arXiv:2603.01773v1 [cs.CL] 2 Mar 2026

AnnoABSA：一款采用检索增强建议的基于网络的方面级情感分析标注工具

Nils Constantin Hellwig[1], **JakobFehle**[1], **Udo Kruschwitz**[2], **Christian Wolff**[1]1雷根斯堡大学媒体信息学组，德国雷根斯堡2雷根斯堡大学信息科学组，德国雷根斯堡
nils-constantin.hellwig@ur.de, jakob.fehle@ur.de, udo.krusch
witz@ur.de,christian.wolff@ur.de

摘要

我们介绍 AnnoABSA，这是首款支持方面级情感分析（ABSA）全方位任务的基于网络的标注工具。该工具高度可定制，可灵活配置情感元素和特定任务的要求。除了人工标注外，AnnoABSA 还提供可选的大语言模型（LLM）检索增强生成（RAG）建议，以人在回路方法提供上下文感知的辅助，同时确保人工标注员始终掌控标注过程。为随时间推移提升预测质量，系统会检索十个最相似的、已完成标注的示例，并将其作为少样本示例加入提示词中，确保随着标注过程的推进，建议愈发准确。AnnoABSA 以 MIT 许可证开源发布，可免费使用，并且易于扩展，适用于研究和实际应用。

关键词： 标注工具、基于方面的情感分析、检索增强生成、大语言模型、人工智能辅助

1. 引言

基于方面的情感分析（ABSA）是一种细粒度的情感分析方法。它通过识别文本中的具体方面并确定与各方面相关的情感倾向，超越了文档级极性分类（Pontiki 等人，2016）。该领域涵盖多种基于方面的情感分析子任务，这些任务在方面识别的粒度上各有不同。它们涉及以下情感要素的不同组合：方面词*a*、方面类别*c*、观点词 *o*和情感极性 *p*。例如，在句子 “披萨很美味。”中，“披萨”是方面词，“食物总体”可以作为相关的方面类别，“美味”是观点词，情感极性则为正面。如果某个方面没有对应的方面词（=隐式方面），则将方面词设为“NULL”，例如 “味道很美味。”。一个句子可能包含多个方面，因此需要标注多个方面。

由于方面级情感分析的粒度较细，创建用于训练和评估方面级情感分析专用模型的标注资源仍然极其耗时且费力（Nasution 和 Onan，2024年；Negi 等人，2024年；Wang 等人，2024年）。低资源语言和特定领域语境中的资源尤其匮乏（Xu 等人，2025年；Le Ludec 等人，2023；Fehle 等人，2025年）。

标注数据集的匮乏也反映在专用标注工具的有限供应上，目前只有少数专门解决方案。迄今为止，只有通用型标注框架，例如 *INCEpTION*[1] (Wo- jatzki 等人，2017年）、*BRAT*[2] (Pontiki 等人，2016，2015年，2014年）和*Label Studio*[3] (Hellwig 等人，2024年；Fehle 等人，2025年）据报道曾用于方面级情感分析标注任务。然而，这些工具缺少某些子任务所需的基本功能。例如，在目标方面情感检测（TASD）中，文本可能包含许多隐式方面，它们并不对应特定的文本片段，但可以归入某个方面类别，并赋予相应的情感极性。上述工具无法处理此类动态条目列表，在此情形下，即隐式观点的方面标注。

鉴于人工数据标注面临的挑战，近期研究日益关注利用大语言模型（LLM）来减少各类自然语言处理领域的标注工作量，包括社交媒体分析（Hasan 等人，2024；Mu 等人，2024；Zhang 等，2024）、（生物）医学（Labrak 等人，2024；Ateia 和 Kruschwitz，2023）以及金融（Deußer 等人，2024；Deng 等人，2024）。对于方面级情感分析，少样本学习在方面级情感分析任务中展现出有竞争力的性能，仅需少量示例即可达到接近微调模型的微平均 F1 分数（Hellwig 等，2025；Zhou 等，2024）。然而，这些方法的性能仍不及专门针对方面级情感分析任务进行微调的模型（Hellwig 等人，2025；Zhang 等人，2024；Zhou 等，2024）所报告的水平。

[1]INCEpTION（Klie 等人，2018）：https://inception-project.github.io/[2]BRAT：https://github.com/nlplab/brat[3]LabelStudio：https://labelstud.io/

在这项工作中，我们介绍**AnnoABSA**，这是一款专为方面级情感分析设计的网页标注工具。该工具提供丰富的自定义功能，并支持文献中记载的所有主要方面级情感分析子任务（Zhang 等人，2022年），包括方面词抽取、方面类别分类、情感极性检测、观点词识别，以及它们的各种组合（成对、三元组和四元组）。

借鉴近期其他耗时较长的标注任务研究（Ghazouali 和 Michelucci，2025年；Kim 等人，2024年；Sahitaj 等人，2025年；Li 等人，2025年），我们旨在整合基础模型的能力，协助标注人员完成标注过程。除了传统的手动标注功能外，AnnoABSA 还可选择性地集成一种基于检索增强生成（RAG）的建议机制，将基于大语言模型的预测与人类专业知识相结合。我们提出的 RAG 机制会从标注过程中先前已标注的实例样本池中检索语义上最相似的示例，以引导大语言模型提供建议。这种混合方法利用了大语言模型的高效性和一致性，同时保留了人工标注人员细致入微的判断力和领域知识，从而兼顾标注速度与质量。

我们的主要贡献如下：

- 我们介绍了首个面向方面级情感分析、全面兼容所有方面级情感分析子任务的开源标注工具，并根据宽松的 MIT 许可证发布于 https://github.com/NilsHellwig/AnnoABSA。

- 我们提出了一种基于检索增强的大语言模型建议机制，该机制利用最相似的已标注示例来提升标注效率，并随着时间推移提高建议质量。

- 系统评估表明，在预测性能方面，基于检索增强生成的建议显著优于随机抽样基线。

- 我们通过一项对照研究提供了证据，表明与无辅助的人工标注相比，在基于检索增强生成的建议辅助下，专家标注人员的标注时间显著减少了30.51%。

2. 系统描述

2.1. 动机

我们介绍 AnnoABSA，这是一款新颖的标注工具，旨在满足支持所有方面级情感分析任务（见附录A）的需求，同时提供便捷直观的用户界面（UI），以最少的用户交互实现高效标注。本节将详细介绍系统架构、用户界面设计决策以及 AnnoABSA 的数据管理。此外，表 1 概述了 AnnoABSA 的核心特性，并将其与以往用于方面级情感分析任务的现有标注工具进行比较。

2.2. 系统架构

AnnoABSA 使用 *React.js*[4], 实现，这是一个专为单页应用设计的前端框架。我们采用 *TypeScript*[5] 作为前端的主要编程语言，它是 JavaScript 的静态类型超集，可增强代码健壮性并降低出错可能性。后端采用*FastAPI*[6],这一 Python RESTful API 框架，从而能够集成*Pandas*[7] 和 *NumPy*[8] 等必要的 Python 包，以实现高效的数据处理与操作。

AnnoABSA 可通过命令行界面启动，只需提供 JSON 或 CSV 格式的待标注文本，以及一份指定标注参数的配置文件。

```
$ ./annoabsa reviews.json –load-config config.json
```

系统提供丰富的自定义选项，可在配置文件中或通过命令行标志进行指定。这些配置参数包括：用于标注的输入文本文件、需要考虑的情感元素（方面词、方面类别、观点词、情感极性）、有效情感极性和方面类别的列表，以及用于指定隐式方面词和/或观点词是否有效的布尔选项。因此，AnnoABSA 支持对任何语言的文本数据进行标注。所有受支持标志的全面概述见附录B。执行命令行工具后，AnnoABSA用户界面将在网页浏览器中打开。

2.3. 用户界面设计

用户界面见图1。界面组件经过精心设计，可在台式电脑和平板电脑上舒适使用。遵循极简设计原则（Sani和Shokooh,

[4]React.js: https://reactjs.org/
[5]TypeScript: https://www.typescriptlang.org/
[6]FastAPI: https://fastapi.tiangolo.com/
[7]Pandas: https://pandas.pydata.org/
[8]NumPy: https://numpy.org/

特性	Label Studio	INCEpTION	BRAT	AnnoABSA
技术要求				
开源	√ (Apache 2.0)	√ (Apache 2.0)	√ (MIT)	√ (MIT)
安装	Docker、Python 3	Docker、Tomcat、Java	Python 2	Python 3
定期更新	√	√	×2021年10月	√
基于网络	√	√	√	√
多用户支持	√	√	×	√
用户界面与易用性				
标注设置	XML 模板	项目设置	项目设置	JSON配置/命令行界面
界面设计	高度可定制（XML 模板）	技术性、功能性	技术性、功能性	现代、直观
语言翻译	√	×	×	√
标注指南	√ 可配置使用 XML	×	×	√ 集成于界面（PDF 格式）
文档	√	√	√	√
支持	社区论坛，GitHub 问题	GitHub 问题	GitHub 问题（不活跃的开发者）	GitHub 问题，电子邮件
数据管理				
导入/导出格式	JSON、TXT、XML；仅支持导出 CSV	UIMA、TSV、JSON、TXT	TXT、ANN（BRAT 站外格式）	JSON、CSV
常规标注功能				
多标签功能	√	√	√	√
关系标注	√	√	√	√
标签自定义	√ XML 模板	√ 项目配置	√ 配置文件	√ JSON 配置
团队协作	√ 全面支持并提供角色管理	×	×	×
词元识别[1]	√	√	√	√
面向 *ABSA* 的功能				
验证[2]	×	×	×	√
基于人工智能的建议	√ 集成单个模型/API	√ 集成单个模型/API	×	√
动态列表[3]	×	×	×	√
分配（多个）类别到一个文本片段	×	×	×	√

表 1： **方面级情感分析标注工具比较**。该表从技术要求、易用性特性、数据管理能力以及方面级情感分析专属功能等方面，对四个标注平台进行了详细比较。此表扩展了Colucci Cante 等人（2024年，第359页）提供的比较。他们依据以下标准评估了非语义文本标注工具：多标签功能、标注建议、关系标注、标签自定义和团队协作。

[1] 词元识别：自动识别文本中的词元边界，以防止标注范围从词语内部开始或结束而导致错误。

[2] 验证：Label Studio、INCEpTION 和 BRAT 不会强制要求在方面元组中关联情感要素。例如，在 TASD 任务中（元组由方面词、方面类别和情感极性组成），标注人员可能会标记一个方面词，却忘记设置相应的方面类别和情感极性。AnnoABSA 通过严格验证来避免此类错误，确保每个元组都包含特定任务所需的确切数量的情感要素。系统也会验证各个情感要素（例如，情感极性必须为“正面”“负面”或“中性”；方面类别必须来自预定义列表；方面词和观点词必须是表示隐式方面或观点的 NULL，或是文本中应被标注的子串）。

[3] 动态列表：方面级情感分析任务的标注可能包含不限数量的方面。例如，对于 ACD任务（考察方面类别与情感极性的配对），工具应提供相应功能，使用户能够为文档分配不限数量的方面类别与情感极性组合。Label Studio，INCEpTION 和 BRAT 无法处理将类别变量组合无限分配给文档的情况。

2016），我们专注于主要内容，避免冗余的视觉元素，并采用扁平化设计美学。

2.3.1. 顶部栏

顶部栏提供了便捷的示例导航功能，可使用方向键移至上一个或下一个实例。此外，双击

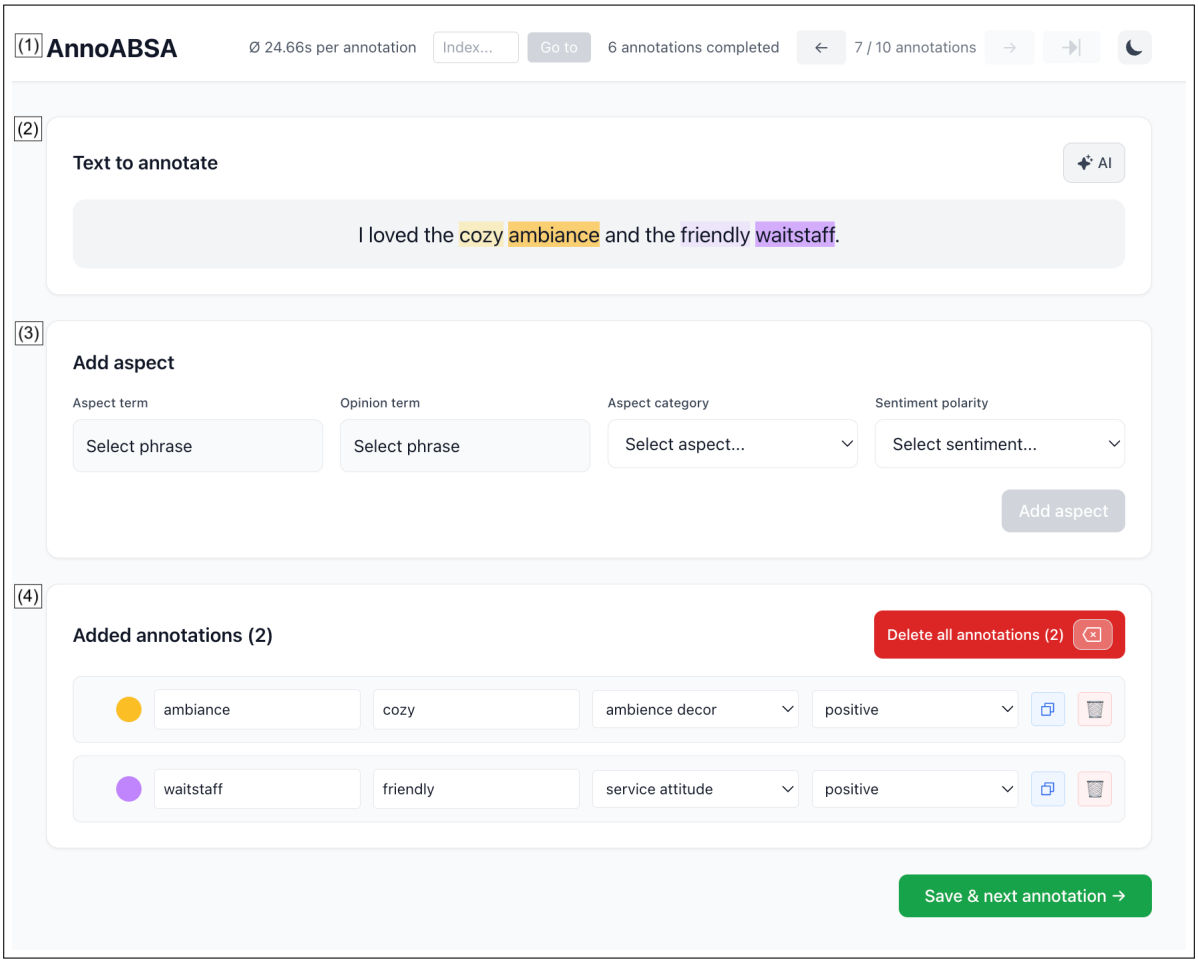


图1：**AnnoABSA 的用户界面**。用户界面由四个部分组成：（1）用于切换当前显示示例的顶部导航栏；（2）带有高亮情感标注的已标注文本；（3）方面添加面板；（4）已添加标注的可编辑列表。

右侧的箭头按钮可让用户直接跳转到索引最高的未标注示例，输入框则支持导航到指定索引。

2.3.2. 添加新方面的表单

用户界面的下一部分用于添加新的方面标注。根据命令行界面规范，会显示不同数量的待标注情感要素。图 1展示了全部四种情感要素，不过也可以配置为显示较少的情感要素。方面类别和情感极性这类分类情感要素可通过下拉菜单选择，而方面词和观点词则各自会打开一个弹出界面。

如图2所示，弹出窗口会两次显示待标注文本：一次用于标注方面词，另一次用于标注观点词。只需标注一种短语类型时，文本只显示一次。我们曾考虑采用工具切换方法（分别提供可选的方面词和观点词选择工具），但这会需要额外点击。我们选择的界面只需标记短语并点击“完成”，从而将用户操作减少到绝对必要的程度。

短语边界通过点击单个字符或（若已配置）词元来定义。标记短语后，文本及其位置会显示出来以供确认。值得注意的是，无论是在标注新方面还是编辑已有方面时，都会出现此弹出界面。指定所有情感要素后，用户点击“添加方面”，即可将该情感要素添加到下方“添加方面”部分所列的已应用标注中。

2.3.3. 标注列表

所有已标注的方面都会显示在列表中，并且可以随时修改。我们为方面添加了复制按钮，以便在方面仅有一个情感元素不同时使用。例如，在句子“这张披萨

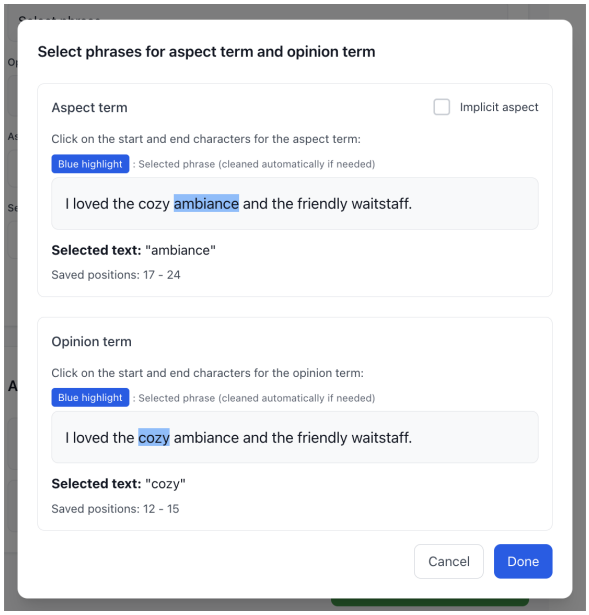


图 2：**用于添加或操作短语标注的弹出窗口**。 如果同时需要方面词和观点词标注，文本会显示两次，每次对应一种短语标注。隐式方面可通过复选框标记。

和汉堡都很美味”中，方面情感四元组预测（ASQP）的金标准标签由两个四元组组成：(’披萨’, ’食物质量’, ’美味’, ’正面’) 和 (’汉堡’, ’食物质量’, ’美味’, ’正面’)。标注时，可以复制一个四元组，并相应修改相关情感元素。如有需要，也可以删除标注。

2.4. 数据处理

包含所有待标注示例的 CSV 或 JSON 文件会在标注过程中被直接修改，标注会随着标注工作的推进而追加。除给定文本中的情感要素和短语位置外，还可选择存储标注时长。使用 JSON 格式可在需要时以最少的代码修改轻松集成到 NoSQL 数据库。

3. 检索增强生成建议

鉴于人工标注员开始为方面级情感分析文本进行标注时，尚无可用于监督模型的已标注训练样例，我们采用基于大语言模型的方法生成标注建议。先前的研究（Hellwig 等人，2025年；Zhou等人，2024年；Zhang 等人，2024年）表明，利用少量固定少样本示例的大语言模型，其性能可接近微调模型，尤其是在低资源场景中中，而且，基于检索增强生成（RAG）的少样本学习在各类方面级情感分析任务中的表现均优于随机抽样。

本节探讨技术考量，包括模型支持和提示词技术，随后对用于生成标注建议的随机抽样与基于检索增强生成的抽样方法进行比较评估。

3.1. 模型支持

AnnoABSA 通过适当的命令行参数灵活集成模型，既支持通过 OpenAI API[9] 接入商业模型，也支持通过本地托管的 Ollama API[10]接入开源模型。选择这些工具包，是因为它们具有平台独立性，并原生实现了结构化输出能力。通过引导式解码生成结构化输出，可确保大语言模型的建议仅限于命令行配置中定义的方面类别和情感极性，从而防止生成幻觉标签。此外，结构化输出还能确保预测短语确实出现在待标注文本中。

3.2. 提示词

我们采用了Hellwig 等人（2025年）和Gou 等人（2023）使用的提示结构，该结构由任务描述、示例和待标注文本组成。我们对提示词所做的一项修改，是将示例标签格式化为 JSON，这与前述强制结构化输出的做法类似。附录C中提供了提示模板示例。

3.3. 示例选择策略评估

尽管Zhou等人（2024年）提出的基于检索增强生成的方法采用k 个语义最相似的训练示例作为少样本示例，取得了高于随机选择的性能得分，但其方法无法直接应用于标注过程。他们的方法假设可以访问一个经过完整标注的训练语料库，以便选择示例；而在标注场景中，人类已标注实例的集合（样本池）会随着标注过程的推进而不断扩大。为解决这一差异，我们在标注框架中对随机抽样和基于检索增强生成的抽样策略进行了性能分析。性能分析中的所有大语言模型调用均在以下条件下进行

[9]OpenAI API: https://openai.com/api/[10]
Ollama: https://ollama.com/

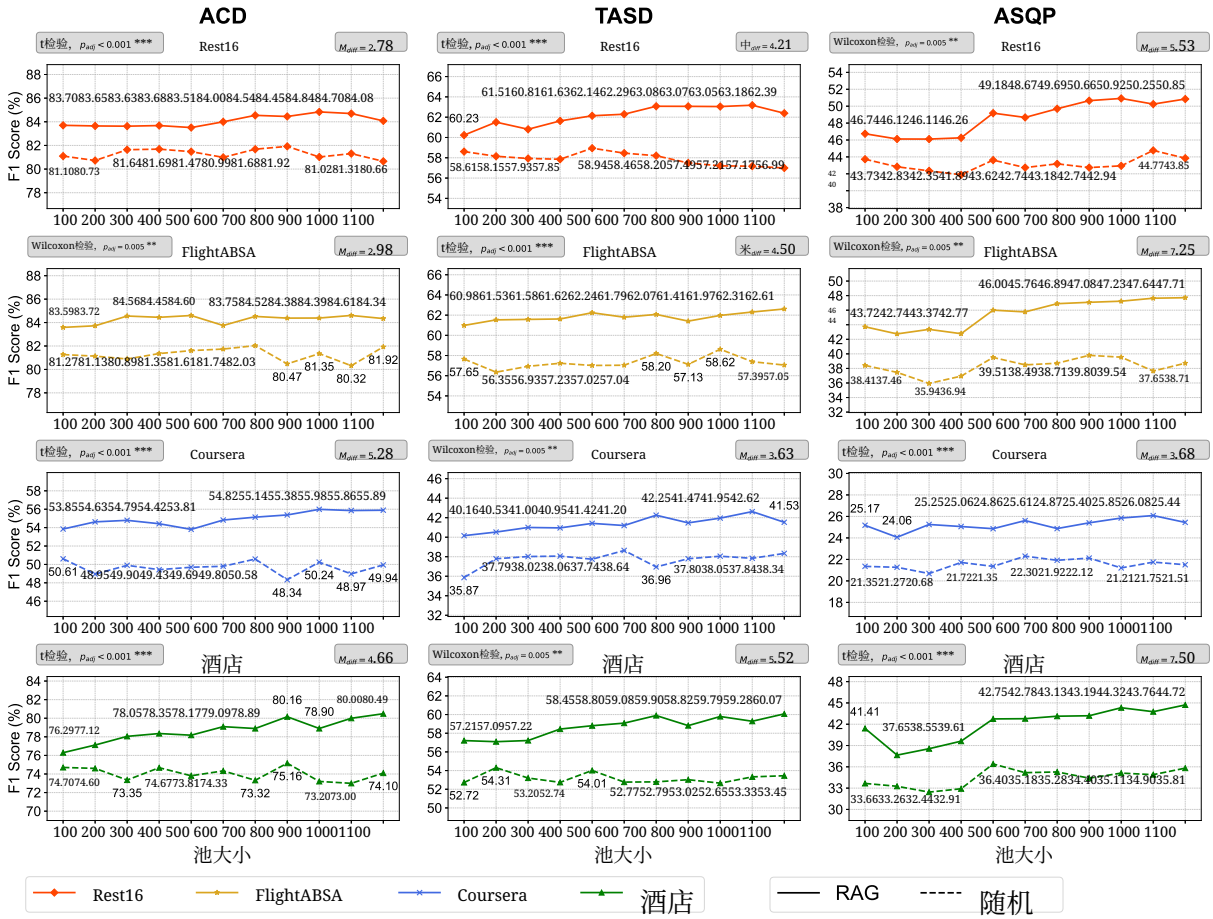


图3：**基于检索增强生成的方法与随机抽样方法的 F1 分数比较**。检索增强生成方法在所有配置下均持续优于随机抽样。统计检验（配对 t 检验或经 Holm-Bonferroni 校正的 Wilcoxon 符号秩检验）证实，所有差异均具有统计显著性。 M_{diff} 显示了平均性能差异。

在配备 96 GB 显存的 NVIDIA RTX PRO 6000 GPU 上。

3.3.1. 方法

选择策略。 遵循Zhou等人（2024年）确立的方法，我们采用 BM25（Robertson等人，2009年）作为检索增强生成的稀疏检索算法，因为它能够快速比较相似度，从而加快建议生成。对于随机抽样，则从样本池中随机选取示例。

样本池大小。为研究可用于少样本示例检索的金标准标注示例数量增加时性能的变化，我们分析了从 0 到 1,100 个可用训练示例的范围，增量步长为 100。例如，当样本池大小为 300 时，少样本示例从 300 个示例中选取。我们将 1,100 个示例视为样本池的最大大小，因为这是所有数据集中可用的最大 100 的倍数。

数据集。我们在以下数据集上评估了性能：这四个数据集涵盖了不同的评论领域，包括餐厅评论（SemEval2016、Rest16）（Pontiki 等人，2016；Zhang 等人，2021）、在线学习课程评论（Coursera）（Chebolu 等人，2024年）、航空公司评论（FlightABSA）（Hellwig 等人，2025年）以及酒店评论（Chebolu 等人，2024年）。我们从相应的训练集中随机选取了 1,100 个样例。

大语言模型配置。与Hellwig 等人(2025年)类似，我们采用了谷歌的 Gemma-3-27B（Team 等人，2025年），并将温度设为 0，以确保在下一词元预测过程中确定性地选择概率最高的词元。提示上下文包含 10 个少样本示例。对于每种样本池大小、任务和数据集的组合，大语言模型均运行五次。每次运行都使用不同的随机种子，以确保从相应训练集中抽取的 1,100 个示例具有不同的选择结果。

任务。我们评估了三项复杂程度各异的任务：一项单方面任务（方面类别检测，ACD）和两项元组预测任务：目标方面情感检测（TASD），以及

方面情感四元组预测（ASQP）。ACD 要求识别给定文本中涉及的所有方面类别。TASD 提取由方面词、方面类别和情感极性组成的观点三元组，而 ASQP 还会提取观点词，是粒度最细的方面级情感分析任务。

评估指标。 每个步骤的评估均在各数据集的完整相应测试集上进行。如方面级情感分析领域的惯例，所报告的评估指标为微平均F1值（Zhang 等人，2022年）。我们发布所有预测标签，并在GitHub仓库中提供宏平均F1分数、精确率和召回率的性能得分。

3.3.2. 结果与讨论

结果（见图3）表明，检索增强生成方法在所有数据集的全部池规模和任务上均持续优于随机少样本选择。值得注意的是，检索增强生成取得了显著的性能提升，在若干情况下与随机抽样相比，差异（M_{diff}）高达六个百分点。

Pool	大语言模型执行时间（秒）					
	ACD		TASD		ASQP	
	RAG	随机 检索增强生成		随机 检索增强生成		随机性
100	0.816	0.851	1.292	1.240	1.708	1.622
200	0.848	0.852	1.291	1.204	1.723	1.626
300	0.846	0.871	1.291	1.211	1.736	1.642
400	0.862	0.856	1.298	1.221	1.743	1.622
500	0.858	0.867	1.314	1.220	1.743	1.647
600	0.868	0.852	1.308	1.219	1.745	1.630
700	0.864	0.884	1.311	1.210	1.739	1.640
800	0.865	0.872	1.320	1.210	1.755	1.638
900	0.868	0.864	1.322	1.204	1.759	1.635
1,000	0.872	0.856	1.310	1.202	1.745	1.636
1,100	0.866	0.859	1.296	1.199	1.756	1.617
AVG	0.858	0.862	1.305	1.213	1.741	1.632

表 2：**方面级情感分析任务的大语言模型推理时间对比**。对 ACD、TASD 和 ASQP 任务，在不同池规模下比较基于检索增强生成的方法与随机抽样策略的平均单次预测执行时间（秒）。

对于服从正态分布的差值，使用配对 t 检验检验统计显著性；对于不服从正态分布的数据集，则使用Wilcoxon符号秩检验，随后采用 Holm-Bonferroni 校正（Holm，1979年）以校正 12 项比较中的多重检验（$\alpha = 0.05$）。对于每种任务与数据集的组合，我们比较了检索增强生成与随机抽样条件下获得的 11 个性能分数（每种池大小对应一个分数）。统计显著性在所有任务与数据集的组合中均有体现。

总体而言，我们的研究结果表明，随着用于抽取少样本示例的样本池不断扩大，基于检索增强生成的方法比随机抽样取得了更高的性能分数；在所有任务和数据集上观察到的差异均具有统计显著性。因此，我们将检索增强生成方法整合到了 AnnoABSA 的最终版本中。

如Hellwig 等人(2025年)此前针对随机抽样所展示的那样，增加少样本示例的数量有望进一步提升性能。然而，这会增加预测延迟，因为需要将更多词元加载到模型上下文中。在我们的 NVIDIA RTX PRO 6000 硬件上，单项任务的预测可在数秒内完成，如表 2所示；但增加少样本示例的数量，将按比例延长处理时间，若使用商用大语言模型，还会增加费用。

3.4. 关于标注速度的用户研究

与其他大语言模型辅助的标注工具一样，AnnoABSA 基于大语言模型的建议旨在缩短标注时间（Kim 等人，2024年）。因此，我们开展了一项用户研究，以评估与不使用人工智能辅助的人工标注相比，AnnoABSA 是否能加快标注速度。我们选择 ASQP 作为目标任务；在所有方面级情感分析任务中，该任务每个元组所包含的情感要素最多（方面词、方面类别、观点词和情感极性），因此是最全面的标注任务。

3.4.1. 方法

我们开展了一项被试内用户研究，共有 8 名专家标注人员参与，其中包括 4 名计算机科学专业的博士生和 4 名硕士生，他们此前均有方面级情感分析标注任务的经验。研究在受控的可用性实验室环境中进行，以尽量减少外界干扰；标注任务开始前，研究主管仅在初始说明阶段在场回答问题。

本研究采用平衡设计，每位参与者在不同日期完成两次标注会话，以消除疲劳效应：一次启用 AI 建议，另一次不启用建议。为控制潜在的顺序效应和学习效应，我们采用拉丁方平衡，在参与者之间系统地调整系统使用顺序（先使用 AI 还是先使用基线系统）

以及数据集分配（子集 A 或子集 B）。

子集 A 和子集 B 均由从餐厅数据集中随机抽取的 50 个示例组成，该数据集作为 SemEval 2016 任务 5 的一部分发布（Pontiki 等人，2016）。参与者获提供相应的标注指南以供参考。在每次会话中标注 50 个测试示例之前，参与者先通过 5 个演示示例完成熟悉阶段，以了解工具界面；期间，他们可以提问，并接受研究主管的指导。

我们采用了与选择策略评估中相同的大语言模型（Gemma-3-27B）和 GPU 配置。鉴于标注者处理每个示例所花费的时间是我们的主要评估指标，我们实施了基于前端的时间追踪。每个示例加载时计时器启动，参与者在界面中打开下一个示例时计时器停止。

3.4.2. 结果与讨论

配对 t 检验显示，人工智能辅助条件下（$M = 24.19, SD = 3.49$）与基线条件下（$M = 34.80, SD = 6.92$）每个示例的平均标注时间存在统计学显著差异，$t(7) = -3.640$，$p = 0.008$。Shapiro-Wilk 检验证实，差值的分布符合正态性假设（$W = 0.944$, $p = 0.647$）。这些发现表明，大语言模型生成的建议能够显著加快人工标注工作流程，使标注时间减少 30.51%。

我们的结果与先前研究一致，这些研究表明，人工智能辅助标注可以减少人工标注任务所需的时间（Kim 等人，2024年；Ni 等人，2024年；Sahitaj 等人，2025年；Li 等人，2025年），尽管我们研究中观察到的时间缩减幅度小于其他任务中报告的结果。值得注意的是，Ni 等人（2024年）证明，通过对 GPT-4 输出进行一致性检查，近一半的事实性主张标注可以完全自动化，从而将专家标注工作量减少约 50%。同样，Sahitaj 等人（2025年）报告称，推文中的宣传内容检测标注时间减少了 73%；而Li 等人（2025年）发现，人工智能辅助预标注 X 射线图像的标注时间减少了 36%。

4. 结论与未来工作

我们介绍了 AnnoABSA，这是首个支持方面级情感分析所有子任务的开源、基于网络的标注工具。AnnoABSA 具有直观且可定制的界面，内置验证机制，并可选择通过少样本提示提供大语言模型驱动的建议。我们的评估表明，集成的基于检索增强生成的方法能够在标注过程中动态利用不断扩大的已标注示例样本池，其表现显著优于随机抽样。用户研究发现，采用人工智能辅助建议可显著缩短标注时间。

尽管本研究展示了 AnnoABSA 的功能能力及其对标注效率的积极影响，但未来研究应系统地考察人工智能辅助如何影响方面级情感分析任务中的标注时间、质量和标注者信心。此类研究可同时考察众包工作者以及具备成熟方面级情感分析专业知识的领域专家。

最后，我们指出，我们的方法无需进行特定任务的模型训练，一经部署即可立即发挥作用，因而能够便捷地适配方面级情感分析之外的多种自然语言处理标注场景。

我们邀请社区通过我们的 GitHub 仓库为 AnnoABSA 的持续开发做出贡献。欢迎以问题和拉取请求的形式提出特性请求，我们承诺及时整合有价值的贡献，以惠及更广泛的研究社区。

5. 伦理声明与局限性

本研究在没有工业资助或商业赞助的情况下开展。我们使用了 AI 编程智能体 Claude Sonnet 4.5[11] 以及 OpenAI 的开源大语言模型 gpt-oss:20b[12] 来提供编程支持。Claude Sonnet 4.5 也被用于协助撰写本出版物。

在解读我们的结果时，应考虑若干局限性。首先，由于计算资源限制，我们对基于大语言模型的建议的评估仅限于 Gemma-3-27B。更大的语言模型或增加少样本示例规模，有可能带来更高质量的建议，但对于商业专有大语言模型而言，这将带来更高的计算或财务成本。我们为评估基于大语言模型的建议执行了 580,690 条提示词，若采用商业模型，将产生可观的费用。不过，如果希望采用其他模型，AnnoABSA 支持任意 Ollama 模型以及兼容 OpenAI 的模型。

最后，尽管我们展示了标注时间显著缩短，但评估仅限于小规模标注任务。这些效率提升能否推广到涉及数千个示例的大规模真实场景，

[11] Claude Sonnet: https://www.anthropic.com/claude/sonnet [12] gpt-oss:20b: https://ollama.com/library/gpt-oss:20b

仍有待验证，尤其是长时间标注会话中标注者疲劳可能产生的累积效应。

6. 参考文献

Samy Ateia和Udo Kruschwitz。2023年。ChatGPT 是生物医学专家吗？载于评估论坛会议与实验室工作论文集（*CLEF 2023*），希腊塞萨洛尼基，*2023*年*9*月*18*日至*21*日，第3497卷*CEUR*研讨会论文集，第73–90页。CEUR-WS.org。

Yoav Benjamini和Yosef Hochberg。1995年。控制错误发现率：一种实用且有效的多重检验方法。英国皇家统计学会会刊。*B*辑（方法论），57(1)：289–300。

Siva Uday Sampreeth Chebolu、Franck Dernoncourt、Nedim Lipka和Thamar Solorio。2024年。Oats：一个用于基于方面的情感分析的观点方面目标情感联合检测挑战数据集。载于*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），第12336–12347页。

Luigi Colucci Cante、Salvatore D’Angelo、Beni- amino Di Martino 和 Mariangela Graziano。2024年。文本标注工具：全面综述与比较分析，第353–362页。施普林格。

Yong Deng、Xintong Zhang、Danping Zhou、De- quan Zhang 和 Boya Huang。2024年。金融领域中自然语言处理的应用：结合大语言模型与思维链推理的协同方法。载于第五届人工智能与计算机工程国际会议论文集，第494–500页。

Tobias Deußer、Cong Zhao、Daniel Uedelhoven、Lorenz Sparrenberg、Lars Hillebrand、Christian Bauckhage 和 Rafet Sifa。2024年。利用大语言模型从财务报告中进行少样本关键绩效指标提取。载于*2024*年*IEEE*国际大数据会议（*BigData*），第4864–4868页。IEEE。

Jakob Fehle、Niklas Donhauser、Udo Kruschwitz、Nils Constantin Hellwig 和 Christian Wolff。2025年。《真实场景中的德语基于方面的情感分析：B2B数据集的创建与跨领域评估》。载于第*21*届自然语言处理会议（*KONVENS 2025*），第9卷，第213页。

Safouane El Ghazouali 和 Umberto Michelucci。2025年。Visiofirm：面向计算机视觉的跨平台 AI 辅助标注工具。

Zhibin Gou、Qingyan Guo 和 Yujiu Yang。2023。MVP：多视角提示词改进了方面情感元组预测。载于第*61*届计算语言学协会年会论文集（第*1*卷：长篇论文），第4380–4397页。

Md Arid Hasan、Shudipta Das、Afiyat Anjum、Firoj Alam、Anika Anjum、Avijit Sarker 和 Sheak Rashed Haider Noori。2024年。使用大语言模型进行零样本与少样本提示：与孟加拉语情感分析微调模型的比较研究。载于*2024*年计算语言学、语言资源与评测联合国际会议论文集（*LREC-COLING 2024*），第17808–17818页。

Nils Constantin Hellwig、Jakob Fehle、Markus Bink 和 Christian Wolff。2024年。GERestaurant：用于基于方面的情感分析的德语餐厅评论标注数据集。载于第*20*届自然语言处理会议（*KONVENS 2024*）论文集，第123–133页，奥地利维也纳。计算语言学协会。

Nils Constantin Hellwig、Jakob Fehle、Udo Kr- uschwitz 和 Christian Wolff。2025年。我们还需要人工标注人员吗？提示大语言模型进行方面情感四元组预测。收录于 首届大语言模型与结构建模联合研讨会（*XLLM 2025*）论文集，第153–172页，奥地利维也纳。计算语言学协会。

Sture Holm。1979年。一种简单的序贯拒绝式多重检验程序。《斯堪的纳维亚统计学杂志》，第65–70页。

Hannah Kim、Kushan Mitra、Rafael Li Chen、Sajjadur Rahman 和 Dan Zhang。2024年。Meganno+：一种人类与大语言模型协作的标注系统。收录于欧洲计算语言学协会第*18*届会议：系统示例论文集，第168–176页。

Jan-Christoph Klie、Michael Bugert、Beto Boullosa、Richard Eckart de Castilho 和 Iryna Gurevych。2018年。INCEpTION平台：机器辅助、知识导向的交互式标注。收录于 第*27*届计算语言学国际会议：系统示例论文集，第5–9页，圣菲，新墨西哥州。

Yanis Labrak、Mickaël Rouvier 和 Richard Du-four。2024年。关于

指令微调大语言模型在临床和生物医学任务中的应用。在*2024*年计算语言学、语言资源与评估联合国际会议（*LREC-COLING 2024*）论文集，第 2049–2066 页。

Clément Le Ludec、Maxime Cornet 和 Antonio A Casilli。2023年。标注的问题：法国与马达加斯加之间的人力劳动与外包。*Big Data& Society*，10(2):20539517231188723。

Yan Li、Hao Qiu、Xu Wang、Na Dong 和 Xinghua Yu。2025年。Rapidx标注器：用于工业射线图像标注与增强的专业软件工具。*SoftwareX*，31:102328。

Yida Mu、Ben P. Wu、William Thorne、Ambrose Robinson、Nikolaos Aletras、Carolina Scarton、Kalina Bontcheva 和 Xingyi Song。2024年。应对零样本分类中的提示复杂度：大语言模型在计算社会科学中的一项研究。载于*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），第12074–12086页，都灵，意大利。ELRA和ICCL。

Arbi Haza Nasution 和 Aytug Onan。2024年。Chat-gpt label：比较人类生成标注与大语言模型生成标注在低资源语言自然语言处理任务中的质量。*IEEEAccess*。

Gaurav Negi、Rajdeep Sarkar、Omnia Zayed 和 Paul Buitelaar。2024年。使用迁移学习进行基于方面的情感分析的混合方法。载于*2024*年计算语言学、语言资源与评测联合国际会议论文集（*LREC-COLING 2024*），第647–658页。

Jingwei Ni、Minjing Shi、Dominik Sta mmbach、Mrinmaya Sachan、Elliott Ash 和 Markus Leippold。2024年。AFACTA：使用可靠的大语言模型标注器辅助事实性主张检测的标注。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第1890–1912页。

Maria Pontiki、Dimitrios Galanis、Harris Papageorgiou、Ion Androutsopoulos、Suresh Manandhar、Mohammad AL-Smadi、Mahmoud Al-Ayyoub、赵妍妍、秦兵、Orphee De Clercq等。2016。SemEval-2016第5项任务：基于方面的情感分析。见第十届语义评测国际研讨会论文集（*SemEval-2016*），第19–30页。

Maria Pontiki、Dimitris Galanis、Haris Papageorgiou、SureshManandhar 和 Ion Androutsopou- los。2015年。SemEval-2015 第12项任务：基于方面的情感分析。载于《第九届语义评估国际研讨会论文集（*SemEval 2015*）》，第486–495页，丹佛，科罗拉多州。计算语言学协会。

Maria Pontiki、Dimitris Galanis、John Pavlopoulos、Harris Papageorgiou、Ion Androutsopoulos和Suresh Manandhar。2014年。SemEval-2014第4项任务：基于方面的情感分析。见第八届语义评估国际研讨会论文集（*SemEval 2014*），第27–35页，爱尔兰都柏林。计算语言学协会。

Stephen Robertson、Hugo Zaragoza 等。2009年。概率相关性框架：BM25 及其扩展。《信息检索基础与趋势》，3(4):333–389。

Ariana Sahitaj、Premtim Sahitaj、Veronika Solopova、Jiaao Li、Sebastian Möller 和 Vera Schmitt。2025年。用于宣传检测的混合标注：整合大语言模型预标注与人类智能。载于《第四届自然语言处理正向影响研讨会（*NLP4PI*）论文集》，第215–228 页。

Somayeh Sani 和 Yeganeh Shokooh。2016年。基于格式塔视觉感知规律的商业网站用户界面极简设计（科技领域三个顶级品牌的案例研究）。第 115–124 页。

Hope Schroeder、Deb Roy 和 Jad Kabbara。2025年。让人类参与其中就够了吗？对用于主观任务的大语言模型辅助标注进行研究。载于《计算语言学协会会议论文集：*ACL 2025*》，第 25771–25795 页。

Gemma 团队、Aishwarya Kamath、Johan Fer- ret、Shreya Pathak、Nino Vieillard、Ramona Merhej、Sarah Perrin、Tatiana Matejovicova、Alexandre Ramé、Morgane Rivière 等。2025年。Gemma 3 技术报告。*arXiv* 预印本 *arXiv:2503.19786*。

An Wang、Junfeng Jiang、Youmi Ma、Ao Liu 和 Naoaki Okazaki。2024年。用于方面情感四元组预测的生成式数据增强。自然语言处理期刊，31(4):1523–1544。

Michael Wojatzki、Eugen Ruppert、Sarah Holschneider、Torsten Zesch 和 Chris Biemann。2017年。GermEval 2017：社交媒体客户反馈中的基于方面的情感共享任务。*GermEval* 论文集，第1–12页。

Hongling Xu、Yice Zhang、Qianlong Wang 和 Ruifeng Xu。2025年。DS^2-ABSA：双流数据合成与标签细化，适用于少样本

基于方面的情感分析。载于计算语言学协会第 *63* 届年会论文集（第 *1* 卷：长篇论文），第 15460–15478 页，奥地利维也纳。计算语言学协会。

Wenxuan Zhang、Yang Deng、Xin Li、Yifei Yuan、Lidong Bing 和 Wai Lam。2021。将方面情感四元组预测视为释义生成。载于*2021*年自然语言处理实证方法会议论文集，第 9209–9219 页。

Wenxuan Zhang、Yue Deng、Bing Liu、Sinno Pan 和 Lidong Bing。2024。大语言模型时代的情感分析：现实检验。载于计算语言学协会会议论文集：*NAACL 2024*，第 3881–3906 页。

Wenxuan Zhang、Xin Li、Yang Deng、Lidong Bing 和 Wai Lam。2022年。基于方面的情感分析综述：任务、方法与挑战。*IEEE*知识与数据工程汇刊，35(11)：11019–11038。

Changzhi Zhou、Dandan Song、Yuhang Tian、Zhijing Wu、Hao Wang、Xinyu Zhang、Jun Yang、Ziyi Yang 和 Shuhao Zhang。2024年。对大语言模型在基于方面的情感分析上的表现进行全面评估。*arXiv*预印本 *arXiv:2412.02279*。

A. 方面级情感分析任务

任务类型	任务名称	元素	金标准标签（示例）
单项	方面词提取（ATE）	*a*	['服务员']
	方面类别检测（ACD）	*c*	['食物质量', '服务速度']
	观点词提取（OTE）	*o*	['美味', '太慢']
复合型	方面情感分类（ASC）	*a*, *s*	[('服务员', '负面')]
	方面-观点对抽取（AOPE）	*a*, *o*	[('服务员', '太慢')]
	端到端方面级情感分析（E2E-ABSA）	*a*, *s*	[('NULL', '正面'), ('服务员', '负面')]
	方面类别情感分析（ACSA）	*c*, *s*	[('食物质量', '正面'), ('服务速度', '负面')]
	方面情感三元组抽取（ASTE）	*a*, *o*, *s*	[('NULL', '美味', '正面'), ('服务员', 'way 太慢了', '负面')]
	目标方面情感检测 (TASD)	*a*, *c*, *s*	[('NULL', '食物 质量', '正面'), ('服务员', '服务速度', '负面')]
	方面情感四元组预测 (ASQP)	*a*, *c*, *o*, *s*	[('NULL', '食物 质量', '美味', '正面'), ('服务员', '服务速度', 'way 太慢', '负面')]
	方面-类别-观点-情感四元组抽取（ACOS）	*a*, *c*, *o*, *s*	[('NULL', '食物质量', '美味', '正面'), ('服务员', '服务速度', '太慢', '负面')]

表 3：**AnnoABSA 支持的方面级情感分析任务概览**。对于每项任务，均展示句子“味道确实很美味，但服务员的速度实在太慢了”对应的金标准标注。方面类别通常从预定义集合中选择（{食物质量，服务速度，...}）。符号说明：*a* = 方面词，*c* = 方面类别，*o* = 观点词，*s* = 情感极性。对于隐式方面，方面词*a* 设为'NULL'。所列任务与Zhang 等人(2022年)在文献综述中报告的任务相同。

B. 命令行标志

选项	描述	默认值
服务器配置		
–后端	仅启动后端服务器	–
–backend-port	后端服务器端口	8000
–frontend-port	前端服务器端口	3000
–backend-ip	后端服务器 IP 地址	本地主机
–frontend-ip	前端服务器 IP 地址	本地主机
会话管理		
–session-id	标注会话的唯一标识符。	None
标注元素		
–elements	要标注的情感要素	方面_term, 方面_category, sen-情感_极性，意见_术语
–极性	有效的情感极性	正面、负面、中性
–类别	有效的方面类别	餐厅领域（13 个类别）
–implicit-aspect	允许使用隐式方面词	已启用
–disable-implicit-aspect	禁用隐式方面词	已禁用
–implicit-opinion	允许使用隐式意见词	已禁用
–disable_implicit_opinion	禁用隐式意见词	已启用
界面与处理		
–禁用_清理_短语	禁用对以下内容的自动标点清理 短语 起始/结束位置	已启用
–disable-save-positions	禁用保存短语位置	已启用
–disable-click-on-token	禁用点击词元功能	已启用
–auto-positions	启动时启用自动填充位置	已禁用
–annotation-guidelines	包含标注指南的 PDF 文件路径	已禁用
分析与计时		
–store-time	存储标注会话的计时数据	已禁用
–display-avg-annotation-time	在界面中显示平均标注时间	已禁用
AI 集成		
–ai-suggestions	启用基于大语言模型的 AI 驱动的预测建议	已禁用
–disable-ai-automatic-prediction	禁用自动触发 AI 预测	已禁用
–llm-model	用于生成建议的大语言模型（例如 gemma3:4b、gpt-4o）	gemma3:4b
–openai-key	用于使用 OpenAI 模型的 OpenAI API 密钥	–
–n-few-shot	大语言模型上下文中少样本示例的最大数量	10
配置管理		
–save-config	将配置保存到 JSON 文件	–
–show-config	在控制台中显示当前配置	–

表 4：**AnnoABSA 的命令行界面选项**。 AnnoABSA 可对服务器设置、标注逻辑和 AI 集成进行广泛定制。

C. 基于 RAG 的建议提示词

```
According to the following sentiment elements definition:- The 'aspect term' is the exact word or phrase in the text that represents a specific feature,
 attribute, or aspect of a product or service that a usermay express an opinion about. The aspect term might be 'NULL' for implicit aspect.
- The 'aspect category' refers to the category that aspect belongs to, and the available categories includes: hotel comfort, rooms design_features,
facilities miscellaneous, rooms cleanliness, food_drinks miscellaneous, food_drinks style_options, facilities comfort, room_amenities general, servicegeneral,
 rooms quality, rooms general, hotel general, food_drinks prices, facilities general, room_amenities quality, facilities quality, roomsmiscellaneous,
 facilities design_features, hotel prices, food_drinks quality, room_amenities prices, room_amenities comfort, rooms prices, hotelcleanliness,
 hotel miscellaneous, facilities prices, rooms comfort, hotel quality, room_amenities design_features, location general, facilitiescleanliness,
 hotel design_features, room_amenities cleanliness.- The 'sentiment polarity' refers to the degree of positivity,
 negativity or neutrality expressed in the opinion towards a particular aspect or feature ofa product or service,
 and the available polarities include: neutral, negative, positive.
- The 'opinion term' is the exact word or phrase in the text that refers to the sentiment or attitude expressed by a user towards a particular aspect or
feature of a product or service.Recognize all sentiment elements with their corresponding aspect terms, aspect categorys, sentiment polaritys,
 opinion terms in the following text in theform of a list of objects, each object having key(s) 'aspect term', 'aspect category', 'sentiment polarity',
 'opinion term'.Here are some examples:Text: This is a budget hotel in a good location central to everything .
Sentiment elements: [('aspect term': 'hotel', 'aspect category': 'hotel prices', 'sentiment polarity': 'positive', 'opinion term': 'budget'), ('aspect
term': 'hotel', 'aspect category': 'location general', 'sentiment polarity': 'positive', 'opinion term': 'good location')]
Text: it was very relaxing good buffet breakfast included and so close to everything .
Sentiment elements: [('aspect term': 'breakfast', 'aspect category': 'food_drinks quality', 'sentiment polarity': 'positive', 'opinion term': 'relaxing'),
('aspect term': 'breakfast', 'aspect category': 'food_drinks quality', 'sentiment polarity': 'positive', 'opinion term': 'good')]
Text: Loved our Spaulding Hotel Very good location , centrally located close to Theaters and shopping .
Sentiment elements: [('aspect term': 'Spaulding Hotel', 'aspect category': 'hotel general', 'sentiment polarity': 'positive', 'opinion term': 'Loved'),
('aspect term': 'location', 'aspect category': 'location general', 'sentiment polarity': 'positive', 'opinion term': 'Very good'), ('aspect term':
'Spaulding Hotel', 'aspect category': 'location general', 'sentiment polarity': 'positive', 'opinion term': 'centrally located close to Theaters and
shopping')]Text: Good restaurants nearby (
Sentiment elements: [('aspect term': 'NULL', 'aspect category': 'location general', 'sentiment polarity': 'positive', 'opinion term': 'Good restaurants
nearby')]Text: well thought out design to incorporate alot of extras .
Sentiment elements: [('aspect term': 'NULL', 'aspect category': 'rooms design_features', 'sentiment polarity': 'positive', 'opinion term': 'well thought
out design')]Text: There are a number of restaurants in walking distance .
Sentiment elements: [('aspect term': 'NULL', 'aspect category': 'location general', 'sentiment polarity': 'positive', 'opinion term': 'number of
restaurants in walking distance')]Text: The owners and staff are extremely accomodating , offering good advice to restaurants
Sentiment elements: [('aspect term': 'owners', 'aspect category': 'location general', 'sentiment polarity': 'positive', 'opinion term': 'extremely
accomodating'), ('aspect term': 'staff', 'aspect category': 'location general', 'sentiment polarity': 'positive', 'opinion term': 'offering good advice to
restaurants')]
Text: Outside is a magnificent view of the river for which the hotel is named and parallel is a cobblestone street filled with a wide range of charming
shops .
Sentiment elements: [('aspect term': 'hotel', 'aspect category': 'hotel design_features', 'sentiment polarity': 'positive', 'opinion term': 'magnificent
view of the river')]
Text: The only thing that the hotel could improve on was seating in the main lobby -  - there is a couple of chairs and a couch and alot of empty space .
Sentiment elements: [('aspect term': 'hotel', 'aspect category': 'facilities comfort', 'sentiment polarity': 'neutral', 'opinion term': 'improve on was
seating in the main lobby')]Text: great location close to amenities , but nice and quiet on an evening .
Sentiment elements: [('aspect term': 'location', 'aspect category': 'location general', 'sentiment polarity': 'positive', 'opinion term': 'great')]
Text: It is in a good location and close to alot of shops and restaurants .Sentiment elements:
```

图 4：**用于基于 RAG 的建议预测的提示词**。该提示词包含任务描述（其中解释了情感要素）、十个上下文示例，以及用于方面预测的目标文本。所示的少样本示例取自 Coursera 数据集，并包含 ASQP 标注