arXiv:2302.04443v1 [cs.LG] 9 Feb 2023

利用预训练语言模型和微调提升电子商务推荐

Nuofan Xu
billxusuzhou@gmail.com

Chenhui Hu
chenhui.hu@gmail.com

摘要—预训练语言模型（PLM）在广泛的自然语言处理（NLP）任务中取得了巨大成功。然而，将其应用于推荐系统领域才刚刚起步。传统推荐算法未能充分利用电子商务数据集中的丰富文本信息，从而阻碍了这些模型性能的提升。我们全面研究了在一个电子商务数据集中，将 PLM 融入传统推荐算法的各种策略所产生的影响，并将结果与基础版推荐基线模型进行了比较。我们发现，应用 PLM 并针对特定领域进行微调，可以提升组合模型的预测能力。这些结果凸显了在电子商务场景中利用文本信息的重要性，并为如何更好地将 PLM 与传统推荐系统算法结合应用提供了参考。本文所用代码已发布于 Github：https://github.com/NuofanXu/bert retail recommender。

一、引言

随着网络技术的快速发展，互联网用户能够获取海量信息，这使用户难以筛选出相关信息，导致决策过程耗时更长、决策质量下降。为解决信息过载问题，各领域的研究人员付出了大量努力，开发能够根据用户偏好生成精选信息的推荐系统，应用范围涵盖在线广告 [Chaudhuri 等人， 2016],音乐选择[Kowald 等人， 2020], 电影评分 [Nguyen 和 Zhu， 2013],产品推荐 [Agarwal 等人， 2018]。通过利用用户与商品（UI）之间的关系，推荐系统有望简化数据处理并实现个性化，且一直是推动打造细致入微的客户体验的关键因素 [Ricci 等人，2015]。因此，在线零售商和网络服务提供商广泛采用推荐系统，以完善产品并获取利润。

电子商务在客户群体中的重要性和受欢迎程度日益提升。尽管推荐系统如今已成为许多行业不可或缺的一部分，但电子商务领域仍面临一些独特挑战，阻碍了推荐系统的广泛应用。在大多数情况下，潜在客户并不清楚自己想买什么。由于内部搜索引擎提供的属性分类体系有限，他们很难准确表达自己的需求。

这项工作于2021年完成。

[Agarwal 等人， 2018] 因此，他们在浏览大量选项时很容易感到气馁。毫无疑问，在线零售商所采用的推荐系统能否捕捉商品内在的物品特征和用户偏好，已成为其区别于竞争对手的关键因素之一，对确保公司的成功至关重要 [Ricci 等人，2015]。

推荐算法大致分为两类。基于内容的算法会分别为用户和商品建立专属档案，通常以向量的形式表示。用户档案描述用户喜欢的内容，商品档案则描述商品包含的内容。创建这些档案需要大量关于用户和商品的辅助信息。随后，通常使用点积这类相似度度量来确定用户偏好。协同过滤方法会分析过往的用户行为，并根据历史记录找出相似的用户和商品。其核心假设是，过去意见一致的用户未来也往往会达成一致。人们还使用深度学习（DL）技术进一步提升这些推荐模型的性能，因为深度学习具有更强的建模能力、更高的灵活性以及特征表示学习能力。Covington 等人 [Covington 等人， 2016]提出了一种基于深度神经网络的推荐系统，用于在 YouTube 上推荐视频。Cheng 等人 [Cheng 等人， 2016] 提出了一种适用于稀疏输入的通用推荐系统宽深模型。Shumpei 等人 [Okura， 2017] 展示了一种基于循环神经网络（RNN）的雅虎新闻推荐系统。事实证明，所有这些基于深度学习的模型都显著优于传统模型。混合模型是指整合两种或更多推荐方法的推荐系统。

尽管现有的大多数推荐算法在建模用户-物品关系方面表现强大，但在利用文本信息方面存在重大局限。具体而言，这些算法无法有效整合电子商务数据集中的文本信息，而这类信息十分丰富，且可能包含大量有价值的内容。未经准备，这些推荐算法也会因在线零售场景中缺乏负反馈而面临严重的相关性问题 [Jiang 等人， 2020]。关于应采用哪些推荐算法以及如何有效使用这些算法的问题，尽管一直备受争论和研究，仍然充满挑战，尚未得到解决。

近年来，BERT（来自变换器的双向编码器表示）及其变体在自然语言处理领域的出现与发展，大幅提升了机器学习模型理解文本信息的能力 [Devlin 等人， 2019]。借助在大规模文本数据集上预训练上下文表示的优势，以及无监督训练目标，BERT 已能在广泛的自然语言处理任务中取得最先进的性能。

然而，在电子商务领域，由于以下原因，人们无法充分发挥 BERT 算法的优势：BERT 并不总能准确理解频繁出现的领域专用短语，因为这些短语未被纳入训练目标；在电子商务语境中，与电影评论或网络文档等其他数据源相比，标签或商品描述等文本信息通常长度非常有限，而对于这类文本，若没有额外的领域专用预训练，BERT 很难进行推理，并且非常容易受到数据集噪声的影响 [Zhang 等人， 2020]。

统一语言表示模型的优势如何帮助推荐模型更好地理解和捕捉影响人们决策的内在规律，以及不同的进一步预训练方案将如何影响最终推荐结果，目前仍不清楚。在本文中，我们针对一种性能优异的预训练语言模型 RoBERTa [Liu 等人， 2019]，探讨了这一问题，并研究了在在线零售数据集上将丰富的文本特征融入传统推荐系统算法的各种方法的有效性。我们考虑两类推荐算法——基于内容的推荐算法（矩阵分解，简称 MF）和基于机器学习的算法（XGBoost 和决策树）。我们评估了引入句子嵌入相较于基线算法所带来的性能提升，并证明无论采用哪种推荐算法，使用 RoBERTa 都能持续提升模型性能。此外，我们还研究了使用基于变换器的去噪自编码器（TSDAE）和掩码语言模型（MLM）对 BERT 进行预训练的效果。对于 MLM，按照 RoBERTa 论文中的做法，我们采用动态掩码而非静态掩码[Liu 等人， 2019]。受这一过程启发，我们进一步分别针对 MF 和 XGBoost 提出了一种我们认为最有效的推荐方案。实验表明，即使用户-商品交互矩阵高度稀疏，在该方案下训练的推荐模型仍优于其基础版对应模型。

总而言之，我们的贡献包括：

- 一项关于融入丰富文本信息的全面研究在电子商务领域中利用这些信息提升推荐算法的性能；
- 对基础版推荐算法和 BERT 增强型推荐模型生成的推荐结果进行全面评估和可视化；
- 研究领域自适应预训练的有效性。

II. 相关工作

1) 矩阵分解： 一种直接的矩阵分解模型会将用户和商品映射到一个维度为 K 的联合潜在因子空间中，使得用户-商品交互在该空间中被建模为用户特征向量与商品特征向量的内积 [Nguyen 和 Zhu， 2013]。然而，纯矩阵分解依赖过去的信息，无法准确捕捉评分值中某些观测变化是由用户或商品相关偏差导致的情况。Xxx 通过采用包含额外偏置项的新训练目标和误差项来解决这一问题 [Tran 和 Lee， 2018]。然而，由于无法处理新商品和新用户，冷启动问题仍然具有挑战性 [Elahi 等人， 2016]。Wang 等人整合了额外的用户属性（人口统计信息）以缓解这一问题 [Wang 等人， 2012],尽管如此，矩阵分解模型的性能仍受到用户-商品交互矩阵高稀疏度的影响。在许多现实应用场景中，我们通常只能获取隐式反馈，即使用点击、点赞或购买等用户-商品交互作为代理，来表示用户对这些商品的偏好或厌恶。基于标准矩阵分解的协同过滤将用户-商品矩阵中的条目视为用户对商品的显式偏好。这些算法假设数据是显式的，即用户使用某种评分尺度对喜欢和不喜欢的事物都进行了评分。它们通常将缺失数据视为未知值，将其设为某个常数（通常为 0），然后使用随机梯度下降训练分解后的矩阵，使其重构整个矩阵，包括未知值。Hu 等人提出了隐式反馈 [Hu 等人， 2008]。实质上，隐式矩阵分解并不直接对评分矩阵建模，而是将数据视为表示用户行为观测强度的数值（例如点击次数，或某人观看电影所花费的累计时长）。这些数值随后用于表示对所观测到的用户偏好的置信程度，而不是用户对商品给出的显式评分。然后，模型尝试寻找可用于预测用户对某件商品预期偏好的潜在因子。

2）预训练语言模型：近期，大量研究表明，预训练语言模型（PLM）能够学习通用语言表示 [Peters 等人， 2018],[Radford 等人， 2018], [Devlin 等人， 2019]。在大型语料库上学习到的这些语言表示无需从头开始训练新模型，并能更快、更好地实现自然语言处理任务。与传统词嵌入模型相比 [Gupta 和 Manning， 2015],预训练语言模型根据完整的输入上下文学习预测和表示词语，因此能够更准确地理解语义。尽管

预训练语言模型已在广泛的自然语言处理任务中取得成功，但仍有许多方面有待改进。人们已付出诸多努力，进一步优化其稳健性和模型性能。SpanBERT [Joshi 等人，2020]提出了随机掩码跨度重建，而非单词重建。然而，这种跨度由随机连续词语组成，可能无法构成短语，因此无法准确捕捉短语级知识。ERNIE-1。0 [Sun 等人，2019]将短语级掩码和实体级掩码融入 BERT。RoBERTa [Liu 等人，2019]提出了对 BERT 预训练流程的改进，从而提升了下游任务性能，使预训练语言模型能够进一步释放其潜力。事实证明，如果在大型语料库上进行训练，预训练语言模型在句子补全任务上表现出色（在 Winograd 挑战赛句子上也取得了一定成绩）。

3）*XGBoost*：在众多强大的机器学习算法中，梯度树提升已被证明是在许多标准分类基准测试中表现出色的方法之一。XGBoost 是一种快速且高度可扩展的树提升实现。它采用新颖的树学习算法处理稀疏输入，并通过近似树学习处理实例权重。随着推荐系统被长期使用，人们已经开发出多种推荐算法。文献表明，扩展基于模型的算法 [Martin 和 Englot，2017]是最有前景的算法类别。在这一类别中，人们会调整多种机器学习技术，使其能够处理用户与商品之间的交互以及交互数据、用户属性和项目属性。在早期研究中被证明非常成功的技术之一是决策树学习。在这种方法中，利用已有观测数据，通过决策树预测商品的目标值。采用梯度提升技术，模型会依次加入，以纠正先前模型的错误，直到无法取得进一步改进。

4) 微调预训练语言模型：微调预训练语言模型（PLM）已成为自然语言处理中进行迁移学习的标准方法。在过去三年中，微调 [Howard 和 Ruder，2018]已取代了通过特征提取使用预训练嵌入的做法 [Peters 等人，2018]与此同时，由于预训练语言模型具有更高的样本效率和性能，因此相比于在机器翻译任务上训练的模型，它们更受青睐 [McCann 等人，2018], 自然语言推理[Conneau 等人，2017], 以及其他任务 [Zhang 和 Bowman，2019]。这些方法在实证上的成功促使人们开发出更大型的模型 [Devlin 等人，2019]，[Raffel 等人，2020]。事实上，近期的模型规模如此之大，以至于即使不进行任何参数更新，也能取得合理的性能 [Brown 等人，2020]。然而，这种零样本设置存在局限性，因此，要取得最佳性能或保持较高效率，在实践中使用大型预训练语言模型时，微调很可能仍将是惯常做法。

III. 数据集与数据准备

本文使用了三种不同的文本数据集。我们以 Online Retail 数据集 [Chen 等人，2015] 作为主要研究对象。该数据集涉及一个范围狭窄的电子商务领域，因为在收集该数据集期间，这家英国在线零售商销售的商品种类有限。我们使用亚马逊数据集作为一个相似但涵盖商品范围更广的数据集，因为二者都属于电子商务数据集。openwebtext 数据集 [Aaron 和 Vanya，2019] 被视为 RoBERTa 训练所用文本数据集的一个近似参照，仅在计算词汇相似度时用于交叉参照。由于其规模远大于另外两个数据集，因此只对其中一部分进行抽样。

该在线零售数据集包含一家在英国注册、无实体店的在线零售企业于2010年12月1日至2011年12月9日期间产生的交易记录。该公司主要销售适用于各种场合的独特礼品，其许多客户都是批发商。每条观测记录包含八个协变量，分别为发票编号、股票代码、商品描述、数量、日期、单价、客户编号和国家。发票编号为一个6位数，由零售企业唯一分配，以便识别不同商品是否属于同一用户。已取消交易以字母“C”开头的发票编号标识。股票代码是一个5位整数，由系统为每种不同的产品唯一分配；客户编号则是一个专属于用户的5位整数。商品描述是对商品内容的简短描述。日期为每笔交易生成时的日期和时间。最后一个字段是国家，显示每位客户居住所在国家的名称。该数据集共涵盖37个国家，范围从北美到欧盟，其中大多数客户来自英国和法国。图1展示了该数据集的前五行。

数据集中有一些值得一提的特殊条目和有趣特征：

- 负数量交易– 这些负值的绝对值与数据集中此前的购买记录相符。– 在在线零售场景中，这意味着有些商品被退回了。

- 零价或负价交易。– 零价交易对分析没有提供有用信息，而且这类交易大多没有任何描述。– 负价交易。从描述中可以看出，这些交易可能是作为更正而进行的。

- 具有极端离群值的长尾数据分布
  – 数量和单位价格的右偏。如图2所示，即使采用对数尺度，数据仍然呈右偏。事实上，50%的用户每次只购买1–3件商品，而一些批发商会一次大量购买同一种商品，数量超过500件。此外，数据集中售出的大多数商品价格都很便宜，

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

图1：Online Retail 数据集的一个片段

高单位价格的情况则很少见。因此，这些数据呈现出极强的右偏。

–离群值。从图3可以看出，大多数商品都是常见的日用品，价格也很正常。然而，该数据集中有些商品价格非常昂贵，这可能会显著影响购买金额。此外，数量的标准差为218.08，数值极大。这是由一些罕见且极端的事件造成的。例如，在记录该数据集期间，曾发生一笔购买数量达80955件的大额订单，但遗憾的是，该订单最终被取消，因此未纳入我们的数据。

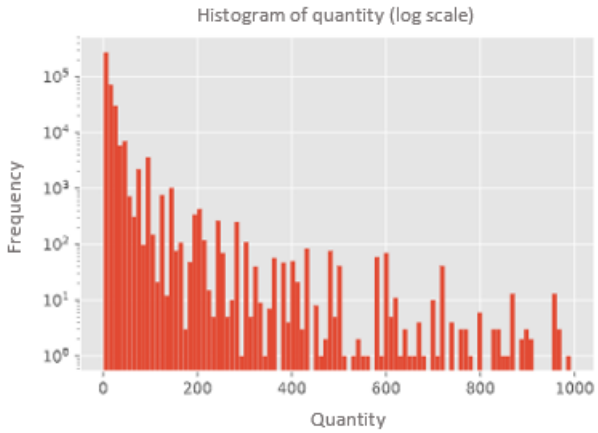


图2：在线零售数据集中商品购买数量的直方图

在未取消的交易中，有1336笔交易的数量为负数；另有1336笔交易的价格为零且数量为负数。这分别占数据集中观测条目总数的1.71%和0.46%，比例并不显著，直接筛除这些问题条目不太可能造成信息损失。该在线零售数据集共有541909个条目，其中有406829个非空客户ID和540455条商品描述，这意味着24.9%的客户 ID和0.27%的商品描述为空，因而无法使用。筛除所有空值条目和退货情况后，共有397924条观测值。这些记录对应4339个不同用户和3665种唯一商品，由此构成的用户-商品矩阵稀疏度为97.5%。

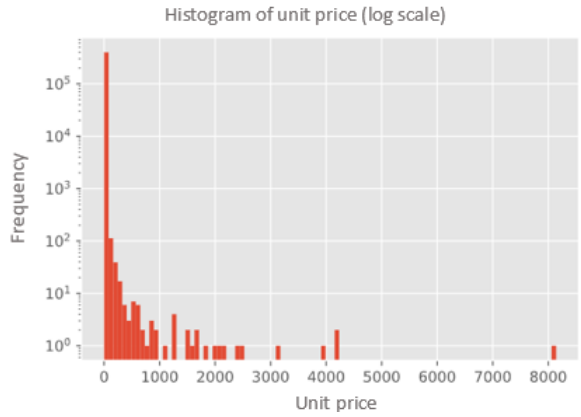


图3：在线零售数据集中商品单价的直方图

为了降低离群值和分布不均对模型性能的影响，通常有两种方法。我们可以通过减去均值并将结果除以标准差来标准化特征，但这会导致特征中出现负值，并可能使矩阵分解产生负输出值。在本文中，我们选择第二种方法，通过创建如图4所示的直方图来均衡分布，并为每个数量范围分配标签，使每个范围大致包含相同数量的产品。由于这种方法具有离散性，每个标签所对应的商品数量不会完全相同，但如图4所示，其分布明显优于图2所示的右偏分布。

IV. 基线模型

我们首先构建不使用商品描述文本信息的基线模型。本文使用了两种主要的推荐模型：

- 矩阵分解推荐器
- XGBoost 分解推荐器

对于每种方法，我们都调查了这些推荐模型三种变体的性能。详情见表I。

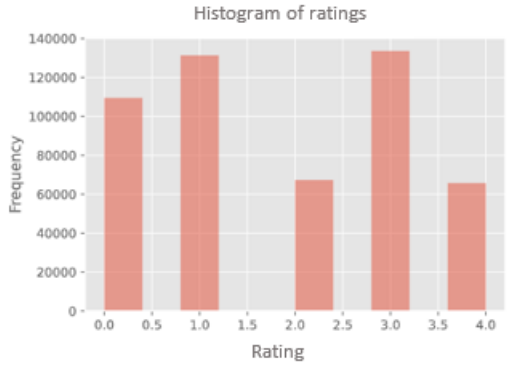


图4：在线零售数据集中数量到评分的转换后评分的直方图

表I：推荐模型及其变体

变体	**推荐器1** 矩阵分解	**推荐器2** XGBoost
1	基础版	基础版
2	广义低秩模型	与随机森林结合
3	隐式	使用嵌入特征

A. 矩阵分解

基于标准矩阵分解的协同过滤将用户-商品矩阵中的条目视为用户对商品给出的显式偏好。这些算法假设使用的是显式数据，即用户按照某种评分尺度对喜欢和不喜欢的事物都进行了评分。它们通常将缺失数据视为未知值，将其设为一个常数（通常为 0），然后使用随机梯度下降（SGD）训练分解后的矩阵，使其重现整个矩阵，包括未知值 [Nguyen 和 Zhu， 2013]。然而，在许多真实世界的应用场景中，我们通常只能获取隐式反馈（例如浏览、点击、购买、点赞和分享）。从本质上说，隐式矩阵分解并非直接尝试对评分矩阵建模，而是将数据视为表示用户行为观测强度的数值（例如点击次数，或用户观看一部电影所花费的累计时长）。这些数值反映的是对观测到的用户偏好的置信度，而非用户对商品给出的显式评分。随后，模型会尝试找出可用于预测用户对某件商品的预期偏好的潜在因子 [Hu 等人， 2008]。

以本文使用的数据集为例，我们可以假设用户购买某件商品就意味着他们喜欢它，但我们没有相应的信号表明用户不喜欢某件商品。隐式数据通常比显式数据更丰富，也更容易收集——即使让用户给出五星评分，这些评分绝大多数也都是正面的，因此无论如何都需要考虑隐式行为。这意味着我们不能简单地将缺失数据视为未知，而必须把用户没有购买某件商品视为一个信号，表明用户可能不喜欢该商品。这给学习因子分解表示带来了几个挑战。

表示。

第一个挑战是如何高效地进行这种分解：如果将未知项视为负例，朴素的实现方式会遍历输入矩阵中的每个条目。这里的维度约为 4K × 3K，也就是说，总共有超过 1,200 万个条目需要考察，而非零条目只有 4 万个。第二个问题是，我们无法确定用户没有听某位艺术家的作品，就意味着他们不喜欢这位艺术家。用户不听某位艺术家的作品可能有其他原因，尤其是考虑到数据集中每位用户仅有播放次数最多的 50 位艺术家。为了处理负向数据可信度不高的情况，这种方法会针对二元偏好采用不同的置信度，学习矩阵分解表示：未出现的商品被视为负例，置信度较低；出现的商品则被视为正例，置信度高得多。

B. XGBoost 和随机森林

XGBoost 每次构建一棵决策树。每棵新树都会纠正此前训练的决策树所产生的错误 [Chen 和 Guestrin, 2016]。由于提升树是通过优化目标函数得出的，原则上，XGBoost 可用于求解几乎所有能够写出梯度的目标函数。这包括排序和泊松回归等任务，而随机森林（RF）较难实现这些任务。不过，XGBoost 模型的建模能力很强，因此当数据含有噪声时更容易过拟合。由于树是按顺序构建的，训练通常比 RF 耗时更长。此外，基于 XGBoost 的模型通常也比 RF 更难调优。几个重要的超参数包括树的数量、树的深度和学习率，并且构建的每棵树通常都较浅。

随机森林（RF）独立训练每棵树，并使用数据的随机样本 [Louppe， 2014]。这种随机性有助于提高模型的稳健性，使其优于单棵决策树。因此，随机森林不容易在训练数据上过拟合。随机森林不相似度有助于在具有强相关特征的高维问题中实现较高的预测准确率，因此已被应用于各种场景。它还有助于降低模型对高噪声数据以及包含大量缺失值的数据的敏感性，例如某些属性是分类变量或半连续变量的情况。与 XGBoost 相比，随机森林的模型调优要容易得多。随机森林有两个主要参数：每个节点要选取的特征数量和决策树数量。随机森林算法的主要局限在于，树的数量过多会导致算法无法快速进行实时预测。对于包含不同层级数分类变量的数据，随机森林会偏向层级数较多的属性。

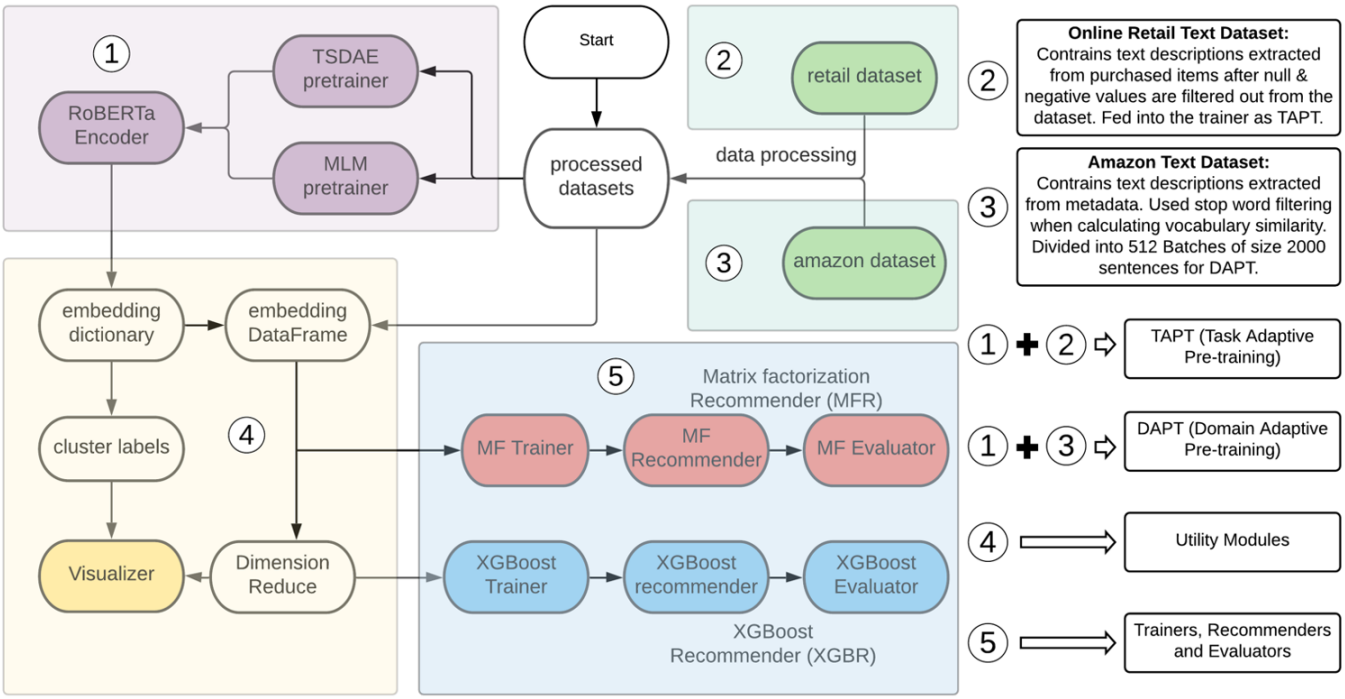


图 5：数据处理和训练流程的流程图

XGBoost 通常用于训练梯度提升决策树和其他梯度提升模型。随机森林使用与梯度提升决策树相同的模型表示和推理方式，但采用不同的训练算法。可以使用 XGBoost 训练独立的随机森林，也可以将随机森林用作梯度提升的基模型。

本文选择 XGBoost 作为推荐算法，对从在线零售数据集中提取的所有特征进行建模，并为每位用户提供推荐结果，再将其与真实购买记录进行评估。随机森林算法用作特征选择器，旨在尽量降低不太有用的特征和底层噪声的影响，避免 XGBoost 模型过拟合。

V. RoBERTa 微调

A. 语料相似度

本文使用两个文本数据集对 RoBERTa 进行微调。

- 在线零售数据集中的商品描述
- 亚马逊元数据集中的商品描述

图 6 展示了这两个文本数据集之间的词语相似度。图 7 和图 8 分别展示了应用停用词过滤后，从在线零售数据集和亚马逊描述数据集中找到的出现频率最高的 10 个常见词。请注意，统计词频时，零售数据集和亚马逊数据集都进行了停用词过滤。相比之下，亚马逊数据集中的描述字符长度要长得多，且包含的关键信息少于

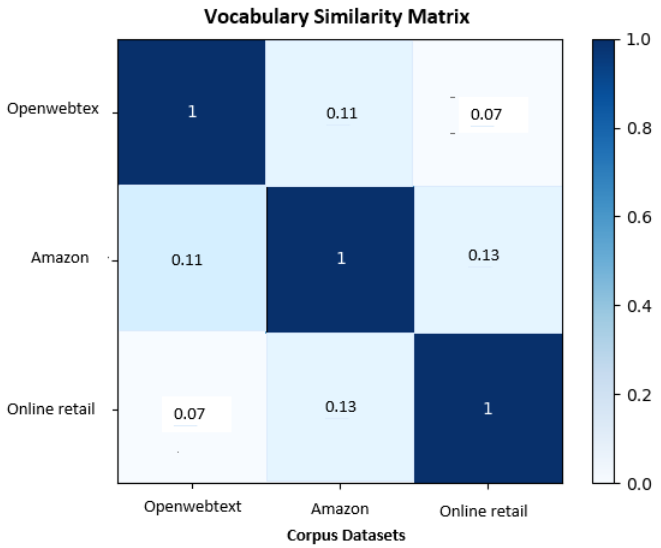


图 6：各语言语料库之间的词汇重叠率（%）。针对每个领域，从该领域抽取的商品描述中选取出现频率最高的 150 个词，以此构建各领域的词汇表。由于用于训练 RoBERTa 的数据集已不再可用，因此选取一份 openwebtext 数据集样本作为数据源，因为它与 RoBERTa 的预训练语料库相似。

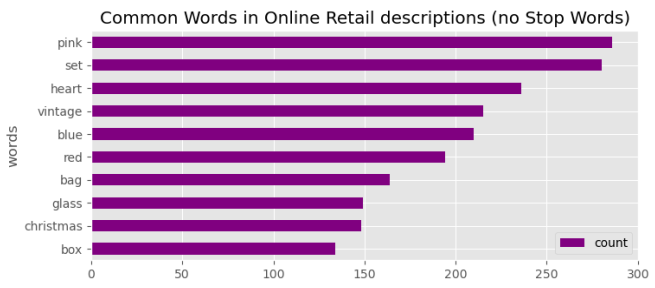


图7：经过停用词过滤后，零售数据集描述中出现频率最高的10个常见词。

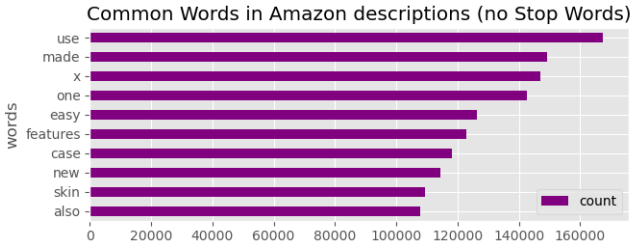


图8：经过停用词过滤后，亚马逊数据集描述中出现频率最高的10个常见词。

零售数据集中的描述。这可能会在训练期间导致图形处理器出现内存问题，并给预训练结果带来不必要的噪声。为节省内存，在预训练过程中，从亚马逊数据集中提取的句子被分为 512 个批次，每批次包含 2000 条，并将字符上限设为 50，以尽量匹配零售数据集中描述的字符长度，并过滤掉这些描述中不太有用的信息。

B. 领域自适应预训练与任务自适应预训练相结合

针对特定目标任务整理的数据集通常只是更广泛领域中文本信息的一个子集。鉴于在线零售数据集的规模相对较小，我们在此提出假设：在线零售数据集是由亚马逊数据集所代表的更广泛电子商务领域的一个定义较为狭窄的子集。我们采用了与 Gururangan 等人所使用的类似预训练方案。 [Gururangan 等人， 2020] 该方案包含三个不同阶段。我们首先进行常规的 RoBERTa 预训练，然后分别使用在线零售数据集和亚马逊评论数据集进行任务自适应预训练（TAPT）和领域自适应预训练（DAPT）。正如 Gururangan 等人所建议的， [Gururangan 等人， 2020]，DAPT 和 TAPT 可能会相互补充，这取决于是否围绕不同粒度层级的任务定义了一个领域谱系。我们研究了同时使用这两种自适应技术的效果，并观察了调整不同预训练阶段顺序所产生的影响。我们从 RoBERTa 开始，然后采用 DAPT 或 TAPT 中的一种作为基线。随后，我们按照表 V 所示，探索两种不同的预训练阶段顺序。我们表明，与基线相比，将领域自适应预训练与任务自适应预训练相结合，能够在所有任务上实现更好的性能

。结果还表明，在 DAPT 之后应用 TAPT 可以获得更高的推荐准确率。

六、实验结果

本文使用了两种主要的推荐模型：矩阵分解和 XGBoost。我们采用多种变体方法，探索提升推荐性能的关键。详细列表见表I。

我们首先使用基础版矩阵分解和 XGBoost 模型建立基线性能。在此基础上，我们进行额外调整，以观察性能是否有所提升，并更好地了解如何将预训练语言模型应用于这些推荐系统。为了直观展示并评估预训练语言模型，我们将文本数据集中的句子输入训练好的模型，生成维度固定为 728 的句子嵌入。然后，将这些高维句子嵌入输入降维网络，根据设置将其维度降至 10 维或 2 维。可选的降维方法有两种：t-SNE 和主成分分析。主成分分析的结果更稳定，运行速度也更快 [Garber 和 Hazan， 2015]，而 t-SNE 的聚类效果更好，但运行速度较慢。t-SNE 在处理高维向量时也表现不佳。本文中的嵌入空间为二维，通过主成分分析进行可视化，相关结果也使用主成分分析生成。我们也测试了 t-SNE，并按照 [Van Der Maaten 和 Hinton， 2008]，先使用主成分分析将维度降至 10 维，再使用 t-SNE 进一步降至 2 维。然而，t-SNE 并未带来更好的结果，且对其性能的分析与本文研究无关。

我们采用蒸馏版 RoBERTa 作为基线预训练语言模型，并可选择按照 Gururangan 等人提出的两种预训练方案进行微调。[Gururangan 等人， 2020],领域自适应预训练（DAPT）和任务自适应预训练（TAPT）。我们的领域专用预训练遵循 RoBERTa 训练所规定的设置。由于内存限制，我们没有选择在 openwebtext 上对 RoBERTa 模型进行进一步预训练。我们改用 RoBERTa 作为基线模型，分别使用亚马逊描述数据集和零售描述数据集，研究 DAPT 和 TAPT 的效果。我们采用了两种不同的预训练方法：基于 Transformer 的序列去噪自编码器（TSDAE）和掩码语言模型（MLM）。研究表明，TSDAE 是一种强大的预训练方法，能够学习句子嵌入，其表现显著优于 MLM 等其他方法。对于需要领域知识的异构领域，与基线 RoBERTa 模型相比，它还具有更好的泛化能力和鲁棒性。我们按不同顺序进行预训练，以比较不同方案的效果。

基于 XGBoost 的算法评估结果				
	准确率	精确率	召回率	F1 分数
基线	71.2 (0.2)	70.1 (0.3)	70.8 (0.4)	70.4 (0.2)
@ 批次 100	72.3 (0.3)	71.4 (0.4)	69.9 (0.9)	70.6 (0.3)
@ 批次 200	72.4 (0.2)	72.2 (0.3)	71.0 (0.6)	71.6 (0.3)
@ 批次 300	73.0 (0.4)	72.6 (0.6)	72.2 (0.4)	72.4 (0.3)
@ 批次 400	73.3 (0.2)	72.8 (0.2)	72.5 (0.3)	72.6 (0.2)
@ 批次 500	73.1 (0.3)	73.0 (0.2)	**72.8** (0.4)	72.9 (0.3)
@ 最终	**73.6** (0.6)	**73.4** (0.2)	72.6 (0.4)	**73.0** (0.3)

表 II：基线模型（RoBERTa）在 5 次试验运行中计算得出的预测准确率、精确率、召回率和 F1 分数及其标准差，以及使用亚马逊数据集通过 TSDAE 进一步预训练的模型的上述指标。上述结果是在同一亚马逊数据集上使用 XGBoost 评估的，但经过 512 个批次训练后的最终模型是在零售数据集上评估的。

基于 XGBoost 的算法评估结果				
	准确率	精确率	召回率	F1 分数
基线	67.4 (0.3)	68.2 (0.2)	67.0 (0.3)	64.8 (0.3)
@ 批次 100	72.2 (0.3)	71.6 (0.2)	69 (0.2)	70.3 (0.3)
@ 批次 200	72.4 (0.2)	72.5 (0.3)	71.0 (0.1)	71.7 (0.4)
@ 批次 300	73.0 (0.2)	72.8 (0.2)	72.2 (0.3)	72.5 (0.1)
@ 批次 400	73.3 (0.3)	73.0 (0.3)	72.5 (0.5)	72.7 (0.4)
@ 批次 500	74.0 (0.1)	73.6 (0.3)	**73.2** (0.2)	73.4 (0.2)
@ 最终	**74.2** (0.2)	**74.5** (0.4)	72.8 (0.3)	**73.6** (0.2)

表 III：基线模型（RoBERTa）的预测准确率、精确率、召回率和 F1 分数及标准差，基于 5 次试验运行计算得出；另一个模型则使用亚马逊数据集通过 MLM 进行进一步预训练。上述结果使用 XGBoost 在同一亚马逊数据集上进行评估，但经过 512 个批次训练后的最终模型是在零售数据集上评估的。

首先，我们分别研究 DAPT 和 TAPT 的有效性。亚马逊句子被划分为 512 个批次，每个批次包含 2000 条，且每条句子允许的最大字符数设为 50。随后，这些批次分别输入两个不同的预训练框架 TSDAE 和 MLM。采用预测准确率、精确率、召回率和 F1 分数作为评估指标，并每训练 100 个批次及训练结束时评估模型性能。结果列于表 II 和表 III。随着训练进行，XGB 推荐系统在所有评估指标上的性能均有所提升，而训练速度随批次数增加而减慢。图 9 所示的训练损失图也支持这一结论。训练损失在最初 100 个批次中快速下降，之后趋于稳定。

接下来，我们探讨 TSDAE 与掩码语言模型预训练的不同组合顺序，以及领域自适应预训练和任务自适应预训练的不同组合，并与前文确立的基线性能进行比较。结果列于表 IV。

我们还研究了 XGBoost 模型生成推荐结果时所用的各类特征的影响。随机森林算法作为一种可选的特征选择方法，用于筛除重要性低于

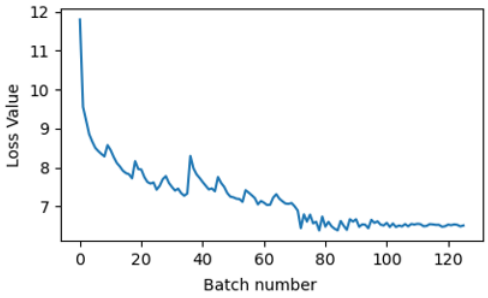


图 9：在亚马逊数据集上评估的掩码语言模型预训练图中，损失值与批次编号的关系

基于 XGBoost 的算法评估结果				
	准确率	精确率	召回率	F1 分数
TSDAE @亚马逊	71.2 (0.2)	70.1 (0.3)	70.8 (0.4)	70.4 (0.2)
掩码语言模型 @亚马逊	72.3 (0.6)	71.1 (0.4)	69.8 (0.4)	70.4 (0.2)
T→M @亚马逊	72.4 (0.3)	72.2 (0.2)	71.4 (0.3)	71.8 (0.3)
M→T @亚马逊	73.0 (0.1)	72.6 (0.3)	72.2 (0.2)	72.4 (0.4)
T→M @零售	73.3 (0.2)	72.8 (0.4)	72.5 (0.6)	72.6 (0.1)
M→T @零售	73.1 (0.4)	73.0 (0.2)	72.8 (0.4)	72.9 (0.3)
掩码语言模型 @D→T	73.6 (0.2)	73.4 (0.4)	72.6 (0.2)	73.0 (0.2)
TSDAE @D→T	73.5 (0.3)	73.2 (0.2)	72.6 (0.1)	72.9 (0.1)
T→M @D→T	74.4 (0.2)	74.3 (0.2)	**73.1** (0.2)	73.7 (0.2)
M→T @D→T	**74.8** (0.2)	**74.5** (0.3)	73.0 (0.4)	**73.7** (0.5)

表 IV：预测准确率、精确率、召回率和 F1 分数及其标准差，均根据 5 次试验运行计算得出。T→M 表示 TSDAE→MLM，D→T 表示 DAPT→TAPT。

0.04 的特征，以尝试降低 XGBoost 模型在不太有用的特征上过拟合的概率。除特征名称后带有 * 的特征外，其他特征均以并行方式添加。详情见表 V。

基于 XGB 的算法的评估结果				
	准确率	精确率	召回率	F1 分数
原始 XGBoost	54.2 (0.3)	52.3 (0.2)	49.5 (0.2)	67 (0.1)
+随机森林	56.1 (0.4)	55.4 (0.2)	51.9 (0.3)	56.4 (0.6)
+簇标签	60.2 (0.3)	61.2 (0.3)	55.1 (0.1)	58.7 (0.3)
+10维嵌入	73.0 (0.2)	72.8 (0.5)	72 (0.3)	73 (0.2)
+2维嵌入	73.3 (0.3)	73.0 (0.4)	72.5 (0.2)	73 (0.5)
最佳任务自适应预训练	74.0 (1.2)	73.6 (0.8)	73.2 (0.9)	73.6 (0.6)
+股票代码	74.2 (0.3)	74.5 (0.4)	72.8 (0.5)	73.9 (0.2)
+发票编号	74.2 (0.6)	74.5 (0.3)	72.8 (0.3)	73.9 (0.6)
DA任务自适应预训练→任务自适应预训练	74.2 (0.3)	74.5 (0.5)	72.8 (0.4)	73.9 (0.4)

表 V：使用不同特征数量时，最终 XGBoost 模型的预测准确率、精确率、召回率和 F1 分数及其标准差（根据总计 5 次试验运行计算）。结果在零售数据集上进行评估。

由于购买商品数量的标签为 0、1、2、3、4，我们还希望检查模型在每个标签上的性能，也就是在每个购买数量范围上的性能。结果如表VI所示。

如表 VII 所示，与精确率 @5 相比，召回率 @5 相当低。这是可以理解的，因为本文所使用的零售数据集具有其自身特点。这家在线零售商的许多客户本身也是零售商

基于 XGB 算法的评估结果				
标签	TSDAE	MLM	T→掩码语言模型	掩码语言模型→T
0	75.3 (0.2)	77.6 (0.4)	80.2 (0.4)	82.0 (0.2)
1	50.4 (0.2)	52.4 (0.2)	53.9 (0.3)	55.4 (0.2)
2	58.7 (0.2)	60.4 (0.2)	60.8 (0.3)	62.6 (0.1)
3	77.8 (0.3)	78.1（0.3）	78.4（0.4）	79.2（0.3）
4	79.2（0.2）	79.4（0.1）	80.0（0.2）	81.4 (0.3)

表VI：采用不同预训练策略的领域自适应预训练后接任务自适应预训练所达到的最高预测准确率，取 3 次试验运行的平均值，并报告标准差。

基于矩阵分解的算法的评估结果				
标签	精确率@5	召回率@5	F1 分数	MAP
朴素矩阵分解	23.4 (0.3)	3.4 (0.1)	5.9 (0.2)	54.6 (0.1)
GLRM	26.4 (0.2)	4.8 (1.2)	8.1 (1.4)	57.4 (0.9)
隐式 MF	33.4 (0.5)	6.2 (1.4)	10.5 (1.1)	58.6 (2.4)
+簇标签	35.6 (0.8)	6.4 (0.7)	10.8 (0.8)	58.9 (0.8)
+嵌入相似度	**36.2** (1.0)	**6.6** (1.2)	**11.2** (1.3)	**60.2** (2.3)

表 VII：采用不同矩阵分解算法进行矩阵分解，在零售数据集上评估的精确率、召回率、F1 分数和 MAP@5。* 表示数据集经过如下筛选：从数据集中移除购买商品少于 5 件的任何用户，以及被购买次数少于 5 次的任何商品，用于训练和评估。

来自其他行业，而且他们通常会一次购买许多不同类型的商品。事实上，有些顾客购买的商品超过 40 件。即使模型给出的 5 条推荐结果全部正确，召回率 @5 也只有 5/40 = 12.5%。因此，召回率 @5 并不是评估矩阵分解模型性能的理想指标，因为这里给出的推荐结果数量通常远少于客户购买的商品数量。更适合关注的是基于精确率的指标，即 F1 分数或 MAP（平均准确率）。

我们推测，在这家在线零售商处购买过商品的不同客户之间并不存在很强的相互关系。客户往往会购买所属行业所需的商品。例如，家具零售商喜欢购买螺钉和其他相关零部件，通常一次购买数量很大。这与另一类用户（例如个人买家）的购买行为并无太大关联。引入顾客 ID 特征后模型性能有所提升，这也支持了这一推测。

基于矩阵分解的算法的评估结果				
标签	精确率@5	召回率@5	F1 分数	MAP
TSDAE @零售	35.8 (0.7)	6.4 (0.7)	10.9 (0.7)	59.2 (0.1)
掩码语言模型 @零售	36.3 (0.3)	6.6 (0.6)	11.2 (0.9)	60.6 (0.6)
T→M @零售	36.3 (0.8)	6.6 (1.2)	11.2 (1.4)	**60.4** (0.9)
M→T @零售	**36.4** (0.3)	**6.7** (0.5)	**11.2** (0.6)	60.2 (0.4)

表 VIII：使用第一列所列不同预训练策略的矩阵分解方法，在零售数据集上评估精确率、召回率、F1 分数和 MAP@5。此处使用的矩阵分解方法是隐式矩阵分解，采用由预训练的任务自适应预训练模型生成的句子嵌入之间的余弦相似度。

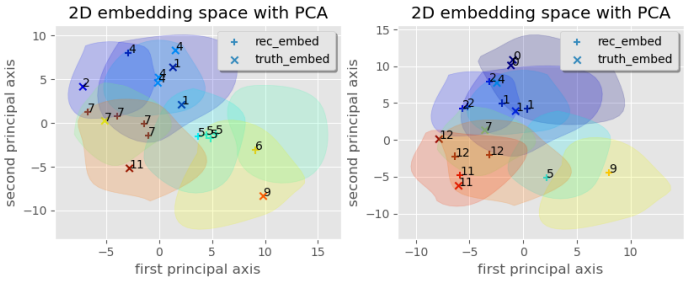
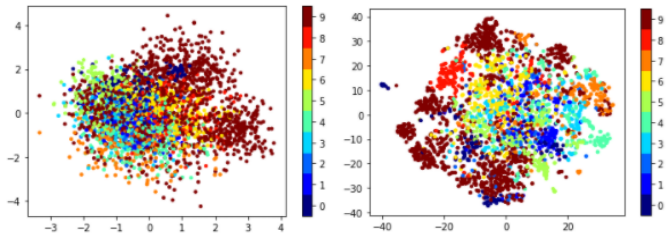


图 10：用户 0（左）和用户 1（右）的嵌入空间可视化



(a) 主成分分析　　(b) t-SNE

图 11：使用主成分分析（左）和 T-SNE（右）生成的嵌入空间，每个点代表特定产品描述的嵌入向量

这里选取了两个模型未能准确推荐商品的用户示例进行可视化，以便更好地理解其中的情况。这些示例通过主成分分析进行可视化，结果如图10所示。我们根据两个标准对其进行评估：预测的簇标签和商品的预测股票代码是否正确。对于左侧示例（详情见表IX），10个簇标签中只有1个预测正确，股票代码则无一预测正确。从图10可以看出，预测的商品大多聚集在簇7和簇5附近，而真实值商品则分散在它们周围。这意味着推荐系统已经捕捉到了一些模式，并试图关注其认为重要的潜在空间维度。然而，由于模型能力不足，它无法充分利用句子嵌入中蕴含的丰富信息。图10右侧的示例也呈现出类似模式。作为参考，图11展示了通过主成分分析（左）和t-SNE（右）将嵌入向量分布可视化为二维结果后的差异。主成分分析得到的结果分布更分散，而t-SNE得到的结果则聚集得更紧密。然而，由于句子嵌入具有高维特性，要在二维空间中准确解读结果仍然很困难，且解读结果会受到可视化方法的影响。

基于矩阵分解的算法评估结果										
	商品 1	商品 2	商品 3	商品 4	商品 5	商品 6	商品 7	商品 8	商品 9	商品 10
推荐	1605	266	1931	1930	749	620	751	1937	2609	2612
簇标签	7	5	5	6	7	4	5	5	7	7
真实值	170	171	1203	1205	1253	1410	1482	1483	1484	N/A
簇标签	7	11	1	1	2	9	4	4	4	N/A
预测正确性	✗	✗	✗	✗	✗	✗	✗	✗	✗	N/A
标签正确性	✓	✗	✗	✗	✗	✗	✗	✗	✗	N/A

表IX：针对模型表现不佳的一个用户示例的推荐、聚类标签预测及评估结果。

VII. 讨论与未来工作

实验表明，直接将句子嵌入用作推荐算法时，基于MF的推荐器表现不佳。研究者尝试将PLM生成的句子嵌入作为项目表示，与MF模型生成的项目特征结合。采用的方法包括：

- 结合 MF推荐器和 PLM推荐器的排名靠前的推荐结果，后者根据其句子嵌入的余弦相似度对商品进行排序。
- 将用户-项目矩阵中的每个条目视为 MF模型中用户向量与项目向量点积和 PLM中的相似度分数的加权和
- 利用 MF模型认为相似的其他用户所购买商品的句子嵌入作为辅助信息，以促进 MF推荐器的运行

然而，上述方法均未带来任何可观察到的推荐性能提升。这表明，由于句子嵌入的高维特性以及该在线零售数据集的特点，MF并不是适合直接用作推荐算法的理想模型。因此，需要一种能够更好地利用这些句子嵌入中丰富信息的模型。然而，正如 LadaBERT[Mao 等人所指出的, 2021]，当矩阵分解与权重剪枝和知识蒸馏结合使用作为模型压缩技术时，可以使 BERT模型减少内存和数据方面的开销。此外，它还能帮助模型在显著减少训练开销的情况下取得更好的性能。本文未涵盖这一方法，而它可以成为一个有趣的未来研究方向。

我们发现，引入基于 Transformer 的预训练语言嵌入模型能够提升传统推荐算法的性能，尤其是 XGBoost 推荐系统。通过使用 TSDAE 和 MLM 对 RoBERTa 模型进行精心设计的微调，还能进一步提高推荐准确率。实证数据表明，先进行 MLM、再进行 TSDAE 能带来更显著的性能提升。我们推测，MLM 预训练为模型提供了初步的领域特定知识，而 TSDAE 则进一步增强了模型在异构领域中的泛化能力和鲁棒性。

亚马逊描述数据集与零售描述数据集之间的词汇相似度相对较低，这可能会影响领域专用预训练的效果。此外，由于图形处理器内存的限制，本文使用的数据集与用于训练 RoBERTa 的语言语料库相比，规模较小。未来可以使用规模更大且结构上更接近在线零售数据集的数据集，来研究这种预训练方法的全部潜力。

还有许多基于深度学习的先进推荐算法可用作推荐模型。本文选择重点研究使用矩阵分解算法或 XGBoost 算法时，将预训练语言模型引入推荐系统的效果。探索预训练语言模型在基于深度学习的推荐系统中的优势，将是一个有趣的未来研究方向。

正如 Zhang 等人所提出并通过实验验证的那样， [Zhang 等人， 2020], 自适应混合掩码（AHM）通过引入一种新的掩码策略扩展了 MLM。具体而言，它设置了两种不同的模式，即词掩码模式和短语掩码模式。前者随机掩盖彼此独立的词语，后者则掩盖领域短语。此外，它能够根据反馈损失在这两种模式之间自适应切换，使模型逐步捕获词语级和短语级知识。与 MLM 相比，若将其选作微调方法，它更有潜力提升模型性能。此外，未来还可以采用新的先进预训练方法来替代本文使用的这两种方法，就像 AHM 一样。

VIII. 结论

近期的实证改进表明，将预训练语言模型和精心设计的微调策略融入传统推荐系统后，丰富、分阶段且无监督的预训练已成为提升模型性能不可或缺的重要环节，尤其是在蓬勃发展的电子商务领域。我们的主要贡献是研究将预训练语言模型与微调方法结合矩阵分解和 XGBoost 推荐算法的多种方式，并提出一种微调方案，使传统模型分别优于其基础版对应模型。

参考文献

[Aaron 和 Vanya， 2019] Aaron, G. 和 Vanya, C. (2019)。Openwebtext 语料库。http://skylion007.github.io/OpenWebTextCorpus。[Agarwal 等人， 2018] Agarwal, P., Vempati, S., 和 Borar, S. (2018)。电子商务时尚领域相似产品推荐的个性化。[Brown 等人， 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., 和 Amodei, D. (2020)。语言模型是少样本学习器。《神经信息处理系统进展》，2020-Decem。[Chaudhuri 等人， 2016] Chaudhuri, S., Theocharous, G., 和 Ghavamzadeh, M. (2016)。个性化广告推荐：应对普遍存在的点击稀疏问题的一种排序方法。[Chen 等人， 2015] Chen, D., Sain, S. L., 和 Guo, K. (2015)。在线零售数据集。https://archive.ics.uci.edu/ml/datasets/online+零售。[Chen 和 Guestrin， 2016] Chen, T. 和 Guestrin, C. (2016)。XGBoost：一种可扩展的树提升系统。ACM SIGKDD知识发现与数据挖掘国际会议论文集，13-17-Augu:785–794。[Cheng 等人， 2016] Cheng, H. T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., Anil, R., Haque, Z., Hong, L., Jain, V., Liu, X., 和 Shah, H. (2016)。用于推荐系统的宽度与深度学习。ACM国际会议论文集，15-Septemb:7–10。[Conneau 等人， 2017] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., 和 Bordes, A. (2017)。基于自然语言推理数据的通用句子表示监督学习。EMNLP 2017——自然语言处理实证方法会议论文集，第670–680页。[Covington 等人， 2016] Covington, P., Adams, J., 和 Sargin, E. (2016)。用于 YouTube 推荐的深度神经网络。RecSys 2016——第十届 ACM 推荐系统会议论文集，第191–198页。[Devlin 等人， 2019] Devlin, J., Chang, M. W., Lee, K., 和 Toutanova, K. (2019)。BERT：用于语言理解的深度双向变换器预训练。NAACL HLT 2019——2019年计算语言学协会北美分会：人类语言技术会议论文集，1(Mlm):4171–4186。[Elahi 等人， 2016] Elahi, M., Ricci, F., 和 Rubens, N. (2016)。协同过滤推荐系统中主动学习的综述。《计算机科学评论》，20(2018年12月):29–50。[Garber 和 Hazan， 2015] Garber, D. 和 Hazan, E. (2015)。通过凸优化快速简便地进行主成分分析。第1–28页。[Gupta 和 Manning， 2015] Gupta, S. 和 Manning, C. D. (2015)。用于引导自举实体分类器的词语分布式表示。NAACL HLT 2015——2015年计算语言学协会北美分会：人类语言技术会议论文集，第1215–1220页。[Gururangan 等人， 2020] Gururangan, S., Marasovi´c, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., 和 Smith, N. A. (2020)。不要停止预训练：让语言模型适应领域和任务。第8342–8360页。[Howard 和 Ruder， 2018] Howard, J. 和 Ruder, S. (2018)。用于文本分类的通用语言模型微调。ACL 2018——计算语言学协会第56届年会论文集（长篇论文），1:328–339。[Hu 等人， 2008] Hu, Y., Volinsky, C., 和 Koren, Y. (2008)。针对隐式反馈数据集的协同过滤。IEEE国际数据挖掘会议论文集，ICDM，第263–272页。[Jiang 等人， 2020] Jiang, Y., Shang, Y., Shen, H., Yang, W.-Y., 和 Xiao, Y. (2020)。微调BERT以实现电子商务非默认搜索排序。[Joshi 等人， 2020] Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., 和 Levy, O. (2020)。SpanBERT：通过表示和预测跨度改进预训练。《计算语言学协会会刊》，8:64–77。[Kowald 等人， 2020] Kowald, D., Lex, E., 和 Schedl, M. (2020)。利用人类记忆过程对流派偏好进行建模，以实现个性化音乐推荐。[Liu 等人， 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. 和 Stoyanov, V.（2019）。RoBERTa：一种稳健优化的 BERT 预训练方法。（1）。[Louppe， 2014] Louppe, G.（2014）。理解随机森林：从理论到实践。（7月）。[Mao 等人， 2021] Mao, Y., Wang, Y., Wu, C., Zhang, C., Wang, Y., Zhang, Q., Yang, Y., Tong, Y. 和 Bai, J.（2021）。LadaBERT：通过混合模型压缩实现 BERT 的轻量级适配。第 3225–3234 页。[Martin 和 Englot， 2017] Martin, J. 和 Englot, B.（2017）。在异方差环境中扩展基于模型的机器人策略梯度。第一届机器人学习会议论文集，78（CoRL）：438–447。[McCann 等人， 2018] McCann, B., Keskar, N. S., Xiong, C. 和 Socher, R.（2018）。自然语言十项全能：将多任务学习视为问答。[Nguyen 和 Zhu， 2013] Nguyen, J. 和 Zhu, M.（2013）。面向推荐系统的内容增强矩阵分解技术。《统计分析与数据挖掘》，6(4)：286–301。[Okura， 2017] Okura, S.（2017）。基于嵌入的新闻推荐主题模块：首次展示。第 1933–1942 页。[Peters 等人， 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. 和 Zettlemoyer, L.（2018）。深度上下文化词表示。NAACL HLT 2018——2018年计算语言学协会北美分会自然语言技术会议论文集，1:2227–2237。[Radford 等人， 2018] Radford, A., Narasimhan, T., Salimans, T. 和 Sutskever, I.（2018）。通过生成式预训练改进语言理解。预印本，第 1–12 页。[Raffel 等人， 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. 和 Liu, P. J.（2020）。探索统一文本到文本 Transformer 的迁移学习极限。《机器学习研究杂志》，21：1–67。[Ricci 等人， 2015] Ricci, F., Shapira, B. 和 Rokach, L.（2015）。《推荐系统手册》，第二版。[Sun 等人， 2019] Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H. 和 Wu, H.（2019）。ERNIE：通过知识集成增强表示。[Tran 和 Lee， 2018] Tran, T. 和 Lee, K.（2018）。利用用户和商品嵌入对矩阵分解进行正则化以实现推荐。国际信息与知识管理会议论文集，第 687–696 页。[Van Der Maaten 和 Hinton， 2008] Van Der Maaten, L. 和 Hinton, G.（2008）。使用 t-SNE 可视化数据。《机器学习研究杂志》，9（2008年11月）：2579–2625。[Wang 等人， 2012] Wang, Y., Chan, S. C. F. 和 Ngai, G.（2012）。人口统计推荐系统在旅游景点推荐中的适用性：TripAdvisor 案例研究。2012年 IEEE/WIC/ACM 网络智能与智能代理技术国际会议研讨会论文集（WI-IAT 2012），第 97–101 页。[Zhang 等人， 2020] Zhang, D., Yuan, Z., Liu, Y., Fu, Z., Zhuang, F., Wang, P., Chen, H. 和 Xiong, H.（2020）。E-BERT：一种融合短语与产品知识增强的电子商务语言模型。[Zhang 和 Bowman， 2019] Zhang, K. 和 Bowman, S.（2019）。语言建模比机器翻译教给你的更多：通过辅助句法任务分析总结的经验。第 359–361 页。