Computer Science and Application 计算机科学与应用, 2026, 16(9), 25-36
Published Online September 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.169286

Hans汉斯

层级评论上下文依赖识别数据集构建与研究

——以小红书旅游评论数据为例

孙杰一

同济大学经济与管理学院，上海

收稿日期：2026年8月1日；录用日期：2026年8月31日；发布日期：2026年9月8日

摘　要

文本情感分析已广泛用于在线评论和旅游体验研究，但小红书内容具有笔记－评论双层结构，评论往往需要结合原笔记才能准确理解。为避免在数据清洗阶段误删有价值评论，或在主题分析中因简单拼接造成噪声混入与主题情感错配，本文构建面向笔记－评论融合的评论上下文依赖识别任务。本文以小红书旅游评论为研究对象，整理有效评论数据30,802条，并依据语义将评论划分为完整语义评论、上下文依赖评论和无有效态度评论三类。该分类将评论处理由简单“保留/删除”转化为“独立分析/融合分析/过滤噪声”的分层策略，使缺少对象但包含情绪态度的评论得以保留，也避免无效互动进入后续主题建模。在双人人工标注与分歧复核基础上，本文比较BERT、ERNIE等八类模型，并通过五次数据划分检验结果稳定性。实验结果显示，预训练语言模型整体优于传统机器学习方法，其中ERNIE-3.0-base取得最高平均Macro F1；同时，上下文依赖评论的识别效果仍明显低于其他类别。进一步的BERTopic下游验证表明，分层处理策略能够提升主题一致性和主题区分度，说明该分类体系具有实际应用价值。本文可为层级评论数据清洗、评论保留策略和笔记－评论融合建模提供参考。

关键词

旅游评论，笔记－评论结构，上下文依赖识别，语义完整性，预训练语言模型

Construction and Analysis of a Hierarchical Comment Dataset for Context-Dependency Recognition

—An Empirical Study Based on RedNote Tourism Comment Data

Jieyi Sun

School of Economics and Management, Tongji University, Shanghai

Received: August 1, 2026; accepted: August 31, 2026; published: September 8, 2026

文章引用: 孙杰一. 层级评论上下文依赖识别数据集构建与研究[J]. 计算机科学与应用, 2026, 16(9): 25-36.
DOI: 10.12677/csa.2026.169286

Abstract

Text sentiment analysis has been widely applied to online reviews and tourism experience research. However, RedNote content has a "post-comment" two-layer structure, and comments often need to be understood in relation to the original post. To avoid mistakenly removing valuable comments during data cleaning, or introducing noise and topic-sentiment mismatches through simple concatenation in topic analysis, this study constructs a comment context-dependency recognition task oriented toward "post-comment" fusion. Taking RedNote tourism comments as the research object, this study organizes 30,802 valid comments and classifies them into three categories according to semantics: complete semantic comments, context-dependent comments, and comments with no valid attitude. This classification transforms comment processing from a simple "retain/delete" operation into a layered strategy of "independent analysis/fusion analysis/noise filtering", allowing comments that lack explicit objects but contain emotional attitudes to be retained, while preventing invalid interactions from entering subsequent topic modeling. Based on dual manual annotation and disagreement review, this study compares eight models, including BERT and ERNIE, and examines result stability through five data splits. The experimental results show that pretrained language models generally outperform traditional machine learning methods, with ERNIE-3.0-base achieving the highest average Macro F1. Meanwhile, the recognition performance for context-dependent comments remains significantly lower than that for the other categories. Further downstream validation using BERTopic shows that the hierarchical processing strategy improves topic coherence and topic distinctiveness, indicating that the proposed classification scheme has practical application value. This study provides a reference for hierarchical comment data cleaning, comment retention strategies, and "post-comment" fusion modeling.

Keywords

Tourism Comments, Post-Comment Structure, Context-Dependency Recognition, Semantic Completeness, Pretrained Language Models

1. 引言

随着移动互联网和内容社区的发展，社交媒体已成为游客获取旅游信息和形成目的地印象的重要渠道。用户生成内容(User Generated Content, UGC)记录了游客在真实旅行与消费过程中的体验、评价和情绪反馈，对旅游决策、目的地形象感知和旅游服务优化具有重要价值[1] [2]。在这一背景下，小红书逐渐成为国内旅游经验分享和消费决策的重要平台。

与传统旅游评论平台相比，小红书旅游内容并非单一评论结构，而是由笔记–评论共同构成的信息空间。笔记正文通常承担攻略展示、经验叙述和观点表达功能，评论区则包含附和、追问、质疑、补充、调侃和情绪表达等互动内容。因此，评论区是理解小红书旅游内容传播与用户反馈的重要组成部分。

然而，这种双层结构也给文本分析带来了新问题。传统评论分析通常默认单条评论具有较完整的评价对象和态度，但小红书评论中大量内容需要结合原笔记才能理解。例如“太真实了”“确实”“笑死”等评论虽包含情绪，却缺少明确对象；而“求攻略”、“地址在哪里”、“蹲一个”等评论则属于信息请求或社交互动。如果简单删除短评论，可能丢失有价值的情绪反馈；如果将所有评论直接拼接进主题模

型，又可能引入噪声并造成主题–情感错配[3] [4]。

基于此，本文提出面向笔记–评论融合的小红书旅游评论三分类任务，将评论划分为完整语义评论、上下文依赖评论和无有效态度评论三类。该任务关注评论是否能够独立理解、是否需要与笔记融合分析，以及是否应在后续主题建模中被过滤。本文基于青岛、成都和厦门三地评论构建数据集，并比较多类文本分类模型的识别效果，以期为层级评论数据清洗、评论保留策略和笔记–评论融合建模提供参考。

2. 相关研究

2.1. 旅游 UGC 与社交媒体评论研究

用户生成内容(User Generated Content, UGC)是旅游体验传播和旅游决策研究中的重要数据来源。游客在社交媒体中发布的笔记、图片、短视频和评论，既记录了真实旅行体验，也影响其他用户对目的地形象、服务质量和出行风险的判断。近年来关于微博、小红书等平台的研究表明，社交媒体旅游内容会影响用户旅行决策，评论区中的反馈、追问和态度表达也会参与旅游信息的再生产过程[1] [2]。因此，旅游评论不只是附属于正文的补充信息，而是理解平台旅游内容传播和用户感知的重要组成部分。

2.2. 笔记–评论结构与主题分析

传统在线评论研究通常将单条评论视为相对独立的文本单位，但小红书内容具有更明显的笔记–评论双层结构。评论区中既有“风景很好”、“酒店很差”等可以独立理解的完整评价，也有“确实”、“太真实了”、“笑死”等必须依赖原笔记才能判断对象的情绪回应，还有“求攻略”、“地址在哪里”等信息请求。在主题分析场景中，短文本本身容易受到语义稀疏影响[3]，而 BERTopic 等基于语义表示的主题模型虽然改善了短文本主题发现能力[4]，但仍然依赖输入文本的语义质量。因此，在主题建模前区分评论是否需要与原笔记融合，是提高后续主题解释质量的重要前提。

2.3. 文本分类模型研究

文本分类方法大体经历了由人工特征和传统机器学习向预训练语言模型发展的过程。传统方法实现简单、训练成本较低，适合作为基础基线，但对语义省略、短句情绪和上下文依赖表达的理解能力有限。近年来，BERT、RoBERTa、ERNIE 和 MacBERT 等预训练语言模型通过上下文语义表示提升了文本分类效果[5]-[8]，ELECTRA 和 TinyBERT 等轻量化模型也为不同计算资源条件下的文本分类提供了可比方案[9] [10]。已有综述指出，深度学习模型已成为文本分类任务中的重要技术路径，但具体效果仍取决于任务定义、数据质量和类别边界清晰度[11]。

2.4. 研究述评

综上，已有研究为旅游短文本主题分析和文本分类方法提供了较充分基础，但对小红书笔记–评论双层结构下评论语义完整性的关注仍不足。现有处理方式往往将评论简单视为有效文本或噪声文本，容易忽略具有情绪态度但缺少明确对象、需要结合原笔记理解的上下文依赖评论。本文据此构建完整语义评论、上下文依赖评论和无有效态度评论三类识别任务，并通过多模型基线实验检验该任务的可识别性与难点。

3. 数据来源与标注体系

3.1. 数据来源与数据规模

本文数据来源于小红书平台中与旅游体验相关的笔记及其评论内容。数据采集阶段使用八爪鱼采集

器进行爬取，以“青岛”、“成都”、“厦门”等城市名称为核心检索词，并结合“旅游”、“出游”、“旅居”、“生活”、“定居”等关键词筛选相关笔记。筛选标准主要包括：第一，笔记标题、正文或标签中包含目标城市名称；第二，笔记内容与旅游、出行、旅居体验、城市生活或定居感受相关；第三，笔记下方存在可用于分析的用户评论。由于本文关注的是小红书旅游评论而非特定事件或短期舆情，因此数据采集不限定具体发布时间范围。原始数据包含笔记信息、评论文本及相关基础字段，如图 1 所示。考虑到本文关注的核心问题是小红书笔记–评论双层结构下评论文本的上下文依赖识别，因此数据整理阶段主要保留评论文本及其所属笔记来源信息，并将评论作为后续分类建模的基本分析单位。经过数据清洗与人工标注后，最终形成 30,802 条有效评论数据，其中青岛 11,364 条、厦门 10,124 条、成都 9314 条，如表 1 所示。

笔记详情		评论列表										
		评论ID	用户名	用户主页	地区	评论时间	评论天数	评论内容	点赞数量	博主点赞	回复数量	互动数量
686f0cde000000002001ad30	https://www.xiaoho	6a36540f000000001500	ShandiHH	6948108400000000370	广东	2026-06-20	4	mark、要去了	0	否	0	0
标题：朋友来成都，我会推荐的	8个好地方💚	6a2815a1000000002b02	智者勇入爱河	5ee99eae00000000010(	四川	2026-06-09	15	全是本地人不去的地方[大	2	否	12	14
用户主页	用户名：天新Daily	6a1fcec6000000027022	克里麦🇫🇷	5b4ae6086b58b75dc287	四川	2026-06-03	21	好地方	0	否	1	1
点赞量：12000	评论量：145	6a1f9cff0000000028036(	泽小曲儿 Better m	59ec9b4911be1043e8f9	四川	2026-06-03	21	这几个地方哪个离大丰近	0	否	1	1
		6a1893440000000002803	M🇲🇲（不回私信	66e94ab5000000001d02	陕西	2026-05-29	26	想住cpl附近好一点的酒店	0	否	0	0
		6a13e663000000002b02	🦋璨金果子🦋	5b1a6d994eacab50c53a	上海	2026-05-25	30	如果在成都一周逛您所说	1	否	0	1
		6a0d647f000000002800(	jojo	58b974d76a6a6952e336	法国	2026-05-20	35	第一次去 一家三口带孩子	0	否	0	0
		69f958d3000000002902	Que sera	57874a6482ec395d2bd8	四川	2026-05-05	50	认可！！	0	否	1	1
		69f8487e000000002a00	安妮王	5b191bfde8ac2b4d5b45	四川	2026-05-04	51	我算是家住真正老城区的	0	否	0	0
		69f4d8c5000000002902(	小小白ss	694155a000000000030	四川	2026-05-02	53	太齐全了	0	否	0	0
		69e99ac6000000001401	高小米	5ae92d14e8ac2b4c4cba	福建	2026-04-23	62	推荐住宿住哪?	0	否	3	3
		69e5bae3000000001401	cheese	59bfce0c44363b43970a	重庆	2026-04-20	65	哪个地方离东安湖体育场	1	否	4	5
		69e32e1700000001401	Esme	5c2e394d000000000503	四川	2026-04-18	67	图一是?	0	否	1	1
		69df28f2000000003501d	prostop	6028d302000000000100	四川	2026-04-15	70	我同意，这些地方镇的很	0	否	0	0
		69d60ce2000000001500	Cam	5bd50ab16d0c4d00010(	广东	2026-04-08	77	有脱口秀推荐吗	0	否	1	1
		69d0c8b9000000000503	啤梨啪啦	5e46017a000000000100	四川	2026-04-04	81	看见硬米酒馆就不会错...	0	否	0	0
		69ca2e7b0000000003002	咚掌柜	5c418e0800000000070(	上海	2026-03-30	86	这些好玩吗	0	否	1	1
		69c91635000000003002	Sukiiᵀ	5d047cdf000000001602	马来	2026-03-29	87	图15是哪里	0	否	1	1
		69c73953000000000d01	maxine	5b5566224eacab4d73bf	四川	2026-03-28	88	蹲个爱拍照的搭子	0	否	0	0
		69c29660000000001500	荷动知鱼飞	5d3c4292000000001603	重庆	2026-03-24	92	只有东顺城街没有去。东	1	否	0	1
		69b8bef3000000000202(	总是成的想你	5a224fe111be106241ba	广东	2026-03-17	99	78点适合去哪	0	否	0	0
		69b580db000000001400	小小D伯格🍾	60e7faf9000000002002	四川	2026-03-14	102	这个也不错哎[哇R]	0	否	0	0
		69b24ccc000000001900	新出炉热乎的红薯	5d031e5a000000001000	北京	2026-03-12	104	这8个是按照推荐由强到弱	0	否	1	1

Figure 1. Original data structure
图 1. 原始数据结构

Table 1. Data sources and scale
表 1. 数据来源与规模

城市	有效评论数量	占比
青岛	11,364	36.89%
厦门	10,124	32.87%
成都	9314	30.24%
合计	30,802	100.00%

3.2. 数据清洗与 Emoji 处理

在基础数据清洗阶段，本文主要针对评论文本中的明显无效内容进行处理。首先，删除空文本、乱码文本和纯数字文本等无法提供语义信息的记录；其次，对于评论中出现的@用户内容，仅删除@及其后连续的用户名片段，以避免用户提及信息干扰评论语义分析；最后，清理 Unicode 形式的普通 Emoji 和无效符号，并对清洗后为空的记录进行删除。需要说明的是，小红书平台中常见的表情符号在爬取数据中以“[xxR]”形式保存，本文并未将其简单视为噪声，而是保留为特定文本格式参与情绪和语义判断。例如，“真不错[微笑 R]”中的“[微笑 R]”能够提供积极情绪线索，“[doge]”则通常具有调侃、反语或自嘲语气。因此，Emoji 在本文标注体系中被视为评论表达的组成部分。

3.3. 标签体系设计

本文的分类目标并不是判断评论情感正负，而是判断评论是否能够脱离原笔记独立理解、是否需要结合笔记上下文，以及是否具有进入后续旅游主题分析的价值。基于这一目标，本文构建了三类标签：Label 0 为完整语义评论，Label 1 为上下文依赖评论，Label 2 为无有效态度评论。该标签体系直接服务于后续笔记–评论融合策略：Label 0 可作为独立评论文本进入主题模型，Label 1 需要与原笔记内容进行融合后再分析，Label 2 则可在后续主题分析前过滤。具体判定标准如表 2 所示。

Table 2. Label system for comment three-class classification
表 2. 评论三分类标签体系

标签	类别含义	判定标准	后续处理方式
Label 0	完整语义评论	评论本身包含明确对象、事件、事实或方面，并表达清晰观点、情绪或态度，可脱离原笔记理解。	可单独进入主题模型。
Label 1	上下文依赖评论	评论包含情绪、态度、附和、调侃或反问，但缺少明确对象或事件，需要结合原笔记理解。	与原笔记内容融合后再进入主题模型。
Label 2	无有效态度评论	评论主要为信息请求、求链接、求地址、关注互动或无意义刷屏，不表达有效旅游体验观点。	在主题分析前过滤。

3.4. 双人人工标注与分歧处理

为保证标注质量，本文采用双人标注方式对评论数据进行三分类标注。两名标注者首先依据统一标注规则分别完成标注，随后对存在分歧的样本进行复核，并形成最终标签。为进一步检验标注结果的可靠性，本文计算标注者间一致性指标。结果显示，双人标注的简单一致率为 84.31%，Cohen’s Kappa 系数为 0.623。需要说明的是，评论数据中 Label 0 完整语义评论占比达到 76.86%，类别分布存在明显不均衡。由于 Cohen’s Kappa 在计算时会扣除由标签基准分布带来的随机一致部分，当某一类别占比较高时，Kappa 值可能低于简单一致率所反映的一致程度。因此，0.623 的 Kappa 值并不意味着标注质量较低，而是同时受到任务边界模糊性与类别不平衡的影响。考虑到上下文依赖评论本身涉及语义完整性、情绪表达与原笔记关系判断，Label 1 类评论具有较强边界性，本文进一步对分歧样本进行人工复核，并以复核结果作为最终标签，以提高数据集标签的可靠性。数据集中三个类别的样本数量与比例如表 3 所示。

Table 3. Sample counts and proportions by category
表 3. 样本类别数量与比例

标签	类别含义	样本数量	比例
Label 0	完整语义评论	23,674	76.86%
Label 1	上下文依赖评论	4575	14.85%
Label 2	无有效态度评论	2553	8.29%

4. 实验设计

4.1. 实验任务

本文实验任务为小红书旅游评论三分类识别。模型输入为单条评论文本，输出为对应类别标签，即完整语义评论、上下文依赖评论或无有效态度评论。该实验任务可以看作笔记–评论融合分析前的数据组织环节，其输出结果将决定评论后续是单独分析、与笔记融合分析，还是作为无效互动内容过滤。

4.2. 数据划分

在完成数据清洗与标注后，本文将 30,802 条有效评论按照训练集、验证集和测试集进行划分。训练集用于模型参数学习，验证集用于选择最优模型，测试集用于最终效果评价。基础实验采用约 70%、15%、15%的划分比例，并尽量保持三个类别标签及不同城市来源在各数据集中的比例一致。为降低单次随机划分对实验结论的影响，本文进一步构建了 5 次数据划分，并对各模型在不同划分下的测试结果计算均值和标准差，以检验模型表现的稳定性。具体划分方式如表 4 所示。

Table 4. Dataset split strategy
表 4. 数据集划分方式

数据集	用途	划分比例
训练集	用于模型参数学习	约 70%
验证集	用于模型选择与调参	约 15%
测试集	用于最终效果评价	约 15%

4.3. 对比模型

为全面比较不同方法在该任务中的适用性，本文设置传统机器学习模型、标准预训练语言模型和轻量化预训练语言模型三类基线。传统机器学习模型包括 TF-IDF + Logistic Regression 和 TF-IDF + SVM，用于衡量词面统计特征在该任务上的基础表现。预训练语言模型包括 BERT-base、RoBERTa-wwm、ERNIE-3.0-base 和 MacBERT-base，用于比较不同中文语义表示模型对上下文依赖识别的建模能力。轻量化模型包括 ELECTRA-small 和 TinyBERT，用于观察模型规模和效率变化对分类效果的影响。具体模型设置如表 5 所示。

Table 5. Comparison model settings
表 5. 对比模型设置

模型类别	模型名称	实验作用
传统机器学习	TF-IDF + Logistic Regression	传统词面统计基线
传统机器学习	TF-IDF + SVM	传统高维稀疏文本分类基线
轻量化预训练模型	TinyBERT	轻量化模型对照
轻量化预训练模型	ELECTRA-small	高效预训练模型对照
预训练语言模型	BERT-base	基础预训练语言模型基线
预训练语言模型	RoBERTa-wwm	BERT 优化模型对照
预训练语言模型	ERNIE-3.0-base	知识增强预训练模型对照
预训练语言模型	MacBERT-base	中文预训练模型对照

4.4. 模型训练与参数设置

传统机器学习模型采用字符级 TF-IDF 表示，n-gram 范围设置为 1 到 3，并使用类别平衡策略缓解标签分布差异。深度预训练模型均采用对应模型的 tokenizer 对评论文本进行编码，最大序列长度设置为 128。分类层输出维度为 3，对应三类评论标签。训练过程中使用交叉熵损失函数，优化器采用 AdamW，学习率设置为 2e−5，训练轮数设置为 3。BERT、RoBERTa、ERNIE 和 MacBERT 的批量大小设置为 16，TinyBERT 和 ELECTRA-small 的批量大小设置为 32。模型训练在 AutoDL GPU 环境中完成，代码由 Python、PyTorch、Transformers 和 scikit-learn 实现。主要训练参数如表 6 所示。

Table 6. Main training parameters
表 6. 主要训练参数

参数	设置
最大序列长度	128
训练轮数	3
优化器	AdamW
学习率	2e−5
BERT/RoBERTa/ERNIE/MacBERT batch size	16
TinyBERT/ELECTRA-small batch size	32
评价重点	Macro F1 与 Label 1 F1

4.5. 评价指标

本文采用 Accuracy、Precision、Recall 和 F1 值评价模型效果。由于三类标签分布并不完全均衡，单纯 Accuracy 可能受到多数类影响，因此本文重点关注 Macro F1。Macro F1 对每个类别的 F1 值进行平均，能够更好反映模型对少数类和难分类类别的综合识别能力。此外，本文单独报告 Label 0、Label 1 和 Label 2 的 F1 值，其中 Label 1 表示上下文依赖评论，是本文最关注的类别之一。

5. 多模型测试结果与分析

5.1. 模型总体表现

如表 7 所示，8 类模型在 5 次数据划分下的测试集平均结果。该表用于呈现模型总体性能。结果显示，预训练语言模型整体优于传统机器学习方法。其中，ERNIE-3.0-base 的平均 Macro F1 最高，达到 0.8265，平均 Accuracy 为 0.9006；MacBERT-base 的平均 Macro F1 为 0.8235，RoBERTa-wwm 为 0.8219，二者与 ERNIE-3.0-base 的差距较小。传统模型中，TF-IDF + SVM 优于 TF-IDF + Logistic Regression，但其平均 Macro F1 仅为 0.7560，仍明显低于主要预训练语言模型。

Table 7. Average test results across five data splits
表 7. 五次数据划分下各模型测试集平均结果

模型	Accuracy	Macro P	Macro R	Macro F1	Label 0 F1	Label 1 F1	Label 2 F1
ERNIE-3.0	0.9006	0.8333	0.8213	0.8265 ± 0.0108	0.9440	0.7042	0.8313
MacBERT	0.8988	0.8359	0.8127	0.8235 ± 0.0063	0.9429	0.7015	0.8262
RoBERTa-wwm	0.8985	0.8326	0.8127	0.8219 ± 0.0073	0.9428	0.7057	0.8173
BERT-base	0.8939	0.8220	0.8051	0.8130 ± 0.0101	0.9402	0.6973	0.8015
ELECTRA-small	0.8832	0.8002	0.7700	0.7841 ± 0.0059	0.9350	0.6824	0.7350
TF-IDF + SVM	0.8597	0.7633	0.7497	0.7560 ± 0.0032	0.9194	0.6262	0.7224
TF-IDF + LR	0.8469	0.7290	0.7834	0.7525 ± 0.0063	0.9101	0.6350	0.7126
TinyBERT	0.8171	0.6062	0.4946	0.5073 ± 0.0306	0.8994	0.5447	0.0777

5.2. 多次划分下的稳定性

图 2 展示了主要预训练模型在 5 次划分中的 Macro F1 波动情况，而不是重复模型总体排名。从标准

差来看，ERNIE-3.0-base、MacBERT-base 和 RoBERTa-wwm 的 Macro F1 准差分别为 0.0108、0.0063 和 0.0073，整体波动较小，说明强预训练模型的效果并非由某一次划分偶然造成。相比之下，TinyBERT 的 Macro F1 标准差为 0.0306，波动相对更明显，说明轻量化模型对数据划分更加敏感。

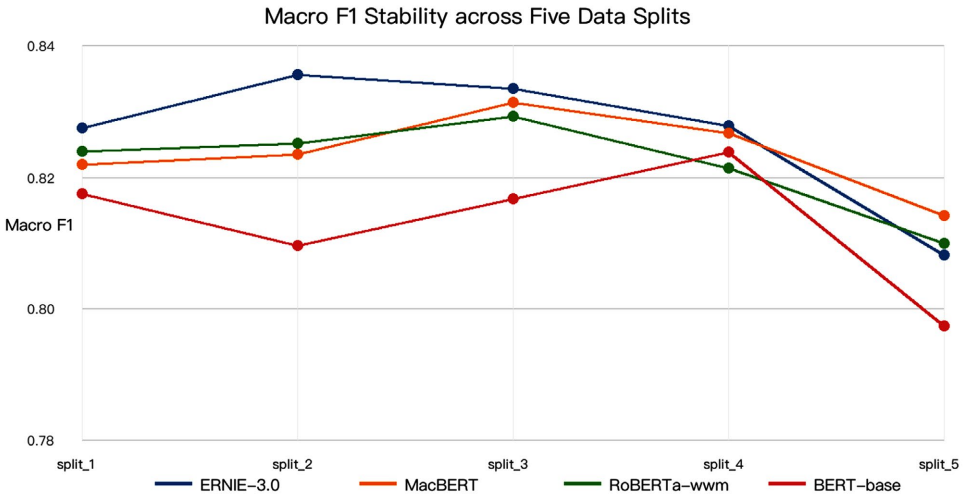

Figure 2. Macro F1 stability of main pre-trained models across five data splits
图 2. 主要预训练模型在五次划分下的 Macro F1 稳定性

5.3. 最优模型的预测结构

平均表现最优的 ERNIE-3.0-base 在 5 次测试集上的平均混淆矩阵如图 3 所示。与总体指标不同，混淆矩阵用于观察模型错误主要发生在哪些类别之间。结果表明，模型对 Label 0 的识别最稳定，因为完整语义评论通常具有明确对象和评价内容；Label 2 也具有较清晰的形式特征，例如信息请求、社交互动和无意义表达。相比之下，Label 1 更容易与 Label 0 混淆，说明上下文依赖评论虽然包含情绪或态度，但缺少明确评价对象，模型仅根据评论本身仍难以准确判断其是否需要结合原笔记理解。

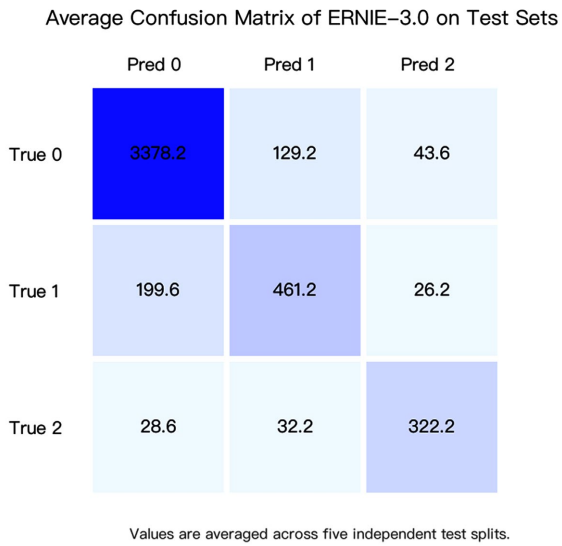

Figure 3. Average confusion matrix of ERNIE-3.0-base across five test sets
图 3. ERNIE-3.0-base 在五次测试集上的平均混淆矩阵

5.4. 结果小结

综合表 7、图 1 和图 2，本文实验得到三个主要结论。第一，预训练语言模型比传统词面特征模型更适合小红书旅游评论上下文依赖识别任务，说明上下文语义表示能力是该任务的关键。第二，ERNIE-3.0-base 在平均性能上最优，MacBERT-base 和 RoBERTa-wwm 表现接近，可作为后续研究的重要基线。第三，Label 1 是最难识别的类别，其难点来自评论文本自身的语义不完整，而不是简单的情感正负差异。这一结果也从实验层面支持了本文对小红书评论进行分层处理的必要性。

6. 下游主题建模验证实验

为进一步验证本文提出的评论分层策略在实际文本分析任务中的应用价值，本文设计了下游主题建模验证实验。该实验以 BERTopic 主题模型为基础，比较不同评论处理方式对主题建模结果的影响。实验目的并非重新评价分类模型本身，而是检验本文提出的三分类体系是否能够为后续旅游评论主题分析提供更有效的数据组织方式。

6.1. 实验目的与数据来源

本文选取大连城市的小红书旅游笔记－评论数据作为外部验证样本。该数据未参与前文分类模型训练，能够在一定程度上反映分类模型和分层策略在新城市数据上的应用效果。首先，本文调用前文训练效果较优的 ERNIE-3.0-base 分类模型，对大连评论数据进行三分类预测。共获得有效评论 9943 条，其中预测为 Label 0 的评论 7721 条，预测为 Label 1 的评论 1318 条，预测为 Label 2 的评论 904 条。随后，本文基于预测标签构建不同主题建模输入数据，并进行 BERTopic 对比实验。

6.2. 对比方案设计

为比较不同评论处理方式对主题建模质量的影响，本文设置三组实验方案。

Baseline 1 仅使用预测为 Label 0 的完整语义评论进行主题建模。该方案对应传统数据清洗中“只保留可独立理解评论”的处理方式。

Baseline 2 将全部评论直接输入 BERTopic，不区分评论是否具有完整语义，也不区分评论是否属于信息请求、社交互动或无效内容。该方案对应“全部评论无差别建模”的处理方式。

Proposed 组采用本文提出的分层处理策略。对于 Label 0 评论，直接作为独立文本输入主题模型；对于 Label 1 评论，将其与所属原笔记内容进行拼接后输入主题模型；对于 Label 2 评论，由于其主要表现为信息请求、社交互动或无有效态度表达，因此在主题建模前予以过滤。需要说明的是，小红书评论中的“[xxR]”和“[doge]”等 Emoji 表情在分类阶段被视为情感表达线索参与判断，但在 BERTopic 的主题词提取阶段，为避免 Emoji 标记本身成为主题词，本文对其进行了过滤处理。

6.3. BERTopic 模型设置与评价指标

主题建模阶段采用 BERTopic 模型。文本语义表示使用中文语义向量模型 BAAI/bge-small-zh-v1.5 生成，随后通过 UMAP 进行降维，并使用 HDBSCAN 完成密度聚类。主题关键词提取阶段使用基于中文分词的 CountVectorizer 以获得各主题的代表性关键词。为保证三组实验结果具有可比性，本文在三组实验中保持相同的模型参数设置，其中 HDBSCAN 的 min_cluster_size 设置为 80，min_samples 设置为 40，UMAP 的 n_neighbors 设置为 30。

实验主要从有效主题数、噪声比例、主题多样性和主题一致性四个方面评价主题建模质量。主题一致性采用 c-NPMI 指标衡量，用于反映同一主题内关键词之间的语义关联程度，数值越高表示主题内部

语义越集中。主题多样性用于衡量不同主题之间关键词的重复程度，数值越高说明主题之间区分度越高。噪声比例表示 BERTopic 中被 HDBSCAN 判定为 outlier 的文本比例，该指标可反映输入文本中难以形成稳定主题的样本占比。

6.4. 实验结果与分析

实验结果如表 8 所示。结果显示，本文提出的分层处理策略在主题一致性和主题多样性方面均优于两种基线方案。Proposed 组的主题一致性达到 0.5503，明显高于 Baseline 1 的 0.2036 和 Baseline 2 的 0.2043，说明经过分层处理后，同一主题内部关键词之间的语义关联更强，主题结构更加集中。与此同时，Proposed 组的主题多样性为 0.9125，高于 Baseline 1 的 0.9000 和 Baseline 2 的 0.8667，说明不同主题之间的关键词重复程度更低，主题区分度更好。

从有效主题数量看，Baseline 1 和 Baseline 2 均仅形成 3 个有效主题，而 Proposed 组形成 8 个有效主题。这说明分层策略并非简单增加文本数量，而是通过将上下文依赖评论与原笔记内容关联，使原本缺少评价对象的情绪评论获得明确语义指向。例如，“太真实了”、“确实”、“不建议”等评论在单独输入主题模型时难以判断其评价对象，而与原笔记拼接后，其语义可以被还原到住宿、交通、景点、消费或城市生活等更具体的主题语境中。因此，BERTopic 能够识别出更加细粒度的旅游体验主题。

从噪声比例看，Proposed 组的噪声比例为 5.31%，高于 Baseline 1 的 0.84%和 Baseline 2 的 0.44%。这一结果需要结合主题数量和主题一致性共同理解，Proposed 组形成的主题数量更多，主题边界更细，HDBSCAN 在聚类过程中会将部分无法稳定归入具体主题的文本判定为 outlier，因此噪声比例有所上升。但与此同时，Proposed 组的主题一致性和主题多样性均显著提高，说明其并非简单扩大噪声，而是在提高主题区分度的同时形成了更具解释力的主题结构。

Table 8. Comparison of BERTopic results under different comment processing strategies
表 8. 不同评论处理策略下的 BERTopic 结果比较

处理策略	输入文本说明	评论数	主题数	噪声比例	主题多样性	主题一致性
Baseline 1	仅输入 Label 0 评论	7718	3	0.84%	0.9000	0.2036
Baseline 2	全部评论直接输入	9812	3	0.44%	0.8667	0.2043
Proposed	Label 0 直接输入，Label 1 与原笔记拼接，Label 2 过滤	9036	8	5.31%	0.9125	0.5503

7. 研究发现与启示

本文围绕小红书旅游内容中的“笔记–评论”双层结构，构建了评论上下文依赖识别任务，并通过数据标注、模型比较和下游主题建模验证，考察了评论分层处理在旅游评论分析中的作用。总体来看，本文的研究发现主要体现在以下几个方面。

7.1. 笔记–评论双层结构下的评论分层价值

本文研究表明，小红书评论不能简单视为笔记正文的附属信息，也不能直接套用传统在线评论的单层文本处理逻辑。小红书评论区具有更强的互动性和语境依赖性，其中既包含可以独立理解的旅游体验评价，也包含大量需要结合原笔记才能理解的情绪回应、附和、调侃和质疑。这类上下文依赖评论虽然缺少完整评价对象，但仍然承载了用户的态度、情绪和群体反馈。

因此，本文提出的三分类体系并不是简单判断评论“有效”或“无效”，而是将评论处理转化为分层组织策略：完整语义评论可作为独立文本直接分析；上下文依赖评论应与原笔记建立关联后再分析；

无有效态度评论则可根据研究目的在后续建模前过滤。该思路有助于避免在清洗阶段误删有价值的情绪评论，也能减少信息请求、社交互动和无意义文本对后续分析的干扰。

7.2. 上下文依赖评论的识别难点

实验结果显示，Label 1 的识别效果普遍低于 Label 0 和 Label 2，说明上下文依赖评论是本任务的主要难点。其原因在于 Label 1 评论并非没有意义，而是意义不完整。例如“确实”、“太真实了”、“笑死”、“值得”等评论能够表达态度，但缺少明确评价对象，必须回到原笔记中才能判断其真实指向。

这也说明，仅依靠评论文本本身进行分类存在天然限制。预训练语言模型能够捕捉情绪和语义线索，但在没有原笔记上下文的情况下，仍难以判断某些短评论究竟是完整评价还是上下文依赖回应。因此，Label 1 的低识别效果不仅反映模型能力边界，也反映了小红书评论文本的结构性特征。

7.3. 对后续主题情感分析的启示

本文研究表明，小红书旅游评论分析应从“单条评论是否有效”的判断，进一步转向“评论与笔记之间如何关联”的判断。对于完整语义评论，可以直接进入主题建模、情感分析或游客感知挖掘；对于上下文依赖评论，应通过与原笔记拼接、建立上下文关联或联合建模等方式恢复其评价对象；对于无有效态度评论，则应根据研究目标进行过滤或单独处理。

下游主题建模验证进一步说明，评论分层策略不仅具有分类意义，也具有应用价值。它能够在保留上下文依赖评论情绪信息的同时减少无效互动内容对主题建模的干扰，从而为小红书“笔记–评论”融合分析提供更合理的数据入口。这表明，本文提出的评论上下文依赖识别任务可以作为后续旅游主题识别、游客感知分析和目的地形象研究的前置模块。

7.4. 研究不足与未来方向

本文仍存在一定局限。首先，当前分类模型主要以评论文本本身作为输入，尚未直接引入原笔记正文、评论位置和回复关系等上下文信息，因此对部分上下文依赖评论的识别仍受到信息不足的限制。其次，上下文依赖评论本身具有一定主观判断空间，边界样本在标注过程中仍可能产生分歧。最后，本文数据主要来自青岛、成都、厦门等旅游城市，虽然补充使用大连数据进行了下游验证，但数据范围仍有进一步扩展空间。

未来研究可从三个方面推进。第一，引入原笔记正文、评论层级关系和互动关系，构建真正的笔记–评论联合建模方法。第二，将本文提出的分层标签体系应用于更多下游任务，如旅游主题识别、游客情感分析和目的地形象感知研究，以进一步验证其适用性。第三，扩展更多城市、不同类型目的地和不同平台数据，检验该分类体系在更广泛旅游场景中的泛化能力。

8. 结语

本文以小红书旅游评论为研究对象，围绕“笔记–评论”双层结构下评论语义是否完整、是否依赖原笔记以及是否具有有效态度表达的问题，构建了评论上下文依赖识别任务，并形成完整语义评论、上下文依赖评论和无有效态度评论三类标签体系。基于 30,802 条小红书旅游评论数据，本文完成了数据标注、标注一致性检验、类别分布分析、八类模型基准实验以及下游主题建模验证。

研究结果表明，预训练语言模型整体优于传统机器学习方法，其中 ERNIE-3.0-base 在多次数据划分下最高平均 Macro F1，但上下文依赖评论仍是三类评论中最难识别的类别。进一步的下游实验表明，基于本文标签体系的分层处理策略能够改善 BERTopic 主题建模的数据输入质量，提升主题结果的语义一致性和可解释性。

总体而言，本文研究的主要价值在于将小红书评论处理从简单的数据清洗问题转化为面向“笔记–评论”融合的语义分层问题。该分类体系既可以用于评论数据清洗和保留策略设计，也可以作为后续主题建模、情感分析和旅游体验研究的数据组织基础。未来随着原笔记内容、评论互动关系和更多城市数据的引入，小红书旅游评论的上下文依赖识别与融合分析仍有进一步拓展空间。

参考文献

[1] Wang, Z., Huang, W.J. and Liu-Lastres, B. (2022) Impact of User-Generated Travel Posts on Travel Decisions: A Comparative Study on Weibo and Xiaohongshu. *Annals of Tourism Research Empirical Insights*, **3**, Article ID: 100064. https://doi.org/10.1016/j.annale.2022.100064

[2] Saini, H., Kumar, P. and Oberoi, S. (2023) Welcome to the Destination! Social Media Influencers as Cogent Determinant of Travel Decision: A Systematic Literature Review and Conceptual Framework. *Cogent Social Sciences*, **9**, Article ID: 2240055. https://doi.org/10.1080/23311886.2023.2240055

[3] Laureate, C.D.P., Buntine, W. and Linger, H. (2023) A Systematic Review of the Use of Topic Models for Short Text Social Media Analysis. *Artificial Intelligence Review*, **56**, 14223-14255. https://doi.org/10.1007/s10462-023-10471-x

[4] Grootendorst, M. (2022) BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. arXiv: 2203.05794.

[5] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805.

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., *et al*. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692.

[7] Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., *et al*. (2019) ERNIE: Enhanced Representation through Knowledge Integration. arXiv: 1904.09223.

[8] Cui, Y., Che, W., Liu, T., Qin, B., Wang, S. and Hu, G. (2020) Revisiting Pre-Trained Models for Chinese Natural Language Processing. *Findings of the Association for Computational Linguistics*: *EMNLP* 2020, 16-20 November 2020, 657-668. https://doi.org/10.18653/v1/2020.findings-emnlp.58

[9] Clark, K., Luong, M.T., Le, Q.V. and Manning, C.D. (2020) ELECTRA: Pre-Training Text Encoders as Discriminators Rather than Generators. arXiv: 2003.10555.

[10] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., *et al*. (2020) TinyBERT: Distilling BERT for Natural Language Understanding. *Findings of the Association for Computational Linguistics*: *EMNLP* 2020, 16-20 November 2020, 4163-4174. https://doi.org/10.18653/v1/2020.findings-emnlp.372

[11] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M. and Gao, J. (2021) Deep Learning-Based Text Classification: A Comprehensive Review. *ACM Computing Surveys*, **54**, 1-40. https://doi.org/10.1145/3439726