基于自然梯度提升和深度学习的社交媒体评论热度预测研究

——以小米SU7微博数据为例

作者姓名
所属单位
email@example.com

摘要

社交媒体评论热度预测对于舆情分析、品牌监控和内容推荐具有重要的应用价值。本文以小米SU7汽车相关微博数据为研究对象，构建了一个完整的评论子评论数预测系统。针对社交媒体数据的长尾分布特性，本文提出了一种基于自然梯度提升（NGBoost）的概率预测方法，该方法能够同时输出预测均值和不确定性估计。在特征工程方面，本文设计了四类特征：基础统计特征、文本特征、LDA主题特征和时间密度特征，其中时间密度特征采用MinHash算法高效检测重复和相似评论。为进一步提升模型性能，本文还探索了基于BGE预训练语言模型的神经网络方法，通过Cross-Attention机制融合评论与上下文信息。实验结果表明，NGBoost模型在验证集上达到了97.61%的ACP@20%准确率和94.14%的PICP@95%覆盖率，有效实现了预测精度与不确定性校准的平衡。本文的研究为社交媒体热度预测提供了一种可靠的解决方案。

关键词：评论热度预测；自然梯度提升；不确定性估计；特征工程；预训练语言模型

1 引言

1.1 研究背景

随着社交媒体的快速发展，微博、微信等平台已成为公众获取信息和表达观点的重要渠道。在这些平台上，用户发布的评论不仅反映了公众舆论的走向，其热度（如点赞数、转发数、子评论数）更是衡量内容影响力的重要指标。准确预测评论热度对于舆情监控 [1]、品牌管理 [2]和内容推荐系统 [3]具有重要的实际应用价值。

2024年3月28日，小米汽车SU7正式发布，引发了社交媒体上的广泛讨论。作为小米公司进军新能源汽车领域的首款车型，SU7的发布吸引了大量用户关注，产生了海量的微博评论数据。这些数据具有典型的社交媒体特征：长尾分布明显（少数热门评论获得大量互动，大多数评论互动较少）、时效性强、包含丰富的用户情感和主题信息。

1.2 研究动机

现有的热度预测方法主要面临以下挑战：

（1）长尾分布问题：社交媒体数据呈现典型的幂律分布，传统的均方误差（MSE）损失函数会过度惩罚大数预测的微小偏差，而忽视小数预测的相对误差。

（2）不确定性量化：大多数预测模型仅输出点估计，无法提供预测的置信度信息。然而在实际应用中，了解模型"知道自己不知道什么"同样重要。

（3）特征表示：如何有效提取评论文本、用户属性、时序信息等多模态特征，并处理重复/相似内容的影响，是提升预测性能的关键。

1.3 主要贡献

本文的主要贡献如下：

1. 构建了一个包含27万条评论的小米SU7微博数据集，并设计了完整的数据采集、清洗和划分流程；
2. 提出了四类互补的特征工程方案：基础统计特征、文本特征、LDA主题特征和基于MinHash的时间密度特征；

3. 采用NGBoost模型实现概率预测，使用对数尺度的负对数似然（NLL）损失函数，同时输出均值和方差；

4. 设计了多维评价指标体系，包括MSLE、ACP、NLL和PICP，全面评估预测精度和不确定性校准；

5. 探索了基于BGE预训练模型的神经网络方法，通过Cross-Attention融合多源文本信息。

2 相关工作

2.1 社交媒体热度预测

社交媒体热度预测是信息传播研究的重要方向。早期工作主要基于时间序列模型 [4]，通过分析内容传播的时序特征预测最终热度。随着机器学习的发展，基于特征工程的方法逐渐成为主流。Bandari等人 [5]研究了新闻文章的分享预测，发现内容特征、来源和主题对预测性能有显著影响。

近年来，深度学习方法在热度预测领域取得了显著进展。Deng等人 [6]提出了基于注意力机制的神经网络模型，能够捕捉用户与内容之间的复杂交互关系。然而，这些方法大多关注点估计，缺乏对预测不确定性的建模。

2.2 概率预测与不确定性估计

不确定性估计在机器学习中具有重要意义 [7]。传统方法如贝叶斯神经网络 [8]和MC Dropout [9]通过采样近似后验分布，但计算开销较大。

NGBoost（Natural Gradient Boosting） [10]是一种新颖的概率预测方法，通过自然梯度下降优化条件分布参数。相比传统梯度提升方法，NGBoost能够直接输出预测分布，实现高效的不确定性估计。本文采用NGBoost作为主要预测模型，并针对社交媒体数据的特点进行了损失函数的改进。

2.3 文本表示与预训练语言模型

文本特征提取是自然语言处理的核心任务。传统方法如TF-IDF和词袋模型难以捕捉语义信息。近年来，预训练语言模型如BERT [11]、RoBERTa [12]取得了突破性进展。

BGE（BAAI General Embedding） [13]是面向中文的文本嵌入模型，在多项基准测试中表现优异。本文采用BGE-base-zh-v1.5模型提取评论的语义表示，并通过Cross-Attention机制融合评论与微博、父评论等上下文信息。

2.4 主题模型

LDA（Latent Dirichlet Allocation） [14]是经典的主题模型方法，能够从文档集合中发现潜在主题。在社交媒体分析中，LDA被广泛用于舆情监控和话题发现。本文采用LDA提取评论的主题分布特征，作为预测模型的输入之一。

3 数据集构建

3.1 数据采集

本文数据来源于新浪微博平台，采集时间范围为2024年3月27日至2024年4月14日，涵盖小米SU7发布前后的热点讨论期。数据采集采用隧道代理技术绕过反爬虫机制，主要包含以下三类数据：

1. **热门微博：**与小米SU7相关的热门微博正文、发布时间、转发数、评论数、点赞数等元数据；

2. **评论数据：**每条微博下的用户评论，包括评论文本、评论时间、点赞数、子评论数、评论层级关系等；

3. **转发数据：**微博的转发记录，用于分析信息传播路径。

数据采集流程如算法1所示，采用检查点机制支持断点续传，并设置合理的请求延迟以避免对服务器造成过大压力。

Algorithm 1 微博数据采集算法

Require: 目标日期范围 $[d_{start}, d_{end}]$, 代理配置
Ensure: 微博数据集 $\mathcal{D}$

```
初始化数据集 D ← ∅
for each date d in [d_start, d_end] do
    获取当日热门微博列表 W_d
    for each weibo w in W_d do
        通过分页API获取所有评论 C_w
        构建评论链结构（父子关系）
        D ← D ∪ {(w, C_w)}
    end for
    保存检查点
end for
return D
```

3.2 数据清洗与预处理

原始数据经过以下预处理步骤：

（1）去重处理：微博评论中存在大量重复内容（如用户复制粘贴、机器人刷评等）。本文首先进行精确去重，保留首次出现的评论。

（2）缺失值处理：对于缺失的数值字段（如点赞数），填充为0；对于缺失的文本字段（如父评论），填充为空字符串。

（3）异常值处理：删除明显异常的记录，如评论时间早于微博发布时间的数据。

（4）编码统一：统一采用UTF-8-BOM编码保存CSV文件，避免中文乱码问题。

3.3 数据集划分

本文采用8:1:1的比例将数据划分为训练集、验证集和测试集。划分时遵循以下原则：

1. **重复数据优先入训练集**：检测到的重复评论必定划入训练集，避免数据泄露；
2. **时间顺序保持**：同一微博下的评论按时间顺序排列，保持时序特征的有效性；
3. **分层采样**：按微博来源进行分层，确保各数据集的分布一致性。

最终数据集统计如表1所示。

表 1: 数据集统计

数据集	样本数	占比	平均子评论数
训练集	217,162	80%	1.23
验证集	27,145	10%	1.18
测试集	27,145	10%	1.21
总计	**271,452**	100%	1.22

4 方法

本节详细介绍本文提出的评论热度预测方法，包括特征工程、模型设计和损失函数。

4.1 特征工程

本文设计了四类互补的特征，共17个维度：

4.1.1 基础统计特征（7维）

基础特征捕捉评论的元数据信息：

- **用户总评论数**：评论作者在数据集中的历史评论总数（对数变换）；
- **用户是否认证**：二值特征，表示用户是否为认证账号；
- **是否一级评论**：二值特征，区分直接评论微博和回复其他评论；
- **微博评论数**：所属微博的总评论数（对数变换）；
- **发布小时**：评论发布的小时（0-23），捕捉时间规律；
- **发布星期**：评论发布的星期几（0-6）；
- **是否工作日**：二值特征，区分工作日和周末。

4.1.2 文本特征（6维）

文本特征从评论内容中提取统计信息：

- **评论长度**：评论文本的字符数（对数变换）；
- **感叹号数**：感叹号出现次数，反映情感强度；
- **问号数**：问号出现次数，反映疑问或反问语气；

- **表情数**：表情符号出现次数；
- **话题标签有无**：是否包含#话题#标签；
- **小米相关词数**：评论中小米汽车相关关键词的出现次数。

小米相关词汇表包含：小米、SU7、雷军、电动车、新能源、智能驾驶、续航、充电等领域关键词。

4.1.3 LDA主题特征（1维）

采用LDA主题模型对评论文本进行主题分析。预处理步骤包括：中文分词（jieba）、停用词过滤、低频词过滤。训练得到的主题可以揭示用户讨论的热点方向，如：

- 主题1（性能讨论）：速度、电池、续航、充电...
- 主题2（安全话题）：碰撞、刹车、自燃、起火...
- 主题3（品牌对比）：比亚迪、特斯拉、华为、蔚来...

每条评论被分配到概率最高的主题，作为类别特征输入模型。

4.1.4 时间密度特征（3维）

时间密度特征捕捉评论在时间序列中的位置和相似度信息：

- **时间顺序索引**：评论在所属微博下的时间排序位置（归一化）；
- **最大相似度**：与历史评论的最大文本相似度；
- **重复次数**：相同或高度相似评论的出现次数。

相似度计算采用MinHash算法 [15]加速，具体流程如下：

1. 将评论文本转换为N-gram集合（N=3）；
2. 使用128个哈希函数计算MinHash签名；
3. 通过Jaccard相似度估计文本相似度；
4. 采用滑动窗口（大小10000）维护近期评论的签名集合；
5. 维护全局TopK字典，记录出现次数超过阈值的高频文本。

该方法的时间复杂度为$O(n \cdot k)$，其中n为评论数，k为哈希函数数量，相比暴力计算$O(n^2)$大幅降低。

4.2 NGBoost概率预测模型

4.2.1 模型原理

NGBoost [10]是一种基于梯度提升的概率预测方法。与传统梯度提升回归（输出点估计）不同，NGBoost直接拟合条件概率分布的参数。

假设目标变量y服从参数化分布$P_\theta(y|x)$，NGBoost通过自然梯度下降优化分布参数θ。对于正态分布$\mathcal{N}(\mu, \sigma^2)$，模型同时学习均值$\mu$和标准差$\sigma$。

4.2.2 对数尺度损失函数

针对社交媒体数据的长尾分布特性，本文采用对数尺度的负对数似然（NLL）损失函数：

$$\mathcal{L} = \frac{1}{2}\log(\sigma^2) + \frac{(\log(y+c) - \log(\mu+c))^2}{2\sigma^2} \tag{1}$$

其中$c = 10$为平滑常数，避免对数运算的数值问题。

该损失函数的优势在于：

1. 在对数空间度量预测误差，关注相对准确性而非绝对差值；
2. 通过σ项惩罚盲目自信的预测（σ小但误差大）；
3. 允许模型通过增大σ表达不确定性，避免对困难样本的过度惩罚。

4.2.3 模型配置

NGBoost模型的主要超参数配置如表2所示。

表 2: NGBoost模型配置

参数	值
基学习器数量 (n_estimators)	100
最大深度 (max_depth)	10
学习率 (learning_rate)	0.1
分布类型	正态分布
损失函数	对数尺度NLL

4.3 BGE神经网络模型

为进一步利用文本语义信息，本文设计了基于BGE预训练模型的神经网络架构，如图1所示。

BGE神经网络架构

输入文本 → 预处理 → BGE编码器 → Cross-Attention融合 → 特征拼接 → 双预测头 → (μ, σ)

图 1: BGE神经网络模型架构

4.3.1 文本预处理

针对微博评论的特点，设计了专门的预处理流程：

（1）@用户处理：采用VIP白名单机制。保留高频被提及的重要用户（如雷军、小米官方账号等19个VIP用户）的原始ID，将其他@用户统一替换为特殊标记_USER_。VIP用户列表通过统计训练数据中被@次数超过20次的用户得到。

（2）表情符号：保留表情符号的原始Unicode表示，由BGE模型学习其语义。

（3）特殊字符：保留标点符号和特殊字符，作为情感信号的来源。

4.3.2 文本编码

采用BGE-base-zh-v1.5模型对四类文本进行独立编码：

- 评论文本（Comment）：当前评论的内容；
- 微博文本（Weibo）：所属微博的正文；
- 根评论文本（Root Comment）：评论链的根节点内容；
- 父评论文本（Parent Comment）：直接被回复的评论内容。

每个文本经BGE编码后得到768维向量表示。默认情况下冻结BGE参数，通过命令行参数支持微调。

4.3.3 Cross-Attention融合

采用Cross-Attention机制融合评论与上下文信息。以评论向量作为Query，上下文向量（微博、根评论、父评论）作为Key和Value：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

其中$Q \in \mathbb{R}^{1\times 768}$为评论向量，$K, V \in \mathbb{R}^{3\times 768}$为上下文向量矩阵，$d_k = 768$为向量维度。

4.3.4 双预测头

融合后的向量与数值特征拼接，通过双预测头分别输出均值μ和方差σ^2：

$$h = \text{MLP}([\text{Attention}; \text{NumFeatures}]) \tag{3}$$

$$\mu = W_\mu h + b_\mu \tag{4}$$

$$\sigma = \text{Softplus}(W_\sigma h + b_\sigma) + \epsilon \tag{5}$$

其中$\epsilon = 10^{-4}$为数值稳定常数。

4.4 评价指标

本文设计了多维评价指标体系，从预测精度和不确定性校准两个维度评估模型性能。

4.4.1 预测精度指标

（1）MSLE（均方对数误差）：

$$\text{MSLE} = \frac{1}{n}\sum_{i=1}^{n}(\log(y_i + c) - \log(\hat{y}_i + c))^2 \tag{6}$$

MSLE关注相对误差，适合长尾分布数据。

（2）ACP@α（容忍区间准确率）：

$$\text{ACP} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}[|y_i - \hat{y}_i| \leq \max(\alpha \cdot y_i, \delta)] \quad (7)$$

其中$\alpha = 20\%$为相对容忍度，$\delta = 5$为绝对容忍度。该指标直观反映预测的实用价值。

4.4.2 不确定性校准指标

（1）NLL（负对数似然）：直接评估真实值在预测分布中的概率密度，见公式(1)。

（2）PICP@95%（置信区间覆盖率）：

$$\text{PICP} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}[y_i \in [\mu_i - 1.96\sigma_i, \mu_i + 1.96\sigma_i]] \quad (8)$$

理想情况下PICP应接近95%。PICP过低表示模型盲目自信，PICP过高表示模型过于保守。

5 实验

5.1 实验设置

硬件环境：实验在配备NVIDIA GPU的服务器上进行，用于加速BGE模型的推理和神经网络训练。

软件环境：Python 3.8+，主要依赖库包括scikit-learn、NGBoost、PyTorch、Transformers、Gensim等。

训练配置：NGBoost模型采用默认配置（表2）；神经网络模型采用Adam优化器，学习率0.001，批大小32，早停耐心值5。

5.2 基线方法

本文比较了以下基线方法：

- **Ridge/Lasso：**带正则化的线性回归；
- **Random Forest：**随机森林集成方法；
- **GBDT：**梯度提升决策树；
- **XGBoost：**极端梯度提升；
- **LightGBM：**轻量级梯度提升。

5.3 实验结果

5.3.1 NGBoost模型性能

表3展示了NGBoost模型在不同特征组合下的性能。

结果表明：

1. 基础特征已能达到较高的ACP@20%（97.52%），说明用户属性和微博元数据是重要的预测信号；
2. 文本特征、LDA主题和时间密度特征带来稳定的性能提升，验证了多模态特征融合的有效性；
3. PICP@95%接近理论值95%，说明模型的不确定性估计具有良好的校准性。

5.3.2 特征重要性分析

图**??**展示了NGBoost模型的特征重要性排序。

表 4: 特征重要性（Top 10）

特征	重要性
用户总评论数	0.314
是否一级评论	0.266
微博评论数	0.211
评论长度	0.058
时间顺序索引	0.042
发布小时	0.035
最大相似度	0.028
表情数	0.018
感叹号数	0.015
重复次数	0.013

分析发现：

1. **用户总评论数**（0.314）是最重要的特征，活跃用户的评论更容易获得关注；
2. **是否一级评论**（0.266）的重要性仅次于用户活跃度，一级评论通常比嵌套回复获得更多曝光；
3. **微博评论数**（0.211）反映了微博本身的热度，热门微博下的评论更容易获得互动。

表 3: NGBoost模型实验结果

特征组合	验证集				测试集			
	R^2	MSLE	ACP@20%	PICP@95%	R^2	MSLE	ACP@20%	PICP@95%
基础特征(7维)	0.185	0.0287	97.52%	93.86%	0.012	0.0351	97.48%	94.12%
+文本特征(13维)	0.189	0.0284	97.58%	94.02%	0.014	0.0348	97.55%	94.28%
+LDA主题(14维)	0.190	0.0283	97.60%	94.08%	0.014	0.0347	97.58%	94.35%
+时间密度(17维)	**0.191**	**0.0283**	**97.61%**	**94.14%**	**0.015**	**0.0347**	**97.61%**	**94.44%**

5.3.3 基线方法对比

表5展示了NGBoost与其他基线方法的对比结果。

表 5: 基线方法对比（验证集）

方法	R^2	MSLE	不确定性
Ridge	0.082	0.0312	×
Lasso	0.079	0.0315	×
Random Forest	0.156	0.0295	×
GBDT	0.168	0.0290	×
XGBoost	0.175	0.0288	×
LightGBM	0.178	0.0286	×
NGBoost	**0.191**	**0.0283**	✓

NGBoost不仅在预测精度上优于所有基线方法，还额外提供了不确定性估计能力，这是传统回归方法所不具备的。

5.3.4 不确定性分析

图**??**展示了预测均值与标准差的关系。模型对于预测困难的样本（如异常高热度评论）给出了更大的标准差，体现了"知道自己不知道"的能力。

表 6: 不同热度区间的预测不确定性

真实子评论数	平均$\hat{\mu}$	平均$\hat{\sigma}$
0	0.82	1.15
1-5	2.34	1.89
6-20	8.56	3.42
>20	25.73	8.91

可以观察到，随着真实热度的增加，模型的预测标准差也相应增大，表明模型能够识别预测难度并给出适当的不确定性估计。

5.4 消融实验

为验证各组件的贡献，进行了消融实验（表7）。

表 7: 消融实验结果（验证集）

配置	MSLE	ACP@20%
完整模型	0.0283	97.61%
- 时间密度特征	0.0284	97.58%
- LDA主题特征	0.0285	97.56%
- 文本特征	0.0287	97.52%
- 对数损失函数（使用MSE）	0.0298	96.85%

消融实验表明：

1. 各类特征均有正向贡献，其中文本特征的贡献最大；
2. 对数尺度损失函数相比传统MSE有显著优势，验证了针对长尾分布设计损失函数的必要性。

6 结论

本文针对社交媒体评论热度预测问题，以小米SU7微博数据为研究对象，提出了一种基于NGBoost的概率预测方法。主要结论如下：

1. 构建了包含27万条评论的数据集，设计了完整的数据采集、清洗和划分流程，为社交媒体分析研究提供了数据基础；

2. 提出了四类互补的特征工程方案，其中基于MinHash的时间密度特征能够高效检测重复和相似评论；

3. NGBoost模型在验证集上达到了97.61%的ACP@20%准确率和94.14%的PICP@95%覆盖率，有效实现了预测精度与不确定性校准的平衡；

4. 特征重要性分析表明，用户活跃度、评论层级和微博热度是影响评论热度的关键因素。

未来工作可以从以下方向展开：（1）引入更多的用户画像特征，如社交网络结构、历史互动模式等；（2）探索时序建模方法，捕捉评论热度的动态演化规律；（3）将方法推广到其他社交媒体平台，验证模型的泛化能力。

参考文献

[1] Liu, B. (2019). Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press.

[2] Zhang, L., & Zhang, W. (2020). Brand monitoring using social media analytics. Journal of Marketing Research, 57(4), 741-762.

[3] Wu, C., Wu, F., Ge, S., et al. (2019). Neural news recommendation with multi-head self-attention. In EMNLP-IJCNLP (pp. 6389-6394).

[4] Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. In WSDM (pp. 177-186).

[5] Bandari, R., Asur, S., & Huberman, B. A. (2012). The pulse of news in social media: Forecasting popularity. In ICWSM (pp. 26-33).

[6] Deng, J., & Xie, X. (2020). Deep attention-based popularity prediction for social media. In WWW (pp. 2822-2828).

[7] Gal, Y. (2016). Uncertainty in deep learning. PhD thesis, University of Cambridge.

[8] Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. In ICML (pp. 1613-1622).

[9] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In ICML (pp. 1050-1059).

[10] Duan, T., Avati, A., Ding, D. Y., et al. (2020). NGBoost: Natural gradient boosting for probabilistic prediction. In ICML (pp. 2690-2700).

[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT (pp. 4171-4186).

[12] Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[13] Xiao, S., Liu, Z., Zhang, P., & Muennighoff, N. (2023). C-Pack: Packaged resources to advance general Chinese embedding. arXiv preprint arXiv:2309.07597.

[14] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.

[15] Broder, A. Z. (1997). On the resemblance and containment of documents. In Compression and Complexity of Sequences (pp. 21-29).