基于 TextCNN 的地震新闻标题分类方法

白 灵[1,2] 黄 猛[1,2] 刘 帅[1,2]
BAI Ling HUANG Meng LIU Shuai

摘 要 当破坏性地震发生后，大量信息产出，网上的地震新闻信息更是大量汇集。高效精准地识别与自动分类地震新闻，可使地震应急部门及时搜集各方面的应急态势，缓解面对海量新闻的压力，减少获取信息的时间成本。首先论述了地震新闻标题数据集的建设，然后实验对比分析了深度学习模型对地震新闻标题文本的分类效果。实验表明，采用 Word2vec 进行文本表示的 TextCNN 分类模型效果比较好，准确率达到了 92.03%。

关键词 地震新闻标题；文本分类；数据集建设；TextCNN

doi：10.3969/j.issn.1672-9528.2024.08.021

0 引言

我国地震频发，且强度大、范围广，其造成的经济损失及人员伤亡居于自然灾害的首位。地震发生后，如果能在黄金救援 72 h 内做出恰当的应急响应，则可降低各种损失、减少人员伤亡。若想做出恰当的响应，需要对各种信息进行综合的研判。突发地震事件信息量巨大，应急人员面对大量的信息，若能在短时间内对各类信息进行分类、分析及研判，则可提升救援效率。

新闻标题蕴含着新闻的关键信息，新闻标题短文本的自动化分类为地震应急部门快速获取及分类地震新闻信息提供了强有力的手段。本文介绍了地震新闻标题数据集建设、模型研究及改进、对比实验及分析等。

1 地震新闻标题数据集建设

针对地震新闻标题带标签数据集建设问题，本文做了以下工作。①获取了 1900 年至 2023 年地震目录中的震中位置信息，分析了地震新闻中和地震应急信息相关的关键字，应

1. 防灾科技学院 河北三河 065201
2. 河北省高校智慧应急应用技术研发中心 河北三河 065201
[基金项目] 廊坊市 2022 年科学技术研究与发展计划项目（2022011029）

参考文献：
[1] 陈婕 , 刘纪平 , 徐胜华 . 增强边缘信息的全卷积神经网络遥感影像建筑物变化检测 [J]. 测绘通报 ,2023(6):61-67.
[2] 柳思聪 , 都科丞 , 郑永杰 , 等 . 人工智能时代的遥感变化检测技术 : 继承、发展与挑战 [J]. 遥感学报 ,2023,27(9):1975-1987.
[3] 徐志红 , 关元秀 , 王善华 , 等 . 融合对象影像分析和 OCNN 的耕地变化检测 [J]. 遥感信息 ,2022,37(5):15-22.
[4] 姚沐风 , 昝露洋 , 李柏鹏 , 等 . 基于 CAR-Siamese 网络的高分辨率遥感图像建筑物变化检测 [J]. 中国科学院大学学报 , 2023, 40(3):380-387.
[5] 王佳 , 刘锦秀 , 李晓民 , 等 . 深度学习支持下的智能化信息提取技术在青海省自然资源变化监测中的应用 [J]. 青海国土经略 , 2021(5):61-67.
[6] 董志鹏 . 基于卷积神经网络的高分辨率遥感影像目标检测方法研究 [J]. 测绘学报 ,2023,52(9):1613.
[7] 毕卫华 , 杨化超 . 基于深度学习和 Cesium 的煤矿区地表变化检测方法研究 [J]. 能源技术与管理 ,2023,48(5):171-174.
[8] 孙斌 , 李俊鹏 , 罗哲轩 , 等 . 基于 GF-3 影像的金沙江堰塞湖电网受灾区域提取 [J]. 航天返回与遥感 ,2021,42(5):96-107.
[9] 麻连伟 , 宁卫远 , 焦利伟 , 等 . 基于 U-Net 卷积神经网络的遥感影像变化检测方法研究 [J]. 能源与环保 , 2022, 44(11): 102-106.
[10] 魏汝兰 , 王洪飞 , 盛森 , 等 . 基于深度学习的卫星影像耕地变化检测方法及系统应用 [J]. 软件导刊 ,2023(11):29-34.

【作者简介】
陈志兰（1974—），女，福建龙岩人，本科，高级工程师，研究方向：遥感影像与自然资源信息化研究和应用。

（收稿日期：2024-05-24）

用震中位置信息 + 地震应急关键字的组合方式作为地震新闻标题的爬取策略。②对爬取获取的原始地震新闻标题数据集进行了数据清洗。③分析了地震应急信息的分类方法，制定了地震新闻信息的分类标准。④依据分类标准对地震新闻标题数据集进行了人工打标签工作。⑤对地震新闻标题数据集进行了可视化分析。本文为垂直领域文本分类数据集的建设提供了思路。

1.1 数据获取及清洗

本文构建的地震新闻标题数据集信息来源有两种。

（1）百度搜索：主流媒体网站较多，本文尝试从百度上以关键字的形式进行检索新闻，这样会爬取到大多主流媒体网站的地震新闻信息。

（2）地震局官网：地震局官网的各个专栏的新闻，如国务院要闻、防震减灾要闻、媒体播报、行业动态、市县工作、政务公开等。

两种来源共爬取数据 20 多万条。机器去重后，共 15 万多条地震新闻标题数据，其中百度搜索 5 万条左右数据，地震局官网 10 万条左右。有一些数据，虽然机器认为是不重复的，但实际上是重复的信息，这类数据只能人工去掉，因此这类数据是在打标签时进一步人工去重的。新闻标题中有很多无用且没有意义的符号，如双引号、省略号、网络标签、中文中括号、其他特殊符号等。

1.2 地震新闻分类及人工打标签

灾害分类根据不同的考虑因素和用途可有多种不同的分类方法 [1]。如文献 [2] 以时间为主线，将地震应急信息分为震前基础背景信息、地震震情灾情信息及震后应急救援信息，实现了对多渠道汇集的文档类应急信息进行自动分类。文献 [3] 依据用户需求差异对地震应急信息进行分类、梳理，将地震应急信息分为基础背景信息、动态综合信息、地震现场指挥信息。不同的应急需求，地震灾情信息的分类方法也不尽相同。信息分类的基本方法有：线分类法、面分类法、混合分类法。文献 [4] 应用线分类法将地震灾情现象分为 4 个大类：房屋破坏现象、生命线工程破坏现象、地质灾害现象、人员伤亡。

地震事件新闻随着时间的推移，有明显的阶段性特征。地震发生后的第一时间往往都是和地震三要素相关的震情信息，接下来就是地震事件带来各类损失及人员伤亡情况，然后各级政府及民间组织的应急救援相关情况的报道等。地震新闻的这种阶段性特征和白仙富研究员提出的地震应急现场信息分类体系类似，该分类体系按照信息内容的本质属性进行分类，依据发生什么事件、产生什么影响、对产生的影响人们作出什么响应、针对响应有何评估或者说有哪些成效这样的思路对地震应急现场信息进行分类 [5]。白仙富研究员将地震应急现场信息分为地震震情信息、灾情信息、应急处置信息、处置效益信息 4 大类。本文结合地震新闻的特点把地震新闻信息分为震情信息、灾情信息、应急救援、其他地震信息 4 类。各类信息包含的详情如表 1 所示。

表 1 地震新闻信息分类

类别	包含信息	地震新闻标题中常见关键字
震情信息	地震测震部门发布的地震三要素信息（时间、震级、经纬度）、震源信息、台站背景数据、强震动台网观测数据	发生 ** 级地震，震源深度，有感，震感强烈
灾情信息	房屋建筑破坏、人员伤亡、生命线震害信息、次生灾害、余震信息、财产损失、烈度信息、地面形变和破坏、预警信息、地震发生时场景等	人员伤亡，受伤，死亡，遇难，流离失所，财产损失，废墟，房屋破坏，房屋倒塌，坍塌，断裂，列车出轨，列车停驶，灾害，灾情，大坝，水库，交通，通信，供水，供气，排水，停电，灾区，核电站，泥石流，滑坡，滚石，海啸，海啸预警，海水倒灌，崩塌，填塞，火灾，爆炸，污染，瘟疫，烈度，受伤，损坏，地震断裂，地裂，震陷，余震频发，现场直击
应急救援	灾民安置信息、应急救援决策信息、应急指挥协调信息、专家解读和研判信息、指挥部工作状态、救灾物资及人员调度信息、幸存者救助信息、生命线修复信息、次生灾害处置信息、建筑物房屋修复信息、地面形变及破坏处置信息、历史救援案例信息、捐助信息、地震谣言平息信息、保险公司理赔等	救援，搜救，救灾，获救，抢修，抗震救灾，赈灾，会商，搜救犬，撤离，疏散，物资运输，扣停列车，收治伤员，出院，救灾物资，捐款，捐赠，募捐，援助，谣言，应急响应，应急预案，驰援，赶赴震中，集结待命，昼夜奋战，复课，恢复运行，灾后重建，恢复用电，安置点，安置灾民，防疫，保险理赔，慰问，默哀，哀悼，埋葬，专家研判，抗震救灾发布会
其他地震信息	地震科普信息，应急演练信息，历史地震信息，地震感人故事回忆录，信息化建设信息，纪念活动，政策法规，政府规划，地震科学项目信息，政府部门及科研院所日常工作，行业动态等	科普，预警系统，科学研究，演习，演练，个人事迹，感人故事，见闻，回顾，启示，如今怎么样，纪念，人事“地震”，管理办法，条例，科普活动，知识讲座，科研项目

参照表 1 中的分类标准进行人工打标签。边打标签，边观察标题信息中包含的关键字，把相应的关键字加入表 1，同时也可以根据总结的关键字提升打标签的效率和准确率。每条数据至少经过两个人打标签。经交叉认证，如果两个人都给出同类标签，则保留；如果给出不同标签，则认为是争议数据。经多人确认，如果数据仍存在争议，则舍弃。地震局官网获取的数据大多为标签 4 的数据，数据量过多，仅电脑随机就选取了 8000 多条。最终，有效数据条数为 4.2 万条，标签 1 的数据为 1.6 万条。为了避免在深度学习过程中由于某一类数据比较多，产生数据倾斜现象，随机去掉了一部分。

1.3 数据集可视化分析

地震新闻标题数据集各类标签的数量为标签 1 数据条数为 10 008 条，标签 2 数据条数为 9613 条，标签 3 数据条数为 9259 条，标签 4 数据条数为 8130 条，共计 37 010 条。

四种数据的占比情况为标签 1 数据占比 27%，标签 2 数据占比 26%，标签 3 数据占比 25%，标签 4 数据占比 22%。各类标签数据占比比较均匀，可用于深度学习相应的监督算法。

从图 1 的新闻标题频度统计中可以看出，大部分新闻标题长度在 15 ～ 35 之间。从图 2 的新闻标题长度累积分布中可以看出长度从 1 ～ 40 的数据占了 99%，所以在本数据集上应用深度学习算法进行建模时，句子长度取 40 比较适合，即新闻标题长度不足 40 的，进行填充；高于 40 的，进行截断操作。

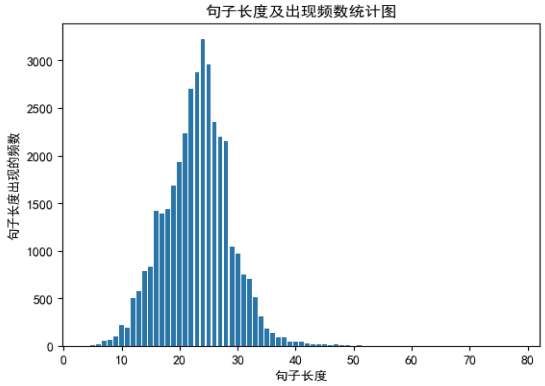

图 1 新闻标题长度频度统计

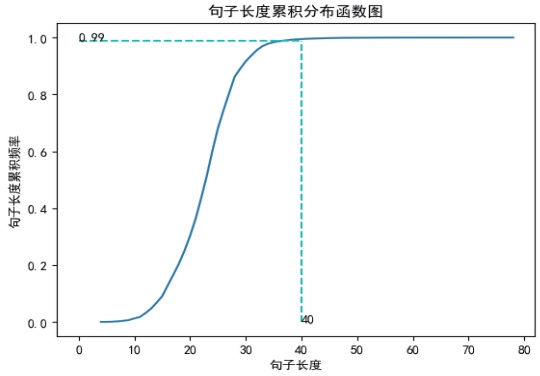

图 2 新闻标题长度累积分布

2 TextCNN 模型研究及改进

TextCNN 是一种基于 CNN 模型改进的适用于文本分类和文本特征提取的深度学习模型 [6]。本文改进的 TextCNN 模型总体框架如图 3 所示。

（1）文本表示：输入文本“上海地震了”经过 Embedding 层进行词嵌入形成文本矩阵。本文使用 Word2Vec 模型进行文本表示，此模型属于分布式文本表示中的一种 [7]。

（2）卷积层：分别使用 2*300、3*300、4*300 的卷积核对文本矩阵进行卷积得到特征向量，每种卷积核的数量为 64。

（3）池化层：卷积得到的特征向量经最大池化处理，提取主要特征值。

（4）全连接层：池化层特征值经连接得到 192 维特征向量，经全连接层并最终经 Softmax 函数进行分类输出。

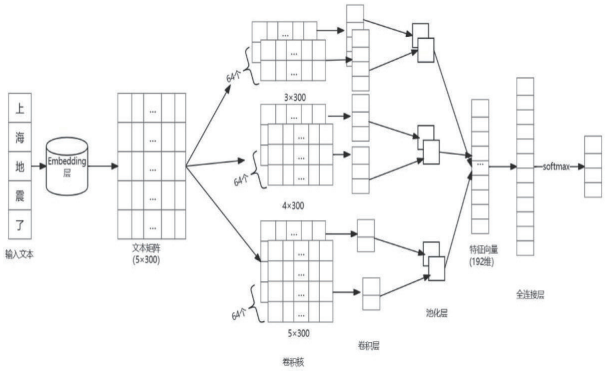

图 3 TextCNN 模型总体框架

3 对比实验分析

3.1 实验环境

本实验的深度学习框架为 PyTorch，使用的编程语言为 Python，版本为 3.7，操作系统为 Windows 10。运行平台的 CPU 为 Intel(R) Core(TM) i7-8550U CPU @ 1.80 GHz 2.00 GHz，内存 16 GB，GPU 为 NVIDIA GeForce MX150。

3.2 参数设置

本文的 Word2Vec 文本表示使用搜狗公布的预训练字向量 [8]，字向量维度为 300。模型参数和训练参数见表 2。实验数据集按 8:1:1 比例来划分训练集、验证集及测试集。

表 2 实验参数

参数名	参数值
优化器 optimitzer	Adam
损失函数 loss	Cross_entropy
学习率 learning_rate	0.001
迭代次数 epoch	10
置零率 dropout	0.1
激活函数 activation	Relu
批尺寸 bach_size	32
文本截断长度	40
词向量维度	300
卷积核尺寸	（2,3,4）

3.3 评价指标

评价指标是评价模型性能的标准，是数据分析中非常重要的部分。本实验采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F_1 值这 4 个评价指标对模型进行分析。评价指标中涉及参数有 TP、TN、FP 及 FN。T 代表判断正

确（True），F 代表判断错误，P 代表正样本（Positive），N 代表负样本（Negative）。

TP：正样本预测为正，判断正确。

TN：负样本预测为负，判断正确。

FP：负样本预测为正，判断错误。

FN：正样本预测为负，判断错误。

各个评价指标的计算如下。

准确率表示预测正确的样本数量与总样本数量的比例，其公式为：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

精确率表示正确预测正样本占实际预测为正样本的比例，其公式为：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

召回率表示正确预测正样本占正样本比例，其公式为：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

F_1 值综合了精确率和召回率，是 Precision 和 Recall 的加权调和平均。F_1 值越高，模型的性能越好，其公式为：

$$F_1 = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

3.4 实验结果与分析

为验证 TextCNN 模型的文本分类性能，做了两组对照实验。

组 1：分别采用随机初始化及 Word2Vec 模型对地震新闻标题进行字嵌入表示，分类模型采用 TextCNN。

组 2：采用 Word2Vec 模型对地震新闻标题进行字嵌入表示，分类模型分别采用 TextCNN 和 LSTM。

实验在迭代 10 次后，在验证集上测得的精确率、召回率、F_1 值、准确率如表 3 所示，评价指标采用 scikit-learn 包中的 classification_report 函数计算所得。

表 3 各模型实验效果对比 /%

Model	Precision	Recall	F_1	Accuracy
random-TextCNN	90.47	90.47	90.42	90.81
Word2Vec-TextCNN	91.70	91.76	91.70	92.03
Word2Vec -LSTM	90.84	90.75	90.74	91.19

从表 3 中的数据可知，各个模型的评价指标均可，基于 Word2Vec-TextCNN 分类模型各个方面均高出 1 个百分点左右，优于其他两个模型。

从图 4 训练过程看，Word2Vec-TextCNN 模型的收敛速度较快，高于其他两个模型。

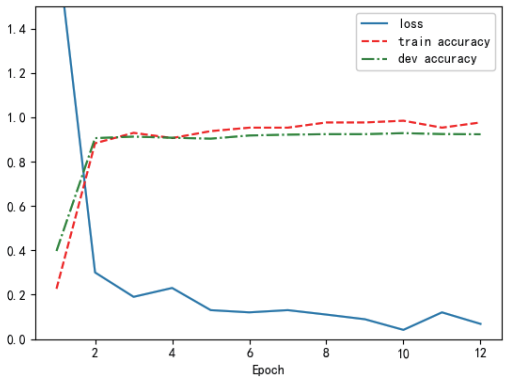

图 4 训练过程损失及精度变化情况

4 结语

Word2Vec-TextCNN 模型在地震新闻标题分类中四个评价指标均比较高，在垂直短文本分类领域有一定的参考价值。在以后的工作中，可以进一步尝试基于 Bert 模型的文本分类方法。

参考文献：

[1] 董旻，杨天青．地震应急灾情信息分类探讨 [J]. 震灾防御技术 ,2014,9(4):937-943.

[2] 王琳，姜立新，杨天青，等．地震应急信息自动分类方法研究 [J]. 震灾防御技术 ,2019,14(4):907-916.

[3] 崔满丰，张晋辉．基于网站的地震应急信息发布技术 [J]. 地震地磁观测与研究 ,2020,41(4):232-238.

[4] 郑向向，帅向华．地震灾情短信编码的初步研究 [J]. 自然灾害学报 ,2012,21(1):92-100.

[5] 白仙富，李永强，陈建华，等．地震应急现场信息分类初步研究 [J]. 地震研究 ,2010,33(1):111-118+120.

[6]KIM Y.Convolutional neural networks for sentence classification[EB/OL].(2014-08-25)[2024-05-10].https://arxiv.org/abs/1408.5882.

[7]MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Distributed Representations of Words and Phrases and their Compositionality.La Jolla, CA:Neural Information Processing Systems,2013:3136-3144.

[8] 预训练中文字向量 [EB/OL].(2021-06-21)[2024-05-12]. https://github.com/Embedding/Chinese-Word-Vectors.

【作者简介】

白灵（1979—），女，黑龙江安达人，硕士研究生，副教授，研究方向：自然语言处理。

（收稿日期：2024-05-29）