

# 基于BGE和交叉注意力机制的 社交媒体评论热度预测

——以小米SU7微博数据为例

李俊凯

浙江工业大学信息工程学院

302023568066@zjut.edu.cn

## 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和预测不确定性量化问题，本文提出了一种基于BGE预训练语言模型和Cross-Attention机制的神经网络方法（BGE-Attention）。该方法通过BGE-base-zh-v1.5模型对评论、微博、根评论和父评论四类文本进行语义编码，利用Cross-Attention机制自适应融合上下文信息，并通过双预测头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主题特征和重复程度特征，其中重复程度特征采用MinHash算法高效检测重复和相似评论。实验结果表明，相比基于NGBoost的基线方法，BGE-Attention模型在测试集上的PICP@95%从94.30%提升至97.75%，MPIW从数千万降至3.0050，在保持97.80%高预测准确率的同时，实现了更精确的不确定性校准。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。我们的代码开源于：[https://github.com/LJK666666666/LLM\\_SU7](https://github.com/LJK666666666/LLM_SU7)。

**关键词：**评论热度预测；预训练语言模型；交叉注意力机制；不确定性估计；特征工程

## 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。

2025年3月27日到4月14日期间，小米汽车SU7的智驾事故引发了社交媒体上的广泛讨论，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的重复程度特征；
3. 提出了BGE-Attention模型，通过BGE预训练模型编码多源文本，利用Cross-Attention机制融合上下文信息，双预测头实现概率预测；
4. 采用对数尺度的负对数似然（NLL）损失函数，有效处理长尾分布数据的预测问题；
5. 设计了多维评价指标体系，包括MSLE、ACP、NLL、PICP和MPIW，全面评估预测精度和不确定性校准。

## 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 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。

### 2.3 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baseline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。

## 2.4 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。

## 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2025年3月27日至2025年4月14日，涵盖小米SU7智驾事故前后的热点讨论期。所有评论的子评论数、点赞数等指标均为2025年12月采集到的数值，距离原始发布时间已超过8个月，可以认为是稳定不变的最终值。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。

数据采集流程采用隧道代理绕过反爬虫机制，使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。

### 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**: 删除用户ID、时间、评论文本等所有特征完全一样的重复数据。
- (2) **缺失值处理**: 对于缺失的文本字段（如根评论/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。

### 3.3 数据集划分

本文采用8:1:1的比例将27余万条数据划分为训练集、验证集和测试集。

### 3.4 评论文本重复数据处理

对于评论文本完全相同的情况，将其时间顺序上首次出现的数据强制划入训练集，避免模型从之后数据中学习到的错误热度信息干扰对首次出现数据的预测。

## 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；
- **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；
- **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；
- **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 重复程度特征（3维）

G	H	I	J	K
用户昵称	评论内容	发布时间	子评论数	点赞数
身骑白马	我这辈子第一次听说，出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的	2025/4/1 23:06	1856	29471
悻然江木	我这辈子第一次听说，出了交通事故在第一时间不找警察，不找保险公司，而是找汽车商家的	2025/4/1 23:22	0	3
超级宇宙哥	我这辈子第一次听说，出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的	2025/4/1 23:24	0	1
宝你铭	我这辈子第一次听说，出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的	2025/4/1 23:29	0	1
临江雪_	是这样的//@身骑白马:我这辈子第一次听说，出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的	2025/4/1 23:57	0	0
伤口woun	@雷军 我这辈子第一次听说，出了交通事故在第一时间不找警察，不找保险，而是找汽车商	2025/4/2 6:03	0	2
璇律66	我这辈子第一次听说，出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的	2025/4/2 6:42	54	429
肖顺东	我的荣放有次跟车过近突然提醒我注意危险，我都没有发现有预碰撞提醒功能，丰田的销售1	2025/4/2 9:10	0	0
快_Ke	我这辈子第一次听说，出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的---	2025/4/2 11:45	1	4
感悟瞬间8	我这辈子第一次听说，出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的	2025/4/2 11:57	2	8

图 1: 评论文案重复出现情况

表 1: 首次发布和跟风发布的指标对比

先后顺序	评论数均值	点赞数均值
首次发布	2.24	15.95
跟风发布	0.17	1.13
比值	13.5	14.2

由图1和表1可见，在文案重复出现的情况下，一般时间顺序上首次发布的顺序会得到极大的热度，后续发布则次之。如果不将这种重复性作为标签输入模型中，模型很有可能感到困惑：为什么同样的文本，这条热度极高，这条热度却极低？

为了解决这个问题，我们构建了重复程度特征，对于每条评论，在其发布之前的所有评论中进行相似度检索，捕捉评论的重复出现情况和与历史评论的相似程度：

- **时间顺序索引：**评论在所有数据中的时间顺序索引；
- **最大相似度：**与历史评论的最大文本相似度；
- **重复次数：**在相同或高度相似评论中是第几次出现。

相似度计算使用Jaccard相似度+N-gram方法，并采用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。

包含“我这辈子”的评论：17 条					
时间索引	BGE相似度	BGE重复	MH相似度	MH重复	评论摘要
32962	0.9920	1	0.0859	0	这...可能是我这辈子开过加速度快售价价值最高的车...
58320	0.9118	2	0.1894	0	我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的...
59338	0.9984	4	0.8594	1	我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险公司，而是找汽车商家的...
59448	0.9981	3	1.0000	2	我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的...
59736	1.0000	7	1.0000	3	我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的...
61065	0.9662	4	0.7500	4	是这样的 //身骑白马 我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的...
62879	0.8693	0	0.5512	5	是的，调查结果还没出来，就让车企来负责 [允悲] 我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的，目的地太明显了，一堆傻子还在哩哩...
66067	0.9189	5	0.7734	6	雷军 我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的...人不能太老实...
66115	0.9820	8	0.8984	7	我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的...
66557	1.0000	12	1.0000	9	我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的...
73041	0.7771	0	0.3438	0	我的老板有一次跟车过马路突然提醒我注意危险，我都没有发现有预碰撞提醒功能，丰田的销售也没有给我介绍。//听龙卷风吹：雷军脾气太好了，换大众丰田试试[允悲] //e旋律66：我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的...
82118	0.8198	0	0.5078	5	我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的----这母亲想用舆论影响结果，耐心等待调查结果吧。中国就是会哭的有奶吃，这个博弈要改改
82736	0.9671	4	1.0000	8	我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的...
93435	0.9141	2	0.8125	7	我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的[doge]...
101921	0.9972	8	1.0000	8	我这辈子第一次听说...出了交通事故在第一时间不找警察，不找保险，而是找汽车商家的...
194886	0.9078	4	0.0703	0	我这辈子最佩服的就是雷总了，亏钱做生意都能做成首富[悲催]...
246524	0.7555	0	0.0312	0	友商们别急哈，正常人可以过几个月半年蹲一下我微博，这边不是网红网黑，放心蹲哈。不正常的等我下班一个个投诉哈。我不是雷总粉丝哈，雷总好基友黑过我偶像，我这辈子都讨厌

图 2: 重复程度检索结果示例

从图2可见，重复程度检索的准确性极高。并且经过对比，比左侧文本嵌入向量+余弦相似度匹配的备选方案精确得多。

## 4.2 不确定性估计的重要性

- 转发/评论/点赞本身具有不确定性和随机性：一条评论得到1000个赞和1005个赞都是有概率出现的，应当对其可能的范围进行估计和界定；
- 语义具有不确定性和模糊性：比如文案为“查看图片”的评论，在缺少图片信息的情况下，语义具有极大的模糊性，模型应当尽可能敏锐地捕捉到这种情况。

## 4.3 基线方法：NGBoost概率预测模型

作为基线方法，本文首先采用NGBoost [10]进行概率预测。NGBoost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 $\theta$ 。对于正态分布 $\mathcal{N}(\mu, \sigma^2)$ ，模型同时学习均值 $\mu$ 和标准差 $\sigma$ 。NGBoost的主要优势在于能够提供不确定性估计，但其局限性在于：(1) 无法充分利用文本语义信息；(2) 标准差估计可能出现极端值，导致置信区间过宽。

## 4.4 对数尺度下NLL损失函数

针对社交媒体数据的长尾分布特性，结合图3中各项热度数据在对数尺度下接近正态分布的情况，本文采用对数尺度的负对数似然（NLL）损失函数：

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

其中 $c$ 为平滑常数。不同 $c$ 值对损失函数的影响如表2所示：

表 2: 平滑常数 $c$ 对不同量级预测误差的影响

平滑常数	预测1实际2的损失	预测110实际120的损失	损失比值
$c = 1$	$ \log 2 - \log 3  \approx 0.406$	$ \log 111 - \log 121  \approx 0.086$	4.7
$c = 10$	$ \log 11 - \log 12  \approx 0.087$	$ \log 120 - \log 130  \approx 0.080$	1.1

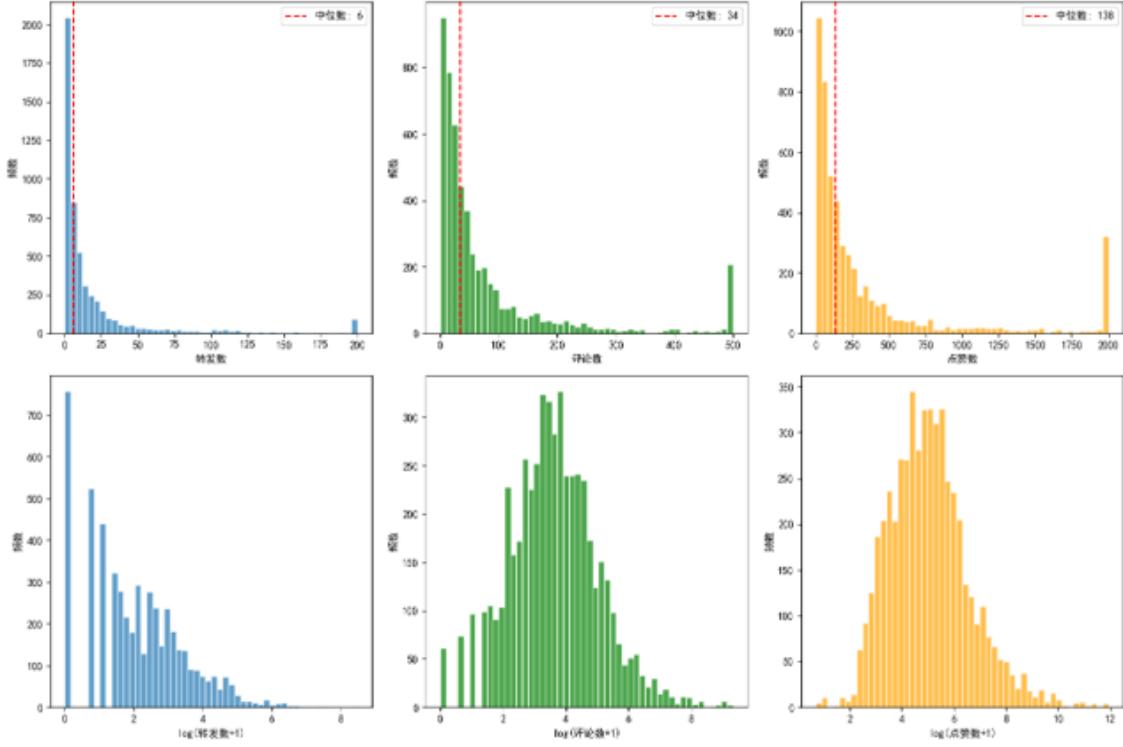


图 3: 转发数、评论数、点赞数分别在原本尺度下和对数尺度下的数据分布图

可见 $c = 1$ 时，小数值的损失是大数值的4.7倍，模型会过度关注小数值样本；而 $c = 10$ 使得不同量级的预测误差惩罚更加均衡，因此本文选择 $c = 10$ 。

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 $\sigma$ 项惩罚盲目自信的预测（ $\sigma$ 小但误差大）；
3. 允许模型通过增大 $\sigma$ 表达不确定性，避免对困难样本的过度惩罚。

#### 4.4.1 模型配置

NGBoost模型的主要超参数配置如表3所示。

表 3: 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

## 4.5 本文方法：BGE-Attention模型

为克服NGBoost无法利用文本语义信息的局限性，本文提出了基于BGE预训练模型的神经网络架构（BGE-Attention），如图4所示。该模型通过预训练语言模型捕捉文本深层语义，利用Cross-Attention机制自适应融合多源上下文信息。



图 4: BGE-Attention模型架构

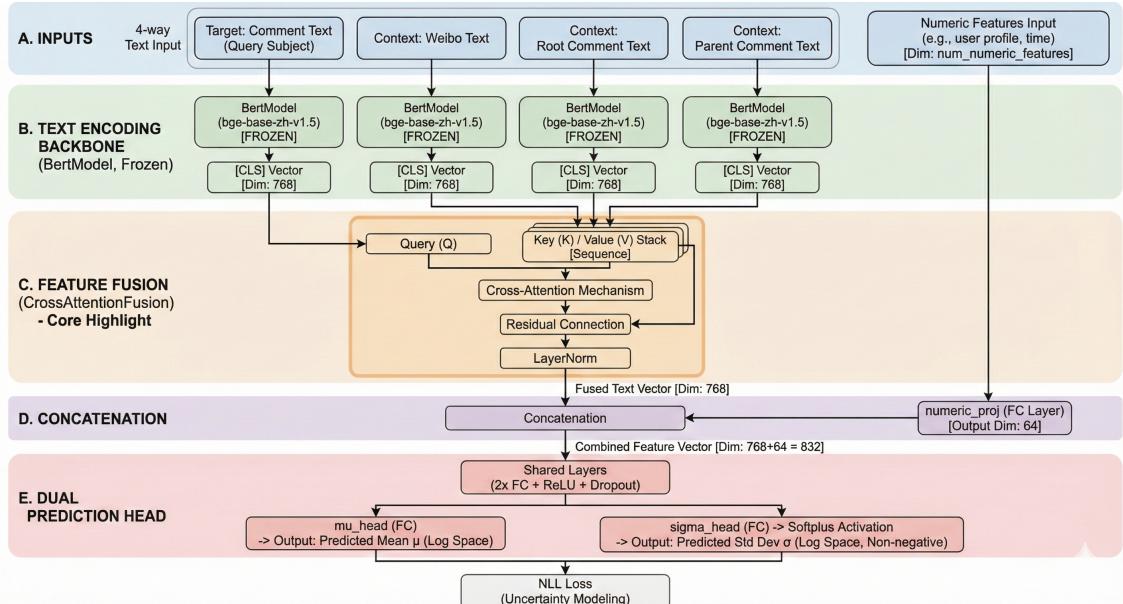


图 5: BGE-Attention模型架构

### 4.5.1 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) **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) **小米相关词汇处理：**比如“理想”在小米SU7相关评论下更有可能代表汽车品牌而非寻常含义，这是预训练的BGE嵌入编码器所难以表征的。再比如“遥遥领先”流行语也具有特殊含义和特殊的自带流量。

(3) **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。

(4) **特殊字符：**保留标点符号和特殊字符，作为情感信号的来源。

### 4.5.2 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。冻结BGE参数防止过拟合。

#### 4.5.3 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:

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

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。

#### 4.5.4 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  $\mu$  和方差  $\sigma^2$ :

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

$$\mu = W_\mu h + b_\mu \quad (4)$$

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

其中  $\epsilon = 10^{-4}$  为数值稳定常数。

### 4.6 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。

#### 4.6.1 预测精度指标

(1) **MSLE** (均方对数误差):

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

MSLE关注相对误差，适合长尾分布数据。

(2) **ACP@( $\alpha, \delta$ )** (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  $\alpha$  为相对容忍度，  $\delta$  为绝对容忍度。本文使用  $\alpha = 20\%$ ,  $\delta = 5$ 。该指标直观反映预测的实用价值。

#### 4.6.2 不确定性校准指标

- (1) 对数尺度下的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) + MPIW (预测区间平均宽度):

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

$$\text{MPIW} = \frac{1}{n} \sum (\text{Upper}_i - \text{Lower}_i) \quad (9)$$

需要注意的是, 对于在对数尺度下进行预测的模型 (如BGE-Attention), PICP和MPIW的计算需要先在对数尺度下构建置信区间 $[\log(\mu+c)-1.96\sigma, \log(\mu+c)+1.96\sigma]$ , 再通过指数变换转换回原始尺度, 即 $[\exp(\text{Lower}_{\log})-c, \exp(\text{Upper}_{\log})-c]$ 。

理想情况下模型应该落在“高覆盖率(PICP) + 低宽度(MPIW)”的黄金区域。

## 5 实验

### 5.1 实验设置

**硬件环境:** 实验在配备NVIDIA A100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 3.10, 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3); 神经网络模型采用Adam优化器, 学习率0.001, 批大小1024, 早停patience为5。

### 5.2 实验结果

#### 5.2.1 基线方法NGBoost性能

表4展示了NGBoost基线模型的实验结果。

表 4: NGBoost基线模型实验结果

数据集	MAE	MSLE	ACP@20%	Log NLL	PICP@95%	MPIW
训练集	0.7879	0.0176	98.15%	-3.6412	97.06%	28685937
验证集	0.9587	0.0284	97.58%	-0.7061	94.30%	11624379
测试集	1.1563	0.0350	97.56%	-0.2952	94.30%	8857315

NGBoost模型的主要问题在于:

- MPIW极大:** 平均预测区间宽度达到数千万, 说明标准差估计不准确, 模型通过极大的 $\sigma$ 来“欺骗”覆盖率指标; 也有可能是模型在某些离群点上效果太差, 导致整体指标偏离严重。
- 泛化性不足:** 验证集和测试集的性能明显低于训练集, 存在过拟合现象。

### 5.2.2 BGE-Attention模型性能

表5展示了本文提出的BGE-Attention模型的实验结果。

表 5: BGE-Attention模型实验结果

数据集	MAE	MSLE	ACP@20%	Log NLL	PICP@95%	MPIW
训练集	0.8717	0.0324	97.91%	-2.9121	97.85%	2.9835
验证集	0.8652	0.0329	97.77%	-2.8979	97.79%	2.9842
测试集	1.0406	0.0347	<b>97.80%</b>	<b>-2.8250</b>	<b>97.75%</b>	<b>3.0050</b>

BGE-Attention模型展现出显著优势：

1. **ACP@20%达到97.80%:** 超过97%的预测值落在真实值20%容忍范围内；
2. **PICP@95%接近理论值:** 97.75%的覆盖率表明模型具有良好的不确定性校准；
3. **MPIW仅为3.0050:** 置信区间窄且准确，相比NGBoost降低了6个数量级；
4. **训练与验证集差距小:** 模型泛化性能优异，无明显过拟合；
5. **训练提升迅速:** 仅训练15个epoch即达到较好性能，受时间限制并未充分训练，可能仍然有提升空间。

### 5.2.3 特征重要性分析

图6展示了NGBoost模型的特征重要性排序。

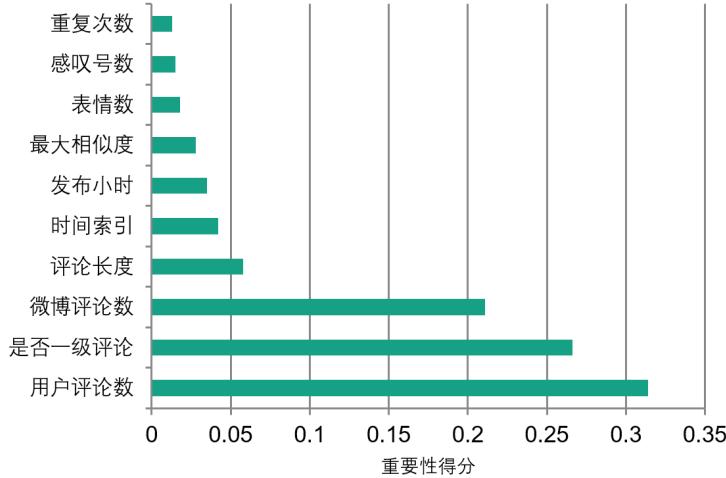


图 6: 特征重要性 (Top 10)

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。
4. 最大相似度 在重复数据仅占2.3%的情况下，该特征的重要性实际上不可忽视。

## 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，提出了基于BGE预训练语言模型和Cross-Attention机制的BGE-Attention模型。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的重复程度特征能够高效检测重复和相似评论；
3. 相比NGBoost基线方法，BGE-Attention模型在测试集上的PICP@95%从94.30%提升至97.75%，MPIW从数千万降至3.0050，在保持97.80%高预测准确率的同时，实现了更精确的不确定性校准；
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。考虑到重复数据所占比例，最大相似度特征的重要性实际上不可忽视。

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).