用及持续使用的关键因素 [25] [11] [12] 。尽管 mHealth 应用程序日益普及，但仍缺乏大规模且公开可用的数据集来捕捉用户对这些数字平台上隐私问题和信任相关因素的真实观点。

为了解决这一缺口，我们引入了 **H**ealth**A**pp**R**eviews for**P**rivacy 和 **T**rust (HARPT)，这是一个源自超过 480,000

条用户评论的大型标注数据集，涵盖了患者门户和远程医疗应用程序。HARPT 独特地专注于根据隐私问题和信任的多个维度对评论进行分类，包括对医疗服务提供者和应用本身的信任。该数据集是通过结合目标关键词过滤、多评分员手动标注、数据增强以及利用基于变换器模型的弱监督的多阶段过程开发的。

除了发布 HARPT 数据集外，我们还对其在监督文本分类中的实用性进行了基准测试。我们评估了多种传统机器学习分类器以及多个变压器模型，为未来隐私和信任检测、医疗应用用户体验及情感分析的研究提供了强有力的基线。通过发布带标签的数据集及其相应的基准结果，HARPT 旨在支持对保护隐私的医疗技术、信任建模及以患者为中心的数字健康创新的持续研究。

2 相关工作

移动健康应用和数字健康物联网系统中的隐私与安全问题已在先前的文献中进行了概念和技术上的探讨。一些研究提出了管理医疗物联网生态系统中隐私风险的框架 [6] [13] ，或使用模糊语言模型来量化对电子健康服务的信任 [28]。更广泛的系统评价综合了用户对于健康数据共享的态度 [3]，并且 Diallo 等人 [9] 审查了发展中国家移动应用中的隐私问题。尽管这些研究非常有价值，但它们主要采用了基于调查、混合方法或理论的方法，并未发布公开数据集。

一个较小但正在增长的研究领域试图直接从用户生成的内容（如应用评论）中提取隐私和信任信号。Mukherjee 等人 [26] 进行了一项大规模的经验分析，涉及超过两百万条移动应用评论，以识别安全和隐私问题。Nema 等人 [27] 考察了跨应用市场的用户对应用隐私的视角。Zhang 等人 [37] 应用主题建模技术自动从应用评论中提取隐私关注主题。Ebrahimi 和 Mahmoud[10] 开发了无监督摘要方法来识别评论中的隐私问题，而 Sorathiya 和 Ginde[32] 设计了混合模型来挖掘应用评论中的道德关切。Hatamian 等人 [16] 使用半监督挖掘技术分析了手机用户在应用商店中的隐私感知。最后，Sabrina 和 Weng[29] 探索了 mHealth 应用评论中的虚假信息分类方法，而 Saheb 等人 [30] 则考察了 COVID 接触追踪应用程序中的隐私方面。

尽管这些工作展示了从应用程序评论中提取隐私和信任信号的兴趣日益增长，但许多研究依赖于专有或有限的数据集、关注非健康领域，或者不公开发布注释数据集以支持可重复的研究。此外，现有的大部分研究集中在通用移动应用上，而移动健康应用程序由于涉及用户健康相关数据的敏感性，带来了独特的隐私挑战。重要的是，先前的研究通常只解决信任或隐私的孤立方面，没有共同考察隐私问题、对应用程序的信任以及对提供者的信任。HARPT 通过在一个统一的标注数据集中捕捉这些维度来填补这一空白。

我们的工作基于文献，引入了一个大规模的公开可用的数据集，包含 480,450 条移动健康应用评论，并对信任和隐私指标进行了弱标注。通过结合手动注释、数据增强和弱监督方法，我们创建了一种新的资源，直接支持未来在 mHealth 应用商店评论中关于隐私和信任的研究。

3 数据集构建

3.1 初始数据收集

我们首先从 Google Play 商店收集了一个初始数据集，包含约 457,165 条用户评论。这些评论是从覆盖患者门户和远程医疗服务的移动健康应用程序中抓取的。为了提高候选评论对人工标注的相关性，我们应用了基于关键词的过滤过程。我们为每个目标类别整理了一份代表性的关键词列表。至少包含其中一个关键词的评论被保留以供进一步标注。这一步骤减少了候选池的大小，同时确保了目标概念的充分代表性。

3.2 注释

从过滤后的池中，随机选择了 4,000 条评论进行手动标注。评论被标注为七个互斥类别之一：能力、可靠性、支持、风险、道德性、数据质量以及数据控制。注释过程遵循了一个三轮审查协议。首先，一个初始标注者标记了一批次的 1,000 条评论。接下来，两个独立审核

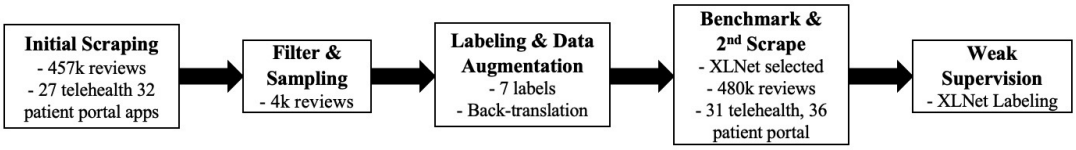

图 1: HARPT 构建管道概述。

表 1: HARPT 分类及其描述和代表性评论摘录。

Label	Description + Example
Data Control	Concerns about consent, third-party access, or control over personal data. "It's great the app lets me review and delete my medical records whenever I want."
Data Quality	Issues with incorrect, outdated, or incomplete health data. "The app is showing someone else's health record!"
Reliability	App crashes, login failures, or inconsistent functionality. "I can't login! Every time I enter my username and password it just loads and loads."
Support	User experience with help services or communication. "Customer support responded to my query within a few hours and helped me resolve my issue."
Risk	Worries about security, breaches, or misuse of data. "I'm worried about how the app stores my personal health records."
Competence	Perceptions of provider professionalism and care quality. "I received tailored advice based on my health data and it helped me manage my condition better."
Ethicality	Judgments of provider fairness, transparency, or responsibility. "I love that the app asks for my consent before collecting my data."

员分别对每批次进行审阅并确认或建议替代标签。最终标签由三个标注者的多数票决定。在所有三位标注者意见不一致的情况下（这种情况仅发生过一次），首席研究员裁定并分配了最终标签。这个多轮次过程确保了高质量的注释，并最大限度地减少了标签噪音。评判者间的一致性使用{**费列斯卡帕**[14]量化，得到值为{**0.877**，这表明标注者之间有几乎完全一致的意见。

3.3 隐私与信任的理论框架

该数据集的标注模式基于应用于医疗技术的隐私和信任理论。我们将标签组织在三个总体结构下：隐私问题、应用中的信任和对提供者的信任。

对于隐私问题，维度是由诸如互联网用户信息隐私关注（IUIPC）模型 [21] 等先前的隐私关注框架工作确定的，该模型将隐私关注概念化为一个包含数据收集、用户控制和隐私实践意识的二级结构。其他如普适计算中的感知控制 [33]，以及特定于医疗保健的信任模型 [8] [35] [34] 与移动健康应用中常见的问题相一致，在这些应用程序中，用户会与敏感的个人健康数据互动，对信息共享的控制有限，并且经常缺乏关于数据使用情况的透明度。这些隐私结构反映了与个人健康信息共享、数据控制、同意和信息质量相关的关注点，这与医疗保健和普适计算领域的先前框架一致。我们的标注方案通过数据控制和数据质量标签来实现这些关注点。

对于对提供商的信任，我们基于组织信任框架，特别是 Mayer 等人提出的整合模型 [22]、Hosmer 阐述的伦理视角 [18] 以及 McKnight 等人开发的信任量表 [24]。这些模型强调提供者的专业能力、道德性、完整性和善意作为信任形成的关键前提，这直接对应于我们的能力和伦理正当性标签。

对于应用中的信任，我们整合了技术信任文献 [23]、电子商务信任模型 [15] 和考虑信任及隐私顾虑的技术接受框架 [8] 的观点。在移动健康应用程序中，用户对技术系统的信任由感知的可靠性、安全性、技术支持以及风险脆弱性所塑造。这些因素体现在我们的模式中的可靠性、支持和风险类别中。

总体而言，这些理论基础确保我们的标注方案能够捕捉到用户对医疗应用程序隐私和信任维度的感知，其中用户同时与医疗服务提供者以及复杂的数字平台进行互动，而这些平台介导着高度敏感的个人数据。

3.4 标注模式和标签定义

每个评审都被分配到七个互斥类别之一，这些类别对应于信任和隐私构建（参见表 1）。

3.5 数据扩增与平衡

人工标注的数据集表现出类别不平衡，某些类别（例如能力）被过度表示。为了解决这种不平衡，我们通过法语、西班牙语和德语的反向翻译应用了数据增强。少数类通过增强进行了过采样，而多数类则下采样至 1000 个实例。最终平衡的训练集包含 7000 条评论。为了评估反向翻译文本的语义保真度，我们计算了原始评论和反向翻译评论之间的{**BLEU 分数**}。这得到了一个分数为{**30.43**}的结果，表明词汇重叠程度适中并且核心意义得以保留。

3.6 地面真相数据集上的模型评估

我们在包含 7,000 个实例的平衡数据集上训练和评估了一组机器学习和基于变压器的模型。传统模型包括支持向量机（SVM）[7]、随机森林 [2]、逻辑回归 [17]、XGBoost[4] 和 LightGBM[19]。基于变压器的模型包括 DistilBERT[31]、ELECTRA[5]、XLNet[36] 和 RoBERTa[20]。这些实验（见图 2）使我们能够评估不同模型架构的分类性能，并为弱监督下的模型选择提供了信息。

3.7 大规模数据集收集

在真实值评估之后，我们进行了第二次、规模更大的数据收集过程，在 67 个移动健康应用中刮取了 480,450 条用户评论。这个更大规模的数据集还包含了患者门户和远程医疗应用程序。根据模型评估结果，我们选择了 XLNet 作为弱监督下表现最佳的模型。训练好的 XLNet 模型被用于将每条评论分类到七个目标类别之一。由此产生的弱标记数据集构成了本文中描述的主要数据集发布。

表 2: 数据集汇总统计

统计量	值
Total Reviews	480,450
Time Range	2011–2025
Unique Apps	67
App Types	64.8% Patient Portal, 35.2% Telehealth
Top 3 Apps	MyChart (19.4%), FollowMyHealth (14.7%), healow (9.8%)
Mean Word Count	16.49
Median Word Count	10
Star Ratings	5 stars: 65.4%, 4 stars: 9.6%, 3 stars: 4.3%, 2 stars: 4.0%, 1 star: 16.7%
Sentiment Labels	competence: 44.3%, reliability: 27.5%, data control: 13.5%, data quality: 9.7%, support: 2.2%, risk: 1.4%, ethicality: 1.3%

4 数据集概述

4.1 描述性统计分析

最终的 HARPT 数据集包含在 2011 年至 2025 年间从 67 个移动健康应用中收集到的 480,450 条独特用户评论。移除重复项和用户标识符后，我们生成了用于基准测试的最终清洗数据集。数据集中的评论平均词数为 16.5 个（标准差=19.46），最长的评论包含 675 个单词。大多数评论包含 4 到 21 个单词，反映了典型应用商店反馈简短的特点。星级评分倾向于积极情绪，有 65.4%的评论获得 5 星评价，而 16.7%的评论获得 1 星评价。

这些评论涵盖了广泛的时间范围，数据收集自 2011 年 7 月到 2025 年 4 月，允许对随着时间推移的信任和隐私问题进行纵向分析。

4.2 情感标签分布

七个标注标签的分布如下：能力（212,950 条评论），可靠性（132,149 条），数据控制（64,962 条），数据质量（46,544 条），支持（10,745 条），风险（7,016 条）和伦理性（6,084 条）。最常见的类别反映了用户对提供商能力和应用程序可靠性的关注。

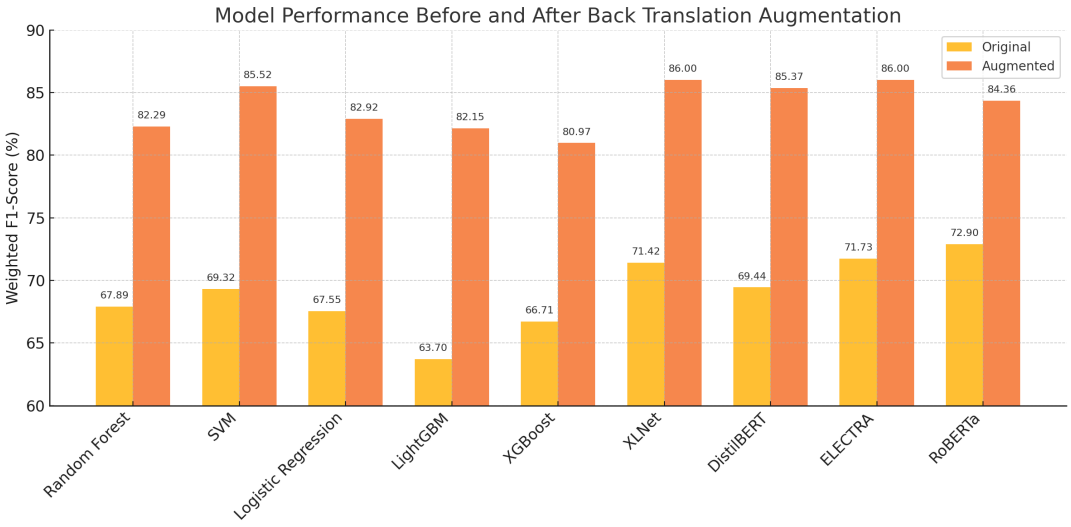

图 2: 模型性能（加权 F1 分数）在各种分类器上进行回译增强前后的对比。

4.3 应用

最常见的被评价的应用程序是MyChart（占评论的 19.4%）、FollowMyHealth（14.7%）、Healow（9.8%）、Practo（9.1%）和 Doctor on Demand（7.3%），这些应用程序共同占据了数据集的 60%以上。该数据集涵盖了患者门户应用和远程医疗平台。大约 65%的评论来自患者门户应用，而 35%来自远程医疗应用。

4.4 数据集发布

HARPT 数据集，包括 480,450 条弱标签评论和 7,000 条真实值子集，在 Dataverse 上公开可用（https://doi.org/10.7910/DVN/TU6OF6F）。在此数据上进行微调的 XLNet 模型可通过 Hugging Face 获得

两个资源均在创意 *commons 4.0* 下发布，以支持可重复的研究。

5 实验

为了在 HARPT 数据集上提供基线性能，我们训练和评估了经典机器学习模型以及最先进的基于变压器的模型。

对于经典模型，我们使用从文本评论中提取的 TF-IDF 向量化特征实现了一个随机森林分类器。超参数包括估计器的数量、最大深度以及每个分割的最小样本数，这些通过网格搜索进行了调整。

对于基于变压器的模型，我们在标注的训练数据上微调了 DistilBERT 和 RoBERTa。微调使用 GPU 进行混合精度训练，并利用提前停止和学习率调度器来优化收敛。所有超参数均通过 Weights and Biases 平台上的 [1] 功能进行了调整。

5.1 结果

HARPT 数据集包括七类：数据控制，数据质量，风险，支持，可靠性，能力和

表 3: 基准性能在 HARPT 数据集上的表现。

模型	准确率	F1	精度	回忆
Random Forest	94.00%	93.96%	94.05%	94.00%
DistilBERT	91.25%	91.27%	91.32%	91.25%
RoBERTa	89.02%	89.04%	89.13%	89.02%

这些结果表明，经典模型和基于变换器的模型都表现出强大的性能。

6 结论

在本文中，我们介绍了 HARPT，一个来自移动健康应用的新型用户评论数据集，重点关注信任和隐私问题。通过结合手动标注、弱监督学习和评论抓取，我们生成了一个包含 480,450 条评论的数据资源，并将其分为七个与隐私和信任相关的类别。我们也对多种分类模型进行了基准测试，为未来的研究提供了初步的基线。我们将手动标注的真实数据以及大规模的弱监督数据集公开提供，以促进移动应用中隐私和信任方面的进一步研究。

7 伦理声明

该数据集是从 Google Play 上的公开用户评论中收集的，不包含任何私人或受保护的健康信息。用户名在预处理阶段已被匿名化。注释工作由经过培训的标注员在多个评审阶段进行。本数据集仅用于研究目的，期望所有使用都尊重用户隐私和伦理研究标准。

8 致谢

第一位作者蒂莫特 · 凯利希望感谢美国国家科学基金会和美国人事管理局网络军团补助金第 #1946619 号对本研究的支持。第二位作者阿卜杜拉基尔 · 科尔库马兹希望感谢土耳其共和国的支持。

References

[1] Lukas Biewald. 2020. Experiment Tracking with Weights and Biases. https://www.wandb.com/.

[2] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.

[3] Fidelia et al. Cascini. 2024. Health data sharing attitudes towards primary and secondary use of data: a systematic review. *BMC Public Health* (2024).

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 785–794.

[5] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather than Generators. *International Conference on Learning Representations (ICLR)* (2020).

[6] Liane Colonna. 2019. Privacy, Risk, Anonymization and Data Sharing in the Internet of Health Things. *Computer Law & Security Review* (2019).

[7] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning* 20, 3 (1995), 273–297.

[8] Devendra Dhagarra, Madhurima Goswami, and Gaurav Kumar. 2020. Impact of Trust and Privacy Concerns on Technology Acceptance in Healthcare: An Integrated Model. *Technological Forecasting and Social Change* 161 (2020), 120332.

[9] Alseny et al. Diallo. 2024. (In)Security of Mobile Apps in Developing Countries: A Systematic Literature Review. *Computers & Security* (2024).

[10] Fahimeh Ebrahimi and Anas Mahmoud. 2022. Unsupervised Summarization of Privacy Concerns in Mobile Application Reviews. *2022 IEEE International Conference* (2022).

[11] Saber Elkefi and Sondes Layeb. 2023. Telemedicine' s future in the post-Covid-19 era, benefits, and challenges: A mixed-method cross-sectional study. *Behaviour & Information Technology* (2023).

[12] Pouyan Esmaeilzadeh. 2020. The impacts of the privacy policy on individual trust in health information exchanges (HIEs). *Internet Research* 30, 3 (2020), 811–843.

[13] Xavier Fadrique. 2019. Privacy and Trust in Healthcare IoT Data Sharing: A Snapshot of the Users' Perspectives. *University of Skövde Thesis* (2019).

[14] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378–382.

[15] David Gefen, Elena Karahanna, and Detmar W. Straub. 2003. Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly* 27, 1 (2003), 51–90.

[16] Majid et al. Hatamian. 2020. Revealing the unrevealed: Mining smartphone users privacy perception on app markets. *Information Processing & Management* (2020).

[17] David W. Hosmer and Stanley Lemeshow. 2000. Applied Logistic Regression. *Wiley Series in Probability and Statistics* (2000).

[18] LaRue Tone Hosmer. 1995. Trust: The Connecting Link Between Organizational Theory and Philosophical Ethics. *Academy of Management Review* 20, 2 (1995), 379–403.

[19] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*. 3146–3154.

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[21] Naresh K. Malhotra, Sung S. Kim, and Jagdish Agarwal. 2004. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research* 15, 4 (2004), 336–355.

[22] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *Academy of Management Review* 20, 3 (1995), 709–734.

[23] D. Harrison McKnight and Michelle Carter. 2011. Trust in Technology: Development of a Set of Constructs and Measures. *MIS Quarterly* 35, 2 (2011), 557–588.

[24] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* 13, 3 (2002), 334–359.

[25] Abdulaziz A. Mohammed and Zoltan Rozsa. 2024. Consumers' intentions to utilize smartphone diet applications: Privacy calculus model. *British Food Journal* 126, 6 (2024), 2416–2437.

[26] Debjyoti et al. Mukherjee. 2020. An Empirical Study on User Reviews Targeting Mobile Apps' Security & Privacy. *arXiv preprint arXiv:2010.06371* (2020).

[27] Preksha et al. Nema. 2021. Analyzing User Perspectives on Mobile App Privacy at Scale. *arXiv preprint arXiv:2104.10096* (2021).

[28] Pekka et al. Ruotsalainen. 2022. Privacy and Trust in eHealth: A Fuzzy Linguistic Solution. *International Journal of Medical Informatics* (2022).

[29] Sabrina and Jiefeng Weng. 2022. Improving Mobile Misinformation Identification through Advanced Information Retrieval and Data Mining. *Journal of Information Science* (2022).

[30] Tayeb et al. Saheb. 2022. Delineating privacy aspects of COVID tracing applications embedded with proximity measurement technologies. *Health Informatics Journal* (2022).

[31] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[32] Aakash Sorathiya and Gouri Ginde. 2023. Beyond Keywords: A Context-based Hybrid Approach to Mining Ethical Concern-related App Reviews. *arXiv preprint arXiv:2411.07398* (2023).

[33] Sarah Spiekermann and Jens Grossklags. 2001. Perceived Control: Scales for Privacy in Ubiquitous Computing. *International Journal of Human-Computer Studies* 59, 1-2 (2001), 193–220.

[34] Ehab Tamime, Hana Saleh, and Khalil Al-Khatib. 2023. Privacy Concerns in Mobile Health Applications: A Systematic Literature Review. In *2023 IEEE TrustCom*. 1–8.

[35] Lex van Velsen, Monique Tabak, and Hermie Hermens. 2017. Measuring patient trust in telemedicine services: Development of a survey instrument and its validation for an oncological context. *International Journal of Medical Informatics* 97 (2017), 52–58.

[36] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems (NeurIPS)* 32 (2019).

[37] Jianzhang et al. Zhang. 2022. Mining user privacy concern topics from app reviews. *Journal of Systems and Software* (2022).