applied sciences

MDPI

Article

Automatic Extraction of Flooding Control Knowledge from Rich Literature Texts Using Deep Learning

Min Zhang [1,2] and Juanle Wang [1,3,4,*]

[1] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[2] College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
[3] China-Pakistan Joint Research Centre on Earth Sciences, Islamabad 45320, Pakistan
[4] Jiangsu Centre for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
* Correspondence: wangjl@igsnrr.ac.cn; Tel.: +86-010-6488-8016

Abstract: Flood control is a global problem; increasing number of flooding disasters occur annually induced by global climate change and extreme weather events. Flood studies are important knowledge sources for flood risk reduction and have been recorded in the academic literature. The main objective of this paper was to acquire flood control knowledge from long-tail data of the literature by using deep learning techniques. Screening was conducted to obtain 4742 flood-related academic documents from past two decades. Machine learning was conducted to parse the documents, and 347 sample data points from different years were collected for sentence segmentation (approximately 61,000 sentences) and manual annotation. Traditional machine learning (NB, LR, SVM, and RF) and artificial neural network-based deep learning algorithms (Bert, Bert-CNN, Bert-RNN, and ERNIE) were implemented for model training, and complete sentence-level knowledge extraction was conducted in batches. The results revealed that artificial neural network-based deep learning methods exhibit better performance than traditional machine learning methods in terms of accuracy, but their training time is much longer. Based on comprehensive feature extraction capability and computational efficiency, the performances of deep learning methods were ranked as: ERNIE > Bert-CNN > Bert > Bert-RNN. When using Bert as the benchmark model, several deformation models showed applicable characteristics. Bert, Bert-CNN, and Bert-RNN were good at acquiring global features, local features, and processing variable-length inputs, respectively. ERNIE showed improved masking mechanism and corpus and therefore exhibited better performance. Finally, 124,196 usage method and 8935 quotation method sentences were obtained in batches. The proportions of method sentence in the literature showed increasing trends over the last 20 years. Thus, as literature with more method sentences accumulates, this study lays a foundation for knowledge extraction in the future.

Keywords: flood control; knowledge extraction; text mining; deep learning; long tail data

Citation: Zhang, M.; Wang, J. Automatic Extraction of Flooding Control Knowledge from Rich Literature Texts Using Deep Learning. *Appl. Sci.* **2023**, *13*, 2115. https://doi.org/10.3390/app13042115

Academic Editor: Xianpeng Wang

Received: 22 December 2022
Revised: 3 February 2023
Accepted: 3 February 2023
Published: 7 February 2023

1. Introduction

With global warming, the frequency and intensity of hydrometeorological disasters have increased, and harm to human society and the ecological environment is becoming increasingly serious. Existing flood disaster research has accumulated rich knowledge resources that focus on risk assessment, early warning, flood frequency, disaster management, impact factor analysis, and other related topics [1]. Data resources for flood disaster risk reduction mainly include monitoring and reporting, real-time remote sensing, historical disasters, socio-economics, and basic geography [2]. Academic literature, as a kind of disaster risk and reduction data containing rich knowledge, should also be effectively utilized. Academic literature is a type of high-quality text data in a standardized format that contains peer-reviewed knowledge, condenses the achievements of scientific researchers,

and is a professional resource with high research value [3]. It can be effectively utilized as a type of disaster risk and reduction data source containing rich knowledge. This knowledge in the manuscripts generally contain research methods, data sources, temporal and spatial attributes of disaster events, disaster losses, and other disaster information of importance. Since literature studies can be repeated and scaled up by other scholars anywhere in the world, they are important long-tail data resources for disaster research. Therefore, the extraction of knowledge related to flood control from these documents is important and meaningful.

However, there is a certain challenge in knowledge extraction based on massive literature. In the digital era, the number of academic papers published has increased dramatically. The number of scientific and technological paper produced worldwide has reached one million and continues to increase at a rate of approximately 3% annually [4]. It is impossible to manually read all of the papers in a certain field to grasp their research status. Academic literature is an example of a classic, unstructured, long-text data. It is time-consuming, laborious, and would ultimately be an incomplete endeavor to master flood control knowledge in academic literature through manual reading. Hence, it is necessary to use intelligent methods for knowledge extraction from a large number of literature documents.

In recent years, artificial intelligence methods based on big data for text mining have been favored by many scholars. The literature is composed of sentences. Therefore, it is appropriate to conduct knowledge extraction research and semantic classification at the sentence level. The current sentence classification methods are divided into three main categories: rule-based methods [5–7], machine learning, and artificial neural network-based deep learning methods. Machine learning algorithms build a model based on sample data, known as training data, to make predictions or decisions without explicit programming to do so. Deep learning is a class of machine learning algorithms based on artificial neural networks that have multiple intermediate layers. Its basic principle is to extract low-level features first, and then arrange and combine low-level features at a higher level to find useful information after combination. The research progress, advantages, and disadvantages of these methods are compared in Table 1.

Table 1. Research status of text classification methods.

Method Type	Research Methods	References	Methods Evaluation
Rule-based methods	Heuristic rules TCL rule base Rules for the detection and classification of emotion signals Sentiment composition rules	Hua et al. [8] Hayes et al. [9] Asghar et al. [10] Tan et al. [11]	The rule-based method has a high accuracy rate, but a low recall rate, and needs to formulate matching rules in advance.
Machine learning methods	Naïve Bayes, complement Naïve Bayes and Decision Tree logistic regression model SVM and CRF CRF SVMs and NB	Widyantoro et al. [12] Zhang et al. [1] Liakata et al. [13] Hirohata et al. [14] Shirsat et al. [15]	Machine learning methods require more time for feature extraction, and the quality of the feature selection also affects model performance.
Artificial neural network-based deep learning methods	LSTM and GRU BERT model BERT-BiLSTM BERT, CNN, LSTM BERT-CNN model, char-CNN, fast Text, DRNN and Bert LR, NB, KNN, SVM, CNN, RNN, FastText, BERT and ERNIE. Word2vec, ELMo and Bert CNN, BiLSTM, RCNN, BiGRU, BERT, ALBERT Bert, XLNet, RoBERTa, and ERNIE	Chung et al. [16] Li et al. [20] Wang et al. [21] Jindal et al. [22] Zheng et al. [23] Zong et al. [24] Xu et al. [25] Xia [26] Wang et al. [27]	Artificial neural network-based deep learning methods have addressed the feature selection problem [17–19]. It can automatically obtain feature expression, replace the artificial feature engineering in machine learning methods, and it also has better performance.

In general, rule-based methods solve problems intuitively, but it takes more time to design rules, with limited coverage, which is not suitable for massive amounts of literature. Machine learning methods incur high costs and considerable time for feature engineering. It also ignores contextual information, making it difficult to learn the semantic information

of vocabulary. Artificial neural network-based deep learning methods avoid the manual design of rules and features and automatically provide the semantic meaning representation of text mining, becoming the current mainstream method. However, the applicability of different deep learning methods remains unclear.

Based on previous research, the purpose of this study was to mine valuable knowledge about flood control from long-tail data found in the academic literature. We used machine learning tool to parse academic literature documents. Various machine learning and artificial neural network-based deep learning methods have been used for flood control knowledge extraction at the sentence level. The advantages, disadvantages, and applicability of various methods were compared from the perspectives of performance and principles. This study enriches literature-based knowledge extraction research at the sentence level.

2. Materials and Methods

2.1. Data Acquisition

Relying on the International Knowledge Center for Engineering Sciences and Technology, the Disaster Risk Reduction Knowledge Service System (DRRKS) is built based on knowledge services to realize the resource connectivity of disaster data, information, and knowledge for knowledge discovery [28,29]. Based on the Scholar module of DRRKS, more than 14,000 articles related to flood disasters over the last 20 years were retrieved for this study. Through thematic classification of abstracts, 5181 literatures texts highly related to flood disasters were screened [1]. The literature data of DRRKS was sourced from the Web of Science database, a comprehensive academic platform. Then, using AceMap (http://acemap.sjtu.edu.cn/, accessed on 1 February 2022), 4840 portable document format (PDF) files were obtained as research data. AceMap is a visual search system oriented toward big academic data that were independently developed by Shanghai Jiaotong University. It utilizes the Ace KG academic knowledge graph as the data structure and a diversified academic atlas as the expression method [30]. It focuses on mining information hidden in academic data and provides intuitive, convenient, and accurate information to researchers and users [31].

2.2. Research Methods

A detailed technical roadmap of this study is shown in Figure 1. Based on literature screening, PDF files of the literature were obtained from AceMap. These documents were parsed with the help of GROBID (https://grobid.readthedocs.io/en/latest/Introduction/, accessed on 1 February 2022), a machine learning tool, to obtain text data in XML format. Manual labeling was performed according to the designed sentence annotation rules to form the corpus for the model training in this study. Next, machine learning models, such as I Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM), and deep learning models, such as Bert, Bert-CNN, Bert-RNN, and Enhanced Representation through Knowledge Integration (ERNIE), were used to train the sentence classification models. The model performance was compared in terms of accuracy, recall, F1, and running time. Finally, a model suitable for this study was selected to extract the flood control knowledge.

2.2.1. PDF Parsing and Text Preprocessing

PDF is a file format for presenting documents that are independent of applications, hardware, and operating systems. The original intention of creating a PDF was to preserve the document layout and content for a consistent presentation on any platform, making it difficult to re-edit. The text data in PDF are stored in the form of characters that record specific location information on the page, rather than in the form of paragraphs or words, so the text data stored in PDF loses semantic structure information [32]. At present, the structural information of academic papers in PDF format must be manually extracted, and batch extraction is difficult. In the process of automated text extraction, there are problems

such as reading protection, embedded images, excessive or insufficient spaces, sorting order, etc. [33]; therefore, PDF-based text extraction is extremely difficult.

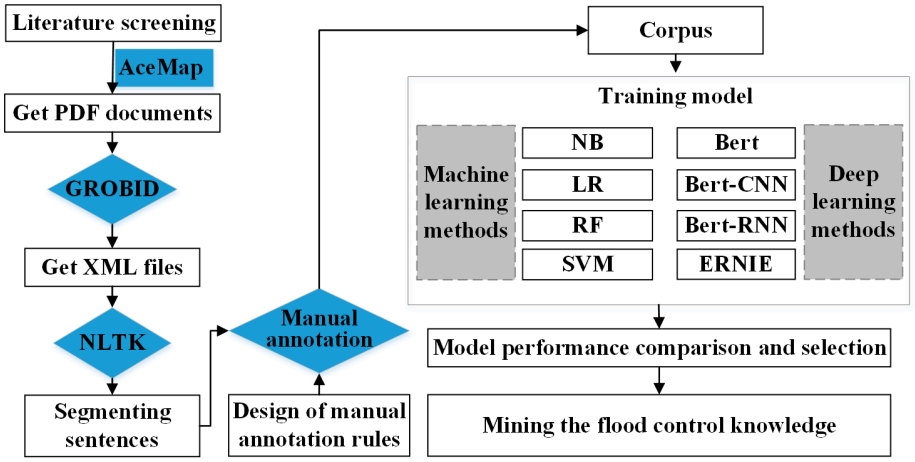

Figure 1. Technical roadmap.

In this study, GROBID was used to extract, parse, and reorganize PDF documents into TEI-encoded XML documents [34]. First, pdf2xml was used to preprocess the PDF document and generate XML files containing text and format information. The wapiti tool and GROBID were then used to process and assemble the XML files in TEL format. After removing the encrypted or unopenable PDF files, 4742 XML files were parsed from the PDF files. Glob is a module used in python for matching and searching. With the help of glob, the text content was parsed from the XML files and stored in the SQLite database (see Figure 2).

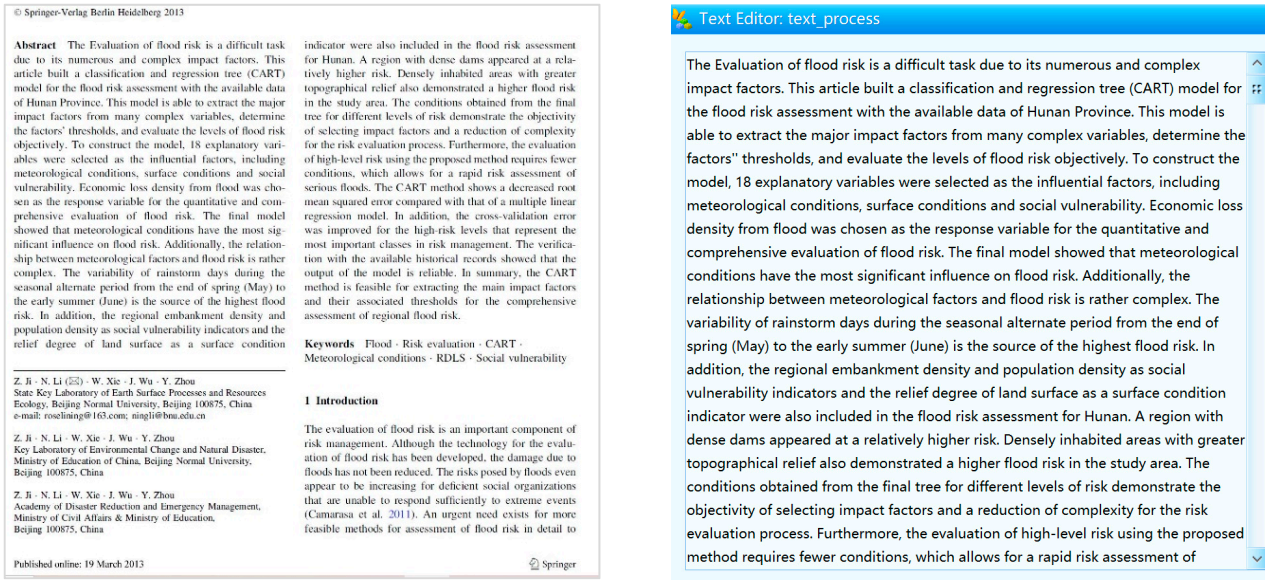

Figure 2. PDF document before and after parsing (**left**: PDF document; **right**: parsed text).

The text content of the academic papers extracted by automated batch-parsing of PDFs contained format errors such as garbled characters and excessive spaces. The text data were cleaned using functions of the database management system such as adding, deleting, modifying, querying, and replacing. The author information, references, and other content were excluded, and only the abstract and full text were retained. The Natural Language Toolkit (NLTK, www.nltk.org, accessed on 1 February 2022) was used for text

segmentation. NLTK is a commonly used toolkit for processing corpora, text classification, language structure analysis, and other tasks.

2.2.2. Manual Annotation Rules for Sentence Classification

Based on the sentence segmentation results, 347 papers of different years from a total of 4742 documents were selected as the sample data for manual annotation. Sentence annotation rules were designed, and 17 graduate students were recruited for data labeling training to perform the labeling work. Referring to relevant research [35,36], the sentences were divided into three types: usage method sentences, quotation method sentences, and non-research method sentence.

The usage method sentence refers to the sentences in which the author clearly points out the algorithm, model, software, tools, data, indicators, etc. proposed or used in their research but does not include the definition and background introduction of the method. This usually appears in the Materials and Methods section. The designed labeling rules of the usage method sentences were as follows: (1) Sentences contain the name of the research method employed in the target text. (2) Sentences contain the research method used in the target text and the problem to be solved by it. (3) Sentences express the execution step; if there is a method entity, it should be labeled as a usage method sentence. An example of usage method sentence is: "This study employed geospatial techniques to create flood inundation maps considering flood hazard indicators calculated by the AHP model." [37]. The quotation method sentences refer to sentences in which the author mentions or quotes the work of others, and these sentences describe the methods proposed and used by others. This usually appears in the Introduction or Discussion sections. The labeling rules for quotation method sentences were set as follows: (1) Sentences containing the names of the research methods created by other scholars in their work. (2) Sentences contain the research methods used by others, and the problems to be solved in their work. Here is an example for a quotation method sentence: "Tung (1985) utilized the Hook-Jeeves Pattern Search (HJ) method in combination with three linear regression (LR), Davidon Fletcher-Powell (PFP), and Conjugate Gradient (CG) methods." [38]. Non-research method sentences do not contain a research method entity, and they encompass all of the sentences in the abstract and text.

Statistically, 347 academic papers, comprising approximately 61,000 sentences, were manually labeled in this study; more than 90% of which were non-research method sentences. To avoid uneven data distribution affecting the training of the model, some labeled data were selected to train the sentence classification model. The number of non-research method, usage method, and quotation method sentences in the training, verification, and test sets used in this study are listed in Table 2.

Table 2. Statistical information from the research method dataset.

Dataset	Total Number of Sentences	Number of Non-Research Method Sentences	Number of Usage Method Sentences	Number of Quotation Method Sentences
Training set	4922	3123	1318	481
Validation set	593	383	161	49
Test set	684	473	160	51

2.2.3. Sentence Classification Method

The sentence classification method used in this study included machine learning methods (such as NB, LR, SVM, and RF) and artificial neural network-based deep learning methods (e.g., Bert, Bert-CNN, Bert-RNN, and ERNIE).

(1) Machine learning methods

NB: The Sklearn Toolkit was used for NB sentence classification. First, the training corpus was preprocessed by punctuation removal and word segmentation, and Term

Frequency–Inverse Document Frequency (TF-IDF) was used for feature selection. Subsequently, a spatial vector was constructed to train the sentence classification model. Finally, the batch prediction results were obtained.

LR: LR sentence classification was implemented using the Sklearn Toolkit. Data preprocessing and the feature selection process were performed the same way as in the NB model. The LR classification model was trained and used for testing and prediction.

SVM: Data preprocessing and feature selection were performed the same way as in the NB model. The training corpus was converted into a space vector model to train the SVM classification model. The trained model was used for testing and batch prediction.

RF: After data preprocessing, the training corpus was converted into a space vector to train the RF sentence classification model. The testing and prediction were performed using a trained model.

(2) Artificial neural network-based deep learning methods

Bert: This is a pre-training language model based on the transformer [39] proposed by Google. It uses the Masked Language Model (MLM) and Next Sentence Prediction (NSP) as model training methods to enhance the ability to capture the characteristics of word level, sentence level, and inter-sentence relationships. The model architecture of Bert is a multilayer bidirectional transformer encoder, in which multiple transformer encoders are stacked. Context semantics were obtained through self-attentions [40]:

$$Out = Transformer(Embedding(Text)) \tag{1}$$

where $Text$ is the input text, $Embedding$ represents the input layer of Bert, $Transformer$ represents the Bert feature extractor, and Out is the text vector.

To fully extract semantic information, the transformer uses attention as the basic unit, and its calculation formula is [39]:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where x is the input, $Q = x \cdot W_Q$, $K = x \cdot W_K$, $V = x \cdot W_V$, $W_i(Q, K, V)$ are random initialization parameters conforming to the normal distribution, and d_k is the data dimension.

The multi-head attention mechanism was used in the transformer; that is, Formula (2) was repeated multiple times, and the results were spliced as Formula (3):

$$Multi\text{-}Head\ Attention(Q, K, V) = concat(Attention(Q_i, K_i, V_i)),\ i = 1, 2, \ldots, t \tag{3}$$

where t was the number of multi-heads.

To speed up the training of the model, normalization was introduced to scale the parameters to a normal distribution. The parameters were calculated using Equation (4):

$$q_{i,j,k} = \frac{x_{i,j,k} - u_{i,j}}{\sqrt{s_{i,j}^2 + \epsilon}}\gamma_k + \beta_k,\ u_{i,j} = \frac{1}{d}\sum_{k=1}^{d} x_{i,j,k},\ s_{i,j}^2 = \frac{1}{d}\sum_{k=1}^{d}(x_{i,j,k} - u_{i,j})^2 \tag{4}$$

where γ_k and β_k were learnable parameters.

To further integrate semantic information, the transformer used a Feedforward Neural Network, whose calculation formula is:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{5}$$

where W_1 and W_2 are learnable random initialization parameters of the normal distribution, and b_1 and b_2 are the bias.

To learn the bidirectional information of context in sentences, the MLM task was conducted to mask some tokens at random and predict the masked tokens according to the unmasked tokens. Specifically, we randomly masked 15% of the words in the sentences

and replaced the masked words using different methods. Of these, 80% were replaced with [MASK], 10% were replaced with random words, 10% were kept unchanged, and the masked words were then predicted [41]. The NSP task predicts whether two sentences in the input are a pair. The model training in this study did not involve NSP.

The data processing flow of BERT for a single-sentence classifier task is shown in Figure 3. Before feeding the word sequences into BERT, 15% of the words in each sequence were replaced with [MASK]. Then, the sum of the token embedding, segment embedding, and position embedding was used as the input embedding of Bert. The Bert-encoder is composed of superimposed multiple transformer encoders. The internal structure of the transformer encoder is illustrated in Figure 4. Finally, the Bert output layer was fully connected to a linear classifier to obtain the classification result. The calculation formula for the classification is as follows:

$$Label = Softmax(Bert(Text) \times W + b),\ Label \in \{+1,\ 0, -1\} \tag{6}$$

where $Text$ is the input text, W and b represent the neural network parameters, and $Label$ is the sentence classification label.

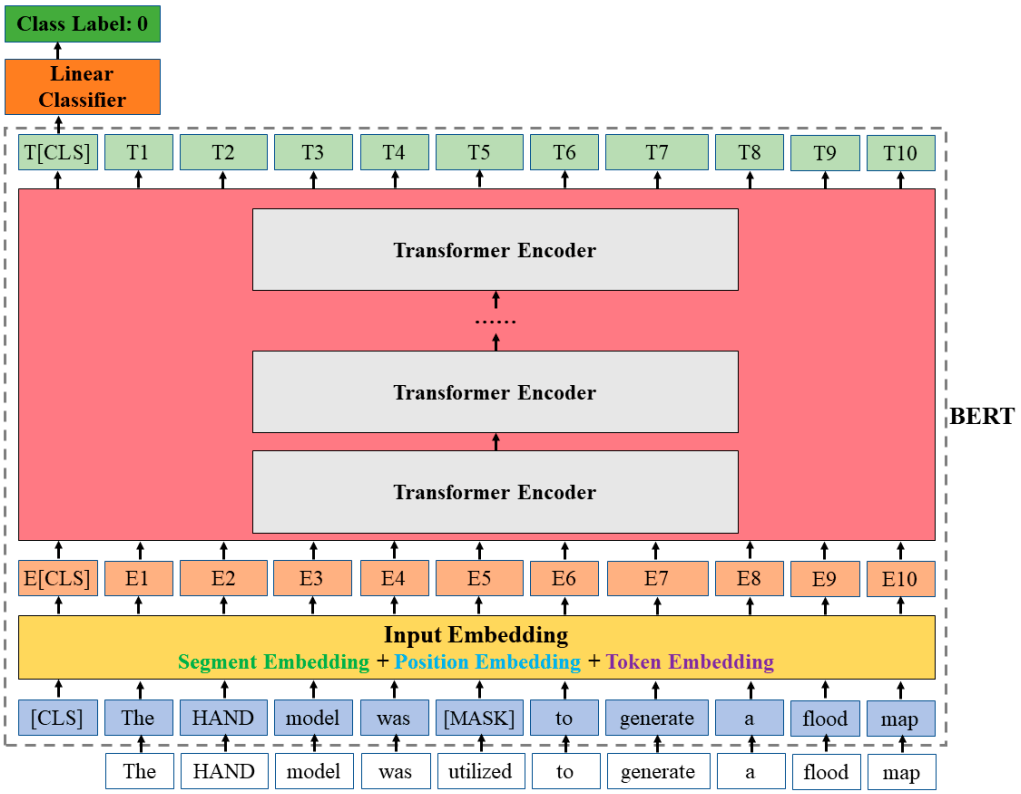

Figure 3. Illustrations of BERT on a single sentence classifier task.

The main parameters were set as follows: batch size = 32, epochs = 20, learning rate = 5×10^{-5}, and hidden size = 768.

Bert-CNN: The Bert model was used to characterize the corpus, and then, the first token vector in all encoder layers of the Bert model was selected for splicing as the input of the CNN model. Finally, convolution kernels of different sizes were used to extract the local semantic features and connect the full connection layer for classification. The main parameters were set as follows: batch size = 32; epochs = 20; learning rate = 5×10^{-5}; hidden size = 768; convolution kernel filter sizes = (2, 3, and 4); and dropout = 0.3.

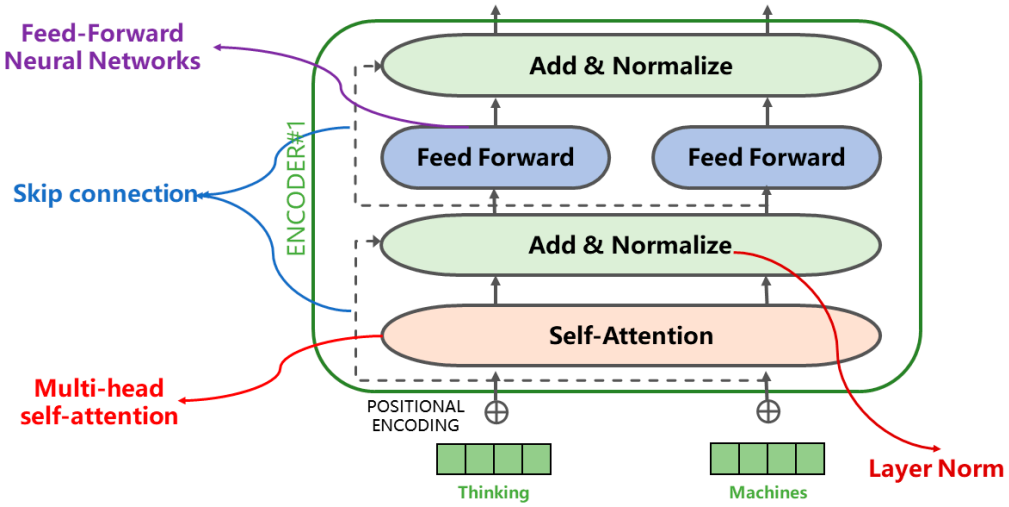

Figure 4. Transformer encoder.

ERNIE: The Bert model mainly focuses on learning and modeling at the word level. ERNIE is an optimization of Bert. The unified modeling of lexical and grammatical structures and semantic information in the training data greatly enhanced the ability of general semantic representation, and greatly surpassed Bert in many tasks. The main parameter settings in this study were batch size = 32, epochs = 20, learning rate = 5×10^{-5}, and hidden size = 768.

3. Results

3.1. Comparative Analysis of Sentences Classification Models

In this study, we attempted to maintain consistency in the process of setting the model parameters. When this was not feasible owing to the difference in the model structure, the optimal parameter setting was adopted. The evaluation criteria included the accuracy rate (Acc), F1, F1 macro-average, F1 weighted average, loss rate, and running time.

The classification results for each method sentence are listed in Table 3. The accuracy distribution range of the machine learning models was [71.94, 76.05], and the range for the artificial neural network-based deep learning models was [73.25, 81.29]. Thus, the accuracy of the machine learning models was lower than that of the deep learning models. However, the training of the machine learning model was completed instantaneously, i.e., in only a few seconds, whereas the training of the deep learning model required 2–6 h. This is because the neural networks in the deep learning model are complex with many different parameters. Among the machine learning models, SVM performed best with an accuracy rate of 76.05%, followed by LR, while the performances of both NB and RF were low but similar to each other. Regarding the deep learning models, Bert was the benchmark model, with an Acc of 77.34%. It was significantly higher than that of Bert-RNN, but lower than those of Bert-CNN and ERNIE. Bert-CNN is an improved model based on Bert, and its evaluation indicators were also better than those of Bert. However, the performance of Bert-RNN was not as good as that of Bert, according to Acc, and the running time was also significantly higher, approximately 5.5 h. The Acc of ERNIE was the highest, reaching 81.29%. The other indicators were also optimal, and the running time was shorter than that of other deep learning models. The indicators Macro_F1 and Weighted_F1 were consistent with Acc in the evaluation of several deep learning models. According to the values of F1@0, F1@1, and F1@2, the classification accuracy of the non-research method sentences was the best, all with a value greater than 80%, followed by the usage method sentences. This finding is directly related to the sample size. The larger the sample size, the better the classification performance.

Table 3. Comparative performance of various classification models.

Model	Acc	F1@0	F1@1	F1@2	Macro_F1	Weighted_F1	Loss	Time
NB	71.94	–	–	–	–	–	–	1 s
LR	73.82	–	–	–	–	–	–	1 s
RF	71.40	–	–	–	–	–	–	1 s
SVM	76.05	–	–	–	–	–	–	5 s
Bert	77.34	60.00	54.69	85.68	66.79	77.37	0.62	3:17:43
Bert-CNN	79.82	64.81	56.47	86.97	69.42	79.51	0.57	2:45:29
Bert-RNN	73.25	61.21	53.16	81.36	65.25	74.55	0.61	5:24:18
ERNIE	81.29	67.67	64.37	87.58	73.21	81.19	0.58	2:41:54

Acc: accuracy rate of sentence classification; F1@0: F1 value of the usage method sentences; F1@1: F1 value of the quotation method sentences; F1@2: F1 value of the non-research method sentences; Macro_F1: F1 macro average; Weighted F1: weighted average of F1; Loss: loss rate; Time: time spent on model training.

3.2. Yearwise Distribution Analysis of Research Method Sentences

Based on 4742 English literature texts on flood disasters, the distribution of research method sentences in the last 20 years was analyzed in this study. It was studied mainly by analyzing the number of usage method sentences, quotation method sentences, and total number of article sentences in different years. The research data included 3203 usage method sentences and 1415 quotation method sentences in the manually annotated dataset, as well as 124,196 usage method sentences and 8935 quotation method sentences obtained through batch extraction.

The proportion of usage method sentences refers to the ratio of the number of usage method sentences to the total number of sentences. The annual distribution is shown in Figure 5. The blue dashed line is the trend line. The distribution of the usage method sentences can be divided into four phases. In the first phase (2000–2003), the overall trend was oscillatory, with a large range of changes. This is because of the small number of documents in this phase, and the number of usage method sentences in a single document significantly affected the overall distribution. In the second phase (2003–2007), an overall upward trend was observed with an increased number of usage method sentences in the papers, indicating an increase in the number of experimental and research papers related to flood disasters. The third phase (2007–2009) exhibited a sharp downward trend and the proportion of usage method sentences decreased significantly, indicating that there were more theoretical studies in this phase. In the fourth phase (2009–2020), there was a fluctuating increasing trend, and the proportion of usage method sentences increased, indicating that experimental and research papers were the majority.

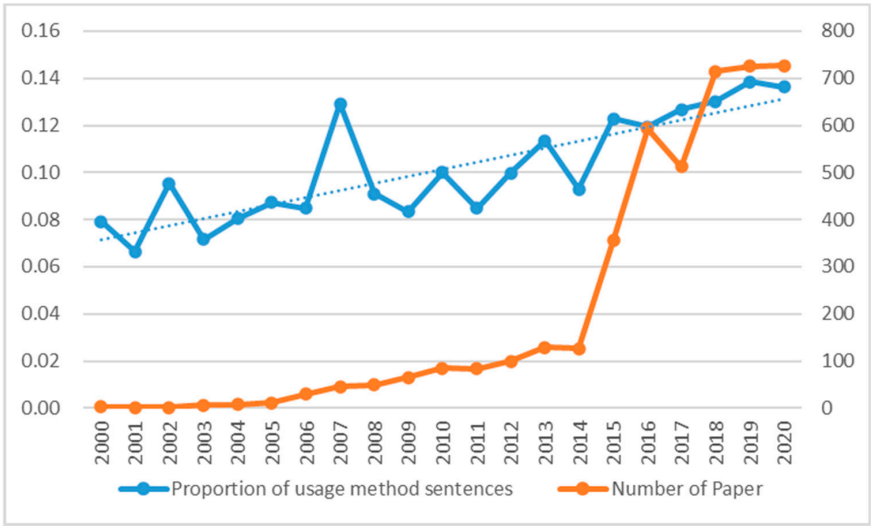

Figure 5. Yearly distribution of usage method sentences in papers.

The proportion of quotation method sentences refers to the ratio of the number of quotation method sentences to the total number of sentences in an article. The annual distribution of quotation method sentences (Figure 6) can be divided into four phases. The blue dashed line is the trend line. In the first phase (2000–2002), the proportion of quotation method sentences in the entire phase was relatively small, which is also related to the small number of articles published. In the second phase (2002–2003), the proportion of quotation method sentences increased sharply, indicating that studies in this phase tended to use the existing methods. In the third phase (2003–2005), a sharp decline in the overall trend was observed, with a decrease in the proportion of quotation method sentences. The fourth phase (2005–2020) showed an overall upward trend, and the ratio of quotation method sentences continued to increase, indicating that papers at this stage were more inclined to use existing methods. It also showed that published papers were mostly research or experimental papers in this phase, which is consistent with the trend in the proportion of usage method sentences.

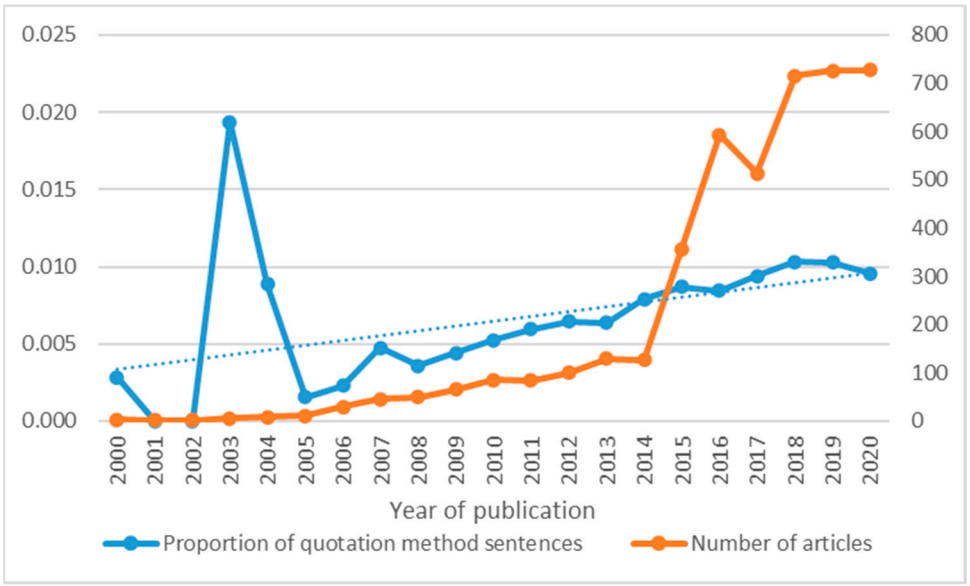

Figure 6. Yearly distribution of quotation method sentences in papers.

4. Discussion

4.1. Comparison of Machine Learning and Deep Learning Methods

As described in Table 3, SVM model performed the best among the machine learning methods. This result is consistent with that of Nomponkrang et al. [42], who used decision tree, NB, K-nearest neighbor, and SVM models to classify sentences and concluded that SVM demonstrated a better performance. He et al. [43] combined a single-layer convolutional neural network with Bert to determine user intentions based on query sentences. The results revealed that the model significantly improved the accuracy of sentence classification, which is consistent with the conclusion of this study that Bert-CNN was superior to Bert. Xu et al. [25] used the word2vec, ELMo, and Bert models to identify telephone fraud text, showing that the Bert model was better than ELMo and word2vec. Xia [26] adopted CNN, BiLSTM, RCNN, BiGRU, Bert, and ALBERT models to classify news texts and found that the Bert and ALBERT models had the best performance. Wang et al. [27] applied pretrained language models, such as Bert, XLNet, RoBERTa, and ERNIE, to study Chinese knowledge graph question answering. The results revealed that ERNIE is more suitable for Chinese question answering tasks. Luo et al. [44] constructed a literature classification model based on Bert and ERNIE and compared the results with machine learning methods (such as NB and SVM) and artificial neural network-based deep learning methods (i.e., CNN and RNN). The results demonstrate the superiority of the methods based on deep pretrained

language models. Overall, the performance of artificial neural network-based deep learning methods was found to be better than that of the traditional machine learning methods. Traditional machine learning is a shallow feature extraction method. It did not have a sufficient understanding of the semantics, structure, sequence, and context of the text, limiting the representation ability of the models.

4.2. Analyzing the Deep Learning Method from the Perspective of Principle

In this study, ERNIE exhibited the best performance based on evaluation indicators. In principle, Bert originates from the encoder part of the transformer [45] (as shown in Figure 4) and comprises several components, such as multi-head attention, skip connection, layer norm, and feed-forward network. Bert-CNN/Bert-RNN was used to replace self-attention with CNN/RNN. The main tasks of Bert were mask language model (MLM) and next sequence prediction (NSP). Bert considered a word piece to be the smallest unit of the mask that lost semantic information. ERNIE is an improved model based on Bert that mainly improves the mask mechanism. This enabled the model to learn the semantic representation of complete concepts through the mask of phrases, proper nouns, entities, and so on. Compared with the Bert model, ERNIE has added more high-quality corpora, making it superior for natural language process (NLP) tasks.

In terms of computational efficiency, the order is ERNIE > Bert-CNN > Bert > Bert-RNN. The Bert-RNN model required the longest training time and had the lowest efficiency, which is related to the dependency structure of the RNN itself. In RNN model training, the calculation at time T depended on the calculation results of the hidden layer at time $T-1$, whereas the calculation at time $T-1$ depended on the calculation results of the hidden layer at time $T-2$, etc., forming a sequence dependency. In other words, the results of the second step can be calculated only after the first step is completed. Therefore, it was difficult for the RNN to have efficient parallel computing capability. However, CNN and Bert did not depend on the input of different time steps in the intermediate state of the network. They have strong parallel computing ability because of their flexibility and freedom. Therefore, the training time of the Bert-RNN model is significantly longer.

From the perspective of comprehensive feature extraction capability, the ranking was ERNIE > Bert-CNN > Bert > Bert-RNN. The input of NLP is generally linear sequence sentences of variable length. The transformer and the CNN set the maximum input length. If the actual length of a sentence was less than the maximum length, padding was used to convert the model input to a fixed length. CNN can efficiently extract local features and semantic information from text. It can achieve excellent results in short-text classification using a pretraining model. However, the feature extraction of this model is distance dependent, and the extraction of some important features beyond a certain text length become difficult. The transformer for long texts presents huge computational complexity, resulting in a sharp slowdown in the computational speed. An RNN is a network structure that can accept variable-length input and conduct linear information transmission from front to back, which is also the fundamental reason for its popularity in the NLP field. However, RNN are limited by their parallel-computing ability.

The research object of this study was sentences in the article, which is a type of short text data. Therefore, the RNN was inferior to CNN and transformers in terms of performance and efficiency. Müller et al. [46] ranked the comprehensive capabilities of the three feature extractors in the order transformer > CNN > RNN. Li et al. [47], in their study, carried out a survey on text classification from shallow to deep learning and they conducted model training based on the same dataset (such as SST-2, IMDB, MR, etc.). Their results showed that the Bert model performed better. The research in [48] also proved that the transformer has a far stronger feature extraction capability than LSTM (an improved RNN model), which is consistent with the conclusions of this study. According to the evaluation indicators such as Acc, F1 macro average, F1 weighted average, and the running time in Table 3, it was seen that Bert-RNN was inferior to Bert in terms of computing efficiency and comprehensive extraction capacity. It showed that the good performance of Bert is not only

multi-head attention was working, but also all components were working together [49] to form systematic engineering.

4.3. Analysis of the Distribution of Research Method Sentences in Flood Disaster Text

Usage method sentences generally appear in research or experimental papers; that is, research or experimental papers use certain methods to conduct experiments or research. If the usage method sentences in a certain period is significantly reduced, there are few research or experimental papers at this stage, and correspondingly, there are more theoretical papers. The annual distribution line chart of the usage method sentences has gone through four phases: oscillation, rising, falling, and rising, and an overall rising trend can be observed. Thus, mostly research or experimental papers were published in the field of flood disasters with relatively few theoretical papers. This also reflects that research on flood disaster risk reduction measures was more focused on practical applications. The number of published papers showed an overall increasing trend, which is consistent with the trend of usage method sentences.

The annual distribution line chart of the quotation method sentences displayed four phases of steady-sharp increase-sharp decrease-rise. This was consistent with the overall trend of the usage method sentences. However, the proportion of quotation method sentences was far less than that of usage method sentences, with distribution ranges of 0–0.02 and 0.06–0.14, respectively, which is consistent with the results in [36]. Although there was a sharp decline in the third phase, the proportions in 2003 and 2004 were still relatively high in the overall trend, which was related to the flood disaster events that occurred in 2003. According to statistics from the Swiss Reinsurance Company, the number of people killed in flood disasters worldwide reached 4280 in 2003. It included floods in India, Bangladesh, Pakistan, Nepal, and other countries, which had the second largest number of disaster casualties after those from earthquakes [50]. During the sharp increase phase in the number of papers since 2014, the proportion of quotation method sentences also increased slightly, but the increase was not as obvious as that of usage method sentences.

5. Conclusions

Academic literature is a high-quality long-tail data resource containing rich knowledge related to flood control and measures. It is necessary and valuable to extract flood control knowledge from literature using intelligent methods. A total of 4742 flood-related literature documents from the last 20 years were obtained through screening. The literature documents were then parsed using the machine learning tool GROBID, and 347 literatures composition sample data points were randomly selected from different years for sentence segmentation (approximately 61,000 sentences) and manual annotation. Traditional machine learning (NB, LR, SVM, and RF) and artificial neural network-based deep learning algorithms (Bert, Bert-CNN, Bert-RNN, and ERNIE) were used for model training and complete sentence-level knowledge extraction in batches. The conclusions are summarized as follows: (1) Artificial neural network-based deep learning methods exhibited better performance than traditional machine learning methods, but deep learning methods took much longer. (2) By comparing several deep learning methods in terms of comprehensive feature extraction capability and computational efficiency, the ranking of the performance of the deep learning methods was observed to be in the order ERNIE > Bert-CNN > Bert > Bert-RNN. (3) Using Bert as the benchmark model, several deformation models based on Bert were shown to have applicable characteristics. Bert was excellent at obtaining global features with satisfactory short-text performance; however, long text introduced extreme computational complexity, resulting in a sharp slowdown of the calculation speed. The Bert-CNN was suitable for obtaining local features, and it performed better for short text classification. Bert-RNN can process variable-length inputs and explore long-term dependencies but is limited by parallel computing capabilities. ERNIE improved the mask mechanism and corpus more, and the performance was better than that of the others. Finally, 124,196 usage method sentences and 8935 quotation method sentences were obtained in batches using ERNIE. A proportion of the usage method

sentences and quotation method sentences exhibited tortuous changes in rising and falling, but the overall trend was shown to be upwards over the last 20 years. This also indicates that with more literature accumulated with more method sentences, this study can lay a deep learning method foundation for the extraction of knowledge from more literature in the future. The next step will use more powerful methods (such as XLNet, RoBERTa, ALBERT, ELMo, etc.) to carry out experiments and compare their performances with ERNIE.

Author Contributions: Data preparation, processing, analysis, and writing—original draft of the manuscript, M.Z.; Research design, analysis, and writing—review, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 42050105; National Key R&D Program of China, grant number 2022YFF0711600; Chinese Academy of Sciences Project, grant number ZDRW-XH-2021-3; Construction Project of the China Knowledge Center for Engineering Sciences and Technology, grant number CKCEST-2021-2-18.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank Wang Xinbing of Shanghai Jiaotong University and his research team for providing data and technical support for this research.

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

References

1. Zhang, M.; Wang, J. Global flood disaster research graph analysis based on literature mining. *Appl. Sci.* **2022**, *12*, 3066. [CrossRef]
2. Li, Y. Construction and Application of Natural Disaster Emergency Knowledge Graph-Taking Flood Disaster as an Example. Ph.D. Thesis, Wuhan University, Wuhan, China, 2021.
3. Jiang, T. A comparative study of term extraction schemes in academic literature. *J. Inf. Resour. Manag.* **2021**, *111*, 112–122.
4. Bornmann, L.; Mutz, R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 2215–2222. [CrossRef]
5. Li, X.; Wang, Z.; Lu, Q. An extraction method for papers via integration of rules with SVM. *Comput. Technol. Dev.* **2017**, *27*, 24–29.
6. Wiebe, J.; Bruce, R.; O'Hara, T. Development and use of a gold standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2000. [CrossRef]
7. Cui, H.; Huang, C. Rule-based implementation of English sentence classification. *Informatiz. Constr.* **2015**, *11*, 180–181.
8. Hua, X.; Xu, F.; Wang, Z.; Li, P. Fine-grained classification method for abstract sentence of scientific paper. *Comput. Eng.* **2012**, *38*, 138–140.
9. Hayes, P.J.; Andersen, P.M.; Nirenburg, I.B.; Schmandt, L.M. TCS: A shell for content-based text categorization. In Proceedings of the Sixth Conference on Artificial Intelligence for Applications, Santa Monica, CA, USA, 5–9 May 1990; pp. 320–326.
10. Asghar, M.Z.; Khan, A.; Bibi, A.; Kundi, F.M.; Ahmad, H. Sentence-Level Emotion Detection Framework Using Rule-Based Classification. *Cogn. Comput.* **2017**, *9*, 1–27. [CrossRef]
11. Tan, L.; Phang, W.; Chin, K.O.; Patricia, A. Rule-based sentiment analysis for financial news. In Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Kowloon Tong, Hong Kong, 9–12 October 2015.
12. Widyantoro, D.H.; Amin, I. Citation sentence identification and classification for related work summarization. In Proceedings of the International Conference on Advanced Computer Science & Information Systems, Jakarta, Indonesia, 18–19 October 2014.
13. Liakata, M.; Saha, S.; Dobnik, S.; Batchelor, C.; Rebholz, D. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* **2012**, *28*, 991–1000. [CrossRef]
14. Hirohata, K.; Okazaki, N.; Ananiadou, S.; Ishizuka, M. Identifying sections in scientific abstracts using conditional random fields. In Proceedings of the Third International Joint Conference on Natural Language Processing, Hyderabad, India, 7–12 January 2008; pp. 381–388.
15. Shirsat, V.S.; Jagdale, R.S.; Deshmukh, S.N. Sentence level sentiment identification and calculation from news articles using machine learning techniques. In Proceedings of the ICCASP 2018, Lonere, India, 26–27 January 2018.
16. Chung, J.; Gulcehre, C.; Cho, K.H.; Bengio, Y. Empirical evaluation of gated recurrent neural networks in sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
17. Kim, Y. Convolutional neural networks for sentence classification. *arXiv* **2014**, arXiv:1408.5882.

18. Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.D.; Ng, A.Y.; Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. Empirical Methods in Natural Language Processing. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013.
19. Anke, L.E.; Schockaert, S. Syntactically aware neural architectures for definition extraction. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 378–385.
20. Li, X.; Zhang, Z.; Liu, H. Automatic recognition of concept definition sentences based on Bert model. *Inf. Sci.* **2022**, *40*, 160–166.
21. Wang, Z.; Li, C.; Huang, M.; Liu, S. Research on Intelligent Classification Method of Seismic Information Text Based on BERT-BiLSTM Optimization Algorithm. In Proceedings of the 2022 IEEE 2nd International Conference on Computer Communication and Artificial Intelligence (CCAI), Beijing, China, 6–8 May 2022; pp. 55–59. [CrossRef]
22. Jindal, A.; Gnaneshwar, D.; Sawhney, R.; Shah, R.R. Leveraging BERT with mixup for sentence classification (Student Abstract). In Proceedings of the Processing National Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
23. Zheng, S.; Yang, M. A New Method of Improving BERT for Text Classification. In Proceedings of the 9th International Conference, IScIDE 2019, Nanjing, China, 17–20 October 2019.
24. Zong, H.; Yang, J.; Zhang, Z.; Li, Z.; Zhang, X. Semantic categorization of Chinese eligibility criteria in clinical trials using machine learning methods. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 128. [CrossRef]
25. Xu, H.; Zhou, J.; Jiang, T.; Lu, J.; Zhang, Z.; Hu, W. Chinese telephone fraud text recognition based on word embedding and hybrid neural work. *Comput. Technol. Dev.* **2022**, *32*, 37–42.
26. Xia, Z. *Research on Chinese Short Text Classification Based on Pre-Trained Language Model*; Chongqing University of Technology: Chongqing, China, 2021.
27. Wang, X.; Li, S.; Yang, Z.; Lin, H.; Wang, J. Chinese knowledge base question answering system based on pre-trained language model. *J. Shanxi Univ. Nat. Sci. Ed.* **2020**, *43*, 955–962.
28. Wang, J.; Kun, B.; Yang, F.; Yuan, Y.; Wei, H. Disaster risk reduction knowledge service: A paradigm shift from disaster data towards knowledge services. *Pure Appl. Geo-Phys.* **2019**, *177*, 135–148. [CrossRef]
29. Wang, J.; Han, X.; Bu, K.; Zhang, M.; Wang, X.; Yuan, Y. Knowledge service system on disaster risk reduction and its application in social media analysis. *J. Glob. Chang. Data Discov.* **2020**, *4*, 25–32. [CrossRef]
30. Zhang, Q.; Zhou, L.; Tang, J.; Fu, L.; Wang, X. Internet of everything: Interconnection, mining and visualization of academic data. *Chin. J. Internet Things* **2018**, *2*, 56–60.
31. Zhang, Y.; Jia, Y.; Fu, L.; Wang, X. AceMap academic map and AceKG academic knowledge graph for academic data visualization. *J. Shanghai Jiaotong Univ.* **2018**, *52*, 1357–1362.
32. Zhou, Y. *Research on PDF Structure Analysis Technology of Academic Papers*; Hunan University: Changsha, China, 2020.
33. Bogdan. What's So Hard about PDF Text Extraction? Available online: https://filingdb.com/b/pdf-text-extraction (accessed on 14 July 2020).
34. Xue, H. *Information Recognition and Extraction from Chinese Periodical Papers Based on Conditional Random Fields*; Chinese Academy of Agricultural Sciences: Beijing, China, 2019.
35. Zhang, Y.; Zhang, C. Methodological and automatic sentence extraction from academic article's full-text. *J. China Soc. Sci. Tech. Inf.* **2020**, *39*, 640–650.
36. Zhang, Z.; Wang, Y.; Wang, R. Constructing the corpus of method in the information science domain. *Sci. Inf. Res.* **2020**, *2*, 30–45.
37. Hadipour, V.; Vafaie, F.; Deilami, K. Coastal flooding risk assessment using a GIS-based spatial multi-criteria decision analysis approach. *Water* **2020**, *12*, 2379. [CrossRef]
38. Akbari, R.; Hessami-Kermani, M.R.; Shojaee, S. Flood routing: Improving outflow using a new non-linear muskingum model with four variable parameters coupled with PSO-GA algorithm. *Water Resour. Manag.* **2020**, *34*, 3291–3316. [CrossRef]
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
40. Song, M.; Liu, Y. Application and optimization of Bert in sentiment classification of Weibo short text. *J. Chin. Comput. Syst.* **2021**, *42*, 714–718.
41. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
42. Nomponkrang, T.; Sanrach, C. The comparison of algorithms for Thai-sentence classification. *Int. J. Inf. Educ. Technol.* **2016**, *6*, 801–808. [CrossRef]
43. He, C.; Chen, S.; Huang, S.; Zhang, J.; Song, X. Using convolutional neural network with BERT for intent determination. In Proceedings of the 2019 International Conference on Asian Language Processing (IALP), Shanghai, China, 15–17 November 2019.
44. Luo, P.; Wang, Y.; Wang, J. Automatic discipline classification for scientific papers based on a deep pre-training language model. *J. China Soc. Sci. Tech. Inf.* **2020**, *39*, 14.
45. Zhang, J. Give Up Fantasy and Embrace Transformer: Comparison of Three Feature Extractors (CNN/RNN/TF) for Natural Language Processing. Available online: https://zhuanlan.zhihu.com/p/54743941 (accessed on 8 October 2022).
46. Tang, G.; Müller, M.; Rios, A.; Sennrich, R. Why self-attention? A targeted evaluation of neural machine translation architectures. *arXiv* **2018**, arXiv:1808.08946.

47. Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.S.; He, L. A survey on text classification: From shallow to deep Learning. *arXiv* **2020**, arXiv:2008.00364.
48. Alec, R.; Karthik, N.; Tim, S.; Ilya, S. Improving Language Understanding by Generative Pre-Training. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 8 October 2022).
49. Domhan, T. How much attention do you need? A granular analysis of neural machine translation architectures. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, NSW, Australia, 15–20 July 2018; Volume 1.
50. Yan, Q. Overview of global catastrophes in 2003. *Insur. Stud.* **2004**, *6*, 4.