Fine-grained Classification of Domain Scholarly Literature Based on Bibliographic Information (Postprint)

Authors: Lei Bing, Liu Xiao, Zhong Zhen

Date: 2023-04-01T16:02:53+00:00

Abstract

[Purpose/Significance] This study constructs a dual-label classification model for domain academic literature based on bibliographic information, categorizing according to "research content" and "research method", thereby providing methodological reference for fine-grained classification of academic literature.

[Method/Process] Taking the Convolutional Neural Network in deep learning as the base model, bibliographic information including title, abstract, keywords, journal name, author, and institution is categorized into explicit and implicit features. Through explicit feature extraction, implicit feature mapping, and other steps, a feature word array is formed, upon which a word vector matrix is generated. After processing through convolutional layers, pooling layers, and a Softmax layer, the classification task is completed.

[Results/Conclusion] Using literature in the e-commerce domain as an example for experimental verification, the results demonstrate that the macro F1-scores of this model for dual-label classification according to "research content" and "research method" are 0.74 and 0.81 respectively, which not only significantly outperforms traditional machine learning methods but also surpasses deep learning classification methods that utilize only explicit features.

Full Text

Fine-Grain Classification Method for Domain Academic Literature Based on Bibliographic Information

Lei Bing[1],[2], **Liu Xiao**[1],[2], **Zhong Zhen**[1],[2] [1]School of Management, Henan University of Technology, Zhengzhou 450001 [2]Business Intelligence and Knowledge Engineering Laboratory, Henan University of Technology, Zhengzhou 450001

Abstract: *[Purpose/Significance]* This study constructs a dual-label classification model based on bibliographic information for domain academic literature, categorized by "research content" and "research method," providing a methodological reference for fine-grain classification of academic literature. *[Method/Process]* Using convolutional neural networks (CNN) in deep learning as the base model, bibliographic information including title, abstract, keywords, journal name, author, and institution is divided into explicit and implicit features. Through explicit feature extraction and implicit feature mapping, a feature word array is formed. Based on this, a word vector matrix is generated and processed through convolutional, pooling, and Softmax layers to complete the classification task. *[Result/Conclusion]* Experimental verification using e-commerce literature demonstrates that the macro-F1 values of this model for dual-label classification of "research content" and "research method" are 0.74 and 0.81, respectively, which not only significantly outperforms traditional machine learning methods but also exceeds deep learning classification methods that use only explicit features.

Keywords: academic literature; subject classification; bibliographic information; deep learning; convolutional neural network

1. Introduction

Subject classification of academic literature is a fundamental task in library and information science that not only improves scholars' efficiency in retrieving academic information but also helps institutions such as science and technology management agencies and literature management platforms more accurately analyze domain development trends, formulate more reasonable policies or rules, and thereby accelerate scientific and technological progress. However, with the continuous refinement of research fields and the rapid growth of academic literature in recent years, traditional classification methods based on bibliographic information and manual or simple machine learning approaches have revealed issues of overly coarse classification and reduced accuracy. Traditional literature classification research primarily focuses on broad categories such as first-level or second-level disciplines, or on subject terms classification. For scholars, however, fine-grain classification is needed. This fine-grain classification typically manifests in two aspects: first, further subdivision of disciplinary directions (domains), such as dividing e-commerce into cross-border e-commerce, rural e-commerce, and e-commerce technology; second, classification dimensions should include not only "research content" but also "research method," meaning each academic literature receives dual labels—for example, a paper classified as "cross-border e-commerce" by research object and "empirical research" by research method.

On the other hand, existing academic literature classification primarily uses explicit bibliographic information such as title, abstract, and keywords. However,

data items like journal name, author, and research institution may not have explicit correlations with research content and method, yet the same journal, author, or institution typically focuses on specific research content and methods, suggesting potential implicit relationships. Therefore, exploring these implicit relationships to improve classification accuracy represents another objective of this study.

In view of this, this paper proposes a deep learning-based approach to explore a fine-grain classification method for domain literature based on bibliographic information. Bibliographic information serves as feature items for classification labels, including not only title, abstract, and keywords that are directly related to classification labels, but also implicit features such as journal name, author, and research institution that are not directly related.

2. Literature Review

Most existing academic literature classification studies are based on abstract, keywords, and title information from bibliographic records. For instance, Wu Jianguang et al. generated knowledge units by using high-frequency feature words from abstracts and manually identified key terms as central words, representing literature as several knowledge units and classifying documents by calculating similarity between knowledge units. H. Chu and Q. Ke applied grounded theory methods to code technical names collected from abstracts to achieve classification of research methods. V. Chakraborty et al. constructed a "document-term" matrix using keywords and abstracts as raw data to represent term frequency in documents. Zhou Lihong and Liu Kan extracted and filtered feature words from titles and abstracts based on part-of-speech tagging, represented literature as feature vectors, and performed classification through association rules. Li Hui and Xuan Hongsheng constructed "document-topic" and "topic-feature word" matrices based on patent literature titles and abstracts to mine technological innovation themes.

Additionally, some scholars have attempted to leverage external resources to construct features and improve classification performance. Li Xiangdong et al. utilized external feature information from CNKI, Wikipedia, and news pages to improve literature classification accuracy. Su Yan et al. used the Medical Subject Headings (MeSH) as a basis to select stem cell domain subject terms as feature vectors representing literature. Pan Donghua et al. selected patent classification codes to construct a technical dictionary, represented literature as vectors based on Derwent Manual Codes (DMC), and constructed a "patent-manual code" matrix to form a technical knowledge graph of patent literature.

Using machine learning algorithms to improve literature classification accuracy is the main technical approach. Common algorithms include Support Vector Machine (SVM), Naive Bayesian Model (NBM), and K-Nearest Neighbor (KNN), which have achieved good results in text classification. S. Baker et al. used SVM-based algorithms for semantic classification of large-scale medical litera-

ture with high accuracy. L. Jiang et al. improved NBM performance through locally weighted approaches. Bai Xiaoming and Qiu Taorong compared KNN and SVM performance for scientific literature classification.

In recent years, with the maturation of deep learning algorithms, studies have shown that Convolutional Neural Networks (CNN) can automatically learn features from text, reducing manual intervention in feature engineering, and typically outperform traditional machine learning algorithms in text classification. B. J. Gutierrez et al. verified that deep learning algorithms outperform traditional machine learning methods when classifying domain literature using multiple algorithms. Guo Limin achieved good results in multi-level classification of 1.7 million documents from the National Index to Chinese Newspapers and Periodicals using a CNN model.

In summary, scholars have conducted in-depth research on academic literature classification using machine learning algorithms based on abstract, keywords, and title information, with continuously improving accuracy. However, bibliographic information includes not only abstracts, keywords, and titles, but also journal names, authors, and institutions, which have rarely been studied. Our preliminary research also found that directly adding journal name, author, and institution data into feature vectors significantly decreased classification accuracy regardless of whether traditional machine learning or deep learning algorithms were used. The main reason is that while abstracts, keywords, and titles contain topic words directly related to research content and method, journal names, authors, and institutions contain almost no such direct topic words. Adding them without processing introduces noise and reduces classification accuracy.

In fact, each journal defines its research domain and "prefers" certain research methods. Each author has their own research domain and preferred methods. Each research institution or team also develops specific research domains and commonly used methods. We infer that implicit relationships with research content and method exist in these data items.

Therefore, this study divides bibliographic information data items into explicit features and implicit features. Abstract, keywords, and title are explicit features, while journal name, author, and institution are implicit features. For explicit features, we directly extract feature words; for implicit features, we perform feature mapping to make them explicit. Based on this, we construct a feature word array, vectorize it, and use it as input to a CNN classification model. In the CNN output layer, we design a "C (research content) × M (research method)" format to achieve dual-label classification. Finally, we validate the method's effectiveness using e-commerce domain literature.

3. Model Construction

This study manually annotates academic literature in the training and test sets according to "research content" and "research method" categories as corpora

for subsequent machine learning.

The basic approach to domain academic literature subject classification is: using bibliographic information as the classification basis, constructing an initial feature matrix through feature extraction, vectorizing it, and then implementing fine-grain classification (dual-label classification by "research content" and "research method") through CNN deep learning algorithms. Academic literature bibliographic information generally includes title, author, institution, journal name, keywords, and abstract, as shown in Table 1 .

Table 1. Example of Academic Literature Bibliographic Information

Title | On the Legal Regulation of "Choose One of Two" Behavior on E-commerce Platforms |
Institution | XX University Law School |
Keywords | Digital Economy; E-commerce; Platform "Choose One of Two"; P2B Regulations |
Abstract | Similar to traditional economies, mandatory "choose one of two" behavior in the digital economy context is not "per se illegal," but if the actor uses this method to severely damage competitors' ability to achieve minimum economies of scale or prevent new enterprises from entering the market, it will significantly hinder market competition. Considering various barriers to market entry including economic, technical, and data obstacles, especially network externalities, China's e-commerce platforms have become highly concentrated. To maintain market competitiveness and enable merchants and consumers to fully experience the benefits and convenience of e-commerce, competition law enforcement agencies should ensure multi-homing of platform merchants—that is, no platform operator has the right to force merchants to trade on only one platform. Meanwhile, considering the characteristics of e-commerce and small and medium-sized merchants' dependence on platform intermediaries, China needs to formulate specialized laws regulating the transaction relationship between intermediary platforms and merchants, and improve Article 35 of the E-commerce Law. |

The classification model proposed in this paper mainly consists of three parts: (1) Construction of feature dictionaries and stop-word dictionaries. Selecting titles, abstracts, and keywords from all literature in the training set to build a "local feature word dictionary" (user_{dict}) and a "stop-word list" (stop_{wordslist}) to improve word segmentation accuracy. (2) Feature matrix construction and vectorization. Dividing bibliographic information into explicit features (abstract, title, keywords) and implicit features (author, journal name, institution). For explicit features, perform word segmentation and stop-word removal; for implicit features, use feature mapping to make them explicit. Based on this, construct a feature word array and perform vectorization as input data for the CNN classification model. (3) Deep learning for literature classification. Classifying literature through the CNN model, with the output layer designed in a "C (research content) × M (research method)" format to achieve dual-label classification. As shown in Figure 1 [Figure 1: see original

paper].

Figure 1. Framework of Domain Academic Literature Subject Classification Method

3.1 Construction of Stop-Word and Feature Word Dictionaries The construction of feature word dictionaries and stop-word dictionaries is a crucial step in data preprocessing. The feature word dictionary serves as a custom dictionary for word segmentation, improving segmentation accuracy. The stop-word dictionary helps filter "noise" from segmentation results, thereby improving deep learning model performance and preventing overfitting. For domain literature, this study designs a method to construct feature word and stop-word dictionaries based on titles, abstracts, and keywords from all literature in the training set.

The feature word dictionary consists of three parts: First, considering the importance of keywords in bibliographic information, all keywords are included in the feature word dictionary. Second, high-frequency words ($ $5) from titles and abstracts are included. Finally, typical vocabulary representing domain literature themes is added based on domain expert knowledge.

For the stop-word dictionary, this study initially used only the Harbin Institute of Technology stop-word lexicon, but the results were unsatisfactory. Analysis revealed two main reasons: First, academic literature contains many formal descriptive words, such as sentence-initial words like "with," "point out," "according to," which can "mislead" machine learning. Second, low-frequency words (<5) with insignificant classification features are prone to overfitting in machine learning. Therefore, in addition to the Harbin Institute of Technology stop-word lexicon, this study's stop-word dictionary also includes sentence-initial words and low-frequency words with insignificant classification features.

3.2 Feature Matrix Construction and Vectorization Although deep learning models can automatically find features from distributed word vectors and have strong transferability and high learning efficiency compared to traditional machine learning algorithms like conditional random fields and support vector machines, the quality of initial features still affects deep learning efficiency. Poor-quality features can lead to overfitting or underfitting. This study divides bibliographic information data items into explicit and implicit features for separate processing.

3.2.1 Explicit Feature Extraction First, keywords are directly added to the keyword feature set K. Second, introducing feature word and stop-word dictionaries, word segmentation tools are used to segment titles and abstracts, forming title feature set T and abstract feature set S, as shown in equations (1)-(3):

$$K = (k_1, k_2, ..., k_r) \tag{1}$$

$$T = (t_1, t_2, ..., t_p) \tag{2}$$

$$S = (s_1, s_2, ..., s_q) \tag{3}$$

where k_r represents the r-th word in keywords, t_p represents the p-th word in the title, and s_q represents the q-th word in the abstract. It should be noted that r, p, and q are variable for each document. To fix the length for subsequent word vectors, three hyperparameters R ($\geq$ r), P ($\geq$ p), and Q ($\geq$ q) are set to fix the lengths of K, T, and S, with insufficient parts filled by "0," as shown in equations (4)-(6):

$$K = (k_1, k_2, ..., k_r, 0, ..., 0) \tag{4}$$

$$T = (t_1, t_2, ..., t_p, 0, ..., 0) \tag{5}$$

$$S = (s_1, s_2, ..., s_q, 0, ..., 0) \tag{6}$$

3.2.2 Implicit Feature Mapping Institutions, journal names, and authors in literature bibliographic information have implicit associations with research content or methods. Taking the e-commerce domain as an example: academic papers published by computer science schools may focus on "information technology applications in e-commerce," while law school papers center on "e-commerce laws and regulations"; agricultural journals may publish papers on "rural e-commerce poverty alleviation," while international trade journals may discuss "cross-border e-commerce." Similarly, specific authors typically use relatively fixed research methods and focus on certain research domains, but when collaborating with other scholars, their research content or methods may change.

Therefore, this study uses feature mapping to make the feature information in institutions, journal names, and authors explicit, then adds them to the initial feature matrix. The specific mapping process is as follows:

(1) Author Feature Processing. Associate authors with domain literature and perform explicit processing of the implicit author feature based on co-occurrence frequency between authors and research content/methods in published literature. Figure 2 [Figure 2: see original paper] shows the generation process of research method labels for different types of authors.

Figure 2. Author Mapping Process

Construct an author co-occurrence matrix based on author collaboration relationships in domain literature. If the co-occurrence frequency (i.e., collaboration frequency) exceeds a specific threshold, the two authors are considered to have a stable cooperative relationship in a certain domain and are treated as collaborative authors for feature mapping; otherwise, only the first author is mapped. Specifically: first, perform frequency statistics on collaborative authors (or first

authors) by research method category to generate an "author-research method" frequency distribution table. Then, calculate the probability value JAP of different research methods used by different authors, as shown in equation (7). A larger JAP value indicates stronger preference of an author for a certain research method. Finally, generate an "author-research method" probability distribution table.

$$JAP_i(j) = \frac{m_{ij}}{\sum_{i=1}^{M} m_{ij}} \tag{7}$$

where M represents the number of research method categories in domain literature, and m_{ij} represents the frequency of author j using the i-th research method. Based on the probability distribution table, map authors to explicit research method features: first set a JAP conversion probability threshold, then select the research method label with the maximum JAP value that meets the threshold. The threshold is a hyperparameter. Assuming through experimentation the threshold is set to 0.7, if an author's JAP value is not lower than 0.7, the author is mapped to that research method label; otherwise, it is replaced by placeholder "0." Table 2 shows examples of authors mapped to research method labels. "Author-1" is mapped to "Research Method-2," "Co-author-2" to "Research Method-1," "Author-5" to "Research Method-3," while "Author-3" and "Co-author-4" are replaced by placeholder "0."

Table 2. Example of "Author-Research Method" Probability Distribution Table

	Research Method-1	Research Method-2	Research Method-3	Research Method-4
Author-1	0.1	0.8	0.1	0
Co-author-2	0.9	0.05	0.05	0
Author-3	0.2	0.3	0.2	0.3
Co-author-4	0.25	0.25	0.25	0.25
Author-5	0.2	0.1	0.7	0

(2) Journal Feature Processing. Similar to author feature processing, associate journal names with domain literature and map them to explicit research content and research method features. Taking research content as an example, the processing flow is shown in Figure 3 [Figure 3: see original paper].

Figure 3. Journal Mapping Process

First, use journal names as objects to count the frequency of different research content in each journal, generating a "journal-research content" frequency distribution table, and calculate the probability value JCP of "research content" labels for each journal, as shown in equation (8). Similarly, a larger JCP value indicates stronger preference of a journal for a certain research content. Then

generate a "journal-research content" probability distribution table based on JCP.

$$JCP_i(j) = \frac{c_{ij}}{\sum_{i=1}^{C} c_{ij}} \tag{8}$$

where c represents the number of research content categories in domain literature, and c_{ij} represents the frequency of journal j for the i-th research content label. Based on the probability distribution table, map journal names to explicit research content features: first set a JCP conversion probability threshold, then convert the journal name to research content labels with values greater than or equal to the threshold. If no labels meet the conditions or the number of labels is insufficient, replace with placeholder "0." Assuming through experimentation the threshold is set to 0.33, if a research content's JCP value is not lower than 0.33, that research content label is added to the journal mapping set. Table 3 shows examples of journal names mapped to research content labels. "Journal-1" mapping set is {Research Content-1, Research Content-4, 0}, "Journal-3" is {Research Content-3, Research Content-4, Research Content-5}, and "Journal-5" is {0, 0, 0}.

Table 3. Example of "Journal-Research Content" Probability Distribution Table

	Research Content-1	Research Content-2	Research Content-3	Research Content-4	Research Content-5
Journal-1	0.5	0.2	0.1	0.4	0.1
Journal-2	0.1	0.1	0.1	0.1	0.6
Journal-3	0.2	0.2	0.4	0.4	0.4
Journal-4	0.1	0.6	0.1	0.1	0.1
Journal-5	0.1	0.1	0.1	0.1	0.1

(3) Institution Feature Processing. Associate research institutions with domain literature and map them to explicit research content and research method features. Taking research content as an example, the processing flow is shown in Figure 4 [Figure 4: see original paper].

Figure 4. Institution Mapping Process

First, process research institutions as follows: If multiple institutions exist in a document, only select the first institution; Use regular expressions to divide first-level and second-level institutions. Considering that first-level institutions like "XX University" basically cannot indicate the research content of domain literature, first-level institutions are deleted and only second-level institutions such as School of Economics and Management or Law School are retained for feature mapping. Then, calculate the probability values of "research content"

and "research method" labels appearing in each institution, and perform feature mapping using the same method as journal mapping.

3.2.3 Word Vectorization Add the processed explicit and implicit features to the feature word array D, as shown in equation (9):

$$D = [K, T, S, A, J, O] \tag{9}$$

where K, T, S, A, J, O represent processed keywords, title, abstract, author, journal name, and institution data, respectively.

Then, convert D into word vectors through Word2Vec to form the initial feature matrix for subsequent deep learning models. Word2Vec is a shallow neural network model that maps words into multi-dimensional digital space, where the position in digital space indicates semantic information. Skip-Gram is a method of the Word2Vec model that can predict the probability of context words appearing from center words. Using pre-trained word vectors can significantly improve CNN model classification performance. Drawing on A. Timoshenko et al., this study sets the sliding window size c to 5 and word vector dimension d to 20, inputs array D into the word vector model, and outputs the word vector matrix $D^* \in \mathbb{R}^{d \times n}$ as input to the CNN model.

3.3 Deep Learning for Literature Classification Compared with traditional machine learning algorithms, deep learning models have achieved better performance in large-scale text classification. Deep learning models can learn from shallow primary features to deep advanced features through neuron connections. For the word vector matrix D^* constructed in this study, the deep learning model CNN can learn both global features and detailed features contained in different bibliographic information.

The CNN model consists of an input layer, convolutional layer, pooling layer, and Softmax layer, using gradient descent methods to adjust weight parameters in reverse. The specific structure is shown in Figure 5 [Figure 5: see original paper]. The CNN input layer is the word vector matrix D^*. The convolutional layer performs convolution operations on the initial feature matrix through multiple convolution kernels to form feature maps. Then, pooling operations are performed on feature maps to reduce dimensions and retain maximum feature values. The pooling layer can filter out useless features and retain important ones. The Softmax layer converts the vector output from the pooling layer into literature subject probability values through a fully connected layer and Softmax function to predict literature categories.

Figure 5. CNN Model Structure

For the research subject set J, this study combines "research content" and "research method" labels in a "C × M" format to achieve dual-label classification, as shown in equation (10):

$$J = C \times M \tag{10}$$

where C represents the research content label set, M represents the research method label set, and J represents the combination of research content and research method labels, thereby achieving dual-label classification. The specific process is shown in Figure 6 [Figure 6: see original paper]. For example, assuming a domain literature has 4 research methods and 8 research content categories, 32 subject labels are needed, labeled Subject Label 1, Subject Label 2, through Subject Label 32. If a document is labeled as "Subject Label 32," its research content and method are "Research Content 8" and "Research Method 4," respectively.

Figure 6. Dual-Label Classification Implementation Process

4. Experimental Validation

To verify the feasibility and effectiveness of the classification model, this study classifies "e-commerce" literature from CNKI by research content and method, and compares it with traditional machine learning algorithms such as Support Vector Machine and Naive Bayes.

4.1 Data Source The data for this study comes from the China Journal Full-text Database of CNKI. Using "e-commerce" as the search term for subject search, journal categories were limited to CSCD, EI, and CSSCI. The search time range was from May 15, 1998 to June 10, 2020, retrieving a total of 8,874 records. Downloaded content included title, research institution, publication journal, keywords, abstract, and other information. After deduplication, noise processing, and missing value handling, 7,647 documents were finally selected for annotation.

Among the 7,647 documents, there were 13,977 keywords, with the highest frequency feature word being "cross-border e-commerce" (136 occurrences); 785 journals, with the highest single journal frequency being 291 (China Circulation Economy), and 560 journals appearing more than twice; 6,785 research institutions, with the highest single institution publication frequency being 82 (Wuhan University School of Information Management), and 1,899 institutions appearing more than twice; 10,568 authors, with the highest single author frequency being 33, and 2,262 authors appearing more than twice. The distribution of high-frequency subject words, journals, and research institutions is shown in Table 4 .

Table 4. Distribution of High-Frequency Subject Words, Journals, and Research Institutions

High-Frequency Subject Words	Cross-border e-commerce
High-Frequency Journals	China Circulation Economy, Library and Information Service, Science and Technology Management Research, Computer Engineering, Productivity Research, Science & Technology Progress and Policy
High-Frequency Institutions	Wuhan University School of Information Management, Jilin University School of Management, Huazhong University of Science and Technology School of Management, Xi'an Jiaotong University School of Management, Chongqing University School of Economics and Business Administration, Xi'an Jiaotong University School of Economics and Finance, University of Shanghai for Science and Technology School of Management, Beijing University of Posts and Telecommunications School of Economics and Management, Renmin University of China Business School, Fudan University School of Management

4.2 Manual Annotation Through preliminary literature research and repeated discussions with multiple e-commerce scholars, drawing on domain literature classification methods by Xiao Lianjie and Zhang Chengzhi et al., 13 classification labels were finally determined, divided into two major categories of "research content" and "research method," covering the main research areas of current e-commerce research, as shown in Table 5 .

Table 5. Domain Literature Subject Labels

Research Content	Business Model; Laws and Regulations; Logistics, Payment, and Finance; Marketing; E-commerce Technology; Rural E-commerce; Cross-border E-commerce; Credit Risk; Other E-commerce
Research Method	Theoretical Research; Empirical Research; Case Study; Technical Research

In the research content subject labels of Table 5, "Other E-commerce" represents niche research areas such as e-commerce talent cultivation. In terms of research methods, this study classifies research that mainly uses qualitative methods to analyze concepts or interpret policies as "Theoretical Research"; research that primarily uses econometrics or industrial economics methods to study or test macro and meso-level cross-sectional or time-series data as "Empirical Research"; research that builds models and uses specific data to analyze cases, generally focusing on organizational behavior, as "Case Study"; and "Technical Research" refers to research using computer technology in e-commerce-related fields.

Considering the professionalism of domain literature and the consistency of annotation standards, this study adopts a small-scale domain expert annotation method rather than the popular crowdsourcing model. Crowdsourced annotation is often completed by many non-domain personnel, which improves efficiency but is unsuitable for highly professional academic literature annotation. Specifically, based on domain expert assistance, this study determines several feature words for each research content label (see Table 6) and performs research content annotation by combining feature word location and frequency. If a document involves two or more research contents, annotation is based on feature word frequency, selecting the research content label with higher feature word frequency.

Table 6. Annotation Features of E-commerce Domain Literature by Research Content Label

Research Content Label	Main Feature Words
Business Model	B2B model, B2C model, online and offline integration, etc.
Laws and Regulations	Tax law, e-commerce law, consumer rights protection law, etc.
Logistics, Payment, and Finance	Payment systems, logistics distribution, P2P, etc.
Marketing	Online reviews, pricing research, consumer preferences, etc.
E-commerce Technology	Recommendation algorithms, cloud computing, etc.

Research Content Label	Main Feature Words
Rural E-commerce	Agricultural product e-commerce, e-commerce poverty alleviation, etc.
Cross-border E-commerce	WTO, trade facilitation, etc.
Credit Risk	Trust crisis, credibility, etc.
Other E-commerce	Other e-commerce domain research issues

Table 7 lists five examples of literature annotation. Taking the third document as an example, feature words related to both "rural e-commerce" and "logistics distribution" appear in the bibliographic information, keywords, and abstract, but "logistics distribution" related feature words appear 5 times while "rural e-commerce" related feature words appear 3 times, thus it is classified as "Logistics, Payment, and Finance." In terms of research method, the words "improved algorithm" appear in the title and keywords, so the research method label is determined as "Technical Research."

Table 7. Annotation Examples

Title	Author/Institution/Journal/Year	Keywords	Research Content Label	Research Method Label
A Simple and Effective Anonymous Fair E-commerce Protocol for Trust Issues among Customers, Merchants, and Third Parties in E-commerce	Zhang XX, XX University School of Business, Symbiosis or Iteration: Re-discussing Cross-border E-commerce and Global Digital Trade, Contemporary Economic Management, 2020	E-commerce, Information Security, Fairness	E-commerce Technology	Technical Research

Title	Author/Institution/Journal/Year	Keywords	Research Content Label	Research Method Label
Rural E-commerce Logistics Distribution Problem, Considering Multiple Distribution Centers within Regions, Dispersed Customer Residences, and Simultaneous Collection and Delivery Needs	Luo XX, XX University School of Information, An Improved Ant Colony Algorithm for Rural E-commerce Integrated Collection and Delivery Vehicle Routing Problem, Systems Engineering, 2019	Common Distribution, Rural E-commerce, Integrated Collection and Delivery Vehicle Routing Problem, Improved Ant Colony Algorithm	Logistics, Payment, and Finance	Technical Research

Title	Author/Institution/Journal/Year	Keywords	Research Content Label	Research Method Label
EC-CDIO E-commerce Talent Cultivation Model Construction	Wang XX, XX University School of Management Science and Engineering, EC-CDIO, Talent Cultivation, Higher Engineering Education Research, 2019	E-commerce, EC-CDIO, Talent Literacy	Other E-commerce	Theoretical Research

Title	Author/Institution/Journal/Year	Keywords	Research Content Label	Research Method Label
Advantages, Challenges, and Countermeasures of Shanghai's Participation in "Belt and Road" Construction—A Study on Rural E-commerce Entrepreneurship Models under Rural Revitalization Strategy	Wang XX, XX University, Rural Revitalization, Rural E-commerce Entrepreneurship Models, Rural E-commerce Entrepreneurship Elements, Agricultural Economy and Management, 2019	Rural Revitalization Strategy, Rural E-commerce Entrepreneurship Models	Rural E-commerce	Case Study

To ensure annotation accuracy, three master's students in e-commerce research independently completed the annotation work according to the above rules. If two or more annotators agreed, the document's label category was determined; if all three disagreed, the document was handled by domain experts.

4.3 Experimental Analysis

4.3.1 Evaluation Metrics Domain literature subject classification is evaluated using Precision (P), Recall (R), and F1 values. This study treats each class label as positive and other categories as negative, constructing a confusion matrix for each label. TP represents samples correctly classified to a certain label, FP represents samples incorrectly classified to this label, TN represents samples correctly classified to other labels, and FN represents samples incorrectly classified to other labels. P, R, and F1 values are calculated as:

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{13}$$

4.3.2 Comparative Analysis E-commerce domain literature classification results are shown in Tables 8 , 9 , and 10 . Table 8 shows the classification accuracy of our method, Table 9 shows classification accuracy based on different initial feature constructions, and Table 10 shows classification accuracy based on other machine learning models.

Table 8. Classification Results of E-commerce Domain Literature Subject Labels

Research Content	P	R	F1
Business Model	0.57	0.60	0.58
Laws and Regulations	0.69	0.41	0.51
Logistics, Payment, and Finance	0.64	0.67	0.65
Marketing	0.70	0.67	0.68
E-commerce Technology	0.50	0.50	0.50
Rural E-commerce	0.97	0.69	0.80
Cross-border E-commerce	0.45	0.54	0.49
Credit Risk	0.50	0.50	0.50
Other E-commerce	0.69	0.45	0.54
Overall	**0.72**	**0.73**	**0.74**

Research Method	P	R	F1
Theoretical Research	0.88	0.80	0.81
Empirical Research	0.75	0.70	0.72
Case Study	0.79	0.78	0.78
Technical Research	0.77	0.78	0.77
Overall	**0.88**	**0.80**	**0.81**

From Table 8, in e-commerce domain literature classification results, "Rural E-commerce" achieves the highest precision at 97%, while "Business Model" has the lowest at 48%. "Other E-commerce" and "Logistics, Payment, and Finance" also have relatively low precision, while other labels are above 70%. Analysis reveals that poorly classified research content has relatively broad scope. For example, literature labeled "Other E-commerce" involves low-coverage research content such as "e-commerce talent cultivation" and "tourism e-commerce." The inconsistency in research content leads to high feature dispersion in "Business Model" and "Other E-commerce" categories, resulting in relatively poor classification. For research methods, except for "Case Study" with lower precision, other methods are above 85%, showing good classification results. For case studies, statistical analysis shows that the proportion of literature using case study methods in the e-commerce domain is small, accounting for only 7.36% of all annotated documents. The small number of documents may lead to poor feature extraction and overfitting for the "Case Study" method.

Table 9. Comparison of Different Data Input and Preprocessing for E-commerce Domain Literature Subject Classification

Input and Preprocessing	Research Content			Research Method		
	P	R	F1	P	R	F1
Our Classification Model	0.72	0.73	0.74	0.88	0.80	0.81
Directly Adding Journal, Author, Institution Names	0.63	0.62	0.62	0.75	0.70	0.72
Title and Abstract Data Only	0.71	0.72	0.72	0.79	0.78	0.78
Only "Harbin Institute of Technology" Stop-word Dictionary	0.69	0.70	0.70	0.77	0.78	0.77

Table 9 shows that the bibliographic information-based classification model achieves 72% precision, 73% recall, and 74% macro-F1 for research content classification, and 88% precision, 80% recall, and 81% macro-F1 for research method

classification. Directly adding raw author, institution, and journal name data to the feature matrix reduces macro-F1 by 9% and 11% for research content and method, respectively. Using only "title" and "abstract" data with the CNN literature classification algorithm employed by other scholars reduces macro-F1 by 2% and 3% compared to our method. Without the domain feature dictionary and using only the "Harbin Institute of Technology stop-word" list for word segmentation, macro-F1 differs by 4% for both research content and method classification. This directly demonstrates that our feature mapping of authors, institutions, and journals and construction of the initial feature matrix help improve model classification performance.

Table 10. Classification Results of Different Models for E-commerce Domain Literature Subject Classification

Model	Research Content			Research Method		
	P	R	F1	P	R	F1
Our Classification Model (CNN)	0.72	0.73	0.74	0.88	0.80	0.81
SVM	0.64	0.65	0.64	0.75	0.67	0.68
NBM	0.63	0.66	0.65	0.75	0.67	0.68
KNN	0.61	0.63	0.62	0.72	0.65	0.68

To verify the effectiveness of CNN algorithms for fine-grain classification of domain literature, this study uses common machine learning algorithms as comparison experiments, including classic SVM, NBM, and KNN algorithms. In experiments, all feature items are identical except for the model. Table 10 shows that our CNN-based classification model achieves the best performance. Among traditional machine learning algorithms, NBM performs relatively well for literature subject classification, but the gap compared to CNN is significant, with macro-F1 differing by 9% for research content and 13% for research method.

Through comparative analysis of Tables 9 and 10, our method improves macro-F1 values for literature classification results, demonstrating that our proposed method is effective for fine-grain subject classification of domain literature.

5. Conclusion

This study constructs a fine-grain literature classification model based on bibliographic information. First, bibliographic information representing literature themes is screened, and feature word and stop-word dictionaries are constructed based on titles, keywords, and abstracts from all training set literature. Then, explicit features (keywords, title, abstract) are extracted, and implicit features (author, journal, institution) are mapped to build a feature array. Next, the feature array is trained as word vectors and used as input to the CNN model. Finally, dual-label classification of "research content" and "research method" for domain literature is achieved through the CNN model. Experimental results

show that our classification model not only significantly outperforms traditional machine learning methods but also achieves higher accuracy than deep learning classification methods using only explicit features from titles.

However, domain literature subject classification based on bibliographic information still faces some issues: First, the complexity of domain literature research topics—single literature may include multiple aspects, but this study only assigns subject labels based on the highest-frequency feature word, which can lead to inaccurate classification. For example, for literature on rural e-commerce logistics, our annotation based on feature word frequency is "Logistics, Payment, and Finance," but the model classifies it as rural e-commerce, indicating the limitation of single-category output. Future research will improve the CNN model output layer to design multi-category output to enhance classification accuracy. Second, literature classification labels depend on domain experts, introducing subjective limitations. Future work will combine machine learning algorithms with domain expert knowledge to obtain research topics and improve automated classification capability. Third, the data scale used in this experiment is relatively small; future work will conduct large-scale domain literature classification experiments to further validate our proposed method.

References

[1] Liu Aijun, Yu Liping. Objective Classification of Bibliometric Indicators and Its Implications: Taking JCR2015 Economics Journals as an Example. *Information Studies: Theory & Application*, 2017, 40(7): 33-37, 49.

[2] Chakraborty V, Chiu V, Vasarhelyi M. Automatic classification of accounting literature. *International Journal of Accounting Information Systems*, 2014, 15(2): 122-148.

[3] Wu Jianguang, Su Yunmei, Yu Qi, et al. Research on Academic Literature Classification Based on Knowledge Elements. *Information Studies: Theory & Application*, 2019, 42(3): 160-165.

[4] Chu H, Ke Q. Research methods: what's in the name? *Library & Information Science Research*, 2017, 39(4): 284-294.

[5] Zhou Lihong, Liu Kan. Research on Scientific Literature Classification Based on Association Rules. *Library and Information Service*, 2012, 56(4): 12-16, 119.

[6] Li Hui, Xuan Hongsheng. Method for Mining Technological Innovation Topics Integrating Multi-Attributes from Patent Perspective: Taking Chip Domain Patents as an Example. *Library and Information Service*, 2020, 64(11): 96-107.

[7] Li Xiangdong, Liu Kang, Ding Cong, et al. Research on Mixed Automatic Classification of Multiple Types of Literature Based on CNKI. *New Technology of Library and Information Service*, 2016(2): 59-66.

[8] Li Xiangdong, Ruan Tao, Liu Kang. Research on Automatic Classification of Multiple Types of Literature Based on Wikipedia. *Data Analysis and Knowledge*

Discovery, 2017, 1(10): 43-52.

[9] Li Xiangdong, Gao Fan, Li Youhai. Research on Cross-Literature-Type Text Automatic Classification Under Common Semantic Space. *Data Analysis and Knowledge Discovery*, 2018, 2(9): 66-73.

[10] Su Yan, Xu Ping, Kong Liangliang, et al. Exploration of Biomedical Classification Subject Headings Reconstruction Based on MeSH: Taking Stem Cell Research Literature as an Example. *Library Journal*, 2015, 34(3): 47-52.

[11] Pan Donghua, Xu Keke. Research on Patent Literature-Based Technical Knowledge Graph Drawing Method. *Journal of the China Society for Scientific and Technical Information*, 2015, 34(8): 866-874.

[12] Baker S, Silins I, Guo Y, et al. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 2016, 32(3): 432-440.

[13] Jiang L, Caiz, Zhang H, et al. Naive Bayes text classifiers: a locally weighted learning approach. *Journal of Experimental & Theoretical Artificial Intelligence*, 2013, 25(2): 273-286.

[14] Bai Xiaoming, Qiu Taorong. Research on Automatic Classification of Scientific Literature Based on SVM and KNN Algorithms. *Microcomputer Information*, 2006(36): 275-276, 265.

[15] Wang S, Huang M, Deng Z. Densely connected CNN with multi-scale feature attention for text classification. *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*. Stockholm: International Joint Conferences on Artificial Intelligence Organization, 2018: 4468-4474.

[16] Gutierrez BJ, Zeng J, Zhang D, et al. Document classification for COVID-19 literature. arXiv preprint arXiv:2006.13816, 2020.

[17] Guo Limin. Research on Automatic Literature Classification Based on Convolutional Neural Network. *Library and Information*, 2017(6): 96-103.

[18] Du Dehui, Li Changling, Xiang Fuzhong, et al. Discussion on Interdisciplinary Related Knowledge Discovery Method Based on Citation Keywords. *Journal of Intelligence*, 2020, 39(9): 189-194.

[19] Yu Yan, Zhao Naizhen. Selection of Domain Stop Words for Patent Topic Analysis Based on Auxiliary Sets. *Data Analysis and Knowledge Discovery*, 2018, 2(11): 95-103.

[20] Xiao Lianjie, Meng Tao, Wang Wei, et al. Research on Intelligence Analysis Method Recognition Based on Deep Learning: Taking Security Intelligence Domain as an Example. *Data Analysis and Knowledge Discovery*, 2019, 3(10): 20-28.

[21] Marco B, Georgiana D, German K. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings*

of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: Association for Computational Linguistics, 2014: 238-247.

[22] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *Computer Science*, 2015(10): 253-263.

[23] Timoshenko A, Hause Jr J. Identifying customer needs from user-generated content. *Marketing Science*, 2019, 38(1): 1-20.

[24] Yan Y, Yin X-C, Yang C, et al. Biomedical literature classification with a CNNs-based hybrid learning network. *PloS One*, 2018, 13(7): 1-31.

[25] Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882, 2014.

[26] Zhang Chengzhi, Li Zhuo, Chu Heting. Research on Automatic Classification of Academic Paper Research Methods Based on Full Text. *Journal of the China Society for Scientific and Technical Information*, 2020, 39(8): 852-862.

[27] Tang Lin, Guo Chonghui, Chen Jingfeng, et al. Research on Domain Ontology Concept Hierarchical Relationship Extraction Based on Chinese Academic Literature. *Journal of the China Society for Scientific and Technical Information*, 2020, 39(4): 387-398.

[28] Willis CG, Law E, Williams AC, et al. CrowdCurio: an online crowdsourcing platform to facilitate climate change studies using herbarium specimens. *New Phytologist*, 2017, 215(1): 479-488.

[29] Zhang Huaxin, Pang Jiangang. Research on Text Classification Based on SVM and KNN. *Modern Intelligence*, 2015, 35(5): 73-77.

[30] Xiao Liming, Yu Kuan, Cai Pin. A Chinese Journal Automatic Classification System Based on Bayes Classifier. *Modern Intelligence*, 2007(4): 146-147, 150.

Author Contributions

Lei Bing: Research design and paper revision; Liu Xiao: Model construction, data testing, and paper writing; Zhong Zhen: Topic selection and paper revision.

Note: Figure translations are in progress. See original paper for figures.