基于时间序列突发特征与 LLaMA 的虚假评论检测

李 东[1] 邓秋菊[2]
LI Dong DENG Qiuju

摘 要 虚假评论的广泛存在已经严重威胁到电子商务平台和在线点评系统的公信力。传统的虚假评论检测方法多依赖于评论文本特征或用户行为特征，尽管在一定程度上提升了检测效果，但在面对协同刷评、模板化评论和混合策略攻击时仍表现不足。基于此，文章提出了一种融合时间序列特征突发性分析与大语言模型文本抽取的虚假评论检测框架。首先，在评论时间轴上构建多维序列，包括评论数量、平均评分以及单次评论用户比例，并提取一系列时序特征，如窗口突发强度、趋势斜率变点和时滞相关峰值，以捕捉群体性刷评的动态特征；其次，利用大语言模型对评论文本进行结构化抽取，生成情感极性、具体性与可核查性、一致性、风格异常等高层次语义特征，从而弥补传统浅层文本特征的不足；最后，将时序特征与文本特征融合输入监督学习模型进行判别。基于 Yelp 和 Amazon 公共数据集的实验表明，所提方法能够在评分异常发生前 9 天提前识别可疑评论窗口，展现出较高的实用性和可解释性，为应对复杂和动态的虚假评论行为提供了新的解决思路。

关键词 虚假评论；时间序列特征；大语言模型

doi：10.3969/j.issn.1672-9528.2026.01.039

0 引言

在电子商务和在线点评平台中，用户评论不仅是消费者了解商品和服务质量的重要参考，也是商家维护声誉与塑造品牌形象的关键因素 [1]。然而，虚假评论的泛滥对平台的生态构成了严重威胁。一方面，商家通过刷评或购买虚假好评的方式夸大产品优势，误导消费者作出不合理的购买决策；另一方面，竞争对手可能通过恶意差评打击对手，造成不公平竞争 [2]。已有研究表明，在部分平台上，虚假评论比例可高达 20% 左右 [3]，其负面影响不仅体现在用户体验上，还会导致平台信誉下降，进而影响长期发展。虚假评论检测一直是学术界和工业界关注的重要问题。早期方法多依赖于文本特征分析，例如利用词汇分布、情感极性和语言模式等构建特征，再通过机器学习分类器进行判别 [4-5]。随着深度学习的兴起，研究者引入卷积神经网络 [6]、集成人工注意力机制的 Transformer 模型 [7] 以及预训练语言模型（如 BERT）[8] 来捕捉评论的语义信息，从而在一定程度上提升了检测的准确性。与此同时，用户行为特征也被用于虚假评论检测，如用户的评论频率、活跃时间模式、评分偏好等。这些方法能够揭示部分异常行为特征，但在面对协同攻击时仍存在不足。另一类重要研究是基于图的集体推断方法 [9]，将用户、评论与商品建模为异构图，在全局范围内识别可疑节点。

尽管上述方法取得了进展，但其在两个方面仍面临挑战。首先是动态性不足。大多数方法对评论的分析仍停留在静态特征层面，难以捕捉短时间内评论突发或多个用户协同刷评的动态特征。事实上，虚假评论往往伴随着明显的时间序列异常，例如在短时间内评论量的集中爆发以及平均评分的突然偏移，而这些动态模式在静态特征下难以被有效识别。其次是语义理解受限。浅层文本特征和小规模模型往往无法充分理解评论中的语义细节，尤其是在虚假评论采用模板化语言或混合噪声策略时，这种局限性尤为突出。针对这些挑战，时间序列突发性分析与大语言模型的结合为虚假评论检测提供了新的方向。一方面，时间序列分析能够从群体行为中捕捉异常模式，如评论数量与评分之间的异常相关性、斜率变点和时滞相关峰值，从而识别潜在的刷评行为；另一方面，大语言模型在自然语言理解方面展现了前所未有的能力，能够提取评论的具体性、可核查性以及与商品元数据的一致性，发现过度模板化和风格异常的评论。两者的结合既能够利用时序特征发现宏观的群体异常，又能通过语义特征分析增强

1. 四川晨豹互联网科技有限公司研发部 四川达州 635000
2. 重庆移通学院公共大数据安全技术重点实验室 重庆 401420
[基金项目] 达州市对外合作项目（18DWHZ0002）；綦江区科技计划项目（2025034）；重庆移通学院校级应用研究项目（KY2024013）

微观的评论判别力，从而在检测准确性与可解释性之间取得平衡。

本文提出了一种结合时间序列突发性特征与大语言模型文本特征的虚假评论检测框架。该框架首先在评论时间序列上提取多维突发特征，包括突发强度、趋势斜率变点和时滞相关峰值，以捕捉潜在的异常行为。随后，利用大语言模型对评论文本进行结构化特征抽取，生成情感极性、具体性与可核查性、一致性、风格异常等高层次语义特征。最后，将时序特征与文本特征通过监督学习模型进行融合判别。基于 Yelp 和 Amazon 公共数据集的实验结果显示，该方法在检测性能上显著优于已有经典方法，并能够提前识别异常窗口，具备较高的实际应用价值。本文的贡献不仅在于提出了一种多模态融合的虚假评论检测框架，还在于探索了时间序列分析与大语言模型在该问题上的互补性，为未来的研究提供了新的思路。

1 相关工作

围绕虚假评论检测，近年来的研究从时序、文本语义建模与图学习与群体协同三个方向持续推进，并逐步向多模态融合演化。首先，在时序与行为维度，研究者开始显式刻画“突发—衰减”的动态过程及其与评分 / 交互的耦合特征。Wang 等 [10] 提出“考虑时序爆发特性的可疑度框架”，使用三维时间序列刻画评论量、评分与行为强度的联动，并基于爆发度与时间权重构造窗口级嫌疑分数，在 Yelp 和 Amazon 上取得稳健提升，证明了“时间突发 + 评分漂移”是一类具有普适性的可解释信号。

在文本语义维度，深度预训练模型成为主流。Refaeli [11] 等提出 BERT 微调的系统评估，在平衡的人为标注数据集上显著优于传统特征 +SVM 基线，说明上下文语义对虚假评论判别的价值；并讨论了跨域泛化与类别不平衡对阈值与校准的影响。同期有应用性研究指出，仅依赖静态特征难以及时发现“短窗注入”，而通过在评论量、评分和行为维度上捕捉时序峰值，可在评分偏移前有效预警 [12]。

此外，图神经网络算法也逐渐成为虚假评论检测的重要方向。Li 等 [12] 提出的 GAS 框架利用图卷积网络建模用户 – 评论 – 商品关系，在大规模数据集上显著优于传统分类器。Cao 等 [13] 提出 REAL 方法，利用图卷积与聚类机制识别协作刷评团伙；Wang 等 [14] 则构建评论者共审关系的马尔可夫随机场进行连环刷评团伙判别。这些方法代表了近年来图模型在群体行为检测中的最新进展。

综上所述，近年来虚假评论检测的研究不断向多模态与深度建模方向发展。时间序列特征能够揭示群体性的异常注入行为，深度语义模型则捕捉评论内容的细粒度特征，而图神经网络则提供了建模用户 – 评论 – 商品交互关系的新途径。本文正是结合时间序列与大语言模型两类特征，力求在准确性、效率与可解释性之间取得平衡。

2 方法

本文的检测框架结合了时间序列突发性特征与大语言模型抽取的文本特征。与已有研究相比，在时序部分延续了评论数量、评分和用户比例等经典特征的建模方式，而主要创新在于引入大语言模型作为结构化特征抽取器，并与时序特征进行多模态融合。这样既能从宏观维度捕捉刷评活动的突发模式，又能从微观维度理解评论文本的细节，从而在检测准确性和可解释性之间取得平衡。

2.1 时间序列特征

为刻画评论行为的动态特征，在滑动窗口内提取以下核心指标。首先是评论数量的突发强度，采用指数加权移动平均（EWMA）来捕捉评论量的偏离程度。设时间步 t 的评论数为 C_t，则：

$$\mu_t=\alpha C_t+(1-\alpha)\mu_{t-1},\sigma_t^2=\alpha(C_t-\mu_t)^2+(1-\alpha)\sigma_{t-1}^2 \quad (1)$$

分数定义为：

$$z_t=\frac{C_t-\mu_t}{\sigma_t+\varepsilon} \quad (2)$$

窗口内的最大 z_t 即为该窗口的突发强度特征。其次是趋势斜率的变点。通过分段线性回归检测评论量或评分曲线的斜率跳变点，若在某一时刻 τ 出现显著变化：

$$\Delta s(\tau)=\widehat{s_{\mathrm{right}}}-\widehat{s_{\mathrm{left}}} \quad (3)$$

则说明评论行为存在异常趋势。最后是评论数与平均评分的时滞相关性。正常情况下两者大体同步，而在刷评场景中往往是评论量先爆发，评分随后才发生偏移。利用互相关函数：

$$\rho_k=\frac{\sum_t(C_t-\overline{C})(R_{t+k}-\overline{R})}{\sqrt{\sum_t(C_t-\overline{C})^2}\sqrt{\sum_t(R_{t+k}-\overline{R})^2}} \quad (4)$$

寻找非零滞后下的相关峰值。如果在 $k\neq 0$ 时 $|\rho_k|$ 显著高于背景水平，则提示存在潜在的“先刷评论再拉高评分”的模式。这三类指标构成了本文的时间序列特征集合，既直观可解释，又能够捕捉虚假评论的群体性动态特征。

2.2 大语言模型的结构化特征抽取

在文本特征的建模过程中，本文采用了“轻量化自然语言处理器与大语言模型相结合”相结合的策略，以兼顾计算效率与语义理解能力。具体而言，首先利用 LTP 工具完成中文分词和依存句法分析，获得评论的基本结构信息，同时通过 Text-CNN 情感分类器生成情感极性初值。这些基础处理结果并非直接作为最终输入，而是作为提示性信息，为大语

言模型的调用提供上下文约束，从而降低模型的计算压力并提升结果的稳定性。

在此基础上，本文引入 LLaMA-2 作为核心的语义特征抽取器，以结构化的方式输出评论的高层次语义表示。与传统的词袋模型或直接端到端分类不同，将 LLaMA-2 作为冻结的语义编码器，直接从其最后一层隐状态中抽取向量表示，再通过轻量化的任务特定输出头生成所需的结构化特征。具体而言，给定评论文本 T_r，首先将其输入至 LLaMA-2 模型，得到 token 级隐状态序列 $\boldsymbol{H}_r \in \mathbf{R}^{L\times d}$，其中 L 为文本长度；d 为隐藏维度。随后对 H_r 进行池化运算以获得评论级表示：

$$h_r=\frac{1}{L}\sum_{i=1}^{L} bH_{r,i} \tag{5}$$

该评论向量在保持上下文语义完整性的同时，避免了额外的提示开销在此基础上，设计了五个轻量化的预测头部，每个头部均为单层线性映射加 Sigmoid 激活，分别输出情感极性 $e(r)$、具体性 $q(r)$、可核查性 $v(r)$ 与模板化 / 风格异常概率 $t(r)$：

$$e(r)=\sigma\left(w_e^{\top} h_r+b_e\right) \tag{6}$$

$$q(r)=\sigma\left(w_q^{\top} h_r+b_q\right) \tag{7}$$

$$v(r)=\sigma\left(w_v^{\top} h_r+b_v\right) \tag{8}$$

$$t(r)=\sigma\left(w_t^{\top} h_r+b_t\right) \tag{9}$$

训练阶段，本文利用规则构造与弱监督信号（如情感词典与 Text-CNN 输出用于 $e(r)$，正则匹配的实体覆盖率用于 $q(r)$，属性匹配规则用于 $v(r)$ 等）为各任务生成伪标签，并采用加权二元交叉熵损失进行多任务联合优化：

$$\mathcal{L}=\lambda_e \mathcal{L}_e+\lambda_q \mathcal{L}_q+\lambda_v \mathcal{L}_v+\lambda_t \mathcal{L}_t \tag{10}$$

式中：λ 为平衡各任务的重要性系数。

实践中，LLaMA-2 的底座参数保持冻结，仅训练任务头部并通过 LoRA 对部分层进行参数高效微调，以保证计算效率与迁移性。

在窗口级别，本文对评论的特征向量进行聚合。连续特征（如情感分数）计算均值、方差与分位数统计，离散特征（如具体性标签）计算比例与熵，语义嵌入则通过窗口内评论向量的余弦相似度分布得到平均值与标准差，用以度量评论集合的聚集度与离散度。最终得到的窗口级文本特征综合反映了该时间段内评论的语义分布特征，可与时序突发特征进行多模态融合。与传统黑箱神经网络判别方式相比，这种基于编码器与轻量输出头的结构化特征抽取方式不仅在性能上稳健，而且因其明确的语义维度划分与可追溯的弱监督信号而具有更高的可解释性。

2.3 多模态融合框架

在获得时序特征和文本特征后，本文提出一个多模态融合框架，将二者结合进行虚假评论检测。核心思路是先利用时间序列特征快速筛选候选窗口，再在这些窗口中调用大语言模型抽取语义特征，最后将两类特征拼接并输入分类器，从而在检测准确性与效率之间取得平衡。

具体而言，候选生成阶段通过突发强度、斜率变点和时滞相关峰值定位潜在异常窗口，减少不必要的 LLM 调用。随后，采用早期融合策略，将时序特征与文本特征拼接为统一向量，送入 XGBoost 分类器。该模型在处理高维特征和类别不平衡时具有优势，并能输出特征重要性以增强可解释性。实验中使用加权交叉熵损失，并通过交叉验证选择参数。模型最终输出窗口级可疑分数，并结合评论特征生成嫌疑评论清单。

该框架的优势在于宏观与微观层面的互补：时序特征捕捉群体刷评的突发模式，文本特征揭示评论内容的异常信号，两者结合能在准确性与可解释性之间达到更优平衡。

3 实验

本文在 Yelp 和 Amazon 两个公开数据集上验证了所提方法。Yelp 数据集由官方过滤器提供虚假评论标签，Amazon 数据集则包含电子产品和服装两个子集，虚假评论标签通过既有研究方法获得。采用滑动窗口提取时间序列特征，并在候选窗口调用“本地抽取器 + 大语言模型”生成文本特征，两类特征拼接后输入 XGBoost 分类器。

为了对比效果，选取了几类常见方法作为基线：（1）朴素贝叶斯（NB）：基于 n-gram 特征的传统文本分类方法；（2）支持向量机（SVM）：利用 TF-IDF 向量训练的线性分类器；（3）LSTM：直接对评论序列建模的深度学习方法；（4）BERT：采用预训练语言模型对评论进行分类的代表性方法。这些方法覆盖了从传统机器学习到深度学习的不同路线，能够全面反映本文方法的优势。评价指标方面，采用 AUPRC、Precision@100 和提前量（Lead Time），分别衡量整体性能、前端检出能力和预警效果。

实验结果如表 1 所示。可以看到，本文方法在两个数据集上均取得最佳表现，其中在 Precision@100 上相较 SVM 提升超过 9%，在 Amazon 数据集上的提前量平均达到 9 天，明显优于基线方法。同时证明，在虚假评论检测过程中，采用时间序列特征与大语言模型文本特征的多模态融合框架不仅在方法设计上具有合理性，在实际实验中同样展现出优于传统方法的检测性能。尤其是在 Precision@100 与提前量等更贴近应用场景的指标上，本方法表现出明显优势，说明其不仅能够识别已发生的虚假评论，还具备提前预警的能力。总体而言，所提出的框架在准确性、效率与可解释性之间取得了较好的平衡，为虚假评论检测提供了一种切实可行且具有

实用价值的解决方案。

表 1 实验结果

方法	Yelp AUPRC	Yelp P@100	Amazon AUPRC	Amazon 提前量 / 天
NB	0.788	73%	0.701	5
SVM	0.756	76%	0.673	6
LSTM	0.768	75%	0.732	8
BERT	0.772	80%	0.721	8
OURS	0.812	82%	0.791	9

4 结语

本文提出了一种结合时间序列突发性特征与大语言模型文本特征的多模态虚假评论检测框架。通过在评论时间轴上提取突发强度、斜率变点与时滞相关峰值等动态指标，并在候选窗口中利用本地抽取器与大语言模型生成结构化语义特征，系统实现了宏观群体行为与微观文本语义的双重建模。实验结果表明，该方法在 Yelp 和 Amazon 数据集上均显著优于传统文本分类和单一时序方法，能够更早、更准确地识别潜在虚假评论行为，并在 Precision@100 与提前量指标上展现出突出的价值。

未来工作可从两个方向展开：一方面，将多模态特征与图神经网络结合，进一步捕捉用户 – 评论 – 商品之间的复杂交互关系；另一方面，探索增量学习与在线检测机制，使系统能够在持续更新的评论流中实时识别虚假评论。总体而言，本文的研究为应对复杂多变的虚假评论行为提供了新的思路与可行路径，对电商平台和在线点评系统的健康生态具有重要意义

参考文献：

[1] 秦益德 . 浅论电商平台的客户评价管理策略 : 如何提升用户信任度 [J]. 电子商务评论 , 2024, 13(4):5740-5745.

[2] 周沫 . 虚假评论对电商平台信誉度的影响及治理对策 [J]. 电子商务评论 , 2025, 14(4):259-265.

[3]LUCA M, ZERVAS G. Fake it till you make it: reputation, competition, and Yelp review fraud[J]. Management science, 2016, 62(12): 3412-3427.

[4]JINDAL N, LIU B. Opinion spam and analysis[C]//Proceedings of the 2008 international conference on web search and data mining. NewYork:ACM,2008: 219-230.

[5] OTT M, CHOI Y, CARDIE C, et al. Finding deceptive opinion spam by any stretch of the imagination[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Brussels:ACL,2011:309–319.

[6] VENKATESH B, RAMNARESH YADAV B V . HACNN: hierarchical attention convolutional neural network for fake review detection[J/OL]. Social network analysis and mining, 2024[2025-12-02].https://link.springer.com/article/10.1007/s13278-024-01380-0.

[7]MOHAWESH R, XU S X, SPRINGER M, et al. Fake or genuine? contextualised text representation for fake review detection[EB/OL].(2021-12-29)[2025-12-20].https://doi.org/10.48550/arXiv.2112.14343.

[8] SUN P F, BI W H, ZHANG Y F, et al. Fake review detection model based on comment content and review behavior[J]. Electronics, 2024, 13(21): 4322.

[9]YU S, REN J, LI S H, et al. Graph learning for fake review detection[J/OL]. Frontiers in artificial intelligence, 2022[2025-12-02].|https://doi.org/10.3389/frai.2022.9225899.

[10] WANG N, YANG J, KONG X F, et al. A fake review identification framework considering the suspicion degree of reviews with time burst characteristics[J]. Expert systems with applications, 2022, 190: 116207.

[11] REFAELI D, HAJEK P. Detecting fake online reviews using fine-tuned BERT[C]//Proceedings of the 2021 5th International Conference on E-Business and Internet.NewYork:ACM, 2021: 76-80.

[12] LI A, QIN Z, LIU R S, et al. Spam review detection with graph convolutional networks[C]//Proceedings of the 28th ACM international conference on information and knowledge management. NewYork: ACM, 2019: 2703-2711.

[13] CAO C, LI S H, YU S, et al. Fake reviewer group detection in online review systems[C]//2021 International Conference on Data Mining Workshops (ICDMW). Piscataway: IEEE, 2021: 935-942.

[14]WANG Z, HU R L, CHEN Q, et al. ColluEagle: collusive review spammer detection using Markov random fields[J]. Data mining and knowledge discovery, 2020, 34:1621-1641.

【作者简介】

李东（1997—），男，四川达州人，本科，工程师，研究方向：自然语言理解。

邓秋菊（1987—），女，重庆铜梁人，本科，副教授，研究方向：图像处理。

（收稿日期：2025-09-04 修回日期：2026-01-05）