文章编号：2095-2163(2023)07-0019-08 中图分类号：TP311，G252 文献标志码：A

社交媒体中应急救援信息分类的影响特征研究

沈洪洲[1,2]，居 玥[1]
（1 南京邮电大学 管理学院，南京 210003；2 南京邮电大学 信息产业融合创新与应急管理研究中心，南京 210003）

摘 要：突发事件应急管理中社交媒体数据质量参差不齐、难以直接为应急救援机构或志愿者的现场救援活动提供帮助，探究有助于从突发事件的社交媒体数据中快速挖掘出应急救援信息的关键特征，从而提升社交媒体数据的严谨性，推动社交媒体数据纳入正式的应急决策过程具有重要意义。以“微博”平台为例，通过对“微博”平台的分析和相关研究文献的总结，确定了 8 个潜在影响微博内容能否支撑应急救援行动的特征。基于“#河南暴雨互助#”话题下的微博内容、传播和用户维度抽取 8 个特征，以决策树模型为基准模型，通过 CART 算法评估各个特征对区分应急救援信息的贡献度。结果表明，信息内容地址信息特征、信息内容语言特征、信息主体特征是社交媒体中的应急救援信息分类的关键特征。

关键词：微博；应急救援信息；基本特征；数据挖掘；决策树

Research on the influence features of emergency rescue information classification in social media

SHEN Hongzhou[1,2], JU Yue[1]
(1 School of Management, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2 Research Center for Information Industry Integration, Innovation and Emergency Management,
Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

【Abstract】 Social media data contains a lot of low-quality information in emergency information management, which makes it difficult to directly provide instant help for the on-site rescue activities of emergency management agencies or volunteers. In order to improve the preciseness of social media data and promote the inclusion of social media data into the formal emergency decision-making process, it is necessary to explore the features that can provide help to quickly mine high-quality emergency rescue information from the social media data in emergencies. Taking the "Weibo" platform as an example, this research, through the analysis of the "Weibo" platform and the summary of relevant research literature, determines eight characteristics that potentially affect whether the microblog content can support emergency rescue operations. Based on the content, communication and user dimensions of Weibo under the topic of "Henan rainstorm mutual assistance", eight features are extracted. Taking the decision tree model as the benchmark model, the contribution of each feature to the help of distinguishing emergency rescue information is evaluated by CART algorithm. The results show that information content address information characteristics, information content language characteristics, and information subject characteristics are the key features of emergency rescue information classification in social media.

【Key words】 Weibo; emergency rescue information; basic features; data mining; decision tree

0 引 言

2021 年郑州 720 暴雨给应急信息管理机构带来了严峻的挑战。在这一过程中，全社会对突发事件的应急反应能力、资源调配能力的表现尤为突出，展现了对传统通信技术和应急信息管理的现有技术能够有效使用的能力，例如地理信息和全球定位系统、遥感技术等当代技术的使用[1-2]；但在此过程中也暴露了这些技术中公民参与度低，难以直接支撑应急救援行动的不足。然而在重大突发事件中，全社会共同参与的动员行动也是必需的，也都会对应急救援行动提供不可或缺的帮助[3]。因此，可以更加主动积极地借助广大民众（尤其是突发事件现场的民众）的群体智慧，来帮助采集、识别、完善和传

基金项目：国家自然科学基金（71974102）；江苏省研究生科研与实践创新计划项目（KYCX21_0833）研究成果之一。

作者简介：沈洪洲（1980-），男，博士，副教授，主要研究方向：社会化媒体、信息资源管理；居 玥（1998-），男，硕士研究生，主要研究方向：社会化媒体。

收稿日期：2022-08-24

播应急管理过程中需要的各类应急救援信息,使其成为应急信息管理的重要信息来源,从而对做出正确的应急决策[4]。

已有学者对相关实践案例进行研究,通过对社交媒体中广大民众发布的信息进行挖掘,进而给应急救援机构提供决策支持[5]。例如,陈茜等学者[6-7]的研究发现突发自然灾害事件背景下的微博所提供的信息可以帮助应急管理部门了解公众情绪走向、认知变化以及公众的态度,并且社交媒体上的用户生成的内容也被广泛用于为紧急救援机构或志愿者开展的现场救援活动提供建议[8]。显然,社交媒体的积极作用已经被应急管理机构承认,但是从社交媒体中挖掘出的应急救援信息的可靠性和有效性仍然不尽如人意,大多应急管理机构没有将社交媒体数据纳入正式的应急决策过程。为在应急信息管理中更加充分地发挥社交媒体综合优势,还需进一步深入研究如何在海量的社交媒体信息中挖掘出真正有价值的应急救援信息。

然而,在突发事件期间,任何一个社交媒体用户都能够发布与事件相关的信息和观点,并进行讨论。由于用户量和数据量巨大[9],研究发现:一方面,在突发事件中不同处境的民众借助社交媒体分布广和传播快的优势,能够提供大量、实时的信息;另一方面,这些不同身份背景的民众由于能力限制,提供的信息质量参差不齐,其中不乏一些低相关性、低质量的内容[10],从而导致难以区分社交媒体中的应急救援信息和普通信息。在此背景下,如何快速挖掘能够帮助识别直接提供态势感知、现场帮助、求助等支撑应急救援行动的社交媒体内容,探讨区分应急救援信息的关键特征从而提升突发事件中使用社交媒体信息的可信度,将社交媒体应急信息纳入正式的应急决策过程,是一个值得深入探讨的研究问题。

本文关注于社交媒体中应急救援信息的收集和挖掘,以微博为具体研究平台,运用数据挖掘方法探究社交媒体中应急救援信息和普通信息之间的特征差别,探讨区分应急救援信息的关键特征,从而帮助应急救援机构更加高效地利用社交媒体数据。

1 相关工作

1.1 社交媒体中应急信息的特征研究

由社交媒体用户生成的信息可以有效地用于不同的场景,包括突发事件中的应急信息管理。几十年前,社交媒体是社交网络的一种技术,而现在则已用作解决问题的工具而不仅仅是技术,并且逐渐成为突发事件应急管理中实时信息获取的重要渠道。Saroj 等学者[11]通过系统的综述,发现突发事件中社交媒体信息主要集中于位置预测、情感分析、求助 & 帮助、时间以及损失伤亡这 5 种类型的信息。对于不同类型的突发事件,信息内容的差异也将导致分类标准的不同。例如,Nguyen 等学者[12]将突发事件信息归类为与事件相关和无关的 2 种粗粒度分类的信息;Derczynski 等学者[13]对突发事件中社交媒体信息进行分类研究,将其粗粒度地分类为信息丰富的和无信息的,尽管区分社交媒体中直接提供态势感知、现场帮助、求助的应急救援信息是粗粒度的分类,不能够帮助进一步理解突发事件的细节,但是却能够有效帮助紧急救援机构及时获取所需要的信息。在社交媒体中应急救援信息的粗粒度分类过程中,本文发现社交媒体在突发事件中主要使用了以下 3 个层面的特征,包括:信息内容、传播和用户特征。

社交媒体的信息内容特征是对突发事件最直观的反映,对社交媒体信息的内容特征进行挖掘分析,能够发现用户表达人物、地点、状态等细节信息,包括对支撑应急救援行动有价值的信息。在突发事件中,社交媒体信息数量和内容还会随时间演变而有明显的变化,同时不同地理位置的社交媒体信息在内容、数量方面也会有着显著差异[14]。而对社交媒体中人类活动的单词频率和位置相关信息的进一步研究,也表明人们的情绪和活动受到暴雨强度的显著影响,验证了社交媒体的内容特征在一定程度上代表着人们的态度和行为[15]。

社交媒体的信息传播特征则是突发事件相关信息在公众中的传播认可的反映,研究发现社交媒体的传播特征能够在一定程度上反映内容的有效性,例如包含态势感知、损失情况和求助位置等能够支撑应急救援行动的关键信息的社交媒体数据,在传播途径中更容易得到社交媒体用户的关注度[16];包含求助、联系和情感的推文等信息特征的微博,其关注度也与信息特征数量成正比[17],吴布林等学者[18]就直接指出了高转发率的社交媒体应急信息更有可能拥有更高的质量。毫无疑问,这些研究都体现了传播特征在一定程度上对于内容质量的反映。

另外,社交媒体的用户特征、即信息主体特征问题也一直是当前推动社交媒体信息纳入正式应急决策过程的关键点。Chen 等学者[19]发现经认证的微博用户往往比未经认证的用户具有更高的社交网络

活动强度和更大的影响力，能够为突发事件的应急管理提供更多可靠有用的信息。并且不同用户所发布信息内容的影响性、权威性、专业性等方面也有着显著差别[20]。

因此本文认为在对社交媒体进行分析时，需要综合考虑社交媒体信息的内容、传播和用户特征，一方面能够传递突发事件中用户的求助、帮助的观点，另一方面也是其他用户对于该条微博态度的传递。因此，从这 3 个维度中抽取相应的特征指标，探究帮助区分社交媒体内容中应急救援信息的关键的特征，从而更加有效地从嘈杂的用户生成内容中提取出应急救援信息。

1.2　社交媒体应急信息质量的评价研究

在突发事件发生时，通过社交媒体进行信息沟通主要有 4 个方向：用户对用户（C2C）、用户对政府机构（C2A）、政府机构对政府机构（A2A）、政府机构对用户（A2C）[21]。在这一沟通过程中，应急机构可以通过收集来自用户的 C2C 和 C2A 信息，来帮助应急救援行动的开展。然而，由于突发事件下社交媒体质量层次不齐，在搜索不太具体的词汇时，数据非常“嘈杂”、缺乏上下文，使得数据质量难有保障，不足以直接帮助应急救援行动[22]。因此，在缺少足够权威评估标准的条件下，部分研究者使用了人类反馈的方法在主观上进行分析判断来评估信息质量[20,23]。除了人类反馈这类主观上评估应急信息质量的方法外，朱益平等学者[24]从测量方法的四要素出发，提出了应急信息质量测量框架。针对应急信息质量评价体系的建立，徐文强等学者[25]从大数据角度下对应急信息质量评估进行研究，从内容质量、描述质量、信息约束这 3 个维度抽取了 8 个指标构建了大数据环境下应急信息质量评估指标体系。另外还有相关研究者在主要的利益相关者、应急服务机构和市民的合作下，开发了一套包含需求、场景、用例的指标来进行突发事件相关社交媒体信息的衡量。其衡量体系由可理解性，相关性、完整性、及时性和可信性这 5 个指标构建[26]。

除了对应急信息质量某一指标和应急信息服务质量评估的探讨之外，吴雪华等学者[27]基于文本向量表示、语言、形式和用户四个维度的特征，采用机器学习对社交媒体应急信息的质量进行自动识别分类。刘校麟等学者[28]使用机器学习识别突发事件中的微博谣言，结果表示机器学习识别谣言的正确率远高于 80%。除此之外，在突发事件中，机器学习方法还被普遍用于突发事件信息抽取[29]、突发事件文本分类[30]、突发事件中情感分析[31]。综上所述，在评估应急信息质量的标准和方法上，未形成统一的质量标准，而机器学习也日渐成为突发事件中社交媒体信息的评价与处理的重要方法。

因此，本文在研究相关理论与实践的基础上，以“微博”为具体研究平台，选取“#河南暴雨互助#”话题为研究数据，从用户生成内容中抽取能够帮助区分应急救援信息的指标，利用机器学习进行应急救援信息分类实验，旨在探究社交媒体中应急救援信息和普通信息之间的特征差别，探讨区分社交媒体中应急救援信息中不同特征的影响程度。

2　研究设计

2.1　研究对象

在对郑州 720 暴雨事件的关注中，研究发现微博“#河南暴雨互助#”话题下的微博内容与应急救援的相关性较高，存在较多的信息能够有效支撑应急救援行动，因此将其作为研究对象。在数据搜集阶段，用 Python 编写关于微博的相关爬虫，爬取微博“#河南暴雨互助#”话题下的原创微博，从 2021.08.20 开始进行数据爬取；通过爬虫程序输入“#河南暴雨互助#”关键词，设置日期为 2021.07.20~2021.08.12，发送到微博搜索引擎，对相关话题下的原创微博爬取数据。由于话题下的原创微博只提供微博的点赞数、评论数、转发数以及微博的相关正文内容，并不足以支撑本文的分析，所以通过爬取发布微博信息的用户主页链接，从而进入用户主页以爬取用户的主页相关内容，如微博数、关注数、粉丝数，以利于后续对相关特征的进一步分析。

获得微博用户信息和正文内容后，对获得的微博内容进行重复性等验证，研究文本内容发现 2021.08.02 之后的微博相关内容对于应急救援的相关性都较低，最终选择 2021.07.20~2021.08.03 期间的微博，删除重复微博后获得的7 979条微博数据，以便进行此后的数据分析。接下来为了获得微博内容是否是应急救援信息，对微博内容进行人工标注数据集，分类为1 936条应急救援信息和6 043条普通信息。

本文研究通过对“微博”网站上应急救援信息的观察分析，并结合对已有的微博应急救援信息内容相关研究文献的整理，在此基础上展开研究论述如下。

2.2　研究方法

本文以数据挖掘为主要研究方法，将评估各个

特征对于社交媒体应急救援信息分类的影响程度，其中使用了 CART 算法作为评估特征贡献度的算法。

研究首先在阅读相应的参考文献以及“微博”平台的数据构成的基础上，确定并筛选了所有可能对社交媒体应急救援信息分类产生影响的特征。最终确定的潜在特征见表 1。

表 1 社交媒体应急救援信息分类潜在影响特征

Tab. 1 Potential impact characteristics of social media emergency rescue information classification

编号	特征名称	特征组成	说明
F_1	信息内容语言特征	文本单词数	微博文本去除停用词后的单词数量
F_2	信息内容语义特征	情感值	每个微博文本的情感值;分为正面、中性、负面情感值
F_3	内容地址信息特征	文本是否拥有精确地址信息	每条微博信息内容是否包含精确信息的特征,有则为 1,无则为 0
F_4	内容联系信息特征	文本是否拥有联系信息	每条微博信息内容是否包含联系信息的特征,有则为 1,无则为 0
F_5	图片信息特征	是否拥有图片	每条微博信息内容是否包含图片信息的特征,有则为 1,无则为 0
F_6	标签信息特征	微博内容包含标签数	每条微博包含的标签数
F_7	信息传播特征	信息评论数、点赞数、转发数	每个微博文本评论数、点赞数、转发数综合得分 0 ~ 100
F_8	信息主体特征	信息发布者微博账号发布的微博数、关注数、粉丝数	每条微博发布者发布的微博数、关注数、粉丝数的综合得分 0 ~ 100

2.2.1 人工标注数据集

社交媒体应急救援信息分类的训练可以看作是一个二分类问题，所以需要选择正样本和负样本。为了确保应急救援信息分类衡量标准的可靠性，需要对研究数据集中的微博内容进行人工标注，即人工判断每一条微博对于应急救援行动是否有用，即能否提供态势感知、现场帮助、求助信息。研究中招募了 8 名大学生志愿者进行人工数据标注，标注过程按照如下步骤进行：

步骤 1 标注要求的培训。对志愿者进行标注要求培训，介绍了数据标注的目的，并详细解释标注的要求和注意点。在志愿者理解数据标注要求后，还进行了试标注，从而确保志愿者们充分掌握数据标注的要求。

步骤 2 数据标注过程。数据人工标注过程由 8 名经过训练的志愿者进行。7 979 条博文数据分为 4 组，每组数据同时被 2 名志愿者分别标注，因此，每条微博都拥有 2 个由不同志愿者标注的结果。志愿者首先通过对微博正文进行仔细阅读并充分理解后，判断微博的文本内容是否对应急救援行动有用，进行标注。每条微博的标注结果分为 3 种，包括：有用、无用、不确定。

步骤 3 核对并确定标注结果。将每条微博的 2 个标注结果进行比对，以形成最终的数据标注结果，最终结果只能是有用或者无用。确定过程如下：

(1)如果 2 个标注结果相同(同时为有用，或同时为无用)，则直接采用该标注结果。

(2)如果一条微博存在 2 个不同的标注结果，即 2 个标注完全相反，或者结果中有“不确定”时，负责标注的 2 人与第三方研究人员共同分析讨论确定最终的标注结果。

正样本为标注为有用、即应急救援信息，标注为无用信息，即普通信息被视为负样本。最终得到 1 936条正样本，6 043 条负样本。

2.2.2 应急救援信息分类特征提取

(1)内容语言学特征提取。微博内容预处理后，采用“Jieba”分词去除文本内容中的停用词，随后统计每条微博内容的单词数量，将文本的单词数量记为内容语言学特征 F_1。

(2)内容语义特征提取。微博内容的情感一般可以分为正向、中性或者负向。本文通过专门的 Python 程序结合成熟的情感词典，计算出该微博文本内容的情感值。首先，对单条微博的文本内容分词后的词汇列表进行遍历，检查出词汇中的程度副词、否定词和情感词，记录相应位置，将积极和消极情感词分别标记为 1、-1。然后，找出程度副词和否定副词的权重，与情感词加权得到情感值得分。计算程序采用了知网 HowNet 情感词典、台湾大学 NTUSD 简体中文情感极性词典以及大连理工大学的中文情感词汇本体库。情感值得分大于 0、小于 0、等于 0 分别代表该微博文本内容表现为正面情感倾向、负面情感倾向以及中性情感倾向。将程序计算得到的情感值记为内容语义特征 F_2。

(3)内容精确地址提取。从信息学角度来说，如果一条文本与其他文本在某个关键属性上差别越大，那么就可以利用这个属性的差别来区分文本的

类别[32]。社交媒体应急救援信息中涉及关键属性，那么其为应急救援信息的概率就越高。而在应急救援信息中，精确的地址信息和联系信息被视为能够帮助救援的关键信息[15]。因此，本文通过自行编写的 Python 程序对微博内容根据文本中的地址特征字进行命名实体识别，提取每条微博存在的精确地址信息，有则为 1，无则为 0。根据由微博内容是否存在精确地址信息形成的一个由 0、1 组成的字典，作为内容地址信息特征 F_3。

(4)内容联系信息提取。在应急救援信息中帮助救援行动开展的关键属性除了精确地址之外，联系信息也被认为是区分应急救援信息的一个重要属性。因此，对每条微博原始博文进行了正则提取联系信息，有则为 1，无则为 0。而根据由微博内容是否存在精确地址信息所形成的由 0、1 组成的字典，则作为内容联系信息特征 F_4。

(5)图片数量提取。由于突发事件中图片能够更为直观地展示受害者以及旁观者的处境，本文将图片数量作为考虑的属性之一。文中是在微博爬取过程进行图片数量的抽取，通过 Python 爬取微博图片并计数，记为图片信息特征 F_5。

(6)标签数量提取。标签是微博话题是否与事件强相关的重要因素，而微博内容中含有的标签数与是否是应急救援信息的关联问题也是亟需探讨的内容。本文采用 Python 程序对微博文本利用正则表达式提取内容中的标签并计数，记为标签数量特征 F_6。

(7)信息传播特征提取。由上文综述可知，社交媒体的信息传播特征能够反映突发事件相关信息在公众中的传播认可，能够有效地评估信息在传播过程中公众的认可度。因此，本文爬取了每条微博的评论数、转发数、点赞数。根据其中位数以及平均数，分别赋值为 0.1、0.1、0.01。3 个维度的最大值分别为 33.33，最终得到的信息传播特征分数为 0~100。将加权得到的信息传播特征得分记为信息传播特征 F_7。

(8)信息主体特征提取。由上文可知，微博发布者的主体特征是信息源可靠性的重要属性。因此本文进入了微博内容发布者主页，爬取了主页中的微博发布数、粉丝数、关注数、微博认证等级。再根据其中位数、平均值、等级数分别赋值为 0.01，0.1、0.1。这 3 个维度的最大值同样为 33.33，最终得到的信息主体特征分数为 0~100。将加权得到的信息传播特征得分记为信息主体特征 F_8。

3　实验评估

为了挖掘出真正能够帮助区分社交媒体中应急救援信息的关键特征项，本部分研究首先通过自行编写的 Python 程序基于研究数据集选择最佳的分类模型，然后根据选定的分类模型对各个特征进行特征贡献度分析，确定能够区分社交媒体中应急救援信息产生重要影响的特征项，并对这些特征项进行讨论。

3.1　社交媒体应急救援信息分类模型选择

在阅读参考文献的基础上，研究确定了朴素贝叶斯、逻辑斯蒂回归、决策树算法作为拟定的初步算法，为了进一步确定最合适的算法，研究拟采用十折交叉验证法分别用朴素贝叶斯、逻辑斯蒂回归和决策树算法结合信息内容、主体、传播这 3 个维度中抽取的 8 个特征进行分类结果比较，以确定最佳分类算法。评估标准拟选定为 $F-score$，实验结果见表 2。

表 2　不同分类算法性能比较

Tab. 2　Performance comparison of different classification algorithms

方法	精度	召回率	F 值
朴素贝叶斯	0.916	0.578	0.709
逻辑斯蒂回归	0.921	0.609	0.733
决策树	0.920	0.626	0.745

从表 2 中分析可知，3 个分类器分类的精度都在 80%以上，综合判断选择了精度和 F 值都较高的决策树算法作为分类模型算法。

3.2　特征的统计以及贡献度分析

为了能够了解数据集分布情况，本文对社交媒体中应急救援信息和普通信息的 8 个数值型指标以及信息传播特征和信息主体特征这 2 个综合指标进行了描述性统计分析，统计其最小值、最大值、平均值、中位数、标准差，统计结果见表 3。由表 3 中可以见得，转发数、评论数、点赞数、发布微博、粉丝数这 5 个指标标准差较大，表明这部分帮助区分社交媒体中应急救援信息的特征数据也不稳定，并且根据分类的实验结果，选择 8 个特征进行分类实验的结果(0.745)也优于 12 个基本指标特征的实验结果(0.726)。因此，本文选择信息内容语言学特征、信息内容语义特征、内容地址信息特征、内容联系信息特征、图片信息特征、标签信息特征、信息传播特征、信息主体特征这个 8 个特征作为基本特征项。

表 3 特征数据的统计差异分析

Tab. 3 Statistical difference analysis of characteristic data

指标名称	答案分类	最小值	最大值	平均值	中位数	标准差
文本单词数	应急救援信息	0	112	51.37	44	1 421.21
	普通信息	0	92	34.32	32	1 316.42
情感值	应急救援信息	−347	426	5.32	3	17.22
	普通信息	−527	314	0.42	0	4.65
转发数	应急救援信息	0	11 575	65.24	4	375.22
	普通信息	0	10 000	36.88	3	6 724.34
评论数	应急救援信息	0	6 077	10.35	1	1 158.00
	普通信息	0	7 134	13.37	0	2 798.31
点赞数	应急救援信息	0	71 359	149.85	1	12 453.81
	普通信息	0	87 642	138.24	0	6 623.22
发布微博数	应急救援信息	1	92 146	2 248.48	494	6 229.31
	普通信息	1	96 728	2 373.32	404	13 131.22
粉丝数	应急救援信息	1	5 724 442	28 690.05	108	213 695.59
	普通信息	1	4 243 221	15 672.21	98	176 553.36
关注数	应急救援信息	1	20 000	371.89	244	2 578.23
	普通信息	1	20 000	624.21	203	567.62
信息传播特征	应急救援信息	0	100	3.016	3	17.22
	普通信息	−527	314	0.42	0	4.65
信息主体特征	应急救援信息	0	100	34.74	9	13.34
	普通信息	0	100	16.73	4	3.27

在分类模型中,2 种较为常用的分类方法是树归纳法和线性逻辑斯蒂回归方法,本文通过使用树归纳法计算基尼不纯度度量来计算特征贡献度[33],采用 CART 算法进行剪枝,即采用一种二分递归分割技术,将分类样本集分为 2 个子样本集,生成的决策树的每一个非叶节点都有 2 个分枝。在 CART 算法中,使用独立于训练样本集的测试样本集对分枝样本集的分类错误进行计算,找出分类错误最小的子树作为最终的分类模型[33-34]。本研究中,利用 CART 算法求得的特征贡献度如图 1 所示。因此,可以确定对社交媒体应急救援信息分类产生重要影响的 3 个特征:内容地址信息特征、内容语言学特征、信息主体特征。

由图 1 可知,内容地址信息特征、即精确地址信息在社交媒体应急救援信息分类中的影响程度最大,本文对这一项数据进行分析。对比应急救援信息和普通信息中是否包含精确地址信息,发现应急救援信息中的精确地址信息约为 80%,而普通信息中精确地址信息只有 25%。并且,包含精确地址信息的普通社交媒体内容有 50%是官方媒体机构对受影响地区和受害者的综合报道。因此,研究认为精确地址信息是区分应急救援信息的关键特征,并且在区分个人发布的社交媒体内容中的应急救援信息方面能够提供更加优越的效果。

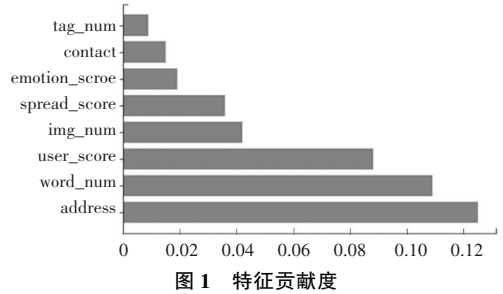

图 1 特征贡献度

Fig. 1 Contribution of different characteristics

信息内容语言特征的文本单词数量就是指去除停用词之后的社交媒体内容分词后的单词数量(47.23)。其中,普通文本的单词数量为 34.32,远低于应急救援文本的单词数量为 51.37。为了避免由于单词数量极值影响实验判断的科学性,进一步分析了文本单词数量的中位数,分别为 32,44。这也表明在一定程度上,社交媒体内容中文本单词数量越多,关于应急事件描述越详细,更有可能是应急救援信息。为了进一步评估关键特征在区分社交媒体

中应急救援信息方面的表现，本文对信息语言学特征（*word_num*）和信息主体特征（*user_score*）进行偏相关关系以进一步阐释其表现，结果如图 2 所示。由图 2 可知，微博内容中单词数量与社交媒体内容是否是应急救援信息有着明显的递增关系，说明微博内容单词数量越多、描述越详细，社交媒体内容就越有可能是应急救援信息。

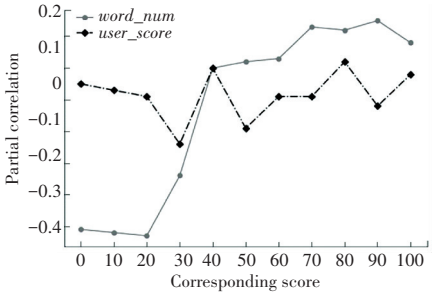

图 2　关键特征与应急救援信息分类偏相关关系

Fig. 2　Partial correlation between key features and emergency rescue information classification

信息主体特征包含微博发布者的微博数、粉丝数、关注数，是社交媒体信息来源可信度的重要特征。分析信息主体得分的平均数，应急救援信息（34.74）明显高于普通信息（16.73），且应急救援信息的信息主体得分中位数（9）也远高于普通信息得分（4）。这就清楚地表明，高质量用户在突发事件中能够提供信息的质量也更高。但是在偏相关关系分析中，微博发布者的用户主体特征并没有与应急救援信息分类表现出明显关系，可得信息主体特征与其他特征共同影响着应急救援信息的分类。

研究中还发现，原以为联系方式信息和图片信息在应急救援信息分类中能够起到关键性作用，然而，在本次实验中却显示出联系信息和是否存在图片没有对区分应急救援信息起到理想的作用。根据初步观察和以往的研究可知，图片信息尽管能够更加直观地展示突发事件的发展状况以及受害者现状，但是在社交媒体中很大一部分图片与社交媒体文本内容呈现为非强相关，这类普通信息对于图片使用的不严谨使得分类器难以依靠图片科学地区分出应急救援信息。与此同时，还进一步发现被归类为应急救援信息的部分内容创作者可能不喜欢、甚至不适应使用图片表达信息，还有部分内容创作者为旁观者，不能够提供高相关性的图片，这也在一定程度上降低了图片的区分应急救援信息的贡献度。另外，对于社交媒体信息中联系信息的使用，根据本次研究与发布者的沟通以及综合分析，有着相当一部分应急救援信息发布者没有意识到除微博本身外联系信息发布的关键性。还有一部分应急救援信息是旁观者的收集，其联系信息的缺失是由于自身能力和获得信息手段的限制。

4　结束语

针对社交媒体内容质量层次不齐、难以有效区分其中应急救援信息的现实问题，本文研究主要基于对“微博”上应急救援信息的分析与相关文献研究，从信息内容、传播、用户三个维度提取了 12 个指标、8 类特征，选择 CART 算法进一步分析这些特征对于区分应急救援信息关键性。研究结果显示内容地址信息特征、信息内容语言特征、信息主体特征对社交媒体中的应急救援信息分类有重要作用。上述探讨发现是对社交媒体中应急救援信息研究的进一步补充。

接下来，将探索如何更高效地在“微博”平台上识别应急救援信息。如何引导用户上传更高质量的应急救援信息，从而帮助应急救援行动快速实施，也是后期需着重探讨与研究的方向。例如，给用户提供更为专业的应急救援信息模板，在社交媒体发布端自动识别与突发事件低相关性的质量信息，给用户提供信息类别的选择上传选项等。

参考文献

[1] 程明睿，高宏. 卫星技术在突发事件中的应用[J]. 中国应急管理，2022(04)：62-63.

[2] 赵春霞. 城市应急管理中 GIS 技术的应用研究[J]. 科技资讯，2022，20(07)：34-36.

[3] 李萍，付绯风. 贝克风险社会理论视阈中的冲突思想及其现实意义[J]. 内蒙古社会科学，2016，37(03)：68-72.

[4] CONRADO S P，NEVILLE K，WOODWORTH S，et al. Managing social media uncertainty to support the decision making process during emergencies[J]. Journal of Decision Systems，2016，25(sup1)：171-181.

[5] YATES D，PAQUETTE S. Emergency knowledge management and social media technologies：A case study of the 2010 Haitian earthquake[J]. International journal of information management，2011，31(1)：6-13.

[6] 陈茜，陈思菁，毛进，等. 突发事件背景下内容添加型转发微博的情绪与认知变化研究[J]. 情报科学，2021，39(11)：51-59.

[7] KIM J，PARK H. A framework for understanding online group behaviors during a catastrophic event[J]. International Journal of Information Management，2020，51：102051.

[8] 邵力，乔墩. 网络热点事件微博评论中的情感冲突分析[J]. 兰州大学学报(社会科学版)，2016，44(06)：62-68.

[9] CHEN Junting，SHE J. An analysis of verifications in microblogging social networks—Sina Weibo[C]//2012 32nd International Conference on Distributed Computing Systems Workshops. Macau：IEEE，2012：147-154.

[10] MIRBABAIE M，STIEGLITZ S，VOLKERI S. Volunteered

geographic information and its implications for disaster management[C]//2016 49th Hawaii International Conference on System Sciences (HICSS). Koloa, HI, USA:IEEE, 2016: 207-216.
[11]SAROJ A, PAL S. Use of social media in crisis management: A survey[J]. International Journal of Disaster Risk Reduction, 2020, 48: 101584.
[12]NGUYEN D, MANNAI K A A, JOTY S, et al. Robust classification of crisis-related data on social networks using convolutional neural networks [C]//Proceedings of the International AAAI Conference on Web and Social Media. Montreal, QC, Canada:AAAI,2017, 11(1):1-4.
[13]DERCZYNSKI L, MEESTERS K, BONTCHEVA K, et al. Helping crisis responders find the informative needle in the tweet haystack[J]. arXiv preprint arXiv:1801.09633, 2018.
[14]TAKAHASHI B, TANDOC E C, CARMICHAEL C. Communicating on Twitter during a disaster: an analysis of tweets during Typhoon Haiyan in the Philippines[J]. Computers in Human Behavior,2015, 50: 392-398.
[15]FANG Jian, HU Jiameng, SHI Xianwu, et al. Assessing disaster impacts and response using social media data in China: A case study of 2016 Wuhan rainstorm[J]. International journal of disaster risk reduction, 2019, 34: 275-282.
[16]SUTTON J, SPIRO E S, JOHNSON B, et al. Warning tweets: serial transmission of messages during the warning phase of a disaster event[J]. Information, Communication & Society, 2014, 17(6):765-787.
[17]LI Lifang, ZHANG Qingpeng, WANG Xiao, et al. Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on Weibo[J]. IEEE Transactions on Computational Social Systems, 2020, 7(2): 556-562.
[18]吴布林,薛冬,杨克. 重大突发公共事件中社交媒体用户信息行为研究[J]. 情报理论与实践,2021,44(10):137-141.
[19]CHEN Junting, SHE J. An analysis of verifications in microblogging social networks—Sina Weibo[C]//2012 32nd International Conference on Distributed Computing Systems Workshops.Macau: IEEE, 2012: 147-154.
[20]谢雨杉,柯青,王笑语,等. 新冠疫情背景下情绪与信息行为的关系及情绪角色的主题分析[J]. 图书情报工作,2022,66(08): 102-112.
[21]MOI M, FRIBERG T, MARTERER R, et al. Strategy for processing and analyzing social media data streams in emergencies [C]//2015 2nd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM). Rennes, France:IEEE, 2015: 42-48.
[22]陆恒杨,范晨悠,吴小俊. 面向网络社交媒体的少样本新冠谣言检测[J]. 中文信息学报,2022,36(01):135-144,172.
[23]RUDRA K, SHARMA A, GANGULY N, et al. Characterizing and countering communal microblogs during disaster events[J]. IEEE Transactions on Computational Social Systems, 2018, 5(2): 403-417.
[24]朱益平,刘春年. 应急信息的数据准确性测量框架研究[J]. 图书馆学研究,2017(13):88-92,58.
[25]徐文强,刘春年,周涛. 大数据环境下应急信息质量评估体系研究[J]. 图书情报工作,2020,64(02):50-58.
[26]MOI M, RODEHUTSKORS N, KOCH R. An ontology for the use of quality evaluated social media data in emergencies[J]. IADIS International Journal on WWW/Internet, 2016, 14(2): 38-57.
[27]吴雪华,毛进,陈思菁,等. 突发事件应急行动支撑信息的自动识别与分类研究[J]. 情报学报,2021,40(08):817-830.
[28]刘校麟,陈蕾. 基于机器学习的突发事件微博谣言识别技术研究进展[J]. 网络安全技术与应用,2022(05):54-56.
[29]孙小川. 面向微博的突发事件抽取方法研究[D]. 北京:中国人民公安大学,2020.
[30]吕龙. 基于深度学习的突发事件新闻文本分类研究[D]. 武汉:武汉理工大学,2020.
[31]杨秀璋,武帅,张苗,等. 基于 TextCNN 和 Attention 的微博舆情事件情感分析[J]. 信息技术与信息化,2021(07):41-46.
[32]吴军. 数学之美[M]. 北京:人民邮电出版社,2012:60-64.
[33]PERLICH C, PROVOST F, SIMONOFF J. Tree induction vs. logistic regression: A learning-curve analysis[J]. 2003,4:211-255.
[34]张艺梅,丁香乾,贺英,等. 逻辑模型树算法性能分析与改进研究[J]. 微型机与应用,2014,33(23):25-28.

(上接第 18 页)

[2] 陈梦莹,王展青. 基于 FRFT 及 HVS 的自适应数字水印算法[J]. 计算机应用研究,2017,34(07):2180-2083.
[3] MENGZhaoxiong, MORIZUMI T, MIYATA S, et al. Design scheme of copyright management system based on digital watermarking and blockchain [C]//42nd IEEE International Conference on Computer Software & Applications. Tokyo, Japan: IEEE ,2018:359-364.
[4] 崔锦泰.小波分析导论[M]. 美国:美国学术出版社,1992.
[5] 陈涛. 基于 DCT 域的数字图像水印算法研究及应用[D]. 济南:山东师范大学,2011.
[6] CHEN W, QUAN C, TAY C J .Optical color image encryption based on Arnold transform and interference method[J]. Optics Communications,2009,282(18):3680-3685.
[7] 熊玮. 一种基于位置变换和灰度变换相结合的数字图像置乱方法[J]. 软件导刊, 2011, 10(11):159-161.
[8] 吴强,彭亚雄. 基于 DWT-FRFT 变换和 QR 分解的盲数字水印算法[J]. 电子科技,2018,31(10):53-55,68.
[9] 华梦, 王雷. 基于 DCT-DWT 的 FRFT 数字水印算法[J]. 南京理工大学学报,2015,39(04):435-439.
[10]尹康康, 石教英, 潘志庚. 一种鲁棒性好的图像水印算法[J]. 软件学报, 2001, 12(05):668-676.
[11]刘锋, 胡晨, 张萌,等. 一种基于小波变换的图像数字水印技术[J]. 电子器件, 2003, 26(01):56-59.
[12]刘挺, 尤韦彦. 一种基于离散小波变换和 HVS 的彩色图像数字水印技术[J]. 计算机工程, 2003, 29(04):115-117.
[13]张阳, 卿粼波, 何小海. 基于 DWT-SVD 鲁棒盲水印算法研究[J]. 智能计算机与应用,2022,12(02): 44-48,53.