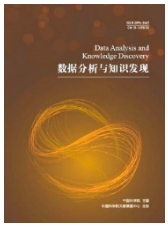


数据分析与知识发现
Data Analysis and Knowledge Discovery
ISSN 2096-3467,CN 10-1478/G2

《数据分析与知识发现》网络首发论文

题目：融合 RF-GA-XGBoost 和 SHAP 的虚假新闻群体互动质量可解释模型
作者：温廷新，白云鹤
网络首发日期：2024-01-17
引用格式：温廷新，白云鹤．融合 RF-GA-XGBoost 和 SHAP 的虚假新闻群体互动质量可解释模型[J/OL]．数据分析与知识发现.
https://link.cnki.net/urlid/10.1478.G2.20240117.1108.018

网络首发时间：2024-01-17 17:56:09
网络首发地址：https://link.cnki.net/urlid/10.1478.G2.20240117.1108.018

融合 RF-GA-XGBoost 和 SHAP 的虚假新闻群体互动质量可解释模型

温廷新，白云鹤
(辽宁工程技术大学工商管理学院 葫芦岛 125105)

摘要：
[目的]良性群体互动在虚假新闻传播过程中具有正面引导作用。为充分发挥社交媒体用户群体互动质量对虚假新闻负面影响的抑制作用，准确判定良性互动的成因及其作用方式，提出一种融合 RF-GA-XGBoost 和 SHAP 的虚假新闻群体互动质量可解释模型。
[方法]以数据集 Weibo21 中的 500 篇虚假新闻及 7029 条评论为研究对象。首先，从评论的内容、形式、情感 3 个维度综合衡量虚假新闻群体互动质量。其次，从这 3 个维度依次提取虚假新闻文本特征。接着，采用随机森林的序列前向搜索策略提取虚假新闻文本的最优特征子集，构建基于 GA-XGBoost 的群体互动质量预测模型，并与 LR、SVM 和 XGBoost 等主流机器学习算法进行实验对比。最后，采用 SHAP 模型对重要特征为群体互动质量带来的影响进行因果解释。
[结果]实验结果表明，GA-XGBoost 模型的 F1-score 和 AUC 值均达到 86%以上，选取的 6 项性能指标均优于其对比模型。此外，虚假新闻文本的内容字符数、词语数量、负面情感词数量等特征是影响虚假新闻社交媒体群体互动质量的重要因素。
[局限]本文未进行多特征交互解释分析，同时也未根据时间戳深入挖掘早期高质量群体互动规律。
[结论]综上，该可解释预测模型能够准确获得各特征对群体互动质量的影响方式，有利于为社交媒体平台在运营策略和功能设计改进方面提供有效决策支持。

关键词：虚假新闻；群体互动质量；GA；XGBoost；SHAP
分类号：TP391，G206
DOI: 10.11925/infotech.2096-3467.2023.0881

An Interpretable Model of Fake News' Group Interaction Quality Based on RF-GA-XGBoost and SHAP

Wen Tingxin, Bai Yunhe
(School of Business Administration, Liaoning Technical University, Huludao 125105, China)

Abstract:
[Objective]Benign group interaction has a positive guiding role in the process of spreading of fake news. To give full play to the inhibitory effect of social media user's group interaction quality on the negative impact of fake news, and accurately determine the causes and ways of benign interaction, an interpretable model of fake news' group interaction quality based on RF-GA-XGBoost and SHAP is proposed.
[Methods]Taking 500 fake news and 7029 comments from the dataset Weibo21 as the research object. Firstly, the fake news' group interaction quality is comprehensively measured from three

dimensions:content, form and emotion of comments. Secondly, the fake news text features are extracted from these three dimensions. Then, the sequential forward search strategy of random forest is used to extract the optimal feature subset of fake news text, and a group interaction quality prediction model based on GA-XGBoost is constructed, and conduct experimental comparisons with other mainstream machine learning algorithms such as LR, SVM and XGBoost. Finally, the SHAP model is used to provide causal explanations for the impact of important features on the group interaction quality.

[Results]The experimental results show that the F1-score and AUC values of the GA-XGBoost model are both above 86%, and the selected six performance indicators are all superior to their comparative models. In addition, the characteristics of false news texts, such as the number of content characters, the number of words, the number of negative emotional words are important factors that affect the fake news' group interaction quality among social media.

[Limitations]This paper does not conduct multi feature interactive interpretation analysis, nor does it dig into the early high-quality group interaction rules according to the timestamp.

[Conclusions]In summary, this interpretable predictive model can accurately obtain the impact of each feature on the group interaction quality, which is conducive to providing effective decision-making support for improving the operational strategy and functional design of social media platforms.

Keywords： Fake news; Group interaction quality; GA; XGBoost; SHAP

1 引言

截至 2022 年 12 月，我国网民规模达 10.67 亿，网络新闻用户规模达 7.83 亿，占网民整体的 73.4%[1]。社交媒体平台大规模兴起，为广大用户提供了发表新闻资讯和撰写文章评论的机会，但同时也为虚假新闻的发布和传播创造了崭新的渠道。如果不对虚假新闻传播加以遏制，不仅会损害媒体和个人的公信力，还会对社会稳定和国家安全产生严重危害，造成一系列重大的负面影响[2]。在社交媒体中，用户互动是网络虚假新闻传播的重要特点，新闻评论是其最直观的体现，新闻作者发布虚假新闻，网民在虚假新闻下发表评论表达和分享对某个虚假新闻事件的态度和看法。新闻作者创造内容和网民发布内容构成了社交媒体中虚假新闻的全部内容。因此，对虚假新闻内容进行挖掘，不仅能控制错误信息和消极情绪继续传播，也有助于社交媒体平台及时把握正确的舆论导向，对最大限度的减少其产生的负面影响至关重要。

虚假新闻的内容挖掘是当前研究的热点问题之一，学者们对此开展了广泛研究，根据研究对象不同，可以分为基于新闻作者创造内容的挖掘和基于公众发布内容的挖掘。基于新闻作者创造内容的挖掘主要是从新闻文本、图像信息等入手挖掘特征检测新闻真实性或抑制其传播路径，为虚假新闻内容真伪识别或传播抑制方式提出优化建议。张国标等[3]以 Twitter 上的 MediaEval2015 数据集为研究对象，提取社交网络中虚假新闻的文本和图像特征信息，进行虚假新闻检测；曾子明等[4]以新浪微博上 2016 年雾霾谣言为研究对象，深入挖掘文本的主题分布特征，将其与用户可信度和微博影响力特征变量作为模型输入进行谣言识别；翟玥等[5]认为大规模的公众参与能使不实信息在传播过程中得到有效抑制；Zhao 等[6]发现虚假信息蔓延的预期危害性会促使公众采取积极应对措施，并可能动员亲朋好友参与其中。基于公众发布内容的挖掘主要是从评论入手应用到虚假新闻检

测、公众情感倾向分析、互动效果引导等研究中。Shu 等[7]捕捉新闻内容和用户评论间的关系，以筛选出有解释性的新闻句子和评论用于假新闻检测；陈娟等[8]以微博上的政府辟谣信息的用户评论为研究对象，构建多元 Logistic 回归模型分析内容、文本和用户特征变量对用户情感倾向的影响；王诣铭[9]提出利用演化博弈模型提高虚假新闻传播事件中社交媒体平台群体互动的质量水平，抑制虚假新闻在社交媒体平台的传播，从而快速降低虚假新闻的负面效应；阮宏飞等[10]运用 Logit 模型实证检验信息互动对传闻形成的影响，结果表明公司与投资者之间的信息互动质量显著降低了公司传闻形成的可能性。

虽然学者们对虚假新闻内容挖掘研究的深入探索取得了丰富的研究成果，但现阶段研究多是从虚假新闻本身或评论某一方面入手，忽略了公众间的良性群体互动可以干扰虚假新闻传播和减小其负面影响这一研究视角。同时，已有研究主要通过检测的方式减小虚假新闻事后影响，却较少有学者对事前治理进行研究，如为抑制虚假新闻负面影响大范围扩散研究良性互动的成因。基于此，本文将虚假新闻及其评论内容有机结合，利用优化智能算法和机器学习算法构建虚假新闻社交媒体群体互动质量模型，再采用 SHAP 解释框架建立不同互动质量结果与各特征变量之间的因果关系，进而为社交媒体平台有效控制虚假新闻负面影响提供决策依据。

2 群体互动质量

2.1 群体互动

虚假新闻传播实质上是互动导向的，伴随着公共观点和群体意见的产生[11]，它由虚假新闻事件的发布者和传播者共同完成，即由新闻作者和公众用户共同实现。杨善林和张大勇等[12-13]认为群体互动是网络群体内部用户间的互动，是信息能够快速、有效传播的关键。此后许多学者也就群体互动在社交媒体领域给出了定义，Berg 等[14]认为群体互动是使虚假新闻事件发展进程产生推动作用的对新闻信息的讨论、补充、质疑、求证和反驳，王诣铭认为社交媒体群体互动主要包括信息互动、情绪互动以及社会资本互动三种[9]。本文所研究的群体互动主要指信息互动和情绪互动，而虚假新闻群体互动包括新闻作者与公众、公众与公众之间的互动，公众集体意见和与新闻作者持有观点间的互动性是构成群体互动质量的基础，如图 1 所示：

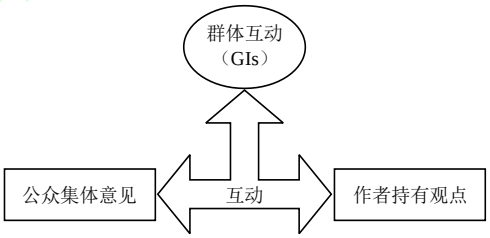


图 1　公众集体意见、作者持有观点和群体互动间关系

Fig1. The Relationship between Public Collective Opinion, Author's Viewpoint, and Group Interaction

2.2 群体互动质量

虚假新闻传播事件中公众间集体意见及其与新闻作者持有观点间可以形成高质互动、低质互动和无关互动，由于无关互动等同于低质互动，最后可以形成

2 种群体互动质量效果：高质互动和低质互动。例如，在埃塞俄比亚客机失事事件中，含有较多常识性、积极性和线索证据性评论的新闻群体互动属于高质互动，含有过度解读或加工的非正面回复和无关推广信息的新闻群体互动属于低质互动。达成群体高质互动的必要条件有两个：（1）公众集体意见达成内部高质互动是首要条件。（2）公众集体意见达成高质互动并与新闻作者持有观点发生良性互动是最终条件。那么达成群体高质互动的路径有且只有一种，其余路径皆为群体低质互动，两种群体互动质量达成路径如图 2 所示。

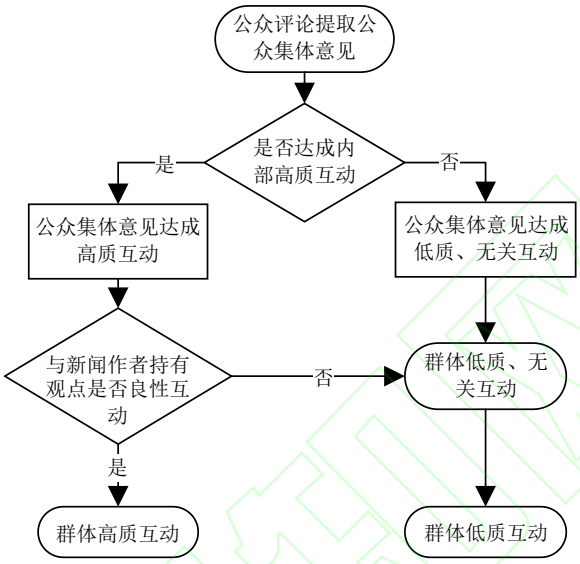


图 2　群体互动质量达成路径

Fig2. Path to Achieving Group Interaction Quality

2.3　群体互动质量判定

Sussman 和 Sanford 等[15-16]提出可以通过互动信息的相关度、时效性、准确性和完整性来判断群体互动质量。李贺等[17]认为群体互动质量不仅表现为公众的互动内容相关度、真实度和及时性，也表现为个体的情感倾向。在自然语言处理领域，学者们常利用问题与回答之间的文本相关度来对回答的质量进行度量。类似地，本文可以通过评论与该虚假新闻事件的相关度度量群体互动质量[18-19]。同时，微博带有集体情绪趋势，评论的情感倾向作为公众对新闻内容最直观的情绪反馈[20-21]，也会对群体互动质量产生影响。另外，阮宏飞[10]等选定回复详细程度作为衡量群体互动质量的指标之一。因此，本文从内容、形式和情感 3 个维度分别选取评论的内容相关度、回复详细程度和情感倾向 3 个指标综合衡量群体互动质量高低。

通常情况下，群体高质互动需要通过上文中达成高质互动的必要条件判断：公众集体意见先达成高质互动，再与新闻作者持有观点发生良性互动。在本文中，新闻文本与评论的内容相关度，可以代表公众与新闻作者互动的内容相关度；新闻的回复详细程度，可以反映出公众对新闻作者的回复详细程度。其中内容相关度和回复详细程度都分为高和低 2 种，内容相关度越高、回复详细度越高，群体互动质量越高[22]。针对新闻评论的情感倾向，包含了众多评论者的主观意见[23]，可以反映出公众对新闻作者或新闻作者所陈述事实的情感倾向，本文仅考虑公众对新闻作者所陈述事实的情感倾向，关于公众对新闻作者所陈述事实的积极评论越多则群体互动质量越高。综上，本文认为某篇虚假新闻评论的内容相关度高、

回答详细程度高和情感倾向积极 3 个条件符合两个及以上时，判定群体互动为高质量，否则为低质量互动。

3 模型设计与实现

3.1 模型整体框架

本文构建基于 RF-GA-XGBoost 的虚假新闻群体互动质量模型，通过 SHAP 可解释框架对微博上不同互动质量类别的各个特征变量重要性进行量化和归因，具体的模型总体框架如图 3 所示。

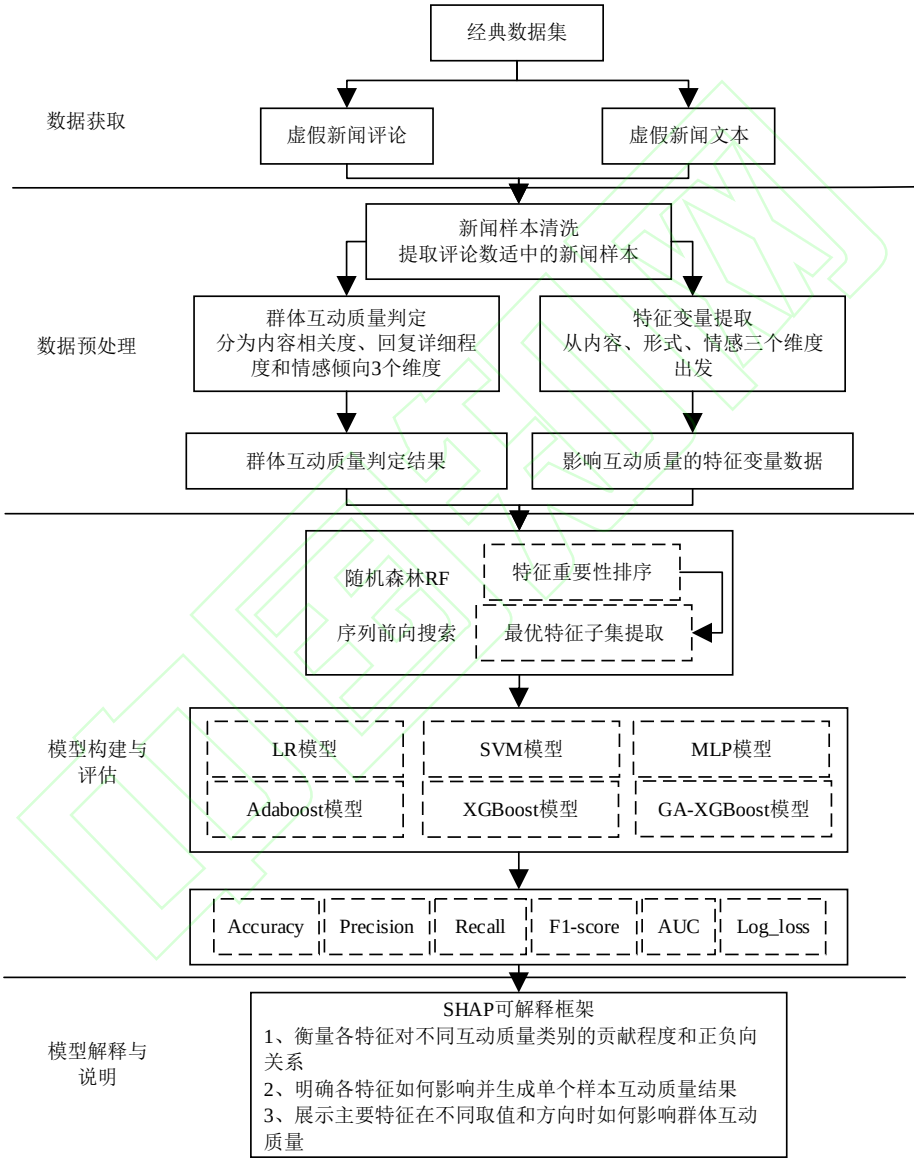


图 3　模型总体框架图

Fig3. Overall Framework Diagram of the Model

3.2 虚假新闻群体互动质量模型设计

根据模型总体框架图，本文的虚假新闻群体互动质量模型的分析与设计过程如下：

（1）数据获取：微博是中文领域中网络新闻发布和传播的主流站点，本文选择微博平台公开的虚假新闻经典数据集为数据源，以一定时间范围内的虚假新闻文本及其评论数据作为研究对象。

（2）数据预处理：分别对虚假新闻评论和文本进行数据预处理。对评论数据进行内容相关度、情感倾向和回答详细程度计算，得到该篇虚假新闻文本的群体互动质量类别。对文本数据进行特征变量提取，新闻文本特征一般分为内容和形式特征变量两类，而有关学者认为社交新媒体中发布和转发的新闻、评论包含着大量的情绪[24-25]，因此有必要从文本中提取情感特征变量。综上，本文从内容、形式和情感特征 3 个维度选取虚假新闻文本特征变量。

（3）模型构建：以上文确定的特征变量为特征变量，以群体互动质量为预测变量，构建基于 RF-GA-XGBoost 的虚假新闻群体互动质量模型。首先，根据各分类样本数量判断是否需要进行样本均衡。其次，随机森林算法是评估特征重要性的常见方法，序列前向搜索能够提高最优特征子集的搜索效率。因此，接下来采用随机森林算法（RF）计算特征重要性并排序，再根据序列前向搜索策略（SFS）提取最优特征子集。之后，为进一步提高分类模型预测的表现能力，需要对模型参数进行优化，遗传算法可以自动化地缩小和确定最优参数组搜索的方向与规模，能高效确定全局最优解。同时为避免遗传算法陷入局部最优，本文先采用随机搜索法缩小参数范围，再利用遗传算法（GA）优化最优单模型 XGBoost 的超参数。最后，将其与虚假新闻研究中常用来解决较大样本二分类问题、且表现效果较好的 5 个经典机器学习模型作实验对比：一方面选择 3 类简单模型与树模型做外比，比如逻辑回归（LR）、支持向量机（SVM）和多层感知机（MLP）；另一方面再选择两个不同的梯度提升树模型来体现树模型选择的层次性，即自适应增强决策树（Adaboost）和极端梯度提升决策树（XGBoost）模型[26-29]。

（4）模型评估：按 8:2 的比例划分数据集后，分别对 6 个分类模型进行训练和分类预测。以准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1-score、AUC 和对数损失（Log_loss）作为评估模型优劣的指标，找出最优模型。

3.3 虚假新闻群体互动质量可解释模型设计

可解释性是决定预测模型能否成功应用的主要因素之一，SHAP 和 LIME 都是经典的事后解释方法，而 SHAP 可以从全局和局部两个层面使众多“黑盒模型”获得合理的解释[30]。易明[26]和刘天畅[27]等融合 XGBoost 与 SHAP 模型分别对公共价值共识和养老平台用户流失进行预测和特征分析，聂卉[28]等结合 LightGBM 和 SHAP 模型进行抑郁症预测和影响因素研究。假设模型基准分（所有样本的目标变量的均值）为 y_{base}，第 i 个样本为 x_i，第 i 个样本的第 j 个特征为 x_{ij}，特征的边际贡献为 ms，边的权重为 w_k，模型对该样本的预测值为 y_i，则第 i 个样本的第 1 个特征的 SHAP 值 $f(x_{i1})$ 如公式（1）所示，同时 SHAP 值要服从公式（2）。

$$f(x_{ij}) = \sum_{k=1}^{n} ms_{i1} w_k \tag{1}$$

$$y_i = y_{base} + \sum_{s=1}^{n} f(x_{is}) \tag{2}$$

虚假新闻可解释性研究已成为未来的热点问题[31]，引入 SHAP 解释框架对最优模型进行解释说明，分析影响群体低质和高质互动的重要因素，并结合

SHAP 三类可视化图从样本和特征两个角度来阐述各特征对群体互动质量的作用方式：①采用 SHAP 特征概要图反映特征对不同互动质量类别的贡献程度和正负向关系；②通过 SHAP 瀑布图和力图反映各特征如何影响单个样本互动质量结果；③通过 SHAP 部分依赖图揭示主要特征对群体互动质量的影响程度和方向。

4 实证分析

4.1 数据获取

微博是中国第一大社交新媒体互动平台，本文选取微博的某个基准数据集，即南琼和曹娟构建的首个中文多领域虚假新闻数据集——Weibo21（https://github.com/kennqiang/MDFEND-Weibo21.）作为数据源。该数据集从微博社区管理中心爬取 2014 年 12 月至 2021 年 3 月的虚假新闻，包括文本内容、配图、时间戳、评论、辟谣信息和所属领域多个维度信息，每条新闻的领域由 10 位专家独立地进行人工注释，如有 8 个以上专家选择同一领域，则可以确定该新闻的最终所属领域，否则专家继续讨论直到达成一致意见。另外，其含有科学、教育和灾害等 9 个领域的假新闻，具有丰富的域，可以更好的进行多域假新闻研究[32]，而本文将所属领域作为研究特征，因此该数据集符合本文的多域研究这一特点。评论数量可以体现互动强度，而其存在两端数据分布稀疏问题，因此为减少互动强度对互动质量结果判定的影响，在剔除时间戳异常、没有互动评论和重复的样本后，最后选择评论互动强度适中的虚假新闻样本，即提取发布评论数量处于中间的 500 篇虚假新闻及 7029 条评论作为研究数据。

4.2 数据预处理

4.2.1 群体互动质量分类

某篇虚假新闻评论的内容相关度为相关新闻评论数占评论总数的比，某篇虚假新闻评论的回复详细程度为新闻评论的平均字符数取自然对数的标准化得分。文本选用相关系数判断标准来度量评论的内容相关度和回复详细度[33]，判断标准认为相关系数大于等于 0.6 时相关强度为强相关，那么当内容相关度大于等于 0.6 时认为内容相关度高，当回复详细度大于等于 0.6 时认为回复详细度高，反之则认为内容相关度低和回复详细度低。应用百度 AipNLP 接口得到每条评论的情感倾向，本文认为当某篇虚假新闻下评论情感极性为 1 和 0 的总数量占比大于等于 0.5 时判定情感倾向积极，反之则认为情感倾向消极。

对 500 篇虚假新闻及 7029 条评论进行内容相关度、回复详细度和情感倾向计算，根据前文描述的高质量群体互动判定规则（内容相关度高、回答详细程度高和情感倾向积极 3 个条件符合两个及以上），对 500 篇虚假新闻评论是否达成高质量群体互动进行编码。若达成，编码为 1；反之，编码为 0。最终群体互动质量分类结果如表 1 所示：

表 1　500 篇虚假新闻文本及评论群体互动质量分类结果

Table1 Classification Results of Group Interaction Quality of 500 Fake News Texts and Comments

序号	群体互动质量分类	分类达成高/低质的新闻数量/篇	分类达成高/低质的新闻评论数量/条
1	内容相关度高/低	386/114	5463/1566
2	回复详细度高/低	126/374	1871/5158

3	情感倾向积极/消极	93/407	1175/5854
4	群体互动质量高/低	154/346	2181/4848

4.2.2 特征变量提取

本文从内容、形式和情感特征 3 个维度依次选取能够代表虚假新闻文本的特征变量。

内容特征是虚假新闻传播的核心内容，是指与文本内容本身相关的特征，主要包括主题类型、所属领域、传播意图、互动方式等语义和传播特征。主题类型和所属领域属于语义特征，两者存在相似性。虚假新闻通常含有目的性，具有特定的传播意图，另外考虑到不同新闻作者在虚假新闻传播时互动方式的不同，最终确定主题类型、传播意图和互动方式作为虚假新闻文本内容特征变量。

形式特征是虚假新闻传播信息的关键手段，是指与文本表现形式有关的特征。劣质信息一般是风格随意、形式单调的，但准确、明晰且规范的优质表达会提高信息传递的质量、效果和效率。可以考虑的形式特征变量包括内容形式、内容字符数、句子数量、词语数量、话题标签数、地名和人名数量、图片数量、是否有视频、是否有链接、是否有@符号、发布年份、发布月份、评论数量。

情感特征是指文本中能够表达情感的特征。一篇文本当中不仅包含情感特征，还包括情绪特征，同时新闻作者为博得公众关注，常增加情感符号的使用[34-35]。因此，最终确定情感属性、正面情感词数、负面情感词数、表情数量、情绪符数量、情绪种类、情绪强度、情绪数量为本文情感特征变量。

最后本文共确定了 24 个特征变量，因各个特征变量存在差异，故需对其进行编码处理，转化为数值型变量。这些特征变量的具体含义及编码方式如表 2 所示：

表 2　虚假新闻群体互动质量特征变量的编码方式

Table2 Encoding Method of Characteristic Variables of Fake News' Group Interaction Quality

特征变量类型	特征变量	具体含义及编码方式
内容特征变量	主题类型	分为科技、军事、教育、事故、政治、健康、财经、娱乐和社会 9 类，分别用 1-9 表示
	传播意图	分为误导公众意图、操纵舆论意图、吸引注意意图 3 类，分别用 1-3 表示
	互动方式	分为反驳、真相、反驳+真相，分别用 1-3 表示
形式特征变量	内容形式	分为文字、文字+图片、文字+视频、文字+图片+视频 4 类，分别用 1-4 表示
	内容字符数	指新闻文本中包含文字和符号的数量，数值统计编码
	句子数量	指新闻文本中包含句子的数量，数值统计编码
	词语数量	指新闻文本中包含词语的数量，数值统计编码
	话题标签数	指新闻文本中包含话题标签的数量，数值统计编码
	地名和人名数量	指新闻文本中包含地名和人名的数量，数值统计编码
	图片数量	指新闻文本中包含图片的数量，数值统计编码
	是否有视频	
	是否有链接	分为“否”和“是”2 类，分别用 0-1 表示
	是否有@符号	
	发布年份	指新闻发布的年份，分为 2014 至 2021 共 8 类，分别用 1-8 表示

	发布月份	指新闻发布的月份，分为 1 月至 12 月共 12 类，数值统计编码
	评论数量	指新闻文本中包含评论的数量，数值统计编码
情感特征变量	情感属性	分为“消极”、“中性”和“积极”3 类，分别用 0-2 表示
	正面情感词数	指新闻文本中包含正面情感词汇的数量，数值统计编码
	负面情感词数	指新闻文本中包含负面情感词汇的数量，数值统计编码
	表情数量	指新闻文本中包含表情的数量，数值统计编码
	情绪符数量	指新闻文本中包含情绪符号（！？等）的数量，数值统计编码
	情绪种类	指新闻文本中包含 7 种基本情绪种类的数量，分别用 0-7 表示
	情绪强度	指新闻文本引起公众情绪波动的程度，分为 10 级，计算公式为 情绪强度=⌈|情感积极概率−0.5|×20⌉，分别用 1-10 表示
	情绪数量	指新闻文本中包含 7 种基本情绪的数量，数值统计编码

4.2.3 最优特征子集选择

如表 1 所示，群体互动质量高为 154 个样本，高质量比低质量的比率接近 1:2，该数据集的分布是不平衡的，目前一般采用欠采样和过采样技术进行样本均衡处理，但由于欠采样会导致样本数据丢失，因此本文采用 SMOTE 过采样法处理不平衡数据。在通过 SMOTE 方法生成合成样本来平衡数据集后，采用基于随机森林的序列向前搜索策略筛选出虚假新闻文本最优特征子集。其中，经过随机森林回归算法计算的特征重要性排序，如图 4 所示。再根据图 4 中的排序结果依次增加特征到特征子集中，并计算出不同特征子集与 Kappa 指数的变化关系，如图 5 所示。从图 5 中可以看出，当特征子集数目为 15 时的 Kappa 指数最大。因此，最终选择随机森林特征重要性排名前 15 的特征，分别是内容字符数、词语数量、负面情感词数量、评论数量、发布月份、情绪强度、主题类型、正面情感词数量、地名和人名数量、句子数量、情绪数量、图片数量、传播意图、话题标签数和情绪符数量。

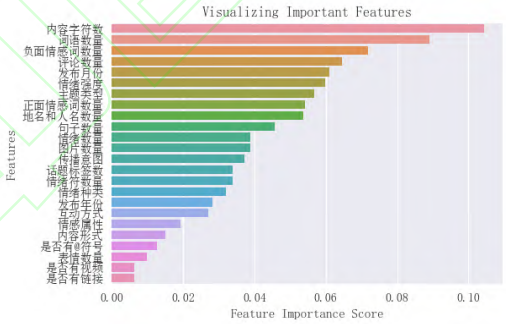


图 4　基于随机森林的特征重要性排序

Fig4. Feature Importance Sorting Based on Random Forest

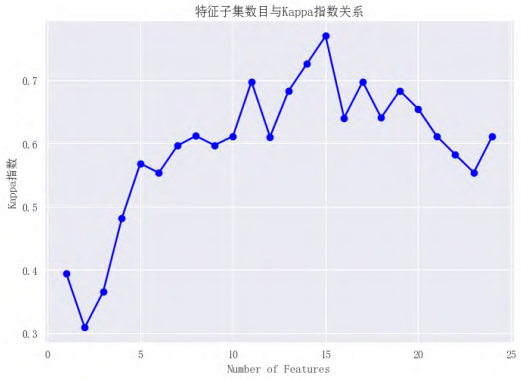


图 5　特征子集数目与 Kappa 指数关系

Fig5. The Relationship between the Number of Feature Subsets and Kappa Index

对选择出的最优特征子集用斯皮尔曼相关系数进行高相关系数滤波，得到 15 个特征的相关系数热力图。如图 6 所示，可以看出仅有 9 对特征相关性大于 0.6，说明筛选出的特征具有一定独立性。这 15 个特征是对群体互动质量最具有显著影响的因素，其特征数量占比为 62.5%，其贡献率达到了 84.3%，根据数量比例与累计贡献率所占比例，说明这 15 个特征具有代表性。综上，筛选出的最优特征子集具有合理性。

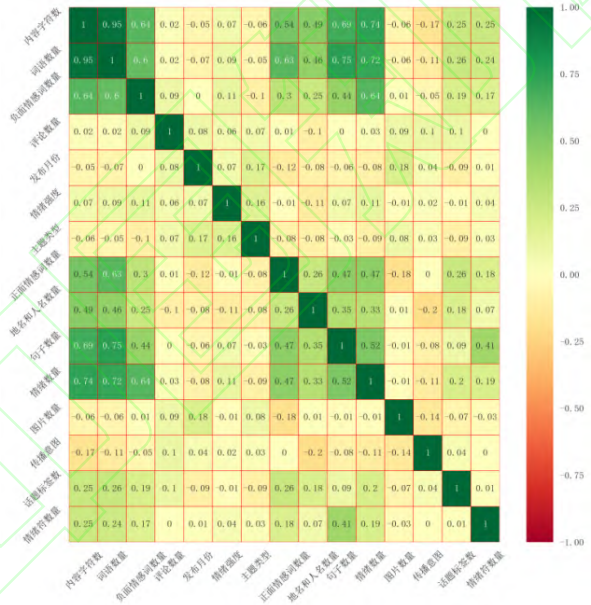


图 6　最优特征子集相关系数热力图

Fig6. Thermogram of Optimal Feature Subset' Correlation Coefficient

4.3　模型构建与评估

按 8:2 的比例划分训练集和测试集，并以上文确定的 15 个变量为特征变量，以群体互动质量为预测变量，构建 LR 模型、SVM 模型、MLP 模型、Adaboost 模型、XGBoost 模型 5 个基本二分类模型进行预测。针对 XGBoost 模型，其树结构对模型结果的影响较大，因此本文先对参数 n_estimators 和 max_depth 进行粗略搜索，再通过遗传算法对其参数进行精确搜索，之后再对参数 learning_rate 和 min_child_weight 用同样的方法进行搜索，进而找到最佳超参数，模型评分参数为 Accuracy，遗传算法适应度函数为 Accuracy，初始种群数量为 100，迭代次

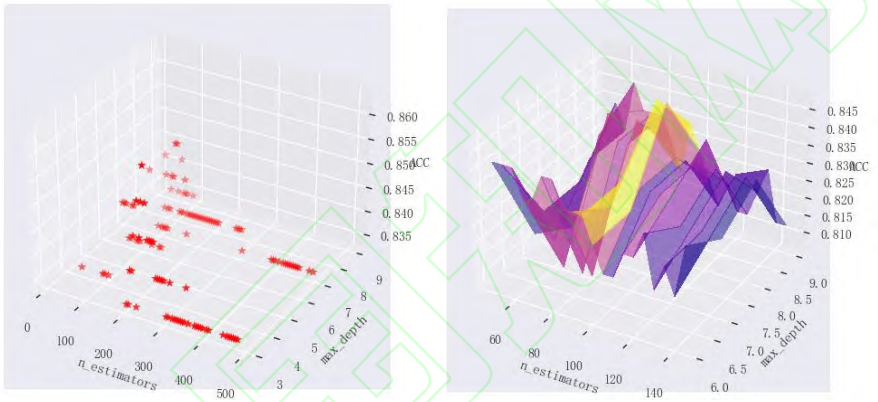
数为 50，每代被保留的个体数量为 4。如表 3 所示，为 4 个超参数含义及粗略搜索范围。

表 3　超参数含义及粗略搜索范围

Table3 Hyperparameter Meanings and Rough Search Scope

超参数	含义	粗略搜索范围
n_estimators	决策树的个数	（20,500,5）
max_depth	每个决策树的最大深度	（1,10,1）
learning_rate	学习率	（0.01,0.5,0.05）
min_child_weight	叶子节点最小权重	（0.1,2,0.01）

针对 n_estimators 和 max_depth 的最佳超参数寻优，首先采用随机搜索方法在较大范围（n_estimators:20-500，max_depth:1-10）进行粗略搜索，粗略搜索图如图 7（a）所示，再根据搜索图缩小搜索范围（n_estimators:50 -150，max_depth:6-10），之后采用遗传算法在该范围内进行精确搜索，调优结果如图 7（b）所示，从图中可以看出，最佳超参数值分别为 95 和 7，随后将其设置为对应参数的最优解。

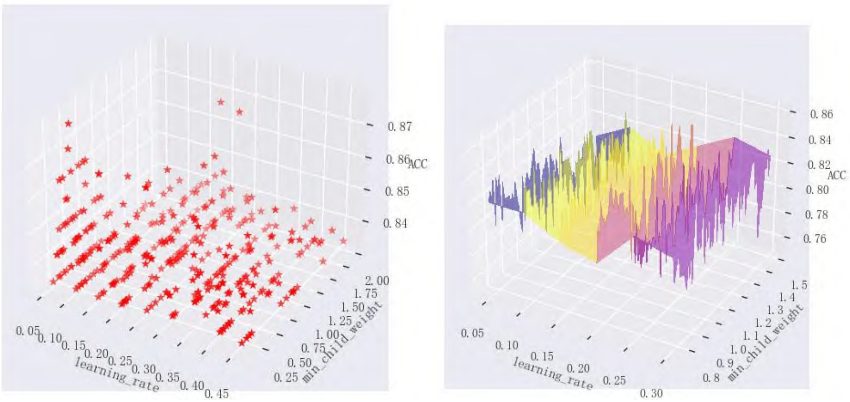


（a）粗略搜索图　　（b）调优结果图

图 7　n_estimators 和 max_depth 的最佳超参数寻优

Fig7. Optimal Hyperparameter Optimization of n_estimators and max_depth

在确定 n_estimators 为 95 和 max_depth 为 7 的基础上，对 learning_rate 和 min_child_weight 的最佳超参数寻优，首先采用随机搜索方法在较大范围（learning_rate:0.01-0.5，min_child_weight:0.1-2）进行粗略搜索，粗略搜索图如图 8（a）所示，再根据搜索图缩小搜索范围（learning_rate:0.05 -0.035，min_child_weight:0.75-1.50），之后采用遗传算法在该范围内进行精确搜索，调优结果如图 8（b）所示，从图中可以看出，最佳超参数值分别为 0.15 和 0.95，随后将其设置为对应参数的最优解。



（a）粗略搜索图　　（b）调优结果图

图 8　learning_rate 和 min_child_weight 的最佳超参数寻优

Fig8. Optimal Hyperparameter Optimization of learning_rate and min_child_weight

以 Accuracy、Precision、Recall、F1-score、AUC 和 Log_loss 作为评估模型优劣的指标，6 种模型的评价指标结果如表 4 所示。从表中可以看出，GA-XGBoost 分类模型的各个指标值都优于其他模型，是最优模型。

表 4　各模型的评价指标结果

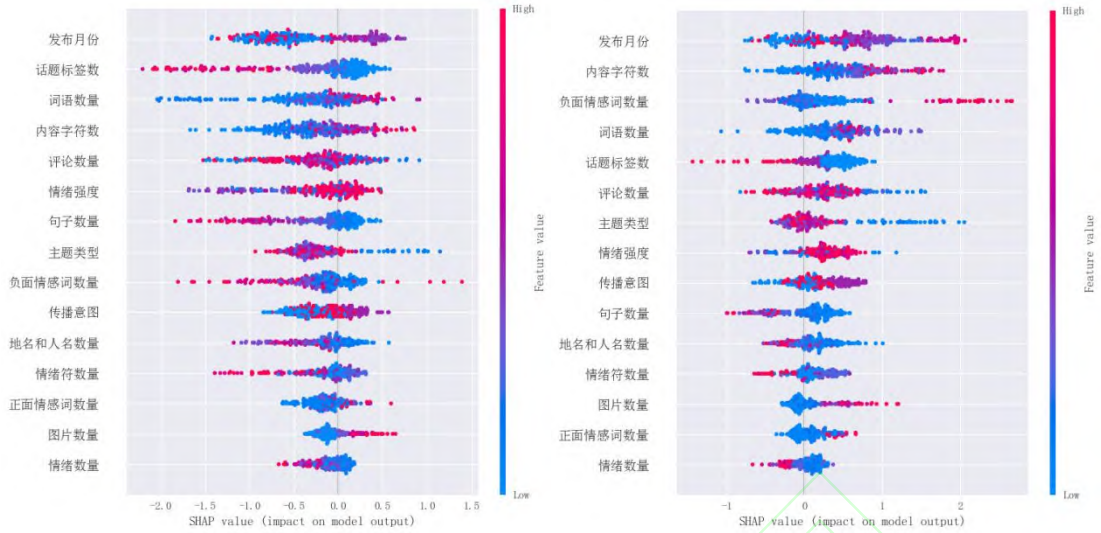
Table4 Evaluation Index Results of Each Model

模型	Accuracy	Precision	Recall	F1-score	AUC	Log_loss
LR	0.5899	0.6094	0.5493	0.5778	0.5908	14.1635
SVM	0.6331	0.7632	0.4085	0.5321	0.6380	12.6726
MLP	0.6043	0.6176	0.5915	0.6043	0.6046	13.6666
Adaboost	0.6835	0.6753	0.7324	0.7027	0.6824	10.9332
XGBoost	0.8201	0.8194	0.8310	0.8252	0.8199	6.2121
GA-XGBoost	**0.8633**	**0.8611**	**0.8732**	**0.8671**	**0.8631**	**4.7212**

4.4　基于 SHAP 模型可解释性分析

4.4.1　全局解释性分析

全局解释性分析采用 SHAP 特征概要图从大小和方向两个方面对解释结果进行可视化，反映特征对预测结果的影响程度、促进或抑制作用。如图 9 所示为该实验的两类样本全局概要图。其中，发布月份、内容字符数和负面情感词数量是影响群体高质量互动达成的重要特征。两种互动质量类别的概要图中排名前 7 的特征为发布月份、内容字符数、负面情感词数量、词语数量、话题标签数、评论数量和情绪强度，说明群体互动质量受形式特征变量影响较大，情感特征变量次之。话题标签数对两种互动质量类别均呈现一定程度的抑制作用，其余 6 个特征与低质量互动呈现显著负相关关系，而与高质量互动呈现显著正相关关系。在低质量互动类别的概要图中，当内容字符数值逐渐增大时，其 Shapley Value 逐渐减小并开始小于 0，而在高质量互动类别中则相反，其值逐渐增大则 Shapley Value 逐渐增大并开始大于 0，说明内容字符数越多对互动质量影响越大。这 6 个特征的较高取值会提高互动质量，说明具有较高数值的这 6 个特征有助于提高群体互动质量。



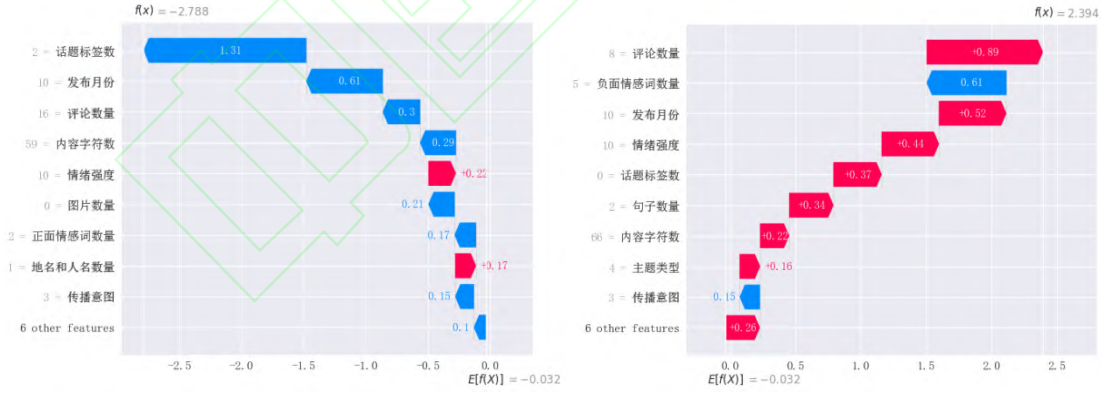
（a）低质量互动类别　　（b）高质量互动类别

图 9　两类样本 SHAP 概要图

Fig9. SHAP Summary Plot of Two Types of Sample

4.4.2　单样本解释分析

单样本解释分析通过 SHAP 瀑布图和力图进行解释结果可视化，反映出单个样本的各特征如何影响并生成该样本结果。如图 10 为在两种互动质量分类中随机抽取的某个样本的 SHAP 瀑布图。对于高质量互动样本 b，特征评论数量=8 表示该特征真实值为 8，右边图形上的数字表示产生了 0.89 的正影响，特征负面情感词数量=5 产生了 0.61 的负影响。对比两个瀑布，两个样本受影响的特征明显不同，说明每个样本的结果由不同的原因造成，这些信息可为平台管理者提供个性化的信息，针对不同的虚假新闻作者制定有针对性的干预策略。



（a）低质量互动类别　　（b）高质量互动类别

图 10　两类样本 SHAP 瀑布图

Fig10. SHAP Waterfall Plot of Two Types of Sample

如图 11 为上文中抽取的两个样本 SHAP 力图，对于 b 样本来说，评论数量、发布月份和情绪强度等粉色特征为正向力量，箭头方向向右增加其预测值，有正向作用力，影响程度随着特征长度的依次减少而减弱，负面情感词数量等蓝色特征则相反。

（a）低质量互动类别

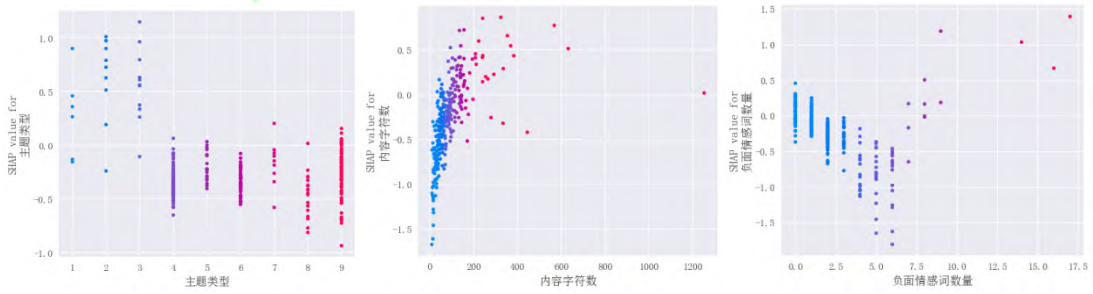
（b）高质量互动类别

图 11　两类样本 SHAP 力图

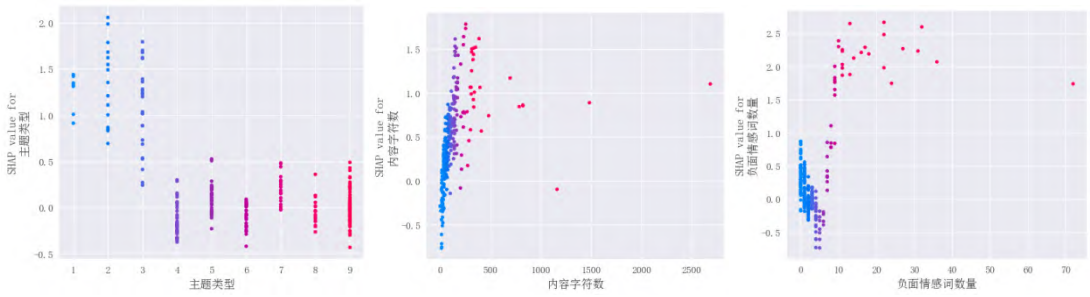
Fig11. SHAP Diagram Plot of Two Types of Sample

4.4.3　单特征解释分析

单特征解释分析通过 SHAP 部分依赖图描述，反映特征对预测结果的边际效应，揭示了单个特征对预测结果的影响程度和方向。如图 12 所示为从内容、形式和情感特征三个方面绘制的两类别样本部分特征依赖图，第一行的三个图表示低质量互动类别的主题类型、内容字符数和负面情感词数量特征依赖图，第二列的三个图则为高质量互动类别的部分特征依赖图。通过纵向比较可以看出每个特征变化对于不同类别的预测结果存在较大差异，以第二列的两个图为例，特征内容字符数对于低质量互动类别群体来说，表现为增长向上的直线，在[0,200]内 SHAP 值逐渐增大，对低质量类别群体的抑制作用逐渐减小为 0 并变为促进，产生了一定负面影响；而对于高质量互动类别群体来说，其表现为斜度更大的向上直线，在[0,500]内其值越大越有利于高质量群体互动。总体来说，内容字符数对于群体互动质量有正向影响。负面情感词数在 2-6 之间会对低质量互动产生很小的抑制作用，对高质量互动产生较大的抑制作用，若其特征值从 2-6 这个频率减少为 1-2 左右，会对高质量互动产生促进作用。总体来说，负面情感词数对群体互动质量有负面影响。将每一类别的不同文本特征进行横向对比，发现情感特征值并不是越大对互动质量越促进，而且三个特征的整体走势截然不同，说明各特征对群体互动质量的影响方向和力度各有差异。通过对六幅图的比较，可以发现对于低质量互动群体来说，有时新闻作者的文本表达并不能有效提高群体互动质量，可能是新闻事件消极或其他原因造成的，对于互动质量更高的群体来说，新闻作者需要保持良好的文本表达形式，但表达的文本特征并不是其值越高越好，更应该关注公众的阅读体验感。



（a）群体低质量互动类别



（b）群体高质量互动类别

图 12　两类样本 SHAP 部分特征依赖图

Fig12. Sectional SHAP Feature dependency Plot of Two Types of Sample

5　结论

本文从探究良性互动及其成因角度出发，构建了基于 RF-GA-XGBoost 的虚假新闻群体互动质量预测模型，并使用 SHAP 解释框架进行全局解释、单样本解释和单特征解释分析，通过提升群体互动质量来抑制虚假新闻负面影响。研究发现虚假新闻文本的内容字符数、词语数量、负面情感词数量等特征是影响群体互动质量的重要因素，各特征对群体互动质量有不同的影响程度和方向。本文的可解释模型增强了人们对群体互动质量分类结果的信任，弥补了虚假新闻检测领域可解释性研究的不足，为辅助社交媒体平台优化运营策略和功能设计提供了可靠依据。

结合以上研究结论，为通过群体高质量互动而最大限度地抑制虚假新闻负面影响，提出如下建议：

（1）为社交媒体平台设定合理力度的审核机制。根据不同特征在不同取值范围时对不同群体互动质量类别的促进或抑制作用，设定相关审核机制规范虚假新闻文本表达。

（2）为社交媒体平台用户设定合适的奖惩力度机制。针对社交媒体平台的评论用户发布的优质内容给予奖励，而对劣质内容给与惩罚干预，比如虚拟金币和积分等。

（3）指引意见领袖发布新闻内容引导舆论。根据高质量互动的新闻文本特点，干预意见领袖发布新闻内容的历史风格，促使其对发布内容进行整合规范，充分发挥意见领袖的正面引导作用。

在未来的研究工作中，本文计划将从以下两个方面行探索：一是进行多特征交互解释分析，分析不同特征之间的交互效应，深入探究文本特征中两两特征交互对互动质量结果的影响；二是根据时间戳特征寻找早期时间节点研究虚假新闻发布早期达成高质量群体互动是否存在规律，此外也会增加公众对新闻作者情感倾向这一因素作为群体互动质量的判定依据。

（致谢：首先我要感谢我的指导老师对我所研究的选题、思路和实验所给与的启发和耐心指导；其次感谢在我遇见难题时为我提供帮助的同学和给予我鼓励的亲人，同时感谢对本文提供重要参考资料和公开数据的学者们，另外也感谢各位审稿人百忙之中提出重要的审阅意见和补充建议，最后感谢国家和省基金项目所提供的资助支持。）

参考文献：

[1] 中国互联网络信息中心. 第 51 次中国互联网络发展状况统计报告[R/OL]. [2023-03-02]. https://www.cnnic.net.cn/n4/2023/0303/c88-10757.html. (China Internet Network Information Center. Statistical Report of the 51rd Chinese Internet Development[R/OL]. [2023-03-02]. https://www.cnnic.net.cn/n4/2023/0303/c88-10757.html.)

[2] 毛震东, 赵博文, 白嘉萌, 等. 基于传播意图特征的虚假新闻检测方法综述[J]. 信号处理, 2022, 38(06): 1155-1169. (Mao Zhendong, Zhao Bowen, Bai Jiameng, et al. Review of Fake News Detection Methods Based on the Features of Propagation Intention[J]. Journal of Signal Processing, 2022, 38(06): 1155-1169.)

[3] 张国标, 李洁, 胡潇戈. 基于多模态特征融合的社交媒体虚假新闻检测[J]. 情报科学, 2021, 39(10): 126-132.(Zhang Guobiao, Li Jie, Hu Xiaoge. Fake News Detection Based on Multimodal Feature Fusion on Social Media[J]. Information Science, 2021, 39(10): 126-132.)

[4] 曾子明, 王婧. 基于 LDA 和随机森林的微博谣言识别研究——以 2016 年雾霾谣言为例[J]. 情报学报, 2019, 38(01): 89-96. (Zeng Ziming, Wang Jing. Research on Microblog Rumor Identification Based on LDA and Random Forest[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(01): 89-96.)

[5] 翟玥, 夏志杰, 王筱莉, 等. 突发事件中公众参与应对社会化媒体不实信息的意愿研究[J]. 情报杂志, 2016, 35(09): 104-110. (Zhai Yue, Xia Zhijie, Wang Xiaoli, et al. Research on Public Participation Willingness in Combating Unconfirmed Information on Social Media in Emergency Events[J]. Journal of Intelligence, 2016, 35(09): 104-110.)

[6] Zhao L, Yin J, Song Y. An Exploration of Rumor Combating Behavior on Social Media in the Context of Social Crises[J]. Computers in Human Behavior, 2016, 58: 25-36.

[7]Shu K , Cui L, Wang S, et al. dEFEND: Explainable Fake News Detectionos[C]. In: Proceedings of the 25th ACM SIGKDDInternational Conference on Knowledge Discovery & Data Mining, Anchorage, AK. New York, USA: ACM, 2019: 395-405.

[8] 陈娟, 刘燕平, 邓胜利. 政府辟谣信息的用户评论及其情感倾向的影响因素研究[J]. 情报科学, 2017, 35(12): 61-65,72. (Chen Juan, Liu Yanping, Deng Shengli. Research on User Reviews of Government Rumor-refuting Information and Factors Influencing Their Emotional Tendencies[J]. Information Science, 2017, 35(12): 61-65,72.)

[9] 王诣铭, 夏志杰, 罗梦莹. 虚假新闻传播事件中提升社交媒体群体互动质量的博弈研究[J]. 情报杂志, 2019, 38(12): 98-106,140. (Wang Yiming, Xia Zhijie, Luo Mengying. A Game Study on Improving the Quality of User Group Interaction in the Spread of Fake News of Social Media[J]. Journal of Intelligence, 2019, 38(12): 98-106,140.)

[10] 阮宏飞, 贾明, 张喆. 信息互动对上市公司传闻治理的影响[J]. 管理科学, 2022, 35(03): 131-146. (Ruan Hongfei, Jia Ming, Zhang Zhe. Effects of Information Interaction on Listed Firms' Rumor Governance[J]. Journal of Management Science, 2022, 35(03): 131-146.)

[11] Sundar S S. Technology and Credibility: Cognitive Heuristics Cued by Modality, Agency, Interactivity and Navigability[M]. Metzger M J, Flanagin A J. Digital Media, Youth, and Credibility. Cambridge, MA: The MIT Press, 2007: 73-100.

[12] 杨善林, 王佳佳, 代宝, 等. 在线社交网络用户行为研究现状与展望[J]. 中国科学院院刊, 2015, 30(02): 200-215. (Yang Shanlin, Wang Jiajia, Dai Bao, et al. State of the Art in Social Network User Behaviors and its Future[J]. Bulletin of Chinese Academy of Sciences, 2015, 30(02): 200-215.)

[13] 张大勇, 许磊, 孔洪新. 社交媒体用户群体互动行为特征研究——以微信用户群分享为例[J]. 情报理论与实践, 2019, 42(10): 97-101,116.(Zhang Dayong, Xu Lei, Kong Hongxin. Users' Interaction Behavior Characteristics of Social Media: An Exploration of Wechat Group's Sharing Behavior[J]. Information Studies: Theory & Application, 2019, 42(10): 97-101,116.)

[14] Berg D V E P, Arentze A T, Timmermans J H. New ICTs and Social Interaction: Modelling Communication Frequency and Communication Mode Choice[J]. New Media & Society, 2012, 14(06): 987-1003.
[15] Sussman S W, Siegal W S. Informational Influence in Organizations: An Integrated Approach to Knowledge Adoption[J]. Information systems research, 2003, 14(01): 47-65.
[16] Sanford B C. Influence Processes for Information Technology Acceptance: An Elaboration Likelihood Model[J]. MIS Quarterly, 2006, 30(4): 805-825.
[17] 李贺，张世颖. 移动互联网用户生成内容质量评价体系研究[J]. 情报理论与实践, 2015, 38(10): 6-11,37.(Li He, Zhang Shiying. Research on Quality Evaluation System of Mobile Internet User Generated Content[J]. Information Studies: Theory & Application, 2015, 38(10): 6-11,37.)
[18] Mihaylov T, Nakov P. Semanticz at Semeval-2016 Task 3: Ranking Relevant Answers in Community Question Answering Using Semantic Similarity Based on Fine-Tuned Word Embeddings[C]. arXiv Preprint, arXiv: 1911.08743.
[19] Nakov P, Hoogeveen D, Arquez L M`, et al. Semeval-2017 Task 3: Community Question Answering[C]. arXiv Preprint, arXiv: 1912.00730.
[20] Xiong J, Feng X, Tang Z. Understanding User-to-User Interaction on Government Microblogs: An Exponential Random Graph Model with the Homophily and Emotional Effect[J]. Information Processing & Management, 2020, 57(4): 102229-102229.
[21] Wang C, Zhou Z, Jin X, et al. The Influence of Affective Cues on Positive Emotion in Predicting Instant Information Sharing on Microblogs: Gender as a Moderator[J]. Information Processing and Management, 2017, 53(3): 721-734.
[22] Charlet D, Damnati G. Simbow at Semeval-2017 Task 3: Soft-Cosine Semantic Similarity Between Questions for Community Question Answering[C]. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval), Vancouver, Canada: ACL, 2017: 315-319.
[23] Sun H, Wang G, Xia S. Text Tendency Analysis Based on Multi-granularity Emotional Chunks and Integrated Learning[J]. Neural Computing and Applications, 2020, 33(14): 1-11.
[24] Fan R, Xu K, Zhao J. An Agent-based Model for EmotionContagion and Competition in Online Social Media[J]. Physica A: Statistical Mechanics and its Applications, 2018, 495: 245-259.
[25] Li C, Bai J, Zhang L, et al. Opinion Community Detection and Opinion Leader Detection Based on Text Information and Network Topology in Cloud Environment[J]. Information Sciences, 2019, 504: 61-83.
[26] 易明，姚玉佳，胡敏. 融合 XGBoost 与 SHAP 的政务新媒体公共价值共识可解释性模型——以“今日头条”十大市级政务号为例[J]. 图书情报工作, 2022, 66(16): 36-47. (Yi Ming, Yao Yujia, Hu Min. An Interpretable Model for New Government Media Public Value Consensus Integrating XGBoost and SHAP: Taking the Top 10 Municipal Government Accounts of the Jinri Toutiao as an Example[J]. Library and Information Service, 2022, 66(16): 36-47.)
[27] 刘天畅，王雷，朱庆华. 基于 SHAP 解释方法的智慧养老服务平台用户流失预测研究[J/OL]. 数据分析与知识发现: 1-18[2023-09-06]. http://kns.cnki.net/kcms/detail/10.1478.g2.20230209.1700.004.html. (Liu Tianchang, Wang Lei, Zhu Qinghua. Research on User Churn Prediction of Smart Senior Care Service Platforms Based on SHAP Interpretation Method[J/OL]. Data Analysis and Knowledge Discovery: 1-18[2023-09-06]. http://kns.cnki.net/kcms/detail/10.1478.g2.20230209.1700.004.html.)
[28] 聂卉，吴晓燕. 结合梯度提升树算法与可解释机器学习模型 SHAP 的抑郁症影响因素研究[J/OL]. 数据分析与知识发现: 1-17[2023-09-06]. http://kns.cnki.net/kcms/detail/10.1478.G2.20230504.1700.006.html. (Nie Hui, Wu Xiaoyan. Combining Gradient Boosting Tree with Interpretation Method SHAP to Detect Factors on Depression[J/OL]. Data Analysis and Knowledge Discovery: 1-17[2023-09-06]. http://kns.cnki.net/kcms/detail/10.1478.G2.20230504.1700.006.html.)

[29] Ekanayake I U, Meddage D P P, Rathnayake U. A Novel Approach to Explain the Black-box Nature of Machine Learning in Compressive Strength Predictions of Concrete Using Shapley Additive Explanations (SHAP)[J]. Case Studies in Construction Materials, 2022, 16.
[30] Lundberg S, Lee S I. A Unified Approach to Interpreting Model Predictions[C]. arXiv Preprint, arXiv: 1705.07874.
[31] Guo B, Ding Y, Yao L, et al. The Future of False Information Detection on Social Media[J]. ACM Computing Surveys, 2020, 53(4): 1-36.
[32] Nan Q, Cao J, Zhu Y, et al. MDFEND: Multi-domain Fake News Detection[J]. arXiv Preprint, arXiv: 2201.00987.
[33] Zhelezniak V, Savkov A, Shen A, et al. Correlation Coefficients and Semantic Textual Similarity[C]. arXiv Preprint, arXiv: 1905.07790.
[34]Jin Z, Cao J, Zhang Y, et al. Novel Visual and Statistical Image Features for Microblogs News Verification[J]. IEEE Transactions on Multimedia, 2017, 19(3): 598-608.
[35] Castillo C, Mendoza M, Poblete B. Information credibility on Twitter[C]. In: Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India: ACM, 2011: 675-684.

通讯作者（Corresponding author）： 白云鹤（BaiYunhe），ORCID：0009-0005-8374-8843，E-mail:15142531981@163.com。

基金项目：本文系“国家自然科学基金项目”基金项目（项目编号：71771111）、“辽宁省社会科学规划基金项目”基金项目（项目编号：L14BTJ004），“辽宁省教改项目”基金项目（项目编号：2021-39）的研究成果之一。
The work is supported by National Natural Science Foundation of China(Grant No.71771111), Liaoning Provincial Social Science Planning Foundation(Grant No.L14BTJ004), Liaoning Province Education Reform Project Foundation(Grant No.2021-39).

作者贡献声明：
温廷新：设计研究方案，论文修改与定稿；
白云鹤：提出研究思路，设计研究方案，数据采集处理与实验，论文起草、修改与定稿。

利益冲突声明：
所有作者声明不存在利益冲突关系。