基于微博舆情的地震影响场可视化平台研究与实现

栗纪娟[1] 黄 猛[1] 田累积[1] 谢颖瑶[1]
LI Jijuan HUANG Meng TIAN Leiji XIE Yingyao

摘 要 震后快速确定地震影响场的分布对地震应急救援工作部署具有非常重要的意义。近年来信息技术的快速发展，微博新闻评论等信息随着地震发生海量增长，也包括很多灾情位置信息，为快速绘制地震影响场提供了可能。但是以上信息存在震感信息量较少、位置信息不精确、文本篇幅短、表述口语化、语义模糊等问题。为了解决以上问题，首先采用震感信息关键词在爬取微博数据时进行筛选，并使用二分类算法提取震感信息。然后采取命名实体识别技术，将震感信息中的地理位置信息进行精准识别。最后选用CNN算法对短文本数据进行分析，使用BIGRU算法解决表述口语化的问题，采用ALBERT模型对语义模糊的文本进行分析，提出一种ALBERT+BIGRU+CNN短文本分类模型，充分提取震感信息的语义特征，结合《中国地震烈度表》作为分类标准，快速准确获取地震影响场数据，并采用Vue+SpringBoot技术构建可视化平台将其绘制到三维地图中，为震后应急救援提供辅助参考。

关键词 地震影响场；ALBERT；震感信息；短文本分类

doi：10.3969/j.issn.1672-9528.2024.02.049

0 引言

地震发生后，第一时间了解受灾区域并有效针对性地开展救援工作至关重要，然而地震烈度图一般由专家去现场对灾区的房屋建筑、地面、山体滑坡等破坏程度调研后绘制而成，时间为一周左右，周期较长，因此能否快速准确地获取地震影响场对地震应急救援有很大的影响。

随着互联网的普及和科学技术的迅猛进步，社交媒体的兴起打破了传统的信息传播途径，以新浪微博为代表，在突发性灾害发生后，微博数据会在短时间内呈几何式增长，海量的微博数据中蕴含了许多有价值的信息，因此本文选取了微博数据作为快速获取地震影响场的数据来源。

近年来人工智能领域发展迅速，机器学习和深度学习在自然语言处理领域的应用越来越广泛，而文本分类是该领域中的重要任务之一。由于传统的机器学习方法提取文本特征方式较浅层，存在无序性、上下文信息丢失和高维稀疏等问题，往往分类效果不佳。而深度学习的兴起解决了以上问题，经典的深度学习模型有：CNN[1]、RNN、LSTM、GRU、BILSTM[2]、BIGRU[3]以及BERT[4]模型系列等。深度学习是一种以神经网络为架构对目标样本进行特征学习的算法，在文本分类任务中显示出优异表现。然而随着深度学习的深入研究，研究人员发现单一的模型不能进一步提升文本分类的准确性，因此提出了混合模型，张磊[5]提出了基于BERT+CNN的产科电子病历短文本分类算法，利用BERT预训练模型对句子进行特征向量化表示，采用卷积神经网络增强了模型提取上层特征的能力，最终模型在产科电子病历中的分类任务中F_1值高达94%。温超东等人[6]提出了一种结合ALBERT[7]和双向门控循环单位的专利文本分类模型，使用ALBERT预训练的动态词向量代替了传统的静态词向量，使用BIGRU充分提取了文本中长距离词之间的关联语义特征，最终模型在不同层级专利文本的分类任务中表现优异。

由此可见，不同组合的深度学习模型在不同的任务中表现不同，因此本文以2022年9月5日四川泸定6.8级地震为例，针对海量的微博数据中震感信息量较少，数据中不提供精确的位置信息，且震感信息中存在文本篇幅短、表述口语化、语义模糊等特点，首先采用震感信息关键词在爬取微博数据时进行筛选，并使用二分类算法清洗后获取震感信息；然后采用命名实体识别（NER）技术，将震感信息中的地理位置信息进行精准识别；最后提出一种ALBERT+BIGRU+CNN短文本分类模型，充分提取震感信息的语义特征，结合《中国地震烈度表》作为分类标准，快速准确获取地震影响场数据，并采用目前所流行的MVVM前后端分离架构构建可视化平台，将分析所得数据可视化到三维地图中，为震后应急救援提供辅助参考。

1. 防灾科技学院 河北三河 065201

1 系统介绍

本系统旨在从微博舆情数据中快速提取并分析震感信息数据，准确获取地震影响场并可视化到平台中。本文获取了2022 年至今破坏性地震 5 次共计 210 604 条舆情数据，主要以 2022 年 9 月 5 日四川泸定 6.8 级地震为例，首先采用震感信息关键词在爬取微博数据时进行筛选，并使用二分类算法提取震感信息；然后采取 NER 技术，将震感信息中的地理位置信息进行精准识别，并结合百度 API 将其转换为经纬度坐标信息；最后提出一种 ALBERT+BIGRU+CNN 短文本分类模型，充分提取震感信息的语义特征，结合《中国地震烈度表》作为分类标准，将震感信息精准分为 VI 度、VII 度、VIII 度、IX 度四类，快速准确获取地震影响场数据。本文基于以上数据，选取 MVVM 前后端分离架构，采用 Vue+SpringBoot 前后端框架构建可视化平台，使用 Echarts+Element+Arcgis 技术将分析所得数据可视化到三维地图中，主要有评论数据管理、词云图和地震影响场可视化模块。

2 系统整体架构概述

系统整体的技术流程图如图 1 所示。

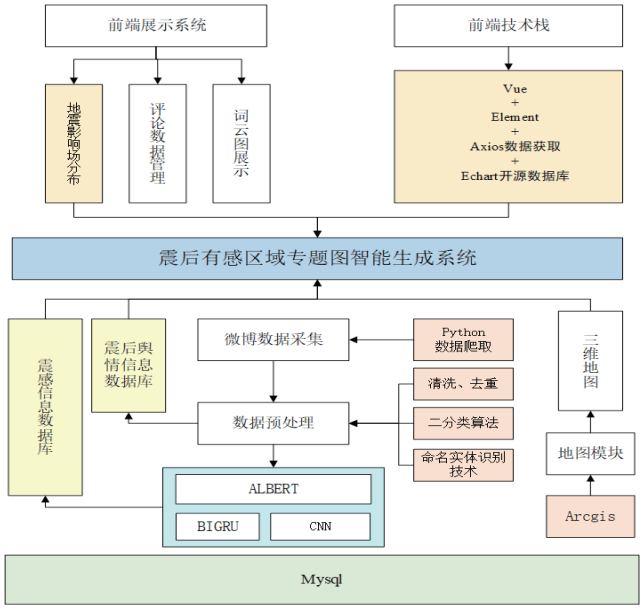

图 1 技术流程图

本系统分为数据采集分析与语料库的构建和前端展示两部分。数据采集分析与语料库的构建主要是采集微博舆情数据并提取分析震感信息构建语料库。系统首先采用震感信息关键词在爬取微博舆情数据时进行筛选，清洗去重后获得震后舆情语料库；然后采用二分类算法获取震感信息，采用 NER 技术将震感信息中的地理位置信息进行精准识别，构建震感信息语料库；最后使用 ALBERT+BIGRU+CNN 模型，充分提取震感信息语义特征，结合《中国地震烈度表》作为分类标准，将震感信息精准分为 VI 度、VII 度、VIII 度、IX 度四类，获取地震影响场数据。

前端展示主要是将数据进行整理分析后可视化到地图中。系统采用 MVVM 的架构方式，使用 Vue 框架的开发平台，以 Spring Boot 作为后台框架，结合 ES6 的相关技术及其规范来进行应用开发；在数据获取方面，根据数据采集分析与语料库的构建模块获取数据并将处理好的信息存入数据库；在数据展示方面，应用 Echarts+Element+ArcGIS 技术和 JavaScript 开发方法进行 Web 前端展示，实现震后快速将地震影响场数据可视化等功能。

3 数据库设计

本次实验主要使用 Python 爬虫技术获取新浪微博中的 2022 年至今破坏性地震 5 次共计 210 604 条舆情数据，搜集了这 5 次地震的烈度分布图，为本次实验提供数据支持，如表 1 所示，由此看出在四川泸定 6.8 级地震获取的数据最多。因此，本文主要以 2022 年 9 月 5 日四川泸定 6.8 级地震为例。

表 1 数据获取

时间	地点	数据量
2022 年 6 月 1 日 17:00	四川雅安市芦山县 6.1 级地震	36 598
2022 年 6 月 10 日 1:28	四川阿坝州马尔康市 6.0 级地震	12 737
2022 年 9 月 5 日 12:52	四川甘孜州泸定县 6.8 级地震	57 192
2023 年 5 月 2 日 23:27	云南保山市隆阳区 5.2 级地震	18 613
2023 年 8 月 6 日 2:33	山东德州市平原县 5.5 级地震	51 591

首先采用 Python 爬虫技术结合地震三要素以及“感”“晃”“摇”“抖”等关键字和起止时间等因素，共获取 57 192 条舆情数据。

然后，本文先使用 Python 技术对原始数据进行去除重复数据和空白数据以及缺失数据后，以 id、用户名、发布内容、发布时间、发布地点等为字段构建震后舆情信息数据库，采用二分类算法使用 01 标注法将数据分为非震感信息和震感信息两类，将无关信息去除，最后保留震感信息共计 30 925 条。再采用 NER 技术使用 BIO 标注法对震感信息进行地理位置实体识别，将没有地理位置的信息去掉，最终获取含有地理位置名称的 9526 条震感信息并使用百度 API 将地理位置信息转化为经纬度坐标信息。

最后，地震烈度评定的因素包括人的感觉、器物的反应、房屋震害程度以及自然环境变化等。本文主要以《中国地震烈度表（GB/T 17742—2020）》作为分类标准，依照我国破坏性地震烈度评定工作的惯例，对收集到的震感信息数据从 VI 度开始标注 [8]，由于本次实验所搜集的地震烈度最大为 IX 度，因此本文将震感信息分为 VI 度、VII 度、VIII 度、IX 度四类，参考云南地震局曹彦波等人 [9] 对震后有感范围的分类，建立如表 2 震感信息分类，结合该表和地震烈度图对震感信

息进行标注后以 id 、用户名、发布内容、发布时间、发布地点、经度、纬度、震感程度等为字段构建震后震感信息数据库。

表 2 震感信息分类

烈度分类	人的感觉	器物的反应	房屋破坏	其他
VI 度	震感明显，多数人有感、惊醒、骑行有感惊慌失措，仓皇逃出	杯子中水振荡、悬挂物或树枝明显摆动、器皿碰撞作响	基本完好，数十间，墙体开裂，梭、掉瓦，填充墙体开裂	路面轻微开裂，生命线设施轻微受损；零星落石、滑坡等
VII 度	震感强烈，惊慌失措，多数人仓皇逃出，站立不稳，骑行不稳	悬挂物剧烈摇摆或损坏坠落、书物掉落、轻家具移动	轻微破坏，数百间，屋架倾斜、脱榫，墙体开裂，梭、掉瓦，填充墙开裂	路面开裂，生命线设施轻微变形、开裂，少量土石滑落、个别滑坡点等
VIII 度	害怕，摇晃颠簸，行走困难	多数家具移动、部分翻到、树干摇动、树枝折断	中等破坏，数千间，局部倒塌，开裂明显，X 型裂缝贯通，填充墙局部倒塌	路基变形，生命线设施开裂、变形、受损，许多崩滑、滑坡、泥石流等
IX 度	坐立不稳，行动的人可能摔跤很害怕，站不稳、坐不稳，跌倒	器物翻倒、树干折断、衣柜等重家具和放置稳当的家具翻倒	严重破坏，数万间，结构严重破坏，较严重的水平或“X”型贯通裂缝	路基下沉，生命线设施断裂、损坏、局部垮塌，大量山体崩塌、滑坡、泥石流等

4 关键技术

为了解决震感信息中文本口语化、语义模糊、文本语义特征提取不充分、对全局语义信息提取不全面、分类不准确等问题，本文提出构建 ALBERT + BIGRU + CNN 模型，模型结构图如图 2 所示。

首先，将文本语料库中的每一条输入信息进行分字操作，将其输入到 ALBERT 的嵌入层后，这里会将每个字符在整个输入序列中位置信息标记出来获取相应标号；其次对每个字符进行向量化表示，获取向量信息序列；然后将该向量序列传送到 transformer 编码器中通过参数共享等方式，获取更深层次的语义信息特征；最后得到每个字符的特征向量表示。将上一层获取的序列特征向量传入双向门控循环网络（BIGRU）层，它由两个独立的门递归单元神经网络（GRU）构成，分别为正向 GRU 和反向 GRU，可以同时从文本序列的前向和后向获取前后文关联特征，最终获取文本序列的全局特征向量表示。t 时刻 BIGRU 的状态值 h_t 由正向 GRU 状态值和反向 GRU 值共同组成，计算公式为：

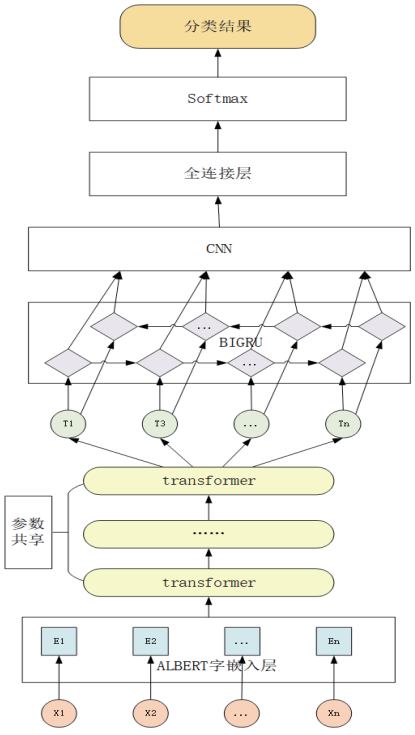

图 2 ALBERT+BIGRU+CNN 模型

$$h_t = w_t\overrightarrow{h_t} + v_t\overleftarrow{h_t} + b_t \tag{1}$$

式中：w_t 和 v_t 表示权重系数矩阵，b_t 表示偏置值。将上层的全局特征向量输入到卷积神经网络（CNN）层，CNN 主要由卷积层和池化层组成，卷积层是 CNN 最重要的部分，它由多个卷积核组成，使用多个不同大小的卷积核对输入序列矩阵进行卷积操作，主要通过数据窗口在整个特征矩阵的平移滑动，计算得到特征矩阵，输出该部分特征信息。卷积计算公式为：

$$c_i = f\left(\sum w_i h_i + b_i\right) \tag{2}$$

式中：w_i 表示卷积核的权重参数，h_i 代表输入矩阵，b_i 代表偏置值。将 CNN 层输出的特征向量输入到最大池化层，这样可以降低最终句子的向量维度，达到降维的目的，并且保留最关键的语义特征信息。将所得特征传入全连接层，对关键信息进行整合，通过 softmax 分类器获得分类结果。计算公式为：

$$y_i = softmax(w_u s_i + b_u) \tag{3}$$

式中：y_i 表示 softmax 分类器预测的类别，w_u 表示权重矩阵，s_i 表示上层输入的特征矩阵，b_u 为相对应的偏置。

5 系统模块介绍

5.1 评论数据管理模块

本模块主要对 Python 爬虫爬取微博评论信息，进行数据清洗去重后，分类到该评论所对应的地震中，使用 axios 发送请求获取数据库评论，采用 ElementUI 组件可视化为表格形式，在该模块中可以看到所有地震的评论信息，采用模糊查询方式，可以按照时间、震级、震中位置进行查询，如图 3 和图 4 所示。

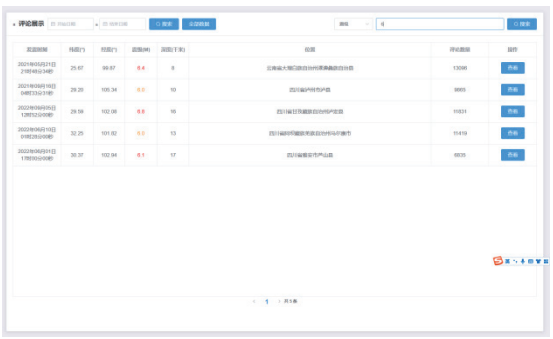

图 3 评论数据管理

图 4 评论展示

5.2 地震影响场三维可视化模块

本平台使用 ArcGIS API 作为地理图层基本展示框架，引入 ArcMap 中的三维地图和二维地图，该框架将三维地图分为 3 层，分别为地理图层、高程图和自定义层。本模块主要以 2022 年泸定 6.8 级地震为例，获取了 57 192 条舆情数据，并采用二分类算法和 NER 等技术最终提取含有地理位置的震感信息 9526 条。将上述数据经 ALBERT+BIGRU+CNN 模型进行分类后，结合地理位置经纬度坐标数据可视化为地震影响场分布图，同时获取微博评论等数据，将其渲染在自定义层进行展示，如图 6 和图 7 所示。对比图 5 官方发布地震烈度图可以看出，本平台地震影响场分布图与地震烈度圈方向和划分等级大致相同。

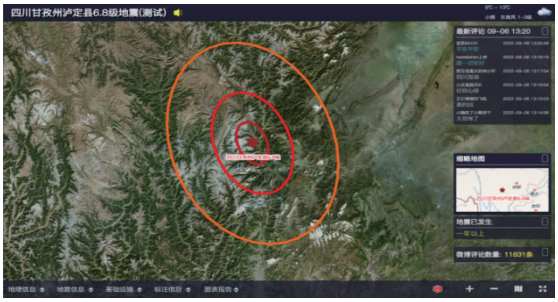

图 6 三维可视化

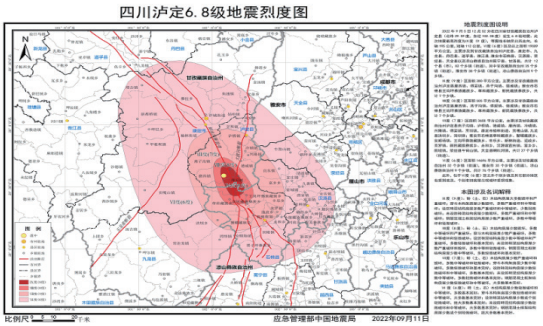

图 7 烈度图

6 结语

破坏性地震发生后，针对海量微博数据中震感信息较少，数据中不提供精确的位置信息，文本篇幅短、表述口语化、语义模糊，语义特征提取不充分等问题，本文主要以获取数据量最多的 2022 年泸定 6.8 级地震为例，首先采用震感信息关键词在爬取微博舆情数据时进行筛选，使用二分类算法清洗后获取震感信息，然后使用 NER 技术将震感信息中的地理位置信息进行精准识别，最后提出一种 ALBERT+BIGRU+CNN 模型结合《中国地震烈度表》作为分类标准，获取地震影响场数据并采用 Vue+SpringBoot 技术构建可视化平台将其绘制到三维地图中。对比图 7 官方发布地震烈度图可以看出本平台地震影响场分布图与地震烈度圈方向和划分等级大致相同，但每一等级所涉及区域大小还需进一步调整。因此在后续的工作中，将针对如何进一步准确绘制地震影响场分布区域等方面展开研究。

参考文献：

[1] KIM Y.Convolutional neural networks for sentence classification[EB/OL].(2014-08-25)[2023-09-26].https://arxiv.org/abs/1408.5882.

[2] TONG J, WANG Z, RUI X. A multimodel-based deep learning framework for short text multiclass classification with the imbalanced and extremely small data set[J].Computational intelligence and neuroscience, 2022(10):178-207.

[3] DEY R , SALEMT F M . Gate-variants of gated recurrent unit (GRU) neural networks[C]//IEEE Pulsed Power Conference. Piscataway:IEEE, 2017:1597-1600.

[4] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human LanguageTechnologies. Piscataway: IEEE, 2019: 4171-4186.

[5] 张磊 . 基于 BERT 和 CNN 的产科电子病历短文本分类算法研究 [D]. 广州 : 暨南大学 ,2021.

[6] 温超东 , 曾诚 , 任俊伟 , 等 . 结合 ALBERT 和双向门控循环单元的专利文本分类 [J]. 计算机应用 ,2021,41(2): 407-412.

[7] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite bert for self-supervised learning of language representations[C]// Proceedings of International Conference on Learning Representations. Piscataway: IEEE, 2019: 1-17.

[8] 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 ,2020.

[9] 曹彦波 , 吴艳梅 , 许瑞杰 , 等 . 基于微博舆情数据的震后有感范围提取研究 [J]. 地震研究 , 2017,40(2):303-310.

【作者简介】

栗纪娟（1999—），女，河北邯郸人，硕士，研究方向：自然语言处理。

黄猛（1976—），男，河南新乡人，硕士，教授，研究方向：GIS、软件工程、机器学习、深度学习、大数据分析。

田累积（2002—），男，内蒙古包头人，学士，研究方向：计算机科学与技术。

谢颖瑶（2003—），女，广东广州人，学士，研究方向：计算机科学与技术。

（收稿日期：2023-11-17）