张岚，李娟，陈静雯，等. 基于微博的强对流天气灾情信息提取与分析：以江苏2021年“4·30”强风雹天气为例[J]. 水利水电技术(中英文)，2025，56(10)：1-16. DOI：10.13928/j.cnki.wrahe.2025.10.001

ZHANG Lan，LI Juan，CHEN Jingwen，et al. Disaster information extraction and analysis of severe convective weather based on Weibo：A case study of severe wind-hail event on April 30，2021 in Jiangsu[J]. Water Resources and Hydropower Engineering，2025，56(10)：1-16. DOI：10.13928/j.cnki.wrahe.2025.10.001

基于微博的强对流天气灾情信息提取与分析：以江苏2021年“4·30”强风雹天气为例

张　岚[1]，李　娟[1]，陈静雯[1]，王啸华[2]

(1. 江苏省突发事件预警信息发布中心，江苏 南京　210041；2. 江苏省气象台，江苏 南京　210041)

摘　要：【目的】强对流天气通常具有突发性和局地性。有效获取社交媒体中的灾情信息，能够弥补强对流天气实况观测密度不足的问题，为灾害应急管理提供信息支撑。【方法】对典型强对流天气数据文本进行中文分词和统计分析，融合气象专家知识，形成适用于强对流天气的灾害主题语料库。将语义信息融入主题模型(LDA)和支持向量机(SVM)分类算法，构建了强对流灾害天气灾情信息提取模型。以江苏2021年“4·30”强风雹天气为例，收集了16 334条原创微博文本信息进行仿真试验。【结果】结果显示：(1)构建的强对流灾情信息提取模型对微博文本中的灾情信息识别与分类效果显著。一次文本主题挖掘提取出天气状况、科普防御、灾情影响、求助救援和其他信息5个主题，对“灾情影响”信息进行二次分类，提取出公共设施、电力通信、车辆交通、农业设施、人员伤亡和其他6种具体灾情信息。经过交叉验证，一次分类平均准确率为92.70%，二次分类平均准确率为90.95%。(2)将强对流天气发生发展过程划分为预警期、突发期和灾后期3个阶段，各类信息均在突发期处于高值区。灾害突发期，公共设施、电力通信信息最多，灾后期，人员伤亡信息讨论度最高。(3)灾情信息数量的空间分布基本与灾害影响严重地区一致。强对流天气发生时，公共设施暴露风险较高，受损现象最普遍。【结论】基于微博的强对流灾情信息提取模型能够有效获取微博文本隐含的灾情信息，反映灾害事件的变化特征和舆论焦点，对灾害监测预警服务和应急决策指挥具有一定参考意义。

关键词：强对流；灾情；社交媒体；微博；信息提取；主题模型；文本分类；降雨

DOI：10.13928/j.cnki.wrahe.2025.10.001　　**开放科学(资源服务)标志码(OSID)：**

中图分类号：X43　　**文献标志码：**A　　**文章编号：**1000-0860(2025)10-0001-16

听语音　聊科研　与作者互动

Disaster information extraction and analysis of severe convective weather based on Weibo：A case study of severe wind-hail event on April 30，2021 in Jiangsu

ZHANG Lan[1]，LI Juan[1]，CHEN Jingwen[1]，WANG Xiaohua[2]

(1. Jiangsu Emergency Early Warning Release Center，Nanjing　210041，Jiangsu，China；
2. Jiangsu Meteorological Observatory，Nanjing　210041，Jiangsu，China)

收稿日期：2025-02-12；**修回日期：**2025-06-28；**录用日期：**2025-06-30；**网络出版日期：**2025-08-22

基金项目：国家自然科学基金项目(42171081)；中国气象局创新发展专项(CXFZ2024J012)；江苏省气象局面上项目(KM202205)

作者简介：张　岚(1987—)，女，高级工程师，硕士，主要从事突发事件预警信息发布及公共气象服务研究。E-mail：499062083@qq.com

通信作者：李　娟(1975—)，女，高级工程师，主任，学士，主要从事突发事件预警信息发布及公共气象服务研究。E-mail：361161860@qq.com

©Editorial Department of Water Resources and Hydropower Engineering. This is an open access article under the CC BY-NC-ND license.

Abstract：[Objective] Severe convective weather is typically characterized by abrupt onset and localized impact. Effectively acquiring disaster information from social media can compensate for the insufficient observation density of severe convective weather and provide information support for disaster emergency management. [Methods] Chinese text segmentation and statistical analysis were performed on text data of typical severe convective weather. By integrating meteorological expert knowledge, a disaster-themed corpus tailored to severe convective weather was developed. Semantic information was incorporated into the latent Dirichlet allocation (LDA) topic model and the support vector machine (SVM) classification algorithm to construct a disaster information extraction model under severe convective weather. Taking the severe wind-hail event in Jiangsu on April 30, 2021 as an example, 16 334 original Weibo text messages were collected for simulation experiments. [Results] (1) The constructed disaster information extraction model for severe convective weather demonstrated remarkable effectiveness in identifying and classifying disaster information in Weibo texts. Through primary topic mining, five themes were extracted: weather conditions, public education on disaster prevention, disaster impact, rescue requests, and other information. The secondary classification was performed on “disaster impact” to extract six specific categories: public facilities, power and communication, vehicle traffic, agricultural facilities, casualties, and others. Cross-validation revealed an average accuracy of 92.70% for the primary classification and 90.95% for the secondary classification. (2) The development process of severe convective weather was divided into three stages: warning stage, outbreak stage, and post-disaster stage. All information categories peaked during the outbreak stage. During the outbreak stage, information on public facilities and power and communication was the most prevalent, while during the post-disaster stage, discussions about casualties were the most frequent. (3) The spatial distribution of disaster information quantity was generally consistent with the regions severely affected by the disaster. During severe convective weather events, public facilities faced a higher risk of exposure with damage being the most common. [Conclusion] The extraction model of disaster information based on Weibo for severe convective weather can effectively extract implicit disaster information in Weibo texts, reflect the variation characteristics of disaster events and the focus of public opinion, and provide valuable reference for disaster monitoring and early warning services as well as emergency response command.

Keywords: severe convective weather; disaster information; social media; Weibo; information extraction; topic model; text classification; rainfall

0　引　言

近年来，极端天气频发，带来的灾害损失极其严重。当前全媒体环境下，灾害发生时，公众通过社交媒体平台实时传递灾情[1]、表达情绪[2-3]、发表观点评论[4]、分配资源[5-6]等，自动参与线上线下灾害信息传播，为受灾者提供支持[7]。目前，国际各界已经认识到社交媒体数据在灾害管理中的应用价值，要求加大利用社交媒体、传统媒体、大数据等为灾害风险传播行动提供支持[8]。国内外学者对社交媒体数据在自然灾害领域的应用研究越来越多。研究表明，社交媒体数据由于其海量、实时、自带时空属性的优势，能够用于灾害事件的监测预测[9-10]。

作为国内主流社交媒体平台，微博具有互动性强、数据量大、传播及时、内容丰富等特点，已经成为了解民意、分析舆情和收集灾情的重要来源。目前，利用微博监测预测突发事件、提取灾情信息的热门研究领域包括洪涝区域识别[11]、洪水风险管理[12]、干旱风险管理[13]、火灾蔓延[14]、地震震情评估[15-16]、台风灾情评估[17]等，针对气象灾害方面的研究较少。

主题识别和文本分类是通过社交媒体平台提取灾情信息的常用方法。苏凯等[18]采用短文本主题模型(BTM)和隐含狄利克雷分布主题模型(LDA)对 2013 年台风“海燕”相关推文进行灾害主题聚类，分析了台风过程中物资医疗需求的空间分布，对比了不同模型的精度和优势；HUANG 等[19]将台风“天鸽”文本划分为 9 个主题，结合微博和腾讯位置数据，分析了公众行为和社会响应特征；XIAO 等[20]提出了基于多用户、多阶段、多元素组合的台风灾害信息主题分类方法，使分类更精细。CHEN 等[21]通过多类分类器对台风文本进行分类，识别并量化了台风灾害的物资损坏信息，为评估社交媒体文本中的台风损伤程度提供了新方法；谢雪苗等[22]运用 LDA 主题模型分析了 2023 年台风“杜苏芮”相关微博数据，分析了灾害不同阶段应急响应和公众需求的变化，证实了社交媒体在灾害管理中的应用潜力和重要价值。梁春阳等[23]基于 LDA 与 SVM 主题分类模型建立了“莫兰蒂”台风的精细化灾情数据库，提出了基于签到点用户活跃度的加权模型，但未检验分类结果的可靠性；杨辰等[24]对 110 报警的气象灾情信息采用 LDA

模型进行主题聚类，识别出 4 类气象灾种，证实了 110 报警数据在灾情分析研究中的价值。王艳东等[25]采用主题分类模型，将北京暴雨相关微博分为交通状况、天气预报、灾情、损失影响等几类主题，为快速提取、定位应急信息提供了思路。黄晶等[26]构建了基于 LDA 和 SVM 的暴雨灾情信息挖掘模型，从 2019 年“4·11 深圳暴雨”数据中提取了 6 种灾情信息：交通影响、人员伤亡、积水、停电、停水和建筑物倒塌，为暴雨灾害应急服务提供了思路，但灾情信息二次分类准确率(82.7%)仍需进一步提高。

以上研究识别了社交媒体数据中的灾害主题，并对灾情信息进行了提取和分类，对相关研究具有重要借鉴意义。但已有研究主要针对台风、暴雨等单一灾害。强对流天气是指伴随雷暴现象的对流性大风、冰雹、短时强降水等，时常多种现象同时发生，形成灾害叠加效应，具有发生突然、天气剧烈、破坏力极大的特点。目前，结合强对流天气的灾害信息提取研究比较少。灾情主题识别一般使用传统的 LDA 主题模型，该模型的应用仍存在改进空间。一方面，语料库是主题模型识别提取信息的重要依据，其质量和数量将直接影响主题分类模型的性能[27]。同样的技术方法，如果使用不同语料库，得到的分类结果可能存在差异。以上研究均未呈现语料库的构建过程和最终样本，不便于后续研究直接借鉴和引用；另一方面，传统的 LDA 主题模型假设词项之间相互独立，忽略了词与词之间的语义关系，导致主题质量下降，影响模型效果[27]。为解决上述问题，本文将结合强对流天气的特征和影响，将客观数据和气象领域知识相结合，构建丰富且精细的强对流天气灾害专用语料库，为后续此类研究提供参考。同时，采用词向量(Word2Vec)将词的语义信息融入主题模型，改进传统 LDA 主题模型的性能。构建基于 LDA 主题模型和 SVM 分类算法的强对流天气灾情信息提取模型，提取灾情时空分布特征，并创新结合气象灾害发生发展的特点和舆情演化规律，分析不同阶段、不同地区的灾情影响和用户关注焦点，为强对流天气的监测服务和应急管理提供信息支持。

1　研究资料和方法

1.1　灾害事件选择

受东北冷涡影响，2021 年 4 月 29 日和 30 日江苏沿江及以北大部分地区遭受大风、冰雹等强对流天气袭击。据江苏省气象局官方数据统计，4 月 30 日全省 13 个市的 630 个乡镇(街道)(占全省 50.4%)日极大风达到 8 级(17.2 m/s)以上，153 个乡镇(街道)(占全省 10.2%)日极大风达到 10 级(24.5 m/s)以上。图 1 为 2021 年 4 月 30 日 05 时—5 月 1 日 05 时江苏省日极大风实况空间分布，前三位分别是南通通州湾 15 级(47.9 m/s)、南通通州三余镇 14 级(45.4 m/s)、海门包场镇东灶港 12 级(39 m/s)。徐州、宿迁、连云港、淮安、盐城、扬州、泰州、南通和常州 9 个市的 23 个乡镇(街道)出现冰雹，最大冰雹直径 3~5 cm。受其影响全省 27 358 人受灾，17 人死亡，转移安置 3 138 人，农作物受灾面积 10 984 hm^2，房屋倒损 12 977 间，直接经济损失高达 1.64 亿元[28]，受到社会舆论高度关注。

1.2　数据获取与预处理

利用新浪舆情平台，以“江苏+大风或冰雹或雷暴或强对流”为关键词，提取 2021 年“4·30”强风雹天气的微博平台数据。信息地域来源设置为江苏，数据时段为 2021 年 4 月 30 日 0 点—2021 年 5 月 1 日 13 点，数据文本总量 59 523 条。人工剔除转发微博和重复、无意义文本，保留 16 334 条原创微博作为语料分析和模型构建的基础数据，数据格式如表 1 所列。其中，4 255 条数据不含地理信息，其他 12 079 条数据含明确的城市地域信息。数据预处理时，删除了标点符号、表情符号、非中文字符、网址链接和停用词等，对文本进行了有效清洗和标准化处理。

1.3　技术方法

1.3.1　中文分词

中文分词是文本挖掘研究的基础。将一段中文句子分成一个个独立的词语，有助于电脑自动识别语句含义。考虑到灾害天气文本包含各类气象预警信息，预警信息文本具有一定的领域专业性，通用分词系统无法满足所需效果，因此，本文分词时引入 Word2Vec 词向量，捕捉词语之间的语义关系，并引用 FCWS_WI 领域纠正器模型[29]，构建预警领域知识的权重哈希双字词典，并利用已有合法预警文本构建人工分词语料，提高预警领域适应性。该算法对输入的字符串标注粗分结果的四词位标签，将字符串和粗分标签输入双向 GRU-CRF 纠正器。计算得到纠正序列，并将粗分四词位标签(B，M，E，S)转化为二词位标签(B，N)，根据纠正序列的转换规则输出分词结果。

1.3.2　LDA 主题模型

隐含狄利克雷分布，简称 LDA(Latent Dirichlet Allocation)，是 BLEI 等[30]提出的一种无监督学习的主题概率生成模型，主要用于文本挖掘相关领域[31]，

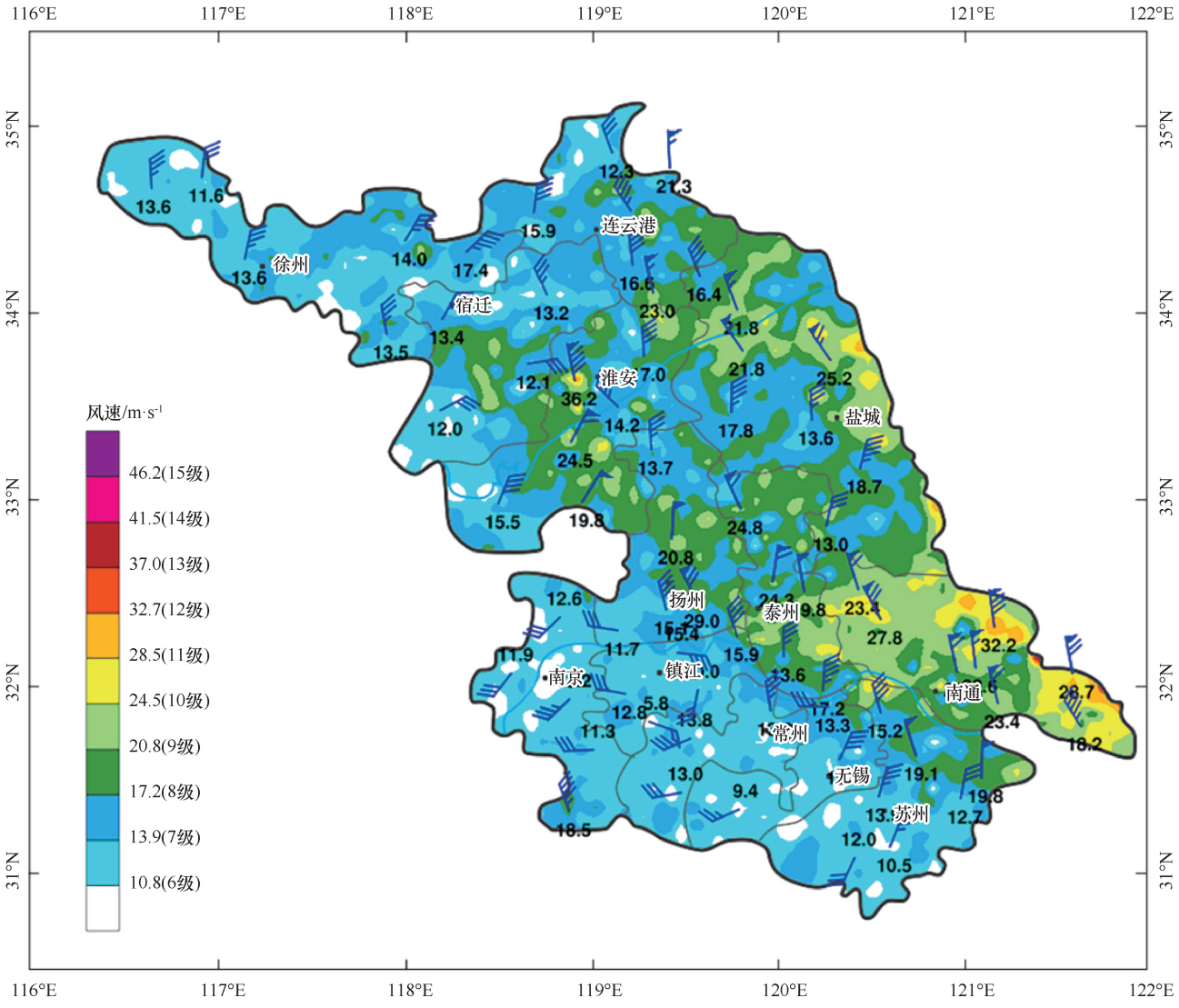

图 1　2021 年 4 月 30 日江苏省日极大风实况分布

Fig. 1　Distribution of daily maximum wind speeds observed in Jiangsu Province on April 30, 2021

表 1　微博文本数据格式示例

Table 1　Example of Weibo text data format

编　号	类　别	内　容
1	原文链接	http://weibo.com/19958352188/...
2	微博内容	刚刚第一次在淮安见到鹌鹑蛋一样大的冰雹
3	来源网站	新浪微博
4	作　者	@用户名称
5	发布日期	××××年××月××日××时××分
6	认证类型	普通用户/个人认证-橙 V/机构认证-政府等
7	信源地域	江苏
8	行业标签	政务/公安/气象/生活/娱乐/其他/旅游等
9	精准地域	江苏、淮安
10	涉及关键词	冰雹

可以识别大量文档集或语料库中隐含的主题信息[32-33]。其核心观点是每个文本由多个主题以多项式分布构成，每个主题由多个单词以多项式分布构成，多项式分布的先验概率分布为狄利克雷分布，按该方法可以生成含有多个文本的数据集[18-26]。该模型训练数据时需输入用户指定的主题个数和语料库，输出以概率分布形式呈现的主题，并人工归纳合并相似主题[26]。本文通过 Matlab 语言，采用关键词识别与 LDA 模型相结合的方法实现主题挖掘。

1.3.3　SVM 分类算法

支持向量机(SVM)是一种广义线性分类器[34]，用监督式机器学习算法对数据进行分类。采用 SVM 对标注好的文档样本集进行训练。训练样本分为 N 份，其中 $N-1$ 份为模型训练样本，剩余 1 份为检验样本，用于验证 $N-1$ 份数据分类结果的精准度[26]。

1.3.4　信息提取模型的性能评价

K 折交叉验证(K-Fold Cross-Validation)是机器学习中一种用于评估模型性能的技术，核心思想是将

数据随机分成多个大小相等的子集，轮流使用不同子集进行训练和验证，确保模型在不同数据集上均表现良好[35-36]。一般 K 值取 5 或 10。将 K 次测试结果的性能指标进行平均作为模型最终评估结果[37]。信息提取模型的性能评价指标一般采用准确率、精准率、召回率、F 值（F1-Measure）。准确率是用被分类正确的样本数除以所有样本数。准确率越高，则分类器越好，该指标最直观。当精准率和召回率出现矛盾时，采用综合评价指标 F 值。F 值是精准率和召回率的加权调和平均。

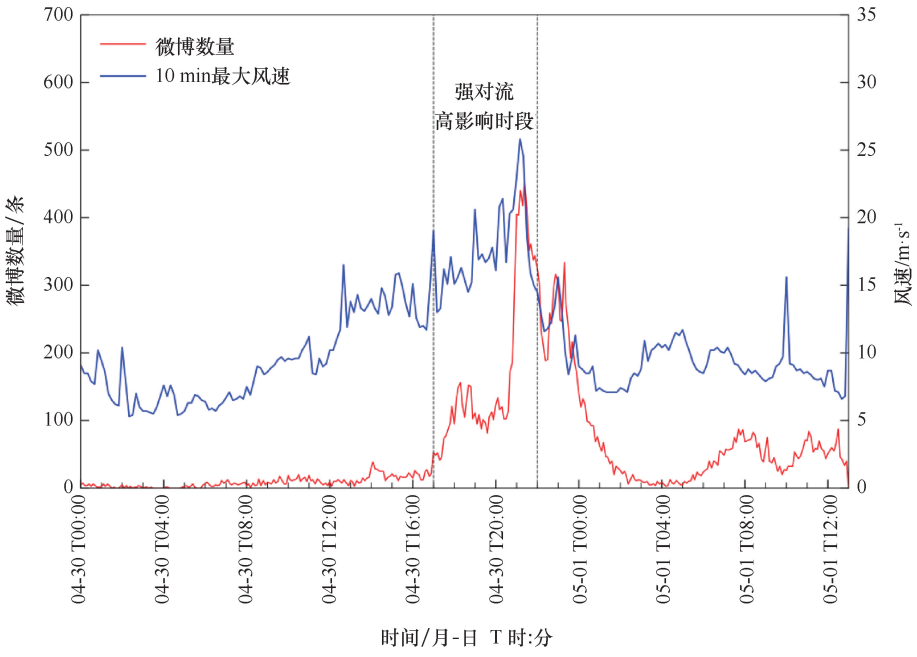

图 2　2021 年“4·30”强风雹天气微博信息数量及全省逐 10 min 最大风速变化

Fig. 2　Variations in Weibo post numbers and province-wide maximum wind speeds every 10 minutes during severe wind-hail event on April 30, 2021

2　强对流天气灾情信息提取模型构建

构建强对流灾情信息提取模型前，统计网络文本中强对流天气的句法表达特征、高频词汇、关键词共现关系等，为语料库提供客观依据。

2.1　强对流天气灾害信息统计分析

2.1.1　时间分布特征

4 月 30 日傍晚到上半夜是此次强对流天气的高影响时段，8 级以上区域性大风在江苏历时 5 h（17—22 时），其中东南部地区出现大范围 12 级以上大风[38]。4 月 30 日江苏全省日极大风速的前十名大部分出现在 21—22 时。图 2 为“4·30”强风雹天气过程微博信息数量及全省逐 10 min 最大风速变化。由图 2 可以看出，4 月 30 日白天微博信息较少，17 时以后大风冰雹的舆论热度迅速攀升，20—21 时前后达到峰值，与当日风力最强时段基本一致；0 时以后，本轮天气过程结束，夜间用户活跃度逐渐下降，微博信息量明显减少；5 月 1 日 6 时以后出现了第二个舆论高峰，信息主要以灾情相关的新闻报道和天气回顾为主。这与杨永清等[39]的研究结果一致，气象灾害舆情一般在灾害发生后短时间内快速达到峰值，且具有明显的长尾效应。灾害过后，媒体和公众对事件仍会持续关注。微博数据可以体现社会公众对强对流灾害事件的关注度和舆论热度。

2.1.2　空间分布特征

带有地理空间描述的数据可以被视为地理空间中的一个点，有助于进行灾害空间分布研究[40]。图 3 为 2021 年“4·30”强风雹天气过程微博信息所属地区的空间分布。南通地区信息量最多，达到 3 989 条，其他城市信息量均不足 1 000 条，其中，苏州、南京相对较多，淮安、镇江信息量最少。气象监测数据显示，本次灾害过程风力最大、影响最严重的地区为南通，因此，南通地区舆论热度最高。气象灾害舆情信息的扩散具有以受灾地区为核心的远距

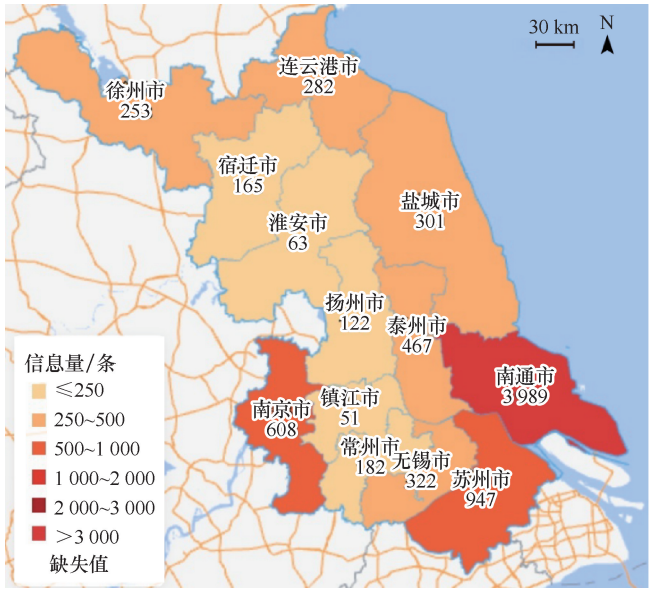

图 3　2021 年“4·30”强风雹天气过程微博信息空间分布

Fig. 3　Spatial distribution of Weibo posts during severe wind-hail event on April 30, 2021

离衰减规律[39]，邻近地区的用户对气象灾害感知明显，信息关注度和传播意愿更强。因此，毗邻南通的苏州、泰州地区受大风影响，也受到舆论关注。微博信息量与灾害影响区域具有较好的一致性。

2.1.3　高频词汇特征

对“4·30”强风雹天气过程微博文本进行分词处理，得到 205 485 个词语。提取前 200 个高频词汇，以词云图的形式展现，结果如图 4 所示。图中每个词语的大小与其出现的频率或次数成正比。大风、南通、14 级、冰雹等词汇频率最高，其次是城市名称、时间描述及气象术语。灾害影响相关词语包括窗户、阳台、玻璃、热水器、太阳能、广告牌、路灯、车辆、高速等承灾体，以及断电、停水、抢修、受伤等灾损信息。公众情绪和情感相关词汇较多，比如吓人、可怕、恐怖、瑟瑟发抖，体现了微博作为社交媒体成为公众在突发事件中情绪表达和意见交换的重要平台。

2.1.4　关键词共现关系特征

一段文本中的关键词存在某种关联关系，这种关联可以用共现频次来表示。一对词汇在同一文本中出现的频次越高，说明这两个主题关系越密切。图 5 为 2021 年“4·30”强风雹天气关键词共现关系图谱。网络节点之间的连线表示两节点间的共现关系强度。线条越粗，颜色越深，表明共现频率越高。节点的大小表示该节点在网络中的重要程度，节点越大，表明重要程度越大。由图 5 可以看出，特征词以“大风”为中心，关联南通、冰雹、14 级等多个主题词，形成了一个完整的灾害舆论体系。包含了公共设施、车辆交通、电力通信、人员等多方面信息。“断电”与“大风”“太阳能”“空调外机”“广告牌”“热水器”等词语存在共现关系，说明大风天气可能造成电力设备受损，导致停电，影响家用电器使用；同时，空调外机、太阳能热水器、广告牌等存在被大风吹落的风险，造成安全事故。“航班”“飞机”“高铁站”“高速”“堵车”等词语，体现了大风天气对交通的影响。词语共现关系可以反映致灾因子与承灾体之间的关系。

2.2　强对流天气灾害语料库构建

语料是任何自然语言研究的基础。语料库是词向量训练的基础，是指与主题相关的大量文本数据的集合[40]。通用语料库一般针对常用词语建立，具有一定普适性，但不包含气象领域知识及特殊词汇。随着新媒体的发展和网络用语的流行，网络用户对灾害事件的语义表达趋向多元化。基于 2021 年“4·30”强风雹天气过程数据的高频词汇及关键词共现关系，

图 4　2021 年“4·30”强风雹天气微博高频词汇词云图

Fig. 4　Word cloud of high-frequency terms in Weibo posts during severe wind-hail event on April 30, 2021

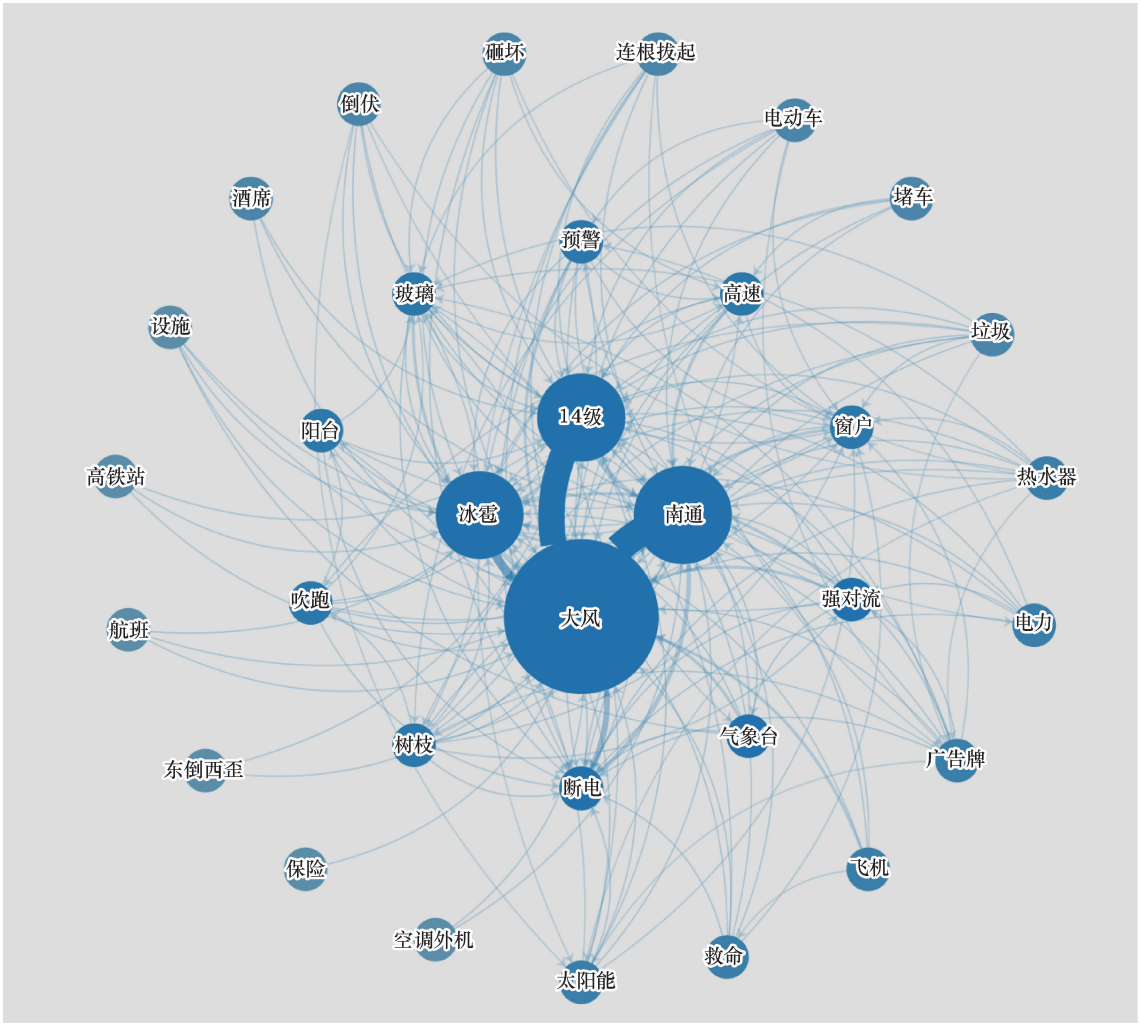

图 5　2021 年“4·30”强风雹天气微博关键词共现关系图谱

Fig. 5　Co-occurrence network of keywords in Weibo posts during severe wind-hail event on April 30, 2021

结合强对流天气灾害领域知识，对大风、冰雹、短时强降水灾害特征词汇进行梳理，采用归纳法，构建了基于社交媒体平台的强对流天气灾害信息主题语料库，如表 2 所列。该语料库是文本主题分类的重要依据，包含了天气状况、科普防御、灾情影响、求助救援及其他信息五个大类，其中，灾情影响信息包含公共设施、车辆交通、电力通信、人员伤亡、农业设施及其他 6 个方面。该语料库不仅考虑了强对流灾害全生命周期相关术语，还融合了气象领域专家知识及网络用语，比如大风、冰雹、龙卷等灾害的方言表达，公众情绪状态的网络语境表述，进一步提高了语料库的质量和可用性，能够为其他强对流天气相关研究提供参考。

2.3　强对流天气灾情信息提取模型

结合 LDA 主题模型和 SVM 分类算法，构建强对流天气灾情信息提取模型的思路如下：首先，用 Word2vec 模型对文本信息进行分词，并用 FCWS_WI 领域纠正器模型进行预警领域词义纠正[29]，获得词向量，进而构建文本向量集合。将社交媒体的文本数据矩阵化并提取文本特征是强对流灾害天气信息主题提取的关键步骤。本文通过以下流程实现：一是充分考虑气象预警信息的领域特性，利用具有领域特征的语料库，将文本信息转化为词向量。二是以词向量为基础，进一步考虑上下文语境，实现更加有效的文本向量提取。然后，引入 LDA 主题模型计算文本向量离散词语共现频率，提取文本向量集合所表达的相关主题，并输出对应主题的词汇分布[26]。最后，对已获取的主题采用 SVM 分类算法训练，构建灾害信息提取模型。当出现新文本时，通过该模型判断该文本的主题类别并分类。根据分类提取结果携带的时间空间信息，制作强对流灾情信息的时间、空间分布图，为灾害天气监测、预报、服务和应急管理提供

表 2　基于社交媒体平台的强对流天气灾害信息主题语料库

Table 2　Disaster information thematic corpus for severe convective weather based on social media platform

编　号	主题类别	主题特征语料
1	天气状况	气象台、发布、天气、预报、预警、警报、消息、信号、早间、午间、晚间、今日、今天、未来、短期、短时、短临、三天、趋势、临近、24 小时、12 小时、6 小时、2 小时、重要、报告、预计、温度、最高、最低、气温、风力、中等、局部、雷暴、雷电、冰雹、严重、雹灾、大风、蓝色、黄色、橙色、红色、最大、极大、瞬时、阵风、平均、偏北、偏东、偏南、偏西、西北、西南、东北、东南、海区、陆上、强对流、飑线、阵风锋、下击、暴流、龙卷、龙卷风、下雨、暴雨、大雨、中雨、小雨、雷阵雨、气旋、冷涡、卫星、云图、雷达、回波、实况、数据、现在、刚刚
2	科普防御	大风、冰雹、暴雨、来袭、风力、影响、科普、知识、防御、收下、收好、指南、应急、攻略、准备、防范、防止、怎么、政府、部署、安置、检查、措施、提醒、加固、加强、减少、注意、回港、关好、不要、玻璃、启动、响应
3	灾情影响	第一次、头一次、出现、冰雹、雹子、下冰雹、电闪雷鸣、闪电、风雨交加、狂风暴雨、疾风骤雨、黑云、忽闪、打雷、打闪、天雷、闷雷、落汤鸡、漏风、起风、刮风、大风、妖风、狂风、强风、台风、龙卷、暴风、风灾、灾难、灾害、雹灾、从天而降、威力、遭遇、袭击、伤亡、威胁、生命、死人、受伤、转移、撤离、人员、万人、余人、井盖、刮倒、掀翻、遭殃、摧残、损坏、毁坏、摧毁、树木、树枝、折断、折倒、倒伏、农业、大棚、顶棚、雨棚、设施、高空、广告牌、空调外机、太阳能、热水器、屋檐、房屋、瓦片、仓库、棚顶、楼顶、吊顶、天花板、阳台、窗户、花盆、户外、坠落、掉落、东倒西歪、砸伤、砸坏、砸烂、击穿、疼、吹走、吹跑、掀起、掀飞、吹翻、席卷、玻璃、碎片、垃圾、灯、路灯、电灯、电缆、电压、高压、电线、断电、停电、没电、电力、抢修、跳闸、通信、线路、交通、路途、封路、中断、拥堵、堵塞、堵车、骑车、开车、高速、大桥、晚点、延误、航班、飞机、起飞、停运、恢复、退票、车票、改签、取消、公交、地铁、电动车、故障、游轮、游船、渔船、划船、轮渡、受淹、淹没、世界末日、终于、来了、噼里啪啦、恶劣、糟糕、崩溃、头疼、天昏地暗、东摇西晃、惨烈、厘米、毫米、直径、半径、鸡蛋、硬币、花生、玉米、乒乓球、鸽子蛋、黄豆、荔枝、汤圆、石子、大小、大冰雹、小冰雹、超大、很大、太猛、冰蛋子、盐粒、米粒、豆子、豆粒、一颗、大颗、弹珠、水深、深度、大片、吓死、吓人、可怕、恐怖、躲雨、猛烈、飞升、体重、抗风、劫难、渡劫、经历、连根拔起、逃生、晃动、摇晃、塌了、碎、渣、瑟瑟发抖、蒙圈、刺激、死神、发誓、雷劈、哐当、吃席、酒席、横幅、高架、脚手架、索命、要命、送走、跑、嘭、淋雨、理赔、赔偿、保险、车险、车窗、被困、回家、声音、挡风、漆黑、屋顶、窟窿、外面、随手、记录、奇形怪状、飞沙走石、图片、视频、头盔、龙门、工地、景点、景区、游客、疏散、关闭、农田、水稻、小麦、蔬菜、瓜果、暴击、洗礼、秧苗、鞭炮、街道、位置、砸晕、晕倒、围栏、横祸、油菜、枯枝败叶、狼藉、主干道、天灾、人祸、震碎、清障队、凌乱、绿化、清洁、环卫、工人、受害者、变形、满天飞、暴扣、群众、头破血流、急诊、红绿灯、意外、事故、医院、救护、防疫、途中、安全
4	求助救援	帮忙、帮助、求救、求助、救命、危险、救出、抢救、急救、急需、紧急、需要、赶紧、立刻、马上、120、医院
5	其　他	爱情、明星、艺人、演员、偶像、友谊、游戏、网游、手游、电影、电视剧、专辑、歌词、歌曲、诗词、古文、古风、美容、植发、晕染、整形、整容、暴风雪、风水、发际线、钱、文艺、风景、名胜、心灵、鸡汤、安宁、甜蜜、温情、祈祷、盼望、希望、祝愿、祈求、感谢、感动、平安、点赞、辛苦、回忆、想起、据说、听说、讨厌、喜欢

信息支撑。灾情信息提取流程如图 6 所示。

3　仿真试验与结果分析

3.1　灾害信息类别划分

提取灾情信息时需要对灾害相关的文本进行主题分类。图 7 为本文构建的灾害信息类别划分标准。对社交媒体数据一次分类时分为“天气状况”“科普防御”“灾情影响”“求助救援”“其他”5 大类主题。天气状况信息包含天气预报、灾害预警及实况数据等，主要由气象部门或气象爱好者、新闻媒体发布；科普防御信息包括气象科普、灾害提醒、防御准备等信息，一般由气象、应急、媒体和科普账号发布。灾情影响信息包括灾害对公共设施、交通、农业、电力、人员等各方面造成的影响；求助救援信息指被困受灾现场的人们的呼吁、求助信息，救援队伍的救灾情况进展等；其他信息指不属于以上几类的信息，以及含灾害相关词语，但并非真实强对流天气的信息，比如与灾害相关的诗词、歌词、广告、游戏等特定语句。

根据灾害影响的承灾体将一次分类提取的灾害影响信息进行二次分类，划分为“公共设施”“车辆交通”“电力通信”“农业设施”“人员伤亡”“其他”6 个小类。公共设施相关灾害主要指树木绿植和门窗围栏受损、景区工地事故、广告牌、空调外机和太阳能热水器坠落等；车辆交通类灾情影响包括水陆空交通延误取消、交通拥堵、道路积水、车辆被冰雹砸坏等；电力通信灾情主要指电线杆倒伏、电力设备受损、网络信号变差、断电、停电等；农业设施灾情信息包括田园果蔬和畜牧家禽受损、设施大棚坍塌、果实被冰雹砸伤坠落、作物减产等；人员伤亡类信息包括人员被困情况、伤亡情况和伤亡数据等。

3.2　灾害信息分类结果

结合 LDA 主题模型和 SVM 分类算法构建强对流天气灾情信息提取模型时，选取 90% 数据进行模型训练，其余 10% 作为测试数据集。通过多次对比试验，初始主题设为 25 个效果较好。LDA 主题模型是无监督学习算法，因此，人工对 25 个主题进行归纳合并[26]，最终保留 5 个主题，如表 3 所列。其中灾

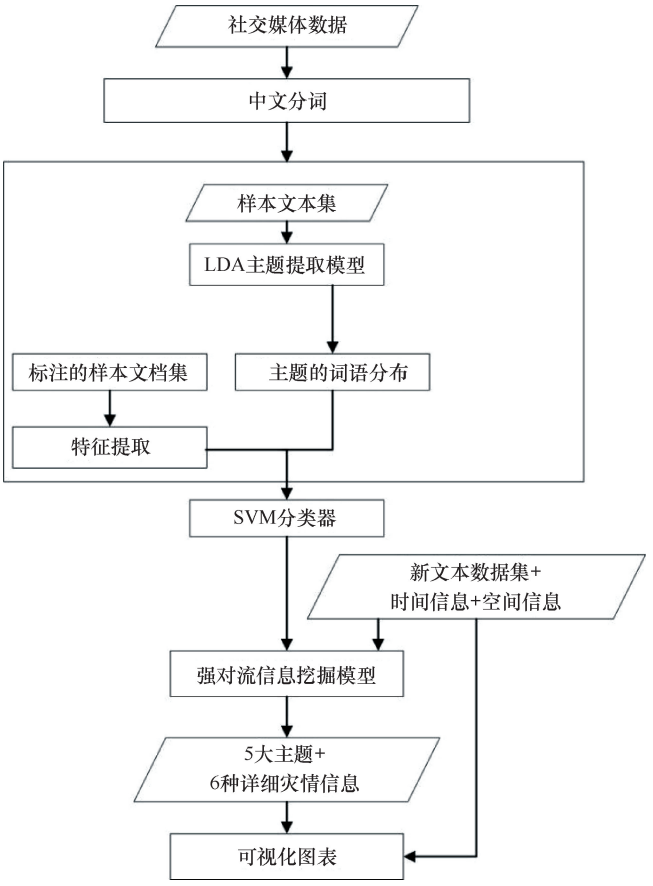

图 6　基于社交媒体数据的强对流灾情信息提取流程

Fig. 6　Disaster information extraction flowchart of severe convective weather based on social media data

表 3　训练数据：一次分类信息统计

Table 3　Training data: statistics of primary classification information

编　号	主　题	微博数量/条	所占比例/%
1	天气状况	4121	28. 03
2	科普防御	109	0. 74
3	灾情影响	7352	50. 01
4	求助救援	121	0. 82
5	其　他	2 998	20. 39
合　计		14 701	100. 00

表 4　测试数据：一次分类主题提取结果

Table 4　Test data: topic extraction results of primary classification

编　号	主　题	微博数量/条	所占比例/%
1	天气状况	68	4. 14
2	科普防御	4	0. 24
3	灾情影响	1 425	86. 84
4	求助救援	12	0. 73
5	其　他	132	8. 04
合　计		1 641	100. 00

情影响信息 7 352 条，占比 50. 01%；天气状况信息 4 121 条，占比 28. 03%；科普防御、求助救援和其他信息相对较少。表 4 为测试数据一次分类结果。经测试数据验证，模型对强对流灾害信息一次分类的准确率为 93. 4%。

二次分类时，将一次分类提取的 7 352 条“灾情影响”信息分为 6 种具体灾情，结果如表 5 所列。

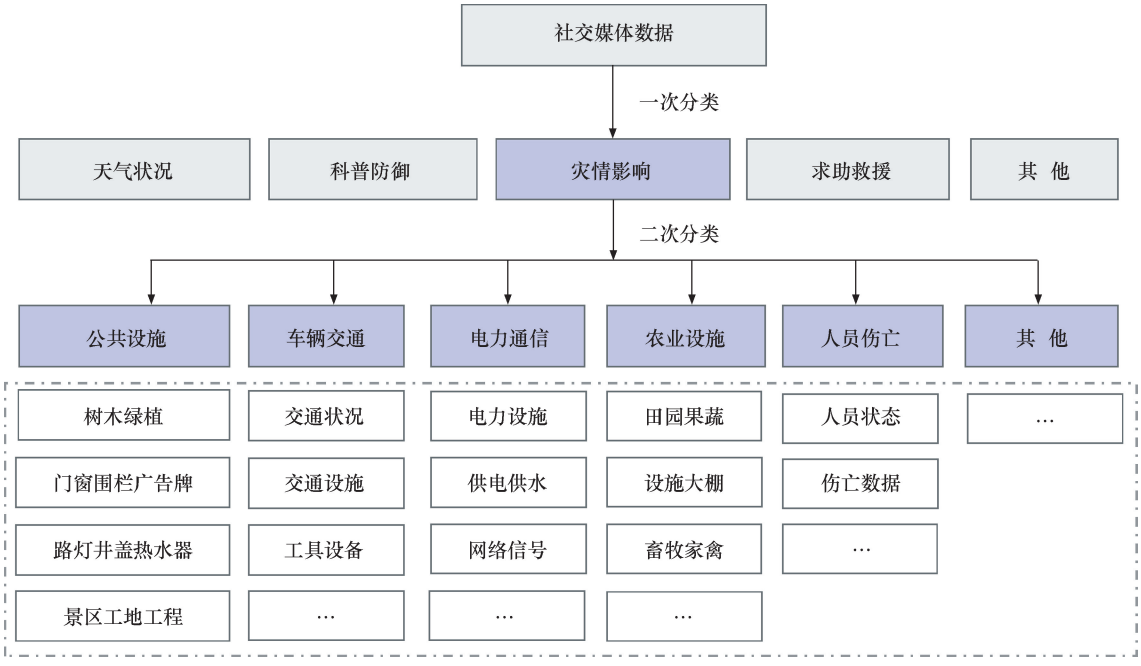

图 7　灾害信息提取流程及类别划分

Fig. 7　Disaster information extraction flowchart and category classification

表 5　训练数据：二次分类 6 种具体灾情统计

Table 5　Training data: statistics of six specific disaster information categories in secondary classification

编　号	具体灾情信息	微博数量/条	所占比例/%
3-1	公共设施	1 838	25.00
3-2	车辆交通	1 007	13.70
3-3	人员伤亡	859	11.68
3-4	电力通信	1 133	15.41
3-5	农业设施	32	0.44
3-6	其　他	2 483	33.77
合　计		7 352	100.00

表 6　测试数据：二次分类 6 种具体灾情提取结果

Table 6　Test data: extraction results of six specific disaster information categories in secondary classification

编　号	具体灾情信息	微博数量/条	所占比例/%
3-1	公共设施	417	29.26
3-2	车辆交通	65	4.56
3-3	人员伤亡	48	3.37
3-4	电力通信	210	14.74
3-5	农业设施	3	0.21
3-6	其　他	682	47.86
合　计		1 425	100.00

“公共设施”类灾情最为普遍，信息量占 25%。电力通信信息 1 133 条，占比 15.41%；提及“车辆交通”和“人员伤亡”相关信息的微博分别占 13.70% 和 11.86% 。农业设施类灾情信息仅占 0.44%。表 6 为测试数据提取的 6 种具体灾情信息，经验证，构建的强对流灾情信息提取模型对灾情影响信息二次分类的准确率为 95%。

为进一步验证模型的稳定性，采用 K 折交叉验证，K 值选为 10。每次按照 9∶1 划分训练集和测试集。一次分类交叉验证结果如表 7 所列。模型平均准确率 92.70%，精准率 96.06%，召回率 95.11%。二次分类交叉验证结果如表 8 所列，平均准确率 90.95%，精准率 92.96%，召回率 93.70%。由此可见，构建的强对流灾情信息提取模型的分类效果较好，相比同类研究，分类效果显著提升。这主要源于本文构建的语料库融合了气象领域知识和网络用语，并将词的语义信息融入主题模型，提高了模型的准确性。

3.3　灾情信息提取结果应用与分析

3.3.1　时间趋势

一次分类结果中，除“其他信息”外的 4 类主题微博数量随时间的变化趋势如图 8 所示。结合气象灾害舆情的特点和本次天气过程的起止时间，将强对流天气发生发展过程划分为预警期、突发期和灾后期三个阶段[41]。总体来看，各类信息均在突发期(4 月 30 日 17 时—5 月 1 日 00 时)处于高值区。预警期内“天气状况”微博数量最多，但关注度较低。灾害突发后，天气状况信息的关注度明显增加。4 月 30 日 18 时以后，强风雹天气对公共设施、电力通信、车辆交通等造成不利影响，各类灾情逐渐显现，“灾情影响”信息数量超过“天气状况”信息，受到广泛关注和讨论。“灾情影响”微博的时间趋势与图 2 的总体时间趋势相似，主要集中在强风雹突发期间，并在 4 月 30 日 21 时前后达到峰值。“救援信息”主要集中在强风雹发生过程中。科普防御信息占比较低，主要由于这类信息的发布主体较少，一般以气象、应急系统账号为主。灾害发生时，媒体和公众对灾情影响的关注度更高。

“灾情影响”二次分类后的 5 种灾情信息时间趋势如图 9 所示。农业设施类信息普遍较少。灾害突发期内，公共设施、电力通信、车辆交通、人员伤亡等各类灾情信息在短时间内迅速攀升，体现了强风雹天气的突发性和高致灾性，尤其公共设施、电力通信受损情况较多。灾后期，5 月 1 日 10 时前后，“@南通发布”微博公布了此次灾害的应急救援和人员伤亡

表 7　一次分类信息 K 折交叉验证结果

Table 7　K-fold cross-validation results of primary classification information

交叉验证	准确率/%	精准率/%	召回率/%	$F1$-Measure/%
Fold-1	93.40	92.83	98.10	95.39
Fold-2	92.50	93.42	97.94	95.63
Fold-3	91.70	96.34	94.39	95.36
Fold-4	94.50	99.59	94.56	97.01
Fold-5	90.90	96.76	91.89	94.26
Fold-6	91.60	92.35	97.82	95.01
Fold-7	92.90	97.24	92.77	94.95
Fold-8	93.20	97.13	94.43	95.76
Fold-9	94.00	98.81	93.75	96.21
Fold-10	92.30	96.91	93.59	95.22
平均值	92.70	96.06	95.11	95.58

表 8　二次分类信息 K 折交叉验证结果

Table 8　K-fold cross-validation results of secondary classification information

交叉验证	准确率/%	精准率/%	召回率/%	$F1$-Measure/%
Fold-1	95.00	94.82	93.61	94.21
Fold-2	92.64	96.94	95.38	96.15
Fold-3	91.67	91.75	95.83	93.75
Fold-4	93.07	96.83	93.28	95.02
Fold-5	89.60	92.04	92.66	92.35
Fold-6	91.71	89.59	99.10	94.11
Fold-7	77.78	98.66	74.16	84.67
Fold-8	89.08	91.09	92.74	91.91
Fold-9	88.03	85.03	95.50	89.96
Fold-10	91.53	92.88	93.79	93.33
平均值	90.95	92.96	93.70	93.33

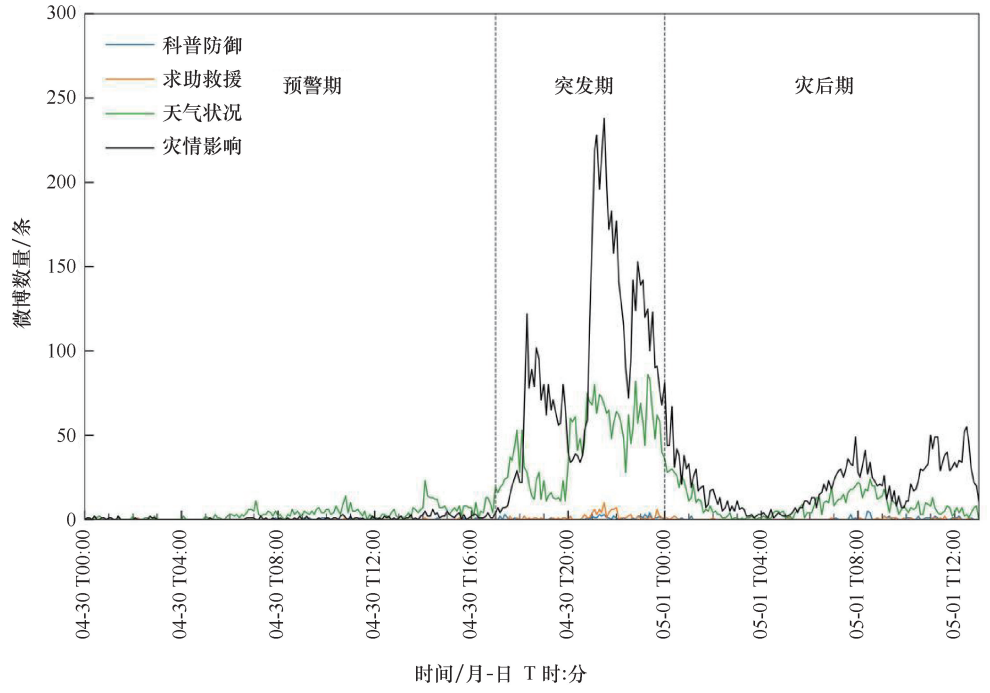

图 8　一次分类信息的时间趋势

Fig. 8　Temporal trends of primary classification information

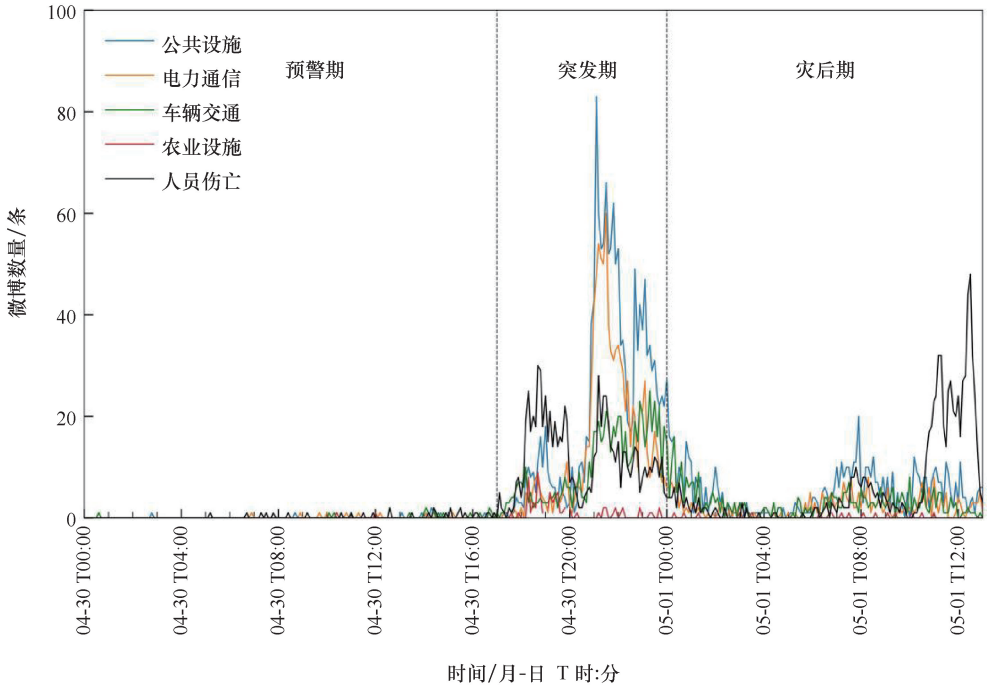

图 9　二次分类信息的时间趋势

Fig. 9　Temporal trends of secondary classification information

情况后，关于人员伤亡的讨论出现第二波高峰，其他信息逐渐减少，灾害不同阶段，用户关注焦点呈现显著变化。

3.3.2　空间分布特征

根据灾情影响信息二次分类结果，南通地区灾情信息最多，其他地区灾情信息较少，与实际情况相符。以南通地区为样本，将仿真试验中二次分类的灾情信息中提及精细地理位置的微博进行统计。图 10 为基于微博数据的南通地区灾损信息空间分布。据统计，如皋地区灾情提及频次最高，达 153 人次。启东

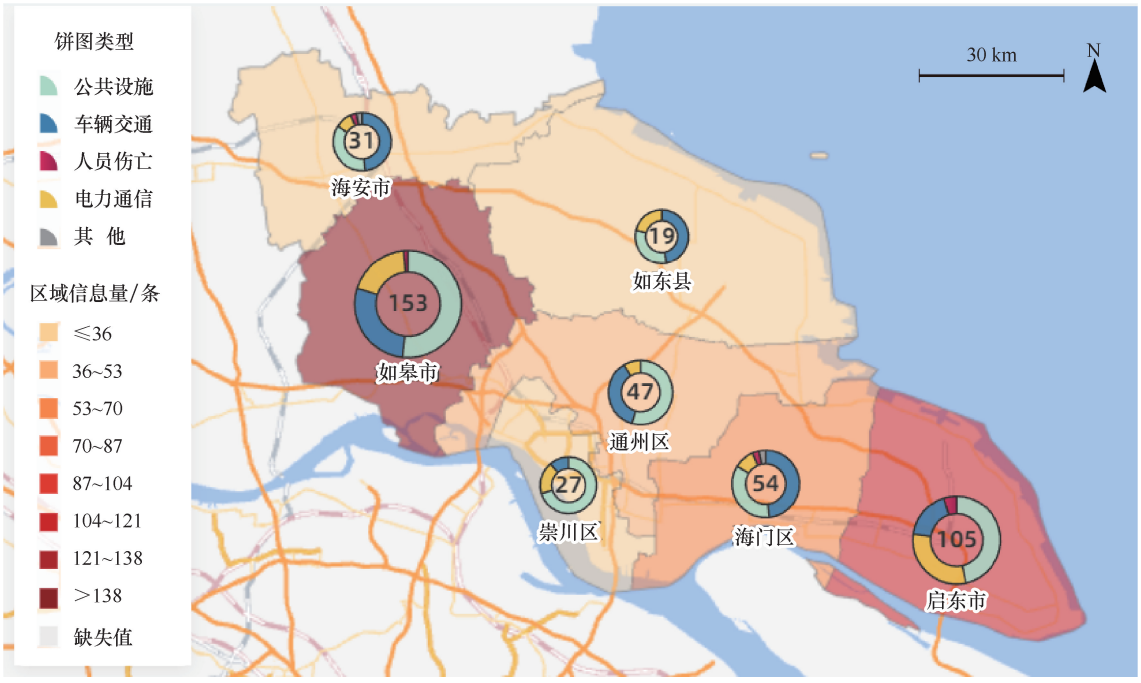

图 10　基于微博数据的南通地区灾损信息空间分布

Fig. 10　Spatial distribution of disaster damage information in Nantong based on Weibo data

105 人次，次之。其他地区灾情信息数量较少。各类灾害中，公共设施受损的信息比例最高。强风雹天气中公共设施因其暴露性而面临较高的受损风险。冰雹可能对建筑物屋顶、外墙、车辆和户外设施造成直接损害。强风可能导致建筑物损坏、树木折断倒塌，影响道路通行和电力通信。精细地理位置信息与用户隐私密切相关，灾害中分享精细地理位置的用户比例较小，分享意愿受多种因素综合影响，因此，该图不能全面反映灾情的空间分布特征，但可以作为参考信息为气象服务和应急管理人员提供灾情线索，为社交媒体数据提取结果的应用提供可视化解决方案。

据江苏省气象局官方微博“@ 江苏气象”公布数据，南通部分地区 4 月 30 日 18 时至 22 时出现冰雹和强雷暴大风天气，全市有 66 个自动站最大风力超过 10 级，风力最强时段集中在 4 月 30 日 20：00—21：00。江苏全省极大风风速前十名如表 9 所列，其中有 8 个站点位于南通，且极大风均出现在该时间段前后。4 月 30 日 20：37—21：13，南通兴东国际机场突发大风天气持续 36 min，最大阵风 33 m/s，达 12 级，为南通机场自 1992 年开航以来极大值[42]。一架停泊的飞机受强风天气影响发生移位，该事件短时间内快速传播，受到媒体和公众的广泛关注。

据南通市委宣传部官方微博“@ 南通发布”通报，截至 5 月 1 日 10 点，南通有 3 000 余人口因强风雹天气受灾，其中，11 人因大树倒伏砸倒房屋、狂风卷入河道等原因死亡；102 人因灾受伤。“苏海门渔 01728”渔船因突遇大风而倾覆，11 名船员落水。受灾严重地区多处房屋倒损，紧急转移安置 3 050 人；多处电力设施受损、树木倒伏、围墙倒塌、车辆受损、热水器掉落等。图 11 为媒体报道的部分现场灾情图片。以上灾情信息在微博高频词汇及灾情提取结果中均有体现。吴洪颜等[43]通过对江苏 2008—2022 年江苏省雷暴大风灾情数据和气象因子的分析证实，天气过程降水量较小时，风速、降水和冰雹对成灾均呈正相关。最大风速与伤亡人口、倒损房屋、直接经济损失呈显著正相关[43]。本文的研究结果

表 9　2021 年 4 月 30 日江苏全省极大风速前十名

Table 9　Top ten maximum wind speeds in Jiangsu Province on April 30, 2021

站　名	极大风速/m·s^{-1}	发生时间
南通通州湾	47.9(15 级)	30 日 20 时 48 分
南通通州湾三余镇三友	45.4(14 级)	30 日 20 时 46 分
南通海门包场镇东灶港	39.0(13 级)	30 日 20 时 47 分
南通通州环本农场	37.9(13 级)	30 日 20 时 42 分
盐城射阳射阳港	37.7(13 级)	30 日 18 时 00 分
南通启东海复镇东元滩涂	37.7(13 级)	30 日 21 时 08 分
淮安市淮安区	36.2(12 级)	30 日 17 时 40 分
南通如东太阳沙	36.1(12 级)	30 日 20 时 33 分
南通吕泗	34.3(12 级)	30 日 20 时 52 分
南通启东近海镇塘芦港	34.2(12 级)	30 日 21 时 08 分

图 11　2021 年 4·30 强风雹天气灾害现场图片(来源：微博平台)

Fig. 11　Photos of severe wind-hail event on April 30, 2021 (Source: Weibo platform)

印证了以上结论，此次灾害降水量较小，日最大风速较大是南通受灾严重的主要原因。

4　结果讨论

本研究构建的强对流天气灾害专用语料库，可以显著提高灾情信息提取和分类的准确性，为相关研究提供参考。基于 LDA 主题模型和 SVM 分类算法的强对流灾情信息提取模型对微博文本隐含的灾情信息识别与分类效果较好，为大数据背景下强对流天气灾情的社会化观测和收集提供了思路。相关部门可以依据此方法从舆情平台实时提取灾情信息，并将提取结果与灾害风险预警地图、气象台站观测数据、人口热力图等信息叠加进行综合展示，为灾害天气监测服务和应急管理提供信息参考。

研究仍然存在不足之处。一是社交媒体数据具有一定主观性以及空间分布不均衡性[1-44]，且受人口密度、通信信号等因素影响，存在部分虚假或干扰信息，需要结合多方面因素综合分析。本研究案例未出现重大谣言，但在其他案例应用时需要考虑不实信息对灾情信息提取结果的干扰，提前剔除谣言和不实信息。二是强对流天气通常会对农作物和农业生产造成显著威胁[45]，本次灾害江苏全省农作物受灾面积 10 984 hm^2。灾情影响信息二次分类时设置了农业设施类别，但仿真试验提取到农业设施类灾情信息比例非常低。这主要由于微博用户以年轻群体为主，且灾害期间社交媒体的内容生成和传播更倾向于关注大城市或人口稠密的区域[46]。农村或偏远地区人口密度小，微博用户相对偏少，社交媒体参与度低，造成微博平台对农村地区灾情信息传播缺失或滞后，农业设施类灾情信息较少。城市地区基础设施更完善、互联网普及率更高，社交媒体的使用率通常高于农村地区[47]。因此，本研究的模型和方法对人口密集、网络发达的城市地区灾情信息提取具有一定优越性，有助于及时获取局地性的灾情信息和公众反馈，不适用于农村及偏远地区。

5　结　论

本文构建了强对流天气灾害专用语料库，以及基于 LDA 主题模型和 SVM 分类算法的强对流天气灾情信息提取模型，以江苏 2021 年“4·30”强风雹天气为例，识别出 5 类主题文本和 6 类具体灾情信息。主要研究结论有以下几点。

(1)通过对“4·30”强风雹天气的数据文本进行统计分析，发现 4 月 30 日 18 点以后江苏大风冰雹的舆论热度迅速攀升，20—21 点前后达到峰值，与风力最强时段一致。强对流天气的高频词汇、词语共现关系可以体现舆论关注焦点。考虑灾害全生命周期，融合气象领域专家知识和网络用语的强对流天气灾害主题语料库，能够为灾害信息提取与分类提供有效支撑。

(2)融合语义信息的 LDA 主题模型和 SVM 分类算法构建的强对流灾情信息提取模型对微博文本中隐含的灾情信息识别与分类效果较好。以“4·30”强风雹天气微博数据进行仿真试验，一次文本主题分类提取出天气状况、科普防御、灾情影响、求助救援和其他信息 5 个主题。二次分类提取出公共设施、电力通信、车辆交通、农业设施、人员伤亡和其他 6 种具体灾情影响信息。经过 10 折交叉验证，一次分类平均准确率为 92.70%，二次分类平均准确率为 90.95%，优于同类已有研究。

(3)强对流灾情信息提取结果的时间趋势可以反映不同时期的灾情变化和用户关注焦点。各类信息均在灾害突发期处于高值区。灾害突发期，公共设施、电力通信信息关注度最高，灾后期，人员伤亡信息讨

论度最高。灾情空间分布基本能反映灾害的高影响区域，“4·30”强风雹天气南通地区灾情信息最多，公共设施受损情况最普遍。日最大风速较大是南通受灾严重的主要原因。

社交媒体信息具有多样性。本文主要进行了文本信息的提取，后续还需深入研究基于社交媒体的图片、视频等多模态灾害信息提取技术和应用场景，进一步提高灾情态势感知的准确性和全面性[48]，为气象灾害监测预警与风险管理提供支撑。同时，结合气象灾害下网络信息传播模式[49]，进一步完善强对流、台风、暴雨等各类气象灾害应急预案，加强自然灾害网络舆情综合治理，维护健康网络生态。

参考文献(References)：

[1] TANG J T, YANG S N, WANG W P. Social media-based disaster research: Development, trends, and obstacles [J]. International Journal of Disaster Risk Reduction, 2021, 55: 102095.

[2] YANG T F, XIE J B, LI G Q, et al. Social media big data mining and spatio-temporal analysis on public emotions for disaster mitigation [J]. International Journal of Geo-Information, 2019, 8(1): 29.

[3] WU K J, WU J D, LI Y. Mining typhoon victim information based on multi-source data fusion using social media data in China: A case study of the 2019 Super Typhoon Lekima [J]. Geomatics, Natural Hazards and Risk, 2022, 13(1): 1087-1105.

[4] AHMED Y A, AHMAD M N, AHMAD N, et al. Social media for knowledge-sharing: A systematic literature review [J]. Telematics and Informatics, 2019, 37: 72-112.

[5] WANG Z Y, YE X Y, TSOU M H. Spatial, temporal, and content analysis of Twitter for wildfire hazards[J]. Natural Hazards, 2016, 83(1): 523-540.

[6] 周义棋，田向亮，钟茂华. 基于微博数据的自然灾害应急救助需求评估[J]. 清华大学学报(自然科学版)，2022，62(10)：1626-1635.
ZHOU Yiqi, TIAN Xiangliang, ZHONG Maohua. Assessment of natural disaster emergency relief based on Microblog data, China[J]. Journal of Tsinghua University(Science and Technology), 2022, 62(10): 1626-1635.

[7] 张谱，张豪，孔锋，等. 基于微博数据的暴雨洪涝灾害舆情特征研究：以2021年中国三场暴雨洪涝为例[J]. 水利水电技术(中英文)，2023，54(2)：47-59.
ZHANG Pu, ZHANG Hao, KONG Feng, et al. A study on public opinion characteristics of rainstorm flooding disasters based on Sina Weibo data: Take the three rainstorm flooding disasters in China in 2021 as an example [J]. Water Resources and Hydropower Engineering, 2023, 54(2): 47-59.

[8] United Nationgs. Sendai framework for disaster risk reduction 2015—2030[EB/OL]. (2015-03-18)[2024-05-02]. https://www.undrr.org/media/16176/download?startDownload=20240502.

[9] 陈梓，高涛，罗年学，等. 反映自然灾害时空分布的社交媒体有效性探讨[J]. 测绘科学，2017，42(8)：44-48.
CHEN Zi, GAO Tao, LUO Nianxue, et al. Empirical discussion on relation between realistic disasters and social media data[J]. Science of Surveying and Mapping, 2017, 42(8): 44-48.

[10] CROOKS A, CROITORU A, STEFANIDIS A, et al. Earthquake: Twitter as a distributed sensor system [J]. Transactions in GIS, 2013, 17(1): 124-147.

[11] SMITH L, LIANG Q, JAMES P, et al. Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework [J]. Journal of Flood Risk Management, 2017, 10(3): 370-380.

[12] 高志国，李毅，张利辉，等. 洪水灾害时空过程模拟可视化表达研究进展与展望[J]. 水利水电技术(中英文)，2023，54(8)：43-53.
GAO Zhiguo, LI Yi, ZHANG Lihui, et al. Research overview on spatiotemporal process simulation and expression of flood disaster scenarios[J]. Water Resources and Hydropower Engineering, 2023, 54(8): 43-53.

[13] TANG Z H, ZHANG L G, XU F H, et al. Examining the role of social media in California's drought risk management in 2014[J]. Natural Hazards, 2015, 79(1): 171-193.

[14] SACHDEVA S, MCCAFFREY S, LOCKED. Social media approaches to modeling wildfire smoke dispersion: Spatio-temporal and social scientific investigations[J]. Information, Communication & Society, 2016, 20 (8): 1146-1161.

[15] WANG Y, RUAN S, WANG T, et al. Rapid estimation of an earthquake impact area using a spatial logistic growth model based on social media data[J]. International Journal of Digital Earth, 2019, 12 (11): 1265-1284.

[16] 薄涛. 基于社交媒体的地震灾情数据挖掘与烈度快速评估应用[D]. 哈尔滨：中国地震局工程力学研究所，2018.
BO Tao. Earthquake Disaster Data Mining and Application of Rapid Intensity Assessment based on Social Media[D]. Harbin: Institute of Engineering Mechanics, China Earthquake Administration, 2018.

[17] KRYVASHEYEU Y, CHEN H, OBRADOVICH N, et al. Rapid assessment of disaster damage using social media activity[J]. Science Advances, 2016, 2(3): e1500779.

[18] 苏凯，程昌秀，Murzintcev N，等. 主题模型在基于社交媒体的灾害分类中的应用及比较[J]. 地球信息科学学报，2019，21(8)：1152-1160.
SU Kai, CHENG Changxiu, Murzintcev N, et al. Application and comparison of topic model in identifying latent topics from disaster-related tweets[J]. Journal of Geo-Information Science, 2019, 21(8): 1152-1160.

[19] HUANG S, DU Y Y, YI J W, et al. Understanding human activities in response to typhoon hato from multi-source geospatial big data: A case study in Guangdong, China[J]. Remote Sensing, 2022, 14(5): 1269.

[20] XIAO C, ZHANG X D, XING Z Y, et al. Investigation of the Expression Method of Theme-Typhoon Disaster Information [J]. ISPRS International Journal of Geo-Information, 2021, 10(3): 109.

[21] CHEN Z, LIM S. Social media data-based typhoon disaster assessment [J]. International Journal of Disaster Risk Reduction, 2021, 64: 102482.

[22] 谢雪苗，邵亦文. 社交媒体数据分析在台风灾害管理中的应用潜力探究：以台风“杜苏芮”对福建省的影响为例[J]. 热带地理，2024，44(6)：1090-1101.
XIE Xuemiao, SHAO Yiwen. Application potential of social media data analytics in typhoon disaster management: Taking the impact of Typhoon Doksuri on Fujian province as an example[J]. Tropical Geography, 2024, 44(6): 1090-1101.

[23] 梁春阳，林广发，张明锋，等. 社交媒体数据对反映台风灾害时空分布的有效性研究[J]. 地球信息科学学报，2018，20(6)：807-816.
LIANG Chunyang, LIN Guangfa, ZHANG Mingfeng, et al. Assessing the effectiveness of social media data in mapping the distribution of typhoon disasters[J]. Journal of Geo-Information Science, 2018, 20(6): 807-816.

[24] 杨辰，潘顺，严岩. 基于自然语言识别的上海市报警灾情数据识别及其气象灾害特征分析研究[J]. 自然灾害学报，2021，30(3)：142-150.
YANG Chen, PAN Shun, YAN Yan. Research on recognition of alarm disaster data and analysis of meteorological disaster features in Shanghai based on natural language recognition algorithm [J]. Journal of Natural Disasters, 2021, 30(3): 142-150.

[25] 王艳东，李昊，王腾，等. 基于社交媒体的突发事件应急信息挖掘与分析[J]. 武汉大学学报(信息科学版)，2016，41(3)：290-297.
WANG Yandong, LI Hao, WANG Teng, et al. The mining and analysis of emergency information in sudden events based on social media[J]. Geomatics and Information Science of Wuhan University, 2016, 41(3): 290-297.

[26] 黄晶，李梦晗，康晋乐，等. 基于社交媒体的暴雨灾情信息实时挖掘与分析：以 2019 年“4·11 深圳暴雨”为例[J]. 水利经济，2021，39(2)：86-94.
HUANG Jing, LI Menghan, KANG Jinle, et al. Mining and analysis of rainstorm disaster information based on social media—Case study of Shenzhen rainstorm on April 11, 2019[J]. Journal of Economics of Water Resources, 2021, 39(2): 86-94.

[27] CHAUHAN U, SHAH A. Topic modeling using Latent Dirichlet Allocation [J]. ACM Computing Surveys, 2021, 54(7): 1-35.

[28] 顾荣直，吴洪颜，吴海英. 基于灰色关联度等方法的江苏雷暴大风灾害风险评估与区划[J]. 暴雨灾害，2025，44(1)：91-99.
GU Rongzhi, WU Hongyan, WU Haiying. Risk assessment and zoning of thunderstorm-gale disasters in Jiangsu Province based on grey correlation degree and other methods [J]. Torrential Rain and Disasters, 2025, 44(1): 91-99.

[29] 陈静雯，马福民，刘新，等. 基于神经网络的预警领域分词仿真算法[J]. 计算机仿真，2021，38(12)：1-6.
CHEN Jingwen, MA Fumin, LIU Xin, et al. Word segmentation simulation algorithm based on neural network in early warning field [J]. Computer Simulation, 2021, 38(12): 1-6.

[30] BLEI D M, NG A Y, JORDAN M I, et al. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[31] 刘丽华. 自然灾害微博舆情的社会计算模型建构研究[D]. 武汉：武汉大学，2018.
LIU Lihua. Research on Social Computing Model Construction of Natural Disasters Weibo Public Opinion based on Sentiment Analysis and Topic Modeling[D]. Wuhan: Wuhan University, 2018.

[32] HANKAR M, KASRI M, HSSANE B A. A comprehensive overview of topic modeling: Techniques, applications and challenges [J]. Neurocomputing, 2025, 628: 129638.

[33] JELODAR H, WANG Y L, YUAN C, et al. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey [J]. Multimedia Tools and Applications, 2019, 78(11): 15169-15211.

[34] CORTES C, VAPNIKV. Support Vector Machine [J]. Machine Learning, 1995, 20(3): 273-297.

[35] OYEDELE O. Determining the optimal number of folds to use in a *K*-fold cross-validation: A neural network classification experiment[J]. Research in Mathematics, 2023, 10(1): 2201015.

[36] NTI I K, NYARKO-BOATENG O, ANING J. Performance of machine learning algorithms with different *K* values in *K*-fold cross-validation[J]. International Journal of Information Technology and Computer Science, 2021, 13(6): 11.

[37] ZHANG X Y, LIU C A. Model Averaging prediction by *K*-Fold Cross-Validation[J]. Journal of Econometrics, 2023, 235(1): 280-301.

[38] 陈圣劼，刘梅，杨梦兮，等. 江苏“4·30”强风雹成因及双偏振雷达特征分析[J]. 气象科学，2022，42(5)：638-649.
CHEN Shengjie, LIU Mei, YANG Mengxi, et al. Analysis on causes of ‘4.30’ severe gales and hails event and associated characteristics of dual-polarization radar echoes over Jiangsu [J]. Journal of the Meteorological Sciences, 2022, 42(5): 638-649.

[39] 杨永清，王鹏博，张媛媛. 气象灾害舆情的时空演化特征及影响因素分析[J]. 晋图学刊，2024(4)：16-29.

YANG Yongqing, WANG Pengbo, ZHANG Yuanyuan. Analysis of spatio-temporal evolution characteristics and influencing factors of public opinion during meteorological disasters[J]. Shanxi Library Journal, 2024(4): 16-29.

[40] 刘淑涵，王艳东，付小康. 利用卷积神经网络提取微博中的暴雨灾害信息[J]. 地球信息科学学报，2019，21(7)：1009-1017.
LIU Shuhan, WANG Yandong, FU Xiaokang. Extracting rainstorm disaster information from microblogs using convolutional neural network[J]. Journal of Geo-information Science, 2019, 21(7): 1009-1017.

[41] 黄晶，吴星妍，李梦晗. 不同主体视角下极端暴雨灾害事件网络舆情演化研究[J]. 水利经济，2023，41(4)：94-101.
HUANG Jing, WU Xingyan, LI Menghan. Research on the evolution of network public opinion of extreme rainstorm disaster events from the perspective of different subjects[J]. Journal of Economics of Water Resources. 2023, 41(4): 94-101.

[42] 张晓蔚，李佳帅，朱亮，等. 南通机场4.30大风天气过程诊断分析[J]. 民航学报，2023，7(3)：59-64.
ZHANG Xiaowei, LI Jiashuai, ZHU Liang, et al. Diagnosis and analysis of a gale weather process at Nantong airport[J]. Journal of Civil Aviation, 2023, 7(3): 59-64.

[43] 吴洪颜，顾荣直，王勇. 基于灾情数据的江苏省大风灾害等级和损失风险评价方法[J]. 气象科学，2024，44(5)：997-1002.
WU Hongyan, GU Rongzhi, WANG Yong. Evaluation method of gale disaster grade and loss risk in Jiangsu Province based on disaster data [J]. Journal of the Meteorological Sciences, 2024, 44(5): 997-1002.

[44] LEYKIN D, LAHAD M, AHARONSON-DANIELL. Gauging urban resilience from social media[J]. International Journal of Disaster Risk Reduction, 2018, 31: 393-402.

[45] ANJUM S, MUSTAQ Z, SARWAR M, et al. Climate change and its impact on agriculture[J]. Journal of Agriculture and Biology, 2024, 2(2): 199-218.

[46] FAN C, ESPARZA M, DARGIN J, et al. Spatial biases in crowdsourced data: Social media content attention concentrates on populous areas in disasters [J]. Computers, Environment and Urban Systems, 2020, 83: 101514.

[47] JAIDKA K, GUNTUKU S C, LEE J H, et al. The rural-urban stress divide: Obtaining geographical insights through Twitter [J]. Computers in Human Behavior, 2021, 114: 106544.

[48] PALA, WANG J, WU Y, et al. Social media driven big data analysis for disaster situation awareness: A tutorial [J]. IEEE Transactions on Big Data, 2023, 9(1): 1-21.

[49] 张岚，艾文文，罗晓春，等. 气象灾害下网络信息传播模式研究：以台风为例[J]. 气象科学，2020，40(6)：868-874.
ZHANG Lan, AI Wenwen, LUO Xiaochun, et al. Research on communication of network information under meteorological disaster: Case of typhoon[J]. Journal of the Meteorological Sciences, 2020, 40(6): 868-874.

（责任编辑　王海锋）