基于微博博文情绪分析的话题热度趋势预测研究

Shanghaitech

鞠屹昂
2021533088
Shanghaitech

张天宇
Shanghaitech

李佳熹
2022533184
Shanghaitech

1. ABSTRACT

微博作为中国最大的社交媒体平台之一，其话题广场每日产生数百万条实时博文，反映了公众对热点事件的即时反应。观察发现，话题热度的生命周期往往与用户情绪波动密切相关：例如某明星绯闻事件初期，博文以好奇情绪为主，热度平缓上升；当当事人回应引发愤怒情绪时，话题讨论量激增；而事件澄清后情绪趋于理性，热度逐渐消退。这种关联性表明，情绪分析可为预测话题热度提供重要线索。

本研究用数据挖掘的研究方法获取、分析数据，建立群体情绪-热度关联模型，预测话题未来 1 小时的热度趋势。

2. Introduction

本研究以微博话题为研究对象，通过分析用户博文中的情绪变化规律，探索其对话题热度演化的影响机制。首先利用数据挖掘方法获取微博话题博文，进行数据预处理后利用自然语言方法量化博文的情绪偏向和情感强度，建立模型构建情绪指数指标，结合博文热度数据（浏览量）构建时间序列数据集；其次提取群体情绪波动特征（如情绪极值点、情绪方差变化等）与热度变化特征的相关性；最终建立情绪-热度关联模型，预测话题未来 1 小时的热度趋势。实验部分通过爬取多个社会事件话题的博文数据进行验证，结果显示情绪突变对热度攀升具有显著预警作用。本方法可为社交媒体平台提供低成本的热度预判参考。

最终文章将划分为数据收集、数据预处理、时间序列模型建模、实验等几个主要部分。

3. Related work

目前国内对于微博热度预测的研究主要可以划分为以下几类：

1.基于特征的方法

国内学者在基于特征的微博热度预测方法上进行了大量研究。王晓萌等[1]提出了基于连接强度的微博消息流行度预测模型，通过分析用户之间的连接强度来预测微博的热度。胡颖等[2]对流行度演化分析与预测进行了综述，总结了多种影响微博热度的特征因素。刘定一等[3]提出了一种融合微博热点分析和 LSTM 模型的网络舆情预测方法，结合了文本分析与时序模型的优势。冯新淇等[4]基于 RLDA 主题模型提取特征，用于微博热度预测。这些研究均表明，通过精心设计的特征工程，可以有效提升微博热度预测的准确性。

2.基于时序的方法

基于时序的热度预测方法在国内也有较多研究。X Chen 等[5]对微博热点事件的平稳和非平稳时间序列进行检测，利用动态线性模型（DLM）进行热度预测。C Xiao 等[6]提出了基于时间敏感性的预测模型，通过潜在因素模型（LFM）构建时间序列。Y Xiao 等[7]通过主成分分析法和混沌理论，提出了基于混沌时间序列的小波神经网络预测模型。这些研究充分利用了微博热度随时间变化的规律，能够对未来的热度趋势进行较为准确的预测。

3.基于用户行为的方法

用户行为对微博热度的影响也受到了国内学者的关注。H Wang 等[8]受经典易感流行病模型的启发，提出了基于扩展传染病 SIS 模型的微博传播行为建模方法。Q Cao 等[9]基于耦合图神经网络方法，对级联效应中的节点激活状态进行建模，有效捕捉了基于底层网络的热度预测的级联效应。F Chen 等[10]提出了一个带时变激励函数的标记自激过程模型（MaSEPTiDE），对博文转发动态进行建模。这些研究从用户行为的角度出发，揭示了用户之间的交互影响和传播行为对微博热度的决定性作用。

4.传统机器学习算法

传统机器学习算法在微博热度预测中得到了广泛应用。WH Tan 等[11]提出了利用内、外部知识相结合的经验贝叶斯方法，预测传播进入稳态时的效果。H Zhang 等[12]基于聚类思想和模糊自相关融合分析方法，提出了一种基于快速 K 近邻的热度预测改进算法。这些方法通过特征工程和传统机器学习算法的结合，能够实现较为稳定的预测效果。

5.深度学习算法

深度学习算法在微博热度预测中逐渐成为主流。G Chen 等[13]提出了一种热度预测嵌入模型，包括时间嵌入模型和用户与词的联合嵌入模型，并引入注意力机制。H Yu 等[14]提出了基于多模态的深度学习方法，实现了标签字符串、社会信息和拓扑网络的多模态嵌入。D Liao 等[15]提出了深度融合时间过程和内容特征（DFTC）的热度预测模型，通过循环神经网络和卷积神经网络实现时间动态过程的建模。这些研究充分利用了深度学习模型的强大特征学习能力，显著提升了微博热度预测的性能。

6.集成学习算法

集成学习算法在微博热度预测中也取得了较好的效果。J

Chen 等[16]采用独热编码和 Word2vec 方法对博文元数据进行编码和词嵌入，通过极端梯度提升回归实现社交媒体热度预测。J Tang 等[17]比较了决策树和随机森林算法在微博热度预测中的表现，发现随机森林算法具有更好的预测效果。这些研究通过集成多个弱学习器，有效地提高了预测的准确性和稳定性。

研究局限与未来方向

尽管国内在微博热度预测方面已经取得了丰富的研究成果，但仍存在一些问题。首先，预测模型向着复杂度更高、可解释性更差的方向发展。其次，数据集的公开性不足，导致不同研究之间的结果难以横向比较。

与本文最为类似的文章是任中杰等[18]，采用情感方差作为舆情演变阶段分析的证据，将舆情演变过程划分为四个阶段：高热期，持续期、反复期、消亡期，针对天津 8·12 事故进行研究。本文将主要与此文章进行比较，在共同的 motivation 下有以下创新：

1. 使用大语言模型+提示词工程替代朴素贝叶斯情感分析，提供更加多元的情绪维度，不仅仅有正向、负向两维。
2. 以 russell 情绪环形模型理论为依托，建模情绪煽动值公式，挖掘额外的信息维度，增强预测算法的鲁棒性。

4. Problem statement

本研究的核心问题是如何建模群体情绪与热度变化之间的联系，我们从心理学领域的二维情绪模型获得启发，从情绪唤起（Arousal）和情绪效价（Valence）两个维度建模其对热度的影响：高唤起情绪（如愤怒、兴奋）通常与话题热度的快速上升相关；高唤起情绪（如愤怒、兴奋）通常与话题热度的快速上升相关；情绪效价（正负性）可以进一步细化预测，例如负效价情绪（如愤怒、悲伤）可能引发更强烈的讨论，而正效价情绪（如快乐、满足）可能引发更广泛的传播。

在完成了特征建模后，我们用支持向量回归捕捉特征与预测目标（热度变化趋势）之间的联系。测试时以准确率和 F1 分数作为评估指标。

5. Methods

5.1 Data mining techniques to be used

数据爬取：通过微博 api 获取特定下的博文内容，编写爬虫从微博数据分析平台（知微知见）爬取话题的历史热度

数据预处理：去除用户名、地理位置等标签，提取博文时间戳，对每小时的历史热度线性插值，博文分词，构建每小时的代表性语料库

情感分析：本地部署 DeepSeek-R1-Distill-Qwen-1.5B-Q8_0 蒸馏模型，结合提示词工程，编写调用接口实现对博文的多维度情绪评估。

特征提取：从情感分析结果中提取情绪特征，如情绪均值、极端情绪占比、情绪分布方差等。经 TF-IDF 方法生成每小时的关键词词云

时间序列分析：建模群体情绪波动特征（如情绪极值点、情绪方差变化等）与话题热度变化特征的相关性。

5.2 Data Preprocess

本研究的预处理流程分为两个主要阶段，确保原始社交媒体数据转化为可用于情绪分析和热度预测的高质量数据集。

原始数据清洗：删除回复中的@用户前缀，仅保留有效评论内容；移除所有@用户标记及话题标签；过滤低价值评论（如"感谢分享"等水军内容）；清除 IP 地址列的冗余前缀（如"来自"等）；保留微博 ID、发布时间、点赞数、转发数等核心字段；排除视频号等非文本类内容。

采用基于 Russell 环形情绪模型的方法量化用户情绪状态，用多标签概率分布编码编码情绪分布，为每条数据新增 26 列情绪概率值。对每条评论进行 26 个情绪占比打分，总和为 1。随后通过 Russell 环形情绪模型，通过 26 个情绪得分以及对应情绪的角度，计算评论的 valence 和 arousal 得分。

5.3 情绪影响力分数(EIS)

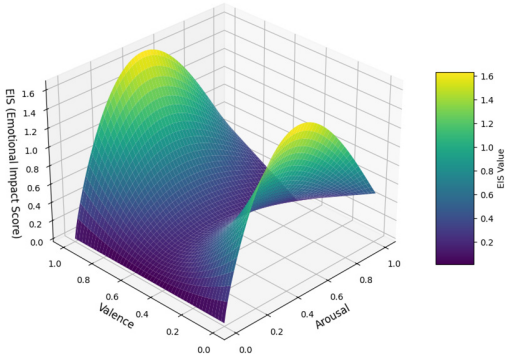
为了更精准地建模情绪刺激对个体的心理影响强度，我们提出情绪影响力分数（Emotional Impact Score, EIS）。该分数基于心理学研究发现，情绪的影响不仅取决于其强度（唤起度，arousal），还与其性质（效价，valence）紧密相关。EIS 整合了这两个维度的非线性效应和交互作用，以反映真实情绪体验的复杂动态。

$$\text{arousal\_effect} = 1.2 \cdot \sin\left(\pi \cdot \text{arousal}^{0.8}\right)$$

$$\text{interaction} = 0.5 \cdot \text{arousal} \cdot |2 \cdot \text{valence} - 1|$$

$$\text{valence\_effect} = 4 \cdot (\text{valence} - 0.5)^2 + 0.2$$

$$\text{EIS} = \text{arousal\_effect} \cdot \text{valence\_effect} + \text{interaction}$$



5.4 Feature Engineering

本热度预测模型旨在通过一系列精心挑选的特征来预测社交媒体（如微博）上的话题热度变化。以下是对所使用的主要特征及其计算方式的详细描述。

首先将每个时段的博文放入同一组，在组内计算：

1. valence_std：将每小时话题博文效价（valence）的标准差，反映群体情绪倾向分布

2. arousal_mean：每小时微博唤醒度（arousal）的平均值，反映群体情绪强度

3. EIS_mean： 综合反映每小时情绪传播的潜在影响力，高值预示更强的传播扩散可能

4. EIS_std： 监测情绪影响力的波动性，异常波动可能触发热度突变

5. arousal_mean_ma3：采用窗口宽度为 3 小时的移动平均对基础序列进行平滑，捕捉用户情绪唤醒度（arousal）的短期趋势变化。3 小时窗口经过实证测试，能平衡短期波动捕获与噪声过滤的需求。通过平滑随机波动，突出唤醒度的持续性变化。

6. Experiments

6.1 Dataset

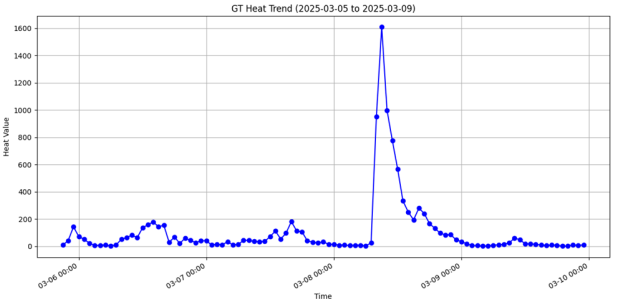
数据集来源：新浪微博话题《何凯文考研英语造假》。制作方式：使用 weibo-search 爬取 2025. 3. 5–2025. 3. 10 话题内相关博文。发现微博数量少于热度数据，分析发现热度数据考虑了评论数量，因此再次利用 Weibo Spider 爬取博文的评论，视作与博文等价的舆论信息。

得到数据集后，去除了总共博文数量小于 10 的时段，并且去除了 1-6 点时段的数据。原因是睡眠时段会影响话题的热度，而时间有限，我们只制作了一个数据集，实验证明数据量不足以让模型充分学习时段特征，因此需要人为排除这一因素。

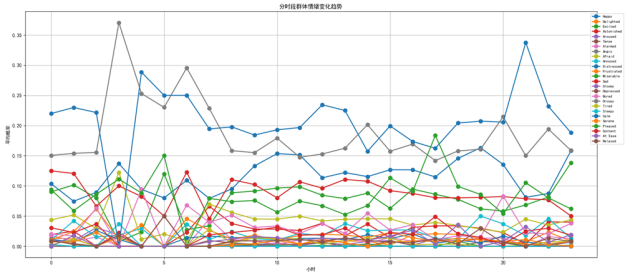
数据集的训练-测试占比为 6:4, 使用 sklearn 划分。

6.3 Visualization & Insights

话题的真实热度变化：

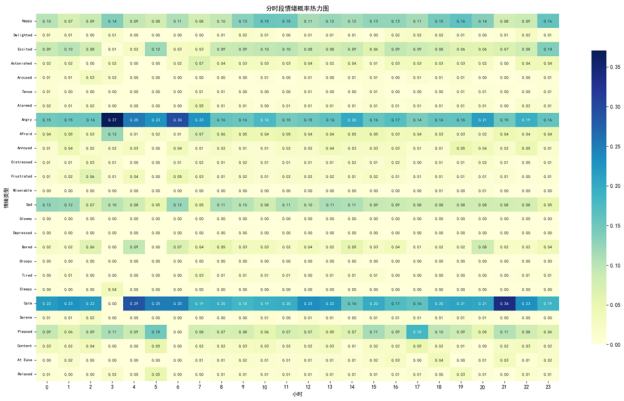


各时段的群体情绪比重变化：

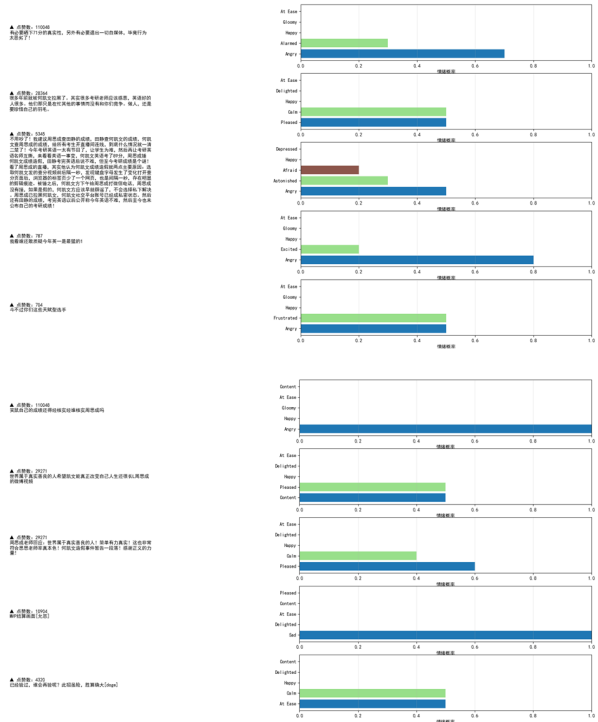


本话题下群体的主要情绪可以总结为：幸灾乐祸、愤怒、悲观。其中愤怒、悲观的情绪高峰与热度的高峰在时间上部分重合，预示了群体情绪与热度的相关性。而无聊、平静、疲惫、放松等情绪强度较低的情绪代表了理性、客观的用户态度，在如此一个热度较高、全民关注的话题中，并没有占较高的比重，进一步说明话题的热度与群众的情绪化十分相关。

归一化后的分时段情绪热力图显示了如愤怒一类情绪强度较高的情绪在话题的初期、高峰期具有较高占比，而如平静一类较平稳的情绪在话题的末期占比更高。



部分高赞评论情绪占比得分可视化：



结果显示本文采用的基于大模型的情绪分析方法对于情感的分解较为准确，为后续的预测步骤提供了较可靠的元数据。

6.2 Model

模型架构 使用 XGBoost 分类器（binary:logistic）作为基础模型，关键参数设置为：决策树数量(n_estimators)=200，最大深度(max_depth)=6，最大深度(max_depth)=6

6.3 Performance & Abalation Study

```
=== 完整特征集 ===
准确率: 0.7778
F1分数: 0.7000

=== 去除EIS特征 ===
准确率: 0.5556
F1分数: 0.5000

=== 去除arousal_mean_ma3特征 ===
准确率: 0.6667
F1分数: 0.6087

完整模型 - Accuracy: 0.7778, F1: 0.7000
              precision    recall  f1-score   support

         0.0       0.88      0.78      0.82        18
         1.0       0.64      0.78      0.70         9

    accuracy                           0.78        27
   macro avg       0.76      0.78      0.76        27
weighted avg       0.80      0.78      0.78        27

模型训练完成，结果已保存
```

7. Conclusions

Evaluation 和 Abalation Study 证明，特征工程获取的特征对于预测的准确率有提升作用。本研究从现象出发，探究用户情绪波动与微博话题热度之间的联系，但实际话题热度变化是一个复杂的过程，因此本研究提出的方法也具有一定局限性。

8. REFERENCES

1. 王晓萌，方滨兴，张宏莉，王星. TSL：基于连接强度的 Facebook 消息流行度预测模型[J]. 通信学报，2019，40(10):1-9.
2. 胡颖，胡长军，傅树深，黄建一. 流行度演化分析与预测综述[J]. 电子与信息学报，2017，39(04):805-816.
3. 刘定一，沈阳阳，詹天明，刘亚军，应毅. 融合微博热点分析和 LSTM 模型的网络舆情预测方法[J]. 江苏大学学报(自然科学版)，2021，42(05):546-553.
4. 冯新淇，张琨，任奕豪，谢彬，赵静. 一种基于 RLDA 主题模型的特征提取方法[J]. 计算机与数字工程，2017，45(10):1980-1985.
5. X Chen, X Lan, J Wan, P Lu, M Yang, L Pancioni. Evolutionary Prediction of Nonstationary Event Popularity Dynamics of Weibo Social Network Using Time-Series Characteristics[J]. Discrete Dynamics in Nature and Society, 2021, 2021(1): 1-19.
6. C Xiao, C Liu, Y Ma, et al. Time sensitivity-based popularity prediction for online promotion on Twitter[J]. Information Sciences, 2020, 525(16): 82-92.
7. Y Xiao, X Xie, Q Li, et al. Nonlinear dynamics model for social popularity prediction based on multivariate chaotic time series[J]. Physica A: Statistical Mechanics and its Applications, 2019, 525(13): 1259-1275.
8. H Wang, Y Li, Z Feng, et al. ReTweeting analysis and prediction in microblogs: An epidemic inspired approach[J]. China Communications, 2013, 10(3): 13-24.
9. Q Cao, H Shen, J Gao, et al. Popularity prediction on social platforms with coupled graph neural networks[C]. Proceedings of the 13th International Conference on Web Search and Data Mining, 2020: 70-78.
10. F Chen, W H Tan. Marked self-exciting point process modelling of information diffusion on Twitter[J]. The Annals of Applied Statistics, 2018, 12(4): 2175-2196.
11. W H Tan, B Chen. Predicting the popularity of tweets using internal and external knowledge: an empirical Bayes type approach[J]. AStA Advances in Statistical Analysis, 2021, 105(2): 335-352.
12. H Zhang. Research on information popularity prediction of multimedia network based on fast K proximity algorithm[J]. International Journal of Autonomous and Adaptive Communications Systems, 2020, 13(2): 103-115.
13. Chen G, Kong Q, Xu N, et al. NPP: A neural popularityprediction model for social media content[J]. Neurocomputing, 2019, 333(11): 221-230.
14. Yu H, Hu Y, Shi P. A prediction method of peak timepopularity based on twitter hashtags[J]. IEEE Access, 2020, 8(1): 61453-61461.
15. Liao D, Xu J, Li G, et al. Popularity prediction on onlinearticles with deep fusion of temporal process and content features[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 200-207.
16. Chen J, Liang D, Zhu Z, et al. Social media popularity prediction based on visual-textual features with XGBoost[C]. Proceedings of the 27th ACM International Conference on Multimedia, 2019: 2692-2696.
17. Tang J, Xu X, Qiu J. Model Construction and Evaluation of Microblog News Popularity Prediction[C]. 2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC), IEEE, 2021:457-464
18. 基于微博数据挖掘的突发事件情感态势演化分析，任中杰 张 鹏 李思成 兰月新 夏一雪 崔彦琛
19. http://www.nlpir.org/wordpress/2018/01/26/500%e4%b8%87%e5%be%ae%e5%8d%9a%e8%af%ad%e6%96%99/
20. http://www.nlpir.org/wordpress/2017/12/03/nlpir%e5%be%ae%e5%8d%9a%e5%86%85%e5%ae%b9%e8%af%ad%e6%96%99%e5%ba%93-23%e4%b8%87%e6%9d%a1/
21. 微博话题爬虫 https://github.com/dataabc/weibo-search
22. 微博评论爬虫 https://github.com/nghuyong/WeiboSpider