arXiv:2603.01212v1 [cs.CL] 1 Mar 2026

可解释人工智能增强的比较意见挖掘：基于方面的评分与语义推理

Ngoc-QuangLe[1⋆]、T. Thanh-Lam Nguyen[1⋆]、Quoc-Trung Phu[1]、Thi-Phuong Le[1,2]、Duy-Cat Can[1,3,4][0000−0002−6861−2893]以及 Hoang-Quynh Le[1][0000−0002−1778−0600]

[1] 越南国家大学工程与技术大学，越南河内纸桥郡春水路144号 22024510,22024516,21020096,phuong.lt,catcd,lhquynh@vnu.edu.vn https://uet.vnu.edu.vn[2] 新加坡管理大学计算机与信息系统学院，新加坡维多利亚街81号，邮编188065 tp.le.2023@phdcs.smu.edu.sg https://scis.smu.edu.sg[3] 沃州大学医院，瑞士洛桑布尼翁街46号，邮编1011 duy-cat.can@chuv.ch[4] 洛桑大学，瑞士洛桑UNIL-Centre区，邮编1015 duy-cat.can@chuv.ch https://www.unil.ch

摘要。 比较意见挖掘涉及比较不同评论中的产品。然而，为此任务设计的基于Transformer的模型往往缺乏透明度，这可能会妨碍用户信任的建立。本文提出 XCom，这是一种增强型基于Transformer的模型，由两个主要模块组成，即 *(i)*基于方面的评分预测，以及 *(ii)* 用于比较意见挖掘的语义分析。 XCom 还集成了一个Shapley加性解释模块，以提供有关模型审慎决策的可解释见解。从实证结果来看，XCom 的表现优于其他基线方法，展现了其提供有意义解释的有效性，使其成为更可靠的比较意见挖掘工具。源代码见： https://anonymous.4open.science/r/XCom。

关键词： 隐式比较意见挖掘 · 可解释人工智能 · 基于方面的情感分析 · Shapley值 · 可解释性

1 引言

在当今市场中，当用户经常面对过多选项时，可能会出现选择过载。评论往往传达个人体验，而非直接比较不同选项。如图1所示，显式比较会清楚地表明人们对不同产品的偏好，而隐式比较则往往分散在多条评论中，需要通过上下文推理来推断用户意图。尽管这类评论提供了有价值的信息，但缺少结构化比较内容

⋆ 这些作者对本研究的贡献相同。

会妨碍客户识别哪些方面最具影响力，或总体而言哪种产品表现更好。比较意见挖掘通过系统地分析和比较不同选项，弥合了这一差距 [16]。它有助于用户辨明关键差异，并做出更明智的决策。

同用户设定也是我们的核心目标，旨在减轻个体行为差异带来的噪声，具体包括语言风格、评分宽松度和个性化方面权重。设想一位“严格”的用户，他通常给出较低评分，撰写的总结也简短而克制；与之相对，“宽松”的用户则以评分较高、评论详尽而闻名。矛盾的是，前者给出的“不错”这一评分，往往意味着比后者给出的“惊艳”更高的质量。将比较限定在同一用户之内，就能消除这些不一致，并以稳定且个性化的基线来捕捉偏好。



图 1：用户评论中显性与隐性比较意见的示意图。**显性：**使用“味道比……更浓郁”等措辞直接进行比较。**隐性：**两条评论来自同一用户。偏好是暗示出来的，而非明言，因此更难检测。**跨用户设定中的信息不足：**两条评论来自不同用户。由于用户标准存在主观差异（例如宽松程度、写作风格），且缺乏用于推断的共同基线，因此比较并不可靠。

先进深度学习技术和大语言模型的日益广泛应用，为提升比较意见挖掘任务的性能带来了巨大潜力。然而，一个重大挑战依然存在：这些模型大多如同黑箱，在给出比较结论时并未提供背后的推理过程 [10]。由于难以理解特定决策背后的依据，这种透明度的缺失可能会削弱用户的信任 [9]。无法解释的比较陈述可能价值不大——尤其是在用户根据特定产品属性做出细粒度决策时。因此，必须具备解释能力，提供有意义的推理，以支持模型的决策。

为应对比较观点挖掘中的这些挑战，我们提出了 XCom，即**X**AI增强的**C**比较**O**观点**M**挖掘。XCom 是一种基于Transformer的模型，它结合了基于方面的评分预测技术和语义分析框架，以解决跨多个方面比较两条评论的问题。此外，我们通过整合

受 SHAP（Shapley加性解释）启发的可解释性阶段 [8]„ 突出显示影响模型决策过程的关键因素。

2 相关工作

Jindal 和 Liu（2006年） [5] 为比较句及其关系的挖掘奠定了基础，开启了识别和分析比较表达、评估产品表现的研究。Le 等人（2021）[7] 随后探索了利用模板或现有比较表达生成比较句的方法。通过综合比较数据，Shi 等人（2023年） [14] 和 Hasan 等人（2025年） [4] 根据用户生成内容判断哪种产品表现更好，为推荐系统和市场分析中的消费者偏好及产品排名提供了见解。

与此同时，Gao 等人（2024年） [3] 和 Le 等人（2024年） [6] 着重从比较句中提取信息。他们的方法通常会识别跨产品进行比较的具体特征（如价格、质量）；Zhao 等人（2024年） [17] 指出，这对电子商务应用尤其有价值。尽管这些方法在受控环境中效果显著，但其在不同产品领域的灵活性和可扩展性有限。同样，Echterhoff 等人（2023年）[2]能够针对单个产品生成比较句，却无法处理任意产品对，因此限制了其在动态现实场景中的应用。

尽管取得了这些进展，现有研究仍面临若干关键局限，例如依赖比较句、难以处理任意产品对，以及缺乏可解释性。我们的框架通过深入挖掘隐含的成对比较意见，并利用 SHAP 提升可解释性，缓解了这些问题。此外，尽管先前的研究已探索如何解释深度学习（Selvaraju et al. (2017) [12]）和基于Transformer的模型（Abnar and Zuidema(2020) [1] and Song et al. (2024) [15]），但将此类方法应用于复杂架构仍然具有挑战性。SHAP 通过分析模型输出变化所对应的词元贡献，为解决这一问题提供了途径。因此，无论模型的内部复杂程度如何，都可以应用 SHAP。

3 方法

图2 展示了所提出的模型，该模型包含三个主要阶段：基于方面的预处理、比较意见分类、预测与解释。

3.1 基于方面的预处理

基于方面的预处理阶段将评论对转换为结构化的、针对特定方面的句子对。该阶段通过规范化、句子切分、词元化和方面分类，对原始评论进行转换，以便进一步分析。

给定用户集合 $\mathcal{U}$ 和数据集 $\mathcal{D} = \{(R_1^u,\ R_2^u) \mid u \in \mathcal{U}\}$，每个比较实例 $(R_1^u,\ R_2^u)$ 由从同一用户 u 处选取的两条评论组成。每个

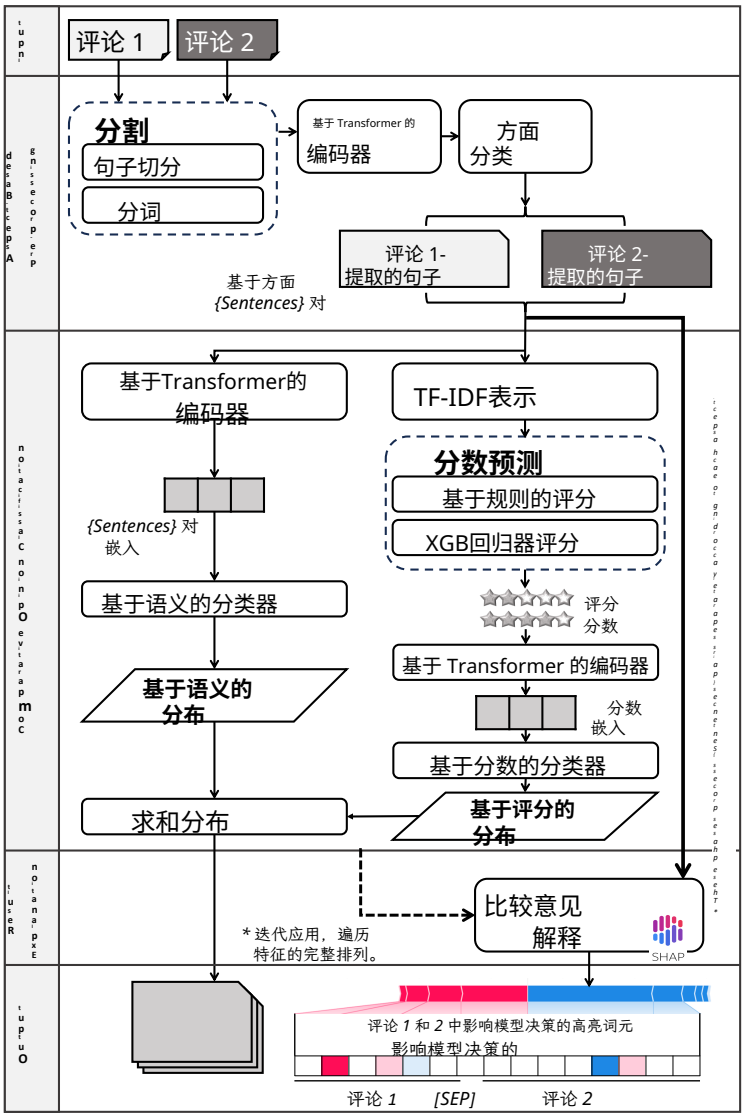


图 2：XCom 架构。

评论 R_i^u 包含若干句子 $R_i^u = [s_1^i, s_2^i, \ldots, s_{N_i}^i]$，其中 s_j^i 是评论 i 中的第 j 个句子，而 N_i 是句子的总数。

我们使用四个基于 BERT 的分类器 f_a，每个分类器都针对一个不同的方面$a \in \mathcal{A}$进行训练。每个分类器在句子对应方面 a时赋值为 1，否则赋值为 0。随后将分类后的句子分组进行比较 $P_a^u = (s_a^1, s_a^2)$，其中 s_a^1和 s_a^2 分别是从第一条和第二条评论中提取的、关于方面 a的句子。

最终输出是一组结构化的基于方面的句子元组：

$$\mathcal{P}^u = \{P_a^u \mid a \in \mathcal{A}\}. \tag{1}$$

3.2 比较意见分类

基于分数的分类器：在基于方面的预处理阶段之后，每个评论对被构造成四个特定方面元组。这些句子对用于计算各方面的评分。

对于给定的方面 $a \in \mathcal{A}$，从评论 R_1^u 和 R_2^u 中提取的句子构成两个集合：S_a^1 和 S_a^2。每个句子的**评分**通过对其形容词的情感分数取平均值来确定：

$$\text{rating score}\left(s_i^k\right) = \frac{1}{N_i^k} \sum_{j=1}^{N_i^k} \text{score}\left(w_j\right), \tag{2}$$

其中，N_i^k是句子 s_i^k中的形容词数量，score (w_j)是形容词 w_j的情感分数。如果 $N_i^k = 0$，即句子 s_i^k 中不包含预定义情感词典中的任何形容词，则计算其 TF-IDF 特征向量。随后将该向量输入特定方面的XGBoost回归器；该回归器经过训练，可在无法直接基于形容词评分时预测评分。预测的评分如下：

$$\text{rating score}\left(s_i^k\right) = f^{\text{XGB}}\left(\text{TF-IDF}\left(s_i^k\right)\right), \tag{3}$$

其中，f^{XGB} 将 TF-IDF 向量映射为评分。由于每个方面都包含多个句子，因此评论的总体方面评分由最低句子评分决定：

$$\text{rating score}\left(S_a^k\right) = \min_{i \in \{1,\ldots,M_a^k\}} \text{rating score}\left(s_i^k\right), \tag{4}$$

其中，M_a^k表示 S_a^k中的句子数量。

对于**分类**，将两条评论的方面评分拼接为结构化输入：$\text{input}^{\text{r}} = [\text{CLS}]$ 评分 (S_a^1) $[\text{SEP}]$ 评分(S_a^2) $[\text{SEP}]$。其中，$[\text{CLS}]$ 是分类标记，$[\text{SEP}]$ 是分隔符。基于 Transformer 的编码器提取评分嵌入：$\mathbf{z}^{\text{r}} = \mathcal{E}^{\text{r}}(\text{input}^{\text{r}})\ |_{\text{CLS}}$，其中$\mathbf{z}^{\text{r}} \in \mathbb{R}^{768}$ 是 CLS 标记的 logits向量。最后，基于分数的分类器处理$\mathbf{z}^{\text{r}}$，以生成三个比较类别上的分布：

$$\mathbf{l}^{\text{r}} = \mathbf{z}^{\text{r}} \cdot \mathbf{W}^{\text{r}} + \mathbf{b}^{\text{r}}, \tag{5}$$

其中，$\mathbf{W}^{\text{r}} \in \mathbb{R}^{768\times 3}$ 是权重矩阵，$\mathbf{b}^{\text{r}} \in \mathbb{R}^3$ 是偏置向量。

基于语义的分类器：此步骤直接处理句子对以生成嵌入，从而捕获语义关系。结构化输入为：$\text{input}^{\text{s}} =$

[CLS] (S_a^1) [SEP] (S_a^2) [SEP] [PAD]。基于 Transformer 的编码器提取句子对嵌入：$\mathbf{z}^{\mathsf{s}} = \mathcal{E}^{\mathsf{s}}(\text{input}^{\mathsf{s}})|_{\text{CLS}}$，其中$\mathbf{z}^{\mathsf{s}} \in \mathbb{R}^{768}$。基于语义的分类器处理$\mathbf{z}^{\mathsf{s}}$，以生成基于评分的分布：

$$\mathbf{l}^{\mathsf{s}} = \mathbf{z}^{\mathsf{s}} \cdot \mathbf{W}^{\mathsf{s}} + \mathbf{b}^{\mathsf{s}}, \tag{6}$$

其中$\mathbf{W}^{\mathsf{s}} \in \mathbb{R}^{768\times 3}$ 是权重矩阵，$\mathbf{b}^{\mathsf{s}} \in \mathbb{R}^3$ 是偏置向量。

比较分类：为获得最终的概率分布，我们首先对基于分数和基于语义的 Logits（$\mathbf{l}^{\mathsf{r}}$ 和$\mathbf{l}^{\mathsf{s}}$）应用 Softmax，然后将得到的概率向量相加：

$$\mathbf{p}^{\mathsf{r}} = \text{softmax}(\mathbf{l}^{\mathsf{r}}), \quad \mathbf{p}^{\mathsf{s}} = \text{softmax}(\mathbf{l}^{\mathsf{s}}),$$
$$\mathbf{p}^{\mathsf{final}} = \mathbf{p}^{\mathsf{r}} + \mathbf{p}^{\mathsf{s}}.$$

ˆ 预测类别 C 对应于最高的概率值：

$$\hat{C} = \arg\max_k \left(\mathbf{p}^{\mathsf{final}}\right)_k, \quad k \in \{1,2,3\} \tag{7}$$

3.3 基于 SHAP 的解释

我们使用 SHAP [13] 通过归因特征重要性来增强模型透明度。它通过测试不同的子集并观察其对预测结果的影响，量化每个特征的贡献。SHAP 通过揭示哪些词元驱动XCom的比较预测，为模型提供可解释性，从而将其从黑箱转变为白盒系统。换言之，SHAP 解释旨在揭示模型内部的决策归因，而非验证与人类认知一致的推理。

SHAP 用可解释线性函数近似模型 f：

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \ , \tag{8}$$

其中 $g(z')$ 近似 $f(x)$，ϕ_j 是特征 j 的夏普利值，$z' \in \{0,1\}^M$ 表示特征存在（$z'_j = 1$）或缺失（$z'_j = 0$）。预测ϕ_0 是模型的平均输出。每个特征的夏普利值计算如下：

$$\phi_i = \sum_{\mathcal{S}\subseteq\mathcal{N}\setminus\{i\}} \frac{|\mathcal{S}|!(|\mathcal{N}|-|\mathcal{S}|-1)!}{|\mathcal{N}|!} [f(\mathcal{S}\cup\{i\}) - f(\mathcal{S})] \tag{9}$$

其中 $\mathcal{N}$ 是所有特征的集合，$\mathcal{S}$ 是排除特征 i 的子集。$f(\mathcal{S})$ 是使用 $\mathcal{S}$ 中特征时的模型输出，而 $f(\mathcal{S}\cup\{i\})$ 包含 i。

在我们的案例中，输入由词元组成，因此每个词元都被视为一个特征。SHAP 通过将某些词元替换为 [MASK] 词元来选择不同的子集，从而有效地将其从输入中移除。随后，它会衡量剩余词元对模型预测结果的影响。

4 实验结果

4.1 数据集

鉴于该任务具有新兴性和挑战性，同时也由于难以获得高度专业化的标注，我们的实验在单个数据集——SUDO 上进行 [11]。在 SUDO 中，每位用户会提供多条评论，然后将这些评论配对，以创建用于隐式观点挖掘的比较实例（有关数据集构建方法的详细信息，请参见[11] ）。每条评论都包含对四个方面的评分： $\mathcal{A} = \{$外观、香气、口感、味道$\}$。该数据集在两个层面进行了标注：

1. **句子-方面级**：如果句子提及某个特定方面，则标记为 1（提及）；否则标记为 0（未提及）。2. **评论比较级**：评论对中的每个方面均分别标记为：-1（较差——第一条评论在该方面的评分低于第二条评论）、0（相似或无法比较）、1（较好——第一条评论在该方面的评分高于第二条评论）或 Null（未涉及该方面）。

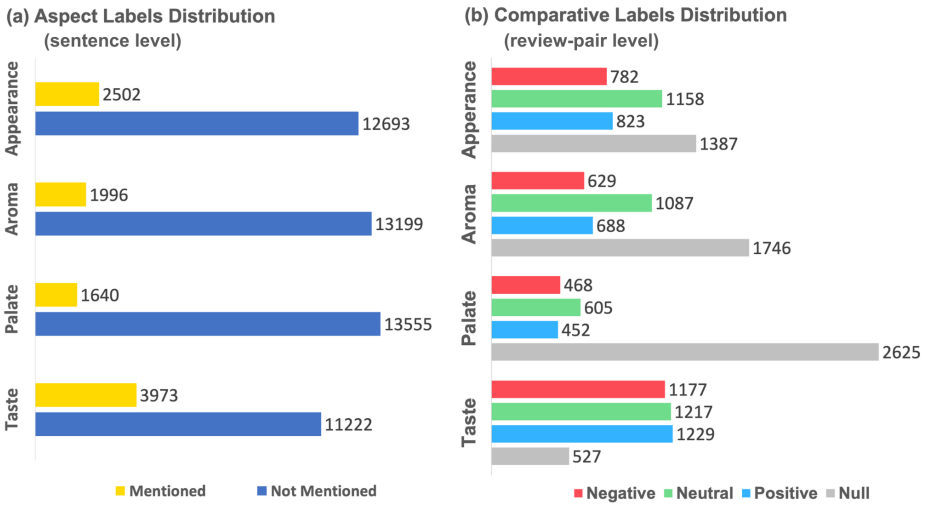


图 3：方面和比较标签的分布。

在句子级，图 3 (a) 展示了各方面类别分布存在显著失衡。未提及某个方面的句子数量远远超过提及该方面的句子。这凸显了在评论中准确识别相关方面的挑战，因此需要采用稳健架构来应对。

在评论级，图 3 （b）显示 `Null` 标签在各各方面中占比很高：味道的比例最低，凸显其在啤酒评价中的核心作用，而 口感 的比例最高，表明其较少被讨论。在非空对中，情感分布显示出

略微倾向于中性，表明当某一方面同时出现在两条评论中时，相关情感往往是一致的。总体而言，这些显著的集中分布给开发有效的监督模型带来了挑战。

4.2 评估指标

我们使用宏平均和微平均下的精确率、召回率和 F1 分数来评估模型性能。宏平均先计算各类别的分数，再取平均；微平均则汇总所有实例进行计算。两种方法均应用于分类类别，以确保评估均衡。

4.3 模型性能

表1：模型总体性能与比较。

模型	微平均			宏平均		
	精确率	Rec.	F1	精确率	Rec.	F1
XCom	**60.51**±0.8	56.40±0.7	**58.38**±0.9	**61.92**±1.0	56.11±0.7	**58.56**±0.8
基于特征的基线方法						
FastText+支持向量机	52.57±1.0	49.15±0.8	50.08±1.1	52.83±0.9	49.76±0.8	51.05±1.0
FastText+XGBoost	52.03±1.1	42.28±1.0	46.65±0.8	53.65±1.0	41.38±0.7	46.22±1.1
基于 Transformer 的基线						
微调后的 BART	49.70±0.9	**61.49**±1.0	54.97±0.9	50.16±1.1	**62.74**±1.0	55.72±1.1
微调T5	55.88±0.7	57.72±0.9	56.79±0.8	55.24±1.0	56.53±0.9	55.86±0.9
通用大语言模型基线						
Llama-3.2-8B-Instruct	24.86±0.9	33.90±0.8	28.69±1.0	25.62±0.7	34.69±1.1	25.61±0.9
Gemini-2.5-Flash	46.06±0.8	48.96±0.9	47.47±0.7	48.97±0.7	53.39±0.6	44.94±0.8
ChatGPT-5.1	47.28±0.0	50.33±0.0	48.76±0.0	47.61±0.0	54.33±0.0	47.24±0.0

注：最佳值以粗体标示，次佳值带下划线。

表 1比较了 XCom与七种基线方法在 宏平均F1 和 微平均F1上的表现，展现了 XCom的稳健性，其在所有指标上均取得最高性能（宏平均F1约为 58.5% 宏平均F1）。基于特征的模型（支持向量机、XGBoost）均使用 fastText词嵌入作为输入特征，难以捕捉复杂的上下文依赖关系；相比之下，基于 Transformer 的基线（微调T5 和 BART）表现更具竞争力。然而，XCom 比排名第二的模型（微调T5）高出近 3 个百分点，验证了我们架构的有效性。通用大语言模型（Llama-3.2-8B-Instruct、Gemini-2.5-Flash、ChatGPT-5.1）的表现明显较低，其微平均F1 得分介于 28% 至49%之间。这进一步说明了我们方法的必要性，因为细粒度的比较

表2：XCom的详细模型结果。

XCom	微平均			宏平均		
	精确率	Rec.	F1	精确率	Rec.	F1
(A) 按方面划分的性能						
外观	61.28±0.9	60.59±0.8	60.93±0.7	63.58±1.0	59.98±0.9	61.18±0.8
香气	60.93±0.8	55.27±0.9	57.96±1.1	61.43±0.9	53.57±0.8	56.72±0.7
口感	56.25±1.0	44.75±1.1	49.85±0.9	64.56±1.1	42.79±0.7	48.23±0.8
味道	61.37±0.7	59.73±0.8	60.54±0.6	61.81±1.0	60.71±0.9	60.98±0.8
（B）分项性能						
方面分类	95.38±0.8	95.38±0.7	95.38±1.0	91.69±1.1	91.10±0.9	91.39±0.8
比较观点	61.21±0.9	61.21±1.0	61.21±0.8	62.58±1.1	60.91±0.7	61.49±0.9
总体	60.51±0.8	56.40±0.7	58.38±0.9	61.92±1.0	56.11±0.7	58.56±0.8

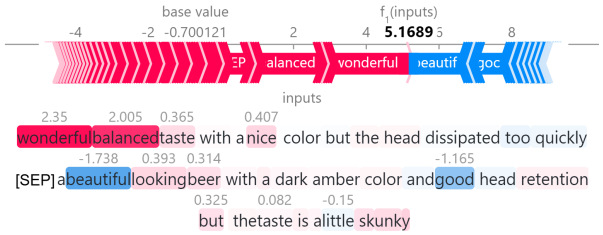
意见挖掘需要结构化理解，而零样本提示无法提供这种理解。

此外，XCom 以显著更低的计算成本取得了这些结果。它基于基于BERT的编码器构建（每个约有 1.1 亿个参数），规模远小于拥有数十亿参数的大语言模型。在我们的实验中，在相同的硬件配置下，训练 XCom 耗时约 10 小时，而由于模型规模和内存占用大得多，对大语言模型进行微调通常需要约 40–80 小时。这凸显了 XCom 的计算效率：它以更低的资源消耗实现了具有竞争力的性能。

表2 展示了所提出模型在四个方面上的性能。外观和味道取得了最高的F1分数，超过60%，表明这些方面包含更清晰、更具区分性的语言线索。相反，口感因口感相关表达微妙且相互重叠而得分最低，而香气表现中等，这受其与味道在词汇上较为接近的影响。分组件结果表明，虽然方面分类模块高度可靠（超过90%，无论是宏平均F1还是微平均F1），后续的比较观点抽取降至约60 % ，在宏平均F1和微平均F1上均如此。因此，这些错误的综合效应通过级联错误传播导致总体结果下降。这些方面级和组件级错误模式以及说明性示例，将在第5.2节中加以考察。

4.4 解释结果

图 4展示了两款产品的两条评论在两个基于方面的比较（味道 和外观）中的解释。对于味道 方面（图4a），第一款产品（评论 1）的评价高于第二款产品（评论 2）。词元‘美妙’和‘均衡’对模型输出产生了非常积极的影响



(a) 口味方面对比的解释示例。



(b) 外观方面对比的解释示例-对比 n.

图4：基于 SHAP 的解释示例，展示口味和外观方面。在我们的基于 SHAP 的可视化中，红色词元表示对模型输出有正向影响，蓝色词元表示有负向影响。箭头长度表示 SHAP 值的绝对值。下方高亮的文本通过标记原文中的关键词元，提供更直观的解释。颜色越浓，表示影响越强。

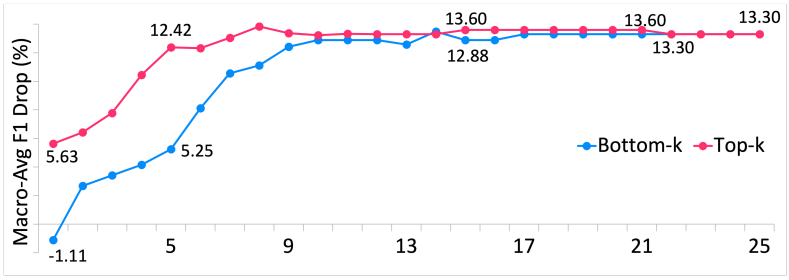


图5：SHAP 忠实度评估：移除前 k 个形容词与后 k 个形容词时宏平均 F1 的下降。

。相反，尽管‘美丽’通常表达积极的情感，但在味道 方面的语境中，它却会产生负面影响，因为它通常不用于描述味道，因此被视为一种误导性特征。对于相同的输入评论，在外观 方面，第一款产品的评价低于第二款产品（图4b）。我们的模型正确识别出‘美丽’指的是啤酒的外观，因此产生了强烈的积极影响。与此同时，‘美妙’的影响较弱，进一步表明模型在使用适当的

.....数据。移除排名靠前的-k个词元与排名靠后的-k 个词元时，概率下降幅度形成鲜明对比，这支持了这一发现。

如图 5所示，我们还通过测量移除按重要性排序的形容词词元后宏平均F1值 的下降幅度，评估了 SHAP 的忠实度。Top-k曲线显示性能立即下降，表明 SHAP 能正确识别预测力最强的词语；而Bottom-k曲线缓慢上升，证实排名靠后的形容词对模型决策影响极小。词元重要性通过对两个输入句子中各词元在所有情感类别上的 SHAP 值绝对值取平均来确定。两条曲线在 $k = 25$处汇合，该值代表任一单独测试样本中形容词数量的最大值；达到这一阈值时，两种策略都会移除所有形容词，留下相同的残余输入，因此相对于基线产生相同的性能下降。

5 讨论

本节通过消融研究量化各个组成部分的贡献，并分析关键失败案例——包括方面歧义、隐含对比和句法复杂性——以阐明当前局限并指导未来改进。

5.1 模型组件贡献分析

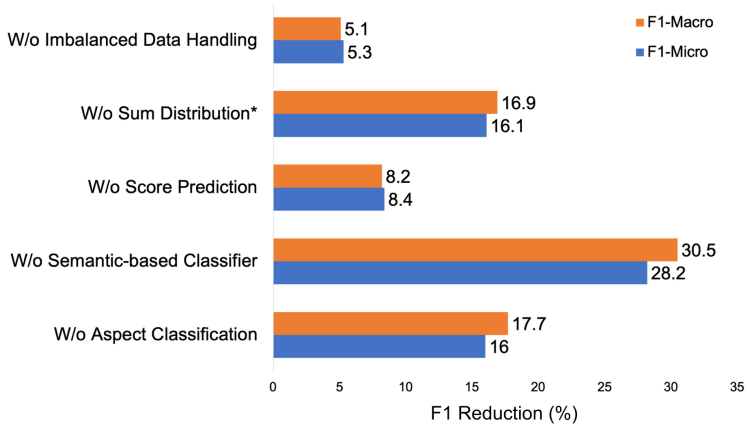


图6：消融研究显示移除每个模型组件时的F1下降。

我们通过消融研究评估每个组件的贡献，系统地移除它们以衡量性能下降。图6 表明

每个组件均带来性能增益，但其影响程度有所不同。基于语义的分类器是识别比较关系的主要驱动力，移除后两项F1指标均急剧下降约30%。方面分类同样不可或缺；若缺少该组件，模型将失去筛选相关观点目标的能力，性能下降近17%。有趣的是，若以单一分类器替代求和分布策略，对拼接后的基于语义和评分的嵌入进行处理，性能下降超过16%——甚至大于移除整个评分模块，这表明不当融合会损害表示有效性。其他消融实验也会降低性能，但程度较轻。

5.2 错误分析

表3：XCom的错误分析与代表性案例。

错误类型	**原因**	**代表性示例**
过度抽取方面相关的句子	模型泛化超出预期范围。	“这款啤酒非常容易入口”——一般性陈述，并非真正针对某个特定方面。
提取不足	方面指代是间接表达的，或依赖语境，而非明确表达。	“我时不时会买它，当我想要来点有格调、**顺口**的东西时**口感顺滑又美味**”——遗漏了隐含方面。
方面混淆	一个词可能指代多个方面类别。	“**啤酒花**处于背景中，但仍可察觉”——在香气和味道之间混淆。
方面评分过估计	词汇复杂度和隐含对比共现方面。	“**泡沫**令人惊艳，让这款啤酒几乎看起来和**品尝**起来一样美好（嗯，倒也没那么好，不过你懂我的意思）”——模型高估了味道所表达的情感。
错误的评分预测	未能正确解读对比性表达。	“它很甜，但**不腻**。事实上，甜味时隐时现……”——模型预测得分较低，未能识别“不腻”。
词汇表外word	该词未被收录在评分词典中——性。	“鼻息中带有**accentuated**一丝丁香的气息”——“Accentuated”未收录于词典中，导致评分错误。

尽管总体表现有效，XCom 模型仍存在多种错误类型，指出了有待改进之处。表3 总结了这些错误、其

可能原因及代表性示例。一个常见问题是过度抽取方面相关句子，即模型纳入了一些泛泛而谈或与特定方面无关的句子（例如：“这款啤酒非常容易入口”）。相反，当方面相关信息是隐含表达而非明确表达时，就会出现抽取不足，导致模型忽略相关线索。例如，“入口顺滑，味道也好”这一短语暗示了口感或味道等方面，但模型未能识别出来。另一项挑战是方面歧义，即某个词（例如“啤酒花”）可能指向多个方面，如味道或香气。对于通过复杂措辞和微妙对比表达多个方面的句子，模型也难以处理；由于其中一个方面具有强烈的正面情感，模型可能会因此高估另一方面的评分。在评分预测中，模型有时会误解“它很甜，但不会甜腻”这类对比表达，从而导致评分预测错误。此外，评分词典中未收录“accentuated”等词，模型无法识别并解读其评分含义，因此可能会给出不准确的分数。这些错误表明仍需进一步改进，包括加强上下文理解、提升模型处理复杂句法结构的能力，以及扩大词汇覆盖范围。

6 结论

本文介绍了 XCom，这是一种比较意见挖掘模型，利用了多个意见挖掘模块，例如方面分类、基于方面的评分预测和比较分类。在第一个模块——方面分类中，我们使用针对每个方面训练的专用基于 Transformer 的分类器，将成对评论中的句子归入结构化的方面句子元组集合，以便进行比较。在第二个模块——基于方面的评分预测中，有两个关键分类器：基于分数的分类器和基于语义的分类器。在最后一个模块——比较分类中，上述两个分类器都会计算概率分布。最后，模型将这两个概率分布相加，并选择概率最高的类别作为预测的比较意见。此外，通过基于 SHAP 的解释，XCom 提升了模型的透明度，超越了传统的黑盒模型。结果表明，我们的模型能够提供可靠且可解释的比较洞见，助力更明智的决策，同时保持较高的计算效率。

局限性 尽管取得了令人鼓舞的结果，我们的方法仍存在若干局限性。下面，我们将讨论这些局限性，并概述未来改进的潜在方向。

首先，我们的模型依赖多个意见模块（即方面分类、基于方面的评分预测和比较分类器）；如果某个模块将错误的预测结果传递到后续阶段，可能会导致级联错误。未来的研究可以探索联合学习框架，以实现端到端优化。

14 Leet 等人。

其次，虽然我们基于 SHAP 的可解释性模块能够揭示影响模型决策的关键因素，但非专业用户并不总是能够轻松理解其结果。未来还需进一步研究如何提升解释的可读性和易用性，尤其是在多方面比较场景中。

第三，我们的 SHAP 解释主要突出重要术语，但并未在不同评论的关键术语之间建立直接比较。虽然这有助于识别每条评论中的影响因素，却无法对不同意见进行结构化对齐。我们正在积极探索一种图式解释方法，直接连接两条评论中的关键术语，以期提供更具结构性和比较性的分析。

第四，SUDO 数据集的规模和领域多样性有限。虽然 SUDO 为啤酒评论领域的隐式比较意见挖掘提供了针对性评估，但要验证其普适性，还需扩展至其他产品类别。扩大实验范围，并将数据集扩展至更多样的领域和用户视角，有助于提升研究发现的普适性。

第五，我们的数据调查显示，用户评论中不同方面之间往往存在相互依赖关系。具体而言，对某一方面给予正面评价的用户，往往也会对其他方面作出类似评价，或表现出正面偏差。我们计划探索跨方面依赖关系，以提升模型性能，并通过纳入这些相互关系来增强解释部分。

最后，我们的模型缺乏专门用于检测因非正式写作风格或俚语而产生的矛盾或误导性评论的机制，而这正是社交领域文本处理中的常见挑战。未来的研究可以探索在分析前采用语言规范化技术，对非正式语言和俚语进行标准化。另一方面，利用基于社交媒体文本的大规模预训练模型，或可提高模型应对用户生成内容中常见非正式语言模式的稳健性。

通过解决这些局限性，我们相信未来的研究能够进一步提升 XCom 在现实世界比较意见挖掘任务中的有效性、可解释性和适用性。

参考文献

1. Abnar, S., Zuidema, W.：量化 Transformer 中的注意力流。在：第58届计算语言学协会年会论文集 2. Echterhoff, J., Yan, A., McAuley, J.：苹果对苹果：从用户评论生成关注方面的比较句。arXiv预印本 arXiv:2307.03691 3. Gao, F., Liu, Y., Fu, W., Zhang, M., Ballard, A., Zhao, L.：将比较意见五元组抽取作为二分集合预测，并采用动态结构剪枝，实现端到端抽取。《Expert Systems with Applications》**245**, 123058（2024年）4. Hasan, E., Rahman, M., Ding, C., Huang, J.X., Raza, S.：基于评论的推荐系统：方法、挑战与未来展望综述 5. Jindal, N., Liu, B.：比较句与关系挖掘。在：第21届人工智能全国会议（AAAI）论文集（2006年）

6. Le, H.Q., Can, D.C., Nguyen, K.V., Tran, M.V.：VLSP 2023–CoMoM共享任务概述：一项基于越南语产品评论的比较意见挖掘数据挑战。arXiv预印本 arXiv:2402.13613（2024年）7. Le, T.H., Lauw, H.W.：基于产品属性比较约束的可解释推荐。在：第14届ACM网络搜索与数据挖掘国际会议论文集。第967–975页（2021年）8. Lundberg, S.：一种解释模型预测的统一方法。arXiv:1705.07874（2017年）9. Mersha, M., Lam, K., Wood, J., AlShami, A., Kalita, J.：可解释人工智能：需求、技术、应用与未来方向综述。《神经计算》第128111页（2024年）10. Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N.：可解释人工智能：全面综述。《人工智能评论》第1–66页（2022年）11. Nguyen, T.L.T., Le, N.Q., Phu, Q.T., Le, T.P., Pham, N.H., Nguyen, P.N., Le, H.Q.：不言而喻的比较：用于从同一用户评论中进行隐式比较意见挖掘的数据集与基准（2026年）12. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.：Grad-cam：通过基于梯度的定位，为深度网络提供可视化解释。在：IEEE国际计算机视觉会议论文集。第618–626页 13 – . Shapley, L.S.：n人博弈札记——II：n人博弈的价值。兰德公司（1951年）14. Shi, J., Zhu, Y., Wang, K., Jing, M., Zang, T., Yu, J., Tang, F.：用于基于评论的推荐的评分-评论图对比学习。在：2023年IEEE第29届并行与分布式系统国际会议（ICPADS）15. Song, L., Cui, Y., Luo, A., Lecue, F., Li, I.：通过凸显重要信息，更好地解释 Transformer。在：计算语言学协会欧洲分会年会（EACL 2024）论文集。第2048–2062页（2024年）16. Varathan, K.D., Giachanou, A., Crestani, F.：比较意见挖掘：综述。《信息科学与技术协会期刊》68829 17 4 811( ), – . Zhao, Z., Zhang, N., Xiong, J., Feng, M., Jiang, C., Wang, X.：提升电子商务推荐：利用BERTFusionDNN从客户评论中发掘洞见。《工程科学理论与实践期刊》4(2), 38–44（2024年）