Journal of Hydrology 603 (2021) 127053

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Extracting the location of flooding events in urban systems and analyzing the semantic risk using social sensing data

Yan Zhang [a], Zeqiang Chen [b,c], Xiang Zheng [d], Nengcheng Chen [a,b,c,*], Yongqiang Wang [e]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China
[b] National Engineering Research Center for Geographic Information System, China University of Geosciences (Wuhan), Wuhan 430079, China
[c] School of Geography and Information Engineering, China University of Geosciences (Wuhan), Wuhan 430074, China
[d] School of Information Management, Wuhan University, Wuhan 430079, China
[e] Changjiang River Scientific Research Institute, Wuhan 430010, China

ARTICLE INFO

This manuscript was handled by Nandita Basu, Editor-in-Chief, with the assistance of Marc F. Muller, Associate Editor

Keywords:
GeoAI
NLP
Deep learning
Social sensing
Urban functional zone
City portrait

ABSTRACT

The aggregation of the same type of socio-economic activities in urban space generates urban functional zones, each of which has one function as the main (e.g., residential, educational or commercial), and is an important part of the city. With the development of deep learning technology in the field of remote sensing, the accuracy of land use decoding has been greatly improved. However, no finer remote sensing image could directly obtain economic and social information and it has a high revisit cycle (low temporal resolution), while urban flooding often lasts only a few hours. Cities contain a large amount of "social sensing" data that records human socio-economic activities, and GIS is a natural discipline with strong socio-economic ties. We propose a new Geo-Semantic2vec algorithm for urban function recognition based on the latest advances in natural language processing technology (BERT model), which utilizes the rich semantic information in urban POI data to portray urban functions. Taking the Wuhan flooding event in summer 2020 as an example, we identified 84.55% of the flooding locations in social media. We also use the new algorithm proposed in this paper to divide the main urban area of Wuhan into 8 types of urban functional zones (kappa coefficient is 0.615) and construct a "City Portrait" of flooding locations. This paper summarizes the progress of existing research on urban function identification using natural language processing techniques and proposes a better algorithm, which is of great value for urban flood location detection and risk assessment.

1. Introduction

The coupling effect of global climate change, sea level rise and urbanization has been accentuated, and hydro-meteorological disaster events are frequent (Chang et al., 2021). According to statistics, the economic losses caused by global floods account for more than 30% of the total losses from natural disasters. Floods are among the most devastating hazards on Earth, posing great threats to a large amount of population in the world (Huang, 2020; Wan and Fell, 2004). China is one of the most severely flooded regions in the world, and the Yangtze River basin is the third largest in the world, with a total basin area of 1.8 million square kilometers, accounting for 18.8% of China's land area. With the economic development of southern China, the economic losses caused by the flooding of the Yangtze River have become more and more serious. 2020 China's Yangtze River basin suffered the most serious flooding since 1998, a total of 70,471,000 people were affected by the disaster, with direct economic losses of 214.31 billion RMB (about 325.966 billion dollars).[1] In this paper, we study the 2020 floods in Wuhan (the largest city in the middle reaches of the Yangtze River), and use a social sensing approach for location extraction and semantic computation of urban flooding events.

There has been considerable research using specifically synthetic aperture radar (SAR) imagery, optical satellite imagery and a digital elevation model (DEM) (Rakwatin et al., 2013), monitoring and analysis of flooding events, such as analyzing the extent of damage and affected population (Sun et al., 2016), integrating remote sensing and deep

* Corresponding author at: State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China.
E-mail address: cnc@whu.edu.cn (N. Chen).
[1] http://www.gov.cn/xinwen/2020zccfh/24/index.htm.

https://doi.org/10.1016/j.jhydrol.2021.127053
Received 6 May 2021; Received in revised form 2 October 2021; Accepted 7 October 2021
Available online 6 November 2021

learning methods to extract the depth of standing water (Hultquist and Cervone, 2020; Syifa et al., 2019; Sharma et al., 2019), etc. However, traditional sensor networks provide data with insufficient spatial and temporal resolution, and satellite observations of the surface are easily obscured by clouds, resulting in longer access cycles (Schnebele et al., 2014). At the same time, urban flooding and waterlogging is characterized by a short impact time, which also places a high demand on the temporal resolution (density) of the data (Xu et al., 2021) .

The social sensing approach represented by the crowd-sourced Volunteered Geographic Information (VGI) data is an important complement to the physical sensing network (Schnebele and Cervone, 2013; Yan et al., 2020), focusing on the socio-economic domain (Cowie et al., 2018), plays an equally important role in natural disaster monitoring and early warning, and is a fast,low-cost,and efficient survey monitoring method (Zou et al., 2018). With the spread of 5G and the extensive coverage of smart devices, these changes have greatly enriched the means and range of data for social sensing, where everyone is a "sensor" providing real-time feedback on real-space events and entities (Liu et al., 2015). Sensor networks are highly accurate and fixed in format, with a high density of available data (Zhang et al., 2021). Compared to it, social sensing data has the advantages of massive data volume, wide coverage area, high observation density and can record human behavior patterns in detail (Forrest et al., 2020).

However, the huge amount of observed data obtained by social sensing methods has a serious noise problem. It has diverse sources, inconsistent structures and low density of valid information. The information describing natural disasters exists mostly in the form of textual descriptions and a large amount of valuable information is submerged in irrelevant messages. This irregular information undoubtedly brings some impact on scientific research, and it is time-consuming and not time-sensitive to filter it manually. It is a popular and challenging research direction to extract disaster information from social sensing data by using artificial intelligence technologies. With the development of natural language processing technology (NLP), the extraction of disaster information from social sensing data has been greatly advanced (Kaufhold et al., 2020). In recent years, many scholars have conducted research on this issue, extracting and analyzing social media information for situational awareness and disaster assessment (Atefeh and Khreich, 2015; Wang et al., 2020); attempting to quantify and regularize social sensing data, constructing knowledge service systems (de Bruijn et al., 2019), improve disaster response capability (Barker and Macleod, 2019; Wang and Ye, 2018). It has been heavily applied in various phases of disaster preparedness, response and recovery (Zahra et al., 2020), such as wildfire hazards (Wang et al., 2016), earthquake (Robinson et al., 2013), Hurricane Harvey (Yang et al., 2019; Zou et al., 2019) and Hurricane Sandy (Wang et al., 2019).

In the face of flooding events, the temporal resolution of physical sensing network represented by remote sensing satellites is too low. After searching open source satellite data from GF and ZY series, LANDSAT, MODIS and Sentinel, we found no high quality remote sensing imagery in Wuhan in early July when the flooding was at its worst in 2020. In addition, the land use interpretation results of remote sensing images do not have economic and social information. These are the two main problems that we set out to solve in this paper.

Our main contributions are twofold:

From an application perspective, this paper proposes a social sensing approach for flooding location identification, distinguishing waterlogged locations from common locations based on semantic information using the BERT-Bilstm-CRF model.

From a methodological perspective, this paper contributes a new GeoSemantic2vec algorithm that can learn the spatial context relationships from POI data to define urban functional zones (the areas assigned to different social and economic activities (Yuan et al., 2014)) through economic and social spatial semantics.

The article is organized as follows. Section 2 introduces the work related to this paper, and Section 3 presents the research methodology, which is divided into two parts: flood area identification and urban function extraction. Sections 4 and 5 are our experimental sections, which introduce our experimental background, flood identification and urban function calculation based on the new method, respectively. The urban function calculation includes the accuracy comparison with the baseline algorithm and identifying urban functions in flooded locations. The last Section concludes the article.

2. Related works

Remote sensing images are often used in urban land use studies. The advantage of remote sensing technology is that it covers a large area and can still obtain more accurate land use information in many inaccessible areas, but remote sensing methods do not directly reflect socio-economic information. In addition, remote sensing images have the same resolution in both suburban and urban areas, but urban areas contain a large amount of "social sensing" data. The most common data source for land use related research using social sensing methods is point of interest (POI) data, which is a kind of mapping of geographic entities. POI information in cities is much richer than remote sensing images (we have collected more than 500,000 POI records within the third ring road of Wuhan city) and has great research potential.

There have been a considerable number of studies using social sensing data such as cab trajectories(Zhang et al., 2016; Wang et al., 2018), cell phone signaling data(Yuan et al., 2012), and bicycle sharing data(Zhang et al., 2018) for studies related to urban functional zoning or land use. Most of them use geospatial analysis methods for trajectory similarity matching or Origin Destination analysis. The use of NLP algorithms to extract potential geospatial features based on POI, social media and other related data, and then identify urban functional zoning (Sun et al., 2016), is another research direction. The earliest and most commonly used mining algorithms are the Term Frequency-Inverse Document Frequency (TF-IDF) method (Aizawa, 2003), Latent Dirichlet Allocation (LDA) method (Blei et al., 2003) and Probabilistic Latent Semantic Analysis (pLSA) (Bosch et al., 2006) method. These unsupervised clustering methods are not restricted by the text form and automatically find the number of topics (urban functional zones). They could generate topic vectors and the distance between vectors indicates semantic similarity (Gao et al., 2017). With the introduction of the Word2vec model by Google, this word embedding method has been well received and widely used in urban feature recognition (Yao et al., 2017; Liu et al., 2020; Zhai et al., 2019). Improved algorithms based on Word2vec represented by Place2vec could consider the (spatial) context and embed the text into tighter vectors (Zhai et al., 2019).

However, these methods have some drawbacks, as words (POI) and vectors are in a one-to-one relationship, it is difficult to solve the problem of multiple meanings of words (one POI may assume multiple functions, and different POIs may belong to the same function such as parks and green spaces, restaurants and eateries). The traditional method represented by LDA is actually a probabilistic model and does not contain semantic information (Angelov, 2020). Besides, Word2vec algorithm is a static approach, although general, but can not do dynamic optimization for specific tasks. With the advancement of natural language processing technology, transformer has successfully replaced the traditional Recurrent Neural Network (RNN) (LSTM/GRU structure) network (Tay et al., 2020). In this paper, we propose a new GeoSemantic2vec algorithm based on the latest Bidirectional Encoder Representation from Transformers (BERT) model.

Unlike the spatial random sampling method (Gao et al., 2017) and the traffic analysis zone (TAZ) delineation method (Yao et al., 2017), we perform a uniform spatial sampling of the city (Intensive point coverage) to obtain as much fine-scale spatial semantic information as possible. Compared with the Place2vec method, the algorithm takes the BERT model as input, and this pre-trained model can better take into account the spatial contextual information, even if the same POI will return different vector embedding results under different spatial contextual

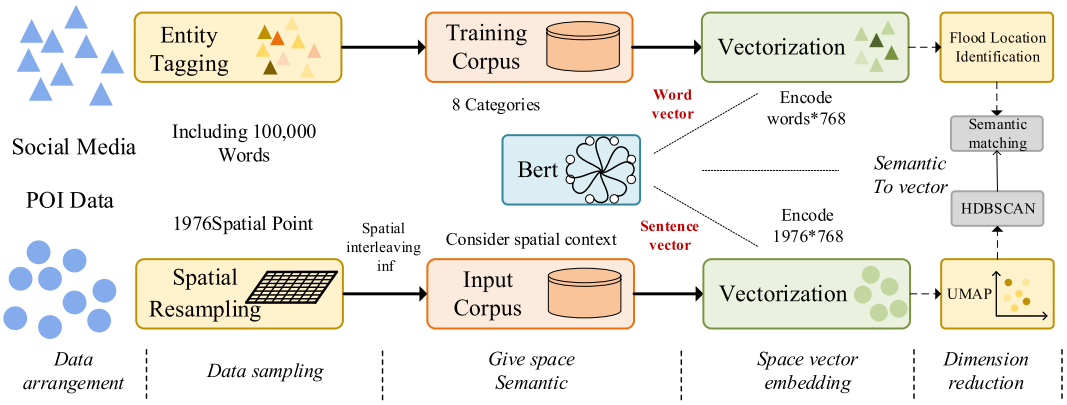

Fig. 1. Technology roadmap for this article.

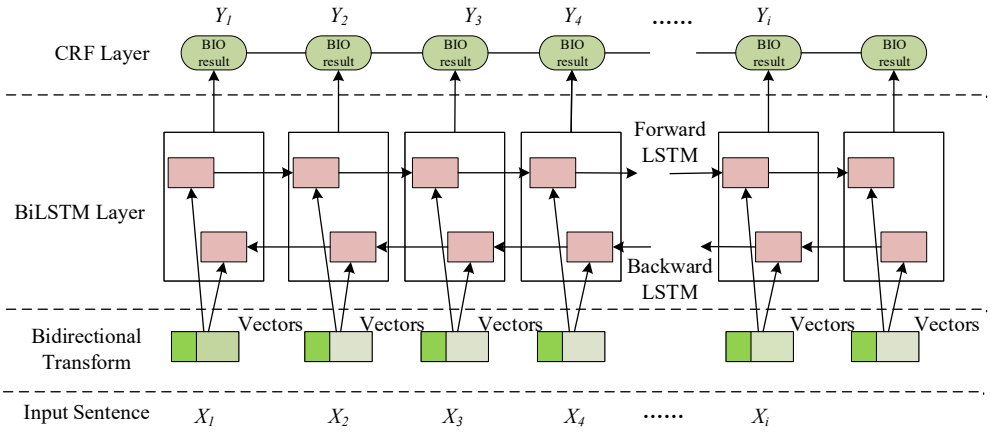

Fig. 2. BERT-LSTM-CRF NER model structure.

relationships. In addition, compared with the k-means clustering method of the Place2vec algorithm, this paper adopts the density-based Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) method, which can effectively eliminate the influence of outliers (Yan et al., 2017; Campello et al., 2013). We use the Uniform Manifold Approximation and Projection (UMAP) algorithm (McInnes et al., 2018) to reduce the dimensionality of the semantic vector. The algorithm proposed in this paper is interpretable and could avoid complicated hyperparameter adjustment (Lu et al., 2011).

This paper belongs to the study of urban flood risk analysis, which is different from the traditional use of flood-related geographical factors (elevation, slope, curvature, distance to river and land use, etc.) to determine flood risk (Löwe et al., 2021; Herbert et al., 2021; Pham et al., 2021; Lei et al., 2021). We perform a data-driven social sensing of urban flooding events and the assessment of their riskiness. It enriches the existing means of flood data collection. The advantage of our approach is that it is based on crowdsourced data extraction of flood locations, but also considers the socio-economic attributes of flood locations, and is based on a data-driven study of the socio-economic risk of floods to humans (Romascanu et al., 2020; Wang et al., 2018).

3. Research methods

As shown in Fig. 1, this paper is organized by two main parts, firstly named entity identification (NER) method to extract the location of waterlogging from social media (Weibo, 'Chinese Twitter') data. Secondly, we propose the GeoSemantic2vec algorithm, which extracts semantic information and clusters different functional areas of the city by spatially sampling the study area and computing the POI spatial context of the sampled points using the BERT model. Then, the semantic computing and socio-economic mining are performed on the location of waterlogging.

Our method could be applied to rapid mapping of flooding disasters, which is faster compared to traditional methods, such as remote sensing and manual census. It fully considers the coupling relationship between people and the environment, and the complex impact of urban flooding on the social environment.

3.1. BERT introduction and named entity identification

BERT is a novel language model based on Transformer architecture (Devlin et al., 2018), which is mainly divided into two steps: pre-training and fine-tune. Compared with traditional neural network novel language models, BERT has achieved best results in several natural language processing tasks such as text classification, text similarity, intelligent question and answer, text labeling, and named entity recognition. The MLM (Mask Language Model) strategy is used to capture semantic connotations during BERT pre-training. The specific approach is: model randomly masks 15% of the words in the corpus (80% of training time), randomly replaces the word with another word 10% of the time, and keeps the original word unchanged for the remaining 10% of the time. BERT uses this noise-injection training approach to improve the model's ability to acquire semantic information, while also making it difficult to acquire the full amount of information, thus ensuring its excellent generalization ability. However, the cost of pre-training is very expensive. In this paper, we use the pre-trained BERT model based on the Chinese wiki released by Google (Number of hidden layers in the Transformer encoder = 12, Size of the encoder layers and the pooler

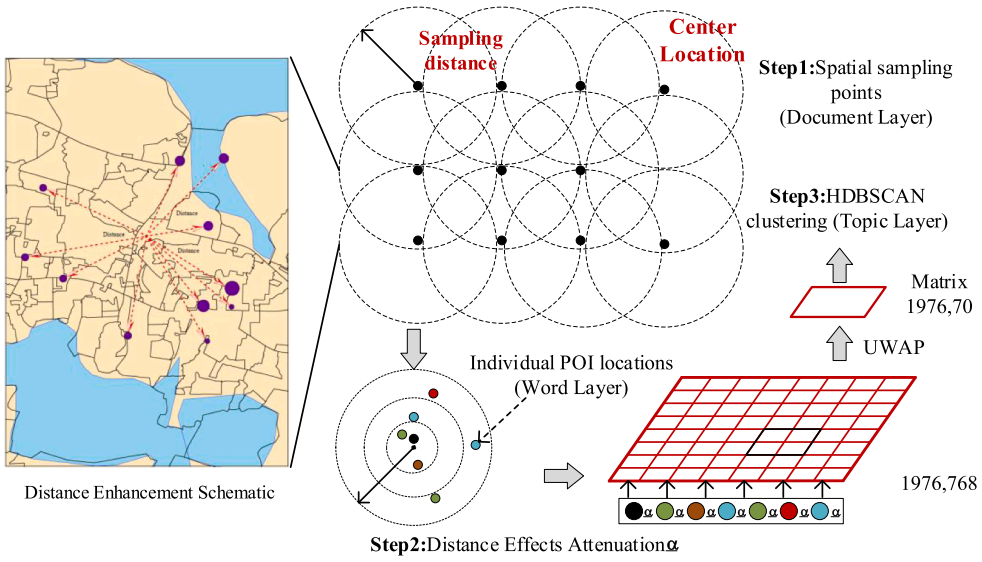

Fig. 3. Algorithm flow and distance enhancement schematic.

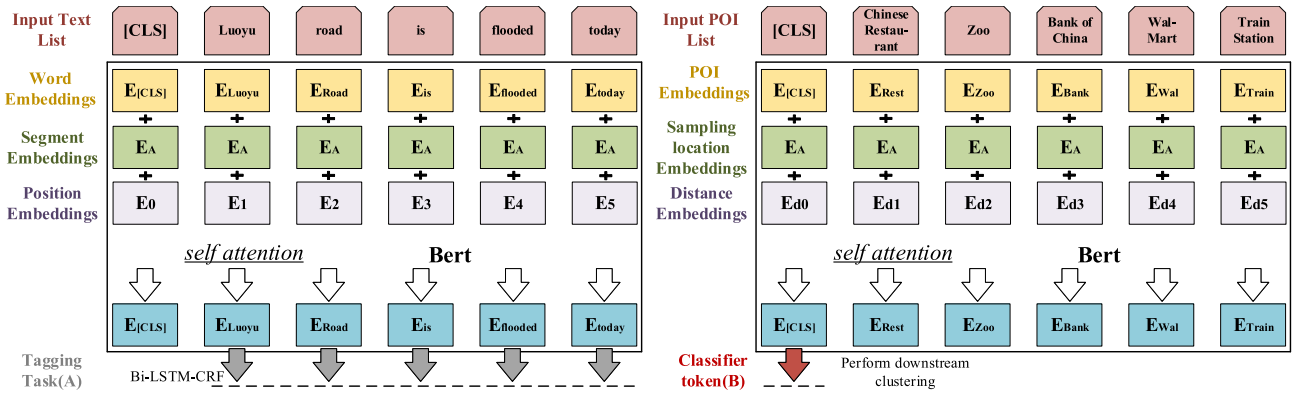

Fig. 4. Schematic diagram of model fine-tuning. (Left side represents NER task, extracts the location of flood from socially-sensing data. Right side represents classification architecture, generates vectorised representations of areas.).

layer = 768, Number of attention heads for each attention layer in the Transformer encoder = 12, total parameters = 110 M, including dictionary of 7,322 Chinese characters) [2], fine-tuning the model to identify flooding locations. Research has been conducted using BERT to extract location from social media data (NER task) (Zhang et al., 2020; Li et al., 2020), the principle of which is shown in Fig. 2. The basic idea is to transform the sequence annotation task (recognizing words in text that describe an address) into a classification task (tagging each word appropriately) at which machine learning excels.

The model inputs are sentences, and BERT encodes each word(X_i) and outputs a word vector. The word vectors are input to the Bi-directional Long Short-Term Memory (Bi-LSTM) layer for learning, which samples the sentences separately according to the inverse order and outputs the probability distribution of the words belonging to each type of entity. This distribution is fed to the Conditional Random Field (CRF) layer for judgment (Tseng et al., 2005) to assign appropriate labels (Y_i) to the words.

3.2. GeoSemantic2vec algorithm

Based on BERT model and urban POI data, we propose a GeoSemantic2vec algorithm for extracting urban functional areas from POI semantic information and location information. Inspiring by text classification tasks in natural language processing research, we perform uniform sampling in the study area to obtain the Sampling Location, generate the buffer of the Sampling Location at a certain distance (Sampling distance), and arrange the POIs (Word) according to the distance from the Sampling Location. The arrangement of POI is the expression of Sampling Location (document). As shown in Fig. 3, we consider that the closer the distance to the Sampling Location, the stronger the effect of the POI on the Sampling Location property, we use the enhancement parameter (Distance Effects Attenuation) to carry out the weighting, and the formula of this factor is:

$$\alpha_{distance}^{l_j} = \left[\frac{1 + \frac{\sum_{k=1}^{|L|} P_{lk}}{|L|}}{1 + d^{\beta}(l_i, l_j)}\right] \quad (1)$$

Among them: The context sampling location is l_j; $|L|$ represents the total number of POIs, $d^{\beta}(l_i, l_j)$ represents the distance between the POI l_i and the sampling location l_j. β represents an inverse distance factor (set to 1 in this paper). The numerator can be regarded as a smoothing constant for a given POI dataset. P_{lk} represents the total POI count associated with the sampling location l_j. As we explained, in this paper, we let $P_{lk} = 1$; therefore, the numerator is a constant (equal to 2) for all

[2] https://github.com/google-research/Bert

Algorithm 1: GeoSenmantic2vec Algorithm

Input: POI Data
Enhancement Factor α
Buffer Distance
$BubbleSort(inputlist, inputlength)$
$UMAP(neighbors, components, matrix)$
$HDBSCAN(eps, min_{samples}, matrix)$
Output: Urban Functional Zone

Document Layer;
$Matrix_{node} \leftarrow$ Spatial sampling matrix;
for $i \leftarrow 1$ **to** *nodes* **do**
    `// nodes is the number of spatial grids;`
    Initialize $Matrix_i \leftarrow \emptyset$;
    for $j \leftarrow 1$ **to** *POIs* **do**
        `// Iteration over all POIs;`
        if *j<Buffer Distance* **then**
            $Matrix_i += \{POI_{ij}\}$;
        end
    end
end
Word Layer;
for $i \leftarrow 1$ **to** $Matrix_{node}$ **do**
    $Matrix_i \leftarrow BubbleSort(Matrix_i, len(Matrix_i))$;
    for $j \leftarrow 1$ **to** $Matrix_i$ **do**
        $Matrix_{ij} \leftarrow Matrix_{ij} * \alpha$
    end
end
Topic Layer;
Semantic embedding of the sampling matrix;
$Embedded\ Matrix \leftarrow Bidirectional\ Encoder(Matrix)$;
$Embedded\ Matrix \leftarrow UMAP(neighbors, components, Embedded\ Matrix)$;
`// Dimensionality reduction matrix;`
$Matrix_{Lable} \leftarrow HDBSCAN(eps, min_{samples}, EmbeddedMatrix)$;
for $index, label \leftarrow 1$ **to** $Matrix_{Lable}$ **do**
    $Urban\ Functional\ Zone_{index} \leftarrow label$
end

POIs. $\alpha^{l_j}_{distance}$ is rounding up to an integer. Our enhancement method is also simple. If the enhancement parameter α value of POI_A is 2, then POI_A will appear twice in the list of sampling location l_j expressions.

After the weighting process, we use the list consisting of POIs as a representation of sampling location. We used the BERT model to generate this expression into a Sentence Vector with 1,976 sampling locations in the study area, generating a sentence vector of 768 dimensions (Size of the encoder layers and the pooler layer). Considering that the total number of POI types is not as rich as that total number of real-world words, we use UMAP to reduce the sentence vector to 70 dimensions (Zhai et al., 2019). Then, we use the HDBSCAN algorithm (McInnes et al., 2017) to cluster the reduced dimensional Sentence Vector and extract the urban themes of Sampling Location.

We provide the pseudo-code of the algorithm with the flowchart

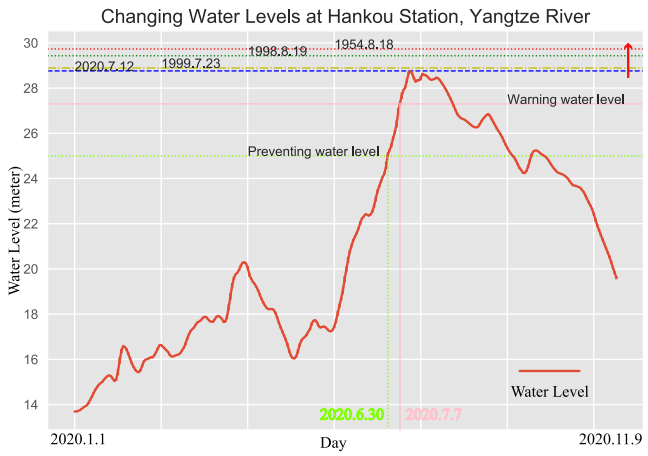

Fig. 5. Water level trends at Hankou Station on the Yangtze River.

model, and the output results are then used for downstream tasks. As shown in Fig. 4A, our sequence annotation task uses a single-word BERT encoding result (Vector), which is input to Bilstm-CRF for flood address extraction, while our urban functional partitioning only requires the encoding result of [CLS] (a 768-dimensional Vector) for the downstream task.

4. Identification of waterlogged locations from social sensing data

4.1. Study area and study time

Wuhan is the largest city in the middle reaches of the Yangtze River, with more than 100 urban lakes, and is known as the "Water City". In the context of rapid urbanization, Wuhan City faces serious urban flood control pressure during the rainy season every year (Zhou et al., 2021). Fig. 5 shows the water level changes at Wuhan Hankou station from January to November 2020. The water level at Hankou station reaches preventing water level on June 30, 2020, warning water level on July 7,

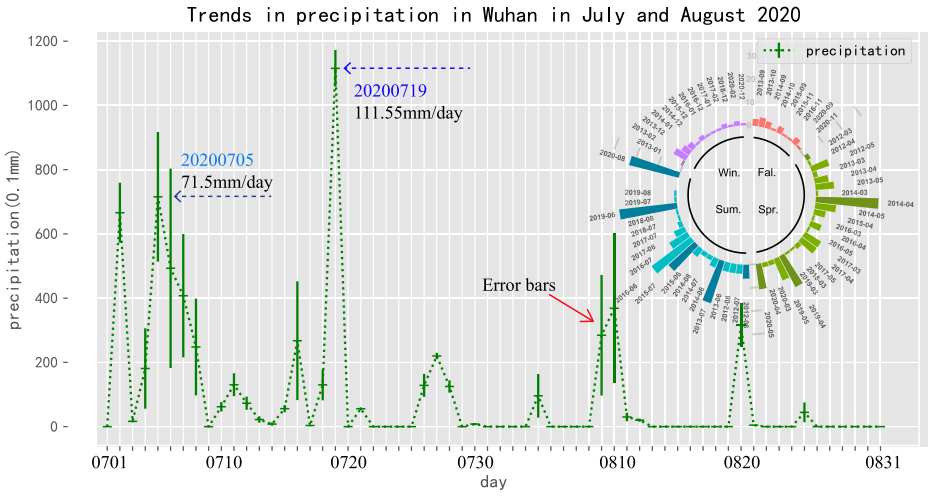

Fig. 6. Precipitation trends and the number of waterlogging notifications in Wuhan.

here[3].

In fact the two applications mentioned in this paper (flood address identification and urban function study) correspond to BERT's annotation tasks (Fig. 4A) Tagging tasks and sentence classification tasks (Fig. 4B). For example in Fig. 4B, the representation of a Sampling Location (Input POI List) is [Chinese Restaurant, Zoo, Bank of China, Wal-Mart, Train Station]. BERT will prefix the sentence (POI List) with [CLS (Classifier)] to generate a new POI List ([CLS,Chinese Restaurant, Zoo, Bank of China, Wal-Mart, Train Station]) for the next step of model training.

As shown in Fig. 4B, we embed the Input POI List, including POI word embedding, sampling location information embedding, and distance embedding. The POI word embedding represents the POI name, the sample location embedding represents the sample location and the distance embedding represents the spatial relationship between POI and sample location. The composite input is "fine-tuned" by the BERT and the water level reaches its peak on July 12.

Similarly, according to National Oceanic and Atmospheric Administration(NOAA data), as shown in Fig. 6, the precipitation in Wuhan reached 71.5 mm/day on July 5 and 111.55 mm/day on July 19, 2020, which means that there is a short period of heavy precipitation and the flood season overlaps with the rainy season, bringing a high risk of flooding in the main city. We searched and obtained all weibos (Similar to tweets in Twitter) about waterlogging in Wuhan after 2013. As shown in the attached figure, there are more weibos in summer and spring, we selected the weibos related to urban flooding in July and August 2020 as our data source to extract the location of waterlogging.

4.2. BERT training process

As mentioned in Sec 4.1, we obtained weibos related to flooding and waterlogging in Wuhan City in July and August 2020, with a total data of about 100,000 words. Social media does not fully cover everyone, and people react and express themselves differently to flooding events. However, if the amount of data is large enough, suitable algorithms can extract reliable information from the big data containing noise (Wang et al., 2015; Chen et al., 2021). We use BIO ternary annotation pattern (B-begin, I-inside, O-outside; B, i.e. Begin, means start, I, i.e.

[3] Step1: Spatially sample the study area, and all POIs in the sampling location buffer are considered as a "document". Step2: Each POI in the buffer is considered as a "word" and arranged according to the spatial relationship between the POIs and the sampling location, and the distance enhancement factor is used to strengthen this relationship. We get a matrix of the number of sampling locations multiplied by 768 (the number of BERT hidden layers). Step3: We reduce the dimensionality of the matrix obtained by Step2, cluster the results, and label each sampling position with the clustering results.

Fig. 7. Schematic of entity annotation process (this paper uses BIO annotation mode, using Yedda tool for annotation.).

Table 1
Model accuracy evaluation results (we treat the waterlogging location description differently from the irrelevant location description to improve the accuracy of flooding address identification).

	precision	recall	F1	Number
Total	51.52%	79.88%	62.64	1822
Area	81.89%	84.55%	83.2	127
Depart	48.57%	77.27%	59.65	35
Level	100.00%	90.91%	95.24	10
Measure	12.79%	49.06%	20.29	813
Rain	74.44%	89.33%	81.21	180
Stain	69.23%	64.29%	66.67	13
Station	68.01%	81.78%	74.26	297
Time	61.38%	86.59%	71.84	347

Intermediate, means middle, O, i.e. Other, means other, is used to mark irrelevant characters (Yang et al., 2017))[4] to annotate entities word by word. For example, a weibo: "serious flooding occurred on Luoyu Road, water depth of 20 mm, good road access on Guangba Road, no waterlogging occurred". Here although there are two locations at the same time, however Guangba Road is obviously not the location we need, we label it verbatim as O O and Luoyu Road as B-Area-begin,I-Area-begin. In addition to waterlogging area, we also labeled Time, Rain (precipitation level), Station (monitoring stations), Depart (response departments), Measure (anti-disaster measures), Stain (waterlogging specifics), Water (Yangtze River level description), and Level (risk level description). Fig. 7 shows our labeling process.

The experimental environment is Centos operating system, Tesla V100 GPU is used for training, and the model is built by Tensorflow.[5] The experimental parameters are set as follows: input dimension (max-seq-length) is 128, train-batch-size is 32, and learning-rate is 2e-5.20% of the data were randomly selected as the test set and the rest of the data were used as the training set. The Precision, Recall, and summation mean $F1-score$ are used as the evaluation criteria of model

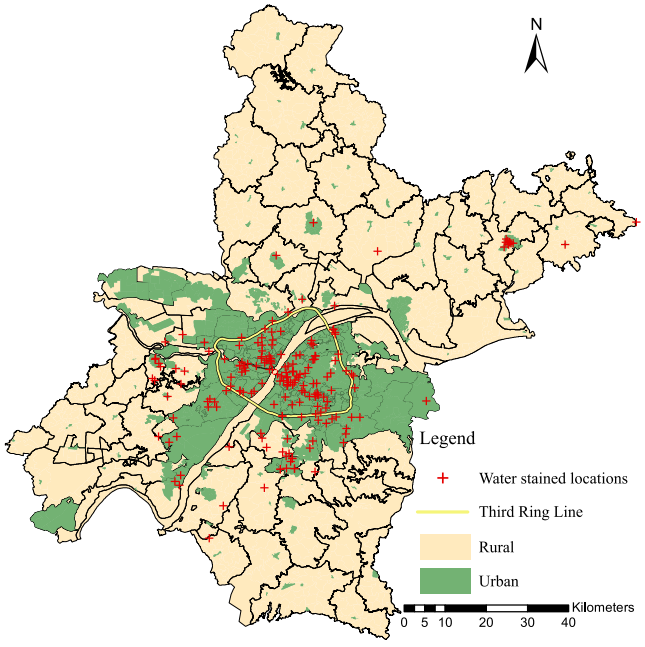

Fig. 8. Waterlogging locations in Wuhan City in summer 2020 (190 waterlogging locations, 2,989 related Weibos).

performance, which is calculated as follows.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Re}call = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = \frac{2 * Precision * \text{Re}call}{Precision + \text{Re}call} \tag{4}$$

where true positive (TP) indicates the case of correctly classifying i class as i class and true negative (TN) indicates the case of predicting j class as j class, both of which are correct predictions. False positives (FP) are the predictions that mark j class as i class, and false negatives (FN) are the cases that predict i class as j class.

[4] https://github.com/jiesutd/YEDDA
[5] https://tensorflow.google.cn.

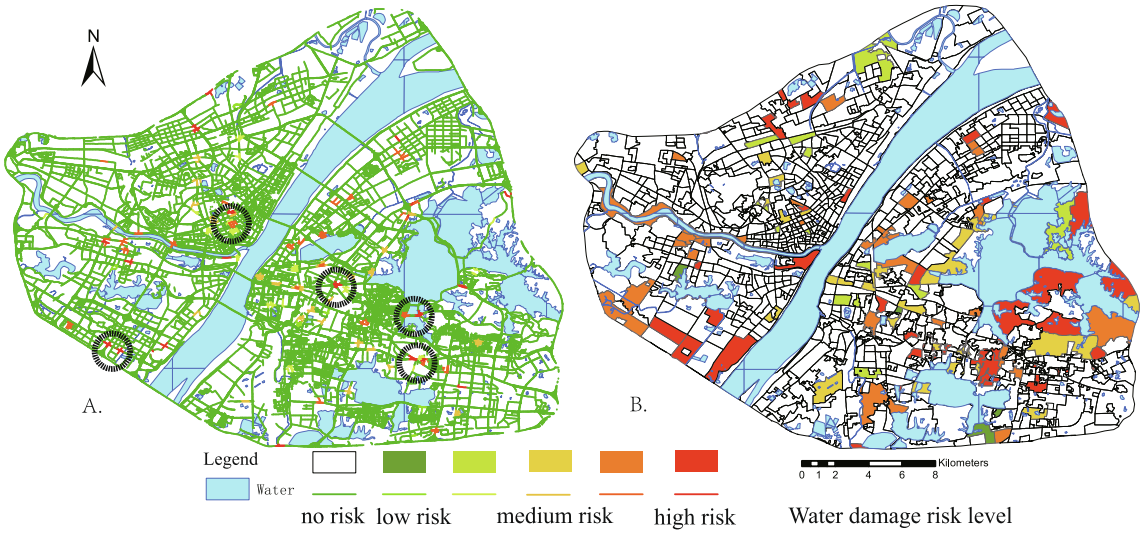

Fig. 9. Risk of waterlogging in Wuhan (Figure A: Road Scale, Figure B: Neighborhood Scale).

Table 2
Details of the POI data of the main city of Wuhan used in this paper.

Code	POI category	Chinese name	Abbreviation	Number
1	Café/Tea Bar	餐饮服务	Caf/Tea	65157
2	Road Facility	道路附属设施	RoadFac	508
3	Address/Location	地名地址信息	Adr/Loc	51410
4	Tourism Attraction	风景名胜	TourAtr	4203
5	Public Facility	公共设施	PubFac	3514
6	Transport services	交通设施服务	TransServ	27749
7	Shopping Mall	购物服务	ShopMal	8654
8	Transportation facilities	通行设施	TrabsFac	26085
9	Bank/Financial	金融保险服务	Bank/Fina	13385
10	Science and Education	科教文化服务	Sci/Edu	32666
11	Motorcycle Service	摩托车服务	MotServ	1006
12	Car service	汽车服务	CarServ	12898
13	Car repair	汽车维修	CarRepa	6213
14	Car sales	汽车销售	CarSale	1430
15	Residence	商务住宅	Residen	30688
16	Living Service	生活服务	LivServ	103579
17	Indoor Facility	室内设施	IndoFac	2675
18	Sports/Recreation	体育休闲服务	Spr/Rec	16910
19	Accommodation service	住宿服务	ComuServ	14806
20	Hospital	医疗保健服务	Hospital	21952
21	Governmental and Public Organizations	政府机构及社会团体	Gov/Pub	22058
22	Factory	公司企业	Factory	33061

Table 1 shows the evaluation results of our model, Number represents the number of entities in the test set. Among the eight recognition targets, the model has better recognition ability for the location of waterlogging. The Precision of it is 0.819, Recall is 0.846, and the summed mean $F1-score$ is 83.2, which means that our model has a good ability to accurately identify the location of inundation from a large number of locations.

4.3. Waterlogging location and geocoding

We manually checked the waterlogging addresses extracted by the NER model and collected 2,989 Weibos related to waterlogging in Wuhan. These Weibos contained 190 waterlogging locations (Area). We geocoded these locations to obtain detailed latitude and longitude coordinates. As shown in Fig. 8, the waterlogging locations are almost all concentrated in the urban area, and the waterlogging phenomenon is more intensive in the third ring road (101 waterlogging locations and 1,774 related Weibos).

We selected the area within the third ring road of Wuhan city as the main study area. The water hazard risk is measured by the number of waterlogging (Number of Weibo notifications), as shown in Fig. 9. The water hazard risk is presented on two scales, with Fig. 9A indicating the road sections prone to waterlogging and Fig. 9B indicating the streets with a high number of waterlogging phenomena. This assessment method does not take into account the road class as well as the socio-economic attributes of the waterlogging location (Darabi et al., 2019).

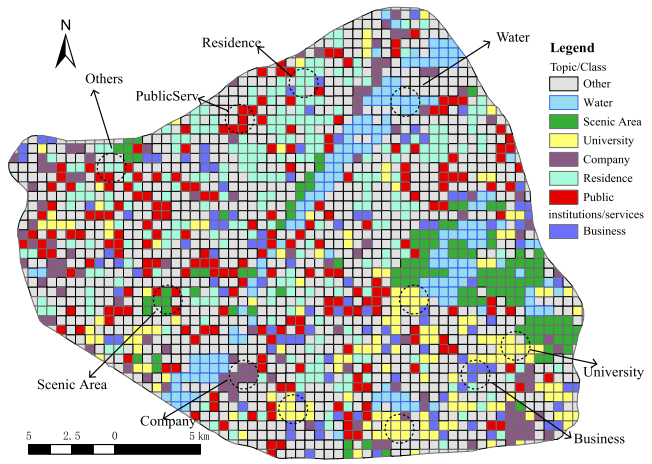

Fig. 10. Calculated city function classification results.

In the next section, we will further explore the geo-semantic information of waterlogging locations using the GeoSemantic2vec algorithm.

5. Semantic calculation of waterlogged locations based on the GeoSemantic2vec algorithm

5.1. City POI dataset

The POI data used in the paper comes from Baidu Maps, which is the largest online map provider in China.[6] Our study area is the main urban area within the third ring road of Wuhan city, with about 500,634 POI data. POI data are classified into 22 basetypes (Caf/Tea, RoadFac, Adr/loc, TourAtr, ShopMal, TrabsFac, Bank/Fina, Sci/Edu, MotServ, CarServ, CarRepa, CarSale, Residen, LivServ, IndoFac, Spr/Rec, ComuServ, Hospital, Gov/Pub, Factory), 220 SUBTYPE, and a more detailed 647 CATEGORY by attributes. We show the distribution of POIs according to basetype classification criteria in Table 2.

5.2. Urban functional zone identification

We use a 500-by-500 grid to divide the land within the 3rd Ring Road of Wuhan into 1,976 regions and set a buffer radius of 500 meters for the centroid of each region (the resolution can be higher if the computing power allows). The buffers of adjacent centroids are overlapped, which allows the model to learn information about the surrounding sampling areas. Based on the distance from the centroid, the POIs in the buffer are ranked and a list of POIs is obtained as an expression of the region (here we take the value of enhancement parameter as 1). Using the GeoSemantic2vec algorithm mentioned earlier, this POI list is sampled and learned to learn the spatial interleaving relationships and contextual semantic information of all POIs in the buffer. It uses the Distance Embeddings layer of the neural network to learn the spatial location relationships, and again learns the semantic information of different POIs through the POI Embeddings layer. Even for POIs with the same name, but with different spatial contextual relationships, their embeddings generate different vectors. This means that "phrases" are formed based on common POI pairings, e.g., Kentucky Fried Chicken(KFC) often appears with retail stores, while retail stores often appear with residential areas, and POIs with the same name (KFC) have different meanings in different pairings (KFC-retail store, KFC-residential area). We use HDBSCAN algorithm to cluster the vector of each sampling location to get the label (functional partition).

According to the clustering results, we divided the 1,976 areas into 8 categories in Fig. 10, namely Water (158, mainly composed of water area), Scenic Area (123, mainly taking on the functions of parks, water-friendly wetlands, etc.), University (162, mainly composed of universities and colleges), Company (114, mainly taking on the functions of industrial parks and factories), Residence (242, mainly for residential land use), Public Service (217, mainly for urban public services), Business (112, mainly for commercial and financial functions), and Other (848, it has no definite relationship).

In addition to that, we also take the vector mean of each functional zone to represent this functional zone and calculate the similarity of different urban functional zones. Table 3 shows the calculation results of Pearson correlation coefficients for different functional zones.

Table 3
Vector similarity of each urban functional zone.

	Other	Water	Scenic	University	Company	Residence	Public serv	Business	Number
Other	1	−0.53634	0.801053	0.838529	0.770569	0.908913	0.965172	0.93716	848
Water	−0.53634	1	−0.61636	−0.66308	−0.7282	−0.63454	−0.67693	−0.62969	158
Scenic	0.801053	−0.61636	1	0.818121	0.691984	0.678163	0.750295	0.841683	123
University	0.838529	−0.66308	0.818121	1	0.715146	0.701391	0.81686	0.779206	162
Company	0.770569	−0.7282	0.691984	0.715146	1	0.693495	0.825015	0.68022	114
Residence	0.908913	−0.63454	0.678163	0.701391	0.693495	1	0.911671	0.934959	242
Public serv	0.965172	−0.67693	0.750295	0.81686	0.825015	0.911671	1	0.909329	217
Business	0.93716	−0.62969	0.841683	0.779206	0.68022	0.934959	0.909329	1	112

The functional zone Other seems to be a multifunctional mixture, with similar vector expressions to the rest of the types of functional zones except Water. Water functional zone is very different from the rest of the functional zones, due to the low density of POI and the single type of this functional area. Scenic area has a similarity to University (0.818), which may be the fact that universities have better forest cover and some historic universities also assume a certain tourism function. University has a high similarity with many types, with the largest difference with Residence (0.701). Company is significantly different from Scenic, Residence and Business, but has a closer resemblance to Public Serv. Residence is strongly correlated with Business (0.935), indicating that mature neighborhoods with more residents tend to have well-established commercial services. Public Serv has the strongest correlation with Residence (0.912), indicating that public services tend to be built in more residential areas. The correlation between Business and Company is the weakest (0.680), indicating that there is spatial heterogeneity between these two types of urban functional areas. Overall, the city is a spatially coherent and multi-functional mixed complex system, and there is spatial heterogeneity as well as similarity(Calafiore et al., 2021).

5.3. Identification and pattern classification

In the previous section we divided the study area into 8 functional zones, and in this section we evaluate the results of this division. We calculate the distribution of POI in different urban functional zones by the following two indicators. Indicator 1 is POI Density (PD), which aims to reveal the density of various types of POI in different functional zones,

[6] map.baidu.com.

Table 4
Calculation results of PD and EF indicators for each functional zone.

	F_{Scenic}		$F_{University}$		$F_{Company}$		$F_{Residence}$		$F_{PublicServ}$		$F_{Business}$		F_{Other}	
Type	PD	EF	PD	EF	PD	EF	PD	EF	PD	EF	PD	EF	PD	EF
Caf/Tea	12.06601	0.684405	53.50603	0.937679	17.17198	0.634808	174.643	1.214939	65.53223	0.935219	243.7238	1.499436	93.44957	1.198794
RoadFac	0.080871	0.588358	0.135085	0.303638	0.226866	1.075693	1.200239	1.070948	0.513403	0.939754	0.994718	0.784924	0.558356	0.918705
Adr/Loc	1.180723	0.084881	1.829791	0.040641	2.600229	0.121828	53.50927	0.471787	9.947184	0.179917	13.30436	0.103738	10.35539	0.168363
TourAtr	42.2634	37.16341	4.015715	1.090978	0.29667	0.170019	4.291264	0.462797	3.392127	0.750468	3.783482	0.360847	3.171838	0.630783
PubFac	2.329097	2.430928	1.535059	0.495008	0.855109	0.581673	4.784513	0.612458	2.9429	0.772803	5.861733	0.663577	3.127263	0.738188
TransServ	12.84238	1.710443	24.53639	1.009662	13.14075	1.14066	60.02016	0.980425	36.46078	1.221795	66.59284	0.961992	38.22862	1.151515
ShopMal	0.582274	0.248669	3.413972	0.450459	2.024339	0.563443	12.39699	0.649328	3.914698	0.420631	14.42342	0.668101	6.20526	0.599338
TrabsFac	7.569564	1.072483	37.98351	1.662711	17.4861	1.614676	53.47639	0.929257	29.28231	1.04384	52.59574	0.808259	33.79462	1.082892
Bank/Fina	1.293943	0.357279	10.4384	0.890489	6.805968	1.224771	39.64898	1.342698	17.67573	1.227944	40.7124	1.219269	19.31255	1.206007
Sci/Edu	7.812179	0.883867	85.10368	2.97485	11.74466	0.86602	65.56921	0.909848	34.5997	0.984908	96.06138	1.17881	39.9295	1.021707
MotServ	0	0	0.073683	0.083634	0.122158	0.292488	1.101589	0.496348	0.953463	0.881303	0.497359	0.198182	0.342521	0.284589
CarServ	0.760191	0.217827	4.273605	0.378342	8.812856	1.645802	31.93786	1.122401	33.31619	2.401886	15.00959	0.466485	15.53309	1.006616
CarRepa	0.355834	0.211669	2.677143	0.49202	2.792192	1.082498	12.63539	0.921832	15.6863	2.347678	6.288041	0.4057	7.774757	1.045957
CarSale	0.145569	0.37622	0.196488	0.156896	1.378645	2.3222	2.786856	0.883371	7.609367	4.948026	1.598655	0.448136	1.205861	0.704839
Residen	3.687736	0.444121	63.44093	2.360552	16.09001	1.262906	59.50225	0.87888	28.10882	0.851713	56.98315	0.744336	41.05559	1.118233
LivServ	17.12857	0.611167	65.60229	0.723203	35.51319	0.82585	290.0632	1.269362	113.4712	1.01867	286.2835	1.107939	142.9392	1.153474
IndoFac	0.097046	0.13408	0.749109	0.319767	0	0	0.279508	0.047362	1.650224	0.573639	36.28946	5.438108	4.321395	1.350295
Spr/Rec	7.472519	1.633179	18.92421	1.27787	6.230078	0.887428	42.18921	1.130896	18.48251	1.016335	57.23183	1.356704	24.07735	1.190127
ComuServ	3.073114	0.767099	13.87694	1.070209	4.781629	0.777896	30.54032	0.934976	15.29208	0.960391	62.36529	1.688482	20.18762	1.139661
Hospital	3.574517	0.601802	15.11726	0.786343	5.793798	0.63573	70.01667	1.445746	22.3422	0.946393	56.25488	1.027253	30.21223	1.150369
Gov/Pub	6.291796	1.05419	12.56292	0.650335	6.928126	0.756542	33.80398	0.694651	29.47484	1.242524	36.34275	0.660454	22.01987	0.834405
Factory	4.852285	0.542426	18.44527	0.637062	47.04843	3.427776	60.0777	0.823687	47.74648	1.342904	95.70612	1.160419	41.14944	1.040342

Note: PD:POI Density(/km^2);EF:Enrichment Factor.

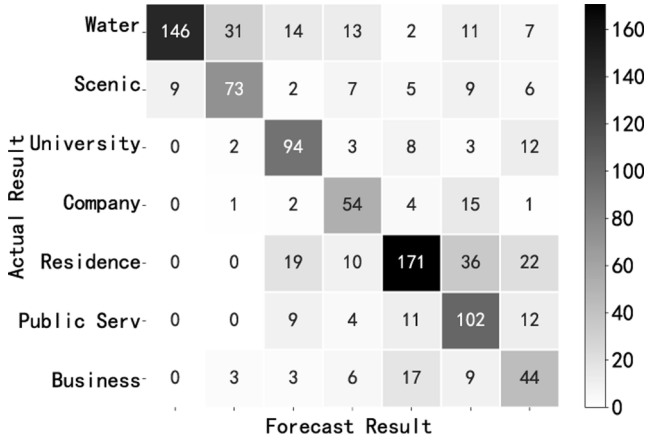

Fig. 11. Classification confusion matrix of GeoSemantic2vec algorithm (kappa coefficient 0.615).

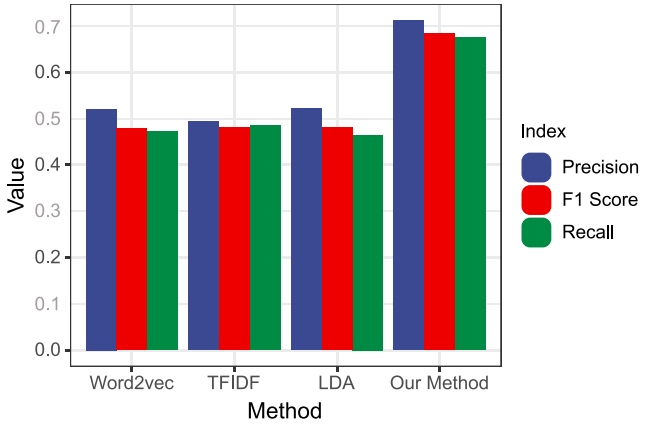

Fig. 12. Accuracy comparison of Word2vec, TFIDF, LDA and GeoSemantic2vec algorithms (after testing, our method Precision is 0.712, recall is 0.676, F1 score is 0.683).

and its calculation formula is as follows:

$$PD_i^q = N_i^q / A_i \tag{5}$$

PD_i^q represents the density of class q POIs in functional area i, N_i^q represents the number of class q POIs in functional area i, and A_i represents the total area of functional area i. As shown in Table 2, the number of different kinds of POIs varies. Some POIs are common and some POIs are rarer, which can lead to biased estimation results of PD metrics. We mimic the TFIDF algorithm in natural language processing and use the indicator2 (the enrichment factor, EF) to overcome this bias, which is calculated as follows:

$$EF_i^q = (N_i^q / N_i) / (N^q / N) \tag{6}$$

where EF_i^q represents the EF value of the class q POI in functional area i, represents the total number of POIs in functional area i, N^q represents the total number of class q POIs. represents the total number of all POIs, which is 500,634 in this paper.

In Table 4, we calculate the PD and EF values for each urban functional zone. The F_{Scenic} functional area has the highest POI density of TourAtr type (PD = 42.263, EF = 37.163), which is very distinctive. We classify these regions as Scenic Area type, which take on the function of recreation in the city. Similarly, the EF value of $F_{University}$ functional area Sci/Edu type POI is the highest, along with the high density of Residen type, LivServ type, TransServ type and Caf/Tea type POI. A college or university is a mixture of multiple types of POIs, and the model extracts this particular mixed structure. In addition to having various living service facilities provided to students and faculty members, education is the most important function undertaken by colleges and universities.

The $F_{Company}$ functional area is also easy to identify, and he has the highest EF value of Factory type POI. The $F_{Residence}$ functional area, as the

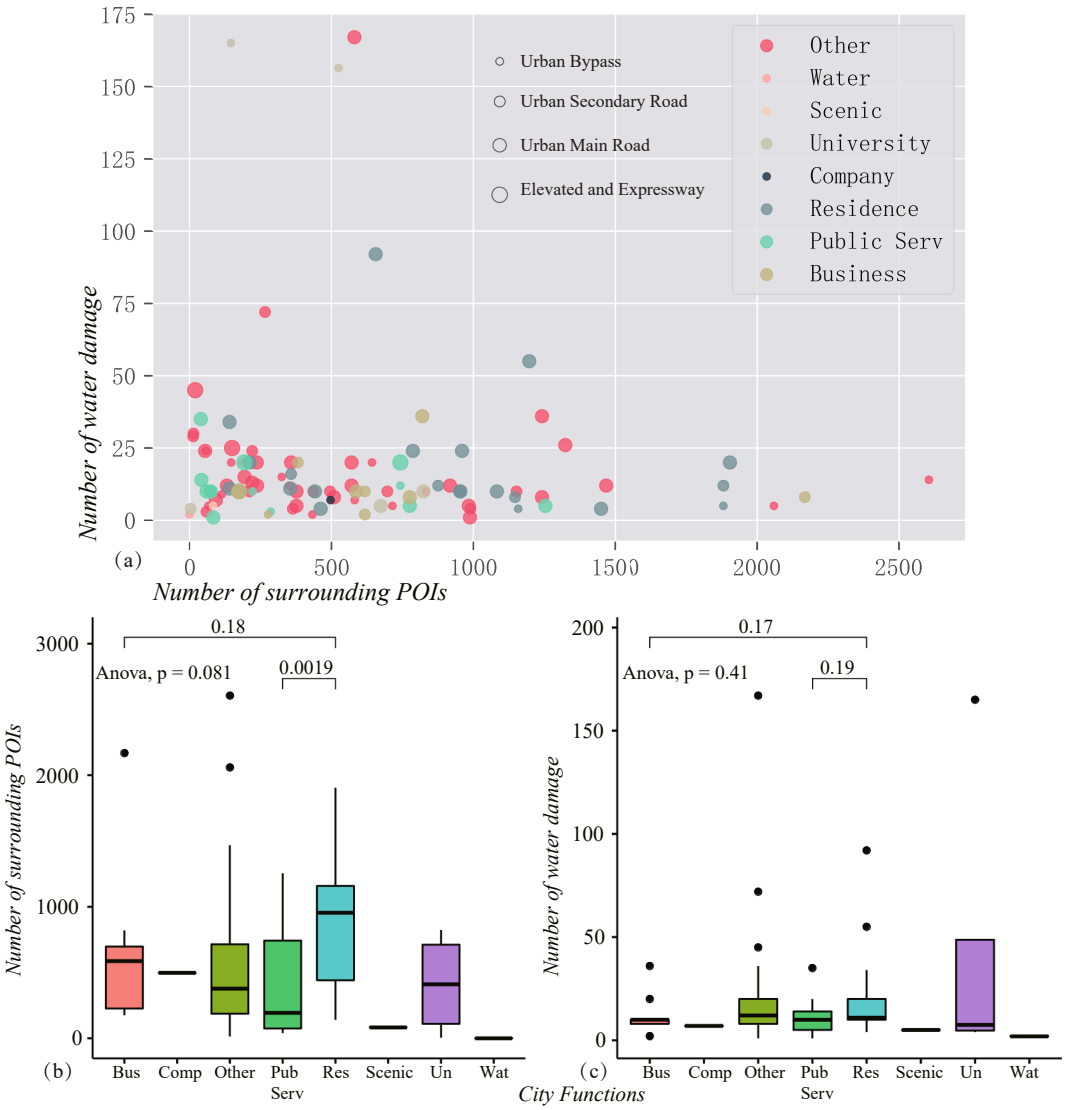

Fig. 13. Statistical results of waterlogged locations (Figure A shows the number of POIs around the waterlogged location on the horizontal axis, the flooding-prone frequency on the vertical axis, the color of the dots represents different urban functions, and the size of the dots represents the level of the road. Figure B shows the prosperity of the surrounding area for different urban functions of waterlogged locations, and Figure C shows the flooding susceptibility of different urban functions of waterlogged locations).

most relevant functional area for residents, is rich in all kinds of living service facilities, and the POI distribution is similar to that of the $F_{Business}$ functional area. The difference is that the $F_{Business}$ functional area has more IndoFac type POIs and less Adr/Loc type POIs. The BERT model could extract the semantic information of POI names. Even if the density of POIs is the same, the difference of semantic information will make the vector embedding representation of the two types of functional areas different. The $F_{PublicServ}$ functional area has the highest EF value of Gov/Pub type POI, and the vast majority of public service institutions are distributed here. Finally, the F_{Other} functional zone does not show significant spatial aggregation characteristics at the spatial scale used in the paper, and the urban functions are more ambiguous.

5.4. Algorithm accuracy comparison and effectiveness evaluation

We use the Word2vec method, TFIDF method and LDA method mentioned in the previous paper as the comparison baseline for feature extraction, and use the k-means algorithm to cluster the feature extraction results to evaluate the accuracy of our algorithm. We used a crowdsourcing approach and invited a dozen of master's students with a background in GIS or urban planning to judge the type of urban functions at the sampled locations. After eliminating locations where it was difficult to accurately determine the function types, we kept 1,012 judgment results and verified them using precision, recall and F1 score commonly used in machine learning.

From the confusion matrix (Fig. 11), we can see that there are some particularly confusing locations such as Water and Scenic, Residence, PublicServ and Business, which may be due to the fact that spatially uniform sampling breaks up functional areas, and one area will contain multiple functions with more information on special types of POI features. In addition, even though we define the attributes of functional areas, there will still be bias in the perception of functional areas by different volunteers. As shown in Fig. 12, our algorithm achieved a test score of 0.712 for Precision, 0.676 for Recall, and 0.683 for F1, which is about 15% improvement compared to the baseline algorithm.

5.5. Socio-economic information mining of waterlogging locations

We plot the road class, the city function, the number of peripheral POIs, and the number of times the water accumulates. As shown in

Fig. 13A, the most serious waterlogging is in the residential and universities areas, which have dense POIs around these waterlogging locations, more developed economies, and higher road grades. The risk caused by waterlogging in these areas is greater, which seriously affect the life of the surrounding residents and traffic travel (Zeng et al., 2020).

We grouped the locations of waterlogging according to urban functions and discussed them using one-way analysis of variance (ANOVA). As shown in Fig. 13B, there is a significant difference in the number of POIs around public serv and residence functional zones containing water hazards ($p = 0.002$). Residence functional areas have more POI information, followed by commercial areas and universities. At the confidence level of $p = 0.081$, there is a significant difference in the number of POIs in the different functional zones of the city. In contrast, the difference in the number of waterlogging in different functional areas was not significant ($p = 0.41$). Even though the number of waterlogging in the two areas is similar, the impact on the city varies greatly. The impact of flooding is not only related to the severity of the hazard, but also to the socio-economic factors surrounding the location of flooding, with flooding in busy areas and major transportation routes causing more serious impacts on the city. It means that mining the socio-economic attributes of different functional areas has more important value for the differentiation and research of flooding risk (Wang et al., 2015).

6. Conclusion and prospect

Urban flooding has a wide range of impact and high intensity in a short period of time, and its risk detection is a very meaningful research direction (Kankanamge et al., 2020). We took the July-September floods in Wuhan as a research object and provided an effective solution for flood risk detection and assessment using social sensing data. The BERT-Bi-LSTM-CRF model is introduced into the identification of flooding locations and irrelevant locations for the first time. It greatly improves the identification accuracy based on the pre-trained prior knowledge and the semantic information of the context in social media texts. Considering the different social attributes of different waterlogged location (Steiger et al., 2016), we proposed a new GeoSemantic2vec algorithm for urban functional zone. The algorithm combines semantic information about the location of flooding and mines it for socio-economic information.

Data such as POI, cab tracks, Twitter, Flickr photos, etc., which record human behavioral activities, are abundant in the urban space. By using these spontaneous VGI data, we can make use of the semantic information to portray the functional areas of the city and build a “City Portrait”. The GeoSemantic2vec method proposed in this paper is suitable for areas with more human activities and more developed socio-economics, where the density of POIs is high and can provide rich semantic information. The method can use the historical POI data of multiple years to analyze the land use change trend of cities; and use the POI data of multiple cities to build a global City Portrait database.

The paper still has certain shortcomings; it does not consider the floor area of POI for weighting; the urban functional areas classified as Other need to continue to be refined; due to the limitation of arithmetic power, the spatial resolution we set is low, and the semantic information obtained is relatively limited; and the results of urban function identification, and the classification criteria of remote sensing images do not fully match, which is pending for new supervised classification algorithms. In the future, we will try to use small samples of urban social sensing data for fast real-time, large-scale flood location detection and risk assessment.

Funding

This research was supported by the National Key R&D Program (No.2018YFB2100500), National Nature Science Foundation of China (Nos. 41971351, 41771422, 41890822), the Fundamental Research Funds for the Central Universities (No. 2042020kf0011), China Scholarship Council (202106270048) and Creative Research Groups of Natural Science Foundation of Hubei Province of China (No.2016CFA003).

CRediT authorship contribution statement

Yan Zhang: Conceptualization, Methodology, Writing – original draft. **Zeqiang Chen:** Writing – review & editing. **Xiang Zheng:** Visualization. **Nengcheng Chen:** Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University. Thanks to Dr. Shen Gaoyun for providing the hydrological data, and with the assistance of Mengtian Wen, Xianghui Liao, Meijuan Yang, Yue Gong, Wenzhe Huang, Jingjing Liu, Hui Lin, and Yingxue Yan for comparing the model accuracy, and Jie Liu for helping with the formatting.

References

Chang, H., Pallathadka, A., Sauer, J., Grimm, N., Zimmerman, R., Cheng, C., et al., 2021. Assessment of urban flood vulnerability using the social-ecological-technological systems framework in six us cities. Sustain. Cities Soc. 102786. doi:10.1016/j.scs.2021.102786.

Huang, X., 2020. Remote sensing and social sensing for improved flood awareness and exposure analysis in the big data era (Ph.D. thesis). University of South Carolina. doi:10.13140/RG.2.2.33524.17281.

Wan, C.F., Fell, R., 2004. Investigation of rate of erosion of soils in embankment dams. J. Geotech. Geoenviron. Eng. 130 (4), 373–380. https://doi.org/10.1061/(ASCE)1090-0241(2004)130:4(373).

Rakwatin, P., Sansena, T., Marjang, N., Rungsipanich, A., 2013. Using multi-temporal remote-sensing data to estimate 2011 flood area and volume over chao phraya river basin, thailand. Remote Sens. Lett. 4 (3), 243–250. https://doi.org/10.1080/2150704X.2012.723833.

Sun, D., Li, S., Zheng, W., Croitoru, A., Stefanidis, A., Goldberg, M., 2016. Mapping floods due to hurricane sandy using npp viirs and atms data and geotagged flickr imagery. Int. J. Digital Earth 9 (5), 427–441. https://doi.org/10.1080/17538947.2015.1040474.

Hultquist, C., Cervone, G., 2020. Integration of crowdsourced images, usgs networks, remote sensing, and a model to assess flood depth during hurricane florence. Remote Sens. 12 (5), 834. https://doi.org/10.3390/rs12050834.

Syifa, M., Park, S.J., Achmad, A.R., Lee, C.W., Eom, J., 2019. Flood mapping using remote sensing imagery and artificial intelligence techniques: a case study in Brumadinho, Brazil. J. Coast. Res. 90(SI), 197–204. doi:10.2112/SI90-024.1.

Sharma, T.P.P., Zhang, J., Koju, U.A., Zhang, S., Bai, Y., Suwal, M.K., 2019. Review of flood disaster studies in nepal: A remote sensing perspective. Int. J. Disaster Risk Reduction 34, 18–27. https://doi.org/10.1016/j.ijdrr.2018.11.022.

Schnebele, E., Cervone, G., Kumar, S., Waters, N., 2014. Real time estimation of the calgary floods using limited remote sensing data. Water 6 (2), 381–398. https://doi.org/10.3390/w6020381.

Schnebele, E., Cervone, G., 2013. Improving remote sensing flood assessment using volunteered geographical data. Natural Hazards Earth Syst. Sci. 13 (3), 669–677.

Xu, Lei, Chen, Nengcheng, Chen, Zeqiang, et al., 2021. Spatiotemporal forecasting in earth system science: Methods, uncertainties, predictability and future directions. Earth-Science Reviews 222. https://doi.org/10.1016/j.earscirev.2021.103828.

Yan, Y., Feng, C.C., Huang, W., Fan, H., Wang, Y.C., Zipf, A., 2020. Volunteered geographic information research in the first decade: a narrative review of selected journal articles in giscience. Int. J. Geogr. Inf. Sci. 34 (9), 1765–1791.

Cowie, S., Arthur, R., Williams, H.T., 2018. @ choo: Tracking pollen and hayfever in the uk using social media. Sensors 18 (12), 4434. https://doi.org/10.3390/s18124434.

Zou, L., Lam, N.S., Cai, H., Qiang, Y., 2018. Mining twitter data for improved understanding of disaster resilience. Ann. Am. Assoc. Geogr. 108 (5), 1422–1441. https://doi.org/10.1080/24694452.2017.1421897.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., et al., 2015. Social sensing: A new approach to understanding our socioeconomic environments. Ann. Assoc. Am. Geogr. 105 (3), 512–530. https://doi.org/10.1080/00045608.2015.1018773.

Forrest, S.A., Trell, E.M., Woltjer, J., 2020. Socio-spatial inequalities in flood resilience: Rainfall flooding in the city of arnhem. Cities 105, 102843. https://doi.org/10.1016/j.cities.2020.102843.

Kaufhold, M.A., Bayer, M., Reuter, C., 2020. Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. Inf. Process. Manage. 57 (1), 102132 https://doi.org/10.1016/j.ipm.2019.102132.
Atefeh, F., Khreich, W., 2015. A survey of techniques for event detection in twitter. Comput. Intell. 31 (1), 132–164. https://doi.org/10.1111/coin.12017.
Wang, R.Q., Hu, Y., Zhou, Z., Yang, K., 2020. Tracking flooding phase transitions and establishing a passive hotline with ai-enabled social media data. IEEE Access. https://doi.org/10.31223/osf.io/yv6g5.
de Bruijn, J.A., de Moel, H., Jongman, B., de Ruiter, M.C., Wagemaker, J., Aerts, J.C., 2019. A global database of historic and real-time flood events based on social media. Scientific Data 6 (1), 1–12. https://doi.org/10.1038/s41597-019-0326-9.
Barker, J., Macleod, C.J., 2019. Development of a national-scale real-time twitter data mining pipeline for social geodata on the potential impacts of flooding on communities. Environ. Model. Software 115, 213–227. https://doi.org/10.31224/osf.io/5d3sr.
Wang, Z., Ye, X., 2018. Social media analytics for natural disaster management. Int. J. Geogr. Inf. Sci. 32 (1), 49–72. https://doi.org/10.1080/13658816.2017.1367003.
Zahra, K., Imran, M., Ostermann, F.O., 2020. Automatic identification of eyewitness messages on twitter during disasters. Inf. Process. Manage. 57 (1), 102107 https://doi.org/10.1016/j.ipm.2019.102107.
Wang, Z., Ye, X., Tsou, M.H., 2016. Spatial, temporal, and content analysis of twitter for wildfire hazards. Nat. Hazards 83 (1), 523–540. https://doi.org/10.1007/s11069-016-2329-6.
Robinson, B., Power, R., Cameron, M., 2013. A sensitive twitter earthquake detector. In: Proceedings of the 22nd international conference on world wide web, pp. 999–1002. https://doi.org/10.1145/2487788.2488101.
Yang, J., Yu, M., Qin, H., Lu, M., Yang, C., 2019. A twitter data credibility framework–hurricane harvey as a use case. ISPRS Int. J. Geo-Inf. 8 (3), 111. https://doi.org/10.3390/ijgi8030111.
Zou, L., Lam, N.S., Shams, S., Cai, H., Meyer, M.A., Yang, S., et al., 2019. Social and geographical disparities in twitter use during hurricane harvey. Int. J. Digital Earth 12 (11), 1300–1318. https://doi.org/10.1080/17538947.2018.1545878.
Wang, Z., Lam, N.S., Obradovich, N., Ye, X., 2019. Are vulnerable communities digitally left behind in social responses to natural disasters? an evidence from hurricane sandy with twitter data. Appl. Geogr. 108, 1–8. https://doi.org/10.1016/j.apgeog.2019.05.001.
Yuan, N.J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., Xiong, H., 2014. Discovering urban functional zones using latent activity trajectories. IEEE Trans. Knowl. Data Eng. 27 (3), 712–725. https://doi.org/10.1109/TKDE.2014.2345405.
Zhang, D., Wan, J., He, Z., Zhao, S., Fan, K., Park, S.O., et al., 2016. Identifying region-wide functions using urban taxicab trajectories. ACM Trans. Embedded Comput. Syst. 15 (2), 1–19. https://doi.org/10.1145/2821507.
Wang, Y., Gu, Y., Dou, M., Qiao, M., 2018. Using spatial semantics and interactions to identify urban functional regions. ISPRS Int. J. Geo-Inf. 7 (4), 130. https://doi.org/10.3390/ijgi7040130.
Yuan, J., Zheng, Y., Xie, X., 2012. Discovering regions of different functions in a city using human mobility and pois. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 186–194. https://doi.org/10.1145/2339530.2339561.
Zhang, Yan, Chen, Nengcheng, Du, Wenying, et al., 2021. Multi-source sensor based urban habitat and resident health sensing: A case study of Wuhan, China. Build. Environ 198. https://doi.org/10.1016/j.buildenv.2021.107883.
Zhang, X., Li, W., Zhang, F., Liu, R., Du, Z., 2018. Identifying urban functional zones using public bicycle rental records and point-of-interest data. ISPRS Int. J. Geo-Inf. 7 (12), 459. https://doi.org/10.3390/ijgi7120459.
Sun, Y., Fan, H., Li, M., Zipf, A., 2016. Identifying the city center using human travel flows generated from location-based social networking data. Environ. Plann. B: Plann. Design 43 (3), 480–498. https://doi.org/10.1177/0265813515617642.
Aizawa, A., 2003. An information-theoretic perspective of tf–idf measures. Inf. Process. Manage. 39 (1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3.
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.
Bosch, A., Zisserman, A., Muñoz, X., 2006. Scene classification via plsa. In: European conference on computer vision. Springer, pp. 517–530. https://doi.org/10.1007/11744085_40.
Gao, S., Janowicz, K., Couclelis, H., 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. Trans. GIS 21 (3), 446–467. https://doi.org/10.1111/tgis.12289.
Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., et al., 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. Int. J. Geogr. Inf. Sci. 31 (4), 825–848. https://doi.org/10.1080/13658816.2016.1244608.
Liu, K., Yin, L., Lu, F., Mou, N., 2020. Visualizing and exploring poi configurations of urban regions on poi-type semantic space. Cities 99, 102610. https://doi.org/10.1016/j.cities.2020.102610.
Zhai, W., Bai, X., Shi, Y., Han, Y., Peng, Z.R., Gu, C., 2019. Beyond word2vec: An approach for urban functional region extraction and identification by combining place2vec and pois. Comput. Environ. Urban Syst. 74, 1–12. https://doi.org/10.1016/j.compenvurbsys.2018.11.008.
Angelov, D., 2020. Top2vec: Distributed representations of topics. arXiv preprint arXiv:200809470. doi:https://arxiv.org/abs/2008.09470.
Tay, Y., Dehghani, M., Bahri, D., Metzler, D., 2020. Efficient transformers: A survey. arXiv preprint arXiv:200906732. doi: Efficient transformers: A survey.
Yan, B., Janowicz, K., Mai, G., Gao, S., 2017. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In: Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 1–10. https://doi.org/10.1145/3139958.3140054.
Campello, R.J., Moulavi, D., Sander, J., 2013. Density-based clustering based on hierarchical density estimates. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, pp. 160–172. https://doi.org/10.1007/978-3-642-37456-2_14.
McInnes, L., Healy, J., Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. doi:10.21105/joss.00861.
Lu, Y., Mei, Q., Zhai, C., 2011. Investigating task performance of probabilistic topic models: an empirical study of plsa and lda. Inf. Retrieval 14 (2), 178–203. https://doi.org/10.1007/s10791-010-9141-9.
Löwe, R., Böhm, J., Jensen, D.G., Leandro, J., Rasmussen, S.H., 2021. U-flood–topographic deep learning for predicting urban pluvial flood water depth. J. Hydrol. 126898.
Herbert, Z.C., Asghar, Z., Oroza, C.A., 2021. Long-term reservoir inflow forecasts: Enhanced water supply and inflow volume accuracy using deep learning. J. Hydrol. 601, 126676.
Pham, B.T., Luu, C., Van Phong, T., Trinh, P.T., Shirzadi, A., Renoud, S., et al., 2021. Can deep learning algorithms outperform benchmark machine learning algorithms in flood susceptibility modeling? J. Hydrol. 592, 125615.
Lei, X., Chen, W., Panahi, M., Falah, F., Rahmati, O., Uuemaa, E., et al., 2021. Urban flood modeling using deep-learning approaches in seoul, south korea. J. Hydrol. 601, 126684.
Romascanu, A., Ker, H., Sieber, R., Greenidge, S., Lumley, S., Bush, D., et al., 2020. Using deep learning and social network analysis to understand and manage extreme flooding. J. Contingencies Crisis Manage. 28 (3), 251–261.
Wang, R.Q., Mao, H., Wang, Y., Rae, C., Shaw, W., 2018. Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. Comput. Geosci. 111, 139–147.
Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. doi:https://arxiv.org/abs/1810.04805.
Zhang, H., Ren, F., Li, H., Yang, R., Zhang, S., Du, Q., 2020. Recognition method of new address elements in chinese address matching based on deep learning. ISPRS Int. J. Geo-Inf. 9 (12), 745. https://doi.org/10.3390/ijgi9120745.
Li, P., Luo, A., Liu, J., Wang, Y., Zhu, J., Deng, Y., et al., 2020. Bidirectional gated recurrent unit neural network for chinese address element segmentation. ISPRS Int. J. Geo-Inf. 9 (11), 635. https://doi.org/10.3390/ijgi9110635.
Tseng, H., Chang, P.C., Andrew, G., Jurafsky, D., Manning, C.D., 2005. A conditional random field word segmenter for sighan bakeoff 2005. In: Proceedings of the fourth SIGHAN workshop on Chinese language Processing.
McInnes, L., Healy, J., Astels, S., 2017. hdbscan: Hierarchical density based clustering. J. Open Source Software 2 (11), 205. https://doi.org/10.21105/joss.00205.
Zhou, Q., Wang, J., Tian, L., Feng, L., Li, J., Xing, Q., 2021. Remotely sensed water turbidity dynamics and its potential driving factors in wuhan, an urbanizing city of china. J. Hydrol. 593, 125893 https://doi.org/10.1016/j.jhydrol.2020.125893.
Wang, D., Abdelzaher, T., Kaplan, L., 2015. Social sensing: building reliable systems on unreliable data. Morgan Kaufmann.
Chen, N., Zhang, Y., Du, W., Li, Y., Chen, M., Zheng, X., 2021. Ke-cnn: A new social sensing method for extracting geographical attributes from text semantic features and its application in wuhan, china. Comput. Environ. Urban Syst. 88, 101629.
Yang, J., Zhang, Y., Li, L., Li, X., 2017. Yedda: A lightweight collaborative text span annotation tool. arXiv preprint arXiv:171103759. doi:10.18653/v1/P18-4006.
Darabi, H., Choubin, B., Rahmati, O., Haghighi, A.T., Pradhan, B., Kløve, B., 2019. Urban flood risk mapping using the garp and quest models: A comparative study of machine learning techniques. J. Hydrol. 569, 142–154. https://doi.org/10.1016/j.jhydrol.2018.12.002.
Calafiore, A., Palmer, G., Comber, S., Arribas-Bel, D., Singleton, A., 2021. A geographic data science framework for the functional and contextual analysis of human dynamics within global cities. Comput. Environ. Urban Syst. 85, 101539 https://doi.org/10.1016/j.compenvurbsys.2020.101539.
Zeng, Z., Lan, J., Hamidi, A.R., Zou, S., 2020. Integrating internet media into urban flooding susceptibility assessment: A case study in china. Cities 101, 102697. https://doi.org/10.1016/j.cities.2020.102697.
Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., Bai, X., 2015. Flood hazard risk assessment model based on random forest. J. Hydrol. 527, 1130–1141. https://doi.org/10.1016/j.jhydrol.2015.06.008.
Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., Kamruzzaman, M., 2020. Determining disaster severity through social media analysis: Testing the methodology with south east queensland flood tweets. Int. J. Disaster Risk Reduction 42, 101360. https://doi.org/10.1016/j.ijdrr.2019.101360.
Steiger, E., Resch, B., Zipf, A., 2016. Exploration of spatiotemporal and semantic clusters of twitter data using unsupervised neural networks. Int. J. Geogr. Inf. Sci. 30 (9), 1694–1716. https://doi.org/10.1080/13658816.2015.1099658.