International Journal of Geo-Information

MDPI

Article

A Semantically Enhanced Label Prediction Method for Imbalanced POI Data Category Distribution

Hongwei Zhang [1], Qingyun Du [2,3,*], Shuai Zhang [4] and Renfei Yang [2,5]

[1] College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China; zhanghongwei@cqut.edu.cn
[2] School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China; youngfly@whu.edu.cn
[3] Key Laboratory of Digital Cartography and Land Information Application, Ministry of Natural Resources, Wuhan University, Wuhan 430079, China
[4] School of Smart City, Chongqing Jiaotong University, Chongqing 400074, China; zhangshuai@cqjtu.edu.cn
[5] Chongqing Geomatics and Remote Sensing Center, Chongqing 401120, China
* Correspondence: qydu@whu.edu.cn

Abstract: POI data play an important role in various location-based services, including navigation, positioning, and local search applications. However, as cities rapidly develop, a substantial amount of new POI data are generated daily, often accompanied by issues with the quality of their labels. Therefore, there is an urgent need to implement intelligent inference and enhancement processing for POI data labels. Conventional neural network models primarily target balanced data distribution, but they fail to address the issue of imbalanced distribution of POI data labels in terms of quantity. Furthermore, most neural network classification models implicitly learn the semantic knowledge of different categories from training datasets, neglecting the explicit semantic information offered by natural language labels. Considering the above problems, several negative samples are introduced for each input to a positive class, thereby transforming the multi-classification task into a binary classification problem. Simultaneously, POI data labels are introduced to provide explicit semantic information, and the semantic relationship between POI data labels and their names is determined using cross-coding. Experiments demonstrate that the $macro - F1$ score for the test dataset, which contains 75 different categories of POI data, reaches 0.84. This result surpasses the performance of traditional methods, highlighting the effectiveness of the proposed method.

Keywords: point of interest (POI); POI classification; imbalanced dataset; semantic enhancement

check for updates

Citation: Zhang, H.; Du, Q.; Zhang, S.; Yang, R. A Semantically Enhanced Label Prediction Method for Imbalanced POI Data Category Distribution. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 364. https://doi.org/10.3390/ijgi13100364

Academic Editors: Wolfgang Kainz and Hartwig H. Hochmair

Received: 6 September 2024
Revised: 13 October 2024
Accepted: 15 October 2024
Published: 17 October 2024

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

Copyright: © 2024 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The term point of interest (POI), widely used in geographic information systems, refers to geographical entities that can be abstracted as points, especially those that are closely connected to everyday life, including but not limited to restaurants, supermarkets, and educational institutions [1,2]. POI data exhibit a wide range of application scenarios and possess robust computational and expressive properties. With the rapid development of internet technology and the widespread use of mobile devices, location-based services based on POI data have gained increasing significance. For example, individuals frequently rely on POI data from apps like Meituan and Ele.me to locate nearby points of interest, such as restaurants, hotels, and tourist attractions [3,4]. However, due to rapid urbanization, a substantial amount of new POI data emerge daily. According to data released by Didi at the end of 2016, approximately 80,000 recommended pick-up points were updated daily using extensive travel data. These recommended pick-up points resulted in a 10% reduction in driver calls and an average decrease of one minute in passenger waiting times. Therefore, meeting the demand for personalized POI recommendation services, addressing the issue

of information overload faced by users, and achieving the automatic classification of POI data have become the focus and challenge of POI data analysis and mining research [5,6].

Identifying POI data labels is a complex challenge, compounded by several factors. Firstly, the volume of POI data is enormous, with a wide range of categories, covering various aspects of life. Baidu Map and Amap offer hundreds of categories of POI data, with more detailed classifications reaching thousands of categories. Secondly, POI data naming is highly inconsistent, lacking a standard convention and often including a mix of Chinese, English, letters, and numbers. Short names may consist of just a few words, or even one word, while longer names may have dozens of words. Thirdly, the presence of complex nesting relationships, abbreviations, and other challenges poses significant difficulties in the analysis and identification of POI data, particularly in government agencies and corporate entities, where such phenomena are commonplace. Furthermore, the same POI name can belong to multiple different semantic categories. For example, "Mingming snacks" can be classified as either "Fast food", "Specialty snacks", or a "Chinese restaurant".

There are two main methods for POI data classification: those based on rich feature information and those based on finite element information. For the former method, researchers utilize geosocial check-in data [7–9], internet engine data [10], and user comment data to classify POI data [11]. Although these studies can extract richer features through external data sources, it can be challenging to match external data sources with the continuously emerging new POI data in practice. In the latter method, researchers frequently rely on POI names, geolocation details, and predefined categories for POI classification [12,13]. This method can effectively alleviate the problem of missing data sources. However, both methods have limitations in addressing the issue of imbalanced distribution of POI data labels. There is significant variation in the quantity of POI data across different categories within the same dataset. For instance, the number of restaurants significantly exceeds the number of cinemas.

To address the issue of imbalanced datasets in classification tasks, extensive research has been conducted. At the data level, two primary techniques for addressing data imbalance have emerged: oversampling [14,15] and undersampling [16,17]. These techniques either employ clustering methods or pursue further undersampling and oversampling following clustering. At the algorithm level, cost-sensitive learning [18,19] and ensemble learning [20–22] have garnered widespread attention. Cost-sensitive learning is employed to handle imbalanced datasets by assigning higher costs to the misclassification of minority categories. Ensemble learning, on the other hand, relies on a collection of sub-classifiers, integrating their outcomes to arrive at the final classification result through defined rules. Furthermore, the challenge of classifying imbalanced datasets can also be addressed through the utilization of one-vs-one [23] and one-vs-rest [24] transformation strategies, which facilitate the conversion of multi-classification tasks into binary classification problems. However, when it comes to POI datasets, the current methods for handling imbalanced datasets often fall short of effectively enhancing classification accuracy.

With the continuous advancement of technology, deep neural networks have achieved remarkable breakthroughs in the field of natural language processing [25–27]. Notably, in 2019, Devlin et al. introduced bidirectional encoder representations from transformers (BERT) [28], which set new benchmarks for various natural language processing tasks. The success of BERT is attributed to its attention mechanism, which can capture the semantic relationships between sentence pairs through cross-coding (also known as cross-attention). Nevertheless, traditional neural network models typically process data uniformly, focusing on learning the overall data distribution. As a result, the resulting feature space is often dominated by categories with larger sample sizes, neglecting the challenge of uneven category distribution. Moreover, most deep neural networks implicitly learn semantic features between different categories from training datasets, overlooking the explicit semantic information provided by natural language labels. The explicit information conveyed by these labels is particularly beneficial for the semantic recognition of POI data in this study. This is because POI data labels often consist of keywords found within their names,

providing valuable explicit semantic information. For example, both "Hangzhou Sanhe Advertising Decoration Co., Ltd." and "Pingtian Advertising Decoration Co., Ltd." belong to the category of advertising decoration companies. It is evident that the label is a key term within their names, carrying explicit semantic information. Therefore, utilizing this explicit information can significantly enhance the accuracy of POI data classification.

To address the aforementioned issues, POI data labels are introduced to provide explicit semantic information. Additionally, a transformation strategy is employed to convert the multi-classification task into a binary classification problem, thereby enhancing the accuracy of POI data category recognition.

In summary, the key contributions of this study are as follows:

1. A transformation strategy is employed to transform the multi-classification task into a binary classification problem. This method introduces several negative samples for each positive class, enhancing the model's generalization ability and allowing it to effectively handle diverse data and scenarios.
2. POI data labels are introduced to provide explicit semantic information, and a cross-coding mechanism is employed to assess the relationship between POI data labels and their names, which improves the accuracy of POI data category recognition.

2. Related Work

In this section, we present a brief overview of two categories of related work: POI classification and handling imbalanced datasets.

2.1. POI Classification

Due to the complexity of POI semantics, a variety of data sources are used for POI classification tasks. McKenzie et al. proposed a method to label unclassified geosocial check-in data based on the similarity of users' check-in time, name, and category [7]. Zhou et al. utilized internet search engines as external knowledge bases to introduce real-time and large-scale text feature information into the original POI text for predicting POI categories [10]. Giannopoulos and Meimaris combined domain knowledge and semantic information extracted from external data sources to classify POI data [29]. Wu et al. employed bidirectional spatiotemporal dependencies and users' dynamic preferences to identify missing POI data labels [30]. Noorian et al. utilized visual labels of street-level images as supplementary information for POI data classification [31]. However, these methods are limited by the absence of accessible external data sources in the face of the new POI data that emerge every day.

Currently, few studies have focused exclusively on utilizing POI data to address these challenging tasks. Li et al. utilized Word2vec to vectorize POI names and employed support vector machines for automatic POI data classification [12]. Duan proposed a random vocabulary iterative model that incorporates significant feature word filtering and local textual semantic relevance segmentation for POI data classification [13]. Choi et al. employed Gaussian weights based on the onomastic and contextual information of POI names, assigning higher weights to words closer to the POI name and lower weights to those farther away [32]. Although these methods adeptly mitigate the scarcity of external data sources, they often fall short of delivering the desired level of classification accuracy.

2.2. Imbalanced Dataset Handling

To address the issue of imbalanced datasets, numerous studies have been conducted by researchers. At the data level, researchers commonly utilize oversampling and under-sampling techniques to address the issue of imbalanced datasets. Ijaz et al. employed the density-based noise outlier detection of spatial clustering to eliminate the outlier data and utilized the synthetic minority over-sampling technique (SMOTE) to oversample the data by inserting synthetic samples for underrepresented categories [33]. Agusta reduced the number of majority category samples by clustering them in the kernel space and combined the AdaBoost method with an undersampling algorithm [34]. Cao et al. utilized an

oversampling method to generate new samples for the minority category and employed a clustering technique to extract samples from sub-clusters of the majority category [35]. However, these techniques can result in a distorted representation of the sample space. To address this issue, Khan et al. proposed a cost-sensitive deep neural network that automatically learns robust feature representations for both majority and minority categories [36]. Du et al. based on the concepts of voting and coordination, divided the training dataset into multiple sub-training datasets, constructed multiple classifiers, and labeled unlabeled data through the voting results of these classifiers [37]. Additionally, employing transformation strategies that convert multi-classification tasks into binary classification problems can effectively address the issue of uneven category distribution within datasets. Smith et al. utilized this transformation strategy to classify cow behavior [38].

Currently, few studies have examined the classification of POI data with imbalanced label distribution. Yu et al. proposed a multi-level label prediction model to address this issue in POI data categories [39]. However, this method only considers 10 categories of POI data and relies on the coordinates of POI data to capture adjacent spatial distribution, which has certain limitations in practical applications. This is because it can be challenging to ensure that new POI data contain coordinate information.

3. POI Classification Model

In this section, we introduce a method to address the issue of imbalanced distributions across POI data categories. We then incorporate POI data labels to enhance the semantic information. Finally, we propose a semantically enhanced label prediction model to tackle imbalanced POI category distribution and establish the evaluation criteria for this model.

3.1. Transforming the Multi-Classification Task into a Binary Classification Problem

Considering a training dataset comprising M categories, the one-vs-one transformation strategy involves the creation of pairwise combinations between distinct categories, resulting in $M(M-1)/2$ individual classifiers. In the testing phase, the test data are simultaneously submitted to all classifiers, and the final classification result is determined through voting or based on the prediction confidence level of each classifier. The one-vs-rest transformation strategy designates one category as the positive class, with the remaining categories aggregated as the negative class, resulting in M classifiers. During testing, if only one classifier predicts a positive class, the corresponding label is taken as the final classification result. When multiple classifiers predict a positive class, the prediction confidence level of each classifier is typically considered, and the label with the highest confidence is selected as the classification output. Importantly, the one-vs-rest strategy employs all categories for each classifier, while the one-vs-one strategy restricts each classifier to just two categories. The one-vs-rest transformation often results in a longer training time when dealing with multiple categories and can become less meaningful when the number of categories becomes excessively large.

Given the extensive range of POI data categories in this study, the two transformation strategies become impractical. To tackle this challenge, we adopt the concept of the one-to-rest transformation strategy. Specifically, one category is designated as the positive class at a time, while a random subset of the remaining categories is sampled as the negative class. However, this method introduces the challenge of multiple classifiers. For this problem, a learning function is defined that consolidates the predictions from multiple classifiers into a single classification decision based on confidence scores. The formula for calculating the relationship between any POI name t and label l_i is as follows:

$$f(l_i, t) = P(True|l_i, t),\ i \in [1, \cdots, M]$$

where i represents the number of labels, and if the text t belongs to the label l_i, the value of f is true, otherwise it is false. In other words, the function accepts a text label and a POI name as input and evaluates the relationship to determine the classification result. For example, the format for input pairs in the POI data classification task is as follows:

Based on the preceding discussion, it is evident that any classification task can be decomposed into the problem's definition, particularly the transformation of a multi-classification task into a series of binary problems—a concept well-documented in the foundational literature of machine learning [40,41]. However, there is a crucial distinction: the introduction of a function approximation method that enables treating the multi-class classification task as a single classifier, ultimately generating a predictive model.

3.2. Semantic Enhancement

Currently, the majority of classification methods implicitly learn the semantic knowledge of different categories from the training dataset, yet they often neglect the explicit semantic information embedded within the natural language labels of these categories. As can be seen from Figure 1, by introducing labels, the category of POI names can be easily determined through terms like "hot pot" or "hotel" within their respective names. POI data labels often act as keywords within their names, and incorporating this explicit semantic information provided by the labels is highly beneficial for POI category recognition. Therefore, this study employs POI data labels to enhance semantic understanding, offering explicit information for the semantic categorization of POIs.

Label	POI name
< hot pot restaurant,	Beijing charcoal hot pot >
< hotel,	Hanting Hotel >

Figure 1. The POI data classification task with a binary tuple input format.

The model structure proposed is shown in Figure 2. The model accepts a pair consisting of a label and a POI name as input, separated by the token "[SEP]". Based on the attention mechanism of the robustly optimized BERT pretraining approach (RoBERTa), the model determines the semantic relationship between the POI data label and the corresponding POI name. It transforms the binary classification problem into a task of evaluating the relationship between two sentences. If the prediction is true, it indicates that the POI name belongs to that category. Otherwise, it does not. For the sequence input to RoBERTa, the output is the vector of each token in the sequence. However, a challenging problem lies in selecting the optimal combination of vectors from different layers to represent the relation between two sentences. Given that BERT frequently employs the "[CLS]" token as a sentence vector to evaluate the relationship between two sentences, and acknowledging that RoBERTa represents an advancement over BERT, this study adopts the "[CLS]" vector from the final layer to represent the relationship between the label and the POI name. Subsequently, the resultant vector is subjected to linear layer mapping, followed by the application of the Softmax function for normalization. Ultimately, the model produces a binary prediction outcome—true or false—based on the calculated probability.

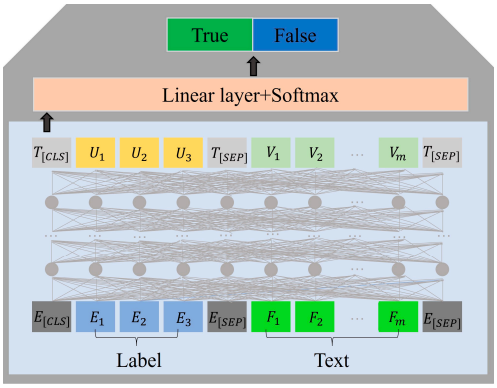


Figure 2. The structure of the semantic recognition model.

3.3. Training and Prediction

This study employs the pretrained Chinese model RoBERTa, which features 12 layers, with each layer containing 12 self-attention heads and 768 hidden nodes. For the POI data classification task encompassing M categories, it is necessary to generate M pairs of <label, POI name> for each input instance. For example, if the label of the i-th input corresponds to the category of the POI name, this results in a positive training example: (<$label_i$, POI name>, True). Conversely, for $(M-1)$ negative instances, the labels do not correspond to the POI name's category, represented as (<$label_k$, POI name>, False), with the stipulation that $k \neq i$. Figure 3 illustrates the scenario with two negative samples.

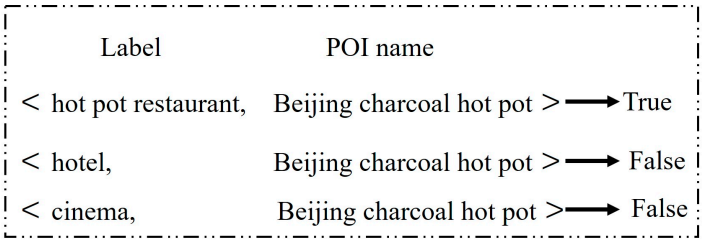


Figure 3. Input Example Demonstration.

By employing the one-to-rest transformation strategy, the training dataset experiences a significant expansion, thereby effectively enhancing the model's generalization ability. This enhancement has been comprehensively validated and confirmed through numerous studies [42,43]. However, considering the complexity of POI data semantics, populating the remaining categories as negative samples is not an optimal method. To reduce the quantity of the training dataset without compromising the model's generalization, a random subset of the remaining categories is selected to serve as negative classes. For label prediction with this model, the resultant output is the scores of all categories. Accordingly, the label with the greatest confidence score is designated as the final prediction.

3.4. Model Evaluation

For natural language classification tasks, three common metrics are precision, recall, and F_1 score. However, the applicability of these metrics is typically limited to binary classification tasks. In contrast, the POI data category recognition in this study is a multi-classification task. Therefore, we employ macro-precision ($macro-P$), macro-recall ($macro-R$), and $macro-F_1$ score to evaluate the model's overall performance. The calculation of these macro metrics is based on the individual precision, recall, and F_1 score for each category. For a given label tag_i in the multi-classification task, the corresponding precision P_{tag_i}, recall R_{tag_i}, and F_{1tag_i} score are calculated as follows:

$$P_{tag_i} = \frac{TP_{tag_i}}{TP_{tag_i} + FP_{tag_i}} \tag{1}$$

$$R_{tag_i} = \frac{TP_{tag_i}}{TP_{tag_i} + FN_{tag_i}} \tag{2}$$

$$F_{1tag_i} = \frac{2 * P_{tag_i} * R_{tag_i}}{P_{tag_i} + R_{tag_i}} \tag{3}$$

where TP_{tag_i} represents the number of labels correctly categorized as tag_i, FP_{tag_i} represents the misclassification of labels from other categories as tag_i, and FN_{tag_i} denotes the number of labels tag_i incorrectly identified as labels from other categories. After calculating the

precision, recall, and $F1$ scores for each category, $macro - P$, $macro - R$, and $macro - F_1$ score can be computed as follows:

$$macro - P = \frac{1}{M}\sum_{i=1}^{M} P_{tag_i} \tag{4}$$

$$macro - R = \frac{1}{M}\sum_{i=1}^{M} R_{tag_i} \tag{5}$$

$$macro - F_1 = \frac{2 * macro - P * macro - R}{macro - P + macro - R} \tag{6}$$

4. Experiments and Results Analysis

In this section, we first introduce the dataset utilized for the POI data classification task. Then, to verify the effectiveness of our proposed method, we conduct several comparative experiments.

4.1. Dataset

In this study, the POI data are classified according to the categorization standard of Baidu Maps. Due to the complexity of POI semantic recognition, we do not apply a fine-grained semantic classification scheme. The selection of POI categories in this study is based on the following two principles: (1) POI names belonging to multiple categories are not sub-categorized further; only the first-level classification to which they belong is recognized. (2) For the phenomenon of multi-semantic intersections, it is treated on a case-by-case basis. As an example, we consider the "Food" category within Baidu Maps to illustrate. For categories like "Chinese restaurant" and "Fast food restaurant" within the secondary classification of food, their names lack distinctive keywords, so they are uniformly categorized into the first-level classification of food. Categories that contain distinctive keywords, such as "coffee shop" and "cake dessert", are further sub-categorized. When encountering multi-semantic intersections during classification, we utilize a manual methodology to assign each instance to the most appropriate category based on its distinct semantic meaning. For instance, "Da Fu Farmhouse" is assigned to the "Leisure and entertainment" category based on its intended meaning. In summary, after processing, the "Food" category is divided into two main levels: the primary classification of food and the secondary classifications, including "Tea house", "Cake shop", "Coffee shop", "Hot pot restaurant", and "Cold drink store". After analyzing the collected POI data in Hangzhou, a total of 75 different categories of POI data are obtained, and their specific classifications are presented in Table 1. Other categories that do not appear in the second-level classification are grouped into the first-level classification, excluding transportation facilities, which are not included in the first-level classification.

Table 1. POI data categories.

First-Level Classification	Second-Level Classification
Food	Tea house, Cake shop, Coffee shop, Cold drink shop, Hot pot restaurant
Hotel	
Shopping	Supermarket, Clothing and leather goods store, Flower market, Home appliance and electronics store, Building materials and hardware market, Furniture store, Kitchen and bathroom store, Lamps and porcelain market, Sporting goods store, Stationery store, Pet supplies store, Children's products store, Bookstore, Tobacco and alcohol store, Optical shop, Fruit market, Bicycle and electric bike store

Table 1. *Cont.*

First-Level Classification	Second-Level Classification
Life service	Photography and printing shop, Ticket office, Repair station, Logistics, Laundry shop, Bath and massage place, Travel agency
Fair lady	
Tourist attraction	
Leisure and entertainment	Cinema, Bar, Chess and card room, Internet cafe, KTV
Exercise and fitness	Basketball venue, Table tennis venue, Billiard hall, Taekwondo venue, Tennis venue, Swimming venue, Badminton venue
Education and training	
Culture and media	
Medical care	Convalescent home, Pet clinic, Drugstore
Automotive service	Gas station, Car charging station
Transportation facility	Bus stop, Parking lot, Bus station
Finance	
Company	Advertising and decoration company, Mechanical and electronic company, Construction company, Commercial trading company, Network technology company, Pharmaceutical company, Agribusiness company, Metallurgical and chemical company, Automotive company, Financial company
Government agency	
Entrance and exit	

In this study, we collected 1,111,554 pieces of POI data from Hangzhou City. Due to the messiness of the POI data, a cleaning process was undertaken that encompassed the following steps: (1) removing the spaces and special symbols (e.g., "/", "—") contained in the POI names; (2) deleting duplicate data by simply removing entries with identical names; and (3) conducting POI data category processing, which included the manual refinement of categories, and filtering based on the POI classification of Baidu Maps and the keywords contained in the names. Following the processing of the POI data, a total of 139,471 unique POI names were selected. The number of different categories is illustrated in Figure 4. As depicted in Figure 4, there is considerable variation in the quantity of POI names among different categories, with some categories encompassing 10,000 POI names while others have less than 200. Categories that have a higher volume of POI names are assigned to the first-level classification, with the objective of improving the accuracy of recognition at this level. Ultimately, the selected data were partitioned into three distinct segments: a training dataset, a test dataset, and a validation dataset, allocated in the proportions of 60%, 20%, and 20%, respectively.

4.2. To Evaluate the Effectiveness of This Study, Three Methods Were Employed for Comparison

RoBERTa–Recurrent Neural Network (RNN): The method first converts POI names into vector representations using RoBERTa. Subsequently, these vectors are input into an RNN to extract their contextual information. Finally, the vectors output from the RNN are passed through a linear layer for the label prediction. This multi-classification method aims to verify the effectiveness of transforming the multi-classification task into a binary classification problem.

BERT (L = 12, H = 768, A = 12): This model is a Chinese model trained by Google, where L represents the number of layers, H represents the number of hidden nodes, and A

represents the number of self-attention heads in each layer. Label prediction of POI data is performed using the method described above.

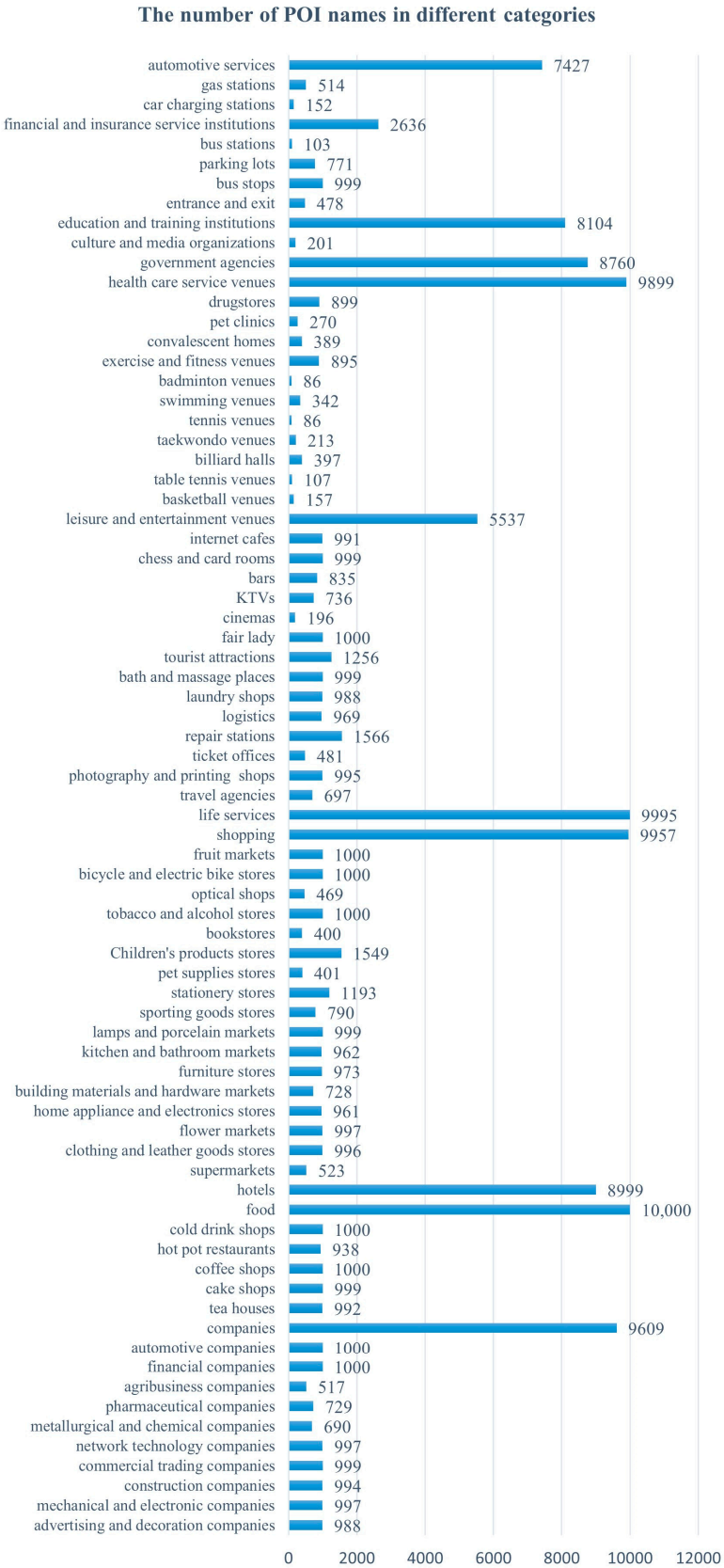


Figure 4. The number of POI names in different categories.

A Lite BERT (ALBERT): BERT, with its extensive number of parameters, demands significant computational resources for efficient execution. To address this challenge, Lan et al. proposed ALBERT [44], an optimized variant of BERT. ALBERT achieves a 40% reduction in model size while maintaining 97% of BERT's language comprehension accuracy. To ensure comparability among different models, this study utilized the Chinese model ALBERT (L = 24, H = 1024, A = 24).

4.3. Experimental Settings

As previously stated, this study has defined 75 distinct POI categories. During the training phase, it is generally recommended to generate 74 negative samples for each positive class. However, due to the intricate nested relationships inherent in POI names, utilizing negative samples to populate the remaining classes is not an optimal method. Therefore, for each positive class of input, the quantities of negative samples are set to 5, 10, 15, and 20 for each positive class. To determine a suitable learning rate, this study employs the method of cyclic learning rates [45], which involves plotting the loss function and learning rate curves to determine the optimal learning rate. Moreover, this method also necessitates the specification of the batch size parameter. Reimers et al. have suggested that, for larger datasets, a batch size ranging from 8 to 32 is advisable [46]. Accordingly, the learning rate range is set to $[10^{-7}, 1]$, and the batch sizes are set to 8, 16, and 32 for further discussion.

As depicted in Figure 5, the relationship between the learning rate and the loss function value across varying parameters is illustrated. The optimal learning rate is determined to lie between the steepest descent of the loss function value and its minimum point. It is evident from the figure that a batch size of 32 and a learning rate of 0.001 are most effective when utilizing five negative samples. With 10 negative samples, the optimal settings are a batch size of 16 and a learning rate of 0.001. When applying 15 negative samples, the optimal batch size is 16 and the learning rate is 0.01. Lastly, for 20 negative samples, the optimal batch size is 8 and the learning rate is 0.001.

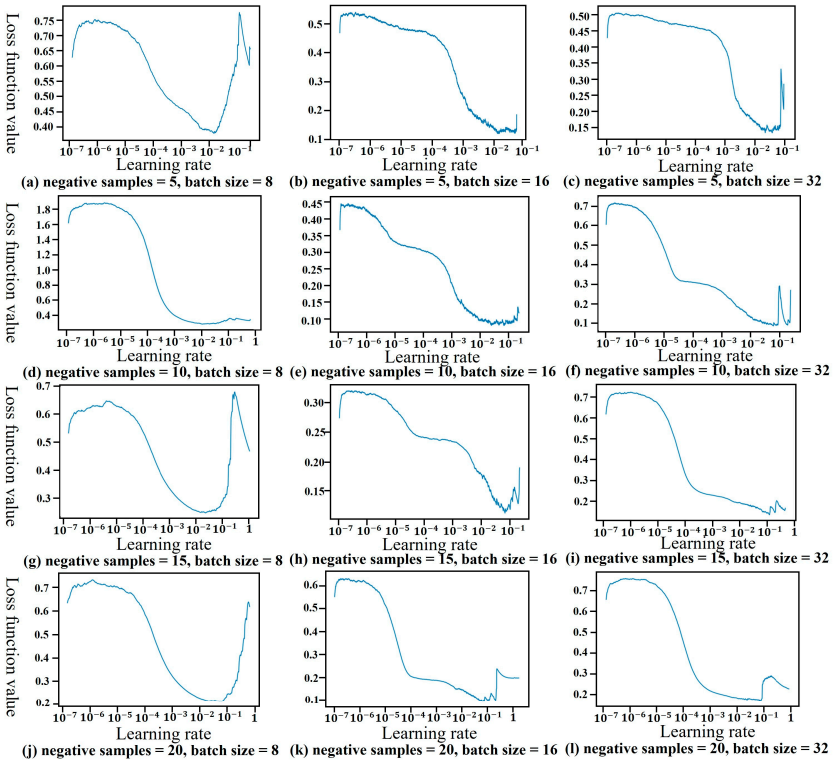


Figure 5. The relationship between loss function value and learning rate under different numbers of negative samples and batch sizes.

Once the optimal hyperparameters for the model under various conditions of negative samples were established, we initiated the training process. As shown in Figure 6, the $macro - F_1$ scores were obtained from the validation dataset across varying numbers of negative samples. For different numbers of negative samples, different epochs were set. When the number of negative samples is 5, 10, 15, and 20, the corresponding maximum scores of $macro - F_1$ on the validation dataset are 0.8317, 0.8334, 0.84, and 0.8389, respectively, achieved at the 49th, 45th, 22nd, and 29th epochs. Therefore, it is recommended to set the number of negative samples to 15.

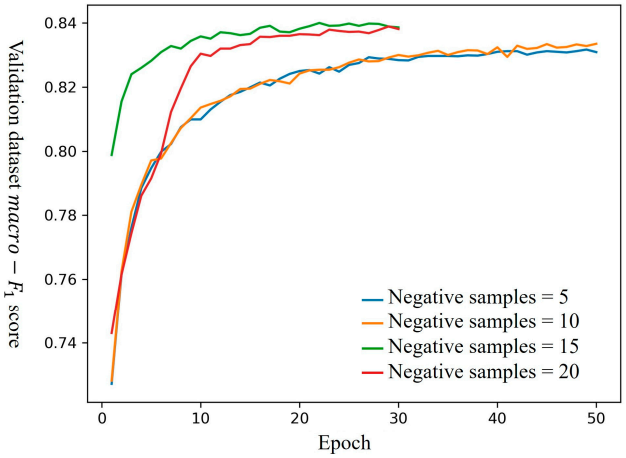


Figure 6. *Macro* $- F_1$ scores on the validation dataset with different numbers of negative samples.

In the same way, the optimal parameter combinations for the comparative methods are shown in Table 2.

Table 2. The optimal parameter combinations of the comparative methods.

Method	Number of Negative Samples	Batch Size	Learning Rate	Number of RNN Layers	Hidden Size
RoBERTa-RNN		32	0.001	1	100
BERT	10	32	0.001		
	15	16	0.001		
	20	8	0.001		
ALBERT	5	32	0.001		
	10	8	0.001		
	15	8	0.001		

Following the training of BERT and ALBERT models using the optimal parameter settings across varying quantities of negative samples, it was observed that BERT achieved the most promising results with 15 negative samples, whereas ALBERT attained optimal performance with 10 negative samples.

4.4. Results Analysis

As can be seen in Figure 7, the validation dataset's $macro - F_1$ scores are presented across different methodologies. The figure clearly illustrates that distinct methods exhibit varied convergence behaviors, with each requiring a different number of epochs to achieve convergence. When considering RoBERTa, BERT, and ALBERT—all of which convert multi-classification tasks into binary classification problems—the optimal $macro - F_1$ scores on the

validation dataset are 0.84, 0.8306, and 0.7858, respectively. Significantly, the corresponding $macro - F_1$ scores for the test dataset are slightly higher, at 0.84, 0.834, and 0.788, respectively. This observation highlights the models' impressive generalization capabilities. The initial two models exhibit markedly higher $macro - F_1$ scores on both the validation and test datasets in comparison to ALBERT, reflecting the relatively weaker capacity of ALBERT in capturing semantic information between sentence pairs. The method proposed in this study outperforms the other two methods in $macro - F_1$ scores in both the validation dataset and the test dataset, thus validating the effectiveness of the method proposed in this study.

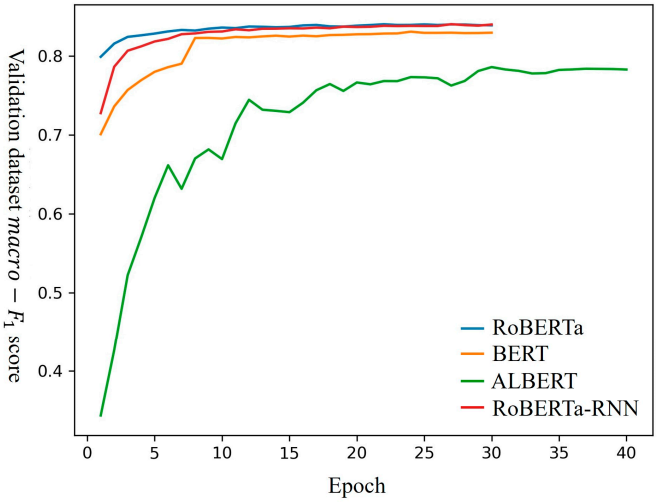


Figure 7. $Macro - F_1$ scores on the validation dataset across different methods.

The RoBERTa-RNN model achieves an optimal $macro - F_1$ score of 0.8401 on the validation dataset, whereas the corresponding score on the test dataset is lower at 0.827. Table 3 lists the 10 POI categories with the poorest performance on the test dataset for both the proposed method and the RoBERTa-RNN model in this paper. It is evident from the table that for POI data with a smaller quantity, the F_1 score of the RoBERTa-RNN model is either zero or exceptionally low, indicating that the RoBERTa-RNN model has hardly captured the characteristics of these POI data. In contrast, the method proposed in this study has obtained significantly higher F_1 scores for POI data with limited quantities, and a $macro - F_1$ score of 0.84 for the test dataset. These results demonstrate that the proposed semantic enhancement method can effectively capture explicit semantic information provided by natural language labels. Simultaneously, converting multi-classification tasks into binary classification problems can effectively address the challenge of data category imbalance.

Table 3. The top 10 POI categories with the poorest performance across different methods on the test dataset.

POI Category	RoBERTa			RoBERTa-RNN			Size of Training Dataset
	Precision	Recall	F_1 Score	Precision	Recall	F_1 Score	
Cold drink shop	0.6345	0.625	0.6297	0.6836	0.605	0.6419	1000
Building materials and hardware market	0.5562	0.6828	0.613	0.547	0.6828	0.6074	728
Shopping	0.5947	0.6389	0.616	0.583	0.6494	0.6144	9957
Sporting goods store	0.7121	0.5949	0.6483	0.697	0.5823	0.6345	790
Tea house	0.6872	0.6768	0.6819	0.743	0.6717	0.7056	992
Convalescent home	0.746	0.6104	0.6714	0.6901	0.6364	0.6622	389
KTV	0.6606	0.7415	0.6987	0.6708	0.7347	0.7013	736
Agribusiness company	0.6495	0.6117	0.63	0.6262	0.6505	0.6381	517
Clothing and leather goods store	0.3492	0.1106	0.1679	0.3652	0.2111	0.2675	996
Cake shop	0.7636	0.6332	0.6923	0.7314	0.6432	0.6845	999
Entrance and exit	1	0.9684	0.984	0	0	0	478

Table 3. *Cont.*

POI Category	RoBERTa			RoBERTa-RNN			Size of Training Dataset
	Precision	Recall	F_1 Score	Precision	Recall	F_1 Score	
Bus station	1	1	1	0	0	0	103
Basketball venue	0.9615	0.8065	0.8772	0	0	0	157
Table tennis venue	1	1	1	0	0	0	107
Tennis venue	0.8947	1	0.9444	0	0	0	86
Badminton venue	0.875	0.8235	0.8485	0	0	0	86
Car charging station	0.9630	0.8667	0.9123	0	0	0	152
Culture and media	0.9231	0.9000	0.9114	0	0	0	201
Pet clinic	0.9375	0.8333	0.8823	1	0.0185	0.0363	270
Cinema	0.8919	0.8462	0.8684	1	0.0769	0.1428	196

Moreover, as shown in Table 3, among the 10 POI categories with the poorest performance on the test dataset using the method proposed in this paper, the RoBERTa and RoBERTa-RNN models outperform each other across five categories. When considering the average F_1 score across the 10 POI data categories, the RoBERTa-RNN model is superior. The sub-optimal recognition performance for these specific POI data categories can be attributed to several factors: (1) a lack of semantic correlation between the labels and the POI names, frequently observed in the "Clothing and leather goods store" category. (2) POI names possess ambiguous semantics, as some POI names in the data omit proper nouns, for instance, the POI name "Jinbei Decoration" could be classified as either an "Advertising decoration company" or a "Shopping store". (3) POI names frequently encompassing semantic elements pertinent to multiple categories, a phenomenon particularly prevalent in the "Shopping" category of POI data. It is evident that the selection of labels is crucial for accurate POI data classification.

5. Conclusions

Most classification methods implicitly extract semantic knowledge from various categories within the training dataset, yet they often overlook the explicit semantic information provided by natural language labels. However, in this study, the explicit information carried by these labels is highly beneficial for the semantic recognition of POI data. Therefore, to enhance the identification of semantic categories in POI data, we introduce explicit semantic information derived from POI data labels. A transformation strategy is employed to convert the multi-classification task into a binary classification problem by introducing multiple negative samples for each positive input class. The correlation between the labels of POI data and their respective names is determined through the cross-coding technique employed by RoBERTa. Specifically, we employ RoBERTa to evaluate the relationship between pairs of sentences.

To assess the effectiveness of semantic enhancement and the benefits of transforming multi-classification into binary classification, we conducted a comparative analysis between our proposed method and RoBERTa-RNN's multi-classification method. The experimental results demonstrate that our proposed method outperforms the latter, achieving a significantly higher $macro - F_1$ score of 1.3% on the test dataset. Furthermore, the RoBERTa-RNN model encounters challenges with imbalanced data categories and has a tendency to overfit. When comparing RoBERTa with BERT and ALBERT in terms of transforming the multi-classification task into a binary classification problem, it is observed that RoBERTa achieves higher $macro - F_1$ score on the test dataset, exceeding those of BERT and ALBERT by 0.6% and 5.2%, respectively. These results indicate that RoBERTa possesses a superior capacity for capturing semantic nuances between sentence pairs, which substantiates the advantages of the method proposed in this study.

Author Contributions: Qingyun Du proposed the original idea and conducted the organization of the content. Hongwei Zhang carried out experiments and analysis of the results. Hongwei Zhang wrote the paper. Shuai Zhang and Renfei Yang put forward some suggestions on the modification of this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Key Laboratory of Digital Mapping and Land Information Application, Ministry of Natural Resources (ZRZYBWD202307); the Natural Science Foundation of Chongqing, China (Grant No. 2024NSCQ-MSX0321); and the Doctor's "Direct Train" research project of Chongqing, China (No. CSTB2022BSXM-JSX0020). The work was also supported by funds from the science and technology research program of Chongqing Municipal Education Commission (Grant No. KJQN202100734).

Data Availability Statement: The analysis datasets for the current study are available from the first author on reasonable request(zhanghongwei@cqut.edu.cn).

Acknowledgments: The authors thank the reviewers and editors for their constructive comments on this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, K.; Yin, L.; Lu, F.; Mou, N. Visualizing and exploring POI configurations of urban regions on POI-type semantic space. *Cities* **2020**, *99*, 102610. [CrossRef]
2. Milias, V.; Psyllidis, A. Assessing the influence of point-of-interest features on the classification of place categories. *Comput. Environ. Urban Syst.* **2021**, *86*, 101597. [CrossRef]
3. Wang, Z.; Zhu, Y.; Zhang, Q.; Liu, H.; Wang, C.; Liu, T. Graph-enhanced spatial-temporal network for next POI recommendation. *ACM Trans. Knowl. Discov. Data* **2022**, *16*, 1–21. [CrossRef]
4. Cai, Z.; Yuan, G.; Qiao, S.; Qu, S.; Zhang, Y.; Bing, R. FG-CF: Friends-aware graph collaborative filtering for POI recommendation. *Neurocomputing* **2022**, *488*, 107–119. [CrossRef]
5. Funke, S.; Storandt, S. Automatic tag enrichment for points-of-interest in open street map. In Proceedings of the Web and Wireless Geographical Information Systems: 15th International Symposium, Shanghai, China, 8–9 May 2017; pp. 3–18.
6. Zhou, J.; Gou, S.; Hu, R.; Zhang, D.; Xu, J.; Jiang, A.; Li, Y.; Xiong, H. A collaborative learning framework to tag refinement for points of interest. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1752–1761.
7. McKenzie, G.; Janowicz, K.; Gao, S.; Gong, L. How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. *Comput. Environ. Urban Syst.* **2015**, *54*, 336–346. [CrossRef]
8. Chang, B.; Park, Y.; Kim, S.; Kang, J. DeepPIM: A deep neural point-of-interest imputation model. *Inf. Sci.* **2018**, *465*, 61–71. [CrossRef]
9. Zhang, Z.; Zhao, X.; Wang, G.; Bi, X.J.C.C. A new point-of-interest classification model with an extreme learning machine. *Cogn. Comput.* **2018**, *10*, 951–964. [CrossRef]
10. Zhou, C.; Yang, H.; Zhao, J.; Zhang, X. POI classification method based on feature extension and deep learning. *J. Adv. Comput. Intell. Intell. Inform.* **2020**, *24*, 944–952. [CrossRef]
11. Fu, C.; McKenzie, G.; Frias-Martinez, V.; Stewart, K. Identifying spatiotemporal urban activities through linguistic signatures. *Comput. Environ. Urban Syst.* **2018**, *72*, 25–37. [CrossRef]
12. Li, X.; Wang, X.; Li, P.; Li, X.; Luo, A. POI Automatic Classification Method based on Word2Vec and Support Vector Machine. *Sci. Surv. Mapp.* **2022**, *47*, 195–203.
13. Duan, L. Textual Classification Retrieval Approach for POI Documents Based on Random Terms Iterative Model. *Appl. Res. Comput.* **2014**, *31*, 3024–3027.
14. Koziarski, M.; Krawczyk, B.; Woźniak, M. Radial-based oversampling for noisy imbalanced data classification. *Neurocomputing* **2019**, *343*, 19–33. [CrossRef]
15. Xiaolong, X.; Wen, C.; Yanfei, S. Over-sampling algorithm for imbalanced data classification. *J. Syst. Eng. Electron.* **2019**, *30*, 1182–1191.
16. Koziarski, M. Radial-based undersampling for imbalanced data classification. *Pattern Recognit.* **2020**, *102*, 107262. [CrossRef]
17. Hoyos-Osorio, J.; Alvarez-Meza, A.; Daza-Santacoloma, G.; Orozco-Gutierrez, A.; Castellanos-Dominguez, G. Relevant information undersampling to support imbalanced data classification. *Neurocomputing* **2021**, *436*, 136–146. [CrossRef]
18. Fu, S.; Tian, Y.; Tang, J.; Liu, X. Cost-sensitive learning with modified Stein loss function. *Neurocomputing* **2023**, *525*, 57–75. [CrossRef]
19. Zheng, W.; Zhao, H. Cost-sensitive hierarchical classification for imbalance classes. *Appl. Intell.* **2020**, *50*, 2328–2338. [CrossRef]
20. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
21. Schapire, R.E. The strength of weak learnability. *Mach. Learn.* **1990**, *5*, 197–227. [CrossRef]
22. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
23. Fürnkranz, J. Round robin classification. *J. Mach. Learn. Res.* **2002**, *2*, 721–747.
24. Rifkin, R.; Klautau, A. In defense of one-vs-all classification. *J. Mach. Learn. Res.* **2004**, *5*, 101–141.
25. Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D.L.; Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **2017**, *33*, I37–I48. [CrossRef] [PubMed]

26. Meishan, Z.; Nan, Y.; Guohong, F. A simple and effective neural model for joint word segmentation and POS tagging. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1528–1538.
27. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.
28. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MI, USA, 2–7 June 2019; pp. 4171–4186.
29. Giannopoulos, G.; Meimaris, M. Learning Domain Driven and Semantically Enriched Embeddings for POI Classification. In Proceedings of the 16th International Symposium on Spatial Temporal Databases, Vienna, Austria, 19–21 August 2019; pp. 241–247.
30. Wu, J.; Hu, R.; Li, D.; Ren, L.; Hu, W.; Xiao, Y. Where have you been: Dual spatiotemporal-aware user mobility modeling for missing check-in POI identification. *Inf. Process. Manag.* **2022**, *59*, 103030. [CrossRef]
31. Noorian, S.S.; Psyllidis, A.; Bozzon, A. ST-Sem: A Multimodal Method for Points-of-Interest Classification Using Street-Level Imagery. In Proceedings of the Web Engineering: 19th International Conference, Daejeon, Republic of Korea, 11–14 June 2019; pp. 32–46.
32. Choi, S.J.; Song, H.J.; Park, S.B.; Lee, S.J. A POI categorization by composition of onomastic and contextual information. In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, Poland, 11–14 August 2014; pp. 38–45.
33. Ijaz, M.F.; Alfian, G.; Syafrudin, M.; Rhee, J. Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest. *Appl. Sci.* **2018**, *8*, 1325. [CrossRef]
34. Agusta, Z.P. Modified balanced random forest for improving imbalanced data prediction. *Int. J. Adv. Intell. Inform.* **2019**, *5*, 58–65. [CrossRef]
35. Cao, L.; Shen, H. CSS: Handling imbalanced data by improved clustering with stratified sampling. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e6071. [CrossRef]
36. Khan, S.H.; Hayat, M.; Bennamoun, M.; Sohel, F.A.; Togneri, R. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *29*, 3573–3587. [CrossRef]
37. Du, H.; Teng, S.; Zhang, y. Transductive Support Vector Machine Based on Cooperative Labeling. *J. Chin. Comput. Syst.* **2016**, *37*, 2443–2447.
38. Smith, D.; Rahman, A.; Bishop-Hurley, G.J.; Hills, J.; Shahriar, S.; Henry, D.; Rawnsley, R. Behavior classification of cows fitted with motion collars: Decomposing multi-class classification into a set of binary problems. *Comput. Electron. Agric.* **2016**, *131*, 40–50. [CrossRef]
39. Yu, W.; Wei, C.; Chen, J. Predicting the Unbalanced Labels of POIs on Digital Maps using Hierarchical Model. *Acta Geod. Et Cartogr. Sin.* **2023**, *52*, 318–328.
40. Allwein, E.L.; Schapire, R.E.; Singer, Y. Reducing multiclass to binary: A unifying approach for margin classifiers. *J. Mach. Learn. Res.* **2000**, *1*, 113–141.
41. Aly, M. Survey on multiclass classification methods. *Neural Netw* **2005**, *19*, 2.
42. Zhang, H.; Ren, F.; Li, H.; Yang, R.; Zhang, S.; Du, Q. Recognition method of new address elements in Chinese address matching based on deep learning. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 745. [CrossRef]
43. Zhang, L.; Liu, Z.; Zhang, W.; Zhang, D. Style uncertainty based self-paced meta learning for generalizable person re-identification. *IEEE Trans. Image Process.* **2023**, *32*, 2107–2119. [CrossRef]
44. Lan, Z.; Mingda, C.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. In Proceedings of the ICLR 2020: International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
45. Smith, L.N. Cyclical learning rates for training neural networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Santa Rosa, CA, USA, 24–31 March 2017; pp. 464–472.
46. Reimers, N.; Gurevych, I. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 338–348.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.