基于 BERTopic 与 GPT 模型的社交媒体虚假信息文本主题内容研究

万宏静[1]　崔　琦[1]　程　谦[2]
（1. 华北电力大学人文与社会科学学院，北京　102206；
2. 北京航空航天大学计算机学院，北京　100191）

摘　要：[**目的 / 意义**] 从不同主题维度对社交媒体虚假信息文本进行内容分析，有利于揭示社交媒体虚假信息传播特点与规律，针对性提升社交媒体的信息质量，进一步推动社交媒体健康发展。[**方法 / 过程**] 基于 BERTopic 主题模型，从公开数据集 MCFEND 及 CHEF 中提取共 26 478 条虚假信息相关主题，并使用 GPT4.0 模型进行主题标签凝练，实现对社交媒体虚假信息主题内容特征的深入分析。首先，利用 BERTopic 模型对预处理后的社交媒体虚假信息文本数据进行 SBERT 文本向量化、UMAP 降维、HDBSCAN 聚类与 MMR 主题优化，并从中自动提取 30 个核心主题。其次，引入 GPT 模型，提供 prompt（提示词）凝练主题标签，提升主题标签的准确性和可解释性。最后，进一步归纳主题词，分析主题提取结果、主题内容强度及主题时间演化特征。[**结果 / 结论**] 研究发现，社交媒体虚假信息传播具有如下特点与规律：社交媒体虚假信息主题内容泛化且跨国传播较为明显；特定虚假信息主题具有较高关注度并关联热点话题；地域差异或文化背景驱动社交媒体虚假信息主题内容出现分化。

关键词：BERTopic 模型　GPT 模型　社交媒体　虚假信息　主题模型
分类号：G206　TP181
DOI：10.31193/SSAP.J.ISSN.2096-6695.2025.03.03

0　引　言

在当今数字化信息传播环境中，网络社交媒体凭借其开放性、互动性和高效性，已成为用户获取信息、表达观点及社交互动的重要媒介。但是，社交媒体在促进信息流通的同时，也为虚假信息的生成、传播与扩散提供了有利条件。Vosoughi 等[1]调查了 2006 年至 2017 年推特（Twitter）发布的所有新闻报道，发现社交媒体虚假信息的传播极具渗透力与扩散性，这种异常传播不仅污染网络信息环境、干扰公众认知，甚至引发群体偏见与对立，从而对社会稳定构成潜在威胁。因

[作者简介] 万宏静，女，硕士生，研究方向为智能与计算传播，Email：120242207032@ncepu.edu.cn；崔琦，女，副教授，研究方向为跨文化传播，Email：cuiqi99@yeah.net；程谦，男，博士生，研究方向为深度学习、大语言模型优化及神经网络压缩等，Email：qiancheng25@buaa.edu.cn。

此，构建科学、准确、高效的主题分析方法以揭示虚假信息的主题内容特征与传播规律，已成为破解社交媒体治理困局的基础性课题。

主题分析能够从大量文本数据中提取和识别出潜在的主题或话题，发现文本数据中的深层语义信息[2]。现有相关研究大多采用内容分析法选取小样本数据进行人工标注，存在效率低下且主观性强等缺陷。近年来，随着自然语言处理（Natural Language Processing, NLP）的主题建模技术在文本信息分析领域取得显著进展，部分研究采用了传统的隐含狄利克雷分布（LDA）[3]、非负矩阵分解（NMF）[4]、潜在语义分析（LSA）[5]及主题分布式表示（Top2Vec）[6]等模型进行主题建模，但这些模型存在预先确定主题数目、无法综合考虑上下文语义关系或在主题识别过程中缺乏对主题标签自动优化机制等局限。相较而言，基于深度学习的 BERTopic（Bidirectional Encoder Representations from Transformers for Topic Modeling）[7]主题建模方法，由于融合了 BERT[8]的语义理解能力，其处理大规模文本数据的同时可以综合考虑上下文语义关系后自动生成最优主题数量，并在主题一致性和主题多样性上具有一定优势。此外，GPT4.0（Generative Pre–trained Transformer）模型[9]作为当前先进的大规模语言模型，凭借其强大的生成式预训练架构，能够进一步在 BERTopic 模型主题建模的基础上，通过人工提供 prompt（提示词）的方式自动凝练主题标签，从而避免人工编码或在预先确定主题数量时出现偏差。因此，本文尝试将 BERTopic 模型与 GPT 模型相结合，对我国公开数据集 MCFEND（Multi–source Benchmark Dataset for Chinese Fake News Detection）[10]和 CHEF（Chinese Evidence–based Fact–checking dataset）[11]中 26 478 条有效的社交媒体虚假信息文本的主题内容进行深入分析，这种技术协同效应在一定程度上突破了传统主题建模方法的双重局限。一方面，基于 BERTopic 模型可以精准高效提取社交媒体虚假信息最优主题数量；另一方面，借助 GPT4.0 模型可以自动凝练主题标签。研究结果不仅有助于深化对社交媒体平台虚假信息文本主题内容的理解，也能为社交媒体虚假信息检测与治理提供支持，进而优化社交媒体内容治理策略，促进社交媒体内容质量的提升，推动社交媒体平台健康发展。

1　相关研究

1.1　社交媒体虚假信息文本主题内容相关研究

社交媒体是允许用户生成和交换内容的互联网应用平台[12]，而社交媒体虚假信息、错误信息、谣言、假新闻等概念则较为复杂。本文参考 Aïmeur 等[13]与 Fallis[14]对“虚假信息”（“disinformation”或“misinformation”）的定义，将“虚假信息”界定为“与事实不符的欺骗性和误导性信息”。同时，借鉴王剑等[15]对“社交媒体平台虚假信息”的定义，将“社交媒体虚假信息”定义为“在社交媒体平台上产生并传播的不真实或误导性信息”。这一定义可以在一定程度上揭示社交媒体虚假信息的内容特征，为进一步探讨其主题内容提供依据。

目前关于社交媒体虚假信息文本主题内容的研究，大多采用内容分析法，选取小样本数据并经过人工标记处理进行主题分析。张恒瑞[16]利用 Excel 将 857 条虚假信息样本归纳为医疗、食品、安全、时事、常识、政策等 13 类主题。彭柳等[17]采用内容分析法对中国互联网联合辟谣

平台所披露的 150 条有关“新型冠状病毒疫情”的谣言信息，从谣言主题、佐证证据、内容主张等 5 个维度进行文本分析。Rosińska[18] 选取 2019 年波兰事实核查网站上的 192 例虚假信息进行文本分析，发现虚假信息主题涉及政治、经济、社会、伪科学等。近年来，自然语言处理技术不断发展，已经实现了从大量未结构化的文本数据中自动识别并提取关键主题信息，社交媒体虚假信息内容研究方法也逐渐从人工判别转向自动化分析。Ahammad 等[19] 利用 LDA、NMF 和 LSA 三种不同的主题建模算法对 10 254 条有关新型冠状病毒的虚假信息进行主题提取，并手动标注主题，发现虚假信息的主题集中在疫苗、犯罪、隔离、医学、政治和社会等方面。李新月等[20] 采用 LDA 主题模型分析了 1166 条健康辟谣信息的内容主题，发现营养价值、食品添加、疾病治疗等是辟谣平台中健康辟谣信息的关注重点。

1.2 BERTopic 与 GPT 相关研究

2022 年，随着大型预训练语言模型的快速发展，BERTopic 作为一种基于无监督深度学习的预训练主题模型被相关学者提出[7]。BERTopic 利用 BERT 模型强大的语义理解能力，能够在无监督的情况下处理大规模文本数据，并自动提取其中的主题，在主题一致性和主题多样性上具有一定优势。目前，BERTopic 主题模型的应用主要集中在自动提取文本数据中最优主题数量以及主题时间演化分析[21-24]。与此同时，基于 Transformer 架构的 GPT 模型在自然语言生成、机器翻译、情感分析、舆情检测、个性化信息获取等领域也得到广泛应用[25-27]。

综上所述，现有社交媒体虚假信息文本主题内容研究主要采用内容分析法和传统主题建模方法，但二者均存在显著局限。内容分析法因受限于人工编码的低效性和主观偏差，难以应对社交媒体时代的海量数据处理需求。传统主题建模方法（如 LDA）不仅需要依赖人工干预确定最优主题数量，更缺乏自动化主题标签生成机制，导致分析结果的解释性和应用价值受限。因此，本文将 BERTopic 主题模型与 GPT4.0 模型相结合，开展社交媒体虚假信息文本主题内容相关研究。其中，BERTopic 弥补了内容分析法无法高效处理海量数据和人工编码的缺陷，突破了传统主题建模方法需人工确定最优主题数量的主观局限性；GPT4.0 模型实现了自动、精准、高效地识别与凝练社交媒体虚假信息内容主题标签。

2 研究设计

基于 BERTopic 主题模型，从公开数据集 MCFEND 及 CHEF 中提取共 26 478 条虚假信息相关主题，并通过提供 prompt 的方式使用 GPT4.0 模型进行主题标签凝练，实现对社交媒体虚假信息主题内容特征的深入分析。具体研究框架如图 1 所示。

2.1 数据收集与预处理

数据收集分别以我国公开数据集 MCFEND 和 CHEF 作为信息来源。其中，MCFEND 是我国用于检测假新闻的多源基准数据集，而 CHEF 是国内第一个社交媒体事实核查数据集。这两组公开数据集中的虚假信息均来源于各类社交媒体、应用程序及传统在线新闻媒体等，并经过中国互联网联合辟谣平台、腾讯较真事实核查平台以及国外的 AFP Fact Check、MyGoPen 等事实核查机构的核查，其数据内容涵盖政治、经济、健康、文化、社会等各领域[10-11]。目前，这两

个数据集多用于训练或测试虚假信息检测模型。例如，Guo 等[28]在实验过程中将 MCFEND 作为虚假信息检测数据集来训练检测模型，Zhang 等[29]使用 CHEF 数据集验证模型 Conv-FFD 的效率。

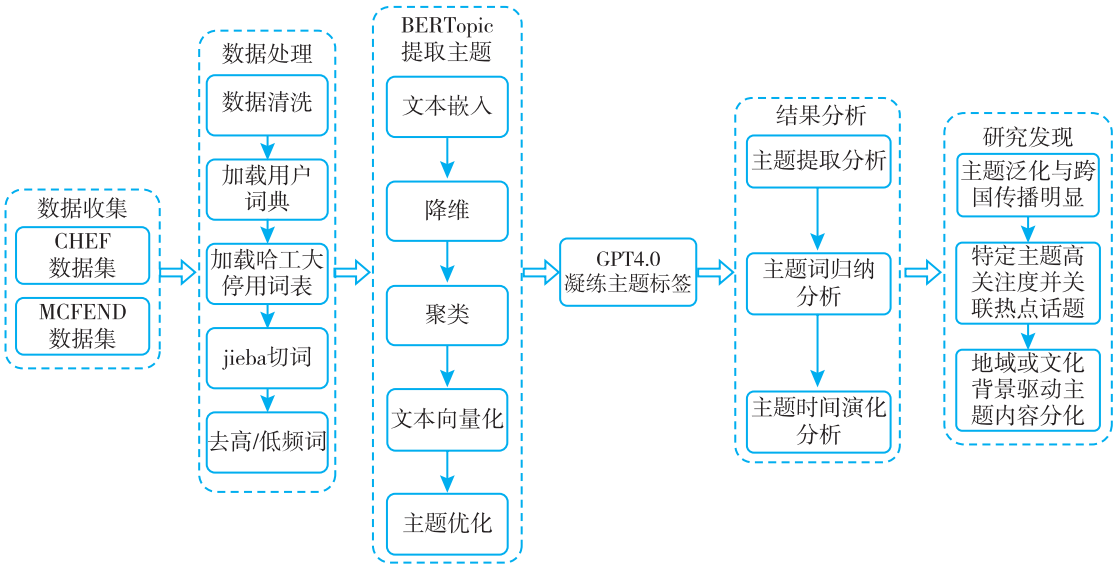

图 1　具体研究框架

虚假信息数据预处理流程包括三个步骤。首先，使用正则表达式对虚假信息文本进行初步文本清洗，删除对文本分析贡献不大且对机器的理解存在干扰的内容，包括不相关的非中文文本、特殊符号和标点符号、空白条目与重复数据等。其次，将“地方新闻”等词汇纳入用户词典，同时在“哈工大停用词表”[30]的基础上，结合文档的词性标注和词频统计，删除高频率低语义词语，如“谣言”“媒体”“网页”等，并构建包含 872 个词的停用词表。最后，使用分词工具 jieba 的精确模式[31]将中文文本切分为词汇单元，并基于语料库词频分布去除高频词与低频词，使模型在处理文本时能够更加聚焦于具有实质性贡献的词汇。如表 1 所示。

表 1　虚假信息数据预处理流程举例

原始文本举例	数据清洗	加载用户词典	加载停用词	切词并去除高 / 低频词
“5G 电信网络正在加速新型冠状病毒的传播，因为手机网络破坏了人们的免疫系统。”	5G 电信网络正在加速新型冠状病毒的传播　因为手机网络破坏了人们的免疫系统	5G 电信网络正在加速新型冠状病毒的传播　因为手机网络破坏了人们的免疫系统	5G 电信网络正在加速新型冠状病毒传播　手机网络破坏免疫系统	5G/ 电信 / 网络 / 加速 / 新型冠状病毒 / 传播 / 破坏 / 免疫 / 系统

在本次收集的原始数据集中，虚假信息文本数据共计 27 715 条，经过初步清洗和去重后，获得文本数据共计 26 478 条。预处理后的虚假信息相关数据情况如表 2 所示。

表 2　预处理后的虚假信息数据情况

数据来源	数据数量	数据占比（%）	原始数据采集时间（年）
微博社区治理	5683	21.4	2011—2023
中国互联网联合辟谣平台	6323	23.8	2019—2023
腾讯较真事实核查平台	1959	7.3	2015—2023
Taiwan FactCheck Center	3458	13.7	2018—2022
MyGoPen	3899	14.7	2015—2022
Factcheck Lab	132	0.4	2020—2021
AFP Fact Check	598	2.2	2019—2023
HKU Annie Lab	290	1	2019—2021
HKBU Fact Check	218	0.8	2020—2021
其他平台	3918	14.7	2015—2021

2.2　基于 BERTopic 模型的主题提取

BERTopic 主题模型能够处理大规模文本数据并从中提取和提炼出核心主题，操作步骤主要包括文本向量化、降维、聚类等。具体建模过程如图 2 所示。

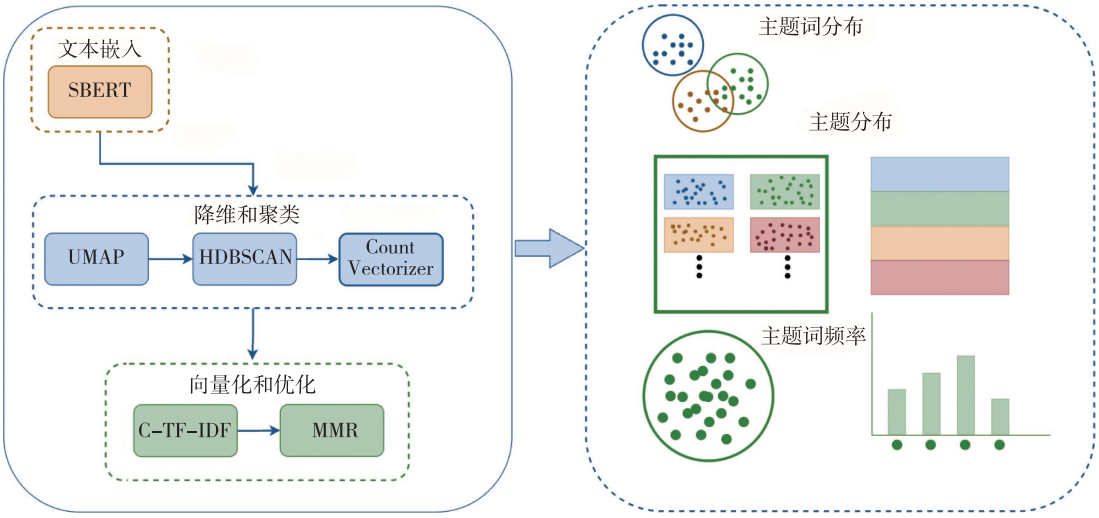

图 2　BERTopic 主题模型具体建模过程

通过微调 BERTopic 模型默认参数，将离群数据文本数量控制在 30% ~ 40% 左右。观察模型的自动聚类结果，在保证聚类结果有效性的基础上，自动提取 30 个社交媒体虚假信息的主题特征作为最终结果，包括主题数量、主题词及主题对应的文本数量等。在模型默认参数的基础上进行多轮微调，得到微调后的模型参数，如表 3 所示。

表 3　微调后的模型参数

模型	参数
UMAP	n_neighbors=15 n_components=5 min_dist=0.0 metric='cosine' random_state=42
HDBSCAN	min_cluster_size=200 min_samples=15 metric='euclidean'
BERTopic	language='Chinese' calculate_probabilities=true ngram_range(3,3)
MMR	diversity=0.3

2.2.1　文本向量化

文本向量化是指在尽可能保留文本原始特征且不丢失语义信息的前提下，将文本转换为高维向量，从而高效地理解文本语言，捕捉词语之间的语义关系，以便进行降维和聚类操作[32]。根据虚假信息数据集的语句特征，将预处理后的文本数据输入到预先训练的 SBERT（Sentence-BERT）模型中的 paraphrase-multilingual-MiniLM-L12-v2 句向量模型[33]中，将其转换为 384 维向量。

2.2.2　降维与聚类

BERTopic 模型默认的 UMAP（Uniform Manifold Approximation and Projection）算法[34]，可以降低上一步创建的语句嵌入维数，为后续应用 HDBSCAN 模型[35]进行聚类提供基础。在 UMAP 模块中，设置邻域参数 n_neighbors=15，在保留局部细节与全局结构之间取得平衡，避免过度聚焦稀疏噪声或遗漏细粒度主题；最小间距 min_dist=0.0 可以增强低维空间的簇内聚合度，以适配虚假信息文本语义密集特性；余弦相似度（metric='cosine'）优于欧氏距离对文本向量的适应性，可以度量高维语义空间距离，有效捕捉文本向量的方向相似性。随后，采用 HDBSCAN 算法对文本进行层次密度的空间聚类，通过观察聚类结果手动调整参数来处理噪声和离群点。通过参数 min_cluster_size=200 和 min_samples=15 分别设定主题最小规模与核心点密度阈值，以过滤偶发噪声干扰并避免碎片化小簇干扰主流主题识别。同时，针对欧氏距离（metric='euclidean'）调整参数，可以保持降维后空间的度量一致性，抑制非显著主题生成。

2.2.3　主题表示与优化

采用 CountVectorizer 算法[36]，通过统计文本中每个词的出现次数，将文本数据转换为词袋模型（Bag–of–Words）[37]，以获取词频分布情况。同时，结合 c-TF-IDF 算法计算每个主题下词汇的 c-TF-IDF 值，确定主题词数量，量化词汇重要性。此外，为实现在特定的语义空间范畴内高效反映各主题中的语义信息量，本文使用最大边际相关性算法（MMR）[38]，通过调节多样性参数（diversity=0.3）降低近义词冗余，利用 BERTopic 模型自动提取 30 个社交媒体虚假信息相关主题作为最终结果。

2.3　基于 GPT4.0 模型的主题标签凝练

为更清晰地理解和解释模型提取的主题，并为后续的分析和应用提供参考，在利用

BERTopic 模型自动获取 30 个社交媒体虚假信息内容主题数量的基础上，使用 GPT4.0 模型进行主题标签凝练。首先，以提供 prompt 的方式给大模型一个示例；然后，分别将文本和主题词导入，得到每个主题标签的凝练结果；最后，通过人工判断标签凝练结果是否准确，并对主题标签和主题词进行简单修正。GPT4.0 指令具体内容如表 4 所示。

表 4　GPT4.0 指令具体内容

指令	具体内容
Prompt1	这是一则虚假信息文本，根据其主题词，将主题标签凝练为“饮食健康”（后附一则虚假信息文本内容，包括主题标签、关键主题词、文本数量等）
Prompt2	请基于以上例子，凝练以下虚假信息文本主题标签（后附全部需要凝练主题标签的虚假信息文本）
Prompt3	在主题标签凝练结果的基础上，对标签进一步优化，并尽量用四个字进行概括

按照表 4 所示的具体指令操作，最后得到主题标签凝练提取结果，文本数量共计 17 471 条，如表 5 所示。

表 5　主题标签凝练提取结果

主题序号	主题标签	主题词	文本数量
0	饮食健康	食品、食物、糖尿病、水果、西瓜、农药	1233
1	宏观经济	金融、央行、亿元、银行	1096
2	疫苗接种	辉瑞疫苗、疫苗接种、病毒影响	1043
3	名人生活	希拉里、安吉丽娜·朱莉、刘翔、名人生活、公众人物	1041
4	社会关注	孩子拐走、转发信息、寻求帮助	987
5	国际关系	中国、中国台湾、日本、中国香港、国际关系、外交问题	920
6	美国政治	特朗普、奥巴马、总统选举、支持者、白宫	830
7	灾害事故	地震、爆炸、火灾、事故处理、自然灾害	752
8	网络安全	手机诈骗、免费充电、用户安全、网络骗局	708
9	中东冲突	叙利亚、伊斯兰、以色列、中东战争	672
10	航空事故	飞机失事、埃塞俄比亚、空难、中国航班	647
11	病毒预防	新型冠状病毒、肺炎预防、感染控制、疫情溯源	646
12	美国大选	投票、选举操控、克林顿、希拉里、选民行为	608
13	国际冲突	俄罗斯、乌克兰、普京、美国、战争、北约	556
14	交通安全	司机行为、安全带、交警执法、交通事故	534
15	疾病预防	运动对心脏的影响、血管健康、心肌梗塞预防、失智症研究	506
16	地方政治	河南、领导、北京、干部、地方新闻	468
17	疫情封城	封城措施、北京、深圳、上海、城市隔离	464
18	教育考试	学生、高考、学校、老师、教育政策	453
19	癌症问题	癌症、致癌物质、牛奶、癌细胞、食物安全	444
20	疫情病例	确诊病例、医院、肺炎、疫情状况	385
21	动物保护	狗肉、动物保护、举报、宠物安全	348
22	政治丑闻	克林顿电子邮件、希拉里调查、美国联邦调查局	300
23	娱乐行业	电影节、国际电影、电视剧、上映计划	285
24	社会保障	退休申请、医保、养老金、福利发放	279

续表

主题序号	主题标签	主题词	文本数量
25	太空探索	地球、机器人、太空探索、外星生命、火星任务	269
26	食品污染	猪肉、猪瘟、瘦肉精、食品安全	232
27	水产安全	小龙虾、螃蟹、寄生虫、海鲜安全	232
28	媒体公信	美国假新闻、虚假信息、信息传播、媒体问题	224
29	个人防护	口罩、防晒、医用口罩、个人卫生	219

3　研究结果分析

3.1　社交媒体虚假信息主题提取结果分析

针对表 5 所示的主题提取结果，分别从主题层次聚类与主题间相似度来分析社交媒体虚假信息的主题内容特征。

3.1.1　主题层次聚类

采用 HDBSCAN 算法，对表 5 所示的提炼主题进行主题层次聚类，聚类结果如图 3 所示。结合表 5 和图 3 进行分析，可以发现几个层次较为明显的主题聚类团簇。在第一个主题团簇中，主题 22“政治丑闻”、主题 12“美国大选”、主题 6“美国政治”、主题 13“国际冲突”、主题 9“中东冲突”、主题 28“媒体公信”聚合在一起，形成一个聚焦于美国政治生态及其国际影响的聚类团簇，涵盖选举争议、国际地缘冲突以及作为传播基础的媒体信任度等议题的虚假信息。在第二个主题团簇中，又包括几个层次的主题团簇。其中，主题 21“动物保护”、主题 3“名人生活”、主题 16“地方政府”、主题 4“社会关注”、主题 18“教育考试”聚合在一起，形成一个

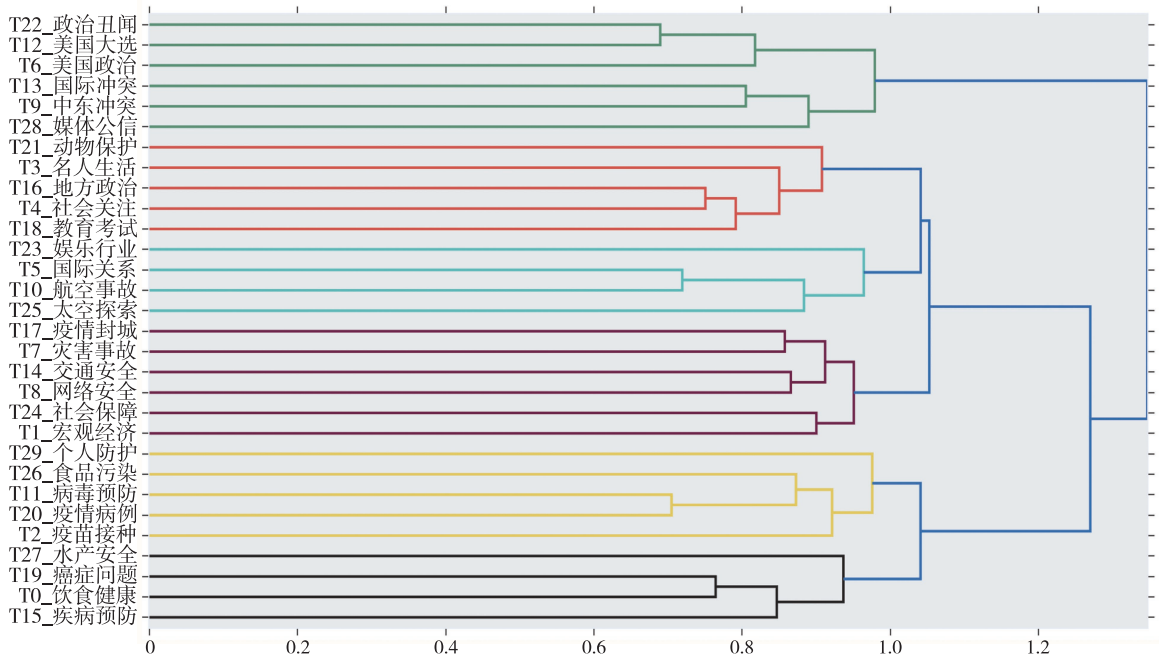

图 3　社交媒体虚假信息主题层次聚类图

聚焦于国内社会民生的聚类团簇，涵盖基层社会治理、高考、儿童走失等议题的虚假信息；主题 23“娱乐行业”、主题 5“国际关系”、主题 10“航空事故”、主题 25“太空探索”聚合在一起，形成一个涉及国际事务与地缘政治的聚类团簇，涵盖国际电影节、中国香港、美日关系等议题的虚假信息；主题 17“疫情封城”、主题 7“灾害事故”、主题 14“交通安全”、主题 8“网络安全”、主题 24“社会保障”、主题 1“宏观经济”聚合在一起，形成一个聚焦于公共安全与社会保障的聚类团簇，涵盖地震、交通事故、网络诈骗等议题的虚假信息；主题 29“个人防护”、主题 26“食品污染”、主题 11“病毒预防”、主题 20“疫情病例”、主题 2“疫苗接种”聚合在一起，形成一个聚焦于公共卫生与疫情的聚类团簇，涵盖新型冠状病毒、病毒传播、疫苗接种等议题的虚假信息；主题 27“水产安全”、主题 19“癌症问题”、主题 0“饮食健康”、主题 15“疾病预防”聚合在一起，形成一个聚焦于健康饮食与疾病预防的聚类团簇，涵盖心脏健康、糖尿病、癌症等议题的虚假信息。

3.1.2　主题间相似度分析

针对表 5 中的主题标签，利用余弦相似度构建了社交媒体虚假信息主题的相似度热力图（图 4）。横纵轴分别列出了 30 个主题，颜色深浅代表两两主题间的相似度分值。整体来看，主对角线均为深蓝色，反映了主题与自身的相似度最高，矩阵呈对称分布，符合相似度矩阵的特征。除对角线外，大多数相似度处于 0.3–0.6 的中低水平，表明各主题间总体差异性较大，但在部分区域存在相对较高的相似度，提示虚假信息在某些主题间可能存在交叉传播。例如：疫情类主题（如主题 2“疫苗接种”与主题 11“病毒预防”）之间的相似度显著偏高；国际冲突类主题（如主

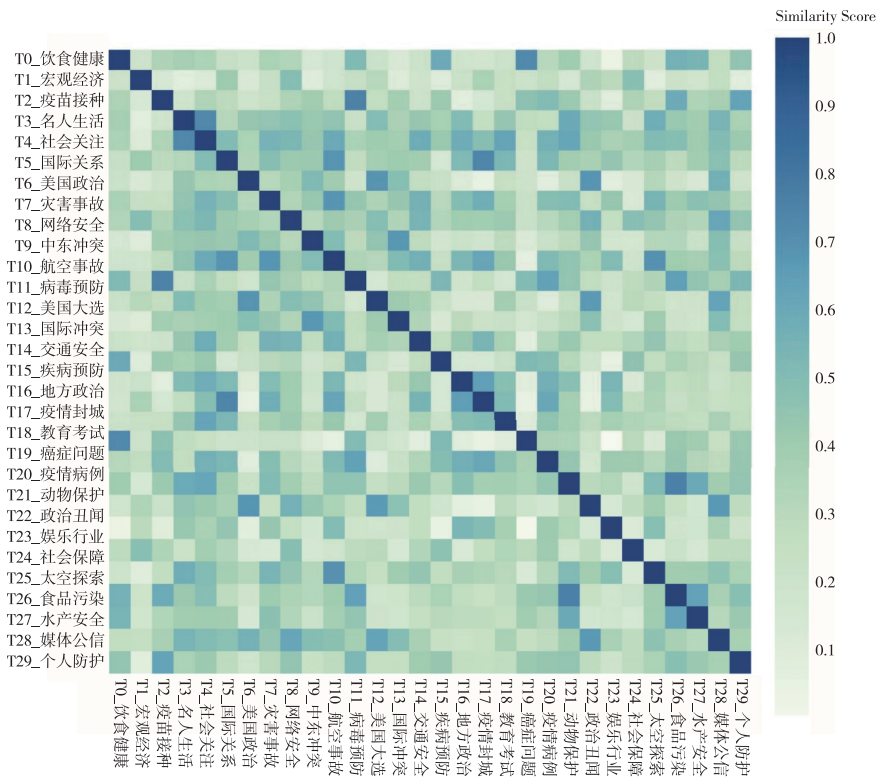

图 4　社交媒体虚假信息主题相似度热力图

题 9“中东冲突”与主题 13“国际冲突”）呈现较强关联；娱乐类主题（如主题 3“名人生活”与主题 23“娱乐行业”）与疫情类、国际冲突类主题的相似度普遍较低。总体而言，该图揭示了虚假信息主题的聚集与分化特征：一方面，部分主题呈现明显的同类聚集，表明在相似领域虚假信息内容更易共享叙事框架；另一方面，不同领域间的低相似度反映了主题传播的相对独立性。这一结果表明，在治理策略上，应针对高相似度的虚假信息主题群组采取联动监测与处置，而对低相似度、独立性强的虚假信息主题则可采用差异化干预措施，以提升治理的针对性与有效性。

3.2 社交媒体虚假信息主题词分析

在进行主题词内容分析时需首先确定主题词数量。按照 c-TF-IDF 分值进行排序，c-TF-IDF 分值越高的主题词越能代表主题反映的内容。根据每个词在该主题中 c-TF-IDF 分值选择主题词，绘制出社交媒体虚假信息主题词语义表征效力折线图（图 5）。大部分主题在超过三个主题词后，主题词的 c-TF-IDF 分值也随之下降，折线变化也趋平缓。这意味着主题的 c-TF-IDF 得分较为接近，主题内部语义结构较为均匀，若此时继续增加主题词数量并不能显著提升表达主题语义信息量，因此在进行主题词内容分析时需重点关注每个主题下前三个主题词的内容。

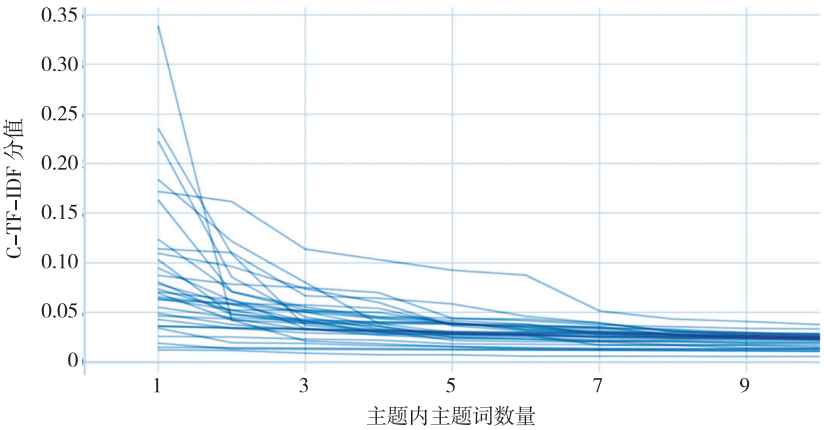

图 5 社交媒体虚假信息主题词语义表征效力折线图

在确定主题词数量的基础上，利用 BERTopic 模型对主题之下的主题词内容进行分析，以更加具体地反映社交媒体虚假信息的内容特征，进一步识别出社交媒体虚假信息的主题焦点。例如，若某主题的主题词包括“食品”“食物”“糖尿病”等，则可以推测这一主题的虚假信息与饮食健康有关。

按照虚假信息文本主题数量排序，对排在前八位的部分主题及其主题内的主题词进行可视化展示（图 6）。从整体上看，社交媒体虚假信息的前八个主题涉及饮食健康、政治经济、社会关注、国际关系、灾害事故类等焦点话题。具体到某个主题，可以看到主题 0“饮食健康”涉及的主题词包括食品、食物、糖尿病、水果、西瓜、农药，这说明“饮食健康”主题所涉及的社交媒体虚假信息大多与饮食、疾病、健康等虚假信息相关。因此，主题词覆盖的具体内容可以反映社

交媒体各类虚假信息的核心范畴，为理解社交媒体虚假信息的主题内容提供较为全面的视角，有助于更好地监测虚假信息传播路径及其可能产生的影响范围。

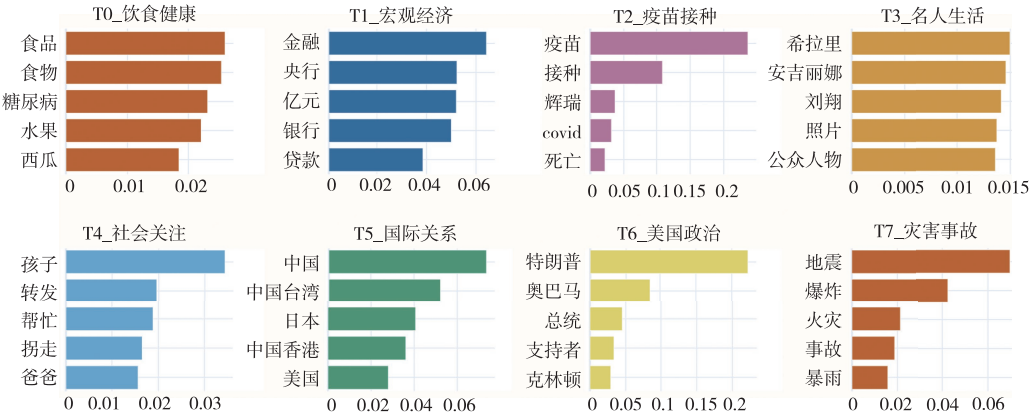

图 6　部分社交媒体虚假信息主题及主题词构成

3.3　社交媒体虚假信息主题时间演化分析

利用 BERTopic 动态主题模型，可以更好地比较分析社交媒体虚假信息不同内容主题的时间演化趋势。图 7 展示了主题 0 至主题 12 部分社交媒体虚假信息主题的时间演化曲线，横轴表示时间跨度，纵轴表示虚假信息文本中每个主题出现的频次，不同颜色折线代表不同的社交媒体虚假信息主题。图中，主题 0“饮食健康”、主题 2“疫苗接种”、主题 11“病毒预防”等健康类虚假信息主题在 2020 年至 2022 年显著增加并出现高峰，且与新型冠状病毒疫情暴发的时间段基本吻合。这一趋势出现的主要原因是疫情期间公众对健康类信息需求激增，但健康类信息的专业性较强，使得受众无法有效辨别真实信息而轻信并传播虚假信息，促使虚假信息在社交媒体算法机制的加持下迅速扩散，这表明虚假信息的高峰期通常与重大社会事件或公共卫生事件相关。此外，主题 1“宏观经济”、主题 6“美国政治”、主题 12“美国大选”等经济类和政

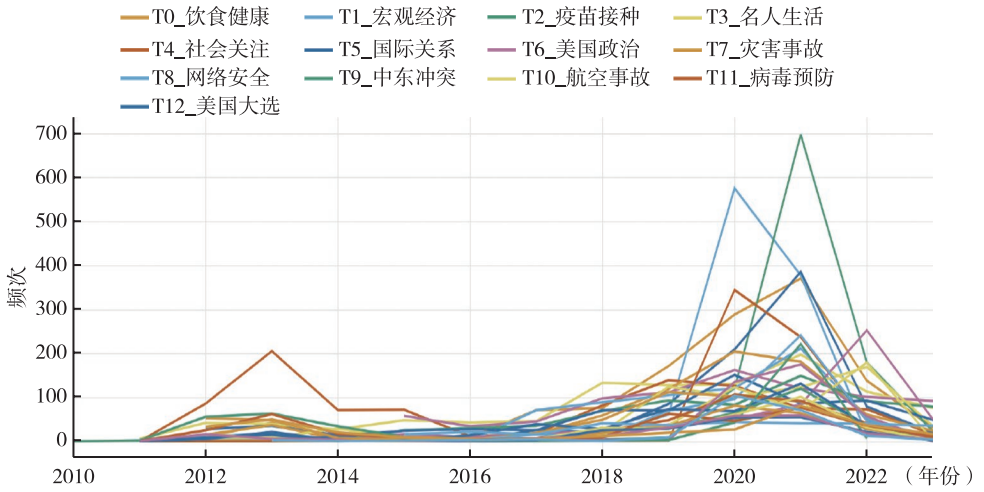

图 7　部分社交媒体虚假信息主题的时间演化曲线

治类虚假信息在时间序列上呈现明显的同步波动，暗示了政策决策和经济预测的交织影响，进一步证实了社交媒体虚假信息不同主题之间的复杂相关性；主题 3“名人生活”、主题 4“社会关注”等社会类虚假信息虽然波动较小，但这类虚假信息主题持续在社交媒体上出现，产生的社会影响也不容忽视。

4　研究发现

4.1　虚假信息主题泛化且跨国传播明显

通过对 26478 条社交媒体虚假信息进行分析，发现虚假信息涵盖的主题包括食品安全、宏观经济、疫苗接种、名人生活、儿童安全、国际关系、自然灾害、网络安全等。这种主题多样性反映出虚假信息不限于特定领域，而是渗透社会生活的各个方面，引发广泛讨论和传播。

在多样性主题的虚假信息中，政治类虚假信息在跨国传播中影响显著，尤其是在重大国际问题或选举过程中，虚假信息的扩散能够迅速影响公众舆论并产生实际社会影响。这与 Starbird 等［39］的研究结论一致，虚假信息的主题多样性及政治类虚假信息跨国传播，不仅是社交媒体信息失序的表现，更是操纵社会认知与行为的复杂机制之一。

4.2　特定主题的虚假信息具有较高关注度并关联热点话题

一些特定主题因其敏感性和高关注度成为虚假信息的集中爆发点。例如，在国际政治领域，有关美国选举、美国政治、国际关系的虚假信息频繁出现，这类虚假信息不但影响国际政治立场，甚至可能对国际关系造成负面影响。此外，重大灾害事故和公共安全领域的虚假信息同样具有高关注度，这类虚假信息利用社交媒体快速传播，极易引发人们的恐慌心理。由此可见，高关注度主题领域的虚假信息易产生严重的社会后果，应当引起高度重视和进一步研究。

4.3　地域特点或文化背景驱动虚假信息的主题内容分化

社交媒体虚假信息的内容大多涉及社会矛盾或政治局势，一般借助当地居民的文化认同感和对外部信息的天然不信任进行传播，这在一定程度上增强了其传播速度及难以被揭穿的特性。因此，地域特点或文化背景在一定程度上驱动着社交媒体虚假信息的主题内容分化。例如，国际政治类虚假信息内容大多围绕某些国家政治动荡、经济危机或社会问题展开，以影响国际舆论或加剧该地区的社会紧张局势。这类信息不仅可能在目标地区引发社会动荡，还可能对国际关系产生不利影响。为有效应对这类虚假信息，各国政府、媒体和公众之间需要加强合作，建立更为透明的舆论环境并提高公众辨别信息真伪的能力，同时促进国际交流与合作，以有效抵御虚假信息带来的危害。

5　结　语

本文将 BERTopic 模型与 GPT 技术相结合，对我国公开数据集 MCFEND 和 CHEF 中的总计 26 478 条社交媒体虚假信息进行内容主题分析。研究发现，社交媒体虚假信息主题呈现多样化特征，且跨国传播现象显著，尤其在政治和国际关系领域表现突出。此外，特定主题如健康医疗、

重大公共事件等因高关注度成为虚假信息传播的集中爆发点，而地域文化差异则进一步驱动了主题内容的分化。本研究不仅有利于对虚假信息传播机制的理解，也为社交媒体平台的虚假信息治理提供了理论依据和实践参考。

受限于数据获取条件，本文未能对文本以外的图片、音视频等多模态社交媒体虚假信息进行分析，也未考虑多语种环境下虚假信息内容主题的差异化特征，这在一定程度上限制了对社交媒体虚假信息的全面理解与分析。未来研究可以结合多模态分析技术，全面捕捉虚假信息的多种传播形式，扩大社交媒体虚假信息的采集范围，从而构建更具普适性的虚假信息识别与理解框架，更加全面地分析社交媒体虚假信息的内容特征。同时，运用深度学习前沿技术研发更为智能化的解决方案，实时监测并有效阻断社交媒体虚假信息传播途径，维护社交媒体平台环境的健康与和谐，将成为社交媒体与网络治理研究的发展方向。

【参考文献】

[1] Vosoughi S, Roy D, Aral S. The spread of true and false news online [J]. Science, 2018, 359(6380): 1146–1151.

[2] 王浩伟，汪璠，王秉琰. 主题视角下生成式人工智能生成内容与用户生成内容的比较 [J]. 情报理论与实践，2023，46（10）：200–207，199.

[3] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993–1022.

[4] Seung D, Lee L. Algorithms for non–negative matrix factorization [J]. Advances in Neural Information Processing Systems, 2001, 13(3): 556–562.

[5] Dumais S T. Latent semantic analysis [J]. Annual Review of Information Science and Technology (ARIST), 2004, 38(1): 189–230.

[6] Angelov D. Top2Vec: distributed representations of topics [J]. arXiv Preprint arXiv:2008.09470, 2020.

[7] Grootendorst M. BERTopic: neural topic modeling with a class–based TF–IDF procedure [J]. arXiv e–prints, 2022: arXiv: 2203.05794.

[8] Rogers A, Kovaleva O, Rumshisky A. A primer in BERTology: what we know about how BERT works [J]. Transactions of the Association for Computational Linguistics, 2021, 8(1): 842–866.

[9] Sanderson K. GPT–4 is here: what scientists think [J]. Nature, 2023, 615(7954): 773.

[10] Li Y, He H, Bai J, et al.Proceedings of the ACM on Web Conference 2024, May13, 2024 [C]. New York:Association for Computing Machinery,2024.

[11] Hu X, Guo Z, Wu G Y, et al. NAACL 2022—2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference,July10–15, 2022 [C]. Seattle:Association for Computational Linguistics (ACL), 2022.

[12] Kaplan A M, Haenlein M. Users of the world, unite! The challenges and opportunities of social media [J]. Business Horizons, 2010, 53(1): 59–68.

[13] Aïmeur E, Amri S, Brassard G. Fake news, disinformation and misinformation in social media: a review [J]. Social Network Analysis and Mining, 2023, 13(1): 30.

[14] Fallis D. What is disinformation? [J]. Library Trends, 2015, 63(3): 401–426.

[15] 王剑，王玉翠，黄梦杰. 社交网络中的虚假信息：定义、检测及控制 [J]. 计算机科学，2021，48

（8）：263–277.

［16］张恒瑞. 社交媒体平台中虚假信息特征分析及治理对策研究［D］. 郑州：郑州航空工业管理学院，2023.

［17］彭柳，陈红飞. 突发性重大公共卫生事件网络谣言文本特征及治理——以中国互联网联合辟谣平台所辟新冠疫情谣言为例［J］.华南师范大学学报（社会科学版），2021（2）：183–192，208.

［18］Rosińska K A. Disinformation in Poland: thematic classification based on content analysis of fake news from 2019［J］. Cyberpsychology:Journal of Psychosocial Research on Cyberspace, 2021, 15(4): 4–5.

［19］Ahammad T, Tanvir A. Identifying hidden patterns of fake COVID–19 news: an in–depth sentiment analysis and topic modeling approach［J］. Natural Language Processing Journal, 2024, 6(1): 100053.

［20］李新月，王莹，韩文婷，等. 健康辟谣信息的内容、质量与优化研究［J］. 情报资料工作，2022，43（3）：84–93.

［21］徐汉青. 融合BERTopic和LSTM的LIS学科AI研究主题演变分析及趋势预测［J/OL］. 情报科学，1–22［2025–02–17］.http://kns.cnki.net/kcms/detail/22.1264.G2.20241218.1456.016.html.

［22］Chong M, Chen H. Racist framing through stigmatized naming: a topical and Geo–Locational Analysis of #Chinavirus and #Chinesevirus on Twitter［J］. Proceedings of the Association for Information Science and Technology, 2021, 58(1): 70–79.

［23］Wang Z, Chen J, Chen J, et al. Identifying interdisciplinary topics and their evolution based on BERTopic［J］. Scientometrics, 2023, 129(11): 7359–7384.

［24］滕广青，江瑶，庹锐. 基于多数据源维度的领域知识演化对比研究：以美国石墨烯领域研究为例［J］. 情报资料工作，2023，44（6）：61–70.

［25］Liu Y, Han T, Ma S, et al. Summary of ChatGPT–Related research and perspective towards the future of Large Language Models ChatGPT ［J］. Meta–Radiology, 2023, 1(2): 100017.

［26］Rathje S, Mirea D M, Sucholutsky I, et al. GPT is an effective tool for multilingual psychological text analysis［J］. Proceedings of the National Academy of Sciences, 2024, 121(34): e2308950121.

［27］李春涛，闫续文，张学人. GPT在文本分析中的应用：一个基于Stata的集成命令用法介绍［J］. 数量经济技术经济研究，2024，41（5）：197–216.

［28］Guo F, Wang X, Xie Y, et al. A survey of datasets for information diffusion tasks［J］. arXiv Preprint arXiv: 2407.05161, 2024.

［29］Zhang Q, Guo Z, Zhu Y, et al. A deep learning–based fast fake news detection model for cyber–physical social services［J］. Pattern Recognition Letters, 2023, 168(4): 31–38.

［30］哈工大停用词表［EB/OL］.［2024–09–03］.https://github.com/LCGCHH/StopWords.

［31］Jieba［EB/OL］.［2024–09–05］.https://github.com/fxsjy/jieba.

［32］Reimers N. Sentence–BERT: sentence embeddings using Siamese BERT–Networks［J］. arXiv Preprint arXiv:1908.10084, 2019.

［33］Sentence–Transformers/Paraphrase–Multilingual–MiniLM–L12–v2［EB/OL］.［2024–09–03］.https://huggingface.co/sentence–transformers/paraphrase–MiniLM–L12–v2.

［34］McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction［J］. arXiv preprint arXiv: 1802.03426, 2018.

［35］Mcinnes L, Healy J, Astels S. Hdbscan: hierarchical density based clustering［J］. Journal of Open Source Software, 2017, 2(11): 205.

［36］Ahmed T, Mukta S F, Mahmud T, et al. 2022 26th International Computer Science and Engineering Conference (ICSEC), December21–23, 2022［C］. Sakon Nakhon: IEEE, 2022.

[37] Zhang Y, Jin R, Zhou Z H. Understanding Bag-of-Words model: a statistical framework [J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1): 43-52.

[38] Carbonell J, Goldstein J. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,August24-28, 1998 [C]. Melbourne Australia: ACM, 1998.

[39] Starbird K, Arif A, Wilson T. Disinformation as collaborative work: surfacing the participatory nature of strategic information operations [J]. Proceedings of the ACM on Human-Computer Interaction, 2019, 3(2): 1-26.

Research on the Topic Content of Social Media Disinformation Texts Based on BERTopic and GPT Model

Wan Hongjing[1] Cui Qi[1] Cheng Qian[2]

(1.College of Humanities and Social Sciences, North China Electric Power University, Beijing 102206, China;
2.School of Computer Science, Beihang University, Beijing 100191, China)

Abstract: [**Purpose/Significance**] Conducting content analysis of disinformation on social media from multiple thematic dimensions can help improve the quality of information on these platforms and promote their healthy development. [**Method/Process**] Based on the BERTopic topic model, a total of 26 478 disinformation related topics were extracted from public datasets MCFEND and CHEF. The topic tags were condensed using the GPT4.0 model by providing prompts, achieving in-depth analysis of the content characteristics of disinformation topics on social media. Firstly, the BERTopic model is used to perform SBERT text vectorization, UMAP conditionality reduction, HDBSCAN clustering, and MMR topic optimization on the preprocessed social media disinformation text data, and automatically extract 30 core topics from it. Secondly, the GPT model is introduced to provide prompt condensed topic labels, improving the accuracy and interpretability of topic labels. Finally, further summarize the topic words, analyze the results of topic extraction, the intensity of topic content, and the temporal evolution characteristics of the topic. [**Result/Conclusion**] The topic of disinformation on social media is generalized and has a more obvious cross-border dissemination. Specific disinformation topics have high attention and are associated with hot topics. Regional or cultural backgrounds drive the differentiation of disinformation themes on social media.

Keywords: BERTopic; GPT; Social media; Disinformation; Topic model

(本文责编：任全娥)