引用格式：杨腾飞,解吉波,李振宇,等.微博中蕴含台风灾害损失信息识别和分类方法[J].地球信息科学学报,2018,20(7):906-917. [Yang T F, Xie J B, Li Z Y, et al. A method of typhoon disaster loss identification and classification using micro-blog information[J]. Journal of Geo-information Science, 2018,20(7):906-917.] **DOI:**10.12082/dqxxkx.2018.180062

微博中蕴含台风灾害损失信息识别和分类方法

杨腾飞[1,2],解吉波[1*],李振宇[3],李国庆[1]

1. 中国科学院遥感与数字地球研究所,北京 100049;2. 中国科学院大学,北京 100049;3. 山东科技大学,青岛 266000

A Method of Typhoon Disaster Loss Identification and Classification Using Micro-blog Information

YANG Tengfei[1,2], XIE Jibo[1*], LI Zhenyu[3], LI Guoqing[1]

1. Institute of Remote Sensing and Digital Earth Chinese Academy of Sciences, Beijing 100049, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China; 3. Shandong University of Science and Technology, Qingdao 266000, China

Abstract: Social media plays a more and more important role in the real-time disaster information distribution and dissemination. During the disaster event, social media usually generates and contains a lot of real-time disaster loss information, which is very useful for the timely disaster response and disaster loss assessment. However, the social media data has many shortcomings, such as high fragmentation of the information, sparsity of the text features, and the lack of annotated corpus and so on, which makes the traditional supervised learning method difficult to be effectively used for disaster information extraction. This paper proposed a fast disaster loss identification and classification method to extract the disaster information from social media data by extending the context features and matching feature words. By this method, we firstly extracted the keywords from a small amount of sample micro-blog text of different disaster loss categories based on Chinese grammar rules and constructed the pairs of feature words collocation. Then, we used the word vector model and the existing lexicon to supplement and expand these pairs of feature words collocation. And the external corpus was introduced to optimize the semantic collocation relationship between feature words according to the rules of the concurrence of Chinese words. At last, we built a classification knowledgebase for identification and classification of disaster loss information related to typhoon disasters included in micro-blog. An experiment system was developed to evaluate the method introduced in the paper. Typhoon "Meranti" landed on 15th September, 2016 was selected as a case study. Results show that this method has a significant effect (each comprehensive evaluation index of different categories is greater than 0.74) on identifying and classifying different categories of disaster loss information from social media. We mapped the spatio-temporal distribution of typhoon influence based on the classification results of disaster loss from social media. The experiment shows that the classification output data and maps could be used for the disaster loss evaluation and mitigation.

Key words: social media; typhoon disaster; short text classification; identification of disaster loss information;assessment of disasters

***Corresponding author:** XIE Jibo, E-mail: xiejb@radi.ac.cn

---

收稿日期：2018-01-18;**修回日期**：2018-03-20.

基金项目：国家重点研发项目(2016YFE0122600);国家自然科学基金项目(41771476)。[**Foundation items:** National Key Research and Development Program of China, No.2016YFE0122600; National Natural Science Foundation of China, No.41771476.]

作者简介：杨腾飞(1988-),男,博士生,研究方向为自然语言处理、灾害信息挖掘。E-mail: yangtf@radi.ac.cn

***通讯作者**：解吉波(1977-),男,博士,副研究员,主要从事地理空间数据基础设施,遥感,地理计算。E-mail: xiejb@radi.ac.cn

摘要：社交媒体在灾害信息的实时发布与传播中发挥着越来越重要的作用。在灾害发生过程中，社交媒体中蕴含的实时灾损信息对灾情及时响应和评估有重要意义。然而，这些涉灾文本具有信息破碎度高、文本特征稀疏、标注语料库匮乏等缺点，使得传统的基于监督学习的方法难以有效提取其中的灾损信息。为此，本文提出了一种通过扩展上下文特征和匹配特征词的方法来快速识别和分类社交媒体中蕴含的不同类别的灾损信息。本方法首先基于中文语法规则，抽取小规模不同灾损类别下微博文本中的涉灾关键词构建特征词搭配对。然后，利用词向量模型和已有词库对这些特征词搭配对进行补充和扩展。同时，根据中文词语共现规则，引入外部语料库优化特征词间的语义搭配关系。最终，以此为基础构建台风灾损分类知识库对灾情文本中蕴含的不同类别灾损信息进行识别和分类。本文以2016年9月15日台风“莫兰蒂”登陆事件作为研究案例，以评估本文方法在灾损信息识别和分类上的效果。结果表明，本文方法对微博文本中蕴含的不同类别风灾损失信息的识别和分类效果显著（各类别综合评价指标都达到了0.74以上）。基于灾损信息分类结果，本文绘制了台风影响的时空分布图，从而进一步说明本文方法在灾害损失评估和减灾救灾方面的效用。

关键词：社交媒体；台风灾害；短文本分类；灾损信息识别；灾情评估

1 引言

近年来，全球自然灾害频繁发生，给人类的生命和财产安全带了严峻威胁。然而，传统的灾害信息收集手段存在着严重的滞后性，已无法满足政府部门及时开展救灾工作的需求。以Twitter[1]、Facebook[2]、微博数据[3]等为代表的社交媒体，其广泛的参与度、多源的传播渠道等特点[4]，已成为政府部门及时了解灾情进展的一项有效手段。在灾害发生的第一时间，公众作为第一接触群体，扮演着动态传感器的角色[5]，能够主动通过上传与灾害有关的事件信息[6]，从而为政府部门提供第一手资料，辅助救灾决策的制定。目前，相关研究已取得显著进展，成果涉及灾情事件检测[7-8]、时空分析探究[9-10]、灾害下社会响应特征研究[11]、灾害发生趋势预测模拟[12-13]等方面，极大地提高了减灾救灾工作的效率。

但现有的研究方法更多的只是对灾害事件本身作识别分析，如仇培元等[14]基于语义相似度抽取微博客中蕴含的地理事件；Qu等[8]利用贝叶斯分类器从海量新浪微博文本中识别地震事件等。而对于灾害描述文本中蕴含的细粒度灾损信息的识别很少提及。这部分信息涉及灾害损失的各个方面且具有很强的时效性，对政府部门及时了解灾情进展并做出具有针对性的救灾行动意义重大。然而，微博文本中蕴含的这些灾损信息破碎程度高、文本特征稀疏，且可用的开放标注语料库匮乏，因此识别和分类难度也较大。

国内外针对短文本信息的识别和分类已开展了较多研究，方法通常包括基于触发词过滤和监督学习[14]，前者是利用对触发词的判断来确定待分类文本是否与主题相关，后者则是通过人工标注的语料库来训练分类模型实现主题分类[15-16]，常见的分类模型包括支持向量机[17-18]、K近邻[19-20]、朴素贝叶斯[21]、随机森林[22]、最大熵[23]等。基于触发词的方法通过提取文本中与主题相关的特征关键词，利用特征关键词来识别和分类目标文本，但该方法对非主题相关的文本分类效果较差[24]。基于监督学习的方法并不受主题相关性约束，但需要人工制备大规模标注语料，该过程费时费力，且各类别语料的数量和质量对于最终的分类结果影响较大。本文旨在从微博短文本中识别和分类不同类别的灾损事件信息，分类粒度较细，信息破碎度高，且同一条文本中通常包含多个灾损类别事件，各灾损事件上下文特征稀疏、表达形式复杂多样，难以建立大规模分类语料库，因此更适合采用基于触发词过滤的方法。例如，某一条微博文本“台风天气，断水停电，窗外一片狼藉，到处都是倒塌的树和砸坏的车”，其中包含了“断水”、“停电”、“倒塌的树”、“砸坏的车”等多个灾损类别事件。

通过上述分析，本文提出了一种基于特征语义扩展和中文词法搭配关系构建灾损分类知识库来识别和分类微博中蕴含的台风灾损信息的方法，并对台风受灾区的微博数据进行了分类测试和灾损评估，验证了本文方法在实际应用中的效果。

2 研究方法

在中文文本分类领域，微博与传统的新闻文本有着显著差异[25]。以本文使用的新浪微博为例，其以短文本为主，最长不超过140个字符，语言表述口语化严重，且同一条微博中通常包含不同类别的灾损事件，信息破碎化程度较高。但分析发现，不同类别灾损信息的表达依然符合汉语的基本语法规则，如主谓、谓宾结构等[14]，且与灾损文本的特征词一一对应。如“整个树被吹倒在地了”，根据语法规则抽取该文本中对应的词性搭配对“名词-动词”即

“树–倒”,可很好地作为文本灾损类别的标识。因此,本文通过构建语法规则,抽取已知灾损类别微博文本中的特征词对作为种子词对。为满足灾损信息的多样性表达,利用词向量模型和《同义词词林》补充和扩展这些种子词对。同时,通过构建外部语料库,利用词频和词语间共现规则优化补充和扩展后的特征词。在此基础上,基于一定的规则构建台风灾损分类知识库对测试文本进行灾损事件的识别和归类。

算法流程如图1所示,包括灾损分类知识库的构建和微博文本蕴含灾损事件信息的识别和分类2部分。具体步骤包括:① 基于训练语料构建规则模板,包括词法规则和否定词约束规则;② 利用词法规则抽取小规模不同灾损类别文本中的特征词搭配对作为种子词对,并利用词向量模型和《同义词词林》补充和扩展种子词对;③ 优化特征词,包括去除低频词和优化词语间搭配关系;④ 根据优化后的特征词对构建分类知识库;⑤ 对测试文本作断句处理,并根据词法规则抽取各个短句的候选特征词与分类知识库以及否定词表匹配,从而完成灾损事件信息抽取和分类。

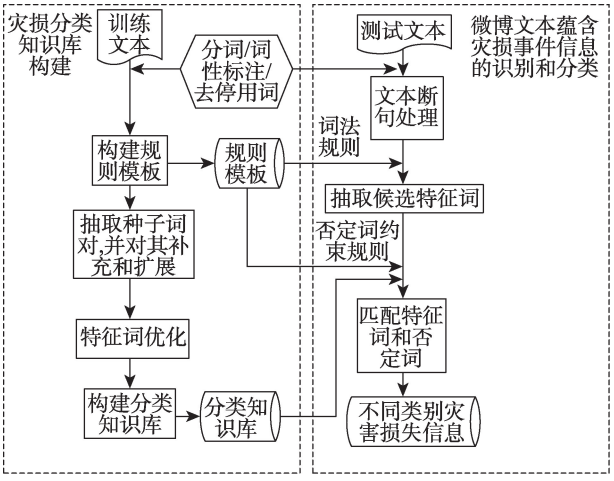


图1 算法流程

Fig. 1 Algorithm flow

2.1 构建规则模板

2.1.1 词法规则

新浪微博文本包含的灾害损失信息破碎程度高、表达形式多样,且同一微博文本可能包含多种类别的风灾损失信息。利用汉语语法结构规则抽取灾情事件上下文特征词可很好的表达原文含义,如短文本“整个树被吹倒在地了”中的名词“树”和动词“倒”构成的特征词搭配对可作为该文本的分类依据。本文词法搭配模式的构建基于风灾事件微博文本的统计和归纳,数据来源于文献[26]所提供的“2017年台风灾害社交媒体数据集”,选取该数据集中5000条台风相关的新浪微博文本,分析其中灾损事件信息的词法特征,得到如表1所示的词法规则模式。

表1 词法规则模式

Tab. 1 Pattern of lexical rule

模式规则	文本样例
v-n	到处都是被打碎的玻璃
n-v	整个树被吹倒在地了
a-n	一地的碎窗玻璃
n-a	道路一直不畅通
d-vi	很快小区就不再供水了
v-vi	即将停止供电
r-v	看见他被树给砸了
v-r	树枝被风吹断刚好砸到他
vi	今天停电一天

注:v为动词;n为名词;a为形容词;d为副词;r为代词;vi为不及物动词

2.1.2 否定词约束规则

根据词法规则抽取的灾损特征词可用于识别文本中包含的灾损信息,而特征词上下文中存在的否定词能够过滤候选文本中的非灾损信息。如“我家的玻璃还好没被吹破”,该文本中特征词对“玻璃–破”之间的否定词“没”标识了该文本的非灾损属性。通常,否定词多属副词,本文对“2017年台风灾害社交媒体数据集”中的9601条台风相关的新浪微博文本进行分词和词性标注,过滤出其中的副词,并提取这些副词中含有的否定词以构建否定词表。同时定义如下使用规则:

规则1:同一上下文中,当否定词位于任一特征词位置之前,则否定词对文本类别属性有反作用。如“树并没有被吹倒”,“台风天一直没有断水”等。

规则2:同一上下文中,当出现双重否定词时,则认为否定词对文本类别属性没有作用。如“台风天航班不会不被影响吧”。

2.2 特征词补充与扩展

基于词法规则抽取小规模标注语料中的特征词搭配对作为分类知识库构建的原始种子词对。以这些种子词对为基础,利用词向量模型和《同义词词林》扩展版丰富特征词搭配信息,以满足汉语

表达方式多样性的需求。

2.2.1 基于词向量模型补充特征词

互联网文本蕴含着丰富的风灾损失事件核心词，可有效补充原有特征词，从而丰富灾损事件的语义表达。本文以抽取的种子词对为基础，从互联网文本中提取与种子词距离相近的词作为补充词。

在自然语言处理领域，通常利用词向量模型计算词语间的距离，距离近的两个词相关度也高。常用的词向量模型包括CBOW和Skip-gram模型[27]，它们是由Mikolov等[28]在神经网络语言模型NNLM（Neural Network Language Model）的基础上改进来的。文献[29]给出了CBOW和Skip-gram模型性能上的详细对比，结果表明Skip-gram模型总体效果要好于CBOW模型。因此，本文在计算词语间相关度上也采用Skip-gram模型。该模型结构包含输入层、投影层和输出层，其原理是通过当前词W(t)来预测该词所在的词组序列的上下文信息，模型结构如图2所示。

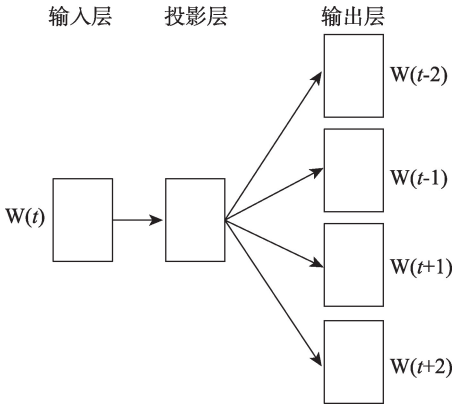


图2　Skip-gram模型结构

Fig. 2　The structure of Skip-gram model

Skip-gram模型旨在使得目标函数G达到最大，如式(1)所示。

$$G=\sum_{w_i\in C}\log p(Context(w_i)|w_i) \tag{1}$$

式中：w_i表示当前词；C表示上下文窗口；$Context(w_i)$表示与当前词w_i距离小于窗口C大小的上下文信息，其条件概率计算如式(2)所示。

$$p\left(Context\left(w_i\right)|w_i\right)=\prod_{u\in Context(w_i)}p(u|w_i) \tag{2}$$

上述公式表明，Skip-gram模型通过引入上下文语境信息来计算词与词之间的相关度，语境相似的词语其相关度较高。因此，可根据计算所得的与当前词w_i相关度最接近的词u作为该当前词的补充。例如，训练语料中出现了较多相似语境的文本如“路灯被台风刮倒了”、“电线杆被台风刮倒了”等，当以特征词“刮倒”作为模型的输入项时，“路灯”和“电线杆”与“刮倒”一词构成的条件概率会得到加强，二者可作为与“刮倒”相关度高的词被模型输出。同时，当以特征词“路灯”或“电线杆”作为输入项时，它们之间会因为和“刮倒”一词的共现关系，相关度也得到加强。表2为词汇相关度计算示例，将种子词对中的“树”和“倒”作为Skip-gram模型的输入项，模型从训练集中找出与“树”和“倒”相关度最高的前10个词。其中，“电线杆”一词作为原种子词对中没有出现过的新词被补充。

表2　词向量模型计算结果示例

Tab. 2　An example of the computational results of the word vector model

树	倒
大树	吹
整棵	大树
折断	断
应声	压垮
断	树干
倒	一棵
根	电线杆
一棵	棵
树枝	42棵
枝干	砸

2.2.2 基于《同义词词林》扩展特征词

词向量模型侧重于同语境新词的补充，而汉语对同一事件的描述用词多样，如“台风直接吹倒了一排树”和“我家楼下新种的小树直接被掀翻”，两文本都是描述台风对林木的影响，特征词搭配对分别为“树–倒”和“小树–掀翻”。其中“树”与“小树”、“倒”与“掀翻”分别构成同义关系。利用同义词在词向量模型的基础上进一步扩展特征词，可满足汉语表达多样化的需求。

一种有效、直接的同义词扩展方法则是利用新版《同义词词林》，该部词林包含了77 492条词语，共分为12个大类，94个中类以及1428个小类，小类下按照同义词划分了词群，词群下包含原子词群，《同义词词林》的结构与用法可参见相关文献[30−31]。本文利用《同义词词林》在补充后的种子词对的基础上作原子词群级别的同义扩展。

2.3 特征词优化

利用词向量模型和《同义词词林》补充和扩展

特征词搭配对，丰富了短文本的语义信息。但在搭配词对生长过程中易出现较多的低频词和错误的搭配关系，这对识别效果影响较大，因此需要作优化处理。

2.3.1 低频词去除

在中文短文本分类中，高频词对于分类有较大的促进作用，而低频词易增加短文本的噪声，降低分类效率[32]。本文以“2017年台风灾害社交媒体数据集”为语料库，匹配特征词词频低于4的所有词并予以去除。图3为低频词处理流程。

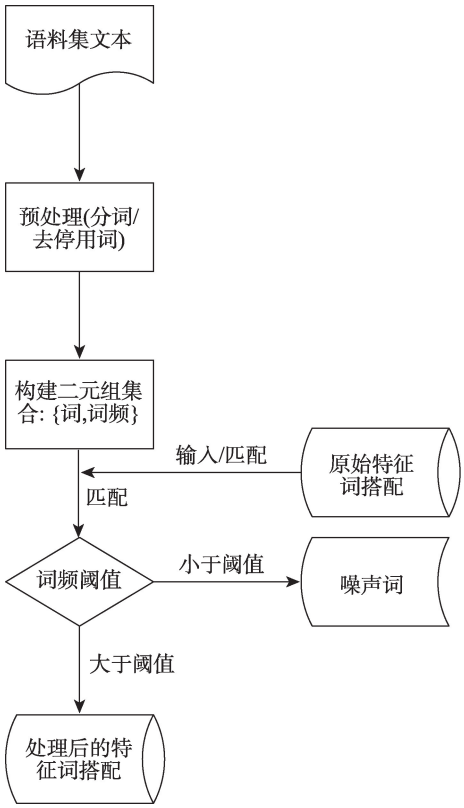


图3 低频词处理流程

Fig. 3 Process of low frequency word processing

2.3.2 词语搭配关系优化

在补充和扩展种子词对的过程中，难免会出现错误的搭配关系。如“树木”和“倒”可构成正确搭配，但“树木”的同义扩展词“树林”并不与“倒”构成正确搭配。

在中文自然语言处理中，两词在文本上下文中共现的次数越多，表明这两词相关度越强，越容易构成正确的搭配关系，因此可利用词语间的共现频率优化特征词搭配。词语间的共现频率通过搜狗实验室中文词语搭配库SogouR（http://www.sogou.com/labs/resource/w.php）和“2017年台风灾害社交媒体数据集”进行计算，其中搜狗实验室中文词语搭配库是搜狗搜索引擎基于全网文本建立的，其格式如下：

二元组1 同现次数1

二元组2 同现次数2

……

二元组N 同现次数N

其中，二元组包含构成搭配的2个词语。

但该中文词语搭配库是2006年10月统计产生的，其中涉及台风灾害相关的词语搭配信息有限。为保证词语间搭配关系的优化效果，本文增加了“2017年台风灾害社交媒体数据集”中的文本信息，并对该数据集作如下处理：

（1）按照标点符号“，”、“。”、“！”、“？”、“；”对文本进行断句，形成短句s。

（2）对各短句进行分词、去停用词和词性标注。

（3）按照词法规则抽取各短句中的相关词语搭配对$s=[w_1,w_2]$。

（4）将词语搭配对按照搜狗实验室中文词语搭配库格式处理，形成新的词语搭配库。

将上述2个中文词语搭配库合并，并与特征词搭配对迭代匹配，清除特征词搭配关系中低于共现频率阈值的词对，从而消除词语间的不正确搭配关系。本文设置共现频率阈值为4。图4为词语搭配关系优化流程。

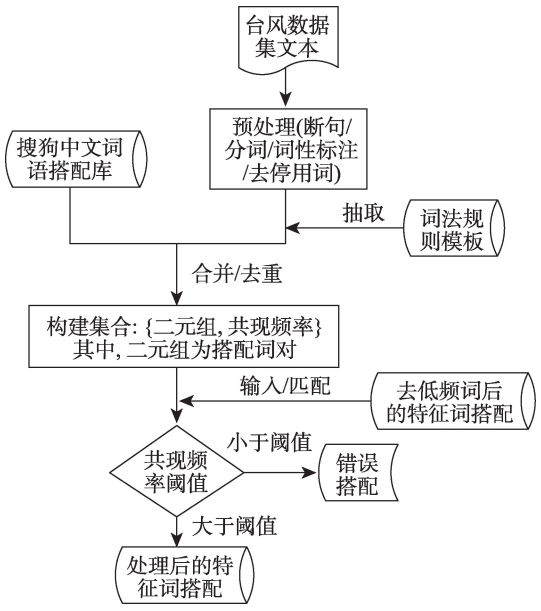


图4 词语搭配关系优化流程

Fig. 4 Optimization process of collocation relationship

2.4 分类知识库构建

基于优化后的各灾损类别特征词搭配对构建分类知识库，该分类知识库的结构参考《同义词词林》。以编码的形式分为4位，第一位为大写字母，按字母表顺序分别代表灾损大类；第二位为小写字母，按字母表顺序代表大类下包含的子类；第三位为数字，按照同一类下词语搭配关系划分，用来表示词群；第四位以w_1或w_2表示搭配词，包含了各词群下的所有原子词。例如：

Ba01w_1={“树木”，“树”，“果树”，“小树”，“大树”，…}

Ba01w_2={“断”，“倒”，“倾倒”，“折断”，“遭殃”，…}

Ba02w_1={“树林”、“林子”、“丛林”、“密林”，“园林”，…}

Ba02w_2={“遭殃”、“摧残”、“损坏”、“毁坏”，“摧毁”，…}

其中，B代表林业影响；a代表林业影响下的子类；01和02代表2种不同的搭配关系；w_1和w_2表示搭配词群，w_1通常表示实体名词，w_2通常表示与之搭配的动词、形容词或副词等。特征词集Ba01w_1与Ba01w_2（或Ba02w_1与Ba02w_2）中的各原子词间可构成搭配关系用于识别和分类林业影响类下的各灾损事件，如Ba01w_1中“树木”与Ba01w_2中的各词构成搭配来匹配待分类文本中的“名词–动词”候选词对。知识库的结构形式如表3所示。

表3　分类知识库结构示例

Tab. 3　An example of the structure of classified knowledge base

	编码位			
	1	2	3	4
符号举例	B	a	01	w_1/w_2
符号性质	大类	小类	词群	原子词群
级别	第1级	第2级	第3级	第4级

2.5 灾损事件抽取与分类

灾损事件通常包含于微博文本的短句中，因此，利用构建的分类知识库对微博文本中的各个短句作识别和分类，具体流程如下：

（1）按照标点“，”、“。”、“!”、“?”、“；”将待分类文本拆为短句集合$D=[s_1,s_2,\cdots]$。

（2）对每个短句文本分词和词性标注，按照词法规则抽取候选特征词搭配对，并记录特征词在短句中的位置，构建四元组$s=[w_1,w_2,i,j]$，其中w_1，w_2表示按照词法规则抽取的特征词，i和j表示特征词w_1和w_2在短句文本中的位置下标。同时根据否定词表匹配该短句文本中是否存在否定词，若存在，记录否定词的位置下标k。

（3）将各四元组s中的特征词对“w_1–w_2”与分类知识库不同灾损类别下的特征词搭配对匹配，同时根据否定词约束规则比较特征词位置下标i、j与否定词位置下标k的关系，从而判断该短句的灾损类别，以确定待分类微博文本的类别属性。

3　实验与分析

3.1 实验语料

本文以“2017年台风灾害社交媒体数据集”中的分类样例作为抽取种子词对的基础语料，词向量模型的训练语料库以该数据集中其余的所有文本构建，并对其中的新闻和微信公众号原始语料进行正文抽取和文本清洗处理，以减少噪声文本对模型的干扰。本文待分类文本来自于2016年9月15日厦门“莫兰蒂”台风当天的新浪微博，共1821条数据。灾损分类标准参考《全国气象灾情收集上报技术规范》（下称《规范》），根据该《规范》对台风灾害损失信息分为6个大类和11个小类（图5），待分类文本各灾损类别的分布如表4所示。

3.2 实验环境

本文基于Java语言研发了“台风灾害损失信息自动识别和分类系统”，用来作为算法测试平台，系统集成了对微博数据实时获取、处理、识别和分类等模块。其中系统的分词和词性标注功能调用NLPIR 2015工具包（http://ictclas.nlpir.org/），Skip-gram模型基于谷歌的词向量模型框架Word2vec实现。各分类结果的评测标准采用准确率（P）、召回率（R）和F-1值（综合评价指标），3个指标的计算公式如式（3）–（5）所示。

$$p=\frac{\text{正确分类的灾损信息条数}}{\text{对应灾损类别下信息识别的总条数}} \tag{3}$$

$$R=\frac{\text{正确分类的灾损信息条数}}{\text{对应灾损类别下信息应有的总条数}} \tag{4}$$

$$F-1\text{值}=\frac{2PR}{P+R} \tag{5}$$

3.3 实验结果与分析

基于召回率、准确率和F-1值的评测标准识别和分类厦门“莫兰蒂”台风当天的新浪微博文本，结果如表5所示。

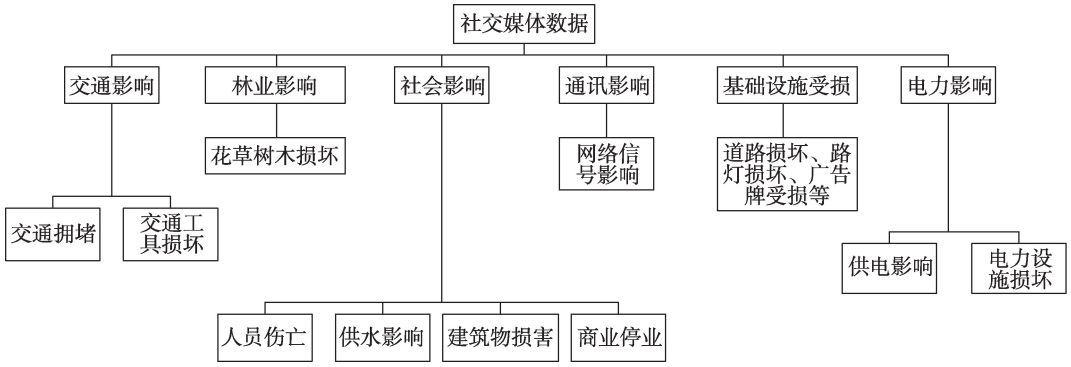


图5 灾损信息类别划分

Fig. 5 Classification of disaster loss

表4 各类别语料分布

Tab. 4 Distribution of different categories of corpus

类别编号	灾损类别	数量/条
1	人员伤亡	34
2	供水影响	337
3	建筑物损伤	154
4	商业影响	63
5	林业影响	181
6	交通受阻	138
7	交通工具损坏	107
8	供电影响	402
9	电力设施受损	138
10	通讯影响	163
11	基础设施损坏	104

表5 实验结果对比

Tab. 5 Comparison of experimental results

类别	评测结果		
	P/%	R/%	F-1值/%
第1类人员伤亡	68.00	89.47	77.27
第2类供水影响	87.32	95.48	91.22
第3类建筑物损伤	76.10	85.14	80.37
第4类商业影响	100.00	75.00	85.71
第5类林业影响	79.00	84.61	81.71
第6类交通受阻	78.74	87.71	82.98
第7类交通工具损坏	74.19	88.46	80.70
第8类供电影响	90.29	93.93	92.07
第9类电力设施损坏	78.54	70.53	74.32
第10类通讯影响	86.95	71.42	78.43
第11类基础设施受损	76.47	72.22	74.28

分类结果显示本文方法在准确率、召回率和F-1值表现较好。与目前常见的短文本分类案例相比，本文涉及分类类别较多，不同类别间有一定的交叉重叠，且同一个短文本涉及多种类别标签，这一定程度上增加了分类难度[16]。同时，实验文本口语化严重、特征词复杂多样、语料信息不均衡等特点，也较大的限制了分类效果[33]。但从现有相关研究成果来看，基于社交媒体的短文本分类在不同背景下的分类效果差别较大，如文献[14]对微博文本中蕴含的地理事件进行提取和分类，并同传统的监督学习方法作了对比，综合评价指标提高了10%以上，但也只达到了71.41%。文献[34]人工标注了大规模分类语料训练SVM模型用于识别微博中的地震事件，虽然综合评价指标达到了89%，但其所涉及的类别单一，且粒度较粗，本文研究之初，也做过相关算法的尝试，效果并不理想。因此，综合来说，本文方法在当前背景下分类效果较好。

本文实验结果在不同类别下的召回率和准确率差别较大。

（1）如图6所示，“基础设施受损”、“电力设施损坏”和“通讯影响”三类的召回率较低，不足75%。分析发现，“基础设施受损”和“电力设施损坏”两类的实体名词包含种类较多，在补充和扩展这类特征词时，并不能完全覆盖。如“小区的小健身广场被毁的不成样了!!!!!!”、“高速路口的收费亭都被掀翻了…”等，其中“健身广场”、“收费亭”并没有在特征词补充与扩展中被收录进分类知识库中。“通讯影响”类中如“手机刚才满格的信号瞬间变成一格了”、“手机突然变成澳门讯号了……什么情况？台风把澳门的讯号刮过来了？”等语义隐含的文本和“4G秒变2G”等不规则特征词的情况，也使得算法很难对其有效。

（2）如图7所示，“人员伤亡”和“交通工具损坏”两类的准确率较低，不足75%，其中“人员伤亡”类的准确率甚至不足70%。本文分析了错误召回的

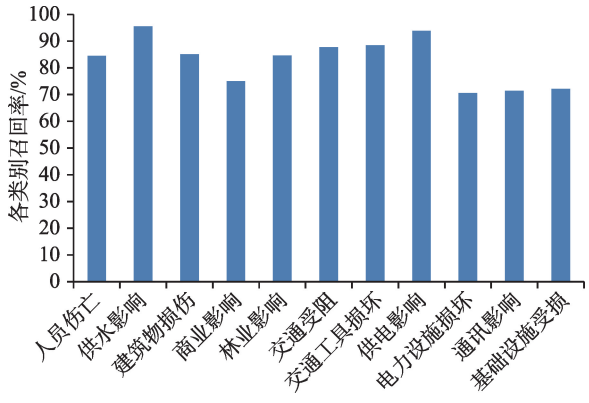


图6　各类别召回率

Fig. 6　Recall rates of various categories

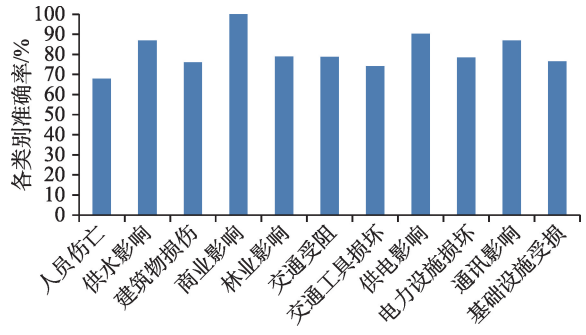


图7　各类别准确率

Fig. 7　Precision rate of various categories

文本，对于“人员伤亡”类别，考虑到在台风发生时，人员伤亡对于救援的重要性，因此，本文算法较多的考虑了“人员伤亡”的召回率，使得错误地召回了一些如“我幼小的心灵受到了伤害”等文本。对于“交通工具损坏”类，则由于一些文本的隐含语义造成错误的召回如“感觉友谊的小船翻了……”等。

4　应用分析

为验证本文方法在灾害实际应用中的效果，本文对实验结果作进一步分析。实验所涉及的数据来源于2016年9月15日当天与厦门“莫兰蒂”台风相关的新浪微博。该数据源包含了15日凌晨3时5分台风登陆至16日凌晨“台风”过境过程中共1821条数据，数据形式包含文本、发布时间、发布位置等信息。

图8展示了本文方法所识别和分类的灾损事件信息与微博发布时间的变化关系。由图8可看出，不同灾损类别的微博数量与总微博数量增幅基本呈正相关，其中“林业影响”、“供电影响”以及“供水影响”3类随时间增幅较大，表明这3类灾损信息在本次“莫兰蒂”台风中受关注度较高，这与事后厦门市官方发布的讯息基本一致（http://www.xm.gov.cn/xmyw/201609/t20160917_1361266.htm）。此外，关系图中表明，从当日16时、17时起，各灾损数据量信息快速增加，根据浙江省水利信息管理中心台风信息发布系统（http://typhoon.zjwater.gov.cn/default.aspx）所显示的台风时间与路径数据可知该时刻台风正继续向西北方向移动并逐渐离开厦门，如图9所示。因此，从16时起，人类活动逐渐增加，使得微博数量曲线显著增长。伴随着人类活动的增加，对于交通状况的关注度也逐渐增强，关系图中“交通受阻”类别曲线在当日16时台风过境后呈现小波峰，表明该时段人类出行受阻严重，所识别的“交通受阻”类别的微博文本增多，同时根据这些微博所提供的位置信息，可呈现图10所示的空间分布。这为政府部门及时制定城市交通快速恢复决策提供了有力的数据支持。此外，随着时间的推移，图11展示了各类别灾损信息在各时间段内的空间位置分布，以可视化的形式向政府部门刻画灾情进展。

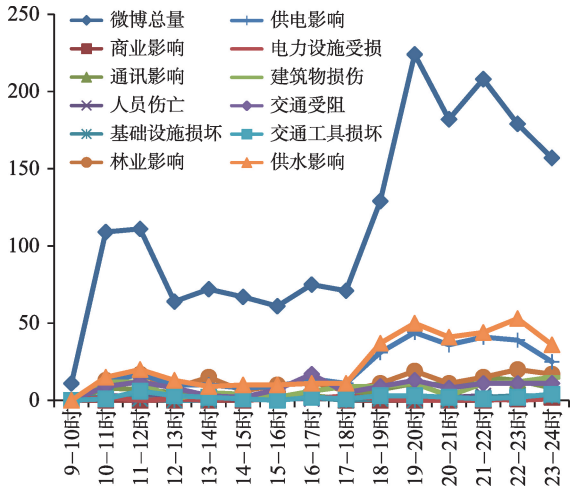


图8　微博量随时间变化关系图

Fig. 8　The variations of the quantity of “Sina-Weibo” with time

至23时，厦门市各类别灾损信息的空间分布情况如图12所示。图中各类别灾损标识信息均通过本文方法从当日厦门市各条新浪微博数据中提取得到。这在灾后第一时间向政府部门提供了受灾区域的整体损失信息。可见在本次台风过境中，厦门市思明区受灾最为严重，同安区和翔安区在供电供水方面受损较大，这些信息与灾后官方发布的灾损评估报告基本一致。

注：该图来自“台风信息发布系统”，网址是http://typhoon.zjwater.gov.cn/default.aspx

图9 台风“莫兰蒂”实时路径图

Fig. 9 Real-time path of typhoon "Meranti"

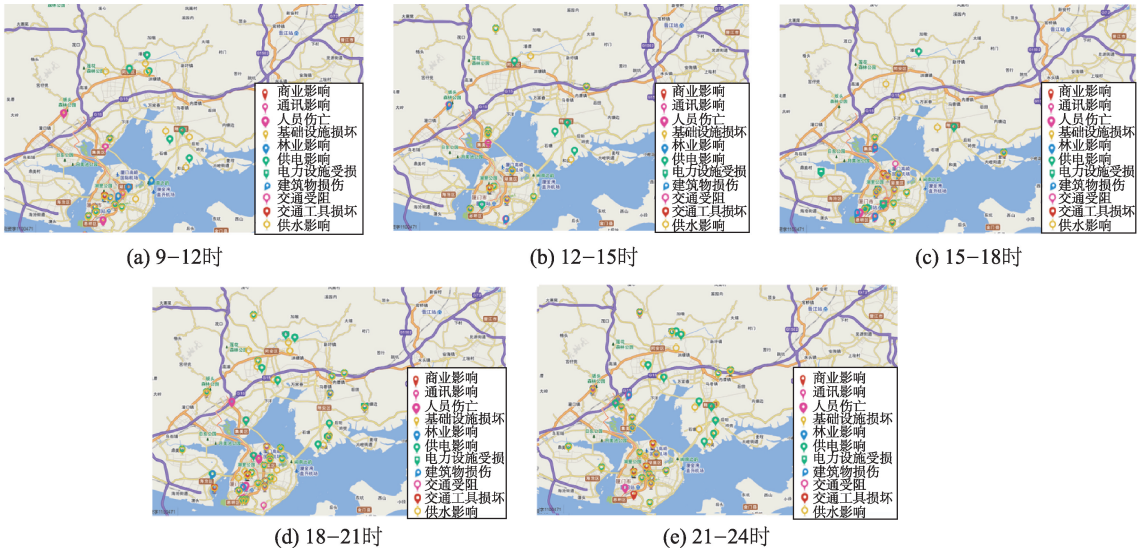
注：底图来源于天地图

图10 “交通受阻”信息空间分布

Fig. 10 Geospaital distribution of the "traffic obstruction" information

5 结语

灾害发生时，基于社交媒体的灾情收集方式为政府部门提供了大量有价值的信息。利用短文本分类技术提取社交媒体文本中包含的灾损事件信息并近乎实时的反馈给相关部门，能够为有针对性的救灾决策的制定提供数据支持。本文基于此需求，以2016年厦门台风“莫兰蒂”过境时的微博数据为基础，针对短文本分类中上下文信息匮乏的缺点，利用词向量模型和《同义词词林》补充和扩展短文本特征词，并在此基础上构建了台风灾损分类知识库，同时基于该分类知识库对微博中包含的风灾损失信息进行识别和分类。最终各类别召回率、准确率和F-1均达到了满意的效果。为进一步验证本



注：底图来源于天地图

图11　各时间段灾损信息空间分布

Fig. 11　Geospatial distribution of disaster loss in each time period

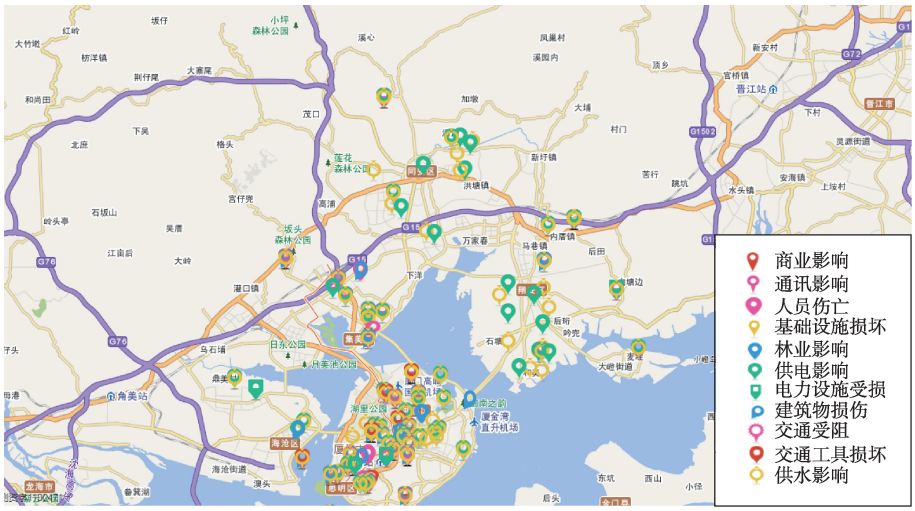


图12　灾损信息整体空间分布

Fig. 12　Overall geospatial distribution of disaster loss information

文方法在实际应用中的可靠性，本文对实验结果作了时空分析，最终结果同灾后官方发布的灾损评估信息基本一致，从而表明本文方法在灾害响应和应急分析上的有效性。但与此同时，社交媒体作为一种灾情获取的辅助手段，其提供的信息也存在一定的局限性，如本文实验所提取的各类别灾损事件的分布量与实际情况存在一定的偏差，这是由于一些公众未上传定位数据，导致一些微博数据地理位置信息缺失。但总体来讲，社交媒体作为新的辅助减灾救灾手段依然发挥着重要作用。

下一步，将尝试本文方法与传统的机器学习模型相结合，利用本文方法在文本特征扩展上的优势和传统机器学习模型上下文语义分析上的优势来进一步提高识别和分类的效果，同时考虑识别灾后各类基础设施的恢复信息从而为政府减灾救灾措施提供实时反馈。

参考文献(References)：

[1] Sakaki T, Okazaki M, Matsuo Y. Tweet analysis for real-time event detection and earthquake reporting system development[J]. IEEE Transactions on Knowledge & Data Engineering, 2013,25(4):919-931.

[2] Bird D, Ling M, Haynes K. Flooding facebook: The use

of social media during the queensland and Victorian floods[J]. Australian Journal of Emergency Management, 2012,27(1):27-33.

[3] 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 Y D, Li H, Wang T, et al. Emergency information mining and analysis of emergency based on social media[J]. Geomatics and Information Science of Wuhan University, 2016,41(3):290-297.]

[4] 彭敏,官宸宇,朱佳晖,等.面向社交媒体文本的话题检测与追踪技术研究综述[J] 武汉大学学报·理学版,2016,62(3):197-217. [Peng M, Guan C Y, Zhu J H, et al. A survey of topic detection and tracking technology for social media texts[J]. Journal of Wuhan University(Science Edition), 2016,62(3):197-217.]

[5] 牟乃夏,张恒才,陈洁,等.轨迹数据挖掘城市应用研究综述[J].地球信息科学学报,2015,17(10):1136-1142. [Mu N X, Zhang H C, Chen J, et al. A survey of urban application research on trajectory data mining[J]. Journal of Geo-information Science, 2015,17(10):1136-1142.]

[6] American Red Cross.Social media in disasters and emergencies.http://i.dell.com/sites/content/shared-content/campaigns/en/Documents/red- cross- survey- social- media- in-disasters-aug-2010.pdf, 2010.

[7] Sakaki T, Okazaki M, Matsuo Y, et al. Earthquake shakes Twitter users: rReal-time event detection by social sensors[C]. International Conference on World Wide Web. ACM, April 26-30,2010,Raleigh,NorthCarolina,USA,2010:851-860.

[8] Qu Y, Huang C, Zhang P, et al. Microblogging after a major disaster in China: A case study of the 2010 Yushu earthquake[C]. ACM Conference on Computer Supported Cooperative Work, CSCW 2011, Hangzhou, China, March. DBLP, 2011:25-34.

[9] Chae J, Thom D, Jang Y, et al. Special section on visual analytics: Public behavior response analysis in disaster events utilizing visual analytics of microblog data[J]. Computers & Graphics, 2014,38(1):51-60.

[10] 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 Z, Gao T, Luo N X, et al. Social media effectiveness to reflect the spatial and temporal distribution of natural disasters[J]. Science of Surveying and Mapping, 2017,42(8):44-48.]

[11] 刘宏波,翟国方.基于社交媒体信息不同灾害的社会响应特征比较研究[J].灾害学,2017,32(1):187-193. [Liu H B, Zhai G F. A comparative study of the social response characteristics of different disasters based on social media information [J]. Journal of Catastrophology, 2017,32(1):187-193.]

[12] Mark A. Stoové, Alisa E. Pedrana. Making the most of a brave new world: Opportunities and considerations for using Twitter as a public health monitoring tool[J]. Preventive Medicine, 2014,63(6):109-111.

[13] Paola Velardi, Giovanni Stilo, Alberto E. Tozzi, et al. Twitter mining for fine-grained syndromic surveillance[J]. Artificial Intelligence in Medicine, 2014,61(3):153-163.

[14] 仇培元,陆锋,张恒才,等.蕴含地理事件微博客消息的自动识别方法[J].地球信息科学学报,2016,18(7):886-893. [Qiu P Y, Lu F, Zhang H C, at al. Containing automatic recognition methods for geo-event micro-blog messages[J]. Journal of Geo-information Science, 2016,18(7):886-893.]

[15] 张春菊.面向中文文本的事件时空与属性信息解析方法研究[J].测绘学报,2015,44(5):590-590. [Zhang C J. Research on the analysis method of event space-time and attribute information for Chinese texts[J]. Acta Geodaetica et Cartographica Sinica, 2015,44(5):590-590.]

[16] 张雪英.基于机器学习的文本自动分类研究进展[J].情报学报,2006,25(6):730-739. [Zhang X Y. Research progress of automatic text classification based on machine learning[J]. Journal of the China Society For Scientific and Technical Information, 2006,25(6):730-739.]

[17] Kumar M A, Gopal M. A comparison study on multiple binary-class SVM methods for unilabel text categorization [J]. Pattern Recognition Letters, 2010,31(11):1437-1444.

[18] Burbidge R, Trotter M, Buxton B, et al. Drug design by machine learning: Support vector machines for pharmaceutical data analysis[J]. Computers & Chemistry, 2001, 26(1):5-14.

[19]]Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, et al. When Is “Nearest Neighbor” Meaningful?[C]. International Conference on Database Theory. Springer, Berlin, Heidelberg, 1999:217-235.

[20] Jiang S, Pang G, Wu M, et al. An improved K-nearest-neighbor algorithm for text categorization[J]. Expert Systems with Applications An International Journal, 2012,39(1):1503-1509.

[21] Sankaranarayanan J, Samet H, Teitler B E, et al. TwitterStand: News in tweets[C]. ACM Sigspatial International Conference on Advances in Geographic Information Systems. ACM, 2009:42-51.

[22] Xu B, Guo X, Ye Y, et al. An improved random forest classifier for text categorization[J]. Journal of Computers, 2012,7(12):2913-2920.

[23] Li R, Tao X, Lei T, et al. Using maximum entropy model for Chinese text categorization[J]. Journal of Computer Research & Development, 2005,42(1):578-587.

[24] 丁效,宋凡,秦兵,等.音乐领域典型事件抽取方法研究[J].中

文信息学报,2011,25(2):15-20. [Ding X, Song F, Qing B, at al. Research on typical event extraction method in music field[J]. Journal of Chinese Information Processing, 2011,25(2):15-20.]

[25] 张剑峰,夏云庆,姚建民.微博文本处理研究综述[J].中文信息学报,2012,26(4):21-27. [Zhang J F, Xia Y Q, Yao J M. Weibo text processing research review[J]. Journal of Chinese Information Processing, 2012,26(4):21-27.]

[26] Yang T, Xie J, Li G. A social media based dataset of typhoon disasters[DB]. Science Data Bank, 2017, DOI: 10.11922/sciencedb.547.

[27] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. ArXiv Preprint arX-iv:13013781,2013.

[28] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003,3:1137-1155.

[29] 熊富林,邓怡豪,唐晓晟. Word2vec的核心架构及其应用[J].南京师范大学学报(工程技术版),2015(1):43-48. [Xiong F L, Deng Y H, Tang X S. Word2vec's core architecture and its application[J]. Journal of Nanjing Normal University, 2015(1):43-48.]

[30] 刘丹丹,彭成 钱龙华,等.《同义词词林》在中文实体关系抽取中的作用[J].中文信息学报,2014,28(2):91-99. [Liu D D, Peng C, Qian L H, at al. The role of synonym in the extraction of Chinese entity Relationships[J]. Journal of Chinese Information Processing, 2014,28(2):91-99.]

[31] 王东,熊世桓.基于同义词词林扩展的短文本分类[J].兰州理工大学学报,2015,41(4):104-108. [Wang D, Xiong S H. Short text classification based on synonym word forest expansion[J].Journal of Lanzhou University of Technology, 2015,41(4):104-108.]

[32] 胡勇军,江嘉欣,常会友.基于LDA高频词扩展的中文短文本分类[J].现代图书情报技术,2013(6):42-48. [Hu Y J, Jiang J X, Chang H Y. Chinese short text classification based on LDA high-frequency word expansion[J]. New Technology of Library and Information Service, 2013(6):42-48.]

[33] 庞观松,蒋盛益.文本自动分类技术研究综述[J].情报理论与实践,2012,35(2):123-128. [Pang G S, Jiang S Y. A survey of automatic text classification technology[J]. Information studies: Theory & Application, 2012,35(2):123-128.]

[34] 白华,林勋国.基于中文短文本分类的社交媒体灾害事件检测系统研究[J].灾害学,2016,31(2):19-23. [Bai H, Lin X G. Social media disaster event detection system based on Chinese short text classification[J]. Journal of Catastrophology, 2016,31(2):19-23.]