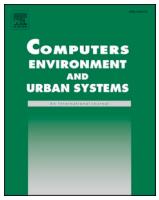
Computers, Environment and Urban Systems 88 (2021) 101629

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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



KE-CNN: A new social sensing method for extracting geographical attributes from text semantic features and its application in Wuhan, China

Nengcheng Chen [a,d], Yan Zhang [a,*], Wenying Du [a,d,*], Yingbing Li [b], Min Chen [b], Xiang Zheng [c]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China
[b] School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China
[c] School of Information Management, Wuhan University, Wuhan 430072, China
[d] National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan 430074, China

ARTICLE INFO

Keywords:
Social sensing
COVID-19
Geographic attributes
Social media
GIS

ABSTRACT

Social sensing is an analytical method to study the interaction between human and space through extracting reliable information from massive volunteered information data. During the ongoing COVID-19 pandemic, there are a large number of Internet social sensing data. However, most of them lack geographic attribute. In order to resolve this problem, this paper proposes a convolutional neural network geographic classification model based on keyword extraction and synonym substitution (KE-CNN) which could determine the geographic attribute by extracting the semantic features from text data. Besides, we realizes the non-contact pandemic social sensing and construct the co-word complex network by capturing the spatiotemporal behaviour of a large number of people. Our research found that (1) mining co-word network can obtain most public opinion information of pandemic events, (2) KE-CNN model improves the accuracy by 5%–15% compared with the traditional machine learning method. Through this method, we could effectively establish medical, catering, railway station, education and other types of text feature set, supplement the missing spatial data tags, and achieve a good geographical seamless social sensing.

1. Introduction

Social sensing is an unsolicited form of crowdsourcing that refers primarily to spatiotemporally tagged big data and the methods and applications based on such big data. Social sensing allows for the observation of human behaviour. These observations may be relevant and useful to research but are not produced for this purpose and unlikely to follow a consistent reporting structure (Cowie, Arthur, & Williams, 2018; Liu et al., 2015). Social sensing can extract the characteristics and dynamics of the social economy, culture, life, and other aspects from large geospatial data (Liu et al., 2015). Every individual in society plays the role of a sensor, providing real-time feedback to the surrounding environment. The popularity of intelligent devices and the development of 5G technology promote the acquisition of "social sensing" data. Compared to traditional data collection methods, online "crowd sourcing" data provided by intelligent devices are large in volume, real-time in velocity, and wide in coverage. Social sensing could show certain aspects of social characteristics by capturing the spatial behaviour patterns and using people as the smallest granularity unit, which primarily includes feelings about the geographic environment, activities and movements in geographical space; and social relations among individuals.

Compared to other data sources, social sensing data have two distinct advantages: 1) they can capture socioeconomic characteristics accurately; and 2) sensors for these data are individual people (Chen et al., 2018). Social sensing data is thus an important complement to high-precision geographic information data and sensor network data. The categories of social sensing data include Twitter, Sina Weibo, Flickr photo sharing data, (Zhang, Sun, Zheng, & Wang, 2020) app client usage, mobile phone signalling, (Lane et al., 2010) taxi trajectory, (Yu et al., 2019) street view of online map, (Zhang, Wu, Zhu, & Liu, 2019) point of Interests (POIs), (Persia et al., 2020; Xing, Meng, Hou, Cao, & Xu, 2017; Zhang et al., 2017) and remote sensing data of lights at nigh (Zhao, Cao, Zhang, Samson, & Chen, 2020). A social sensing method could extract reliable information from large Internet data collected from unknown and possibly unreliable sources (Wang, Abdelzaher, &

* Corresponding author at: State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China.
E-mail addresses: sggzhang@whu.edu.cn (Y. Zhang), duwenying@whu.edu.cn (W. Du).

https://doi.org/10.1016/j.compenvurbsys.2021.101629
Received 14 October 2020; Received in revised form 14 March 2021; Accepted 16 March 2021
Available online 8 April 2021

Kaplan, 2015).

Existing studies have demonstrated that social sensing can observe natural hazards or public health events and detect real-world events using location-based social media (LBSM) data (Arthur, Boulton, Shotton, & Williams, 2018; Huang, Cervone, & Zhang, 2017). Social sensing has also been used in disaster response and recovery such as earthquakes, (Avvenuti, Cresci, La Polla, Marchetti, & Tesconi, 2014) floods, (Arthur et al., 2018) typhoons, (Fan, Jiang, & Mostafavi, 2020; Gu et al., 2014) influenza (Allen, Tsou, Aslam, Nagel, & Gawron, 2016; Corley, Cook, Mikler, & Singh, 2010; Gao, Wang, Padmanabhan, Yin, & Cao, 2018) and urban poverty (Meng, Xing, Yuan, Wong, & Fan, 2020).

COVID-19 poses a risk to people gathering together and going out into public spaces. Therefore, governments have issued various orders that prohibit people-to-people contacts, which increases the difficulty of field investigations, and produces time discontinuities in disaster resistance data (e.g. community infection data, mask supply, the attitude of residents towards the pandemic, the real-time distribution of population and so on). Traditional sensing and assessment methods, such as surveys and interviews, are typically costly, risky, time-consuming, and small in scale (Eyre, De Luca, & Simini, 2020). To solve these problems, we must make use of social sensing, a fast, economic, and efficient survey method, to provide first-hand knowledge of the public's reactions to emergent events (Zou, Lam, Cai, & Qiang, 2018). However, masses of social sensing data (this paper mainly refers to geosocial media data) do not have geographical location information, making the data unusable.

Geographical attributes refer to the location information embedded behind social media texts, and this location information is not detailed address information, but rather functional locations. Unlike the traditional methods of extracting detailed location information from texts based on dictionary matching or called entity recognition, this paper proposes a neural-network-based approach to discover the geographical semantic information embedded in texts. Even if relevant words describing a location do not exist in the social media text, the proposed method can still infer the type of address (geographical attributes) behind the words from their probability of occurrence in the text or the probability of co-occurrence of different words. The proposed method treats each word and the combination between different words as an observation and makes inferences based on these observations (Di Rocco et al., 2020; Sengstock & Gertz, 2012).

The research questions this paper answers are as follows:

RQ1. How the emotional moods of Wuhan residents were affected by the lockdown?

RQ2. Considering the increasingly stringent privacy protection on social media platforms, how can the geographic attributes of social media texts be inferred from the frequency of word occurrence and co-word relationship when there is no address description in the text?

The rest of the paper is organized as follows. In Section 3, we introduce the study object, study area and study data. In Section 4, we introduce the model constructed in this paper. We make an empirical analysis and accuracy evaluation of our model in Section 5. Finally, we conclude and prospect our research work in Section 6.

2. Related works

Using social sensing, different types of potential themes in a city can be identified, and thus, the functional areas of a city can be identified (Gao, Janowicz, & Couclelis, 2017; Li et al., 2020). Additionally, social sensing could also be used to investigate the spatiotemporal characteristics of human mobility by exploiting explicit location footprints and mining latent demographic information implied on a social media platform (Luo, Cao, Mulligan, & Li, 2016). In addition to scientific research, commercial companies also provide products that use social sensing. Social sensing has been most widely recognized as an active tool for providing high-frequency updates of land-use types, real-time sensing of surrounding traffic congestion status (such as Google map, Tencent YiChuXing population thermal map), and evaluation of finer granularity of population distribution (Yang et al., 2019; Zhang, Li, Yang, Zheng, & Chen, 2020).

Many scholars have made insightful attempts to prevent and control the pandemic using social-sensing methods, (Rashid & Wang, 2021; Zhang, Li, et al., 2020) such as conducting timely research on topics of concern to different user groups, extracting early warning information from social media data, identifying confirmed COVID-19 cases quickly using social media, identifying urban risk areas based on pandemic data, and quickly promulgating information via the Internet (Chen, Lerman, & Ferrara, 2020; Jahanbin & Rahmanian, 2020; Li et al., 2020; Su et al., 2020; Zhang, Sun, et al., 2020). Certain scholars have built real-time social sensing and analytic systems that gather and circulate COVID-19 propagation data, and provide risk alerts (Rashid & Wang, 2021).

To analyse the social media data acquired by social-sensing systems, the latent Dirichlet allocation (LDA) topic modelling approach, which parses social media texts into a three-layer structure of word document topics, is most commonly used. Certain researchers argue that each document feature can be considered as an individual sensor. In this paper, the corpus is represented by a high-dimensional matrix after being trained by deep learning. Each dimension of the matrix can be considered to be a sensor to measure a signal of (and possibly abstract, unknown, or meaningless) geographic semantics (Sengstock & Gertz, 2012). Topic models represented by LDA can automatically discover the implied topic structure from many documents without prior knowledge (Huang, Li, & Shan, 2018). Geotagged tweets are analysed with clustering algorithms and topic models to study the spatiotemporal patterns of events and identify semantic content (Resch, Usländer, & Havas, 2018). Similarly, topic models are used to describe the spatiotemporal characteristics of catastrophic events and thus perform risk assessment. The same approach is used in the study of urban functions, where each POI is considered to be a word, the POIs around the sampled location are considered to be a document, and different functional attributes are assigned to the sampled location (Gao et al., 2017). In addition, certain scholars have explored the spatiotemporal and semantic clustering of Twitter data using unsupervised neural network method (Steiger, Resch, & Zipf, 2016) Wang & Stewart, 2015). The extraction of semantic information from web documents (news reports) has been studied by constructing ontology models with gazetteers. Studies have shown that social events can be perceived in temporal, spatial, and textual dimensions from social media data (Zhu et al., 2019). Also, other studies treat these three dimensions as a tuple, and knowing two elements of the tuple allow the inference of the other elements (Diaz, Poblete, & Bravo-Marquez, 2020).

In terms of extracting geographic attributes from social media texts, the aforementioned studies have two shortcomings: LDA methods are an unsupervised classification model, and the extraction results of urban function differ from the real classification criteria. Additionally, these methods require a library of pre-prepared place names that identify them in social media texts (Di Rocco et al., 2020). In this paper, we propose a convolutional neural network geographic classification model based on keyword extraction and synonym substitution (Keyword extraction and synonym substitution-Convolutional Neural Networks, KE-CNN) to overcome the two previously mentioned drawbacks and extract geographic attributes based on semantic features.

3. Study object and data

We used Wuhan as an example and tried to classify COVID-19-related social media data from a geographical viewpoint. Wuhan is the largest city in central China, with 13 districts covering an area of 8569 km^2 and a permanent population of 10,890,000. Among the 13 districts, Wuchang, Hongshan, Jianghan, Jiang'an, Qingshan, Hanyang, and Qiaokou are located in the primary urban area. Wuhan has been the most severely affected city in China during the COVID-19 outbreak, with

the largest number of confirmed cases, the longest closure time, and the most intense public discussion. Sina Weibo, also known as "Chinese Twitter", is one of China's largest social media platforms. To date, it has approximately 462,000,000 active users who share and express their opinions by sending short texts within 400 words. After COVID-19 broke out, there were also heated discussions on Weibo, which brought about a large amount of public opinion data with geographic information. To mine useful information from the data, we collected some COVID-19-related original Weibo data from Wuhan users from January 10 to February 17, 2020. First, we developed an advanced Weibo crawler script based on the Scrapy crawler framework and used MongoDB for data storage. Then, we selected keywords that are related to COVID-19, such as "pneumonia", "pandemic situation", and "virus", to increase the volume of retrieved data. Finally, after duplicate removal of records based on the text field, we captured a total of 213,869 original Weibo records. A corpus of 52,776 words was constructed from these Weibo records, and this corpus was used to train the generated word vector model.

The location information contained in social media data is typically in the form of address descriptions; thus, we geocoded them to obtain the specific latitude and longitude. In this study, we infer the real location of a Weibo check-in record using statistics and matching with POIs, which were obtained through an application programming interface (API) provided by Baidu Map Services, the most widely used map service provider in China. All coordinates in this paper are transformed into the World Geodetic System 1984 (WGS84) datum.

In total, we obtained about 30,000 Weibo check-in records with geographic descriptions. However, the geographic descriptions of certain check-in locations, such as "Wuhan City" and "Hubei Province", were too vague to be used and were thus removed. Additionally, we want to learn the characteristics of the Weibo text in different geographical locations. To avoid chance errors, we chose 9 location types that had more than 50 tagged Weibo data points. After filtering, approximately 4313 Weibo records with detailed locations were retained and were distributed over 1033 POI locations, where each location may have multiple Weibo records. We classify Weibo records as located at a residence, tourism, business, medical, railway station, road traffic, education, catering, shopping and other types.

4. Study method

As mentioned in Section 2, the results of this study are expected to be interpretable rather than generate clusters using unsupervised training methods. Faced with disambiguated text, we must to map the data from the original space of symbolic words into a feature space by encoding various aspects of the relatedness in the text (e.g., lexical, syntactic, co-occurrence relations, frequency of occurrence, and semantics). Feature engineering is often an onerous task and may require external knowledge that are not always available or are difficult to obtain (Severyn & Moschitti, 2015). Traditional text feature engineering has N-gram methods (Brown, Della Pietra, Desouza, Lai, & Mercer, 1992; a tuple of N words that counts the probability of combining different N words together, with the total number of n-grams being the Nth power of the total number of words in the corpus) and one-hot methods (generate a feature vector of length of the total number of words in the corpus). Traditional methods can only learn local information; however, deep learning methods represented by a convolutional neural network (CNN) can learn more global information and could obtain weights and weighted sums of different word combinations for classification.

Combining the advantages of the word vector algorithm, term frequency inverse document frequency (TFIDF) algorithm and CNN algorithm, we propose a KE-CNN geographic classification model based on keyword extraction and synonym replacement. The method is divided into the following steps: (1) Graph-structured representation of social sensing data. (2) Knowledge embedding of social sensing data. (3) Geographical properties of fused knowledge. (4) Training the KE-CNN classifier. (5) Geographic Inference and Complementation. The advantage of this method is that it can use the powerful feature extraction ability of the CNN network to extract the detailed features of the corpus and words, and match the geographic location types corresponding to different features. Additionally, this method can also convert a corpus with an indefinite length into a unified format.

As shown in Fig. 1, the experimental process of this paper is mainly divided into four parts, data preprocessing and corpus construction (Section 3), corpus vectorization (Section 4.1), co-word network construction (Section 4.2) and semantic based geographic classification (Section 4.3). We mainly do the data capture and preprocessing, and encoding Weibo into a corpus with words as the smallest composition In Section 3, which is the basis of the next part. In Section 4.1, we mainly focus on the vectorization of corpus and the result of this part will be the model input of the Section 4.3. Based on the preprocessed corpus, the Section 4.2 will build a co-word complex network, which is the contrast and supplement of the Section 4.3.

Through the collection of the COVID-19 pandemic-related Weibo, we constructed the corpus using text segmentation, synonym merging, sentiment judgement, geographical label classification, etc. Additionally, we showed the sentiment attitude and Internet activity of different regions; extracted keywords of concern to the public; and transformed Internet-based information into a high-dimensional vector via a word embedding algorithm. Based on to the co-occurrence relationship of words, we constructed a complex network under the theme of COVID-19. Finally, we developed a matching model between the user's Weibo information and the ground object type, which could determine the geospatial entity category of text mapping by extracting word features.

4.1. Corpus vectorization

There are certain meaningless inflections, auxiliaries, punctuation, etc. in the corpus, and we filter the data using a dictionary of commonly used stop words and a user-defined dictionary. Then, based on the results of word segmentation, we used the word bag method to transform the text data into vectors and used a vector, in which all elements are 0 except for only one, which is equal to 1, to uniquely represent a word. However, this method will produce a serious problem called "the curse of dimensionality": when managing a long piece of text, it takes a long vector to discriminate all words in the text, which yields an extremely sparse text matrix with low information density. To mitigate this problem, we used Google's open source deep learning tool word2vec (Google, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) to train the input corpus and build embedding vectors. The Word2vec model considers the context relationship between words, sets a certain size of sampling window c, and performs sliding sampling on the input corpus in order. This study used the continuous bag of words (CBOW)-based word2vec model (Mikolov et al., 2013; Yao et al., 2017) that could predict the probability of word w_t occurrence based on context. The mathematical model of the CBOW method is as follows:

$$P\left(w_t|w_{t-c}^{t+c}\right) = P(w_t|\tau(w_{t-c}, w_{t-c+1}, \cdots, w_{t+c-1}, w_{t+c}))\quad (1)$$

τ where refers to a given of operation. In word2vec, this type is the word vector addition operation that adds the vectors of all adjacent words in the window.

The input corpus was transformed into the vector of a specified dimension using the word2vec tool, in which words with similar meanings have closer spatial distances. Based on the word vector, we can determine the famous king and queen formula $\vec{k}ing - \vec{m}an + \vec{w}oman = \vec{q}ueen$. In this paper, we took all 213,869 Weibos (the processing result of segmentation and filtering), which include 52,776 words as model inputs.

The word vector dimension after training was set to 100, and the output matrix (52,776 × 100 in size) was obtained. Each dimension of a word vector represents an attribute and words with similar attributes

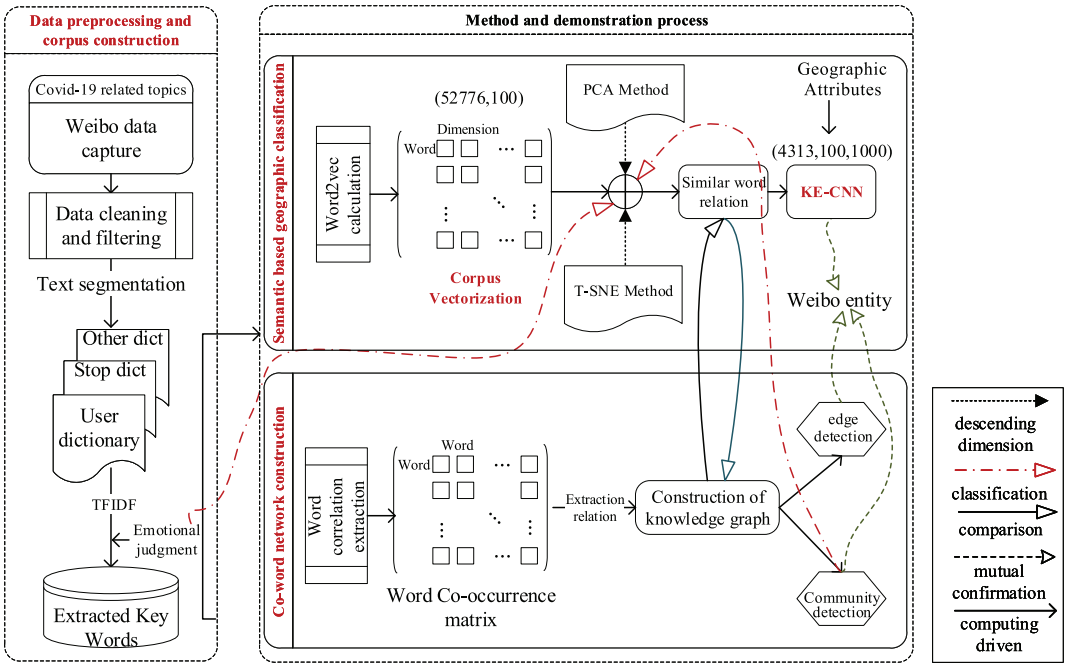


Fig. 1. KE-CNN based geographic attribute identification of social sensing data.

have similar values in such positions.

4.2. Co-word network construction

Co-word analysis is a social network analysis method that was first used in the field of bibliometrics, and has been widely used in various studies to explore hot topics and trends (Zheng, Wang, Zheng, & Liu, 2019). Co-word analysis does not rely on any a priori knowledge, making detection results more objective. For the words in Weibo texts, we displayed them using various dimensionality reduction methods and constructed the co-word network of keywords. Concurrently, related studies also demonstrate that co-word analysis can be used to observe

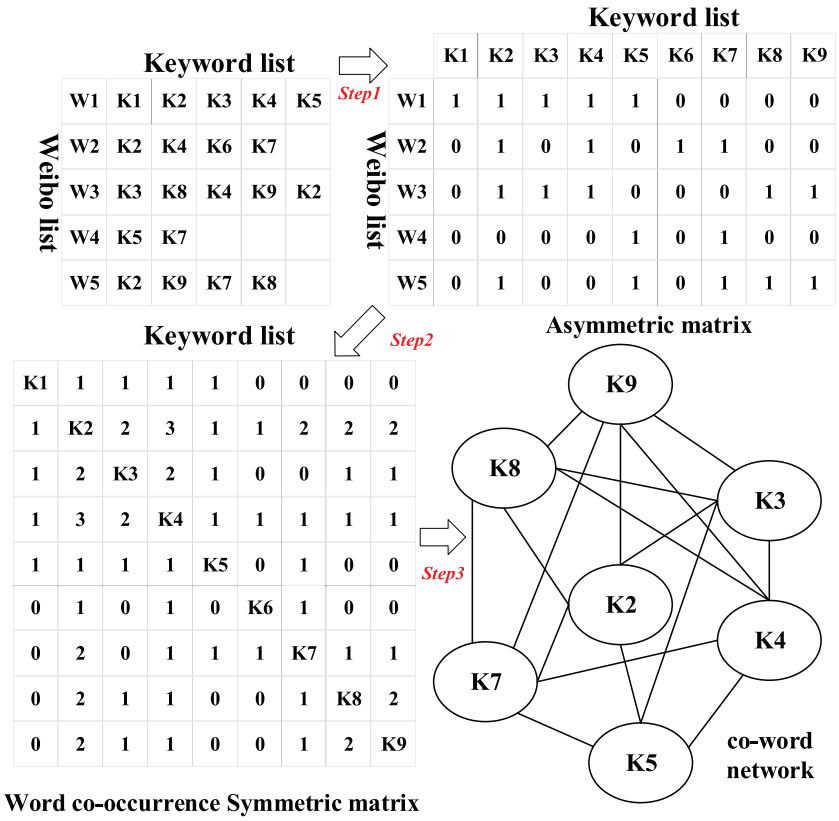


Fig. 2. Construction of co-word networks based on co-occurrence relations of Weibo words.

collective human reactions to extreme events using social media data, which is a unique analytical perspective that can help us improve our understanding of human sentiment states at different stages of major events and also demonstrate the potential causes of events and improve the efficiency of event mitigation (Li, Wang, Gao, & Shi, 2017). As an outbreak event, the COVID-19 pandemic generated many discussions on the Internet during the study period. As the event evolved, the importance and strength of the association of different words in the co-word network changed frequently, and the community of words (acquired using a community discovery algorithm) thus required frequent reconfiguring.

Depending on the word co-occurrence relationship between words, we constructed the co-word relationship matrix. If two words appeared together, the edge relationship between the two words would be developed. The more frequently the two words appeared together, the stronger their edge relationship would become. Based on the constructed co-word relationship matrix, we screened this type of edge relation using a given threshold and generated a complex network model Figure 2. The network describes the public's attitudes towards COVID-19, and the edges imply the changes and trends of the event, which provides us with a unique perspective for analysing society's reaction to the pandemic.

A complex network is the topological abstraction of real complex systems. A graph G can be used to build the model, which is composed of nodes and connecting edges with different weights. The nodes of the complex network are the collection of selected keywords in the corpus, and the edges of the complex network are the sets of co-occurrence relations:

$$G = \begin{bmatrix} N_{11} & \cdots & N_{1m} \\ \vdots & \ddots & \vdots \\ N_{m1} & \cdots & N_{mm} \end{bmatrix}_{m \times m} \tag{2}$$

where m is the number of network word nodes and $N_{i,\,j}$ represents the degree of connection from word node i to word node j. The higher the value $N_{i,\,j}$ is, the stronger the co-occurrence between words becomes.

The community structure is an important topological feature of complex networks, and a complex network is composed of several communities. The frequency of word co-occurrence in the same community is higher than that in different communities. In this paper, the Louvain algorithm was used to analyse the community structure of the complex network, and there are two steps: modularity optimization and network aggregation. Modularity is an important evaluation index that is used to define the degree of community structure and its mathematical definition is as follows: (Newman, 2003)

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{i,j} - \frac{k_i k_j}{2m}\right]\delta(C_i, C_j) \tag{3}$$

where $A_{i,\,j}$ represents the co-occurrence intensity between word i and word j; k_i represents the degree of node (sum of weighted edges of word node connection); m represents the total number of nodes in a complex network; and C_i represents the community with node i. When $C_i = C_j$, the value of function δ is 1; otherwise, it is 0. The closer the modularity Q is to 1, the better the effect of community division is.

The detailed steps of the algorithm are as follows:

(1) Each community was considered to be an independent community, and the number of initial communities was the same as that of node words. (2) All nodes were fused and agglomerated based on modularity gain until the local maximum value of modularity was reached (i.e.,no node could improve the network modularity, and the community structure no longer changed). (3) The community results were compressed; the nodes and the weights of the internal nodes were transformed into new nodes and weights; and the weights of the edges between the original word nodes were transformed into weights between the new communities.

The evaluation indices of complex networks are as follows:

- Node degree: in an undirected graph, the degree of a node is the sum of the edge weights connected with the node;
- Graph density: in complex networks, the higher the graph density is, the closer the network connection is;
- Modularity: the greater the modularity is, the more obvious the community structure is;
- Network diameter: the smaller the network diameter is, the better the accessibility between the nodes becomes.

4.3. Geographic attribute classification method

This section describes the core of the KE-CNN algorithm, which is based on the data augmentation method and text semantic feature extraction method to construct a geographic attribute classification model.

First, keywords are extracted from the corpus based on the TFIDF algorithm. Term frequency (TF) refers to the number of times a given word appears in a Weibo. The more frequently a word appears in a Weibo, the value of its TF increases. Conversely, the value of the inverse document frequency (IDF) increases as the word appears in fewer Weibo. A small IDF indicates that the word has a good classification ability. The IDF of a specific word can be obtained by dividing the total number of files by the number of files containing the word and then taking the logarithm of the quotient. The TFIDF of a word is the product of the TF and the IDF. The mathematical formula for calculating word weight by the TFIDF method is as follows:

$$F_{i,j} = \frac{x_{i,j}}{\sum\limits_{k} x_{k,j}} \times lg\frac{|W|}{1 + |\{j : z_i \in w_j\}|} \tag{4}$$

where, $x_{i,\,j}$ is the frequency of the entry (word) in Weibo w_j; $\sum\limits_{k} x_{k,j}$ is the total frequency of all the words in Weibo w_j; $|W|$ is the total number of Weibo in the corpus;and $|\{j : z_i \in w_j\}|$ is the total number of Weibo containing words z_i.

The TFIDF algorithm can reduce the weight of common words (i.e., the words that appear frequently in each type of document), and consider word frequencies. In this paper, the first 1000 words with the highest TFIDFs were selected as keywords. A matrix with a size of 100 × 1000 was used to represent a Weibo, in which 100 is the dimension of the word vector, and the number of keywords is 1000. Replacement using similar word vectors is an effective data augmentation scheme (Coulombe, 2018). If the words are not within the range of the extracted 1000 words, the closest words in the keyword vectors are calculated and replaced (synonym merging).The similarity between different words $R\left(\overleftarrow{d_i}, \overleftarrow{d_j}\right)$ is calculated by the cosine formula of the included angle:

$$\left(\overleftarrow{d_i}, \overleftarrow{d_j}\right) = cos\left(\overleftarrow{d_i}, \overleftarrow{d_j}\right) = \frac{\overleftarrow{d_i} \cdot \overleftarrow{d_j}}{\left|\overleftarrow{d_i}\right| \times \left|\overleftarrow{d_j}\right|} = \frac{\sum\limits_{k=1}^{n} wt_{ik} \times wt_{jk}}{\sqrt{\sum\limits_{k=1}^{n} wt_{ik}^2} \times \sqrt{\sum\limits_{k=1}^{n} wt_{jk}^2}} \tag{5}$$

We want to extract the text features of Weibos in the same location type. The data for the proposed study are 4313 Weibos distributed in 9 geographic attribute locations, which include estate, tourism, business, medical, railway, road & transport, education, catering, and shopping types. As shown in Table 1, there are more than 50 Weibos data for each type of location. Thus, we developed the relationship between the Weibo text and the real space, and these texts were used as model inputs for the next step of the study. Finally, the relatively dense (4313,100,1000) dimension matrix was generated as the input of the KE-CNN model.

In this study, 2 × 2 and 3 × 3 convolution kernels were used to build a 64-layer CNN model, and 9 types of locations were considered to be the targets of model classification. The model consists of three convolution layers, two max pooling layers, one flatter layer, and two dense layers. The input data were predicted using the softmax classifier and the ReLU

Table 1
Weibo location classification criteria.

Type	Explanation	Number
Residence	Residence-related POIs, such as residential areas, dormitories.	1472
Tourist	Tourism-related POIs, such as parks, museums, scenic spots,temples.	359
Business	Business-related POIs, such as stores,banks.	589
Medical	Medical-related POIs, such as hospitals, clinics, pharmacies,etc.	265
Railway station	Station-related POIs, such as train stations.	205
Road	Road traffic related POI,such as Jianghan Road.	958
Education	Education-related POIs, such as universities, research institutes.	212
Catering	Restaurant-related POIs, such as Chinese restaurants, cafes,snack bars.	50
Shopping	Shopping-related POIs, such as shopping centers, supermarkets.	203

Note: The 213,869 original Weibos contained 52,776 words as a corpus for word vector training. Of them, only 4313 had available detailed location information, distributed among 1033 POIs.

activation function. The total number of parameters of the model is 21,216,329. The corpus data were split into a training dataset (90%) and a testing dataset (10%) after shuffling. Fig. 3 shows a detailed and specific structure graph of this model.

In this experiment, we encountered the problem of model overfitting. Due to the relatively small number of training samples, as the number of training epochs increases, the loss function in the training set decreases, while the loss function in the verification set increases. These results indicate that overfitting occurred in the model. To mitigate this problem, we added two dropout layers (with parameters set to 0.2, losing 20% of the connections at random each time) before the two dense layers to make the model more robust and exhibit better generalizability. Adding the dropout layers cause the neural network eliminate certain connections randomly, and these connections will thus not participate in the calculation.

5. Results and discussion

5.1. Overall sensing results of COVID-19 in Wuhan, China

Calculating sentiment is typically performed with a plain Bayesian classifier. We randomly selected 1000 Weibos from a corpus of 200,000 Weibos for manual sentiment type discriminations, which were divided into negative, positive, and neutral. The labelled data were fed into a plain Bayesian classifier for calculation to learn the probability of occurrence of different words in Weibos for different sentiment types. The trained model scores the remaining Weibo to determine the probability that they belong to the positive type. In general, a probability value greater than 0.6 indicates that a Weibo exhibits a positive sentiment, below 0.4 indicates that a Weibo exhibits a negative sentiment, and otherwise, a Weibo exhibits a neutral sentiment.(Saranya & Jayanthy, 2017).

As mentioned in Section 3, we collected 1033 check-in locations and geographically decoded them into 9 feature types. We consider the mean value of the positive probability of all Weibos' sentiments at each location as the sentiment value of the location point. The higher the sentiment value, the higher the positive sentiment probability of Weibo users in that region; the lower the sentiment value is, the lower the negative sentiment probability value of users in that region. Using the widely distributed check-in points, we could sense the city's sentiment attitude geographically, and these results are shown in Fig. 4. Within the study interval, Jiang'an District and Jianghan District show a wide degree of discussion intensity. The number of extremely positive and extremely negative positions in the city is roughly equal, and the heat is not high. Additionally, there are certain ground feature nodes with more than 300 discussions in the study area, and the Weibo text typically exhibits neutral and positive sentiment. There is a heterogeneous mix of sentiments in the study area, with locations that exhibit high sentiment values with a less pronounced aggregation effect and locations that exhibit low sentiment values, thus indicating a staggered distribution in space.

5.1.1. Data distribution in different regions

After data normalization, we compare the attributes of seven urban districts across six dimensions: the average number of comments on a single Weibo, the average forwarding number of a single Weibo, the positive probability of Weibo, the number of confirmed cases, the number of permanent residents in the statistical yearbook in 2018, the number of confirmed COVID-19 cases, and the total number of Weibos. As shown in Fig. 5, the average number of comments on a single Weibo in Qingshan District is the highest, and the Weibos as a whole indicate negative sentiment. However, there is a difference between the total number of Sina Weibos and the permanent residents. Jiang'an District has the highest total number of Weibo discussions and the number of confirmed cases is second only to Wuchang District; however, the

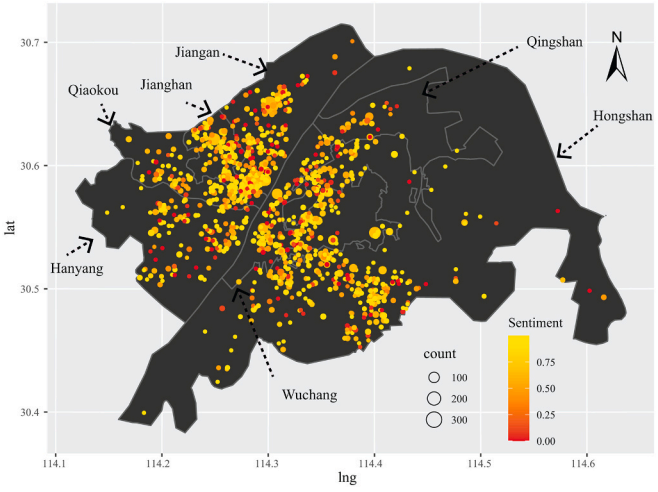


Fig. 4. Sentiment distribution of social sensing location in the main urban area of Wuhan.

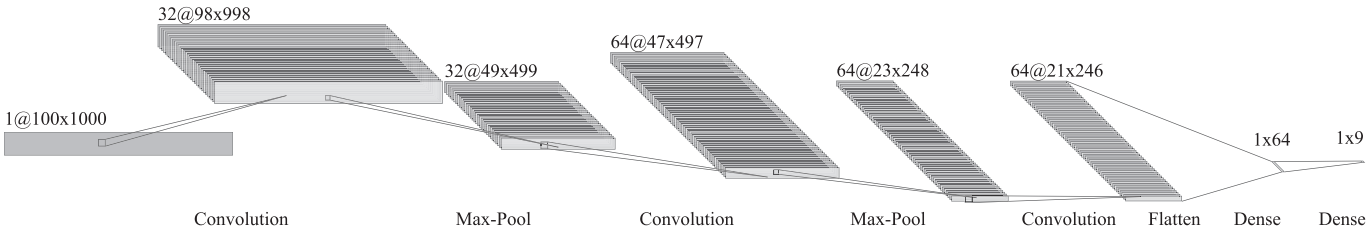


Fig. 3. Model Structure for Geographic Attribute Classification of Weibo Using KE-CNN.

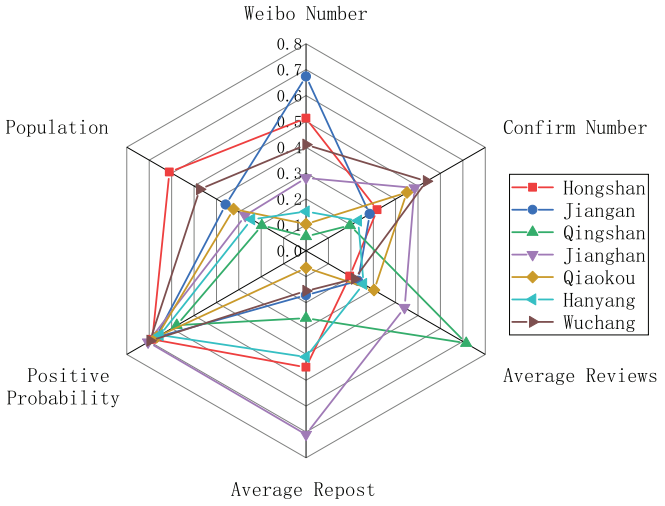


Fig. 5. Radar chart of normalized indicators for each region of Wuhan city. (Weibo Number, Population, Positive Probability, Average Repost, Average Reviews, Confirm Number.)

permanent population in Jiang'an District is only half of Hongshan District, which indicates that users are particularly concerned about the COVID-19 pandemic in these regions.

5.1.2. Different types of data distribution

Because a given location typically has multiple Weibos, we plotted Fig. 6 to show the difference in the number of Weibos at different locations. As shown on the x-axis of Fig. 6, there are different types of geographic properties, and the y-axis represents the number of Weibos made at the same location (Number of check-ins). Locations with the geographic attribute "railway station" have a much higher number of Weibos than other types in terms of the median and upper quantile, which is partly because COVID-19 exploded during China's Spring Festival, when approximately 3 billion were travelling. Conversely, Wuhan is the largest transportation hub in central China, with three major railroad passenger stations (Wuchang Station, Hankou Station, and Wuhan Station) carrying large volumes of passengers. The upper quantile of the commercial area is significantly higher than the other POI types; however, the gap between the median data and other geographic types is not wide, which shows that certain popular business districts' passenger flows consider the vast majority of the total business district passenger flow.

In the study area, residences (with 499 POIs) are the most widely distributed location type, accounting for more than half of the total

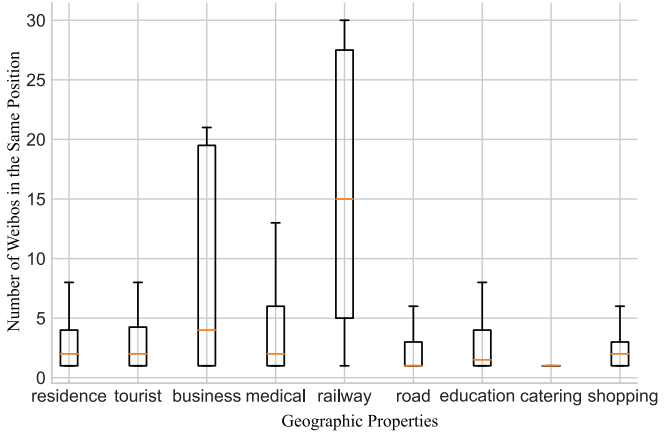


Fig. 6. Distribution of Weibo with different attribute POIs.

check-in locations. However, the number of check-in times at a single location is relatively low. The median residence type occurs approximately twice, which is similar to tourism, medical, and shopping. The median of restaurants and road and transport is 1; the distribution of check-in times of restaurant type is flatter. The type of medical institution has the upper quartile second only to business districts and railway stations, indicating that certain medical institutions have extremely high geographic heat.

This section discusses the spatial distribution of Weibo topics related to the pandemic situation in Wuhan by sampling data and also perceives the sentiment situation in Wuhan. The proportion of confirmed cases in Jianghan District is relatively high, and the discussion on the Internet is also more intense. Additionally, during the pandemic period, hot business districts, railway stations and medical institutions had more signs in times.

5.2. Situation analysis based on word resolution

5.2.1. Network modelling

We used the first 100 keywords extracted through the TFIDF method for visual display. The results of the complex network modelling of words are shown in Fig. 7. The line width of an edge between two words indicates the intensity of word co-occurrence:the higher the intensity, the more likely these two words appear in the same Weibo. The connection strengths between Wuhan-cheer up, Wuhan-pandemic, pneumonia-virus, medical-hospital, and Wuhan-materials are high. Additionally, the size of the node represents the importance of words in the network. As shown in Fig. 7, the top 100 keywords are divided into three communities with a graph density of 0.227, a modularity of 0.194, and a network diameter of 2, which indicates that the maximum distance between any two words is no more than 2.

The theme of community 2 uses "blessing", "victory" and "donation" as core words; community 1 uses "Wuhan", "pandemic situation", "cheer up", "masks", and "medical care"; and community 3 uses "pneumonia", "virus", "coronavirus", "infection", "cure", "new addition", and "diagnosis". The association of words between different communities (the probability of simultaneous occurrence) is lower than that of words in the same community.

5.2.2. Word properties

Neural networks have been successfully applied to many NLP tasks. However, neural networks are vector-based models, which makes it difficult to build sentence meaning from the meanings of words and phrases (Li, Chen, Hovy, & Jurafsky, 2015). We can extract sparse latent features via dimension reduction to obtain the geographical features of social media data (Sengstock & Gertz, 2012)]. For words in Weibo texts, we displayed them using various dimensionality reduction methods and constructed the co-word network of keywords. In NLP tasks, T-Distributed Stochastic Neighbor Embedding (T-SNE) methods have been shown to be more interpretive than traditional Principal Components Analysis (PCA) methods with Locally Linear Embedding (LLE) methods to obtain good clustering of high-dimensional sentiment, numerical, and community attributes on a two-dimensional plane (Li et al., 2015).

We used the word2vec method to train the corpus and obtained the 100-dimensional vector of each word. We tried using PCA, T-SNE (Maaten & Hinton, 2008), and LLE to reduce the dimension of the top 100 keywords. The sentiment attributes (positive, negative, neutral) and the community attributes (A, B, C) were used to find the word distribution rules. The word distribution after dimensionality reduction to two dimensions is shown in Fig. 8, which indicate that the majority of words are grouped together. Among the aforementioned dimensionality reduction methods, the T-SNE method yields better classification efficiency.

This section constructs the co-word network of pandemic data and extracts its community attributes. The primary concerns of Wuhan citizens can be divided into three types; the corpus is sampled by the

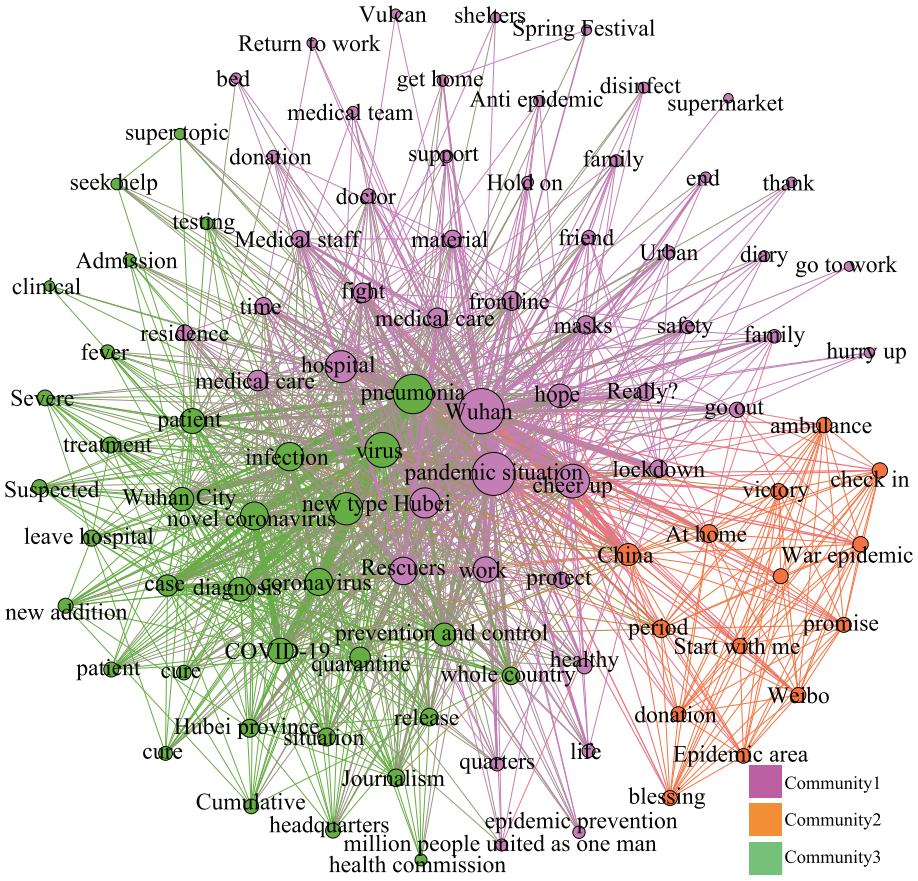


Fig. 7. Community attribute and interaction structure of co-word network.

Word2vec method; and the keywords extracted from TFIDF are reduced in dimension. From the dimensionality reduction results, we can see that the community group attributes of words are better classified than the sentiment attributes of words, which implies that it is possible to extract geographic attributes using co-word relations. In addition, certain edge words in the figure have more obvious classification features. In the next section, we extract these outliers and use text features to distinguish the geographic attributes of Weibos.

5.3. Geographic classification based on word features

We used the model proposed in this paper to draw the confusion matrix of the prediction results. The sum of each row of the confusion matrix represents the Weibo samples with a certain feature attribute. For example, 1339 Weibos in the residence class are correctly classified into the corresponding land-use type; however, 75 Weibos are mistakenly classified into the road and transport type.

Fig. 9 shows the classification results of all 4313 Weibo data, and the kappa coefficient of the matrix is 0.756. The prediction precision of the model for residence type is the lowest, followed by the prediction of road and transport. Thus, the prediction ability of the model for road and transport, residence, and business is poor, which indicates that the Weibos of users in these areas do not show similar text features. For example, Optics Valley Pedestrian Street (one of the most famous shopping districts in Wuhan) exhibits multiple feature attributes. Jianghan Road and Fanhu railway stations have the attributes of road and transport, business, and residence, which undoubtedly affected the classification effect of the model. Therefore, choosing a wider range of ground features as a classification standard can improve model accuracy (Gao et al., 2017). The prediction accuracy of the model proposed in this paper is high for medical, catering, railway station, and education types, which indicates that Weibos with these geographic features contains rich text feature attributes.

Based on a user's Weibo or other social platform content online, we could mine the accurate geographic location information hidden behind the text, which is convenient for pandemic prevention and control as well as scientific research. After statistics, only approximately 5% of the more than 200,000 Weibos crawled in this paper contain geographic attributes; however, using the classification model proposed in this paper, we can effectively label the remaining 95% of Weibos with geographic attributes.

5.4. Model comparison and accuracy evaluation

In the binary classifier, accuracy evaluation is relatively simple. In this paper, we primarily use a multivariate classifier and four evaluation indices: accuracy, precision, recall and f_1. After each class is calculated, the weighted average value is calculated to determine the model accuracy. We use true positive (TP) to represent the probability of correctly classifying class 1 to class 1 and true negative (TN) to represent the probability of predicting class 2 to be class 2, both of which are correct predictions. False positive (FP) describes the statistic of occasion, in which class 2 is predicted to be class 1, and False negative (FN) indicates that class 1 is predicted to be class 2. The proposed definition of the accuracy evaluation indicators is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

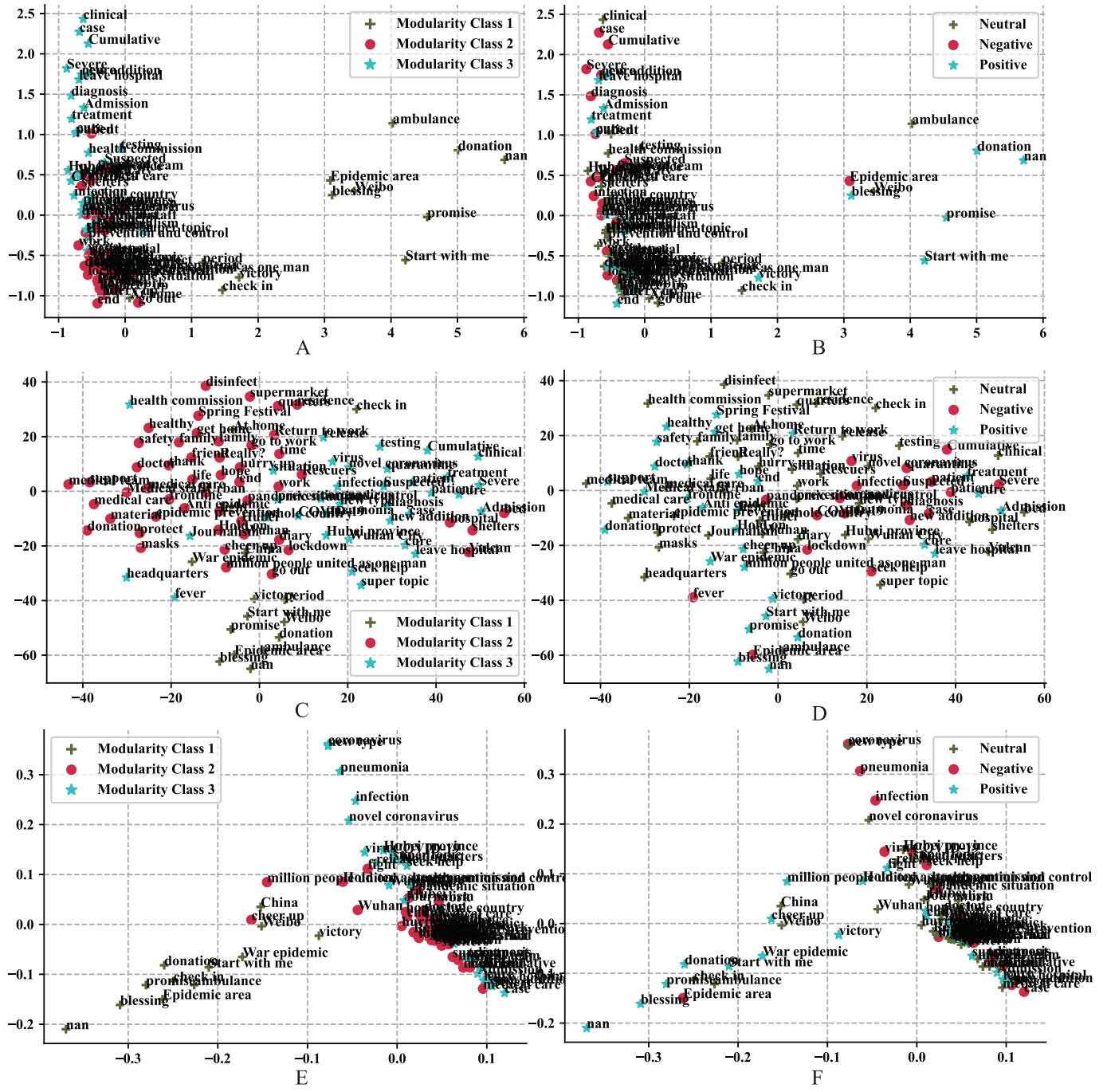


Fig. 8. The dimensionality reduction distribution of word vectors. (A) PCA method + community attributes; (B) PCA method + sentiment attributes; (C) T-SNE method + community attributes; (D) T-SNE method + sentiment attributes; (E) LLE method + community attributes; (F) LLE method + sentiment attributes.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$f_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

The experiment is performed with an NVIDIA GTX-1080, and the Keras deep learning framework based on TensorFlow of Graphics Processing Unit (GPU) version was used to extract Weibo text features of different geographic types. We chose traditional machine learning methods, such as K-Nearest Neighbor (KNN), Support Vector Method (SVM), naive Bayes classification model, and Classification and Regression Trees method (CART), to perform one hot coding of the test dataset and then took four indicators (f_1, *Recall*, *Precision*, *Accuracy*) as evaluation indices for model efficiency. The training results of traditional methods and the proposed method are shown in Fig. 10.

The model proposed in this paper performs best among the 5 models, yielding approximately 5–15% better performance than the traditional machine learning methods in terms of all four evaluation indices. The classification effect of SVM is the worst, and that of naive Bayes is better.

5.5. Applicability, advantages and disadvantages of the proposed model

Using the method of social sensing to perform situation awareness and knowledge fusion in geographical space can effectively extract changes in direction and the regional distribution of social public opinion, as shown in Figs. 4–7. Community-scale social sensing could be

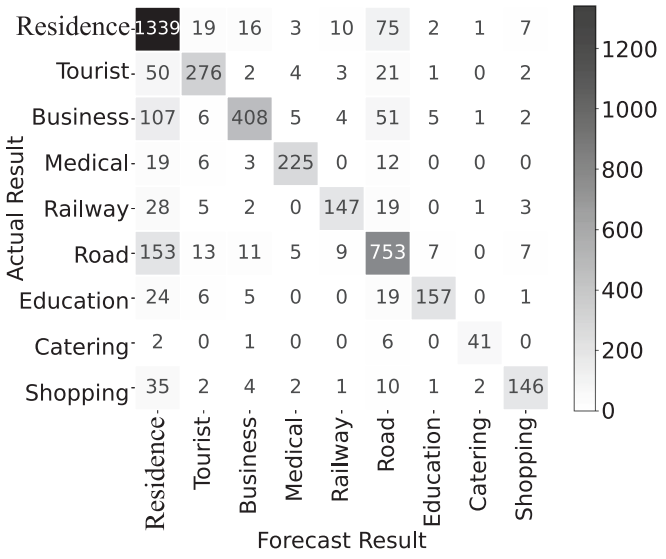


Fig. 9. Confusion matrix of KE-CNN model classification results.

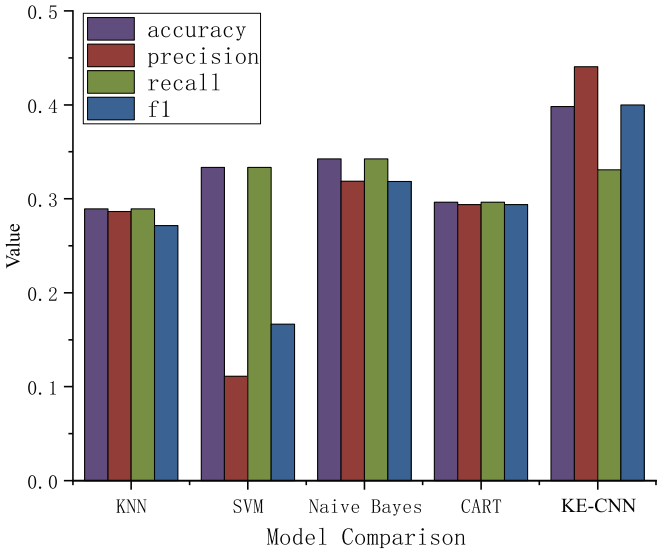


Fig. 10. Comparison results of different models.

combined with remote sensing big data, traffic big data, real-time heat of population and other methods to achieve multiscale seamless perception (Levin et al., 2020). Additionally, we propose a semantic geographic classification model based on the pandemic-related topics discussed in Weibos with geotags as the training corpus, which can achieve a good classification effect for certain feature types. The experiment shows that it is feasible to learn geographic information from semantic features. We generated a pretrained model by training a large-scale and full-scale corpus to achieve better robustness and mobility. When encountering a new social sensing environment, researchers could use transfer learning to speed up training and improve the performance of deep learning models. However, the proposed model still has certain limitations, including not considering the context of words during optimization and not solving the polysemy problem.

6. Conclusion and prospect

Under the various inconvenient circumstances caused by COVID-19, we could resort to geosocial sensing methods for non-contact, large-scale, and deep-level sensing of the social and economic operation. Based on the background of public health events in Wuhan, this paper studied the temporal and spatial distribution of online topic networks, hot spots of residents' attention, and the location information behind the online discussion. The experiment results show that (1) social sensing has the ability to discriminate the sentiment attitude of citizens in different locations and to extract the keywords of mainstream opinions (2) the prediction of ground feature type based on Weibo text is possible, and it shows good precision in certain types (medical type, catering type, railway station type, and education type) using our proposed KE-CNN method.

Social sensing can extract reliable information from a large amount of data that contain noise; however, the information extracted from social sensing can be biased (Wang et al., 2015). The whole narrative of this paper is developed under the framework of social sensing based on social media data. But in fact, social media does not complete the coverage of the whole age group, i.e., young people familiar with smart devices are more inclined to express their opinions on social platforms, so biases inevitably exist in our analysis. Additionally, strong denoising of social media data is another challenging research problem.

In June 2019, Twitter closed the tweet precise geographic location interface (Support, 2019), and this paper proposed a method to infer the geographic attributes behind tweets based on tweet information. The proposed method mitigates the lack of access to geographical location tags in online public opinion data by extracting text features, and completes missing fields for the remaining 95% of the online data. The classification method proposed in this paper has great potential to be studied in more detail (Stock, 2018) and will be used with the BERT model (Sánchez Villegas, Preoţiuc-Pietro, & Aletras, 2020) and coarse-grained surface feature types for future prediction research.

Funding statement

This research was supported by the National Key R&D Program (no. 2018YFB2100500), the Fundamental Research Funds for the Central Universities (no. 2042020kf0011), National Natural Science Foundation of China program (no.41890822) and Creative Research Groups of Natural Science Foundation of Hubei Province of China (no. 2016CFA003).

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.

Acknowledgements

The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Allen, C., Tsou, M. H., Aslam, A., Nagel, A., & Gawron, J. M. (2016). Applying gis and machine learning methods to twitter data for multiscale surveillance of influenza. *PLoS One, 11*, Article e0157734.

Arthur, R., Boulton, C. A., Shotton, H., & Williams, H. T. (2018). Social sensing of floods in the UK. *PLoS One, 13*, Article e0189327.

Avvenuti, M., Cresci, S., La Polla, M. N., Marchetti, A., & Tesconi, M. (2014). Earthquake emergency management by social sensing. In *2014 IEEE international conference on pervasive computing and communication workshops (PERCOM WORKSHOPS)* (pp. 587–592). IEEE.

Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics, 18*, 467–480.

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance, 6*, Article e19273.

Chen, W., Huang, H., Dong, J., Zhang, Y., Tian, Y., & Yang, Z. (2018). Social functional mapping of urban green space using remote sensing and social sensing data. *ISPRS Journal of Photogrammetry and Remote Sensing, 146*, 436–452.

Corley, C. D., Cook, D. J., Mikler, A. R., & Singh, K. P. (2010). Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health, 7*, 596–615.

Coulombe, C. (2018). *Text data augmentation made simple by leveraging nlp cloud apis*. arXiv preprint arXiv:1812.04718.

Cowie, S., Arthur, R., & Williams, H. T. (2018). @ choo: Tracking pollen and hayfever in the UK using social media. *Sensors, 18*, 4434.

Di Rocco, L., Dassereto, F., Bertolotto, M., Buscaldi, D., Catania, B., & Guerrini, G. (2020). Sherloc: A knowledge-driven algorithm for geolocating microblog messages at sub-city level. *International Journal of Geographical Information Science, 35*, 84–115.

Diaz, J., Poblete, B., & Bravo-Marquez, F. (2020). An integrated model for textual social media data with spatio-temporal dimensions. *Information Processing & Management*, 102219.

Eyre, R., De Luca, F., & Simini, F. (2020). Social media usage reveals recovery of small businesses after natural hazard events. *Nature Communications, 11*, 1–10.

Fan, C., Jiang, Y., & Mostafavi, A. (2020). Social sensing in disaster city digital twin: Integrated textual–visual–geo framework for situational awareness during built environment disruptions. *Journal of Management in Engineering, 36*, Article 04020002.

Gao, S., Janowicz, K., & Couclelis, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS, 21*, 446–467.

Gao, Y., Wang, S., Padmanabhan, A., Yin, J., & Cao, G. (2018). Mapping spatiotemporal patterns of events using social media: A case study of influenza trends. *International Journal of Geographical Information Science, 32*, 425–449.

Google. (2013). Word2vec-tool for computing continuous distributed representations of words [EB/OL] https://code.google.com/p/word2vec/ Accessed Jul 30, 2013.

Gu, S., Pan, C., Liu, H., Li, S., Hu, S., Su, L., , … Govindan, R., et al. (2014). Data extrapolation in social sensing for disaster response. In *2014 IEEE international conference on distributed computing in sensor systems* (pp. 119–126). IEEE.

Huang, Q., Cervone, G., & Zhang, G. (2017). A cloud-enabled automatic disaster analysis system of multi-sourced data streams: An example synthesizing social media, remote sensing and wikipedia data. *Computers, Environment and Urban Systems, 66*, 23–37.

Huang, Y., Li, Y., & Shan, J. (2018). Spatial-temporal event detection from geo-tagged tweets. *ISPRS International Journal of Geo-Information, 7*, 150.

Jahanbin, K., & Rahmanian, V. (2020). Using twitter and web news mining to predict covid-19 outbreak. *Asian Pacific Journal of Tropical Medicine, 13*.

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., & Campbell, A. T. (2010). A survey of mobile phone sensing. *IEEE Communications Magazine, 48*, 140–150.

Levin, N., Kyba, C. C., Zhang, Q., de Miguel, A. S., Román, M. O., Li, X., Portnov, B. A., Molthan, A. L., Jechow, A., Miller, S. D., et al. (2020). Remote sensing of night lights: A review and an outlook for the future. *Remote Sensing of Environment, 237*, 111443.

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2015). *Visualizing and understanding neural models in nlp*. arXiv preprint arXiv:1506.01066.

Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T., … Wang, F. (2020). Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. *IEEE Transactions on Computational Social Systems, 7*, 556–562.

Li, X., Wang, Z., Gao, C., & Shi, L. (2017). Reasoning human emotional responses from large-scale social and public media. *Applied Mathematics and Computation, 310*, 182–193.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., … Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers, 105*, 512–530.

Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via twitter: A case study of Chicago. *Applied Geography, 70*, 11–25.

Maaten, L.v.d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research, 9*, 2579–2605.

Meng, Y., Xing, H., Yuan, Y., Wong, M. S., & Fan, K. (2020). Sensing urban poverty: From the perspective of human perception-based greenery and open-space landscapes. *Computers, Environment and Urban Systems, 84*, 101544.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review, 45*, 167–256.

Persia, F., Pilato, G., Ge, M., Bolzoni, P., D'Auria, D., & Helmer, S. (2020). Improving orienteering-based tourist trip planning with social sensing. *Future Generation Computer Systems, 110*, 931–945.

Rashid, M. T., & Wang, D. (2021). Covidsens: A vision on reliable social sensing based risk alerting systems for covid-19 spread. *Artif Intell Rev, 54*, 1–25. https://doi.org/10.1007/s10462-020-09852-3.

Resch, B., Usländer, F., & Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science, 45*, 362–376.

Sánchez Villegas, D., Preoţiuc-Pietro, D., & Aletras, N. (2020). *Point-of-interest type inference from social media text*. arXiv arXiv:2009.14734.

Saranya, K., & Jayanthy, S. (2017). Onto-based sentiment classification using machine learning techniques. In *2017 international conference on innovations in information, embedded and communication systems (ICIIECS)* (pp. 1–5). IEEE.

Sengstock, C., & Gertz, M. (2012). Latent geographic feature extraction from social media. In *Proceedings of the 20th international conference on advances in geographic information systems* (pp. 149–158).

Severyn, A., & Moschitti, A. (2015). Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 373–382).

Steiger, E., Resch, B., & Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of twitter data using unsupervised neural networks. *International Journal of Geographical Information Science, 30*, 1694–1716.

Stock, K. (2018). Mining location from social media: A systematic review. *Computers, Environment and Urban Systems, 71*, 209–240.

Su, Y., Xue, J., Liu, X., Wu, P., Chen, J., Chen, C., … Zhu, T. (2020). Examining the impact of covid-19 lockdown in Wuhan and lombardy: A psycholinguistic analysis on weibo and twitter. *International Journal of Environmental Research and Public Health, 17*, 4552.

Support, T. (2019). Twitter removes support for precise geotagging because no one uses it [EB/OL] https://twitter.com/TwitterSupport/status/1141039841993355264/ Accessed Jun 19, 2019.

Wang, D., Abdelzaher, T., & Kaplan, L. (2015). *Social sensing: Building reliable systems on unreliable data*. Morgan Kaufmann.

Wang, W., & Stewart, K. (2015). Spatiotemporal and semantic information extraction from web news reports about natural hazards. *Computers, Environment and Urban Systems, 50*, 30–40.

Xing, H., Meng, Y., Hou, D., Cao, F., & Xu, H. (2017). Exploring point-of-interest data from social media for artificial surface validation with decision trees. *International Journal of Remote Sensing, 38*, 6945–6969.

Yang, X., Ye, T., Zhao, N., Chen, Q., Yue, W., Qi, J., … Jia, P. (2019). Population mapping with multisensor remote sensing images and point-of-interest data. *Remote Sensing, 11*, 574.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., & Mai, K. (2017). Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. *International Journal of Geographical Information Science, 31*, 825–848.

Yu, B., Lian, T., Huang, Y., Yao, S., Ye, X., Chen, Z., … Wu, J. (2019). Integration of nighttime light remote sensing images and taxi gps tracking data for population surface enhancement. *International Journal of Geographical Information Science, 33*, 687–706.

Zhang, F., Wu, L., Zhu, D., & Liu, Y. (2019). Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. *ISPRS Journal of Photogrammetry and Remote Sensing, 153*, 48–58.

Zhang, X., Sun, Y., Zheng, A., & Wang, Y. (2020). A new approach to refining land use types: Predicting point-of-interest categories using weibo check-in data. *ISPRS International Journal of Geo-Information, 9*, 124.

Zhang, Y., Li, Q., Huang, H., Wu, W., Du, X., & Wang, H. (2017). The combined use of remote sensing and social sensing data in fine-grained urban land use mapping: A case study in Beijing, China. *Remote Sensing, 9*, 865.

Zhang, Y., Li, Y., Yang, B., Zheng, X., & Chen, M. (2020). Risk assessment of covid-19 based on multisource data from a geographical viewpoint. *IEEE Access, 8*, 125702–125713.

Zhao, N., Cao, G., Zhang, W., Samson, E. L., & Chen, Y. (2020). Remote sensing and social sensing for socioeconomic systems: A comparison study between nighttime lights and location-based social media at the 500 m spatial resolution. *International Journal of Applied Earth Observation and Geoinformation, 87*, 102058.

Zheng, L., Wang, F., Zheng, X., & Liu, B. (2019). Discovering the relationship of disasters from big scholar and social media news datasets. *International Journal of Digital Earth, 12*, 1341–1363.

Zhu, R., Lin, D., Jendryke, M., Zuo, C., Ding, L., & Meng, L. (2019). Geo-tagged social media data-based analytical approach for perceiving impacts of social events. *ISPRS International Journal of Geo-Information, 8*, 15.

Zou, L., Lam, N. S., Cai, H., & Qiang, Y. (2018). Mining twitter data for improved understanding of disaster resilience. *Annals of the American Association of Geographers, 108*, 1422–1441.