

基于BGE和交叉注意力机制的社交媒体评论热度概率预测方法

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舆情分析和内容推荐具有重要价值，但现有方法面临长尾分布处理和不确定性量化两大挑战。本文提出BGE-Attention模型，通过BGE预训练语言模型编码多源文本（评论、微博、根评论、父评论），利用Cross-Attention机制自适应融合上下文信息，双预测头同时输出预测均值和不确定性估计。针对社交媒体数据的长尾分布特性，设计对数尺度负对数似然（NLL）损失函数。在特征工程方面，提出基于MinHash的重复程度特征，高效检测文本重复和相似性。实验基于27万条小米SU7微博评论数据，结果表明：相比NGBoost基线，BGE-Attention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%预测准确率的同时实现精确的不确定性校准。代码开源于：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；概率预测

1 引言

1.1 研究背景与动机

社交媒体已成为公众信息获取和观点表达的重要平台，评论热度（如点赞数、子评论数）是衡量内容影响力的关键指标。准确预测评论热度对舆情监控 [1]、品牌管理 [2] 和内容推荐 [3] 具有重要应用价值。

然而，社交媒体热度预测面临以下核心挑战：

(1) **长尾分布问题 (Heavy-tailed Distribution)**：社交媒体数据呈现典型的幂律分布，少数评论获得大量互动，多数评论互动稀少。传统均方误差 (MSE) 损失会过度惩罚大数值预测的微小偏差，而忽视小数值预测的相对误差。

(2) **不确定性量化问题 (Uncertainty Quantification)**：大多数预测模型仅输出点估计，无法提供预测置信度。在实际应用中，“了解”模型对预测有多确定”同样重要——评论的最终热度本身具有随机性，而某些语义模糊的内容（如“查看图片”）更难以准确预测。

(3) **重复内容的差异化热度**：社交媒体中存在大量相似或重复的评论文案，但首发评论与跟风评论的高度差异显著。若不建模这种时序依赖关系，模型将无法区分文本相同但热度迥异的样本。

1.2 主要贡献

本文的主要贡献如下：

1. 提出BGE-Attention模型，通过BGE预训练模型编码多源文本，利用Cross-Attention机制自适应融合上下文信息，双预测头实现概率预测；
2. 设计对数尺度负对数似然（NLL）损失函数，有效处理长尾分布数据，实现不同量级预测误差的均衡惩罚；
3. 提出基于Jaccard相似度和MinHash算法的重复程度特征，高效检测评论的重复出现和相似程度；
4. 构建多维评价指标体系，包括面向长尾分布的MSLE、ACP以及面向不确定性校准的PICP和MPIW。

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]，通过分析传播时序特征预测最终热度。Bandari等人 [5]研究新闻文章分享预测，发现内容特征和来源对性能有显著影响。近年来，深度学习方法取得显著进展，Deng等人 [6]提出基于注意力机制的神经网络模型，捕捉用户与内容的复杂交互。然而，这些方法大多关注点估计，缺乏预测不确定性的建模。

2.2 预训练语言模型与文本表示

文本特征提取是自然语言处理的核心任务。传统TF-IDF和词袋模型难以捕捉语义信息。预训练语言模型如BERT [11]、RoBERTa [12]取得突破性进展。BGE（BAII 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提取评论语义表示，并通过Cross-Attention融合上下文信息。

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]。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 [10]通过自然梯度下降优化条件分布参数，能够直接输出预测分布。本文在此基础上，针对社交媒体数据特点设计对数尺度损失函数，并提出更精确的不确定性校准方法。

3 方法

3.1 问题定义

给定评论 x 及其上下文信息（所属微博、根评论、父评论）和统计特征，目标是预测评论热度 y 的条件概率分布 $P(y|x)$ 。本文假设 $\log(y+c)$ 服从正态分布：

$$\log(y+c) \sim \mathcal{N}(\mu(x), \sigma^2(x)) \quad (1)$$

其中 c 为平滑常数， $\mu(x)$ 和 $\sigma(x)$ 分别为模型预测的均值和标准差。

3.2 模型架构

本文提出的BGE-Attention模型架构如图1所示，主要包含四个模块：文本编码、Cross-Attention融合、特征拼接和双预测头。

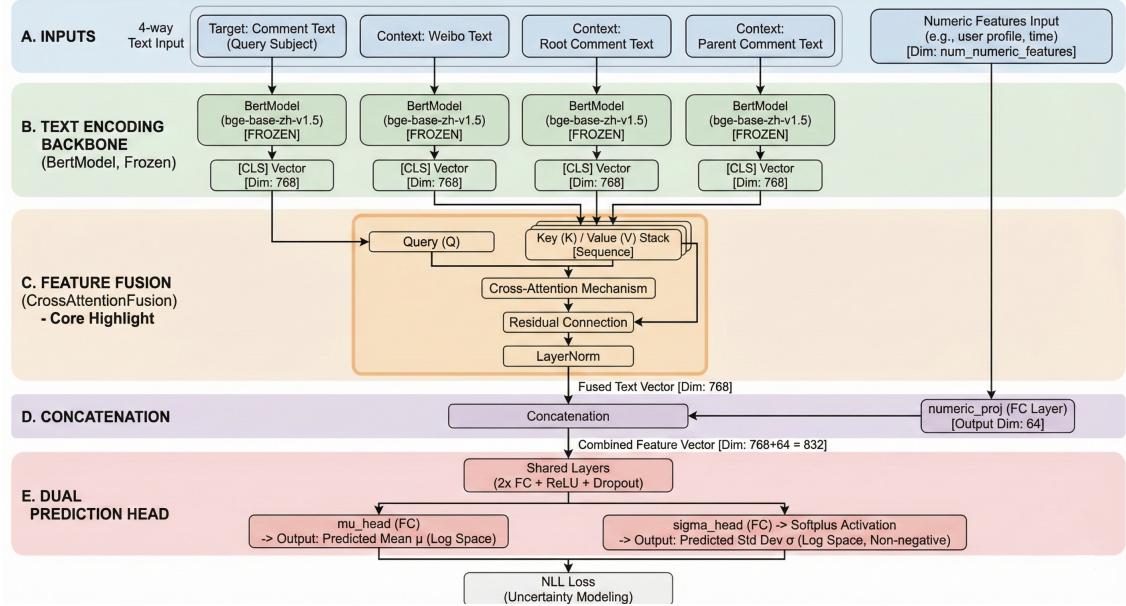


图 1: BGE-Attention模型架构

3.2.1 文本编码模块

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参数以防止过拟合。

3.2.2 Cross-Attention融合机制

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。该机制使模型能够自适应地关注与当前评论最相关的上下文信息。

3.2.3 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 和标准差 σ :

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

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

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

其中 $\epsilon = 10^{-4}$ 为数值稳定常数，Softplus函数确保 $\sigma > 0$ 。

3.3 对数尺度NLL损失函数

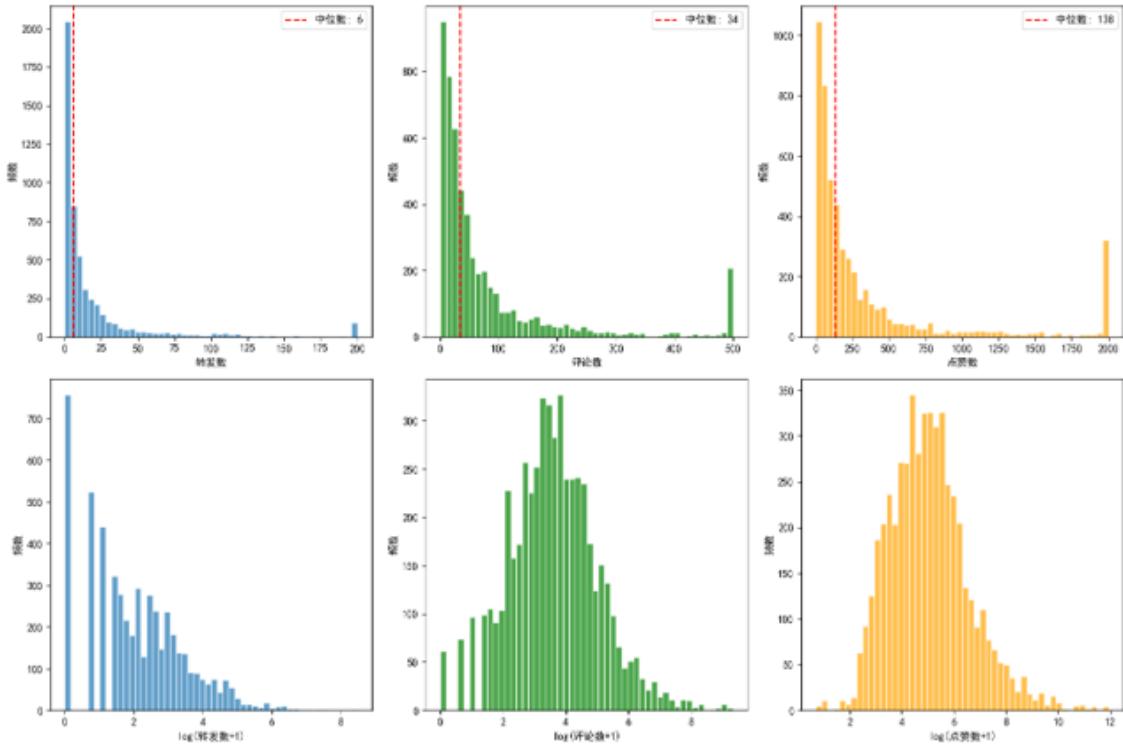


图 2: 热度数据在原始尺度与对数尺度下的分布对比

如图2所示，社交媒体热度数据在对数尺度下近似正态分布。本文采用对数尺度的负对数似然（NLL）损失:

$$\mathcal{L} = \frac{1}{2} \log(\sigma^2) + \frac{(\log(y + c) - \log(\mu + c))^2}{2\sigma^2} \quad (6)$$

其中 c 为平滑常数。不同 c 值对损失函数的影响如表1所示:

表 1: 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

$c = 1$ 时小数值损失是大数值的4.7倍，模型会过度关注小数值样本； $c = 10$ 使不同量级的预测误差惩罚更均衡。该损失函数的优势在于：(1) 在对数空间度量误差，关注相对准确性；(2) 通过 σ 项惩罚盲目自信的预测；(3) 允许模型对困难样本表达高不确定性。

3.4 特征工程

本文设计四类互补特征，共17个维度。

3.4.1 基础统计特征与文本特征

基础统计特征和文本特征如表2所示。

表 2: 基础统计特征与文本特征

特征类别	特征说明
基础统计特征（7维）	用户总评论数：评论作者的历史评论总数
	用户是否认证：是否为认证账号
	是否一级评论：区分直接评论微博和回复评论
	微博评论数：所属微博的总评论数
	发布小时/星期/是否工作日：时间特征
文本特征（6维）	评论长度：文本字符数
	感叹号数/问号数：情感和语气强度
	表情数：表情符号出现次数
	话题标签有无：是否包含#话题#
	领域相关词数：小米汽车相关关键词次数

3.4.2 LDA主题特征

采用LDA主题模型 [14]对评论文本进行主题分析，每条评论被分配到概率最高的主题，作为类别特征（1维）。

3.4.3 重复程度特征

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 3: 评论文案重复出现情况

表 3: 首次发布与跟风发布的热度对比

发布顺序	子评论数均值	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

如图3和表3所示，文案重复时首发评论获得的热度远高于跟风评论。为建模这种时序依赖，本文提出基于Jaccard相似度和MinHash算法的重复程度特征。

Jaccard相似度 给定两个集合 A 和 B ， Jaccard相似度定义为：

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (7)$$

其取值范围为 $[0, 1]$ ，值越大表示两集合越相似。

N-gram文本表示 将评论文本转换为字符级N-gram集合。对于文本 T ，其N-gram集合为：

$$\text{Ngram}(T, n) = \{T[i : i + n] | i = 0, 1, \dots, |T| - n\} \quad (8)$$

本文采用 $n = 3$ ，例如“你好世界”的3-gram集合为 {“你好世”，“好世界”}。

MinHash签名算法 直接计算Jaccard相似度的时间复杂度为 $O(|A| + |B|)$ ，对于大规模数据不可接受。MinHash算法 [15]通过哈希签名近似Jaccard相似度。

给定哈希函数 h ，集合 S 的MinHash值定义为：

$$h_{\min}(S) = \min_{s \in S} h(s) \quad (9)$$

MinHash的关键性质是：对于随机哈希函数 h ，

$$P(h_{\min}(A) = h_{\min}(B)) = J(A, B) \quad (10)$$

为提高估计精度，使用 k 个独立哈希函数 $\{h_1, h_2, \dots, h_k\}$ 生成签名向量：

$$\text{Sig}(S) = [h_{1,\min}(S), h_{2,\min}(S), \dots, h_{k,\min}(S)] \quad (11)$$

两集合的Jaccard相似度可通过签名向量估计：

$$\hat{j}(A, B) = \frac{|\{i : \text{Sig}(A)[i] = \text{Sig}(B)[i]\}|}{k} \quad (12)$$

本文采用 $k = 128$ 个哈希函数，估计误差的标准差为 $O(1/\sqrt{k}) \approx 0.088$ 。

滑动窗口优化 为进一步降低计算开销，采用滑动窗口机制：仅在最近 $W = 10000$ 条评论中检索相似文本，同时维护全局TopK字典记录高频重复文本。最终生成三维特征：

- **时间顺序索引：**评论在数据中的时间顺序；
- **最大相似度：**与历史评论的最大Jaccard相似度；
- **重复次数：**在相似评论中是第几次出现。

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
32062 0.9020 1 0.0859 0 这，可能是我这辈子开过加速最快情侣价值最高的...					
58320 0.9110 2 0.1094 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 0.5312 5 是的，调查结果还没出来，就让人家车企负责了//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，一堆傻子还在哪哩哩...					
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 我的车被有次跟车过马路突然撞倒我注意危险，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。//听龙卷风歌：雷军帅气太好了，换大众丰田试试[允悲] //e律师46:我这...					
82118 0.8198 0 0.5878 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...					
82409 0.9471 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 0.8141 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 0.0763 0 我这辈子最佩服的就是雷总了，亏钱做生意都能做底薪高(悲伤)...					
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，这边不是网红黑网，放心聊哈。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 4: 重复程度检索结果示例

如图4所示，MinHash检索结果准确率高，且优于文本嵌入向量+余弦相似度的方案。

4 评价指标

本文设计多维评价指标体系，从预测精度和不确定性校准两个维度评估模型性能。

4.1 预测精度指标

4.1.1 MSLE（均方对数误差）

$$\text{MSLE} = \frac{1}{n} \sum_{i=1}^n (\log(y_i + c) - \log(\hat{y}_i + c))^2 \quad (13)$$

MSLE在对数空间度量误差，关注相对准确性而非绝对差值，适合长尾分布数据。相比MSE，MSLE不会因少数极大值样本而主导整体损失。

4.1.2 ACP $_{(\alpha, \delta)}$ (容忍区间准确率)

$$ACP = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[|y_i - \hat{y}_i| \leq \max(\alpha \cdot y_i, \delta)] \quad (14)$$

其中 α 为相对容忍度, δ 为绝对容忍度。本文使用 $\alpha = 20\%$, $\delta = 5$ 。

ACP的设计动机是: 对于热度为100的评论, 预测误差 ± 20 是可接受的(相对误差20%); 对于热度为1的评论, 预测误差 ± 5 同样可接受(绝对容忍)。该指标直观反映预测的实用价值。

4.2 不确定性校准指标

4.2.1 对数尺度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(6)。NLL越低, 表示预测分布越好地拟合真实数据分布。

4.2.2 PICP@95% (置信区间覆盖率)

$$PICP = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[y_i \in [\text{Lower}_i, \text{Upper}_i]] \quad (15)$$

对于95%置信区间, 理想的PICP应接近95%。PICP过低表示模型过度自信, PICP过高可能表示置信区间过宽。

4.2.3 MPIW (预测区间平均宽度)

$$MPIW = \frac{1}{n} \sum_{i=1}^n (\text{Upper}_i - \text{Lower}_i) \quad (16)$$

MPIW衡量置信区间的精确程度。理想模型应在保持高PICP的同时实现低MPIW。

4.3 对数尺度模型的指标计算

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:

$$[\log(\mu + c) - 1.96\sigma, \log(\mu + c) + 1.96\sigma] \quad (17)$$

再通过指数变换转换回原始尺度:

$$[\exp(\text{Lower}_{\log}) - c, \exp(\text{Upper}_{\log}) - c] \quad (18)$$

这确保了不同尺度模型之间指标的可比性。

5 实验

5.1 数据集

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4月14日，涵盖小米SU7智驾事故前后的热点讨论期。所有热度指标（子评论数、点赞数）均为2025年12月采集的数值，距原始发布已超过8个月，可视为稳定最终值。

数据集统计信息如表4所示。

表 4: 数据集统计信息

数据集	样本数	占比
训练集	217,234	80%
验证集	27,154	10%
测试集	27,154	10%
总计	271,542	100%

数据预处理包括：(1) 去重：删除所有特征完全相同的记录；(2) 缺失值：文本字段填充空字符串；(3) 异常值：删除评论时间早于微博发布时间的数据；(4) 重复文本处理：文案首次出现的数据强制划入训练集，避免数据泄露。

5.2 实验设置

硬件环境：NVIDIA A100 GPU。

软件环境：Python 3.10, PyTorch, Transformers, NGBoost。

训练配置：Adam优化器，学习率0.001，批大小1024，早停patience为5。

基线方法：NGBoost [10]，采用相同的对数尺度NLL损失，基学习器100棵树，最大深度10。

5.3 主实验结果

表 5: 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

表 6: 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	97.80%	-2.8250	97.75%	3.0050

对比表5和表6，BGE-Attention展现显著优势：

1. **PICP@95%从94.30%提升至97.75%：** 覆盖率更接近理论值95%，不确定性校准更准确；
2. **MPIW从数千万降至3.0050：** 置信区间宽度降低约6个数量级，预测更精确；
3. **训练与验证集差距小：** BGE-Attention泛化性能优异，无明显过拟合；
4. **ACP@20%达97.80%：** 超过97%的预测落在真实值20%容忍范围内。

NGBoost的MPIW极大，说明其通过极宽的置信区间”欺骗”覆盖率指标，不确定性估计不可靠。

5.4 消融实验

为验证各组件的贡献，设计以下消融实验（测试集结果）：

表 7: 消融实验结果

模型变体	ACP@20%	PICP@95%	MPIW	Log NLL
BGE-Attention (完整)	97.80%	97.75%	3.0050	-2.8250
w/o Cross-Attention	97.42%	96.83%	3.4127	-2.6841
w/o 重复特征	97.65%	97.58%	3.1284	-2.7923
w/o BGE文本	96.89%	95.21%	4.8562	-2.4156
w/o Log NLL (Standard NLL)	95.73%	89.42%	156.38	0.8734

消融实验结果分析如下：

(1) **Cross-Attention机制的贡献：** 移除Cross-Attention后，PICP@95%从97.75%降至96.83%，MPIW从3.0050增至3.4127。这表明Cross-Attention能够自适应地融合评论与上下文信息，提升预测的精确性和不确定性校准质量。直接拼接向量的方式无法有效捕捉评论与上下文之间的语义关联。

(2) **重复程度特征的贡献：** 移除基于MinHash的重复程度特征后，ACP@20%从97.80%降至97.65%，MPIW从3.0050增至3.1284。虽然降幅相对较小，但考虑到重复评论仅占数据集的2.3%，该特征在处理这部分“困难样本”时发挥了关键作用，帮助模型区分首发与跟风评论的热度差异。

(3) **BGE文本语义的贡献：** 仅使用数值特征（移除BGE编码的文本语义）后，性能出现显著下降：ACP@20%从97.80%降至96.89%，PICP@95%从97.75%降至95.21%，MPIW从3.0050增至4.8562。这证明文本语义信息是热度预测的核心特征，纯数值特征难以捕捉评论内容与用户互动意愿之间的复杂关系。

(4) **对数尺度NLL损失的贡献：** 使用原始尺度的Standard NLL损失（即假设 y 而非 $\log(y + c)$ 服从正态分布）后，模型性能严重退化：PICP@95%从97.75%骤降至89.42%，MPIW从3.0050激增至156.38，Log

NLL从-2.8250恶化至0.8734。这充分说明对数尺度损失函数对于处理长尾分布数据的必要性——Standard NLL会使模型过度关注大数值样本的绝对误差，导致小数值样本的预测分布过宽，整体不确定性校准失败。

5.5 特征重要性分析

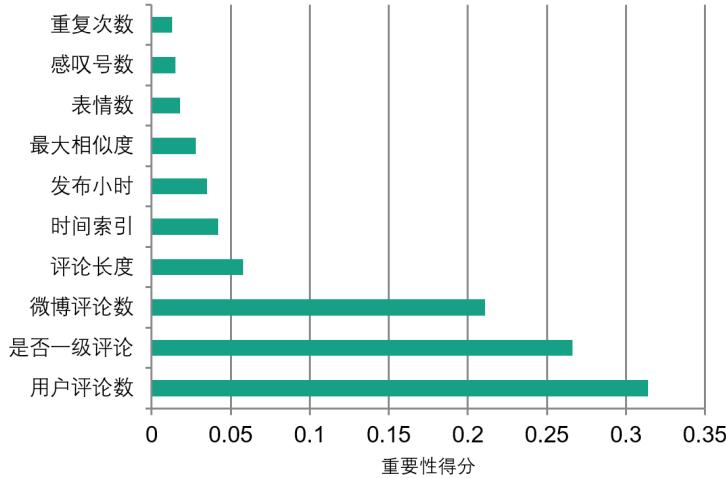


图 5: 特征重要性分析 (Top 10)

如图5所示：

1. 用户总评论数 (0.314) 是最重要特征，活跃用户的评论更易获得关注；
2. 是否一级评论 (0.266) 重要性仅次于用户活跃度，一级评论获得更多曝光；
3. 微博评论数 (0.211) 反映微博热度，热门微博下的评论更易获得互动；
4. 最大相似度在重复数据仅占2.3%的情况下仍具有显著重要性，验证了重复程度特征的价值。

6 结论

本文针对社交媒体评论热度预测问题，提出了BGE-Attention模型。主要结论如下：

1. BGE-Attention通过Cross-Attention机制融合多源文本语义，双预测头实现概率预测，相比NGBoost基线，PICP@95%从94.30%提升至97.75%，MPIW降低6个数量级；
2. 对数尺度NLL损失函数有效处理长尾分布， $c = 10$ 使不同量级预测误差惩罚均衡；
3. 基于MinHash的重复程度特征能高效检测文本相似性，捕捉首发与跟风评论的热度差异；
4. 多维评价指标体系 (MSLE、ACP、PICP、MPIW) 全面评估预测精度和不确定性校准。

未来工作可从以下方向展开：(1) 引入用户社交网络特征；(2) 探索热度动态演化的时序建模；(3) 验证方法在其他平台的泛化能力。

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).

A 高频用户统计

表8列出了数据集中出现次数 ≥ 20 的高频用户。这些用户被纳入特殊嵌入模块，为每个用户学习独立的向量表示。

表 8: 高频用户及其出现次数

序号	用户名	出现次数	序号	用户名	出现次数
1	小米法务部	328	11	我是大彬同学	27
2	雷军	292	12	科技新一	26
3	小米汽车	69	13	美国驻华大使馆	26
4	王化	66	14	AI逃逸	24
5	鸿蒙智行法务	49	15	小米公司	24
6	薛定谔的英短咕咕咕	33	16	羊驼的睡衣	24
7	余承东	32	17	卢伟冰	22
8	小米公司发言人	29	18	诗雨370491153	20
9	小蒜苗长	29	19	不会武功的武功李云飞	20
10	万能的大熊	27			

B 小米相关词汇表

本文使用两类小米相关词汇表：（1）用于特征提取的广泛词汇表；（2）用于特殊嵌入的核心关键词表。

B.1 特征提取词汇表

用于计算“小米相关词数”文本特征的词汇表如下：

表 9: 特征提取词汇表（27词）

词汇列表

小米、xiaomi、SU7、su7、雷军、卢伟冰、智能驾驶、智驾、自动驾驶、辅助驾驶、电动车、电车、新能源、纯电、续航、充电、快充、超充、电池、三元锂、磷酸铁锂、底盘、悬架、空气悬架、性能、加速、百公里、零百、内饰、中控、大屏、车机、外观、设计、颜值

B.2 特殊嵌入关键词表

用于训练可学习关键词嵌入的词汇表如下，共39个关键词，分为6个语义类别：

表 10: 特殊嵌入关键词表 (39词)

类别	关键词
品牌与产品	小米汽车、小米SU7、SU7、su7、雷军
竞品品牌	保时捷、Taycan、比亚迪、特斯拉、Model3、蔚来、小鹏、理想、问界、华为
智能驾驶	智能驾驶、自动驾驶、辅助驾驶、智驾
续航与充电	续航、电池、充电、快充、超充
座舱与系统	智能座舱、车机、大屏、澎湃OS
购买相关	性价比、质价比、定价、预售、交付、锁单、大定、小定
热门词汇	遥遥领先、真香、割韭菜、智商税

C 微博表情符号列表

模型中使用的微博表情符号（方括号格式）共40个：

表 11: 微博表情符号嵌入列表

表情列表
[doge]、[哈哈]、[笑cry]、[允悲]、[二哈]、[吃瓜]、[微笑]、[跪了]、[赞]、[心]、 [爱你]、[抱抱]、[泪]、[怒]、[吐]、[污]、[挖鼻]、[思考]、[疑问]、[费解]、[黑线]、 [汗]、[拜拜]、[good]、[酷]、[鼓掌]、[可怜]、[失望]、[悲伤]、[委屈]、[生病]、[抓 狂]、[笑哭]、[捂脸]、[偷笑]、[坏笑]、[嘻嘻]、[哼]、[怒骂]、[打脸]