面向突发公共事件网络舆情分析的领域情感词典构建研究*

李长荣 纪雪梅
（山东理工大学科技信息研究所，淄博 255049）

摘要：为了对突发公共事件网络舆情中的公众情感进行分析，本文构建了一种具有较好准确性和可靠性的面向网络舆情分析的领域情感词典。首先，基于现有通用情感词典在大规模网络舆论语料中进行情感词的识别和修正，将情感词分为7个大类和21个小类，并对情感词进行极性和强度标注，得到情感种子词典；其次，在情感种子词典的基础上利用Word2Vec模型和余弦相似度计算进行情感词扩展，得到新增情感词；再次，对新增情感词进行分类、极性和强度标注，最终构建一个领域情感词典；最后，选取新冠肺炎疫情事件的微博评论作为语料进行实验验证。结果，本文构建的词典对情感词的识别准确率为0.85，召回率为0.90，*F*1值为0.87，能够有效用于识别突发公共事件网络舆论中的情感类型和强度。

关键词：突发公共事件；情感词典；网络舆情；Word2Vec模型
中图分类号：G353.1　　DOI：10.3772/j.issn.1673-2286.2020.09.005
引文格式：李长荣，纪雪梅. 面向突发公共事件网络舆情分析的领域情感词典构建研究[J]. 数字图书馆论坛，2020 (9)：32-40.

《国家突发公共事件总体应急预案》对突发公共事件进行了说明，指出突发公共事件是突然发生，造成或者可能造成重大人员伤亡、财产损失、生态环境破坏和严重社会危害，危及公共安全的紧急事件。突发公共事件发生后，公众会通过社交媒体、论坛等网络平台发布事件相关的帖子、评论等舆论文本。这些文本不仅包含了事件相关的话题信息，同时也包含了人们对于人物、事件、不同观点等对象的情感倾向性，如喜爱、赞扬、愤怒和批评等。基于突发公共事件舆论文本的公众情感识别能够对突发公共事件下公众情绪的类型、正负面极性和强度进行自动分析，挖掘公众对突发公共事件的态度和情感倾向，有助于舆论走向的把握、情感的引导以及对事件的回应。

目前，文本情感分析的方法主要包括基于情感词典的情感分析方法、有监督的机器学习方法和弱监督的深度学习方法等。其中，基于情感词典的情感分析方法能够对公众情感表达的方式、用词、情绪的细分类型等进行准确分析。情感词典作为一种重要的情感资源，在词语、短语、句子及篇章等不同文本粒度的情感分析任务中起着重要的作用[1]。情感词典是进行公众情感自动分析的基础，在情感词典的基础上可提高文本分词的准确性；通过情感词典也可对公众使用的情感词进行识别，并进一步通过上下文语境进行情感类型和强度的计算。目前常用的情感词典多为通用情感词典，在对突发公共事件进行网络舆情分析时专用性不足，并且随着新的情感表达方式和情感词的不断出现，构建领域情感词典将可以大幅提高网络舆情情感分析的准确性。本文利用大规模突发公共事件舆论文本，结合通用情感词典和深度学习方法对领域情感词及情感词的类型和强度进行识别，旨在构建一个面向突发公共事件网络舆情分析的领域情感词典。

*本研究得到国家社会科学基金青年项目“突发事件情境下社交媒体用户情感表达行为的特征与驱动因素研究”（编号：16CTQ027）资助。

1 研究综述

目前，常用的开放中文情感词典主要有HowNet情感分析用词语集[2]、台湾大学自然语言处理实验室构建的情感词典NTUSD[3]和大连理工大学信息检索研究室发布的情感词汇本体库[4]。这些情感词典通用性较好，但其领域适应性较差。目前，情感分析主要应用于产品评论分析和突发公共事件网络舆情分析两个领域。有学者构建了不同商品领域的情感词典，如邓淑卿等[5]基于句法依赖规则和词性特征的情感词识别模型构建手机领域情感词典；蒋翠清等[6]使用AMVR投票集成规则构建汽车领域情感词典；郭顺利等[7]基于改进的SO-PMI算法构建中文图书评论情感词典。总结目前相关研究，领域情感词典的构建方法主要有两种，即基于语料库的方法和基于语义知识库的方法。

1.1 基于语料库的方法

基于语料库的方法，主要是根据语料中词语之间的共现信息、上下文信息来计算词语的情感极性。Hatzivassiloglou等[8]最先提出了利用句法连接来识别情感词并判断其极性，通过大量实验数据证明了连词前后词的极性关系。Turney等[9]基于一个词与其邻近词的情感趋于一致的思想，采用逐点互信息（PMI）和潜在语义分析（LSA）来估计关联程度，通过与正面或负面种子词的统计关联来识别词语极性。Gamon等[10]扩展了Turney的方法，增加了一个假设，即情绪相反的情感词往往不会在句子层面共同出现。Huang等[11]利用连词判断单词间的极性关系，并结合单词形态上的否定形式，构建情感极性约束矩阵，再利用逐点互信息，判断单词的情感极性。杨春明等[12]使用逐点互信息来反映词语间的相关关系，并用非负矩阵分解（NMF）的方法来构建语料中情感词语之间、情感词语与评价对象之间的关系矩阵，然后利用此关系矩阵结合词语的语义、语素关系构建图模型来构造情感词典。钟敏娟等[13]首先利用关联规则挖掘算法抽取与识别体现领域特征的情感词，然后基于PageRank模型和混合相关关系判别情感词极性。

目前，使用深度学习的方法构建情感词典已经成为一种趋势。杨小平等[14]利用Word2Vec工具从大规模中文语料中提取词向量，研究情感类别划分并选取种子词，基于转换约束集得到候选词的情感极性和情感强度，得到多维汉语情感词典SentiRuc。王仁武等[15]结合Word2Vec词向量技术构建产品特征词和情感词词库，进一步构造情感概念对情感评分，并将其用于分析品牌产品特定特征的用户情感。胡家珩等[16]利用词向量方法将文本信息映射到向量空间，借助已有的通用情感词典，自动标引训练语料，使用Python构建深度神经网络分类器，判断特定领域候选情感词的情感极性，构建情感词典。

1.2 基于语义知识库的方法

基于语义知识库的方法，是指在已有专家标注词典的基础上，利用词语之间的词义联系（如同义词、反义词等）来计算词语的情感极性。Kamps等[17]假设同义词具有相同的极性，并将同义词库提供的同义词连接起来构建词汇网络，词语极性通过网络中与种子词（“好”和“坏”）的距离来确定。Hu等[18]扩展了Kamps的方法，利用WordNet词典构建情感词典，不仅使用了同义词关系，而且考虑了反义词的作用。Liu等[19]基于Open Mind Commonsense数据库识别基本情感，并将其分为高兴、悲伤、愤怒、恐惧、厌恶和惊奇6个基本类别。Lu等[20]利用同义词词林和双语词典构建词汇图，然后使用半监督图模型从种子词中得到更多的正面及负面情感词。周咏梅等[21]提出基于HowNet和SentiWordNet的情感词典构建方法，将中文词语进行义元分解得到对应的英文义元，再通过SentiWordNet计算义元的情感倾向值，分别得到中文词语的正面、负面情感倾向值。衣丽霞等[22]将Hu的方法进行了改进，基于词典WordNet3.0，提出POAE算法自动扩展极性副词，除了同义关系和反义关系，还使用了WordNet词典中的近义关系和又见关系。

作为情感分析的重要工具之一，情感词典目前在网络文本情感分析中得到较好应用，但在突发公共事件的情感分析中，该方法还处于探索阶段。同时，有些情感词在不同领域具有不同的情感倾向，甚至在同一领域，当修饰不同产品特征时也具有不同的情感倾向[23]。因此，构建面向突发公共事件网络舆情分析的领域情感词典，并将其运用于网络舆情分析中，有助于提升突发公共事件网络舆情的监督和应对能力。

2 研究设计与流程

本文设计的领域情感词典构建流程主要分为四步。第一步，构建突发公共事件网络舆论语料库。语料库包括突发公共事件的微博评论语料和新闻评论语料。第二步，构建自定义基础词典。词典主要包括现有基础情感词典、网络流行词、领域词等。同时，结合自定义基础词典对语料库中的数据进行预处理，主要包括分词和词性标注。第三步，构建情感种子词典。基于现有基础情感词典，对突发公共事件网络舆论语料中的数据进行情感词匹配，并对相关情感词进行修正，形成情感种子词典WordSet1。第四步，情感词扩充及领域情感词典的构建。基于Word2Vec模型和余弦相似度算法，对种子情感词典WordSet1进行近义词扩充，并对新词进行情感类型和强度标注，形成最终的领域情感词典WordSet。

2.1 突发公共事件网络舆论语料库的构建

人民网舆情监测室发布的《2015年互联网舆情报告》指出“两微一端”（微博、微信、移动客户端）成为很多中国人了解新闻时事的第一信息源[24]。由于微信朋友圈数据私密性较强，难以采集，本文主要采集新浪微博评论数据和移动客户端新闻评论数据，作为情感词识别和匹配的语料来源。首先，根据国务院制定的《国家突发公共事件总体应急预案》中对突发公共事件的分类，将突发公共事件分为自然灾害、事故灾害、公共卫生和社会安全四类[25]。然后，基于2011—2017年《中国社会舆情与危机管理报告》，为每种类型的突发公共事件选取相应检索词，见表1。

表1 四类突发公共事件检索词选取

事件类型	检索词
自然灾害	洪灾、台风登陆、雪灾、特大暴雨灾害、沙尘暴袭击、雾霾“穹顶之下”、地震、山体滑坡、泥石流灾害、风暴潮灾害、森林火灾
事故灾害	煤矿爆炸、沉船事故、危险化学品爆炸、建筑施工事故、公交车事故、校车事故、动车事故、踩踏事件、辐射事故、纵火、输油管道爆炸
公共卫生	传染病、感染病毒、问题疫苗、毒奶粉、禽流感、毒胶囊、瘦肉精、地沟油、苏丹红
社会安全	暴恐、坠楼、虐童、砍杀学生、贪腐、涉毒、逃税、传销

如表1所示，自然灾害事件选取的检索词有洪灾、台风登陆、地震等；事故灾害事件选取的检索词有煤矿爆炸、沉船事故、公交车事故等；公共卫生事件选取的检索词有传染病、问题疫苗、毒奶粉等；社会安全事件选取的检索词有暴恐、虐童、逃税等。

一方面，以新浪微博为采集平台，以四类突发公共事件的检索词作为关键词对原创微博进行检索，爬取每种类型突发公共事件的原创微博信息，经校对筛选后，得到突发公共事件相关联的原创微博共计42 020条；然后对微博评论进行采集，采集时间为2020年1月29日—2月10日。对含有网址链接、无效评论等影响情感分析的内容进行删除后，微博评论语料库共包含841 128条微博评论数据。另一方面，以四类突发公共事件的检索词作为关键词，对来自搜狐新闻、网易新闻、腾讯新闻、百度新闻、凤凰新闻五家知名移动新闻客户端的新闻数据进行检索和采集，经校对后共采集突发公共事件相关新闻86 571条；然后对新闻评论进行采集，共采集新闻评论838 016条，采集时间为2020年2月5—16日。剔除重复评论、网址链接等，新闻评论语料库共包含704 155条新闻评论数据。

2.2 自定义基础词典的构建及语料库预处理

为了提高对语料库文本进行分词处理的准确性，且能结合突发公共事件舆论语料构建情感种子词，本文首先构建一个囊括基础情感词典、突发公共事件领域词典、网络新词和流行词典的自定义基础词典。其中，基础情感词典选择大连理工大学情感词汇本体库中的27 466个情感词，该词典将情感分为7大类21小类，情感强度分为1、3、5、7、9五档（9表示强度最大），能够满足深入分析情感类型的需求。

突发公共事件领域词、网络流行词和网络新词通过搜狗输入法细胞词库进行选取。搜狗输入法的新词词典和领域词典涵盖了自然灾害、事故灾害、公共卫生、社会安全方面的术语用词。通过对搜狗输入法细胞词库进行格式处理，将scel格式转化为txt后，选取其中的气象灾

害词库、公安词库、环保词库、交通事故处理词库、传染病词库等，形成领域词典。接着，将搜狗输入法细胞词库中的《网络流行语》和《网络流行新词》两种词库导入自定义词典，形成网络新词和流行词典。

为了提高情感词典构建的准确度，需要对网络舆论语料进行预处理。中国科学院计算技术研究所开发的NLPIR分词工具能够从较长的文本内容中，基于信息交叉熵自动发现新特征语言，并自适应测试语料的语言概率分布模型，实现自适应分词，功能强大[26]。因此，将构建好的自定义基础词典导入NLPIR分词工具中，对语料库进行分词处理，并标注词性。

2.3 情感种子词典的构建

首先，结合网络舆论语料库对基础情感词典中的情感词进行识别与修正。网络舆论语料经过上述预处理后，共识别出91 656个词语。将识别出的词语与基础情感词典中的情感词相匹配，共匹配到9 837个情感词，出现在300 723条舆论文本中。

利用基础情感词典，即大连理工大学情感词汇本体库，对现有的9 837个情感词进行情感分类和强度标注。情感词的属性主要有4个：情感分类、词性、强度、极性。在突发公共事件网络舆论语料库中，有些词语的情感分类与大连理工大学情感词汇本体库并不相同，如“呵呵”在大连理工大学情感词汇本体库中被划分为褒义词汇，但是它在大部分舆论文本中表示贬义，如“偷工减料、短斤少两、以次充好，呵呵，奸商不管干哪个行业都是同样的套路”“呵呵，这种毒瘤难道不应该被枪毙吗”。因此，还需结合舆论文本的语境，对上述情感词的情感分类和极性进行人工修正。人工修正方面，将每个情感词所在的舆论文本随机分配给2位不同的标注者，当标注结果相同时，将标注结果保存到数据库中；当标注结果不同时，把舆论文本分配给第3位标注者，然后选择多数一致的标注结果。表2列出了情感词“骄傲”的情感分类与极性进行人工修正的详细过程。

表2 情感分类与极性的人工修正示例

步 骤	步骤说明	处理结果
1	情感词汇本体库标注	骄傲[ND，adj，5，2]
2	舆论原文	牺牲的官兵更应得到抚恤与尊重，他们才是祖国的骄傲！
		逝者如斯夫，英雄一路走好，你们每一个人都是中国的骄傲。
		你们是中华的好儿郎，人民的骄傲，致敬英雄！
3	分词结果	牺牲/v的/u官兵/n更/d应/v得到/v抚恤/v与/c尊重/v，/wp他们/r才/d是/v祖国/n的/u骄傲/a！/wp
		逝者/n如/v斯夫/n，/wp英雄/n一路/m走/v好/a，/wp你们/r每/r一个/m人/n都/d是/v中国/ns的/u骄傲/a。/wp
		你们/r是/v中华/nz的/u好/a儿郎/n，/wp人民/n的/u骄傲/a，/wp致敬/v英雄/n！/wp
4	情感修正	骄傲[PH，adj，5，1]

大连理工大学情感词汇本体库中将情感强度分为1、3、5、7、9五档，将情感极性分为0、1、2，其中0代表中性，1代表褒义，2代表贬义，使用该词典进行情感分析时，过程较为复杂。因此，为便于进一步进行文本情感值计算，本文将通过情感词汇本体库标注的情感词的强度与极性相结合，将极性标注为2的情感词的情感强度用负数表示，将极性标注为1的情感词的情感强度用正数表示，对于极性标注为0的情感词，则结合情感词所在语境人工修正划分了词语的褒贬倾向，故不再有中性词。基于此，本文构建的词语情感强度共划分为十档，即情感极性强度集S={−9，−7，−5，−3，−1，1，3，5，7，9}，分别是贬义（高、中、低）、褒义（低、中、高），数值的绝对值表示强度级别。在进行情感强度判断时，有些词语包含两种情感倾向，为更加准确地进行情感分析，选取情感强度大的情感倾向作为主要情感。如“坚守”包含“尊敬”和“赞扬”两种情感，所以分别在两个情感的相应分量上用5和7表示。对于“坚守”来说，在“尊敬”上的等级为5，在“赞扬”上的等级为7，表明主要情感是赞扬，其情感强度为7。

利用以上规则，经过修正，将最终得到的情感词集定义为情感种子词典WordSet1。最终得到情感种子词7 697个。每一个情感种子词都由以下三元组进行表示，即WordEmo（W_i）= [C_i，N_i，S_i]。

其中，W_i为情感种子词；C_i为所属情感类别，

该类别参照大连理工大学情感词汇本体库将情感分为乐（PA、PE）、好（PD、PH、PG、PB、PK）、怒（NA）、哀（NB、BJ、NH、PF）、惧（NI、NC、NG）、恶（NE、ND、NN、NK、NL）、惊（PC）7大类21小类；N_i为情感词词性，即名词（noun）、动词（verb）、形容词（adj）、副词（adv）、网络词语（nw）、成语（idiom）、介词短语（prep）；S_i为情感强度，即S_i={−9，−7，−5，−3，−1，1，3，5，7，9}。表3列出部分情感种子词及其极性强度编码。

表3 部分情感种子词及其极性强度

情感种子词	极性强度编码	情感种子词	极性强度编码
如意	[PA，adj，3]	问心有愧	[NH，idiom，−5]
踏实	[PE，adj，7]	恐惧	[NC，adj，−5]
温暖	[PB，adj，5]	羞耻	[NG，adj，−5]
尊崇	[PD，verb，9]	鄙夷	[ND，verb，−7]
相信	[PG，verb，7]	伤天害理	[NN，idiom，−9]
悲愤	[NA，adj，−7]	质问	[NL，verb，−9]
绝望	[NJ，adj，−9]	不可思议	[PC，idiom，5]

2.4 领域情感词扩展

为了丰富情感词典，解决数据稀疏问题，采用Word2Vec进行情感词扩展。Word2Vec是Google在2013年推出的一款用于训练词向量的工具，其原理是基于深度学习算法，通过训练，可以把对文本内容的处理转换为K维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似度[27]。

本文采用Python的gensim模块提供的Word2Vec工具包进行训练[28]。训练过程中，本文采用CBOW模型将处理后的舆论语料构建词向量，词向量维度size设定为100，词语近邻窗口window设定为5，采用Hierarchical Softmax算法，即hs设定为1，计算词向量的最小词频min_count为3。Word2vec计算的是余弦值，距离范围为0~1，值越大代表两个词关联度越高，其计算过程如公式（1）所示。

$$\cos(w1, w2) = \frac{\sum_{i=1}^{n} w1_i w2_i}{\sqrt{\sum_{i=1}^{n} w1_i^2}\sqrt{\sum_{i=1}^{n} w2_i^2}} \quad (1)$$

其中，$w1$，$w2$分别表示两个词或词组，利用Word2Vec将词映射成n维向量，n表示维度数，$w1_i$与$w2_i$分别表示第i个维度上的取值。

新增情感词的极性强度的判断主要是通过计算候选词与基准词语的语义相似度来确定，上文中已得出候选词与基准词语之间的余弦距离，其夹角余弦值越大，候选词是新情感词的概率就越大。新增情感词极性强度的确定，如公式（2）所示。

$$SentiScore(word) = \max[\frac{1}{N_j}\sum_{set_p \in set_j}\cos(word, set_p)] \quad (2)$$

其中，$word$表示新增情感词，set_j表示第j类情感的种子词集合，set_p表示第j类情感种子词集合set_j中的第p个情感词，N_j（$1 \leqslant j \leqslant 21$）表示第$j$类情感种子词集$set_j$中种子词的数量。然后按照$SentiScore$值进行排序，新情感词类型及其极性强度的确定取决于其最大$SentiScore$值基准词语的极性强度。

构建面向突发公共事件网络舆情分析的领域情感词典的步骤：①应用初始构建的情感种子词典WordSet1中的情感词作为基准词语得到词语W的向量表示，并将其存入vector.bin文件中；②如果能在情感基准词典WordSet1中找到词语W，则可直接跳入步骤⑤，标注W的情感极性强度，否则，跳入步骤③；③在Word2Vec中执行“./distance vector.bin”，在突发公共事件舆论语料库中查找与词语W最接近的10个词作为候选词，其阈值设定为0.7，相似度大于0.7的候选词作为新情感词；④用公式（2）计算新情感词的极性强度；⑤将W存入面向突发公共事件网络舆情分析的领域情感词典中。

最终共识别出未在情感词汇本体库中收录的新增情感词2 604个。本文的情感词典共分为7大类、21小类，情感强度S_i={−9，−7，−5，−3，−1，1，3，5，7，9}，含有情感词共计10 301个。其各情感类别中包含的情感词个数及代表性词语，见表4。

如表4所示，一些人类的基本情感，如快乐、喜爱、悲伤、烦闷、憎恶，是包含情感词较多的几种情感。另外，在本文构建的情感词典中，赞扬、贬责包含的情感词最多，说明面对突发公共事件，民众在宣泄内心不满的同时也会传播正能量。对于构建出的情感词典，本文采用改进的TF-IDF方法对各类突发公共事件中出现权重较高的情感特征词进行统计，其计算过程如公式（3）、公式（4）所示。

表4 突发公共事件舆论7类情感词举例

一级类	二级类	编 码	情感词个数	情感词及其三元组举例
乐	快乐	PA	583	阳光[adj，PA，5]；哈哈哈[nw，PA，7]；幸福[adj，PA，7]
	安心	PE	225	安好[adj，PE，1]；问心无愧[adj，PE，3]；知足[adj，PE，5]
好	尊敬	PD	332	佩服[verb，PD，3]；孝敬[verb，PD，3]；尊重[verb，PD，5]
	赞扬	PH	3 113	公平[adj，PH，5]；赞赏[verb，PH，5]；有良心[adj，PH，7]
	相信	PG	252	肯定[verb，PG，3]；信奉[verb，PG，5]；公认[verb，PG，7]
	喜爱	PB	399	重视[verb，PB，3]；稀罕[adj，PB，5]；珍惜[verb，PB，7]
	祝愿	PK	106	希望[verb，PK，5]；立志[verb，PK，5]；保佑[verb，PK，5]
怒	愤怒	NA	140	我靠[nw，NA，−5]；你丫[nw，NA，−5]；活该[adj，NA，−3]
哀	悲伤	NB	454	不幸[noun，NB，−9]；毒瘤[noun，NB，−7]；可惜[adj，NB，−5]
	失望	NJ	106	死心[adj，NJ，−9]；一蹶不振[idiom，NJ，−7]；遗憾[adj，NJ，−5]
	疚	NH	44	悔恨[adj，NH，−9]；道歉[verb，NH，−5]；内疚[adj，NH，−3]
	思	PF	72	缅怀[verb，PF，7]；惦念[verb，PF，7]；牵肠挂肚[idiom，PF，9]
惧	慌	NI	119	慌乱[adj，NI，−7]；惊慌[adj，NI，−7]；魂不守舍[idiom，NI，−7]
	恐惧	NC	248	恐慌[adj，NC，−7]；恐怖[adj，NC，−5]；可怕[adj，NC，−3]
	羞	NG	50	丢人[adj，NG，−7]；羞愧[adj，NG，−5]；难为情[adj，NG，−3]
恶	烦闷	NE	476	崩溃[verb，NE，−9]；TMD[nw，NE，−9]；发牢骚[verb，NE，−3]
	憎恶	ND	611	无耻[adj，ND，−9]；该死[adj，ND，−7]；恶心[adj，ND，−5]
	贬责	NN	2 828	禽兽[noun，NN，−9]；狗屁[noun，NN，−7]；SB[nw，NN，−5]
	妒忌	NK	10	嫉妒[adj，NK，−7]；吃醋[verb，NK，−5]；眼馋[verb，NK，−5]
	怀疑	NL	52	怀疑[verb，NL，−9]；将信将疑[idiom，NL，−7]；质疑[verb，NL，−5]
惊	惊奇	PC	81	奇怪[adj，PC，5]；不料[verb，PC，7]；目瞪口呆[idiom，PC，9]

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}-n_{i,j}} \tag{3}$$

$$idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|} \tag{4}$$

其中，$n_{i,j}$表示情感词i在文档d_j中出现的次数，$\sum_k n_{k,j}-n_{i,j}$表示情感词i在其他文档中所有词语的出现次数之和，$|D|$表示语料库中的文档总数，$|\{j:t_i\in d_j\}|$表示包含词语t_i的文档数目，如果该词语不在语料库中，就会导致公式没有意义，因此一般情况下使用$|\{j:t_i\in d_j\}|+1$，然后$\text{TF-IDF}=tf_{i,j}\times idf_i$。各类突发公共事件中TF-IDF值排名前五位的褒义及贬义情感词见表5。

表5 各类突发公共事件情感词示例

事件类型	褒 义	贬 义
自然灾害	公祭；反腐倡廉；驰援；秉公；雄风	铺张浪费；求全责备；吹毛求疵；凶多吉少；伤痕
事故灾害	告慰；防微杜渐；雷厉风行；敬爱；关爱	无序；偷窃；欺瞒；痛惜；傻眼
公共卫生	施恩；耿直；治病救人；无毒；庇护	病毒；做假；变质；失效；谋财害命
社会安全	无微不至；大吃一惊；生机；清官；福分	逃税；窝点；诬告；猥亵；偷税

如表5所示，各类突发公共事件情感词具有较强的领域性特点，自然灾害类事件特有的褒义情感词有公祭、反腐倡廉、驰援等，贬义情感词有铺张浪费、求全责备、吹毛求疵等；事故灾害类事件特有的褒义情感词有告慰、防微杜渐、雷厉风行等，贬义情感词有无序、偷窃、欺瞒等；公共卫生事件特有的褒义情感词有施恩、耿直、治病救人等，贬义情感词有病毒、做假、变质等；社会安全类事件特有的褒义情感词有无微不至、大吃

一惊、生机等，贬义情感词有逃税、窝点、诬告等。

3 实验分析

3.1 测试集的构建与标注

为了检验本文所构建领域情感词典在识别情感词方面的效果，本文选择新冠肺炎疫情事件作为研究案例。新冠肺炎疫情，是新中国成立以来在我国发生的传播速度最快、感染范围最广、防控难度最大的一次重大突发公共卫生事件[29]。2019年12月31日，武汉市卫健委发布通告称近期部分医疗机构发现接诊的多例肺炎病例与华南海鲜市场有关联，引发了较为广泛的社会关注。2020年1月20日，钟南山院士指出“新型冠状病毒具有传染性，已经出现人传人现象”，成为微博热议话题，网民讨论热度不断升高。2020年1月22日，国务院新闻办公室举行新闻发布会。1月23日，湖北省人民政府新闻办公室举行新闻发布会，介绍新冠肺炎防控工作的有关情况，舆情不断升温。

针对新冠肺炎疫情事件，选取新型冠状病毒、新冠肺炎作为主题词。爬取的时间段为2020年1月1日—3月31日。此时间段在微博大V的转发和大量有关问责主管部门的舆情推动下，网民讨论、转发活跃度极高。采集以上时间段期间与该事件相关的热门微博及其评论微博，共计72 497条，形成测试语料库。

经过预处理和数据清洗后，随机选取其中的5 000条文本进行实验。本文采用三人独立标注法，识别文本中的情感词，为使标注结果有效，只有当3个人的标注结果一致时才将标注结果输出。通过人工标注，在给定语料的5 000条文本中，共有词元13 211个，其中标注为情感词的有2 080个。词典判定方面，利用上文中提及的HowNet情感分析用词语集、台湾大学自然语言处理实验室构建的NTUSD词典、大连理工大学信息检索研究室的情感词汇本体库3个通用词典和本文2.4节得到的情感词典，对5 000条文本的情感词进行情感标注。

3.2 实验指标

为验证本文所构建的情感词典的有效性，需采用合适的指标对词典进行评价。情感分析中常用的评价指标有准确率（Precision）、召回率（Recall）和$F1$值（F1-measure）[30]。准确率（P）计算过程如公式（5）所示，召回率（R）计算过程如公式（6）所示，$F1$值计算过程如公式（7）所示。

$$P=\frac{n1}{n2}\times 100\% \tag{5}$$

$$R=\frac{n1}{n3}\times 100\% \tag{6}$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\% \tag{7}$$

其中，公式（5）中$n1$表示正确判断出情感极性的词语数，即被词典和人工标注一致的词语数，$n2$表示被词典识别出情感极性的词语数；公式（6）中$n3$表示舆论文本中识别出情感极性的词语数。将实验结果分别带入上式，即可计算出准确率（P）、召回率（R）和$F1$值。

3.3 实验结果

采用准确率（P）、召回率（R）、$F1$值3个评估指标评估采用本文构建的情感词典的性能，经计算结果见表6。只有当采用本文构建的面向突发公共事件网络舆情分析的领域情感词典在准确率（P）与召回率（R）的得分上优于上文中提及的HowNet情感分析用词语集、台湾大学自然语言处理实验室构建的NTUSD、大连理工大学信息检索研究室的情感词汇本体库时，方可认为该情感词典符合突发公共事件网络舆情分析的要求。

表6 各词典情感分类效果性能评估

	准确率（P）	召回率（R）	$F1$值
NTUSD词典	0.56	0.65	0.60
HowNet情感分析用词语集	0.69	0.72	0.70
情感词汇本体库	0.74	0.81	0.77
领域词典	0.85	0.90	0.87

从表6可以看出，本文构建的情感词典进行情感判别的准确率为0.85，召回率为0.90，$F1$值为0.87。在突发公共事件舆论文本的情感识别中，本文构建的情感词典的表现要优于3个通用词典。所以，总体看来，本文中提出的领域情感词典构建方法具有较高的准确性和可利用性。

4 结语

本文提出了一种面向突发公共事件网络舆情分析的领域情感词典构建方法，该方法充分利用语料库和语义知识库的优点，在大规模网络舆论语料的基础上结合现有情感词典进行种子词提取，通过深度学习中的Word2Vec模型训练词向量，进行情感词的扩展，并根据语义相似度计算获得候选情感词，从而生成领域情感词典。通过准确率和召回率验证，本文提出的构建方法具有较好的准确性和可靠性。这种情感词典的构建方法同样也可以推广应用于其他领域情感词典的构建。

不过，本研究还存在一定的不足。为了保证所构建情感词典的准确性，本研究在种子词构建、情感词扩展和新增情感词强度判断过程中都加入了人工判别，由于文本情感表达的不确定性，人工判断文本情感也难免会有偏差，未来可结合多种语境和专家判别进行情感词类型和强度的修正。此外，用户评论中的表情符号也影响情感类别的判定，未来的研究可结合表情符号进行情感类型的判定。突发公共事件类型多样，不同的事件会有不同的情感表达特征，后续研究需要进一步考虑特定事件情感表达特征的识别。

参考文献

[1] 王科，夏睿. 情感词典自动构建方法综述 [J]. 自动化学报，2016，42（4）：495-496.

[2] 知网情感分析用词语集 [EB/OL]. [2020-01-06]. http://www.keenage.com/html/c_index.html.

[3] 台湾大学自然语言处理实验室. NTUSD [EB/OL]. [2020-01-08]. http://nlg.csie.ntu.edu.tw/.

[4] 徐琳宏，林鸿飞，潘宇，等. 情感词汇本体的构造 [J]. 情报学报，2008，27（2）：180-185.

[5] 邓淑卿，李玩伟，徐健. 基于句法依赖规则和词性特征的情感词识别研究 [J]. 情报理论与实践，2018，41（5）：137-142.

[6] 蒋翠清，郭轶博，刘尧. 基于中文社交媒体文本的领域情感词典构建方法研究 [J]. 数据分析与知识发现，2019，3（2）：98-107.

[7] 郭顺利，张向先. 面向中文图书评论的情感词典构建方法研究 [J]. 现代图书情报技术，2016，32（2）：67-74.

[8] HATZIVASSILOGLOU V，MCKEOWN K. Predicting the semantic orientation of adjectives [J]. Proceedings of the Acl，1997：174-181.

[9] TURNEY P D，LITTMAN M L. Measuring praise and criticism [J]. ACM Transactions on Information Systems，2003，21（4）：315-346.

[10] GAMON M，AUE A. Automatic Identification of Sentiment Vocabulary：Exploiting Low Association with Known Sentiment Terms [C] //Proc of ACL Workshop on Feature Engineering for Machine Learning in NLP. Michigan，USA，2005：57-64.

[11] HUANG S，NIU Z，SHI C. Automatic construction of domain specific sentiment lexicon based on constrained label propagation [J]. Knowledge Based Systems，2014（56）：191-200.

[12] 杨春明，张晖，何天翔，等. 具有共现关系的中文褒贬词典构建 [J]. 计算机工程与应用，2016，52（9）：164-169.

[13] 钟敏娟，万常选，刘德喜. 基于关联规则挖掘和极性分析的商品评论情感词典构建 [J]. 情报学报，2016，35（5）：501-509.

[14] 杨小平，张中夏，王良，等. 基于Word2Vec的情感词典自动构建与优化 [J]. 计算机科学，2017，44（1）：42-47，74.

[15] 王仁武，宋家怡，陈川宝. 基于Word2vec的情感分析在品牌认知中的应用研究 [J]. 图书情报工作，2017，61（22）：6-12.

[16] 胡家珩，岑咏华，吴承尧. 基于深度学习的领域情感词典自动构建——以金融领域为例 [J]. 数据分析与知识发现，2018，2（10）：95-102.

[17] KAMPS J，MARX M，MOKKEN R，et al. Using WordNet to Measure Semantic Orientations of Adjectives [C] //Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon，Portugal，2004：1115-1118.

[18] HU M，LIU B. Mining and summarizing customer reviews [C] // Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York，2004：168-177.

[19] LIU H，LIEBERMAN H，SELKER T. A Model of Textual Affect Sensing Using Real-World Knowledge [C] //International Conference on Intelligent User Interfaces. 2003：125-132.

[20] LU B，SONG Y，ZHANG X，et al. Learning Chinese polarity lexicons by integration of graph models and morphological features [C] //Proceedings of the 6th Asia Information Retrieval Societies Conference. Taipei，2010：466-477.

[21] 周咏梅，杨佳能，阳爱民. 面向文本情感分析的中文情感词典构建方法 [J]. 山东大学学报（工学版），2013，43（6）：27-33.

[22] 衣丽霞，王辉，籍晓红. 情感分析中极性副词的自动扩展 [J]. 计算机应用研究，2013，30（7）：1958-1960，1970.

[23] 郁亚辉. 产品评论中领域情感词典的构建[J]. 中文信息学报，2016，30（5）：137.

[24] 人民网舆情监测室. 2015年互联网舆情分析报告[EB/OL]. [2020-01-18]. http://yuqing.people.com.cn/GB/392071/401685/index.html.

[25] 中华人民共和国中央人民政府. 国家突发公共事件总体应急预案[EB/OL]. [2020-01-22]. http://www.gov.cn/yjgl/2006-01/08/content_21048.htm.

[26] 大数据搜索与挖掘实验室. NLPIR[EB/OL]. [2020-02-23]. http://ictclas.nlpir.org/.

[27] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[J]. Computer Science，2013（1）：28-36.

[28] gensim. models. word2vec-Word2vec embeddings[EB/OL]. [2020-03-07]. https://radimrehurek.com/gensim/models/word2vec.html.

[29] 中华人民共和国中央人民政府. 习近平：在统筹推进新冠肺炎疫情防控和经济社会发展工作部署会议上的讲话[EB/OL]. [2020-03-19]. http://www.gov.cn/xinwen/2020-02/24/content_5482502.htm.

[30] 刘兵. 情感分析：挖掘观点、情感和情绪[M]. 北京：机械工业出版社，2017.

作者简介

李长荣，女，1995年生，硕士研究生，研究方向：网络舆情分析，E-mail：17865919378@163.com。
纪雪梅，女，1985年生，博士，副研究馆员，研究方向：信息计量与网络舆情分析。

Construction of Domain Sentiment Lexicon for Online Public Opinion Analysis in Public Emergencies

LI ChangRong JI XueMei
(Institute of Scientific & Technical Information, Shandong University of Technology, Zibo 255049, China)

Abstract: In order to analyze the public sentiment in online public opinion of the public emergencies, this paper constructs a domain sentiment lexicon for online public opinion analysis. Firstly, based on the existing general sentiment lexicon, the emotional words are identified and corrected through the large-scale public opinion corpus. The emotional words are divided into 7 major categories and 21 subcategories, and they are marked with polarity and intensity to construct a seed lexicon. Then, the Word2Vec model and cosine similarity algorithm are used to expand the number of emotional words on the basis of emotional seed dictionary. Thirdly, the classification, polarity and intensity of new sentiment words are marked, and a domain sentiment lexicon is constructed. Finally, the microblog comments of the COVID-19 was selected as the corpus for experimental verification. The precision of the dictionary constructed in this paper is 0.85, the recall is 0.90, and the F1-measure is 0.87, which can be effectively used to identify the type and intensity of emotions in the online public opinion of the public emergencies.

Keywords: Public Emergencies; Sentiment Lexicon; Online Public Opinion; Word2Vec

（收稿日期：2020-07-20）