带有置信加权微调的语义保留增强用于方面类别情感分析

Yaping Chai, Haoran Xie, Joe S. Qin

***Abstract*—大型语言模型（LLM）是一种有效解决低资源情境下数据稀缺问题的方法。最近的研究设计了手工制作的提示词，引导 LLM 进行数据增强。我们为属性类别情感分析（ACSA）任务引入了一种数据增强策略，该策略保留了原始句子语义并具有语言多样性，具体通过为 LLM 提供结构化的提示模板以生成预定义内容。此外，我们采用后处理技术进一步确保生成句子与原始句子之间的语义一致性。增强的数据增加了训练分布的语义覆盖率，使模型能更好地理解属性类别与情感极性之间的关系，增强其推理能力。此外，我们提出了一种置信度加权微调策略，鼓励模型生成更有信心且准确的情感极性预测。与强大且最近的工作相比，我们的方法在四个基准数据集的所有基线中始终实现最佳性能。**

***Index Terms*—Aspect Category Sentiment Analysis, Large Language Models, Data Augmentation, Confidence-weighted Fine-tuning**

I. 引言

基于方面的情感分析（ABSA）因其可以提供更细粒度的情感表达能力，受到了自然语言处理（NLP）领域研究人员的广泛关注。ABSA 的目标是识别实体的特定方面并确定与每个方面相关的情感极性。这一粒度级别对于电子商务、社交媒体和客户反馈分析等应用至关重要。ABSA 包含多个子任务，其中三个是方面类别检测（ACD）、方面类别情感分类（ACSC）和方面类别情感分析（ACSA）。ACD 涉及识别给定文本中提到的预定义方面类别。ACSC 通过为每个检测到的方面类别分配一个情感极性来扩展 ACD。ACSA 则是一个端到端任务，它结合了 ACD 和 ACSC，联合预测方面类别及其对应的情感极性。图 1 展示了每个子任务的一个例子。

ABSA 的一个挑战是缺乏高质量标注的训练数据 [1], [2], [3], [4] 。然而，人工标注既耗时又需要大量的人力资源。由于大语言模型（LLM）展示了显著的文本生成能力，越来越多的研究开始利用 LLM 进行数据增强，以解决在资源匮乏的情况下的 ABSA 任务。最近关于 ABSA 数据增强的工作 [3], [4] 生成了大量数据，但无法保证合成句子与原句之间的语义一致性。其他工作，例如 [2], [1] ，通过替换词语获得增强数据，但它们大多保留了原句结构，限制了生成数据的表达多样性。此外，[1] 引入了语义一致性，发现保留标注数据语义的生成数据集对于提高模型性能有益。

This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (R1015-23), and the Faculty Research Grant (SDS24A8) and the Direct Grant (DR25E8) of Lingnan University, Hong Kong. *(Corresponding author: Haoran Xie.)*

Yaping Chai, Haoran Xie, and Joe S. Qin are with the School of Data Science, Lingnan University, Hong Kong (e-mail: yapingchai@ln.hk; hrxie@ln.edu.hk; joeqin@ln.edu.hk).

基于上述研究结果，我们旨在解决 ACSA 任务中的数据稀缺问题。具体而言，我们设计了特定的提示模板，引导 LLMs 生成在语义上与原始文本显著不同的数据，同时保留其语义。为了确保合成数据的质量，我们利用后处理技术来过滤掉语义上过于偏离原始句子的数据。此外，微调是一种有效的方法，可以在保留预训练阶段的通用知识的同时，使预训练模型适应下游任务。我们提出了一种置信加权微调方法，其核心思想是在模型做出正确预测时，基于预测的置信水平给予额外权重，从而鼓励模型做出更加自信和准确的预测。

总之，我们工作的主要贡献如下：

- 我们介绍了一种新颖的数据增强方法，该方法利用 LLMs 在保持原句语义的同时引入语言多样性。
- 我们提出了一种加权信心的微调策略，以促进模型做出更自信和准确的预测。
- 我们对实验设置中的关键超参数变化进行了全面分析。我们提出的方法在 ACSC 和 ACSA 任务的所有四个基准数据集上都始终实现最佳性能。

II. 相关工作

A. 方面类别情感分析

早期的研究将 ACSA 视为一个分类任务。AAGCN [5] 通过 softmax 层的概率分布输出获得最终的情感分类结果。AC-MIMLLN [6] 通过聚合关键实例的情感来分类句子中涉及的方面类别的情感。PBJM [7] 将情感极性预测视为一个二元分类任务，并利用自然语言提示模板执行方面类别检测 (ACD) 和方面类别情感分类 (ACSC) 任务。

为了克服传统分类的局限性，传统分类无法直接利用预训练知识，也无法更好地推广到新领域，BART-generation [8] 首次提出将 ACSA 视为一个文本生成任务。这样可以更好地利用 BART [9] 在输入语义层面总结方面的优势。PyACSA [10] 开发了一个 ACSA 框架，并使用 Flan-T5 [11] 作为生成模型，将 ACSA 任务转换为文本到文本的格式。

最近的研究者开始探索大型语言模型（LLMs）在情感分析中的能力，并使用它们来统一 ABSA 的各种子任务。LEGO-ABSA [12] 提出通过控制由多个元素组成的任务提示类型，在一个统一的生成框架中解决不同的 ABSA 任务。ChatABSA [13] 是一个统一的框架，可以将五个复杂的子任务转换为提示格式。

B. 基于提示的技术

提示工程作为指导语言模型输出预期内容的一种技术，自从早期预训练模型如 BERT [14] 的引入以来，已经经历

c1: ambience c2: food

Sentence **S**: *while it was large and a bit noisy, the drinks were fantastic, and the food was superb.*

s1: negative s2: positive

Sub-task	Input	Output
Aspect Category Detection (ACD)	$\boldsymbol{S}$	$\{c_1,c_2\}$
Aspect-Category Sentiment Classification (ACSC)	$\boldsymbol{S}$,c_1	$\{s_1\}$
	$\boldsymbol{S}$,c_2	$\{s_2\}$
Aspect Category Sentiment Analysis (ACSA)	$\boldsymbol{S}$	$\{(c_1,s_1),(c_2,s_2)\}$

Figure 1: ACD、ACSC 和 ACSA 子任务的示例。c、s 和 S 分别表示方面类别、情感倾向性和输入句子。

了显著的变化，它具有完形填空风格的提示，并影响了后来的填空任务的发展。

随着大型语言模型的巨大突破，基于提示的学习已成为扩大大型语言模型能力的革命性技术。[15] 已经证明，零样本和少样本提示显著影响模型性能。

诸如链式思维提示 [16] 和自洽性提示 [17] 等技术通过显式构建多个步骤进一步增强了 LLM 的推理能力。在 ABSA 任务中，RVISA [18] 通过为 LLM 提供推理提示来解决隐式情感分析，从而获得令人信服的理由。

C. 数据增强

最早广泛使用的数据增强方法包括替换同义词、打乱词序和随机删除单词 [19] 。

随着大型语言模型 (LLMs) 的蓬勃发展，许多研究利用基于提示的技术进行数据增强，通过设计精心制作的提示解决数据稀缺问题。在这些研究中，在 ABSA 任务中，RSDA [3] 在跨领域任务中表现良好，但未考虑生成句子与原始句子相比语义偏差过大的问题。SPDAug-ABSA [2] 则提出通过考虑文本序列中每个词的重要性来保留原始句子的语义。然后，他们使用两种替换策略来替换不重要的标记。然而，这仍然是一个标记级转换，不能为整个句子带来多样性。

为了解决情感分析中数据增强的上述限制，我们利用 LLM 的卓越生成能力来生成富有表现力的句子，同时确保与原句在语义上的一致性。

III. 方法

为了通过数据增强获得与原句语义一致且有表达差异的句子，我们保留了原始句子的类别及其对应的情感极性，设计了一个用于 ACSA 任务的提示模板，并引导 LLM 根据预定义的提示模板生成句子。为了确保数据质量，我们使用后处理机制过滤掉与原句语义偏离严重的样本。最终，我们获得了一个包含类别-极性对的合成数据集。此外，微调在通用预训练模型和特定应用的独特需求之间建立了桥梁。我们提出了一种新的微调策略，引入了置信加权损失函数。其核心思想是在模型做出正确预测时，根据预测的置信度提供额外的权重，从而鼓励模型做出更自信和准确的预测。整体架构如图 2 所示。

A. 任务定义

方面类别检测（ACD）的目标是识别与给定文本关联的预定义方面类别 [8] 。与识别明确提到的方面词项的方面词项提取不同，ACD 专注于将文本中讨论的隐含或明确的方面进行分类。类别是预定义的，并特定于领域。给定一个输入句子 S ，目标是预测一组方面类别 C ，其中 $C = \{c_i\}_{i=1}^{m}$ 是由所有 c_i 组成的类别集，c_i 代表句子中的一个类别，而 m 是句子中所有类别的总数。

方面类别情感分类（ACSC）通过为每个检测到的方面类别 c_i 分配一个情感极性 s_i 来扩展 ACD。情感极性可以分为三个主要类别：积极、中性和消极。

方面类别情感分析（ACSA）是结合了 ACD 和 ACSC 的端到端任务。其目标是联合预测方面类别及其对应的情感极性。给定一个输入句子 S ，目标是预测一组类别-情感对 $\{c_i, s_i\}_{i=1}^{m}$ 。

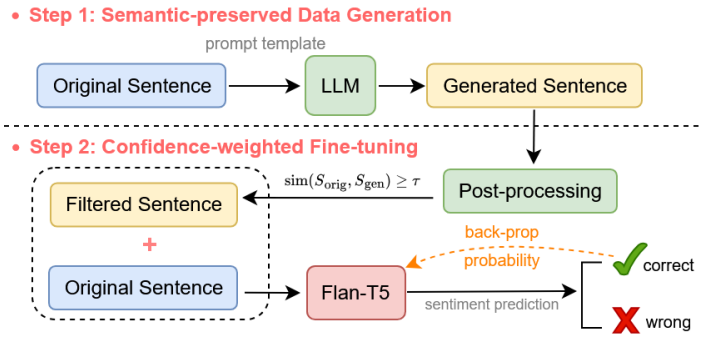

Figure 2: 我们方法的概述，包括数据增强、后处理和置信度加权微调。

B. 语义保留增强

1) 增强句子生成： ACSA 的挑战之一是数据匮乏，特别是由于需要对多个类别及其相关情感的句子进行细粒度标注。然而，收集和标注大规模高质量数据集既昂贵又耗时 [19] 。传统的数据增强方法，如同义词替换和回译，无法保持原句的语义或缺乏语言多样性。为了解决这些限制，我们利用 LLMs 进行文本数据增强，以生成多样化且语义一致的句子。用于句子生成的结构化提示模板如图 3 所示。

给定一个标记样本 $x = (S_{orig}, y)$ ，其中 S_{orig} 表示原始句子，$y = (c_i, s_i)_{i=1}^{n}$ 是由方面类别 c_i 和情感极性 s_i 组成的 n 类别-情感对的集合。我们保留原始句子的所有 $(c_i, s_i)_{i=1}^{n} \in y$ ，并明确指导模型生成一个富有表达力且在构建的提示模板中保留原始句子意义的合成句子 S_{gen} 。将

由 LLM 生成的 S_{gen} 与原始标签 y 配对，以形成一个增强样本 $x' = (S_{gen}, y)$ 。最后，我们获得了一组生成的句子及其类别-极性对。

```
PROMPT_TEMPLATE = """You need to generate a sentence that is
significantly different from the original sentence while preserving all
aspect categories and their associated sentiments.

Original sentence: {text}
Original aspect categories and sentiments:
{categories_with_sentiments}

Requirements:
1. The generated sentence MUST contain ALL of the above aspect
categories with the SAME sentiments
2. Make the expression style completely different from the original
3. Maintain natural language fluency
4. Only respond with the generated sentence, no other text

Generated sentence:"""
```

Figure 3: 用于生成句子的结构化提示模板。

2) 语义一致性过滤： LLM 生成的数据不能完全保证高质量，因此采用精细化策略来增强数据有助于模型训练 [19] 。为了进一步确保生成数据集的质量，我们采用后处理技术来过滤掉与原始句子有显著语义偏差的句子。具体来说，我们利用 Sentence-BERT (SBERT) [20] ，一个句子嵌入模型，来测量原始句子和合成句子之间的语义相似度。SBERT 将句子映射到一个密集的高维向量空间，其中语义相似的句子被放置在彼此更接近的位置。给定一个原始句子 S_{orig} 和一个生成句子 S_{gen} ，SBERT 计算它们的嵌入表示如下：$v_{\text{orig}} = \text{SBERT}(S_{\text{orig}})$， $v_{\text{gen}} = \text{SBERT}(S_{\text{gen}})$ 。这些向量表示不仅捕获词汇信息，还捕获句子的更深层次的上下文和语义特性。

我们计算它们的余弦相似度以量化原始句子和生成句子之间的语义一致性：

$$\text{sim}(S_{\text{orig}}, S_{\text{gen}}) = \frac{v_{\text{orig}} \cdot v_{\text{gen}}}{\|v_{\text{orig}}\| \|v_{\text{gen}}\|}$$

为了消除语义不一致和低质量的句子，我们定义了一个相似度阈值 τ 。如果相似度得分低于 τ ，则生成的句子被认为与原句语义距离过大，因此被丢弃：

$$S_{\text{gen}} \text{ is retained if } \text{sim}(S_{\text{orig}}, S_{\text{gen}}) \geq \tau$$

阈值 τ 是基于验证实验经验性确定的，以确保我们在滤除过度偏离的输出时保留多样但语义上忠实的句子。

C. 置信加权微调

微调可以显著影响 NLP 模型的性能，将通用语言模型转化为具备增强能力的特定领域和任务的专用工具。我们建议使用一种信心加权的方法来微调模型。

给定一个训练样本 x_i ，y_i 表示样本 x_i 的真实标签，$p_i^{(y_i)}$ 表示分配给 y_i 的预测概率。样本 x_i 的标准交叉熵损失为：

$$\mathcal{L}_{\text{CE}}(y_i) = -\log p_i^{(y_i)} \tag{1}$$

为了增强模型生成可靠且准确情感预测的能力，我们提出了一种置信加权微调策略，该策略结合了基于预测正确性和概率的置信值，使模型能够学习更自信和准确的预测。置信度值确保正确的预测根据它们的置信水平为模型训练做出贡献。错误的预测不提供置信度，使用标准的交叉熵损失来训练模型。置信度值 v_i 表示为：

$$v_i = \begin{cases} \max_c p_i^{(c)}, & \hat{y}_i = y_i \\ 0, & \hat{y}_i \neq y_i \end{cases} \tag{2}$$

，其中 $\hat{y}_i$ 是样本 i 的预测情感极性，$p_i^{(c)}$ 表示模型在类别 c 上对第 i 个样本的预测概率，$\max_c p_i^{(c)}$ 表示模型的最大预测概率（即其置信度）。

为了在训练过程中纳入置信值，我们提出了一种置信加权的损失函数，如下式所示：

$$\mathcal{L}_i^{\text{v}} = \mathcal{L}_{\text{CE}}(y_i) \cdot (1 + \alpha \cdot v_i) \tag{3}$$

，其中 $\alpha \in \mathcal{R} > 0$ 是一个超参数，用于控制置信度的影响。当 $v_i > 0$ 时，置信值影响损失的权重，从而提高正确预测的概率。当 $v_i = 0$ 时，损失变为标准的交叉熵损失。

整个数据集 $\mathcal{D}$ 的总体训练目标被计算为置信度加权损失的平均值：

$$\mathcal{L}_{\text{v}} = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \mathcal{L}_i^{\text{v}} \tag{4}$$

IV. 实验

在实验中，我们使用了四个公共数据集：来自 SemEval-2015 任务 12 的 Rest15 和 Lap15，以及来自 SemEval-2016 任务 5 的 Rest16 和 Lap16。每个数据集的统计信息如表 I 所示。我们使用精确率（P）、召回率（R）和 F1 分数（F1）来评估模型。

Table I: 用于实验的每个数据集的统计信息。

		Rest15	Lap15	Rest16	Lap16
Training set		1120	1399	1708	2039
Testing set		582	644	587	573
Category	Train	12	12	81	81
	Test	12	12	58	67
Positive annotations	Train	1198	1103	1657	1637
	Test	454	541	611	481
Negative annotations	Train	403	765	749	1084
	Test	346	329	204	274
Neutral annotations	Train	53	106	101	188
	Test	45	79	44	46

在实验中，我们利用 GPT-4o [1] 作为数据增强的生成模型。像 Flan-T5-xl [11] 这样的编码器-解码器模型预训练于大量语料库，使它们对句法和语义有更强的理解。此外，Flan-T5-xl [11] 已对大量带有自然语言指令的任务进行微调，使其更好地处理隐含情感和领域特定的表达。我们使用 Flan-T5-xl [11] 进行微调。

对于 ACSC 任务，我们在 Rest15、Rest16、Lap15 和 Lap16 数据集上使用的参数 α 分别为 0.4673、0.3045、0.3045 和 0.6123。对于 ACSA 任务，我们使用的参数 α 是 0.1713、

[1] https://platform.openai.com/docs/models/

Table II: ACSC 任务的实验结果。† 表示实验结果均来自 ECAN [21] 。标记为 ___ 的结果在所有方法中表现最佳。标记为 ‡ 的分数在性能上排名第二。

Methods	Rest15			Rest16			Lap15			Lap16		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CAER-BERT †	71.27	63.75	66.25	77.69	68.39	73.6	58.34	51.86	60.78	65.36	54.73	60.79
SCAN †	71.25	67.05	69.09	73.87	76.25	75.05	71.01	45.49	55.45	61.06	45.42	52.08
AC-MIMLLN †	71.45	65.18	68.17	75.66	74.89	75.27	59.06	53.88	55.29	68.08	55.89	61.39
LC-BERT †	72.08	65.76	68.77	76.34	74.16	75.75	66.84	50.78	57.72	71.05	53.11	55.58
EDU-capsule †	72.67	67.34	69.9	75.93	77	74.81	73.98	52.27	65.43	67.02	49.2	58.01
Hier-BERT †	70.42	65.75	68.01	76.99	73.88	75.41	70.19	48.41	57.3	66.01	53.71	59.23
Hier-GCN-BERT †	75.12	66.47	71.94	77.62	75.61	77.68	75	66.84	70.69	68.32	62.54	65.09
MvP †	67.8	68.63	68.21	73.76	75.49	74.62	-	-	-	-	-	-
ECAN †	84.38	74.43	79.09	84.92	79.65	82.2	83.56	75.34	79.24	76.11	65.45	70.45
Ours (w/o DA)	89.75	90.53	89.46	92.41 ‡	92.67 ‡	91.88 ‡	90.81 ‡	91.46 ‡	90.74 ‡	88.97 ‡	89.39	89.17 ‡
Ours (DA)	86.79 ‡	88.28 ‡	86.72 ‡	93.84	93.83	93.44	92.92	93.36	92.99	90.27	89.14 ‡	89.64

Table III: ACSA 任务的实验结果。* 表示实验结果均来自 PBJM [7] 。标记为 ___ 的结果在所有方法中表现最佳。标记为 ‡ 的得分在性能中排名第二。

Model	Rest15			Rest16			Lap15			Lap16		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Pipeline-BERT *	38.12	70	49.35	43.62	79.06	56.21	36.91	51.62	43.02	31.92	51.56	39.42
Cartesian-BERT *	72.02	49.15	58.42	74.96	63.84	68.94	73.06	21.18	32.83	64.99	27.4	39.54
Addonedim-BERT *	68.84	55.86	61.67	71.75	67.95	69.79	64.13	39.57	48.94	58.83	39.49	47.23
AS-DATJM *	66.35	50.52	57.21	70.88	60.35	65.19	58.91	40.28	47.76	57.29	36.7	44.71
Hier-BERT *	67.46	57.98	62.36	70.97	69.65	70.3	65.47	41.26	50.61	59.51	41.93	49.19
Hier-Trans-BERT *	70.22	59.96	64.67	73.25	73.21	73.45	65.63	51.95	57.79	58.06	48.29	52.52
Hier-GCN-BERT *	71.93	58.03	64.23	76.37	72.83	74.55	71.9	54.73	62.13	61.43	48.42	54.15
PBJM *	75.07	61.48	67.58	76.53	73.6	75.03	72.23 ‡	54.96	62.41	62.63	50.08	55.62
Ours (w/o DA)	79.73	75.29	77.44	79.26 ‡	79.89 ‡	79.58 ‡	68.96	63.44 ‡	66.08 ‡	64.14	59.18	61.56
Ours (DA)	76.16 ‡	72.88 ‡	74.48 ‡	81.58	83.75	82.65	72.74	68.6	70.61	62.77 ‡	58.30 ‡	60.45 ‡

0.5903、0.2526 和 0.5378，这些参数是通过使用 Optuna [2] 调优的。在所有实验中，相似性阈值 τ 为 0.7。正如我们在第 **??** 节讨论 τ 对模型性能的影响时所提到的，当 τ =0.7 时，我们的模型在大多数基准测试中达到了最佳结果。为了测试我们模型在 ACSC 和 ACSA 任务中的有效性，我们根据不同的任务将其与多个基线进行比较。

对于 ACSC 任务，我们将其与以下方法进行比较：CAER-BERT [22] 、SCAN [23] 、AC-MIMLLN [6] 、LC-BERT [24] 、EDU-capsule [25] 、Hier-BERT [26] 、Hier-GCN-BERT [26] 、MvP [27] 、ECAN [21] 。在这些基线中，ECAN [21] 在 ACSC 任务中超越了其他基线。

对于 ACSA 任务，我们将其与以下方法进行比较：Pipeline-BERT, Cartesian-BERT, Addonedim-BERT [28] , AS-DATJM [29] , Hier-BERT [26] , Hier-Trans-BERT [26] , Hier-GCN-BERT [26] , PBJM [7] 。其中，PBJM [7] 在 ACSA 任务中取得了最佳表现。

A. 主要结果

表 II 和表 III 展示了我们提出的方法在数据增强（DA）和没有数据增强（w/o DA）情况下的性能，并与一组基线在四个基准数据集 Rest15、Rest16、Lap15 和 Lap16 上的 ACSC 和 ACSA 任务进行了比较。所有基线结果分别来自 ECAN [21] 和 PBJM [7] ，它们是各任务基线中最优秀的模型。值得注意的是，我们提出的方法的两种变体在所有数据集上始终优于其他所有基线模型，获得了最高的实验评估结果。

1) 方面类别情感分类的性能： 在 Rest15 数据集上，我们的模型在没有数据增强的情况下实现了最佳的整体性能，F1 得分为 89.46 %，比 ECAN 高出 10.37 %。值得注意的是，它还实现了更高的精确度（89.75 % 对比 84.38 %）和召回率（90.53 % 对比 74.43 %）。尽管我们的模型在使用数据增强（DA）时的性能略有下降，F1 得分为 86.72 %，但仍比 ECAN 高出 7.63 %。这表明，虽然数据增强在该数据集上没有提高性能，但核心模型架构依然有效。

在 Rest16 数据集上，使用数据增强的模型获得了 93.44 % 的最高 F1 分数，显著超过了 ECAN 的 +11.24 % 。精确率（93.84 % ）和召回率（93.83 % ）均达到了非常高的结果，表明数据增强在 Rest16 数据集上的有效性。即使没有数据增强，我们的方法（w/o DA）的 F1 分数也达到了 91.88 % ，比 ECAN 高出 +9.68 % 。同样地，在 Lap15 和 Lap16 数据集上，我们的 DA 增强模型再次取得了最佳 F1 分数，精确率和召回率都非常高。我们的模型即便在没有增强的情况下也显著优于 ECAN。这些结果突显了我们方法在 DA 和 w/o DA 设置上的优越性。

2) 方面类别情感分析的表现： 在 Rest15 数据集上，我们的方法在没有数据增强的情况下，Ours (w/o DA) 达到了 77.44 % 的最佳 F1 分数，比 PBJM (67.58 %) 高出 +9.86 % F1。虽然我们经过数据增强的模型的表现稍低，F1 分数为 74.48 %，但它仍然比 PBJM 高出 +6.90 % F1。尽管在该数据集中数据增强未能进一步提高性能，但它保

[2]https://github.com/optuna/optuna

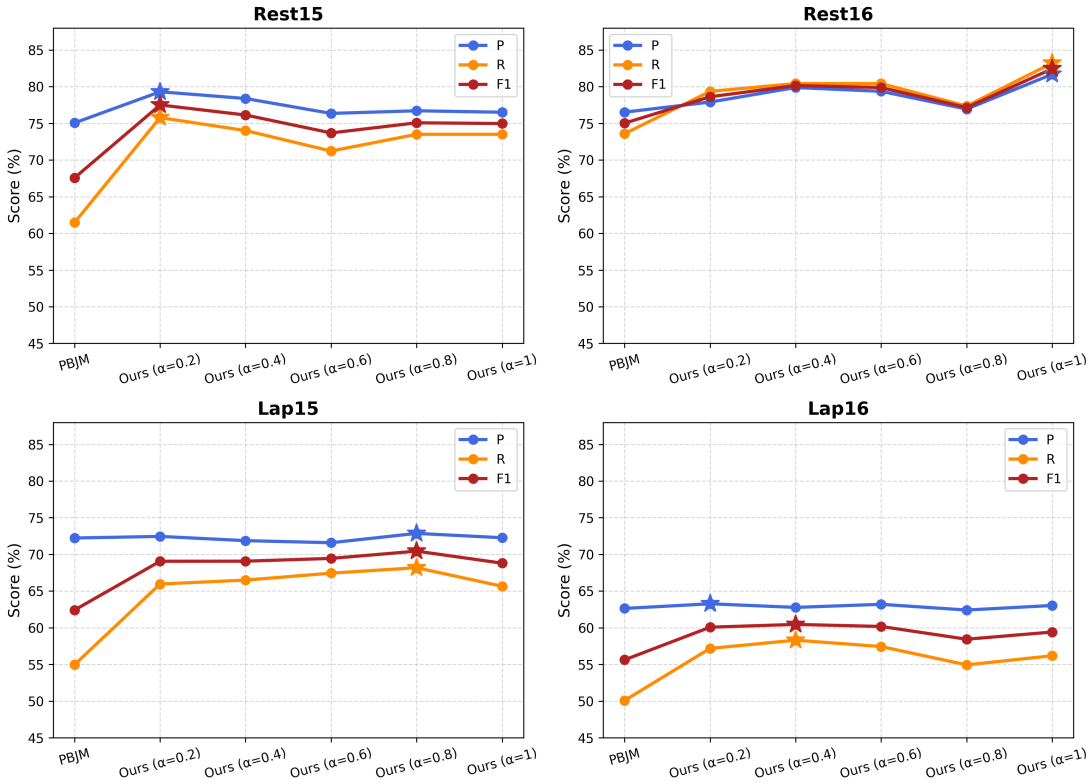

Figure 4: 不同 α 值下在四个数据集上的精确度、召回率和 F1 分数用于 ACSA 任务。PBJM [7] 是基线。最佳的指标值用星号 $\star$ 标出。

持了竞争力的结果。与 Rest15 相似，在 Lap16 数据集上，我们没有数据增强的模型（61.56 % F1）比数据增强版本（60.45 % F1）表现更好。然而，这两种版本分别比 PBJM（55.62 % F1）高出 +5.94 % 和 +4.83 % F1。

与 Rest15 和 Lap16 相比，我们的 DA 增强模型在 Rest16 数据集上实现了最高的 F1 得分 82.65 %，比 PBJM（75.03 %）高出 7.62 % F1。没有 DA 的变体也表现良好（79.58 % F1），比 PBJM 高出 4.55 % F1。这些结果表明，我们的数据增强策略在该领域中特别有效，持续提高精准率和召回率，导致了优秀的 F1 表现。在 Lap15 数据集上，我们的模型表现出类似的实验现象，如上所述，数据增强获得了最佳的 F1 得分。我们的无数据增强方法获得了 66.08 % 的 F1 得分，仍然比 PBJM 高出 3.67 %。

B. 参数分析

我们分析了在四个基准数据集 Rest15、Rest16、Lap15 和 Lap16 上的 ACSA 任务中超参数 α 、τ 和模型大小的影响。详细的实验结果如图 4 和 5 所示。在几乎所有的数据集和指标中，我们的方法在精确度（P）、召回率（R）和 F1 得分方面都优于最佳的基线模型 PBJM [7] 。

在 Rest15 数据集中，与 F1 得分为 67.58 % 的 PBJM 相比，我们的方法在所有 α 的值上均表现更好，其中最佳结果为在 α =0.2 时达到 77.51 % 。具体来说，Precision 在 α =0.2 时最高，为 79.31 % ，同样，Recall 也在 α =0.2 时达到顶峰，为 75.79 % 。性能在超过 α =0.4 时稳步下降，

Table IV: 我们的方法在不同模型大小下对 ACSA 任务的评估结果。

Model	Rest16			Lap16		
	P	R	F1	P	R	F1
Ours (220M)	63.42	50.33	56.12	36.38	23.85	28.81
Ours (770M)	68.43	60.32	64.12	60.18	50.56	54.95
Ours (3B)	81.58	83.75	82.65	62.77	58.3	60.45

这表明增加信心的影响（高 α ）可能会过度强调正确预测，从而牺牲泛化能力。这意味着对于 Rest15 数据集，较低的 α 值（例如，0.2–0.4）是最优的。

在 Rest16 数据集中，PBJM 的 F1 得分为 75.03 % ，而我们的方法在 α =1.0 时达到了 82.45 % ，超过了所有其他方法。在 α =1.0 时召回率达到最大值（83.22 % ）。该数据集受益于较大的置信权重。

在 Lap15 数据集上，PBJM 表现不佳，F1 值为 62.41 %，而我们的方法将其提高到 70.44 %，其中 α =0.8。当 α 达到 0.8 时，模型性能稳步提升，随后在 α =1.0 时略有下降，最佳的 α 范围在 0.6 到 0.8 之间。

最具挑战性的数据集是 Lap16 数据集。PBJM 的 F1 得分仅为 55.62 %。我们的模型在 α =0.4 时将性能提升至 60.45 %，但召回率和准确率相对较低。当 α =0.4 时，我们的模型达到最佳性能，超过此值后，准确率和召回率都会下降。

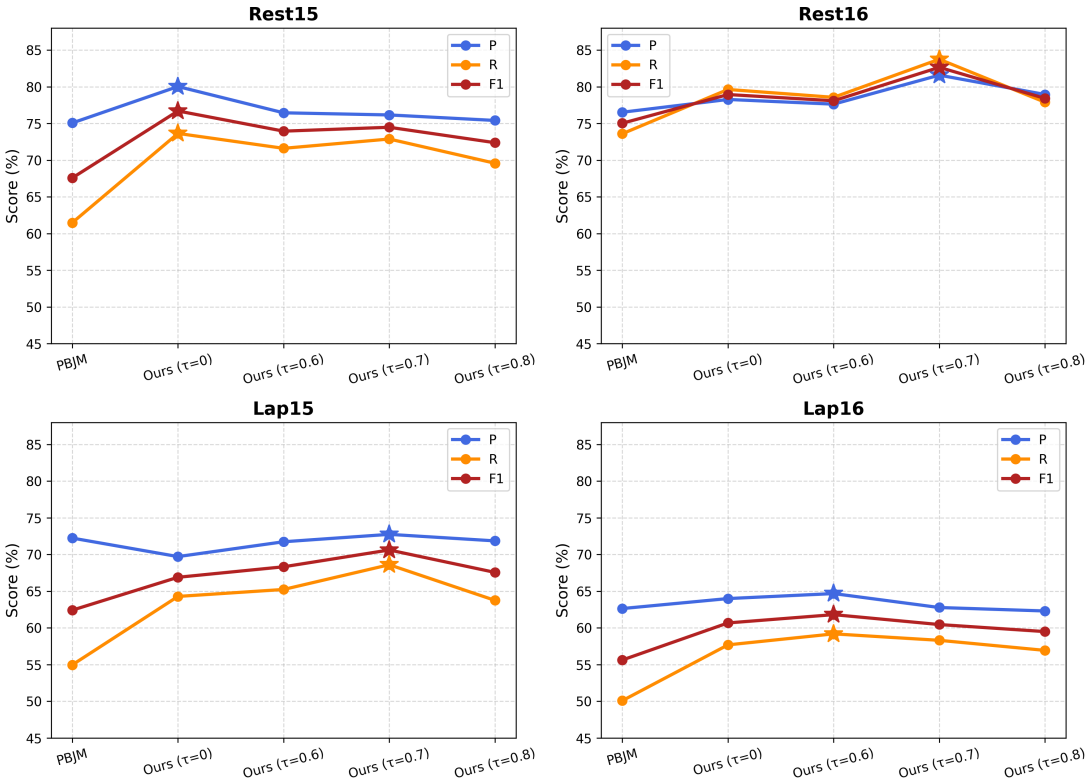

Figure 5: 四个数据集上在不同语义一致性阈值 τ 下的精确率、召回率和 F1 分数用于 ACSA 任务。PBJM [7] 是基线，$\tau = 0$ 表示没有过滤。最佳指标值用星号 $\star$ 标出。

为了研究数据增强中语义一致性过滤的效果，我们使用不同的阈值 $\tau \in \{0, 0.6, 0.7, 0.8\}$ 进行实验。参数 τ 控制基于生成句子与原始句子之间的语义相似度的后处理。具体来说，我们保留余弦相似度高于阈值 τ 的句子。$\tau = 0$ 表示不进行任何过滤，使用所有生成的句子。

从图 5 中可以观察到，我们的方法在所有指标和数据集上均优于 PBJM 基线，这证明了我们增强框架的有效性。值得注意的是，在 Rest16、Lap15 和 Lap16 数据集上，我们使用过滤过的增强数据（$\tau > 0$ ）训练的模型要优于未过滤的版本（$\tau = 0$ ），这表明去除语义不一致的样本可以获得更高质量的训练数据。

在 Rest15 数据集上，我们的模型与 $\tau = 0$ 一起表现最佳，而进一步过滤稍微降低了性能。这表明在某些领域，激进的过滤可能会丢弃增强样本中的有用多样性。尽管如此，即便在这种情况下，我们所有带有 $\tau > 0$ 的模型仍然优于 PBJM 基线。

我们的分析表明，中等阈值 $\tau = 0.7$ 提供了最佳的权衡，并在所有数据集上表现出色。

1) 模型大小的影响： 为了研究模型规模对性能的影响，我们使用基于 Flan-T5 的三种不同模型规模（220M、770M 和 3B 模型）进行实验。我们以 Rest16 和 Lap16 为例，这两个数据集的结果如表 IV 所示。

随着模型规模的增加，我们在所有评估指标上观察到了一致的改进。在 Rest16 数据集上，F1 分数从 220M 模型的 56.12 % 提高到 770M 模型的 64.12 %，然后再提高到 3B 模型的 82.65 %。Lap16 数据集也有类似的趋势，F1 分数随着模型规模的增加而改善。

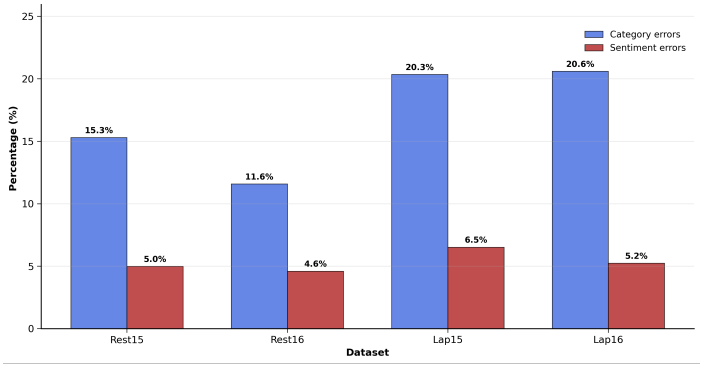

Figure 6: 类别错误分布和情感错误分布，分别表示正确预测情感但错误预测类别，以及正确预测类别但错误预测情感。

结果清楚地表明，扩大模型的规模可以显著提高性能。较大的模型在捕捉方面类别情感分析中的复杂语义时更为有效。

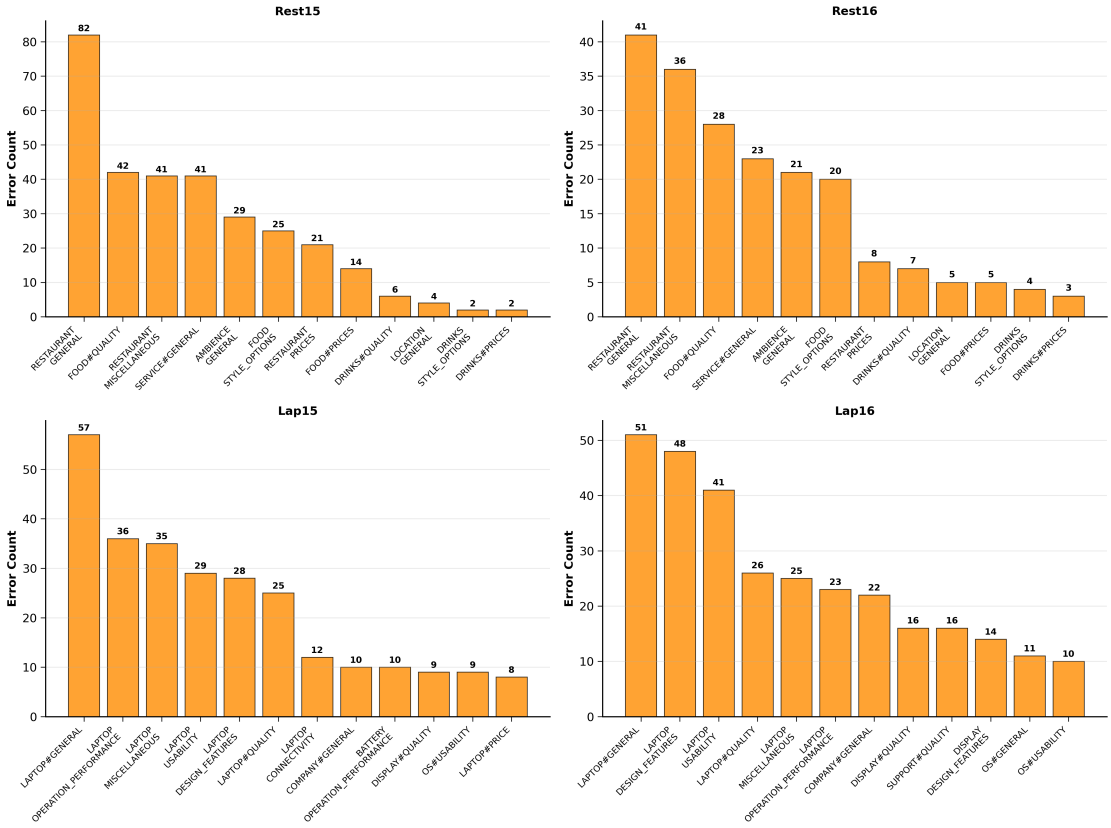

Figure 7: 四个基准数据集中错误预测的方面类别的错误计数。

C. 误差分析

ACSA 是一项端到端的任务，它要求同时正确预测类别和情感，才被视为真正准确的预测。我们分析了在四个 SemEval 基准测试中正确预测情感但错误预测类别，以及正确预测类别但错误预测情感的错误样本分布，如图 6 所示。

1) 识别正确的类别比识别其情感更困难。: 在所有基准数据集中，类别错误超过情感错误，而在餐馆数据集中，类别错误几乎是情感错误的三倍。在笔记本电脑数据集中，类别错误是情感错误的四倍。这表明，即使在类别有限且明确的领域中，比如 FOOD # QUALITY、LOCATION # GENERAL 和 RESTAURANT # PRICES，当多个方面类别同时出现在一个句子中时，识别正确的类别并预测它们的情感仍然具有挑战性。

从餐馆到笔记本电脑领域使得类别错误率增加了 5 个百分点，而情感错误率仅略有增加。如表 I 所示，笔记本电脑数据集增加了细粒度类别的数量，并且类别之间具有更高的语义重叠，例如 BATTERY # GENERAL, BATTERY # MISCELLANEOUS 和 BATTERY # OPERATION_PERFORMANCE。相比之下，餐馆数据集的类别中存在较少的歧义，如 FOOD # QUALITY 和 RESTAURANT # PRICES。更丰富和更细粒度的类别增加了模型在预测方面类别时的挑战。

我们还对四个基准数据集的类别级错误进行了细粒度的分析。图 7 展示了每个数据集的错误方面类别预测的分布。从图 7 中，我们可以获得以下观察结果。

2) 一般类别在错误中占主导地位。: RESTAURANT # GENERAL 和 LAPTOP # GENERAL 错误数量始终位居错误来源的前列。这表明，当模型无法检测句子中的细粒度类别时，这些通用类别被过度预测。

具有高语义重叠的类别，例如 FOOD # QUALITY 和 FOOD # STYLE_OPTIONS，或 LAPTOP # DESIGN_FEATURES 和 LAPTOP # USABILITY，通常同时出现在错误列表中。这表明，当类别的语义表达相似时，模型很难做出准确的预测。此外，观察到在 Laptop 数据集中，类别之间的语义重叠更广泛。相比之下，在 Restaurant 数据集中，错误更多地集中在少数类别中。

V. 结论

我们介绍了一种数据增强方法，该方法利用大型语言模型生成与原始句子在语义上一致且具有语言多样性的数据集。为了进一步确保数据质量，我们引入了一种后处理技术，用于过滤偏离原句语义过多的生成句子，旨在解决 ACSA 任务中的数据稀缺挑战。微调使预训练语言模型能够适应特定领域的情感模式。它使模型能够捕捉目标数据集中方面类别与情感极性的相关性。为了使模型更好地适应 ACSA 任务，我们提出了一种信心加权的微调策略来训练模型，鼓励其学习具有高度信心和正确的预测。我们在四个公开数据集上的实验结果均优于基准模型，有效提高了 ACSA 任务的现有性能。

VI.

致谢 本文描述的研究由中国香港特别行政区研究资助局的资助（R1015-23）、岭南大学的院校研究基金（SDS24A8）和直接资助（DR25E8）支持。

References

[1] Q. Cheng, J. Huang, and Y. Duan, "Semantically consistent data augmentation for neural machine translation via conditional masked language model," in *Proc. Int. Conf. Comput. Linguist. (COLING)*. Int. Comm. Comput. Linguistics, 2022, pp. 5148–5157.

[2] T. Hsu, C. Chen, H. Huang, and H. Chen, "Semantics-preserved data augmentation for aspect-based sentiment analysis," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*. Assoc. Comput. Linguistics, 2021, pp. 4417–4422.

[3] H. Wang, K. He, B. Li, L. Chen, F. Li, X. Han, C. Teng, and D. Ji, "Refining and synthesis: A simple yet effective data augmentation framework for cross-domain aspect-based sentiment analysis," in *Findings of the Association for Computational Linguistics (ACL)*. Assoc. Comput. Linguistics, 2024, pp. 10 318–10 329.

[4] H. Xu, Y. Zhang, Q. Wang, and R. Xu, "Ds^2-absa: Dual-stream data synthesis with label refinement for few-shot aspect-based sentiment analysis," *arXiv preprint arXiv:2412.14849*, 2024.

[5] B. Liang, H. Su, R. Yin, L. Gui, M. Yang, Q. Zhao, X. Yu, and R. Xu, "Beta distribution guided aspect-aware graph for aspect category sentiment analysis with affective knowledge," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2021, pp. 208–218.

[6] Y. Li, C. Yin, S. Zhong, and X. Pan, "Multi-instance multi-label learning networks for aspect-category sentiment analysis," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2020, pp. 3550–3560.

[7] Z. Ping, G. Sang, Z. Liu, and Y. Zhang, "Aspect category sentiment analysis based on prompt-based learning with attention mechanism," *Neurocomputing*, vol. 565, p. 126994, 2024.

[8] J. Liu, Z. Teng, L. Cui, H. Liu, and Y. Zhang, "Solving aspect category sentiment analysis as a text generation task," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2021, pp. 4406–4416.

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*. Assoc. Comput. Linguistics, 2020, pp. 7871–7880.

[10] L. D. Quilio and F. Fioravanti, "A comprehensive framework for aspect-category sentiment analysis," in *Proc. 8th Workshop Natural Lang. Artif. Intell. (NL4AI) at Int. Conf. Italian Assoc. Artif. Intell. (AI*IA)*, 2024.

[11] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *J. Mach. Learn. Res.*, vol. 25, no. 70, pp. 1–53, 2024.

[12] T. Gao, J. Fang, H. Liu, Z. Liu, C. Liu, P. Liu, Y. Bao, and W. Yan, "Lego-absa: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis," in *Proc. Int. Conf. Comput. Linguistics (COLING)*, 2022, pp. 7002–7012.

[13] Y. Bai, Z. Han, Y. Zhao, H. Gao, Z. Zhang, X. Wang, and M. Hu, "Is compound aspect-based sentiment analysis addressed by llms?" in *Findings Assoc. Comput. Linguistics: EMNLP*, 2024, pp. 7836–7861.

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*. Assoc. Comput. Linguistics, 2019, pp. 4171–4186.

[15] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.

[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022.

[17] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*. OpenReview.net, 2023.

[18] W. Lai, H. Xie, G. Xu, and Q. Li, "Rvisa: Reasoning and verification for implicit sentiment analysis," *IEEE Trans. Affect. Comput.*, 2025.

[19] Y. Chai, H. Xie, and S. J. Qin, "Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities," *CoRR*, vol. abs/2501.18845, 2025.

[20] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proc. Conf. Empirical Methods Nat. Lang. Process. and Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*. Assoc. Comput. Linguistics, 2019, pp. 3980–3990.

[21] J. Cui, F. Fukumoto, X. Wang, Y. Suzuki, J. Li, N. Tomuro, and W. Kong, "Enhanced coherence-aware network with hierarchical disentanglement for aspect-category sentiment analysis," in *Proc. Joint Int. Conf. Comput. Linguistics, Lang. Resources and Eval. (LREC-COLING)*, 2024, pp. 5843–5855.

[22] B. Liang, J. Du, R. Xu, B. Li, and H. Huang, "Context-aware embedding for targeted aspect-based sentiment analysis," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*. Assoc. Comput. Linguistics, 2019, pp. 4678–4683.

[23] Y. Li, C. Yin, and S. Zhong, "Sentence constituent-aware aspect-category sentiment analysis with graph attention networks," in *Proc. Int. Conf. Natural Lang. Process. and Chinese Comput. (NLPCC)*, ser. Lecture Notes in Comput. Sci., vol. 12430. Springer, 2020, pp. 815–827.

[24] Y. Wu, Z. Zhang, Y. Zhao, and B. Qin, "Locate and combine: A two-stage framework for aspect-category sentiment analysis," in *Proc. 10th CCF Int. Conf. Natural Lang. Process. Chinese Comput. (NLPCC)*, ser. Lect. Notes Comput. Sci., vol. 13028. Springer, 2021, pp. 595–606.

[25] T. Lin, A. Sun, and Y. Wang, "Edu-capsule: Aspect-based sentiment analysis at clause level," *Knowl. Inf. Syst.*, vol. 65, no. 2, pp. 517–541, 2023.

[26] H. Cai, Y. Tu, X. Zhou, J. Yu, and R. Xia, "Aspect-category based sentiment analysis with hierarchical graph convolutional network," in *Proc. 28th Int. Conf. Comput. Linguistics (COLING)*. Int. Comm. Comput. Linguistics, 2020, pp. 833–843.

[27] Z. Gou, Q. Guo, and Y. Yang, "Mvp: Multi-view prompting improves aspect sentiment tuple prediction," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2023, pp. 4380–4397.

[28] M. Schmitt, S. Steinheber, K. Schreiber, and B. Roth, "Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2018, pp. 1109–1114.

[29] P. Gu and Z. Zhang, "Dual-attention based joint aspect sentiment classification model," in *Int. Conf. Web Eng. (ICWE)*. Springer, 2022, pp. 252–267.