International Journal of Geo-Information

MDPI

Article

Automatic Extraction and Cluster Analysis of Natural Disaster Metadata Based on the Unified Metadata Framework

Zongmin Wang [1], Xujie Shi [1], Haibo Yang [1,*], Bo Yu [2] and Yingchun Cai [1]

[1] School of Water Conservancy and Transportation, Zhengzhou University, Zhengzhou 450001, China; zmwang@zzu.edu.cn (Z.W.); sxj_014499@gs.zzu.cn (X.S.); yccai@zzu.edu.cn (Y.C.)
[2] CEC Guiyang Exploration and Design Research Institute Co., Guiyang 550081, China; yub_gyy@powerchina.cn
* Correspondence: yanghb@zzu.edu.cn

Abstract: The development of information technology has led to massive, multidimensional, and heterogeneously sourced disaster data. However, there's currently no universal metadata standard for managing natural disasters. Common pre-training models for information extraction requiring extensive training data show somewhat limited effectiveness, with limited annotated resources. This study establishes a unified natural disaster metadata standard, utilizes self-trained universal information extraction (UIE) models and Python libraries to extract metadata stored in both structured and unstructured forms, and analyzes the results using the Word2vec-Kmeans cluster algorithm. The results show that (1) the self-trained UIE model, with a learning rate of 3×10^{-4} and a *batch_size* of 32, significantly improves extraction results for various natural disasters by over 50%. Our optimized UIE model outperforms many other extraction methods in terms of precision, recall, and F1 scores. (2) The quality assessments of consistency, completeness, and accuracy for ten tables all exceed 0.80, with variances between the three dimensions being 0.04, 0.03, and 0.05. The overall evaluation of data items of tables also exceeds 0.80, consistent with the results at the table level. The metadata model framework constructed in this study demonstrates high-quality stability. (3) Taking the flood dataset as an example, clustering reveals five main themes with high similarity within clusters, and the differences between clusters are deemed significant relative to the differences within clusters at a significance level of 0.01. Overall, this experiment supports effective sharing of disaster data resources and enhances natural disaster emergency response efficiency.

Keywords: metadata extraction; UIE; natural disaster; Word2vec-Kmeans clustering

Citation: Wang, Z.; Shi, X.; Yang, H.; Yu, B.; Cai, Y. Automatic Extraction and Cluster Analysis of Natural Disaster Metadata Based on the Unified Metadata Framework. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 201. https://doi.org/10.3390/ijgi13060201

Academic Editors: Wolfgang Kainz, Christos Chalkias, Marinos Kavouras, Margarita Kokla and Mara Nikolaidou

Received: 20 March 2024
Revised: 20 May 2024
Accepted: 12 June 2024
Published: 14 June 2024

1. Introduction

Since the beginning of the 21st century, the frequency and quantity of major global disaster events have been increasing due to climate change [1]. Natural disasters such as heavy rainfall and floods, droughts, landslides, and earthquakes have become more frequent, widespread, and intense, posing significant challenges to emergency decision-making, response, and assessment. Integrating vast, multidimensional, and heterogeneous dispersed information resources, establishing unified standards and specifications for describing information resources, and achieving effective sharing of data resources are urgent problems to be addressed in natural disaster emergency response. Metadata [2,3], as the core information describing data characteristics, content, quality, and structure, is crucial for maximizing the utilization of disaster data. It not only forms the foundation for effective disaster information management but also serves as an indispensable basis for scientific decision-making and emergency response. The various types of data required for emergency response to earthquakes, floods, landslides, and other natural disasters necessitate the establishment of a unified metadata management approach under a unified metadata framework to achieve consistency and standardized management of these data resources. However, the efficient extraction and matching of massive heterogeneous disaster

data and information require methods beyond traditional manual processing, which are increasingly inadequate to meet the demands of modern disaster management [4–6].

The research and practice of natural disaster metadata standards is an important part of the field of disaster management and response. There are many organizations and individuals worldwide dedicated to the research of comprehensive disaster metadata standards [7–9]. For example, the Geoscience Australia Metadata, the digital geospatial metadata content standard FGDC-CSDGM, and the emergency disaster database (EM-DAT), one of the most important internationally, are all designed to serve disaster prevention and mitigation. In addition, in China, Chen Ke et al. [10] have designed natural disaster metadata standards, which have been applied in the development of natural disaster loss databases in the Yangtze River Delta region. In terms of specific natural disaster types, research on metadata standards for earthquakes [11,12], landslides [13], and floods [14] has made some progress. In addition, in the field of relevant information, the development of standards such as the Dublin Core Metadata Standard and ISO 19115 [15] promotes data sharing and communication [16]. At present, the research on the comprehensive metadata standards of natural disasters is still in the stage of continuous development and improvement. There are some metadata standards applicable to different types of disasters, but the metadata of most projects are maintained in the form of documents, lacking a unified metadata standard format and management mode.

How to extract massive multi-source and heterogeneous metadata information according to the metadata standard framework is also a key problem. With the rapid development of machine learning and deep learning, support vector machines (SVM) [17], decision tree, random forest [18], and other machine learning algorithms are used to extract structured information from text. Deep learning techniques such as recurrent neural networks (RNNs) [19], long short-term memory networks (LSTMs) [20], transformer, etc., have also made significant progress in metadata extraction tasks [21]. In particular, the emergence of pre-trained models [22] such as BERT and GPT make it so the metadata extraction task can be modeled and trained in an end-to-end way, which greatly improves the accuracy and efficiency of extraction. However, these models face the challenges of high demand for computational resources, large model parameters, and low interpretability [23]. In contrast, UIE [24], a unified information extraction framework, customizes extraction goals through natural language to achieve out-of-the-box use and meet various information extraction needs. The model can support the extraction of key information without limiting industry fields and extraction targets, achieve rapid cold start of zero sample (zero-shot), and have excellent fine-tuning ability of small samples (few-shot) to quickly adapt to specific extraction targets [25]. Its open-domain information extraction technology reduces the dependence of annotation data, thereby improving development efficiency and simultaneously lowering costs.

In summary, how to realize the automatic extraction of multi-source and heterogeneous disaster metadata information on the basis of constructing a unified natural disaster metadata standard is a problem worthy of in-depth exploration. Therefore, this paper establishes a unified natural disaster metadata model framework, adopts UIE and the Python parsing library to achieve automatic metadata extraction and evaluate the extraction results, and conducts cluster analysis on the extraction results using the Word2vec-Kmeans clustering algorithm. The model selected for the study improved the extraction effect significantly, and the metadata model framework constructed in this study demonstrates high-quality stability.

The innovative points of this article are as follows:

(1) We construct a unified metadata model framework for natural disasters based on core metadata and complete metadata.
(2) UIE and the Python analysis library are used to realize automatic extraction of unstructured and structured disaster metadata information, and corresponding constraint rules are formulated to establish an evaluation system from the three dimensions of consistency, completeness, and accuracy.

(3) The Word2vec-Kmeans clustering algorithm is used to realize cluster analysis of the extraction results.

The findings can help to provide more effective support for efficient disaster management.

2. Research Design

This paper builds a unified metadata standard framework suitable for natural disasters based on the metadata standard of the survey. According to the framework, for metadata stored in structured formats, different Python libraries are used to parse the corresponding data format and extract relevant metadata within the framework. For metadata stored in unstructured formats, UIE is utilized to extract the required information and the parameters are adjusted based on the loss curves and F1 scores to achieve optimal extraction results. Subsequently, the extracted optimal results are stored in the database, and corresponding constraint rules are established for evaluation from three aspects: consistency, completeness, and accuracy. Finally, the Word2vec-Kmeans clustering algorithm is used to analyze the extraction results, which improves the efficiency and accuracy of retrieval and recommendation by classifying the text into different groups. The technical route of this study is shown in Figure 1.

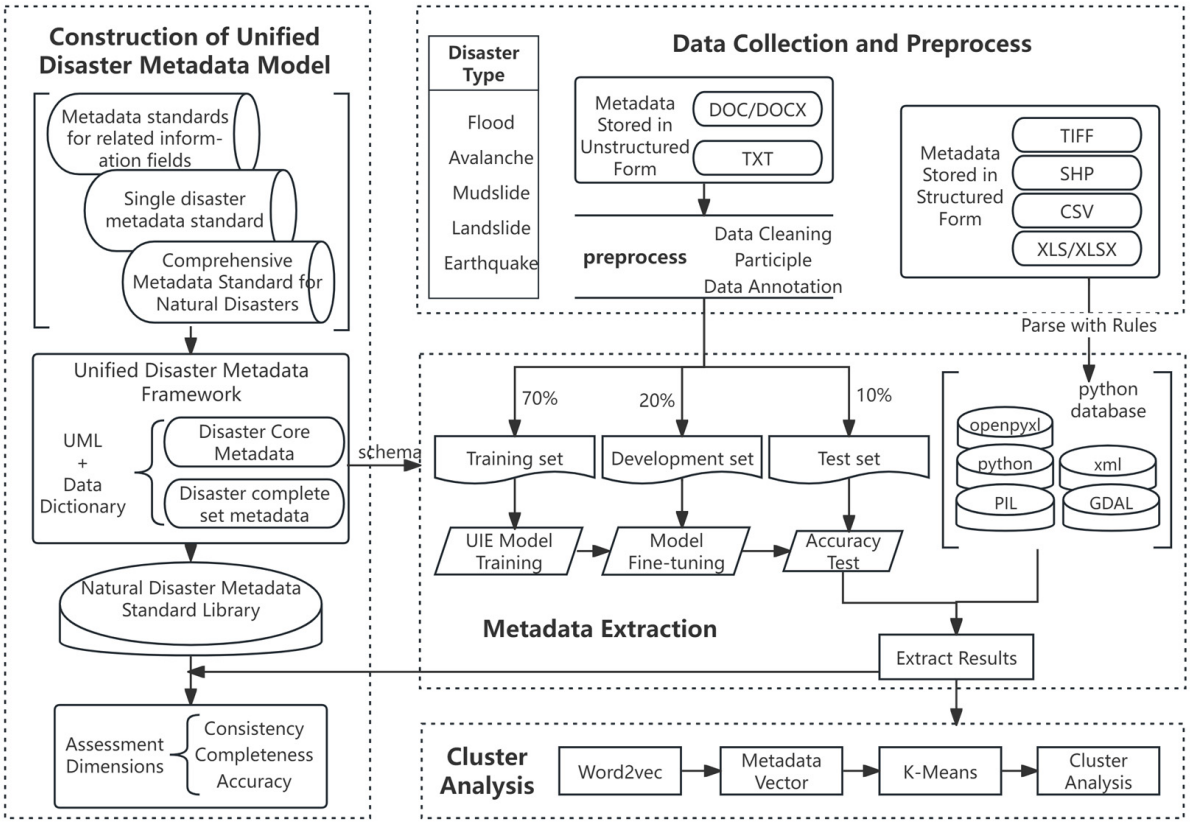

Figure 1. Technical flow chart of this study.

2.1. Experimental Setup and Data Pre-Processing

As shown in Table 1, the data sources selected in this study are the National Earth Observation Science Data Center and the National Comprehensive Earth Observation Data Sharing Platform, which are responsible for operating the Chinese Remote Sensing Satellite Ground Station of the Academy of Space and Space Information Innovation, Chinese Academy of Sciences. The experimental data cover five types of natural disasters, including floods, earthquakes, avalanches, landslides, and mudslides, and the disaster processes include pre-disaster, during-disaster, and post-disaster stages. In terms of the temporal dimension, the data cover small-, medium-, and large-scale disaster events that occurred

across China between 1900 and 2022. In the spatial dimension, the data cover disaster locations nationwide. The data are stored in various formats, including XLS/XLSX, SHP, TIFF, DOC/DOCX, CSV, and TXT, totaling over 65 GB. Among them, the disaster metadata information stored in unstructured formats includes 268 documents, with a data volume of 24.5 MB.

Table 1. Experimental data.

Disaster Type	Time	Data Format	Data Size	Data Source	Counts of Events
Flood	2013–2021	XLS/XLSX, SHP, TIFF, DOC/DOCX, CSV, TXT	5.25 GB	ChinaGEOSS Data Portal (china-geoss.cn) [26]	42
Earthquake	1900–2022		713 MB		76
Landslide	1995–2022		58 GB		56
Mudslide	2005–2022		429 MB		47
Avalanche	2006–2019		854 MB		47

Data preprocessing includes data cleaning, word segmentation, and data annotation for the dataset description document.

(1) Clean the original text data, including removing special characters and some meaningless characters to reduce the impact of noise on model training, removing consecutive repetitive characters and redundant blanks in the text to simplify the text structure and improve the efficiency of the analysis, and detecting and deleting duplicate records or duplicate text in the text data to ensure the uniqueness of the data.

(2) Cleaned text is subjected to the segmentation process, which slices the text into sequences of words, converts them into machine-readable forms, and filters out some common stop words. For some datasets, a comparison of segmentation between jieba and Natural Language Processing and Information Retrieval (NLPIR) is conducted. From Table 2, it can be seen that the accuracy of jieba is 65.54%, while the precision and recall of NLPIR are relatively lower. When processing disaster metadata text information, jieba segmentation performs better, with much higher accuracy than NLPIR and relatively fewer cases of text loss.

Table 2. Comparison of word segmentation results.

The Name of the Word Tokenizer	P	R	F1
NLPIR	0.298	0.380	0.334
Jieba	0.655	0.731	0.691

(3) After tokenizing the words, perform lemmatization and part-of-speech tagging on the words. Utilize the Doccano tool for annotation work to determine entity, relationship, and other information corresponding to each text. Simultaneously, convert the data into the JSON format acceptable by the model.

The preprocessed dataset was split into three groups: 70% for training, 20% for development, and 10% for testing to better ensure that the model was able to fully learn the features and patterns of the data and reduce the risk of overfitting and ultimately improve the performance of the model. The UIE model was trained using the training dataset and the model parameters were fine-tuned based on its accuracy on the validation dataset. Subsequently, the accuracy of the best-performing model on the test dataset was evaluated to assess its performance.

In this study, the hardware environment is a computer configured with NVIDIA geforce RTX 3050ti GPU, 64 GB memory, Intel Core i7-12700k processor and 1 TB SSD storage. The software environment includes the Windows 11 operating system, PaddePaddle 2.6.1 as the main deep learning framework, Python 3.8.5 as the programming language, and dependencies such as PaddleNLP, NumPy, and Matplotlib, among others.

2.2. Unified Disaster Metadata Model Construction

As shown in Table 3, this study surveyed specific standards in three aspects: comprehensive metadata standards for natural disasters, metadata standards for single disaster types, and metadata standards in related information fields. We propose a unified metadata model for natural disasters, adopting a bottom-up approach. First, common metadata elements are selected from the models of the three domains, and common element names are summarized and their frequencies calculated as basic elements. Second, the higher-level categories to which common elements belong are organized, and category names are standardized. Third, the categories that have been classified are summarized and generalized, grouping categories of similar functions to form a system and thus ultimately refining a subset of the metadata model.

Table 3. Related information field metadata.

Natural Disaster Comprehensive Metadata	Individual Natural Disaster Metadata		Related Information Field Metadata
EM-DAT	Earthquake eML	GB/T 24888-2010 [27]	Dublin Core Metadata Standard (DC1.1) [28]
EU-MEDIN RDF Schema	TWML	DB/T 41-2011 [29]	ISO19115 [30]
FGDC Content Standards for Digital Geospatial Metadata	Metadata Standard for Seismic Mitigation and Disaster Prevention Planning	Geological Disaster Emergency Information Resource Metadata Standard	Core Metadata Standard for Earth System Science Data Sharing
Geoscience Australia Metadata	General Metadata Standard for Emergency Field	Debris Flow Disaster Emergency Metadata Standard	GB/T 19710-2005/ISO 19115:2003, MOD [31]
Metadata Standard for Natural Disasters	Core Metadata for Earthquake Data Resources	Core Metadata for Geological Disaster Monitoring Dataset	
	CWML		

As shown in Table 4, by analyzing the metadata structure and the selection of metadata standard elements, it can be seen that the subsets with higher proportions are identification information, data quality information, content information, and other subsets among the 21 metadata standards selected for this survey, which are shown in the table. Based on the analysis results of existing disaster metadata standards, a disaster metadata structure framework that meets the requirements of disaster data and is also generic was constructed. On this basis, starting from core metadata and full metadata, a unified disaster metadata model standard suitable for disaster data is proposed in this study. It should be noted that it is necessary to match them with existing standards as much as possible under the premise of meeting their own usage requirements to ensure the sharing and interoperability of disaster metadata standards. Based on the comprehensive statistical analysis results and the actual needs of disaster data, some subsets with low frequencies can exist as secondary subsets of subsets with higher frequencies. For example, the storage information subset can be classified into content information, responsible unit information, or maintenance information, and contact information can be classified into reference and responsible person information, ultimately forming the basic primary subsets.

Disaster core metadata are the metadata information that disaster dataset producers must provide when providing data. Based on the above research and analysis, this study designed a disaster metadata standard that defines disaster metadata consisting of 30 core metadata elements to meet the needs of disaster metadata management, including disaster types, disaster processes, and other descriptive disaster characteristics. Table 5 provides a detailed explanation of disaster core metadata content in the form of a data dictionary. The "Constraint/Condition" column indicates whether the metadata entity or metadata element must be selected, including mandatory (M), optional (O), and conditionally mandatory (C) options.

Table 4. Metadata standard structure.

Structural Elements	Description of Sub-Elements or Corresponding Structural Elements	Abbreviations
Metadata identification information	The basic information needed to uniquely identify data resources	MDID info
Content information	Information describing the content of the dataset	Cont info
Data-quality information	Evaluation information regarding the quality of the dataset	DQ info
Restriction information	Information containing restrictions on access and use of resources	Restr info
Distribution information	Description of the distributor of the dataset and methods for obtaining data, providing reference material names and dates, as well as responsible unit names, duties, contacts, and other information	Distr info
Metadata reference information	Contains descriptions of the metadata standards themselves, including metadata standard names, versions, etc.	MDRef info
Reference system information	Provides spatial reference system and temporal reference system information	RS info
Extended information	Provides extension information for implementation when specialized standards need to be established and the required metadata elements or entities are not present in this standard	Ext info
Citation and responsible party information	Provides information about responsible units and individuals related to the data, as well as materials, datasets, models, or literature used for referencing or referring to the dataset	CRP info
Coverage information	Defines and describes metadata for the spatial and temporal coverage of resources	Cov info

Table 5. Disaster core metadata.

Name/Role Name	Constraint/Condition	Name/Role Name	Constraint/Condition
Use restrictions	O	Dataset keywords	M
System unique identifier ID	M	Dataset identifier/ID	C
Dataset type	M	Dataset creation time	M
Dataset title	M	Dataset contact information	M
Dataset thumbnail	O	Dataset character set	O
Dataset summary	M	Dataset access restrictions	O
Dataset subject category	M	Data-quality report	C
Dataset security restriction level	M	Data log	C
Dataset language	M	Data format	M

Disaster complete metadata describe all metadata information related to disaster information, including mandatory and optional metadata entities and elements (UML attributes), as illustrated in Figure 2. It encompasses entities such as MDID info, Cont info, DQ info, Restr info, Distr info, MDRef info, RS info, Ext info, CRP info, and Cov info. Among them, MDID info provides necessary details to uniquely identify resources, and MDRef info pertains to dataset metadata. These two entities exist uniquely. Other entities are not essential for metadata information and may appear zero to n times or zero to one time based on specific data usage requirements. The system unique identifier is an element uniquely identifying metadata and occurs exactly once. The character set specifies the character set code used by metadata and this element is required when character coding and UTF-8 are not used.

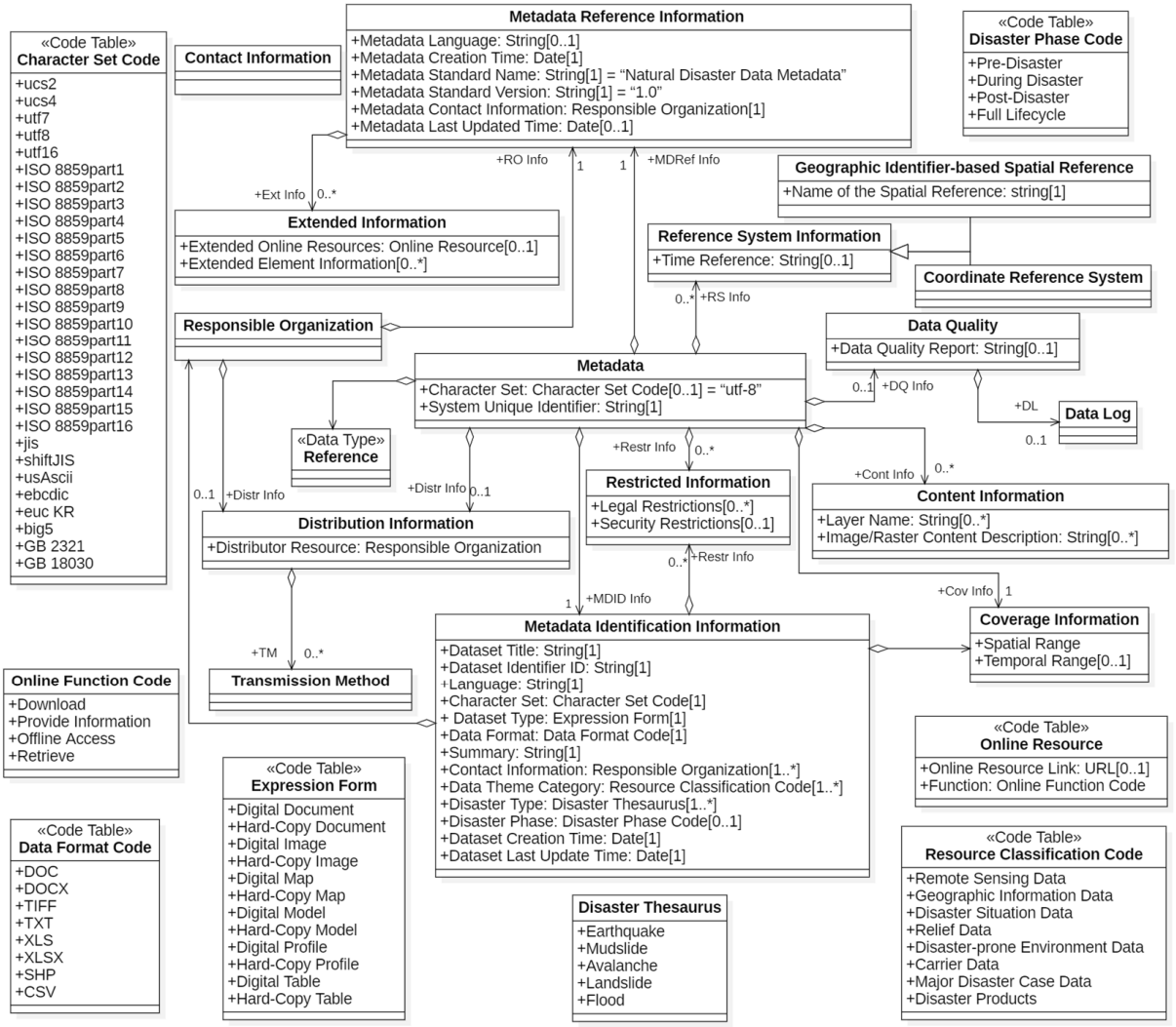

Figure 2. UML structure of disaster metadata.

2.3. Disaster Metadata Extraction

2.3.1. UIE-Based Extraction of Disaster Metadata Stored in Unstructured Form

The storage forms of disaster metadata usually include structured, unstructured, and semi-structured. This study focuses on the metadata information stored in structured and unstructured ways. Disaster metadata stored in unstructured form lack a fixed pattern or organizational structure and needs to be managed and analyzed with the help of deep learning techniques. Information extraction (IE) is a basic natural language processing (NLP) task that converts unstructured or semi-structured descriptions of natural language text into structured features. UIE is a unified framework for general information extraction proposed by Lu et al. [24] in ACL-2022 based on IE. Formally, UIE takes the given structural schema instructor (s) and the text sequence (x) as input and generates the linearized structured extraction language (SEL) (y), which contains the extracted information from x based on schema s:

$$y = UIE(s \oplus x) \tag{1}$$

where $x = [x_1, \ldots, x_{|x|}]$ is the text sequence, $s = [s_1, \ldots, s_{|s|}]$ is the structural schema guide, and $y = [y_1, \ldots, y_{|y|}]$ is an SEL sequence that can be easily converted to the extracted information record.

In the information extraction results of the UIE model, probability refers to the confidence or probability of the model in predicting specific information (such as entities, relationships, etc.). This probability value is usually calculated from the softmax function of

the model in the last layer, reflecting the degree of confidence of the model in the correctness of its prediction. The calculation process is performed as follows:

- Model forward propagation: Input the text of the given metadata into the trained UIE model for forward propagation. During the forward propagation process, the model processes the text and generates various predictions, including entity boundaries (start and end positions), entity types, and relationships between entities.
- Output layer: The model's output layer typically includes a softmax function. For each predicted category, the softmax function normalizes the predicted scores, converting them into probability form. This ensures that the probabilities of all categories sum up to 1.
- Probability calculation: For each prediction, the softmax function of the model output layer calculates a probability value indicating the confidence the model believes is correct.

In the above calculation process, entities refer to various types of metadata information, such as geographic locations, time, events, people, organizations, etc. Relationships describe the connections or dependencies between different entities. Taking time information metadata as an example, entities may include dates, time periods, etc., and relationships can represent the sequential relationships between times, such as before, after, simultaneous, etc.

The benchmark model in this paper was implemented using PaddleNLP, which borrows the framework from UIE. The model was constructed using Python 3.8 in Task Flow. The model parameters of the UIE are shown in the Table 6. To study the best combination of UIE parameters, certain parameters were adjusted for the experiments within the possible values based on the prediction accuracy of the validation set [32], including learning_rate and batch_size, which are also the most sensitive parameters of the deep learning models in most cases [33].

Table 6. UIE model parameters.

Parameter	Description	Experimental Setting
Learning rate	The time interval for updating model parameters per training epoch; values range from 0 to 1.	Adjusting from 1×10^{-3} or 1×10^{-4} to 3×10^{-4}
Batch_size	Batch size	Adjusting from batch sizes of 16 or 32 to 64
num_epochs	Number of training epochs	Setting the maximum iteration rounds to 400
Model	Model selection: program performs model fine-tuning based on the selected model	UIE-base

2.3.2. Analysis of Disaster Metadata Stored in a Structured Form

Structured data are often organized in a well-defined format and structure, with each data field having a specific meaning and data type. The data formats for this study include SHP, TIFF, XLS/XLSX, and CSV. XLS/XLSX and CSV files are typical forms of structured data. The SHP file itself stores geospatial data, but the property table data associated with it usually exists in a structured form. In TIFF files, attribute information is stored in "tags", and each tag has a unique numerical identifier and corresponding data value. From this point of view, attribute information in TIFF files can be considered structured. In this study, pyshp library, GDAL library, XML parsing library, PIL library, and other third-party libraries in Python were used to read the basic information of SHP, TIFF, and XLS/XLSX files. The first row of a CSV file usually contains column labels or table headers. The metadata information is used as part of the column labels or table headers. When parsing the file, these labels were checked, and if comments existed in the labels, regular expressions were used to extract metadata from the comments. The specific parsing process for each type of data is shown in Figure 3.

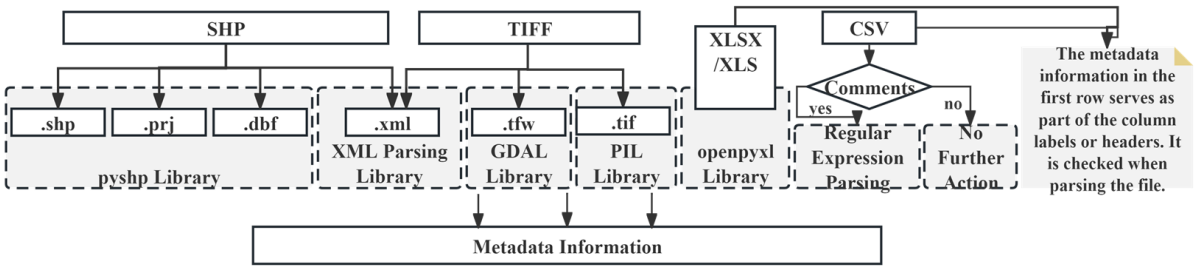

Figure 3. Schematic diagram of metadata extraction stored in a structured form.

2.4. Quality Assessment of Disaster Metadata

Wang R.Y et al. [34] conducted an analysis and survey of 118 attributes to address user requirements for data. As a result, they proposed 20 commonly used evaluation dimensions, such as integrity, consistency, and accuracy. Building upon this foundation, this study established corresponding constraint rules to assess the quality of disaster metadata in terms of consistency, completeness, and accuracy after being stored in the database [35–37]. The completeness constraint rule is one of the most fundamental rules, serving as a prerequisite for ensuring data entry into the database. The consistency constraint rule describes the semantic consistency among different attributes in the data table, including both internal within a data table and across multiple data tables. The accuracy constraint rule describes the accuracy of data values in terms of both form and content. The details are shown in Tables 7–9.

Table 7. The constraint rules of completeness.

Completeness Constraint Rules	Content
Primary key constraint rule	Primary key attribute values must exist and be unique.
Composite primary key constraint rule	A primary key composed of two or more fields must exist and be unique.
Not null constraint rule	Values must exist and cannot be null (non-primary key).
Unique constraint rule	Values must be unique and cannot have duplicates (non-primary key).
Continuity constraint rule	Values must be continuous.
Candidate key constraint rule	Values must exist and be unique (non-primary key).

Table 8. The constraint rules of consistency.

Consistency Constraint Rules	Content
Foreign key constraint rule	The values of the foreign key attribute column in the relation table must be consistent with the attribute values of the associated primary key. That is, the values of the foreign key attribute column must be referenced by the primary key.
Equality consistency constraint rule	Values must be calculated based on one or more attribute columns in one or more relation tables.
Logical consistency constraint rule	Values must have a logical relationship with one or more attribute columns in one or more relation tables.
Existence consistency constraint rule	Values must have a matching relationship with another attribute column.

Table 9. The constraint rules of accuracy.

Accuracy Constraint Rules	Content
Data-type constraint rule	All value types must satisfy the data type defined under the attribute column.
Length constraint rule	String lengths must meet the given length constraint.
Precision constraint rule	Floating-point values must satisfy the given precision constraint.
Data format rule	Values must satisfy the given data format.
Value range rule	Values must be within the given value range.
Fixed-value constraint rule	Values must be in the given set.

(1) Completeness assessment

Given a relation R containing N tuples, with the attribute set $A = \{A_1, A_2, \ldots, A_m\}$, primary key constraint A_1 has a null value count M_1, union key constraint set $B = \{B_1, B_2, \ldots, B_n\}$ has a null value count M_2, and non-null constraint set $C = \{C_1, C_2, \ldots, C_t\}$ has a null value count M_3. Here, $B_i, C_i \in A\{i = 1, 2, \ldots, n; j = 1, 2, \ldots, m; t < m\}$, and C_j are singleton sets. Additionally, in all constraint rule metadata, violating the non-null constraint rule indicates a violation of data completeness. Therefore, the completeness function L_1 on relation R can be defined as:

$$L_1 = \left(1 - \frac{M_1 + M_3}{(1 + t) \times N}\right) \times 100\% \tag{2}$$

(2) Consistency assessment

In all constraint rules, violations of the name, alias, and dimensional consistency constraint rules would indicate violations of the consistency of the data and, therefore, the consistency evaluation L_2 be defined as:

$$L_2 = \left(1 - \frac{S_{cm} - S_{cr}}{S_{cm}} \times \frac{S_{cr}S_{cc} - S_{id} - S_{c1} - S_{c2} - S_{c3}}{S_{cr}S_{cc} - S_{id}}\right) \times 100\% \tag{3}$$

where S_{cm} represents the total number of records in the data, S_{cc} represents the number of attribute columns, S_{c1} represents the number of data points violating the equality consistency constraint rule, S_{c2} represents the number of data points violating the existence consistency constraint rule, S_{c3} represents the number of data points violating the logical consistency constraint rule, S_{c4} represents the number of data points violating the foreign key constraint rule, S_{c5} represents the number of data points violating the equality dependency constraint rule, S_{c6} represents the number of data points violating the logical dependency constraint rule, S_{c7} represents the number of data points violating the code constraint rule, S_{id} represents the number of empty data points in the problem records, and S_{cr} represents the size of the problem record set.

(3) Accuracy assessment

In all constraint rules, violations of length, precision, minimum, maximum, and fixed values indicate the accuracy of the data, so the accuracy evaluation algorithm L_3 is defined as:

$$L_3 = \left(1 - \frac{S_{cm} - S_{cr}}{S_{cm}} \times \frac{S_{cr}S_{cc} - S_{id} - S_{a1} - S_{a2} - S_{a3}}{S_{cr}S_{cc} - S_{id}}\right) \times 100\% \tag{4}$$

where S_{a1} represents the number of data points violating the value range constraint rule, S_{a2} represents the number of data points violating the data type constraint rule, S_{a3} represents the number of data points violating the data format constraint rule, S_{a4} represents the number of data points violating the fixed-value constraint rule, and S_{a5} represents the number of data issues violating the precision constraint rule.

(4) Model accuracy assessment

The accuracy measures used in this study include precision *(P)*, recall *(R)*, and F1 scores (*F1*) [38]. *P* refers to the proportion of samples where the model is actually positive. *R* refers to the proportion of samples that are actually positive and that are predicted to be positive. F1 scores are the harmonic mean of *P* and *R*.

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{7}$$

where *TP* is the correct quantity in the extracted sample, *FP* is the quantity that is not accurately extracted, and *FN* is the incorrect quantity actually extracted in the split word sample.

2.5. Cluster Analysis of Disaster Metadata

Cluster analysis groups disaster data with similar features or attributes into clusters, aiding in identifying the associations and connections between different data. By visually displaying the distribution and characteristics of different categories of data, users can gain a more intuitive understanding and analysis of disaster data. This helps in discovering the interactions, influencing factors, and complex relationships among disaster data, which provides important clues for further data analysis and application. Word2vec is a simple yet effective word-embedding technique that is easy to understand and implement. Its context independence makes it particularly suitable for tasks such as computing semantic similarity and building word recommendation systems. Compared to more complex models like BERT, Word2vec consumes fewer resources.

Word2Vec [39,40] is a type of word-embedding model based on deep learning. It maps words into high-dimensional space, representing words in text as real-valued vectors. Word2Vec includes the continuous bag of words (CBOW) model and the skip-gram model. CBOW [41] is characterized by its fast-training speed, low demand for dense representations, and suitability for frequently occurring words compared to skip-gram. The CBOW model predicts the current semantic unit w based on the context (w). Firstly, the words in the context are mapped to a common semantic space using one-hot encoding at the input layer. Secondly, the resulting matrix is multiplied by a shared matrix *W* and summed, then averaged to update the hidden layer vector *h*. The vector *h* is then multiplied by the shared matrix *W′* and normalized using the softmax activation function to update the output layer weights. This completes the forward propagation process of the CBOW model. Subsequently, the cross-entropy cost function is used as the loss function, and the process of continuously reducing the loss is the backward propagation process [2]. The objective function is the negative logarithmic loss function.

$$-\sum_{t=1}^{T} logP\left(w^{(t)} \middle| w^{(t-m)}, \ldots, w^{(t-1)}\right), w^{(t+1)}, \ldots, w^{(t+m)} \tag{8}$$

where the text sequence length is *T*, the word $w^{(t)}$ index is *t*, and the window size is *m*.

Word vectors are typically high-dimensional. To visualize word vectors, dimensionality reduction techniques are employed to transform them into lower-dimensional vectors. t-Distributed stochastic neighbor embedding (t-SNE) [42,43] is a nonlinear dimensionality reduction algorithm. Its primary objective is to preserve the local similarities between high-dimensional data points and maintain these relationships as much as possible in the lower-dimensional space. This allows the visualized data points to better reflect the structure of the original data. By applying the t-SNE algorithm, high-dimensional data can

be mapped into a two-dimensional space, where each data point is represented by two coordinates (x and y).

The coordinate results after dimensionality reduction are taken as input, and the K-means algorithm is used to cluster them. K-means is a commonly used distance-based clustering algorithm [44–47]. The principle is to divide the text data into pre-specified K clusters, and the center of each cluster represents the average of all the samples in the cluster. The K-means algorithm iteratively assigns text to the nearest cluster and updates the center of the cluster until the convergence condition is reached.

$$J(c, \mu) = \sum_{i=1}^{k} x^{(i)} - \mu_{c^{(i)}}{}^2 \quad (9)$$

where $\mu_{c^{(i)}}$ represents the mean of the *i-th* cluster.

3. Results

3.1. UIE Model Optimization

Figure 4 illustrates the impact of adjusting the *learning rate*(*lr*) and *Batch_size* parameters on the model under different iterations. As shown in Figure 4a,b, when the *lr* was set to 1×10^{-3}, the training loss decreased very slowly with increasing iterations, indicating no convergence, and the F1 score was unstable, suggesting that the model lacked learning ability on the training set. When the *lr* was set to 1×10^{-4}, the training loss converged, and the F1 score fluctuated around 0.95. When the *lr* was set to 3×10^{-4}, the training loss converged the fastest, and the F1 score remained above 0.96. Therefore, in this study, the *lr* was set to 3×10^{-4}. As shown in Figure 4c,d, when the *Batch_size* was 16, 32, or 64, the training loss converged to below 0.5. When the *Batch_size* was 16 or 32, the F1 score on the test set fluctuated slightly within a small range above 0.96, and the convergence speed of the model with a *Batch_size* of 32 was slightly faster than that with a *Batch_size* of 16. When the *Batch_size* was 64, the amplitude of the F1 score was larger. Larger batch sizes usually accelerate the training speed, but larger batch sizes consume more memory. Therefore, a *Batch_size* of 32 was chosen for subsequent experiments.

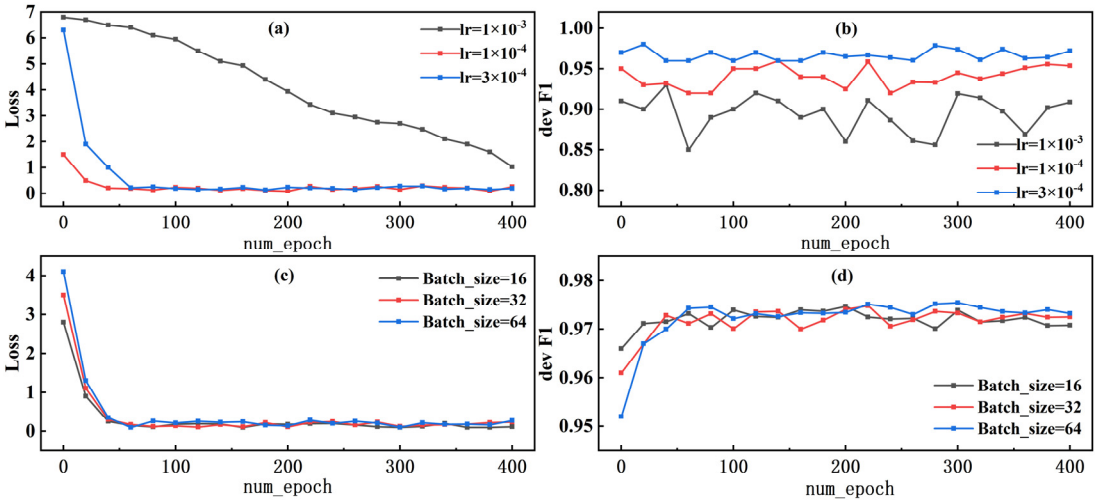

Figure 4. Training losses and F1 scores with different learning rates and batch sizes on the development set. (**a**) Training losses with *lr* ranges 1×10^{-3},1×10^{-4} to 3×10^{-4}; (**b**) F1 scores with *lr* ranges 1×10^{-3},1×10^{-4} to 3×10^{-4}; (**c**) training losses with *batch_size* ranges of 16, 32, and 64; (**d**) F1 scores with *batch_size* ranges of 16, 32, and 64.

3.2. Disaster Metadata Extraction and Quality Assessment

3.2.1. Results of Disaster Metadata Extraction

From Figure 5, it can be observed that before model optimization, the probability values of UIE for dataset metadata extraction of five types of disasters—floods, earthquakes,

landslides, mudslide, and avalanches—ranged between 0.50 and 0.60. Specifically, the average probability value (ave_pro) for flood domain metadata was 0.59, for earthquake domain metadata ave_pro was 0.58, for landslide domain metadata ave_pro was 0.56, for landslides domain metadata ave_pro was 0.57, and for avalanche domain metadata ave_pro was 0.58. After model optimization, the probability values ranged between 0.80 and 1.00. Specifically, the average probability value after optimization (t_ave_pro) for flood domain metadata was 0.89, for earthquake domain metadata t_ave_pro was 0.89, for landslide domain metadata t_ave_pro was 0.90, for landslide domain metadata t_ave_pro was 0.90, and for avalanche domain metadata t_ave_pro was 0.88. The extraction performance improved by 53.16%, 54.46%, 60.25%, 57.36%, and 53.22%, respectively, all exceeding 50%. It can be seen that the model performed best in extracting flood data before tuning, but the improvement in extracting flood metadata was relatively minima after tuning. Model tuning may have resulted in a more balanced handling of different categories of data, reducing over-reliance on flood metadata. So, the improvement in extracting flood metadata was relatively small. As a result, the experiments indicate that the model showed superior performance in extracting metadata for the five types of disasters, which makes it more suitable for extracting metadata information in the field of natural disasters after training and parameter optimization.

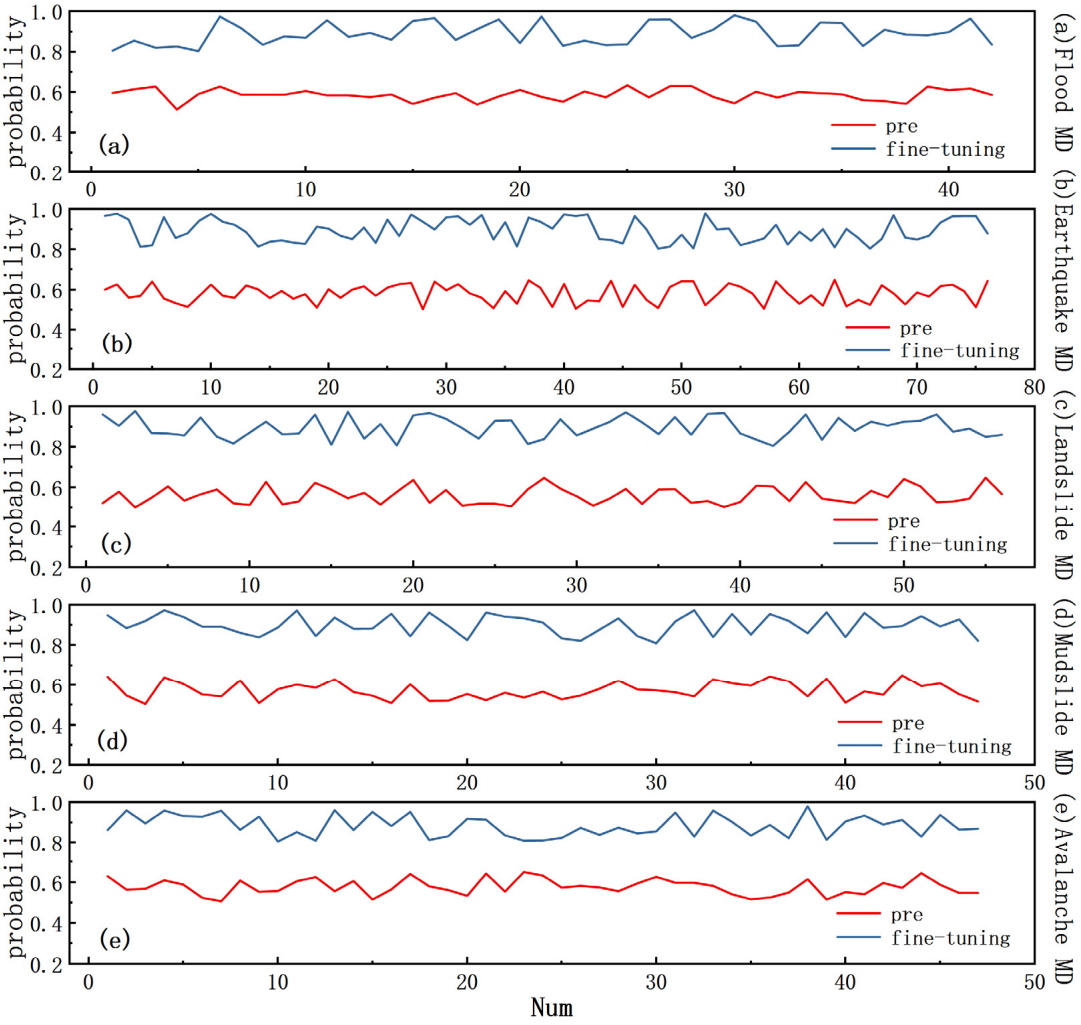

Figure 5. Comparison of the extraction effects of five disaster types before and after UIE model tuning. (**a**) Flood metadata; (**b**) earthquake metadata; (**c**) landslide metadata; (**d**) mudslide metadata; (**e**) avalanche metadata.

3.2.2. Quality Assessment of Extraction Results

Based on the extraction results, Figure 6 provides an evaluation of the data table, focusing on consistency, completeness, and accuracy. The quality assessment results of three dimensions were all above 0.80 for ten tables, with mean values of 0.86 for consistency, 0.94 for completeness, and 0.88 for accuracy. This indicates that the data in the tables maintained consistency among different records, with few missing data points in the tables and relatively complete key information, which refers to the constructed metadata standards having high quality and reliability. Meanwhile, the evaluation results of completeness were slightly higher than those of consistency and accuracy, indicating that the metadata information of the experimental source dataset matched well with the metadata standard framework. As for each table, the mean values of consistency, completeness, and accuracy for Cont info, RS info, CRP info, and Cov info were all above 0.90. Moreover, the variances of consistency, completeness, and accuracy among tables were 0.04, 0.03, and 0.05, indicating that the metadata model framework constructed in the experiment had high-quality stability.

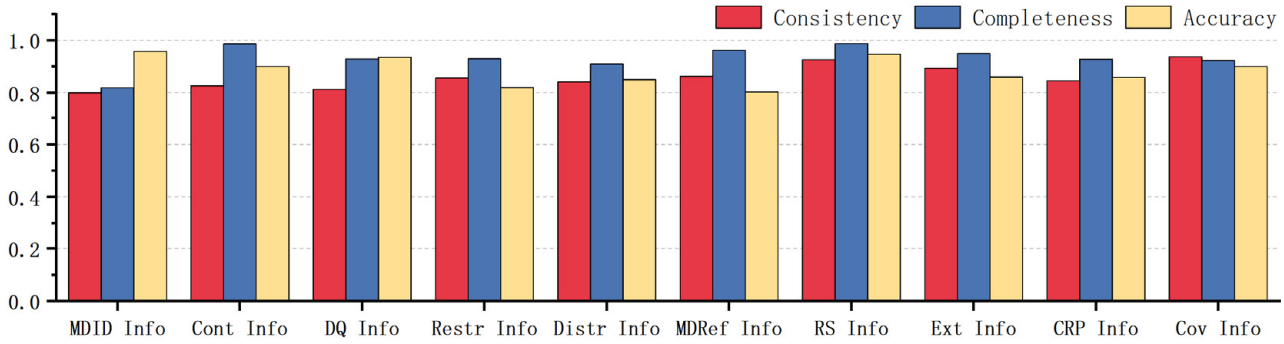

Figure 6. The consistency, completeness, and accuracy evaluation results of the data table dimension after being imported into the database.

Based on the extraction results, Figure 7 provides an evaluation of the data items in the data tables, focusing on consistency, completeness, and accuracy. The results were all above 0.80 and the results for the completeness assessment of data items in each table were higher than those for consistency and accuracy. Among them, the mean consistency of all data item metadata was 0.89, the mean completeness was 0.95, and the mean accuracy was 0.89. The overall trend of the results is consistent with that of the table level. The maximum value for consistency in the evaluation results was the "metadata standard name" item (MDSN = 0.95) in the "MDRef info" table, the maximum value for completeness was the "coordinate reference system" item (CR = 0.99) in the "RS info" table, and the maximum value for accuracy was the "character set" item (CS = 0.95) in the "MDID" table. The minimum value for consistency was the "description" item (De = 0.81) in the "Cov info" table, the minimum value for completeness was the "metadata creation time" item (MDCT = 0.90) in the "MDRef info" table, and the minimum value for accuracy was the "responsible unit" item (RN = 0.82) in the "CRP info" table. Overall, the data item quality was optimal in the "reference system information" table and relatively poor in the "Restr info" table, consistent with the results at the table level. The experimental results reflect the soundness of the overall database architecture, indicating that this study adopted effective metadata management strategies to ensure high-quality metadata.

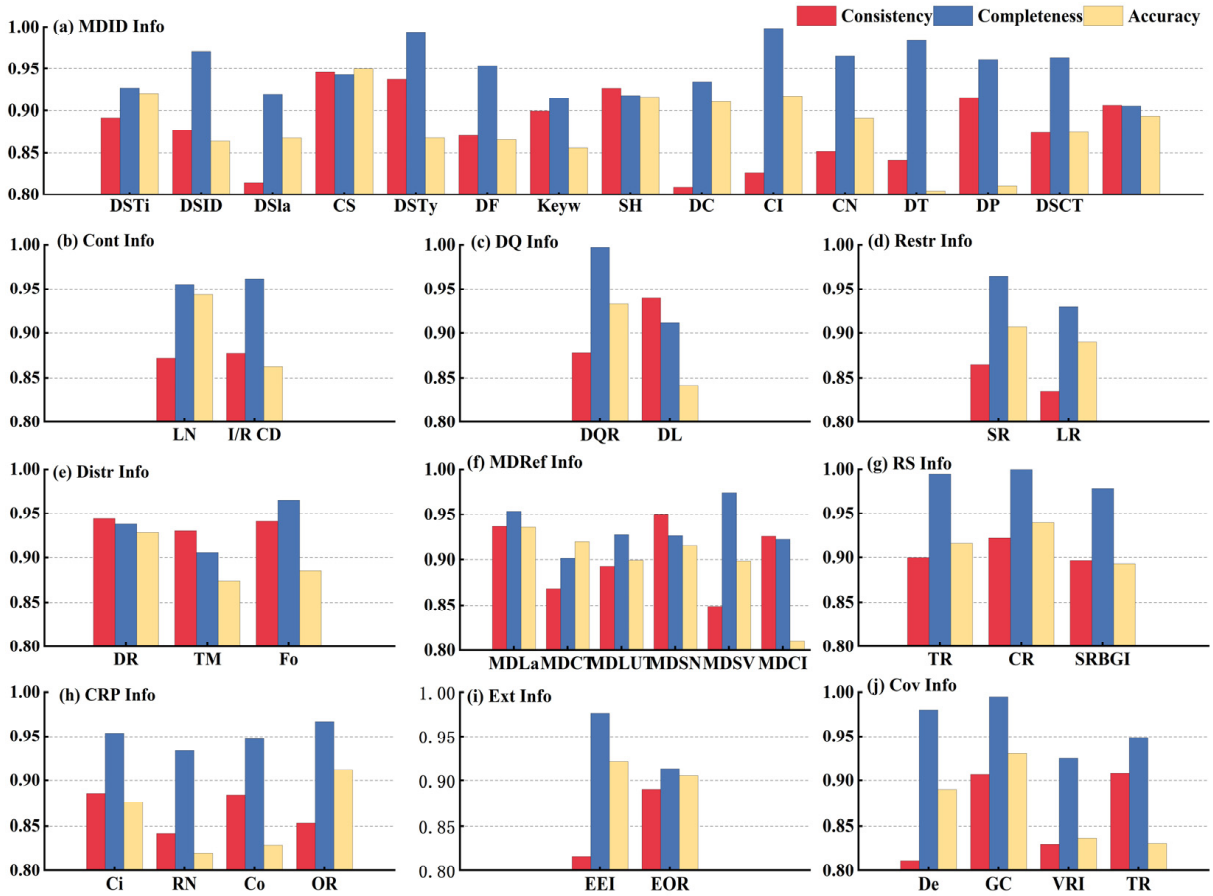

Figure 7. The consistency, completeness, and accuracy evaluation results of the data item dimensions of each data table after being transferred to the database. (**a**) DSTi is Dataset Title, DSID is Dataset Identifier ID, DSla is Dataset language, CS is Character set, DSTy is Dataset type, DF is Data format, Keyw is Keywords, SH is Subject headings, DC is Data category, CI is Coverage information, CN is Contact unit, DT is Disaster type, DP is Disaster process, DSCT is Dataset creation time, DSLUT is Dataset last update time; (**b**) LN is Layer name, I/R CD is Image/raster content description; (**c**) DQR is Data Quality Report, DL is Data Log; (**d**) SR is Security restrictions, LR is Legal restrictions; (**e**) DR is Distributor Resource, TM is Transmission Method, Fo is Format; (**f**) MDLa is Metadata language, MDCT is Metadata creation time, MDLUT is Metadata latest update time, MDSN is Metadata standard name, MDSV is Metadata standard version, MDCI is Metadata contact information; (**g**) TR is Time reference, CR is Coordinate reference, SRBGI is Spatial referencing based on geographical identifiers; (**h**) Ci is Citation, RN is Responsible Unit, Co is Contact, OR is Online Resources; (**i**) EEI is Extended element information, EOR is Extended online resources; (**j**) De is Description, GC is Geographic coverage, VRI is Vertical range information, TR is Time range.

3.3. Disaster Metadata Cluster

Taking flood data as an example, the extracted disaster metadata were transformed into corresponding word vectors using a pre-trained Word2Vec model after removing duplicate words, and served as input for the K-means clustering algorithm. To determine the optimal K-value in the K-means algorithm, the experiment combined the elbow method, the Calinski–Harabaz index, and the Davies–Bouldin index for evaluation. The elbow method reflects the clustering error SSE. From Figure 8a, it can be observed that the inflection point between rapid and slow decline occurred at n_cluster = 5. Additionally, from Figure 8b,c, it is evident that at n_cluster = 5, the Calinski–Harabaz index was highest, while the Davies–Bouldin index was lowest. This indicates better intra-cluster similarity and greater inter-cluster dissimilarity, respectively. Based on the experimental results, the optimal number of clusters was determined to be 5.

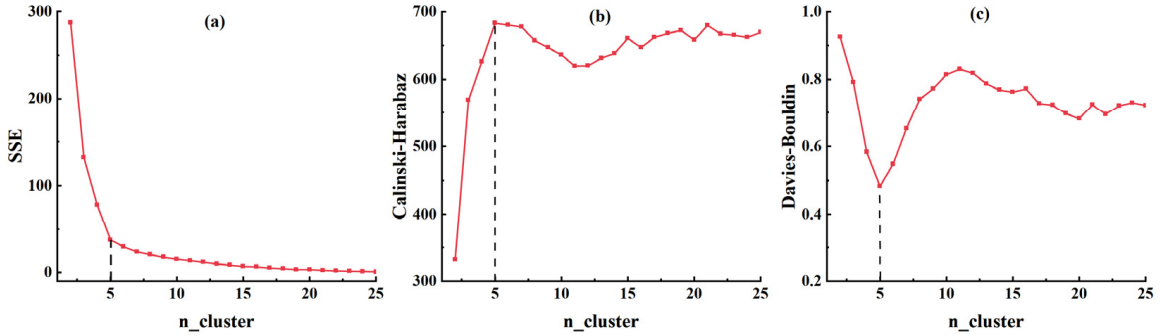

Figure 8. Determining the optimal K-value based on various metrics. (**a**) SSE; (**b**) Calinski–Harabaz; (**c**) Davies–Bouldin.

Based on Matplotlib, the clustering results were plotted as shown in Figure 9. In the flood dataset, the 359 extracted metadata information items were divided into five clusters. With a confidence level of 90% for the centroids, the majority of the information was covered. According to Figure 9, each cluster contained metadata information that categorized the flood dataset into five main themes: Cluster 1 represents contact information, cluster 2 represents location information, cluster 3 represents time information, cluster 4 represents format information, and cluster 5 represents content information.

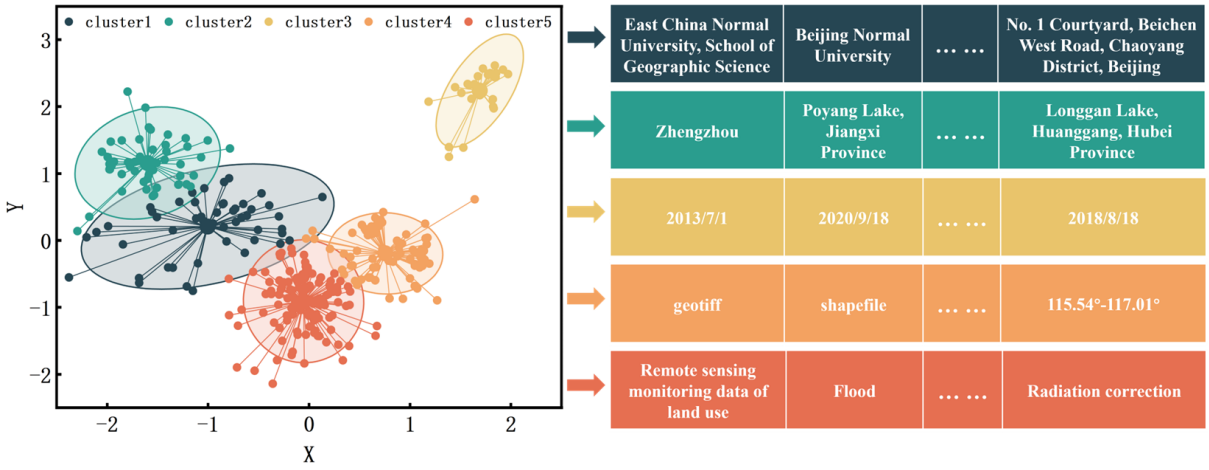

Figure 9. Word2vec–Kmeans cluster analysis results: Take flood data as an example.

Table 10 presents the variance analysis table of the clustering results. In this table, cluster sum of squares (CSS) represents the sum of distances from all sample points to their respective cluster centers, while error sum of squares (ESS) indicates the uncertainty or dispersion of the clustering results. The F-value represents the ratio of between-group variation to within-group variation, and the *p*-value was used to test the significance of the F-value. From Table 10, it can be observed that the sample points are close to their respective cluster centers within groups, indicating high similarity among samples within the clusters and good compactness of the clustering. There is minimal variation within clusters. However, between groups, the differences between clusters were considered significant relative to the differences within clusters at a significance level of 0.01. Cluster 3 shows the best separation from other clusters, while slight overlaps exist between other clusters. The proximity between cluster 1 and cluster 2 may be due to the inclusion of address information in the contact information, which overlaps with location information in the dataset. Similarly, the proximity between cluster 4 and cluster 5 may be attributed to certain format information being classified under content information, leading to overlap.

Table 10. Analysis of variance table.

	CSS	ESS	F-Value	*p*-Value
x	73.60	0.10	714.40	4.83×10^{-168}
y	48.79	0.08	603.90	1.14×10^{-156}

4. Discussion

In order to evaluate the effectiveness of the proposed self-training method for disaster metadata information extraction, multiple sets of comparative experiments were conducted. The proposed method was compared with existing metadata information extraction methods based on test set prediction accuracy. Our method used the optimal combination of hyperparameters obtained as explained in the preceding section. The experimental groups for comparison included traditional regular expressions [48], a pre-trained BERT model [49], a pre-trained UIE model (without fine-tuning), and a self-trained UIE model (fine-tuned using the optimal parameter combination obtained in this study). The extraction results of metadata stored in unstructured form in the field of natural disasters using different information extraction methods are shown in Table 11.

Table 11. The evaluation results of different metadata extraction methods.

Methods	P	R	F1
Regular expressions	0.655	0.731	0.691
BERT	0.802	0.811	0.806
UIE	0.842	0.794	0.778
Self-trained UIE	0.926	0.911	0.918

For simple text patterns, regular expressions typically have very fast matching speed, can be flexibly customized and modified, and support almost all programming languages and text processing tools. But they rely on complex syntax structures; thus, for complex text processing needs, regular expressions cannot provide enough flexibility and functionality, and also have difficulty handling tasks that require consideration of context-related information [50,51]. As shown in Table 11, the effectiveness of regex-based approaches is far inferior to current pre-trained models when extracting metadata stored in unstructured form in the field of natural disasters. The BERT model has strong capabilities in understanding context, multitask learning, and semantic representation; hence, it has been widely applied and recognized in the field of natural language processing. In our experiments, the UIE model without fine-tuning outperformed BERT. This is because the BERT model relies on a large amount of annotated text data during application, and the limited sample size restricts the semantic perception ability of the model [52,53]. In addition, the BERT model has high complexity and consumes significant computational resources, which may not be suitable for resource-constrained environments. Although the precision of UIE reached above 0.8, when applying the UIE model with fine-tuning for metadata information extraction, all indicators exceeded 0.9. These results indicate that the self-trained UIE model improved the prediction accuracy on the test set, effectively enhancing the ability of metadata information extraction and thereby improving the efficiency of disaster data management.

In information extraction tasks, regular expressions are suitable for simple text pattern matching and efficient processing of large-scale text data, especially for handling fixed formats or highly regular information extraction tasks, such as extracting specific format information from structured text. On the other hand, BERT is suitable for complex information extraction tasks that require consideration of context relationships and semantic understanding, and it is particularly adept at handling unstructured text or situations requiring deep semantic understanding, such as extracting semantic relationships or entity relationships from natural language text. Furthermore, metadata information from disaster data comes from diverse sources, with multiple dimensions and diverse formats and structures. Given these data characteristics, self-trained UIE models have better extrac-

tion results compared to traditional regular expressions. In contrast to the BERT model, UIE models allow for faster iteration with limited annotated data, accelerating the model optimization process and thereby achieving greater benefits at lower costs.

5. Conclusions

Based on the design of the natural disaster metadata model architecture, this study investigates the extraction of massive, multidimensional, heterogeneous natural disaster metadata. The conclusions are as follows:

(1) The UIE and Python parsing libraries were utilized to extract disaster metadata information stored in structured and unstructured forms automatically. The experimental results show that the extraction performance of UIE for five types of natural disasters (floods, earthquakes, landslides, mudslides, and avalanches) all improved by more than 50% when the learning rate was set to 0.0001 and the batch size to 32, which achieved optimal extraction results for disaster metadata information.

(2) Under the three dimensions of consistency, completeness, and accuracy, the metadata standards and unified disaster metadata model framework designed in this study showed good applicability in the field of natural disasters (floods, earthquakes, landslides, mudslides, and avalanches) in terms of both the data table dimension and the data item dimension. Furthermore, the completeness dimension was slightly better than consistency and accuracy.

(3) Combining the Word2vec model and K-means algorithm to cluster analyze the metadata of the flood dataset, the clusters were clustered into five main themes: contact information, location information, time information, format information, and content information. Moreover, at a confidence level of 90% for centroids, the clustering results covered most of the information. In terms of intra-group analysis, there was high similarity among samples within clusters, indicating low internal dissimilarity, which suggests a relatively concentrated distribution of text cluster data. For inter-group analysis, significant differences existed between groups compared to within groups at a significance level of 0.01, while there was slight overlap between some clusters. Overall, the clustering effect was good.

The experiment provides a foundation for the exchange, sharing, and utilization of disaster data, which helps to strengthen disaster management and response capabilities, facilitates more efficient management and utilization of various types of natural disaster data, and improves the level of prevention and mitigation of natural disasters in society.

The proposed disaster metadata standard is only a small-scale application case, and more tests are needed to determine its reasonableness and generalizability. The weights of the words can better reflect the semantic information of the text and thus improve the quality of clustering, but this was not considered in this study. Future research can verify the generalizability of the proposed disaster metadata standard and the validity of the model extraction results through more extensive experiments, as well as improve the accuracy of clustering by introducing methods such as weighting considerations in the text clustering process.

Author Contributions: Conceptualization, Zongmin Wang, Xujie Shi and Haibo Yang; methodology, Zongmin Wang, Xujie Shi and Haibo Yang; validation, Xujie Shi; investigation, Zongmin Wang, Xujie Shi, Haibo Yang and Bo Yu; writing—original draft preparation, Zongmin Wang, Xujie Shi; writing—review & editing, Zongmin Wang, Xujie Shi, Haibo Yang, Bo Yu and Yingchun Cai; supervision, Zongmin Wang and Haibo Yang; project administration, Haibo Yang; funding acquisition, Haibo Yang. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Key Research and Development Program of China (grant number 2022YFC3004402), the Henan provincial key research and development program (221111321100).

Data Availability Statement: The authors do not have permission to share data.

Acknowledgments: We are grateful to the editors and anonymous reviewers for their thoughtful comments.

Conflicts of Interest: Dr. Bo Yu is from company. All authors have declared that there is no conflict of interest.

References

1. Shi, K.; Peng, X.; Lu, H.; Zhu, Y.; Niu, Z. Application of Social Sensors in Natural Disasters Emergency Management: A Review. *IEEE Trans. Comput. Soc. Syst.* **2023**, *10*, 3143–3158. [CrossRef]
2. Ji, S.H.; Satish, N.; Li, S.; Dubey, P.K. Parallelizing Word2Vec in Shared and Distributed Memory. *IEEE Trans. Parallel Distrib. Syst.* **2019**, *30*, 2090–2100. [CrossRef]
3. Liao, Y.; Li, B.; Lv, X.; Cheng, C. Method of Multi-type Disaster Data Organization and Management Based on GeoSOT. *Geogr. Geo-Inf. Sci.* **2013**, *29*, 36–40.
4. Jony, R.I.; Woodley, A.; Perrin, D. Flood Detection in Social Media Images using Visual Features and Metadata. In Proceedings of the 2019 Digital Image Computing: Techniques and Applications (DICTA), Perth, WA, Australia, 2–4 December 2019. [CrossRef]
5. Tian, Y.; Li, W. GeoAI for Knowledge Graph Construction: Identifying Causality Between Cascading Events to Support Environmental Resilience Research arXiv. *arXiv* **2022**, arXiv:2211.06011.
6. Molina, D.E.; Datcu, M. Data mining and knowledge discovery tools for exploiting big earth observation data. In Proceedings of the 36th International Symposium on Remote Sensing of the Environment (ISRSE), Berlin, Germany, 11–15 May 2015; pp. 627–633.
7. Eichler, R.; Giebler, C.; Gröger, C.; Schwarz, H.; Mitschang, B. Modeling metadata in data lakes-A generic model. *Data Knowl. Eng.* **2021**, *136*, 101931. [CrossRef]
8. Wang, S.; Wang, J.; Zhan, Q.; Zhang, L.C.; Yao, X.C.; Li, G.Q. A unified representation method for interdisciplinary spatial earth data. *Big Earth Data* **2023**, *7*, 136–155. [CrossRef]
9. Chen, Z.G.; Yang, Y.P. Semantic relatedness algorithm for keyword sets of geographic metadata. *Cartogr. Geogr. Inf. Sci.* **2020**, *47*, 125–140. [CrossRef]
10. Ke, C.; Jiahong, W.; Lizhong, Y. Design and construction of natural disaster metadata standards. *Geomat. Spat. Inf. Technol.* **2013**, *36*, 4–8.
11. Babaie, H.A.; Babaei, A. Developing the earthquake markup language and database with UML and XML schema. *Comput. Geosci.* **2005**, *31*, 1175–1200. [CrossRef]
12. Yu, E.; Acharya, P.; Jaramillo, J.; Kientz, S.; Thomas, V.; Hauksson, E. The Station Information System (SIS): A Centralized Repository for Populating, Managing, and Distributing Metadata of the Advanced National Seismic System Stations. *Seismol. Res. Lett.* **2018**, *89*, 47–55. [CrossRef]
13. Hong, J.H.; Shi, Y.T. Integration of Heterogeneous Sensor Systems for Disaster Responses in Smart Cities: Flooding as an Example. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 279. [CrossRef]
14. Xiang, Z.R.; Demir, I. Flood Markup Language-A standards-based exchange language for flood risk communication. *Environ. Modell. Softw.* **2022**, *152*, 105397. [CrossRef]
15. Di, L.P.; Shao, Y.Z.; Kang, L.J. Implementation of Geospatial Data Provenance in a Web Service Workflow Environment with ISO 19115 and ISO 19115-2 Lineage Model. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 5082–5089. [CrossRef]
16. Goncharov, M.V.; Kolosov, K.A. The principles of extended metadata formation in RNPLS&T's Single Open Information Archive. *Nauchnye Tek. Bibl.* **2023**, *1*, 84–98. [CrossRef]
17. Wu, Y.; Liu, F.G.; Zheng, L.L.; Wu, X.J.; Lai, C.Q. CSR-SVM: Compositional semantic representation for intelligent identification of engineering change documents based on SVM. *Adv. Eng. Inform.* **2023**, *57*, 15. [CrossRef]
18. Al-Fuqaha'a, S.; Al-Madi, N.; Hammo, B. A robust classification approach to enhance clinic identification from Arabic health text. *Neural Comput. Appl.* **2024**, *36*, 7161–7185. [CrossRef]
19. Yan, D.C.; Li, G.Q.; Li, X.Q.; Zhang, H.; Lei, H.; Lu, K.X.; Cheng, M.H.; Zhu, F.X. An Improved Faster R-CNN Method to Detect Tailings Ponds from High-Resolution Remote Sensing Images. *Remote Sens.* **2021**, *13*, 18. [CrossRef]
20. Luo, J.; Du, J.; Nie, B.; Xiong, W.; Liu, L.; He, J. TCM text relationship extraction model based on bidirectional LSTM and GBDT. *Appl. Res. Comput.* **2019**, *36*, 3744–3747.
21. Islam, M.S.; Kabir, M.N.; Ab Ghani, N.; Zamli, K.Z.; Zulkifli, N.S.A.; Rahman, M.M.; Moni, M.A. Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach. *Artif. Intell. Rev.* **2024**, *57*, 79. [CrossRef]
22. Skondras, P.; Zotos, N.; Lagios, D.; Zervas, P.; Giotopoulos, K.C.; Tzimas, G. Deep Learning Approaches for Big Data-Driven Metadata Extraction in Online Job Postings. *Information* **2023**, *14*, 19. [CrossRef]
23. Qiao, B.; Zou, Z.Y.; Huang, Y.; Fang, K.; Zhu, X.H.; Chen, Y.M. A joint model for entity and relation extraction based on BERT. *Neural Comput. Appl.* **2022**, *34*, 3471–3481. [CrossRef]
24. Lu, Y.J.; Liu, Q.; Dai, D.; Xiao, X.Y.; Lin, H.Y.; Han, X.P.; Sun, L.; Wu, H. Unified Structure Generation for Universal Information Extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Acl 2022), Dublin, Ireland, 22–27 May 2022; Volume 1, pp. 5755–5772.

25. Jie, Z.; Suwen, L.; Junhui, L.; Lifan, G.; Haifeng, Z.; Feng, C. Interpretable Sentiment Analysis Based on UIE. *J. Chin. Inf. Process.* **2023**, *37*, 151–157.
26. ChinaGE-OSS Data Portal. Available online: https://www.chinageoss.cn/datasharing (accessed on 4 January 2024).
27. *GBT 24888-2010*; Technical Requirements of Data Share Foremergency Command in Earthquake Occurrence Site. Standard Press of China: Beijing, China, 2010.
28. Dublin Core. Dublin Core™ Metadata Element Set, Version 1.1. Available online: https://www.dublincore.org/specifications/dublin-core/dces/ (accessed on 4 January 2024).
29. *DB/T 41-2011*; Earthquake Data Metadata. China Earthquake Administration: Beijing, China, 2011.
30. *ISO19115*; Geographic Information—Metadata. ISO: Geneva, Switzerland, 2014.
31. *GB/T 19710-2005*; Geographic information—Metadata. Standard Press of China: Beijing, China, 2005.
32. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
33. Breuel, T.M. The Effects of Hyperparameters on SGD Training of Neural Networks. *arXiv* **2015**, arXiv:1508.02788.
34. Wang, R.Y.; Strong, D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]
35. Reiche, K.J.; Höfig, E. Implementation of Metadata Quality Metrics and Application on Public Government Data. In Proceedings of the IEEE 37th Annual Computer Software and Applications Conference (COMPSAC), Kyoto, Japan, 22–26 July 2013; pp. 236–241.
36. Nogueras-Iso, J.; Lacasta, J.; Ureña-Cámara, M.A.; Ariza-López, F.J. Quality of Metadata in Open Data Portals. *IEEE Access* **2021**, *9*, 60364–60382. [CrossRef]
37. Kuzma, M.; Moscicka, A. Metadata evaluation criteria in respect to archival maps description A systematic literature review. *Electron. Libr.* **2020**, *38*, 1–27. [CrossRef]
38. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2011**, arXiv:2010.16061.
39. Rong, X. word2vec Parameter Learning Explained. *arXiv* **2014**, arXiv:1411.2738.
40. Ma, L.; Zhang, Y.Q. Using Word2Vec to Process Big Text Data. In Proceedings of the IEEE International Conference on Big Data, Santa Clara, CA, USA, 29 October–1 November 2015; pp. 2895–2897.
41. Fesseha, A.; Xiong, S.W.; Emiru, E.D.; Diallo, M.; Dahou, A. Text Classification Based on Convolutional Neural Networks and Word Embedding for Low-Resource Languages: Tigrinya. *Information* **2021**, *12*, 17. [CrossRef]
42. Dimitriadis, G.; Neto, J.P.; Kampff, A.R. t-SNE Visualization of Large-Scale Neural Recordings. *Neural Comput.* **2018**, *30*, 1750–1774. [CrossRef] [PubMed]
43. Atzberger, D.; Cech, T.; Trapp, M.; Richter, R.; Scheibel, W.; Dollner, J.; Schreck, T. Large-Scale Evaluation of Topic Models and Dimensionality Reduction Methods for 2D Text Spatialization. *IEEE Trans. Vis. Comput. Graph.* **2024**, *30*, 902–912. [CrossRef] [PubMed]
44. Hu, C.X.; Wu, T.; Liu, S.Q.; Liu, C.S.; Ma, T.; Yang, F. Joint unsupervised contrastive learning and robust GMM for text clustering. *Inf. Process. Manag.* **2024**, *61*, 17. [CrossRef]
45. Xu, Q.; Gu, H.; Ji, S.W. Text clustering based on pre-trained models and autoencoders. *Front. Comput. Neurosci.* **2024**, *17*, 13. [CrossRef] [PubMed]
46. González, F.; Torres-Ruiz, M.; Rivera-Torruco, G.; Chonona-Hernández, L.; Quintero, R. A Natural-Language-Processing-Based Method for the Clustering and Analysis of Movie Reviews and Classification by Genre. *Mathematics* **2023**, *11*, 26. [CrossRef]
47. Liu, X.D.; Tian, Y.Z.; Zhang, X.Q.; Wan, Z.Y. Identification of Urban Functional Regions in Chengdu Based on Taxi Trajectory Time Series Data. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 19. [CrossRef]
48. Cao, Q.; Wang, S.; Chen, Z.; Li, G.; Li, J. The Method of Extracting Names of Geo-science Data based on Regular Expressions. *J. Geo-Inf. Sci.* **2023**, *25*, 1601–1610.
49. Evans, M.T.C.; Latifi, M.; Ahsan, M.; Haider, J. Leveraging Semantic Text Analysis to Improve the Performance of Transformer-Based Relation Extraction. *Information* **2024**, *15*, 91. [CrossRef]
50. Bartoli, A.; De Lorenzo, A.; Medvet, E.; Tarlao, F. Inference of Regular Expressions for Text Extraction from Examples. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 1217–1230. [CrossRef]
51. Fagin, R.; Kimelfeld, B.; Reiss, F.; Vansummeren, S. Document Spanners: A Formal Approach to Information Extraction. *J. ACM* **2015**, *62*, 51. [CrossRef]
52. Gong, Y.; Mao, L.; Li, C.L. Few-shot Learning for Named Entity Recognition Based on BERT and Two-level Model Fusion. *Data Intell.* **2021**, *3*, 568–577. [CrossRef]
53. Bello, A.; Ng, S.C.; Leung, M.F. A BERT Framework to Sentiment Analysis of Tweets. *Sensors* **2023**, *23*, 506. [CrossRef] [PubMed]