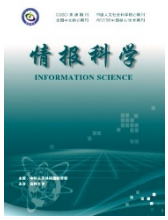

情报科学
Information Science
ISSN 1007-7634,CN 22-1264/G2

《情报科学》网络首发论文

题目：基于 LDA-BERTopic 模型的突发事件短视频用户评论主题挖掘与演化研究
作者：曾金，马全，陈玲
网络首发日期：2025-12-23
引用格式：曾金，马全，陈玲．基于 LDA-BERTopic 模型的突发事件短视频用户评论主题挖掘与演化研究[J/OL]．情报科学.
https://link.cnki.net/urlid/22.1264.G2.20251222.1629.002

网络首发时间：2025-12-23 09:00:10　网络首发地址：https://link.cnki.net/urlid/22.1264.G2.20251222.1629.002

基于 LDA-BERTopic 模型的突发事件短视频用户评论主题挖掘与演化研究*

曾金[1,2,3]　马全[2,3]　陈玲[2,3]

（1.武汉大学信息管理学院，武汉，430072；
2.湖北经济学院信息管理学院，武汉，430205；
3.湖北经济学院大数据与数字经济研究院，武汉，430205）

摘要：【目的/意义】在突发事件网络舆情传播中，短视频发挥着关键作用。本研究提出了一种混合主题建模方法，实现对用户生成内容视频评论的主题特征精准提取，旨在深入分析与系统揭示舆情动态，为科学制定舆情引导策略提供理论依据。**【方法/过程】**首先采用 LDA 模型对数据进行主题建模，提取各个主题的关键词。接着，基于这些主题词筛选原始数据，剔除不含相关主题词的文档，从而减少噪声干扰。最后，使用 BERTopic 模型对处理后的数据进行二次主题建模，以优化主题识别效果。**【结果/结论】**本研究针对抖音短视频平台“湖北冻雨”舆情事件展开分析。研究结果表明，相较于传统的 BERTopic 模型，本文提出的 LDA-BERTopic 模型在抖音短视频用户评论短文本主题识别中展现出更优的主题识别效果。该方法为突发事件网络治理与构建清朗稳定的网络空间提供了新的技术路径和方法支撑。**【创新/局限】**本研究提出了一种结合 LDA 模型与 BERTopic 模型的混合主题建模方法，系统分析了短视频“湖北冻雨”事件的主题演化。研究揭示了用户生成内容的核心关注点，包括事件影响、公众态度及应急响应策略等，进而深化了对短视频在舆情传播中作用的理解。需要指出的是，本研究尚未实现与视频关键帧信息的融合，这在一定程度上限制了研究结果的全面性。

关键词：突发事件；湖北冻雨；网络舆情；LDA 模型；BERTopic 模型

0 引言

在突发公共事件频发的社会背景下，网络舆情会呈现出高度复杂性[1]。特别是在短视频平台场域中，该类媒介凭借其制作便捷性、内容多元性、传播即时性及社交互动性特征，已然发展成为舆情传播的核心场域[2]。相较于传统社交媒体，短视频通过多模态信息呈现方式，在公众情感建构与舆情传导层面展现出更强的渗透力[3]。平台用户评论数据作为情感载体，能够有效反映特定事件的社会认知地图[4]。然而，短视频用户生成内容具有典型的数据稀疏性特征[5-6]，海量评论与目标事件的语义关联度较低，以及特定事件时间序列中的评论数据分布离散，这对突发事件主题演化规律的深度解析与舆情态势的精准研判构成显著挑战。

针对以上问题，为了更准确地挖掘突发事件中短视频用户的短文本语料主题，本研究提出了一种结合 LDA 模型[7]与 BERTopic 模型[8]的混合主题建模方法。该方法通过结合 LDA 模型在主题概率分布建模方面的理论优势，以及 BERTopic 模型在上下文语义表征方面的深度学习特性，从而实现了突发事件短视频语料主题轨迹的动态追踪，有效强化了突发事件主题识别与舆情传播分析。

基金项目：国家社会科学基金一般项目“突发公共卫生事件用户画像构建与舆情演化机制研究”（21BTQ045）
作者简介：曾金（1982-），男，湖北武汉人，博士，副教授，主要从事数据挖掘，信息检索研究；马全（2004-），男，江苏常州人，学士，助理研究员，主要从事文本挖掘，舆情大数据，通讯作者：663244693@qq.com.；陈玲（1991-），女，湖北武汉人，博士，副教授，主要从事政府数据开放研究。

1 相关研究

1.1 突发事件短视频舆情相关研究

突发事件短视频舆情指在突发公共事件发生后，公众借助短视频平台发布相关视听信息，或通过观看追踪事件动态、表达观点及情感的舆情传播现象[9]。根据《国家突发公共事件总体应急预案》分类标准，突发公共事件涵盖自然灾害、事故灾难、公共卫生事件及社会安全事件等类型[10]。目前，现有研究在公共卫生领域取得研究进展，国外学者 Southwick 等[11]通过 TikTok 平台新冠相关视频的实证分析，系统揭示了公众认知观点、信息类型分布和误导信息传播特征，并构建了公共卫生信息传播优化模型。Li 等[12]学者基于 TikTok 视频的用户参与度指标体系，提出了健康信息传播效能提升策略。国内学者李小军等[13]对抖音平台的新冠疫情视频进行了多维度分析与内容解构，揭示了突发公共卫生事件中短视频传播的时空演化规律与传播特点。在自然灾害领域，严玲艳[10]通过社会网络研究法论证了短视频平台在灾情传递、社会救援协同等方面的重要作用，为灾情救助和风险治理提供了相关启示。沈洪洲等[14]采用 Logistic 回归模型定量解析了积极与消极情感型应急知识对短视频传播效能的差异化影响机制。

突发事件短视频舆情相关研究主要关注突发事件短视频在舆情传播中的特点和影响。当前突发事件短视频舆情研究主要聚焦以下三个维度：在传播特征层面，学者们通过内容分析法解构视频文本特征、传播路径及用户行为模式；在社会影响层面，研究着重探讨短视频对公众风险认知、情绪传播及群体行为的影响机制；在治理策略层面，基于传播规律提出舆情引导、信息纠偏及社会协同治理的优化路径。这些研究成果为构建突发事件短视频舆情监测体系、完善应急管理机制提供了理论支撑，对维护公共安全与社会秩序稳定具有重要实践价值。

1.2 BERTopic 主题建模相关研究

Grootendorst[8]提出了 BERTopic 算法，该算法通过结合 BERT[15]的深度语义编码机制和 c-TF-IDF 的加权策略，实现了文本数据深层语义主题的有效识别。杨思洛等[16]利用 BERTopic 模型对我国信息资源管理学科的研究主题及其演化路径进行了系统分析，验证了 BERTopic 在学术趋势分析中的应用潜力与价值。张家惠等[17]基于 BERTopic 模型和 LSTM 深度学习模型进行新型主题预测研究，通过深度语义特征提取与时序预测的协同机制，显著提升了新兴主题预测的准确率与时效性。吴应强等[18]将 BERTopic 应用于政府数据开放政策分析领域，结合扎根理论的质性研究方法，系统解构了数据开放实践与国家战略目标之间的协同关系，并构建了政策匹配度评估指标体系。此外，李豪等[19]提出了一种融合 BERTopic 和 Prompt 的学者研究兴趣生成模型，并在计算机科学领域实现了基于主题建模与提示优化的个性化研究方向生成机制。

综上所述，突发事件情境下短视频舆情研究主要聚焦于危机事件触发后公众通过短视频平台进行信息生产与传播行为规律。研究范畴涵盖自然灾害、事故灾难、公共卫生事件及社会安全事件等多元类型，重点解析短视频内容特征、传播路径以及用户参与行为指标体系，并深入探讨其对公众认知建构与社会舆论场形成的多维度影响机制。特别是在新冠疫情等重大公共卫生事件和典型自然灾害中，学者们通过研究分析揭示了短视频平台的信息传播效能，据此提出了危机沟通策略优化方案与舆情引导机制。

综合现有研究成果，本研究构建了结合 LDA 模型与 BERTopic 模型的混合主题建模方法，应用于突发事件短视频用户评论的主题挖掘。通过结合传统概率主题模型与深度语义建模的优势，旨在建立混合模型下的主题演化路径，系统揭示危机事件背景下短视频舆情的主题演化规律及传播趋势。

2 基于 LDA-BERTopic 模型的突发事件短视频主题演化的方法

本研究主要包括三个部分：首先，进行数据采集与预处理，收集与特定突发事件相关的短视频用户评论数据，并提取评论时间、评论 IP 和评论内容等相关字段，同时对数据进行清洗处理；其次，构建结合 LDA 模型与 BERTopic 模型的混合主题模型，以更精准地识别和提取文本数据中的主题；最后，通过分析事件的时间戳，追踪不同时段内主题的演化，揭示舆情发展的趋势，并提出相应的应对策略。

整体研究框架如图 1 所示。

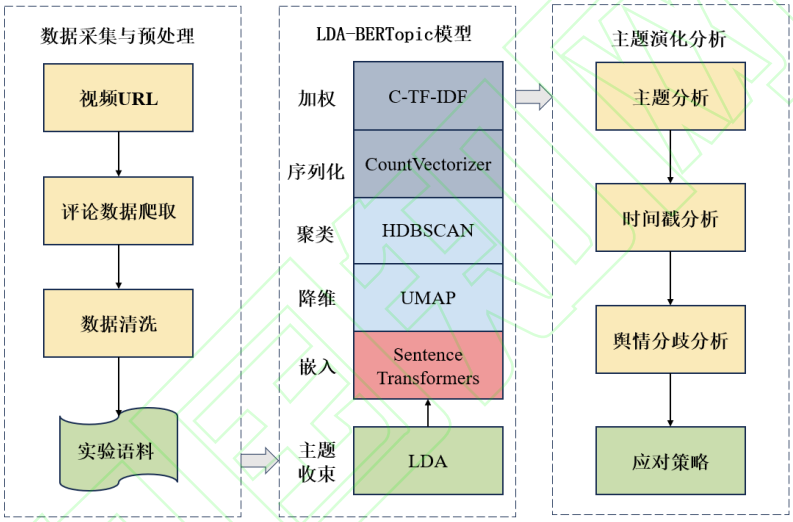

图 1 整体研究框架

Figure 1 Overall research framework

2.1 数据采集与预处理

在短视频平台（如抖音、快手等）上，根据特定突发事件进行相关短视频的检索，并获取对应的唯一视频 ID，从而构建一个短视频 ID 数据集。短视频的 URL 可由基础网址与视频 ID 拼接而成。接下来，利用 Python 中的第三方库“requests”，向这些视频的 URL 发送 HTTP 请求，以获取响应内容。在解析响应数据时，重点关注几个关键字段：首先是评论时间，确保获取评论的准确发表时间，便于后续的筛选与分析；其次是评论的 IP 地址，记录评论者的 IP 信息；最后是评论内容，提取用户的评论文本，这是分析的核心数据。通过以上步骤，能够系统地收集和处理短视频平台上的用户评论数据，并将其整理为 json 格式，以便后续分析与处理。

在获取评论数据后，根据事件发生的时间范围对数据进行筛选，目的是排除往年视频的评论以及事件结束后的评论，以确保数据的时效性和相关性。通过对比评论时间与事件发生时间的关系，可以有效地进行过滤，从而保留与事件紧密相关的评论数据。

数据预处理是数据分析的重要步骤，确保所使用的数据质量高且准确。预处理过程包括去除重复数据（通过 Python 中的第三方库“pandas”识别并删除重复的评论，以减少冗余信息）、处理空值（检查并处理缺失的评论内容，以确保后续分析的完整性），以及清理表情符号（评论数据中可能包含各种表情符号，如“[微笑]”，这些符号对主题识别无实际意义。可以利用 Python 的正则表达式模块“re”构建匹配规则，快速排除这些非文本内容，从而提取出纯文本评论）。

在完成数据清理后，使用 Python 中的中文分词库“jieba”对每条用户评论进行分词处理。根据实际需求，采用动态停用词优化方法来提升分词效果。在此过程中，可以根据实验需求调整停用词表[20]，去除一些常见但对分析无实际意义的词语（如“的”“是”等）。通过这样的处理，可以获得更精准的词汇统计，为后续的主题识别与舆情分析打下基础。

2.2 LDA-BERTopic 模型构建

LDA（Latent Dirichlet Allocation）模型[7]是一种广泛应用于文本挖掘和信息检索领域的概率主题模型。该模型通过将文档集合分解为多个主题，并假设每个主题由一组相关词的分布表示，同时每篇文档可以包含多个主题，从而揭示文本数据中的潜在主题结构。LDA 模型的核心思想是，假设文档是通过多个主题的混合生成的，每个主题由词的分布构成。通过对文本数据进行预处理（如分词和去除停用词），并训练模型来推断文档中的主题分布以及每个主题中的词分布。分析每个学到的主题的词分布，有助于理解文本数据中的隐藏信息和模式，为进一步的文本分类、主题推荐和内容分析提供了基础。尽管 LDA 模型需要大量的预处理工作，并对数据质量有一定要求，但它能够有效地揭示文本数据的结构和内容之间的关系。

在本研究中，首先利用 LDA 模型进行主题提取，对评论数据进行主题建模，并获取每个主题的词分布。在此过程中，困惑度被广泛认可为确定最佳主题数目的重要指标[21]，其公式如下：

$$Perplexity(D)=\exp\left(-\frac{1}{N}\sum_{d=1}^{D}\log p\left(w_d\left|\theta,\beta\right.\right)\right) \tag{1}$$

其中 D 是文档的总数，N 是文档的词汇总数，w_d 是文档 d 中的词，$p\left(w_d\left|\theta,\beta\right.\right)$ 是在给定主题分布 θ 和词汇分布 β 的条件下，生成词 w_d 的概率。

LDA 模型的生成过程涉及多个随机变量和概率分布，可以将其汇总为一个综合公式来表达文档集合中单词的生成过程。该过程可以通过以下公式表示：

$$p\left(w,z,\theta,\beta\left|\alpha,\eta\right.\right)=\prod_{d=1}^{D}p\left(\theta_d\left|\alpha\right.\right)\prod_{n=1}^{N_d}p\left(z_{d,n}\left|\theta_d\right.\right)p\left(w_{d,n}\left|z_{d,n},\beta\right.\right) \tag{2}$$

其中 w 是文档中的词，z 是文档中所有词的主题分配，a 是主题分布的超参数，η 是词汇分布的超参数。

在使用 LDA 模型进行主题提取时，模型会为每个文档分配初始的主题分布。接着，利用每个主题的词分布对原始数据进行处理。在遍历每条评论时，如果发现某个分词不在主题的词分布中，则将其视为偏离主题的噪声数据，并予以剔除。

BERTopic 模型[8]是一种基于 BERT（Bidirectional Encoder Representations from Transformers）[15]的主题建模方法，能够高效捕捉词语之间的深层语义关系和上下文信息[22]。该模型利用预训练的 Sentence-BERT 模型[23]，将文档中的每个

词转换为富含语义信息的嵌入向量。这些嵌入向量不仅能够体现词语的重要性，还能反映其在文档中的语境和语义角色。

本研究采用了中文预训练的“thenlper/gte-large-zh”模型[24]进行句嵌入，以获取文档的语义向量表示。随后，利用 UMAP（Uniform Manifold Approximation and Projection）降维算法[25]对句向量进行降维处理，进而实现数据的有效可视化与分析。在此基础上，应用 HDBSCAN（Hierarchical Density-Based Spatial Clustering of Applications with Noise）[26]聚类算法对降维后的文档嵌入向量进行聚类，将主题相似的文档归为同一组。

在聚类过程中，聚类簇中的代表性文档被视为该簇的主题，从而赋予了每个主题可解释性和直观的含义，使得研究结果更加具有洞察力和实用价值。为了进一步提升主题的准确性，采用了改进的 c-TF-IDF（Class-based Term Frequency-Inverse Document Frequency）算法[8]进行主题表示。与传统的 TF-IDF 算法不同，c-TF-IDF 更加精确地考虑了词汇在特定类别中的重要性，有效地筛选出每个聚类中的代表性词汇，从而形成更加精准的主题。此方法不仅提高了主题的可识别性，还增强了聚类主题的表达能力，为后续的分析提供了更加丰富和可靠的基础。最后，基于 BERTopic 模型对 LDA 模型处理后的文档进行二次主题建模，进一步提升了主题提取的深度和理解精度。

LDA-BERTopic 模型的流程示意图如图 2 所示。

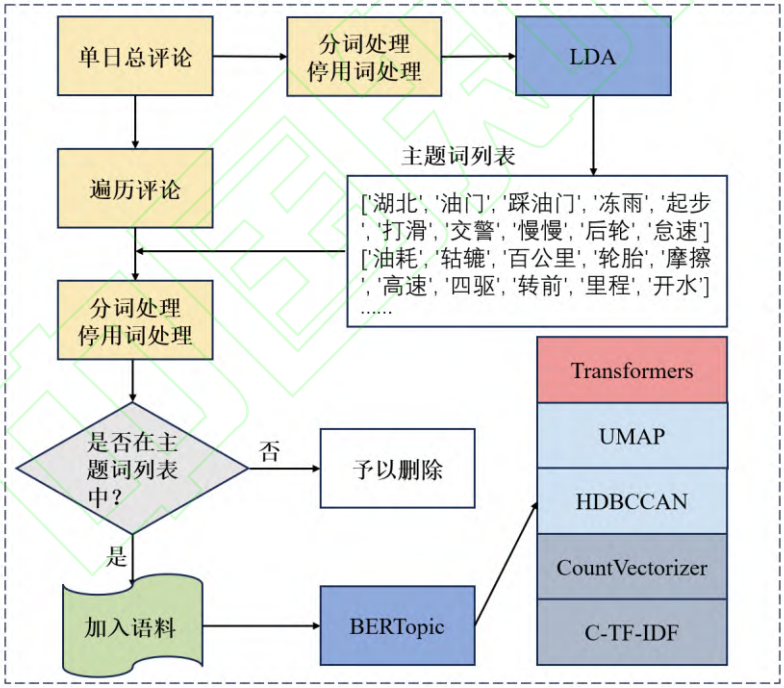

图 2 LDA-BERTopic 模型框架图

Figure 2 LDA-BERTopic model framework diagram

2.3 主题演化分析

在构建 LDA-BERTopic 模型后，依据时间戳对每日的用户评论数据进行了检索并通过模型生成每日的主题词图。这些图表有效地反映了当日用户评论中的主要话题和关键信息。分析每日的主题词图能够帮助及时识别和理解用户关注的热点话题，从而迅速制定相应的应对策略。通过这种实时分析与响应机制，可以有效避免舆情分歧或恶化，确保公众情绪得到妥善管理，并维护稳定的舆论环境。

3 实证分析

3.1 数据采集与预处理

在抖音短视频平台上，以“湖北冻雨”为关键字进行相关短视频 ID 的检索，从而构建短视频 ID 数据集。接下来，将基础的 URL 与每个视频 ID 结合，生成视频的完整 URL。随后，使用 Python 中的“requests”库向抖音视频的 URL 发送 HTTP 请求，获取页面的 HTML 内容。通过 Python 中的“json”库解析并存储用户评论数据，提取与视频相关的信息，包括评论时间、IP 属地和评论内容。整个过程通过遍历视频 ID 列表来实现，确保每个短视频的评论数据都能被准确抓取。

获取评论后，将对数据进行数据预处理。预处理步骤包括时间戳划分、去除重复数据、处理空值和清理表情符号等。首先，根据事件发生的时间范围对数据进行筛选并划分。然后，使用 Python 中的“pandas”库识别并删除重复的评论，以减少冗余信息。接着，检查并处理缺失的评论内容。随后，利用 Python 中的正则表达式模块“re”构建规则，去除评论中的表情符号（例如“[微笑]”等），以提取纯文本评论。最后，将处理后的评论数据以字典形式存储至 json 文件中，便于后续的分析与研究。

通过上述步骤，本研究成功采集了与“湖北冻雨”相关的 204 个短视频，共计 386327 条评论数据。在事件发生的时间范围内（2024 年 2 月 2 日至 2 月 12 日），经过筛选与去重处理，最终获得了 90088 条有效评论数据。

“知微事见”平台是一个专注于舆情监测与分析的先进平台[27]。该平台通过实时采集各类社交媒体和网络平台的数据，进行全面的舆情分析与趋势预测。其中，本研究所采集的每日评论数据与该平台统计的每日抖音短视频数量变化如图 3 所示。

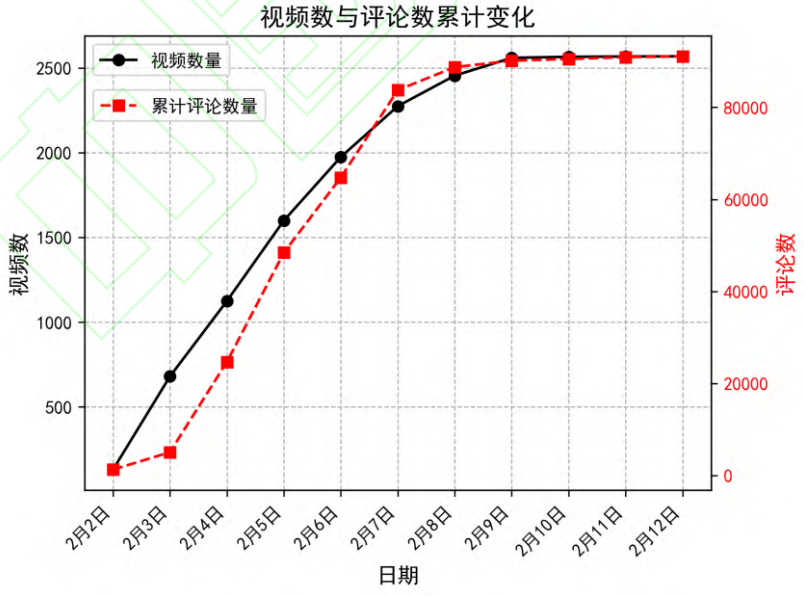

图 3 视频数与累计评论数累计变化

Figure 3 Cumulative change in number of videos and cumulative number of comments

从图 3 中可以看出，该平台统计的每日视频数量与本研究收集的用户评论累计数据之间呈现出较高的拟合度。这表明，所采集的数据能够有效地反映事件的进展，有助于确保实验结果的准确性和可靠性。

3.2 LDA-BERTopic 模型结果对比

在数据采集与预处理之后，对收集到的评论文本进行了中文“jieba”分词处理，利用“jieba”分词将连续的文字串切分为具有实际意义的词语。随后，采用 LDA-BERTopic 模型对每日评论数据进行了主题建模分析。LDA-BERTopic 模型结合了潜在狄利克雷分配（LDA）的主题抽取能力和 BERTopic 的上下文嵌入特征，能够更精准地捕捉评论文本中的潜在主题结构。

为了验证模型的效果，以 2024 年 2 月 2 日的数据作为实验样本，将 LDA-BERTopic 模型的实验结果与 BERTopic 模型进行了对比分析。通过这一对比，不仅揭示了两种模型在主题识别上的异同，还深入探讨了 LDA-BERTopic 模型在处理文本数据时所带来的改进与提升。同时，通过可视化手段直观展示了主题模型的效果，使得分析结果更加清晰易懂。

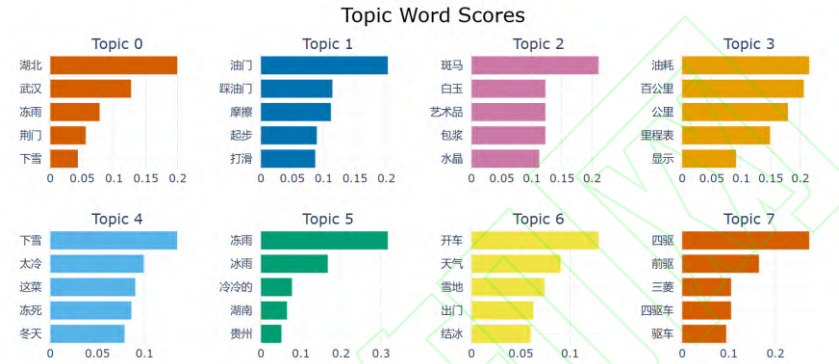

图 4 BERTopic 模型主题词图

Figure 4 BERTopic model subject word diagram

图 4 展示了使用 BERTopic 模型生成的主题词图，从中可以看出，用户评论的主要主题包括“湖北冻雨”“开车打滑”“油耗增加”以及“天气寒冷”。这些主题词反映了由于湖北武汉出现冻雨天气，导致道路打滑现象，进而影响了交通出行。特别地，“油耗”和“公里”这两个词，委婉地表达了在冻雨路面行驶的困难，以及因复杂的交通状况而导致油耗增加的情况。

然而，由于用户评论数据本身具有稀疏性，导致出现了一些与主要主题无关的词汇，如“斑马”“白玉”“艺术品”和“水晶”等。这些词汇可能是用户在评论中随意发表的闲聊或个人看法，并未直接反映核心主题。因此，在分析评论数据时，需要特别注意区分这些背景信息与实际的主要观点，从而更精准地把握用户的意见和需求。

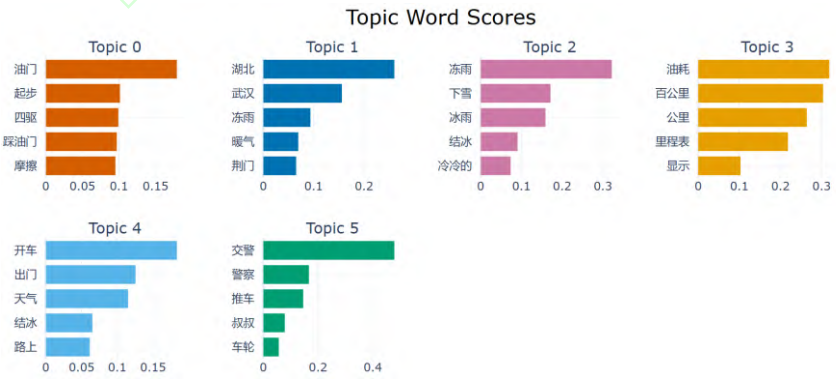

图 5 LDA-BERTopic 模型主题词图

Figure 5 LDA-BERTopic model subject word diagram

图 5 展示了使用 LDA-BERTopic 模型生成的主题词图。与图 4 相比，可以明显看到与主要主题无关的词汇数量有所减少。这表明 LDA-BERTopic 模型在主题抽取方面具有更高的有效性，能够更加聚焦于核心主题。在话题 5 中，提到了交警在路面上的协助，反映出用户对交通管理的支持，同时也暗示了当时交通管理面临的挑战。

以下是 BERTopic 模型与 LDA-BERTopic 模型主题词二维分布可视化，如图 6 和图 7 所示。

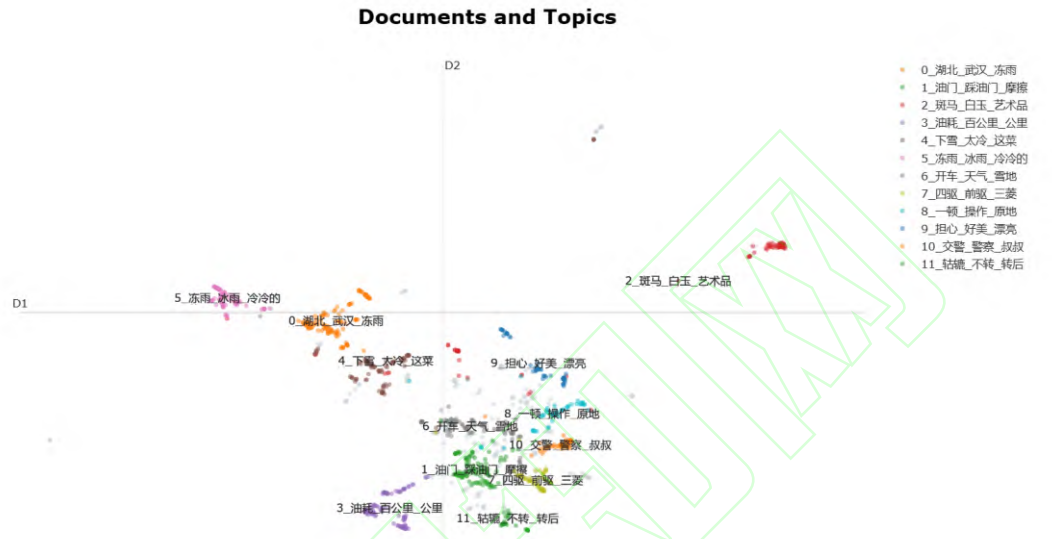

图 6 BERTopic 模型主题词二维分布

Figure 6 BERTopic models 2D distribution of topic words

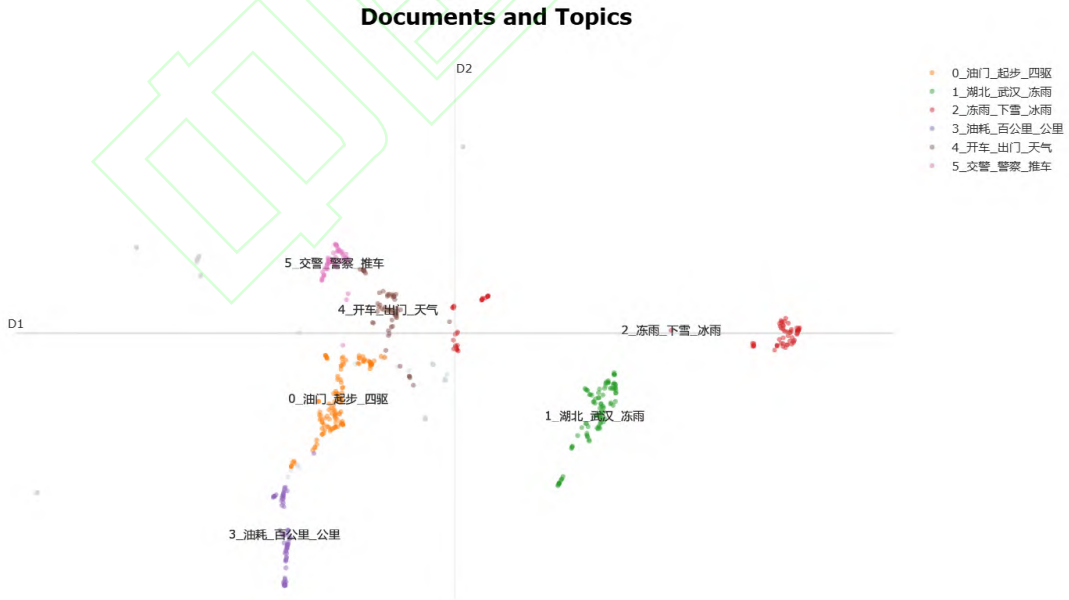

图 7 LDA-BERTopic 模型主题词二维分布

Figure 7 LDA-BERTopic models 2D distribution of topic words

根据 BERTopic 与 LDA-BERTopic 模型的主题词二维分布显示，LDA-BERTopic 在主题词的二维分布上通常更为准确。LDA-BERTopic 能够生成更加

精细且富有解释性的主题，使得主题之间的关系以及主题词的分布得以更加清晰地呈现。

以下是 BERTopic 模型与 LDA-BERTopic 模型主题聚类可视化，如图 8 和图 9 所示。

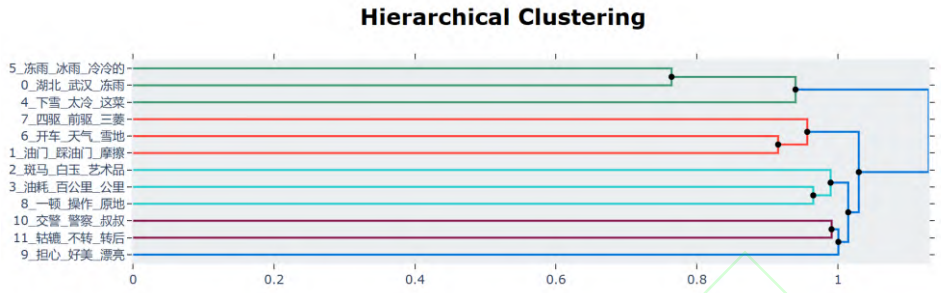

图 8 BERTopic 模型主题聚类

Figure 8 BERTopic model topic clustering

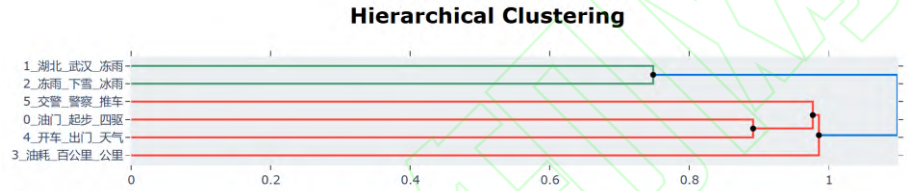

图 9 LDA-BERTopic 模型主题聚类

Figure 9 LDA-BERTopic model topic clustering

通过对比 BERTopic 模型和 LDA-BERTopic 模型在主题聚类方面的表现，可以明显看出，引入 LDA 模型后，无关主题的分支显著减少。这是因为 LDA 模型在预处理阶段有效剔除了与主要主题无关的噪声数据，从而使 BERTopic 模型在后续的聚类过程中能够更专注于核心话题。这样的模型结合策略不仅提高了主题聚类的准确性，还增强了对评论数据中关键信息的捕捉能力，使得最终生成的主题更加紧凑且具有代表性。

3.3 主题演化分析

对 2 月 2 日至 2 月 12 日期间的数据进行了 LDA-BERTopic 模型的主题建模，并从提取的所有主题中总结出了前五个关键词，如表 1 所示。

表 1 2 月 12 日至 2 月 12 日主题词表

Table 1 Topic list of 12 February to 12 February

日期	2 月 2 日	2 月 3 日
主题	油门_起步_四驱_踩油门_摩擦	武汉_湖北_高速_高冷_冻雨
	湖北_武汉_冻雨_暖气_荆门	解决_开水_热水_试试_搞定
	开车_出门_天气_结冰_路上	出门_下雪_开车_上班_冻雨
	油耗_百公里_公里_里程表_显示	车门_冻住_隐藏式_打不开_设计
	交警_警察_推车_叔叔_车轮	冻雨_结冰_衣服_冻住_天气
日期	2 月 4 日	2 月 5 日
主题	湖北_冻雨_下雪_武汉_疫情	湖北_冻雨_疫情_武汉_天气
	冻雨_下雪_结冰_出门_晚点	火车_高铁_内燃机_接触网_影响

	老表_不准_不许_天灾_敢骂 天灾_控制_理解_传递_视频 武汉_多灾多难_英雄_城市_不容易	回家_过年_平安_到家_回来 湖北_天灾_武汉_喷子_骂人 晚点_武汉_高铁_停运_退票
日期	2 月 6 日	2 月 7 日
主题	冻雨_清理_北方_下雪_雨夹雪 湖北_加油_热干面_人民_感动 湖北_骄傲_好样的_人民_点赞 回家_过年_辛苦_平安_希望 发声_谢谢_视频_湖北_出省	湖北_加油_人民_武汉_点赞 冻雨_清理_北方_下雪_南方 湖北_冻雨_河南_理解_科普 撒盐_结冰_融雪剂_提前_压实 中国_团结_人民_力量_善良
日期	2 月 8 日	2 月 9 日
主题	冻雨_湖北_高速_下雪_清理 湖北_点赞_人民_感谢_武汉 中国_团结_人民_善良_最美 感动_温暖_热泪盈眶_评论_画面 加油_湖北_收到_湖南_人民	冻雨_撒盐_北方_科普_清理 湖北_冻雨_河南_高速_天气 视频_中国_加油_东北_传递 湖北_加油_人民_发声_点赞 喷子_湖北_河南_湖南_评论
日期	2 月 10 日、11 日、12 日	
主题	湖北_冻雨_加油_武汉_河南 冻雨_科普_清理_北方_下雪 湖北_人民_点赞_善良_加油	-

2 月 2 日的讨论集中在湖北武汉地区的恶劣天气，包括冻雨和降雪。分析了在这样的天气条件下驾驶所面临的困难和危险，例如轮胎打滑和油耗增加。同时，也提到了交警在这种天气中维持秩序的负责态度。整个讨论涵盖了从天气状况到驾驶挑战，再到交警职责，展现了公众对突发天气事件的关注以及对公共安全的信任。2 月 3 日的讨论继续聚焦于武汉的冻雨天气及其对日常生活的影响。分析了冻雨导致的结冰问题，以及因天气恶劣所引发的出行困难，尤其是上班所面临的挑战。讨论的焦点从天气现象延伸至其对个人生活的深远影响，突显了恶劣天气对城市运作和个人生活节奏的显著影响。2 月 4 日的主题分析继续关注武汉地区的冻雨和降雪，同时引入了疫情因素，突显了湖北武汉所面临的多重挑战。讨论中还涉及了公众对天灾的理解以及传播正能量的态度。疫情的加入不仅彰显了对公共卫生的关注，也反映了社会心态逐渐向积极面对困境并传播正能量的转变。2 月 5 日的主题分析聚焦于冻雨和疫情，同时提到了交通问题，包括高铁晚点和停运。这些问题反映了极端天气对交通系统的冲击，并凸显了公众对节假日期间出行安全的担忧。讨论中，交通问题成为主要焦点，展示了人们对恶劣天气影响下出行状况的高度关注。2 月 6 日的主题分析讨论了清理冻雨后的应对措施，并强调了湖北人民在困难时期展现出的积极态度。特别是，地方特色食物“热干面”被视为精神象征，体现了地方文化的独特性。讨论的焦点转向了对地方精神的赞扬和鼓励，反映了社会情绪的积极转变。2 月 7 日的主题分析强调了湖北的积极应对措施，以及全国人民的团结支持，特别提到了撒盐融雪等应对策略。讨论中突出全国范围内的协作与支持，反映了集体主义精神，并对应对措施进行了科普推广。2 月 8 日的主题分析继续关注湖北的积极应对，表达了对湖北人民的深切感激，并强调了全国人民的团结与善良。持续传递正能量，进一步增强了全国人民与湖北的情感联系。2 月 9 日的主题分析讨论了撒盐清理冻雨的科学方法，再次强调了湖北人民的积极态度以及全国的支持。焦点重新聚焦于具体的应对措施，

继续强化正面情绪的传递。2 月 10 日至 12 日的主题分析中，湖北及冻雨问题仍然是讨论的焦点，强调了人民的善良、团结和加油鼓劲的口号。这几天的讨论主题趋于稳定，形成了对湖北的持续支持和积极的舆论氛围。

在整个突发冻雨事件中，舆情的分歧主要体现在应对措施的有效性和及时性上。为了分析用户对事件处理的反应，本文通过 Python 编写正则表达式，检索用户评论中包含“骂”字样的内容，并合并重复出现的字词（例如，将多次出现的“骂”字合并）。通过生成词云图，可以更准确地呈现用户评论中的关键词，从而深入洞察公众的情绪和观点（见图 10）。

图 10 用户评论词云图

Figure 10 User comment word cloud map

3.4 结果分析

本文讨论了 2024 年初湖北武汉地区遭遇冻雨和降雪天气的多重影响，并深入分析了公众对这一突发自然灾害的反应。通过每日评论数量的变化，可以将该事件划分为三个阶段，划分结果如图 11 所示。这一阶段划分有助于更加清晰地分析公众情绪的变化趋势及其对事件发展的反应。通过梳理不同阶段的讨论，本文从自然灾害的影响和短视频用户生成内容这两个角度，进一步探讨了公众情绪、应对措施以及社交媒体中视频内容的传播与影响。

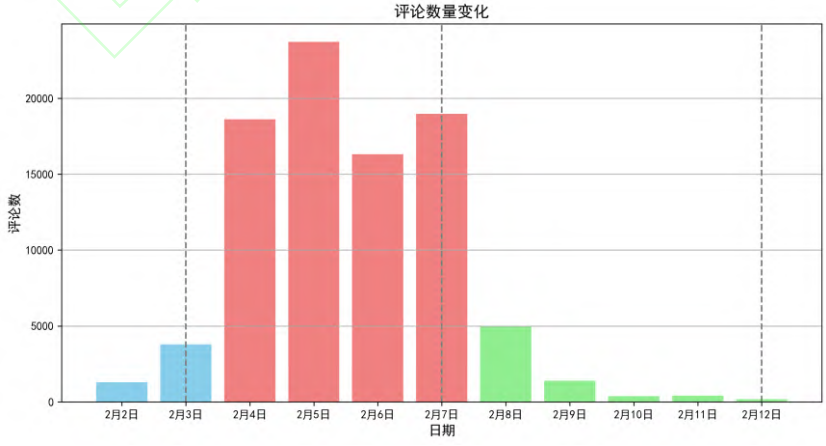

图 11 湖北冻雨事件阶段划分

Figure 11 Stages of freezing rain events in Hubei

（1）突发事件中自然灾害的影响

从 2 月 2 日到 2 月 12 日的讨论来看，湖北武汉地区遭遇的极端天气引发了公众广泛关注，尤其是在初期（2 月 2 日—2 月 3 日），焦点集中在冻雨和降雪对日常生活的直接影响。公众对天气引发的出行困难，尤其是对交通安全的担忧，体现在对车辆性能、油耗以及路面状况的讨论中。这一阶段的讨论呈现出明显的焦虑情绪，反映了公众对自然灾害直接后果的深切关注。例如，讨论中提到的车辆打滑、车门冻结等问题，揭示了人们在极端天气下所面临的实际困难。

随着讨论的深入，公众对湖北及武汉应对恶劣天气的措施表现出了更多的理解与支持。在关注期（2 月 4 日—2 月 7 日），天气对个人生活节奏的影响逐渐成为讨论的焦点，尤其是在春节临近时，公众对“回家过年”的安全问题表现出了深切关切。疫情的叠加效应使得此次讨论不仅聚焦于自然灾害本身，还扩展到了社会心理层面。武汉和湖北人民展现出的坚韧与团结精神成为舆论的焦点，许多人在评论中表达了对当地居民的支持与感动。

进入热度下降期（2 月 8 日—2 月 12 日），公众对湖北人民清理冻雨的积极应对表示赞扬，强调了团结与协作的重要性。可以看出，在面对天灾时，公众情绪逐渐转向积极，更多关注灾后恢复、公共服务保障以及全社会的共同努力。通过这一系列讨论，不仅体现了自然灾害带来的物理困扰，还深刻影响了社会心理，促进了集体主义和互助精神的表达。

（2）突发事件中短视频的用户生成内容

突发事件中的短视频内容，特别是通过社交媒体平台传播的用户生成内容（UGC），在此次自然灾害的讨论中发挥了重要作用。分析显示，短视频不仅展示了湖北人民如何应对极端天气（如撒盐清理道路、解冻车辆等），还传递了人们在天灾面前展现出的坚韧与乐观情绪。用户生成的内容在网络上传播时，激发了社会集体情感的共鸣。例如，视频展示了武汉和湖北人民在困难面前如何团结互助，传播了正能量。通过这些内容，公众不仅关注了灾害本身，还加深了社会对湖北人民的情感联结。

社交媒体平台上的短视频传播凸显了情感与信息的双重作用：情感上，短视频增强了公众的情感共鸣，并激发了对灾后恢复的支持；信息上，用户生成内容为受灾地区的应对措施提供了可视化展示，有效促进了知识的传播与公众的参与感。

然而，在社交平台的内容中，也伴随着虚假信息的传播。例如，一些误导性信息，如“湖北冻雨的懒作为”，在广泛讨论中可能对公众对事件的认知产生负面影响（如图 10 词云图中出现的“骂”字），并进一步加剧信息传播的恶化和舆情的剧烈波动。这类虚假信息不仅可能加剧公众的焦虑情绪，还可能影响决策，妨碍整个社会对事件的有效应对。因此，相关部门必须迅速且透明地通过微博、抖音等社交媒体平台发布真实的处理进展，确保公众及时获得准确信息。通过积极主动地沟通与信息引导，可以有效遏制虚假信息的传播，减少舆情的负面影响，从而维护社会稳定与公众信任。

4 结语

本研究提出了一种结合 LDA 模型与 BERTopic 模型的混合主题建模方法，并将其应用于突发事件短视频用户评论的主题挖掘任务中。BERTopic 模型凭借其在自然语言处理领域的先进性和深度学习能力，成为分析大规模评论数据的强

大工具。为了进一步优化分析效果并应对稀疏数据带来的挑战，本研究巧妙地结合了 LDA 模型，从而增强了对稀疏数据的处理能力。通过这种模型结合的应用，提升了主题识别的精确度，确保在分析突发事件相关评论时能够更准确地捕捉关键主题，为理解公众反应和舆情走势提供了有力的数据支持。

研究表明，短视频用户评论与视频内容特征及文本数据之间存在显著关联[28]。然而，本研究在数据采集阶段受到技术限制，未能同步获取多模态视频数据，这在一定程度上制约了研究的深度。鉴于此，后续研究亟须突破多模态技术瓶颈，通过构建多模态深度学习框架，结合计算机视觉，实现对视频内容特征与文本的联合表征学习。这种多模态融合范式可有效建立视觉特征与用户评论文本之间的语义映射，进而形成一个多维度的 UGC（用户生成内容）分析框架。该方法不仅有助于深化对短视频传播机制的理解，还可以通过特征重要性归因分析，揭示不同模态对用户情感表达的驱动机制，为构建更完善的网络舆情预警体系提供理论支持与实践路径。

参考文献：

[1] 徐海玲,侯亚娟.突发网络舆情事件态势感知模型及其应用实证[J].情报科学,2024,42(5):77-84.

[2] 魏宏程,朱恒民,魏静,等.基于短视频网络的互联网舆情演化研究[J].数据分析与知识发现,2024,8(5):113-126.

[3] 董娜.基于用户生成内容的短视频网络舆情传播生态系统构建[J].图书馆,2022(4):73-81.

[4] 徐孝娟,赵泽瑞.非遗短视频用户信息需求特征及其参与行为研究——以“黄梅戏”短视频在线评论为例[J].现代情报,2022,42(8):74-84.

[5] 胡凯茜,李欣,王龙腾.基于BERTopic模型的网络暴力事件衍生舆情探测[J].情报杂志,2024,43(7):146-153.

[6] YAN X, GUO J, LAN Y, et al. A biterm topic model for short texts[C]//Proceedings of the 22nd international conference on World Wide Web. 2013: 1445-1456.

[7] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[8] GROOTENDORST M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure[DB/OL]. (2022-03-11)[2025-11-19]. https://arxiv.org/abs/2203.05794.

[9] 陈璟浩,王有峰,聂卉梓.突发事件短视频舆情演化分析模型研究[J].信息资源管理学报,2022,12(3):152-164,180.

[10] 严玲艳.群聚、联结与破圈：突发自然灾害事件中短视频平台的应急传播实践[J].福建师范大学学报(哲学社会科学版),2024(3):78-87.

[11] SOUTHWICK L, GUNTUKU S C, KLINGER E V, et al. Characterizing COVID-19 content posted to TikTok: public sentiment and response during the first phase of the COVID-19 pandemic[J]. Journal of Adolescent Health, 2021, 69(2): 234-241.

[12] Li Y, GUAN M, HAMMOND P, et al. Communicating COVID-19 information on TikTok: a content analysis of TikTok videos from official accounts featured in

the COVID-19 information hub[J]. Health education research, 2021, 36(3): 261-271.
[13] 李小军,吴晔,胡璠.社交媒体上新兴传染病的危机传播——对抖音平台新冠肺炎短视频的计算内容分析[J].新闻知识,2020(10):55-67.
[14] 沈洪洲,朱佳,黄仕靖,等.应急知识短视频传播效果研究：基于不同发布端类型的分析[J].情报理论与实践,2024,47(11):101-110.
[15] DEVLIN J. Bert: Pre-training of deep bidirectional transformers for language understanding[DB/OL]. (2018-10-11)[2025-11-19]. https://arxiv.org/abs/1810.04805.
[16] 杨思洛,于永浩.基于BERTopic模型的国内信息资源管理研究主题挖掘与演化分析[J].情报科学,2024,42(8):12-21.
[17] 张家惠,丁敬达.基于BERTopic和LSTM模型的新兴主题预测研究[J].情报科学,2025,43(1):98-105,126.
[18] 吴应强,李白杨,费巍,等.我国政府数据开放研究与国家战略所需的匹配度分析——基于BERTopic模型与扎根理论[J].情报科学,2025,43(1):117-126.
[19] 李豪,张柏苑,邵蝶语,等.融合BERTopic和Prompt的学者研究兴趣生成模型——以计算机科学领域为例[J].情报科学,2025,43(1):127-136,160.
[20] LO R T W, HE B, OUNIS I. Automatically building a stopword list for an information retrieval system[C]//Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR). 2005, 5: 17-24.
[21] 佘硕,林雅玲.基于LDA主题模型的我国突发公共卫生事件应急管理主题热度与趋势分析[J].中国应急管理科学,2024(6):66-85.
[22] Abuzayed A, Al-Khalifa H. BERT for Arabic topic modeling: An experimental study on BERTopic technique[J]. Procedia computer science, 2021, 189: 191-194.
[23] REIMERS N. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks[DB/OL]. (2019-08-27)[2025-11-19]. https://arxiv.org/abs/1908.10084.
[24] LI Z, ZHANG X, ZHANG Y, et al. Towards general text embeddings with multi-stage contrastive learning[DB/OL]. (2023-08-08)[2025-11-19]. https://arxiv.org/abs/2308.03281.
[25] MCINNES L, HEALY J, MELVILLE J. Umap: Uniform manifold approximation and projection for dimension reduction[DB/OL]. (2018-02-09)[2025-11-19]. https://arxiv.org/abs/1802.03426.
[26] Malzer C, Baum M. A hybrid approach to hierarchical density-based cluster selection[C]//2020 IEEE international conference on multisensor fusion and integration for intelligent systems (MFI). IEEE, 2020: 223-228.
[27] 王楠,杜豪,谭舒孺,等.基于深度学习的事件特征提取与舆情反转预测[J].情报杂志,2025,44(3):107-118.
[28] 晋良海,王抒情,王昕煜.基于暴雨灾害短视频的多模态情感特征研究[J].中国安全科学学报,2024,34(7):219-228.

A Study on the Topic Evolution of Short Video User Comments on Breaking News Based on LDA-BERTopic

ZENG Jin[1,2,3] MA Quan[2,3] CHEN Ling[2,3]

(1.School of Information Management, Wuhan University, Wuhan 430072, China;
2.School of Information Management, Hubei University of Economics, Wuhan 430205, China;
3.Institute of Big Data and Digital Economy, College of Information Management, Hubei University of Economics, Wuhan 430205, China)

Abstract: [Objective/Significance] Short videos play a crucial role in shaping public opinion during emergencies. This study introduces a hybrid topic modeling approach to accurately extract the thematic features from user-generated content in video comments. The aim is to conduct a thorough analysis and provide a systematic understanding of public opinion dynamics, offering a theoretical foundation for the development of effective public opinion guidance strategies. **[Methods/Processes]** Initially, the LDA model is applied to extract the themes and keywords from the data. The original dataset is then filtered based on these keywords, removing irrelevant documents to minimize noise interference. Lastly, a secondary topic modeling process using the BERTopic model is employed to further enhance the accuracy of topic recognition. **[Results/Conclusions]** This study examines the public opinion event of 'Freezing Rain in Hubei' on the Shake video platform. The results demonstrate that the LDA-BERTopic hybrid model proposed in this study outperforms the traditional BERTopic model in recognizing themes from short text user comments on ShakeEn video. The method offers a novel technical approach and provides methodological support for online governance during emergencies, contributing to the creation of a more transparent and stable cyberspace. **[Innovation/Limitations]** This study presents a hybrid topic modeling technique combining the LDA and BERTopic models, providing an in-depth analysis of the thematic evolution in the 'Freezing Rain in Hubei' event. It identifies the core concerns in user-generated content, including the event's impact, public attitudes, and emergency response strategies, which enhance the understanding of short videos' role in public opinion dissemination. However, it is worth noting that the study has yet to incorporate video keyframe information, which limits the comprehensiveness of the results.

Keywords: emergency; Hubei freezing rain; online public opinion; LDA model; BERTopic model

作者贡献声明：曾金，提出选题，研究设计，论文修改与最终版本修订。马全，数据采集，数据分析与论文撰写。陈玲，提出修改意见与论文指导。