International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-12

https://doi.org/10.70917/ijcisim-2026-1006
Article

Innovation and Practice of the Ethical Supervision Mechanism for Computer-Assisted Publishing under the Integrated Publishing Model

Yongguo Hu [1], Li Dai [1], Zonghui Wu [1,2,*], Wenjie Huang [1], Jie Liu [1], Yahui Li [3] and Yongsong Yan [4]

[1] Editorial Department of Health Medicine Research and Practice, Southwest University, Chongqing, 400715, China
[2] Southwest University Hospital, Chongqing, 400715, China
[3] Editorial Department of Computer Science, Chongqing, 401121, China
[4] Editorial Department of Nano Materials Science, Chongqing University, Chongqing, 400044, China
* Correspondence author: yghu_editor@163.com

Abstract: With the development of artificial intelligence, the ethical regulation of computer-aided publishing is strongly challenged. This paper establishes the process of ethical regulation of computer-aided publishing, calculates the text similarity, and uses the K-Means algorithm to cluster and analyze the subject keywords. The LDA model is established on the basis of Dirichlet distribution to mine the subject features of computer-aided publishing, and the classification key results are based on plain Bayes, so as to discuss the ethical misconduct of computer-aided publishing. The results show that 73.62% of the keywords in the field of computer-aided publishing, which accounted for the largest percentage, belong to verbs. The mean value of similarity of near-synonyms under publishing keywords reaches 0.864, which helps to sort out the association of different attribute word classes. Also the dendrogram of keywords obtained by clustering has the highest distance of 1.15. The computer-aided publishing regulatory mechanism established in this paper helps to adapt to the convergent publishing model.

Keywords: computer-aided publishing; text mining; ethical regulation; K-Means algorithm; LDA model; plain Bayesian

1. Introduction

In the era of digital intelligence, China's publishing industry, which has steadily improved its overall strength and quality and efficiency, is experiencing profound changes in the midst of its vigorous development [1]. For example, the application of Artificial Intelligence (AI) technology realizes automatic screening and content classification of publications, big data analyzes readers' needs to adjust the publishing market plan, and ChatGPT and AIGC technology realizes language checking, manuscript proofreading, data verification, and literature searching in digital publishing through human-computer interaction, etc. [2-3]. However, in the current publishing industry, which is dominated by commercialized operation, the publishing professional subjects are also facing the ethical dilemma deepened by digital and intellectual technologies [4]: on the one hand, they need to abide by the ethics of technology, assume social responsibility and pay attention to the public interest, and on the other hand, they also need to apply the emerging technologies to maintain survival and development, and pursue private profits. Publishing practitioners are both members of publication production and sales organizations and ordinary individuals who are citizens of the society, and it is inevitable that publishing ethical misconduct occurs under multiple value standards [5-6]. In this context, how to

identify these types of differentiated patterns and reveal their inner mechanisms has become a theoretical proposition that needs to be answered urgently for the development of convergent publishing.

Academic attention to convergent publishing has gone through an evolutionary process from phenomenal description to mechanism exploration. Early research mainly centered on the necessity and feasibility of digital transformation, focusing on the strategic significance of traditional publishing to embrace new technologies [7]. With the in-depth advancement of convergence practice, the research focus has gradually shifted to the exploration of specific paths and strategies, and started to pay attention to operational issues such as the reengineering of the content production process and the integration of channels and platforms, etc. Chen and Sun pointed out that the convergence publishing model refers to the new publishing paradigm of the in-depth integration of publishing theories and practices driven by digital technologies such as 5G, big data, AI, blockchain and so on, and that the key lies in breaking the academic barriers between theory and practice, and forming a new publishing paradigm of the theory and practice. The key lies in breaking the academic barriers between theory and practice, forming a disciplinary ecology in which theory guides practice and practice feeds theory, and building a more cohesive publishing academic community [8]. Leminen et al. constructed an integrative framework for analyzing the evolution of convergent business models in the Finnish book, newspaper, and magazine publishing industries based on the dimensions of “organizational change goals” and “belief system strength” [9]. Besancenot and Vranceanu pointed out that open access is no longer just a payment arrangement, but has evolved into a market signaling device for academic quality, where publishers can influence the “selection-separation” equilibrium of papers of different qualities through pricing strategies, thus reshaping the ecology and competitive landscape of academic journals [10]. Tondi analyzes the linkages and complementarities between traditional book trade models and new publishing models, and explores the mechanisms by which their hybrid business models operate in contemporary cultural production arenas; traditional commercial publishers, Unbound, and the broader cultural ecosystem are not confined to a direct competitive relationship, but rather work together to maintain the dissemination and viability of books through market symbiosis and functional complementarity [11]. Silva and Borges look at the definition and publishing domain of technical books to analyze their formal structure, functional operation and design methodology in a convergent publishing context. The study focuses on the design procedures and methodologies of technical books in a hybrid model, exploring possible models for their production, distribution and reading segments [12]. Pieterse cites the University of South Africa's Unisa Press as a prime example of a publisher that is exploring convergence pathways by simultaneously running multiple sustainable economic publishing models within a single publishing organization and incorporating existing book and journal publishing processes and partnerships [13].

In addition, the lack of clarity in the formulation structure of the publishing ethics norms themselves, and the lack of norms arising from the incomplete implementation mechanism have also indirectly triggered the conflict of values in the publishing ethics normative system [14]. Regarding the research on the regulatory mechanism of publishing ethics in the digital and intelligent era, Rahimi and Abadi pointed out that priority should be given to establishing applicable ethical principles and usage rules. While leveraging the AI's assistance in scientific research writing, one should adhere to the ethical norms and academic integrity mechanisms of scientific publishing. The core of publishing ethics governance lies in establishing a transparent framework for "human-machine collaboration" to ensure the coordinated evolution of technological innovation and academic ethics [15]. Hosseini et al. point out that the use of generative AI (e.g., ChatGPT and Big Language Models) in the ethical regulation of publishing in the age of AI raises ethical dilemmas about responsible authorship and the attribution of moral responsibility [16]. Kocak proposed a publishing ethics supervision mechanism based on a three-layer framework of "preventive disclosure + negative list + post-event accountability", aiming to maintain the integrity, traceability and responsibility chain of scientific literature in the context of generative AI deeply integrating into academic creation. This approach emphasizes the dynamic adaptation of technological application and ethical supervision, providing systematic publication ethics operation guidelines for publishers, editors, reviewers and authors [17]. Da Veiga proposed the distinctiveness principle for AI research ethics guidelines. This regulatory framework emphasizes the principle that "the human-machine responsibility chain must not be broken". By differentiating usage scenarios, clarifying the human-led responsibility and implementing a mandatory transparency disclosure mechanism, it seeks a synergistic path between promoting technological innovation and maintaining scientific integrity, and promotes the integration of publishing ethics norms from fragmentation [18]. Kasani et al. emphasize the importance of establishing dynamic, adaptive institutional arrangements that are in sync with ethical standards while AI technology is being rapidly deployed, to promote the transformation of AI chatbots from a source of ethical risk, to an aid that truly

enhances the value of scholarly publishing [19]. Moy points out that authors are obliged to safeguard the accuracy and completeness of citations. This mechanism embodies the ethical framework of "unbreakable human-computer responsibility" and "transparent disclosure obligation", which requires humans to use AI tools in a limited and traceable manner while retaining full academic responsibility to protect the authenticity, originality, and verifiability of academic records [20].

Facing the ethics of computer-aided publishing, this paper proposes a text mining framework that includes the processes of text acquisition-preprocessing-text mining-visualization-evaluation system, as well as the corresponding technical methods. The relevant text data for the regulation of computer-aided publishing ethics are acquired, and the TF-IDF is used to calculate the text similarity and realize the association of different text features. Calculate the average value of each topic cluster of assisted publishing based on K-Means to obtain the best clustering center. Use LDA topic model combined with plain Bayes to mine the topics of computer-aided publishing related texts. Discuss different publishing ethical regulatory mechanisms based on text mining results combined with computer-aided publishing process.

2. Text Mining for Ethical Regulation of Computer-Assisted Publishing

Through text mining, the ethics of computer-aided publishing is analyzed and the corresponding regulatory mechanism is proposed.

(1) Text acquisition is mainly obtained through web resources, and the crawler program grabs the needed network information to obtain text data.

(2) Preprocessing Tasks Preprocessing is a preprocessing operation on text. Text belongs to the computer unrecognizable language, which needs to be processed into machine-recognizable form and then core mining. The preprocessing task mainly includes text representation and feature selection and extraction.

(3) Core Mining Operations Core mining is mainly performed on the preprocessed text data in the core mining operations such as classification, clustering and extraction of textual information, relationships, and correlation analysis between texts.

(4) Results Display The display form of mining results includes the graphical user interface used by this system, the search and analysis interface and the analysis results display interface and the visualization tools used by the system.

(5) Evaluation system mining results assessment, you can evaluate the system through the evaluation system for a review, and then amend and improve the mining algorithm.

2.1. Text similarity calculation

In the application fields of text management, text clustering and information retrieval, it is based on similarity analysis, and the calculation results of similarity have a direct impact on text mining results. Similarity coefficient is an evaluation index that describes the similarity of sample points, described by 0~1, the larger the similarity coefficient the more similar the sample.

Structured data describes the similarity of unstructured text, which can make the analysis results more readable. Currently calculating the similarity of unstructured documents, typical similarity coefficients for calculating similarity methods are: the Jaccard coefficient that calculates the similarity according to the set model, the Dice coefficient, and the CSM that uses the vector space model.

In similarity computation, a document generally consists of mutually independent keywords $\{t_1, t_2, \cdots\cdots, t_m\}$, such that $D = (d_1, d_2, \cdots, d_n)$ denotes the document set consisting of n documents, where $d_i = (t_{i1}, t_{i2}, \cdots\cdots, t_{im})$ denotes the i th document vector consisting of m keywords, and the i th document vector's feature vector can be expressed as shown in Eq. (1):

$$V(d_i) = (t_{i1}, w_1(d_i); t_{i2}, w_2(d_i); \cdots; t_{im}, w_n(d_i)) \tag{1}$$

where d_i denotes the i th document, t_{ij} denotes the j th keyword in the i th document, and $w_j(d_i)$ denotes a function of the frequency of occurrence of the j th keyword in the i th document $tf(d_i)$, i.e. $w_j(d_i) = \varphi(tf(d_i))$.

The common methods for the calculation of the φ function are the square root function shown in Equation (2), the logarithmic function shown in Equation (3), and the TF-IDF function shown in Equation (4):

$$\varphi = \sqrt{tf(d_i)} \tag{2}$$

$$\varphi = \log\left(tf(d_i)+1\right) \tag{3}$$

$$\varphi = tf(d_i)\times \log\frac{N}{n_i} \tag{4}$$

The total number of documents in TF-IDF is denoted by N, and the number of documents containing the lemma t_i is n_i.

There are similarity calculation methods such as inner product function, cosine similarity, etc. This system adopts the TF-IDF function to process the data and applies the vector space modeling method to map a document into a feature vector $V(d)=\left(t_1,w_1(d);t_2,w_2(d);\cdots;t_n,w_n(d)\right)$, where $t_i\left(i=1,2,\cdots n\right)$ is a column of lexical items that are not the same as each other, and $w_i(d)$ is the weight of t_i in d, which is defined as the frequency of occurrence of t_i in d. $tf_i(d)$ as a function of $W_i(d)=\psi tf_i(d)$. The cosine similarity calculation is defined as equation (5):

$$S_T(d_i,d_j)=\cos(T_i,T_j)=\frac{\sum_{k=1}^{n} w_k(d_i)w_k(d_j)}{\sqrt{\sum_{k=1}^{n} w_k(d_i)^2}\sqrt{\sum_{k=1}^{n} w_k(d_j)^2}} \tag{5}$$

2.2. K-Means Clustering

K-Means is an algorithm that is based on distance and classifies or groups objects based on attribute features into K clusters, belongs to the unsupervised mode and defines the prototypical clustering algorithm with center of mass [21]. The number of clusters and the initial clustering center need to be specified for clustering, and an iterative approach is used to achieve better clustering results.

The flowchart of the algorithm is shown in Figure 1.The basic principle of K-Means: when clustering, it is necessary to randomly select K centers as the center of the cluster, at this time, calculate the distance of all the objects except K centers to each center, compare and select the nearest center as the center of clustering, and then calculate the average value of each cluster as a new clustering center, and cycle until the center of the clusters no longer change, the clustering is over. Clustering ends.

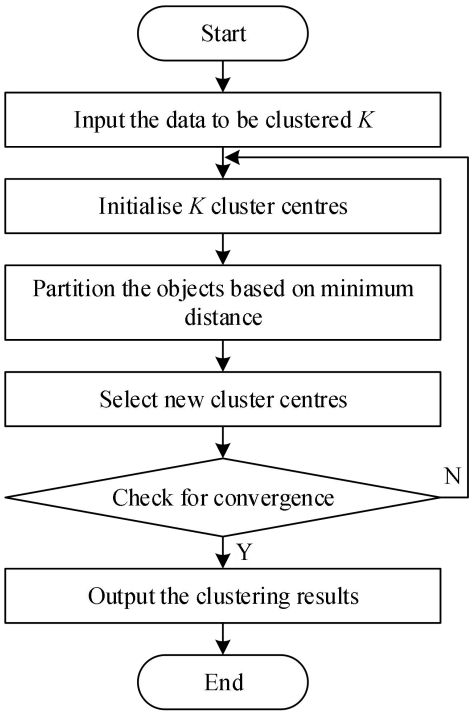

Figure 1. The process of the k-means algorithm

In clustering, the optimal number of clusters needs to be determined in order to optimize the clustering effect. Among the evaluation methods, the elbow method can analyze the optimal number of

clusters, and its core metric is the sum of squared errors SSE, as shown in Equation (6):

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \left| x - \overline{x}_i \right|^2 \tag{6}$$

The SSE in the formula is the sum of the error sum of squares for each sample, where C_i is the i th cluster, x denotes a given data object and is a sample point in C_i, and $\overline{x_i}$ is the center of mass of cluster C_i. The degree of sample delineation increases with the number of clusters K, the sample delineation will be more accurate, the higher the degree of aggregation will be, and the SSE value will be smaller and smaller.

The relationship between SSE and K in the elbow method: if K is smaller than the true number of clusters, the value of K is increasing SSE decline becomes larger; when the real value of the clusters is reached, as K increases, the SSE decreases, and finally tends to level off. At this point will appear similar to the shape of an elbow, the elbow position of the elbow corresponding to the k value is the optimal number of clusters of the sample data.

2.3. LDA topic modeling

LDA model is built on the basis of Dirichlet distribution and probabilistic graphical model, which is an unsupervised learning method based on generative modeling [22]. LDA can be regarded as a generative model, which assumes that each textual data is composed of a mixture of multiple topics, and that each topic in turn consists of multiple words. Specifically, LDA assumes that when generating a text, first, a topic is selected in the topic distribution, then a word is selected in the word distribution corresponding to that topic, and this process is repeated to generate each word and finally a text.The main idea of LDA is to infer the topic distribution of each text and the word distribution of each topic through this generative process.

In the LDA model, the generation process of a document is as follows:

Generate a random topic distribution for document i from the Dirichlet distribution.

For each word in the document i, a topic from the topic distribution is randomly selected as the topic for that word.

For each topic, generate a random distribution of words for that topic from the Dirichlet distribution.

For each word, a word is randomly generated from the distribution of words for its corresponding topic.

The specific analysis process of the LDA topic clustering model is shown in Figure 2.

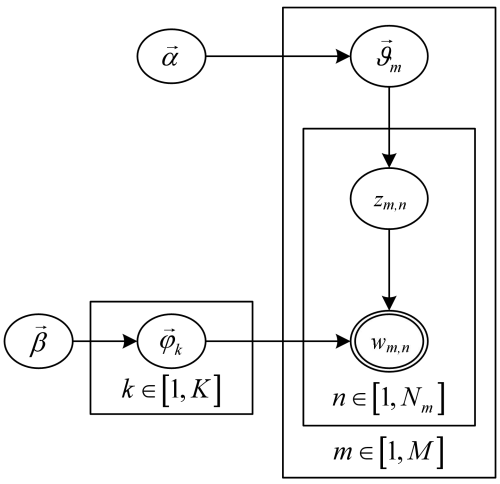

Figure 2. LDA Model Structure

In the above process, the LDA model uses the Dirichlet and polynomial distributions as conjugate prior probability distributions, similar to the Beta distribution which is the conjugate prior probability distribution of the binomial distribution.The graphical model structure of the LDA model is similar to the Bayesian network structure.

The process of LDA outputting the subject distribution is as follows:

(1) Determine the optimal number of subjects k based on the degree of confusion.

(2) Determine the parameters α and β that obey the Dirichlet distribution.

(3) Determine the lexical distribution ϕ_k of the number of topics k of the text layer from the Dirichlet distribution of the parameter β.

After determining the lexical distribution of the text layer, the topic distribution θ is determined for each comment text in a Dirichlet distribution with the parameter α, thus determining the topic Z, and iterating repeatedly over all the utterances of the whole text layer, thus generating all the lexicals of the document W.

LDA models need to be evaluated for their strengths and weaknesses after they are built and optimized for model parameters or other issues. In this paper, perplexity is chosen to evaluate the quality of LDA model and to determine the optimal number of topics. Perplexity is a measure used to evaluate the effectiveness of a model or probability distribution in predicting outcomes. In LDA modeling, the perplexity degree can reflect an evaluation of the effectiveness of learning the distribution of topic word mappings. The lower the perplexity degree, the higher the probability on the sentence, the better the overall performance of the topic grouping model, and the better the ability to predict new text.

The formula for calculating the perplexity degree is as follows:

$$Perplexity(D) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{7}$$

Applied to the LDA model, the $p(w_d)$ value is:

$$p(w) = \sum p(z \mid d)\, p(w \mid z) \tag{8}$$

In this equation, D is the test set in the corpus, and there are a total of M documents, N_d represents the number of words in each text d, w_d represents the text of d a word, $p(w_d)$ is the probability of the document w_d produced.

The proposed LDA model provides new ideas and methods for text data mining, which can not only effectively realize the tasks of text classification, clustering and retrieval, but also be used in the application areas of sentiment analysis, intelligent push, and public opinion analysis.

2.4. Plain Bayes

Plain Bayesian classifier is a classification algorithm commonly used for text categorization, the core idea of which is to compute a posteriori probabilities based on Bayes' theorem and the assumption of independence between features to achieve classification [23]. Plain Bayesian classifiers are usually implemented using polynomial models, Gaussian models, or Python or Numpy based programs. When dealing with discrete features, smoothing is required to avoid situations where the probability is zero, and then the polynomial model is used for classification.

The advantages of plain Bayes are fast convergence, easy to achieve the classification goal, high classification accuracy, fast classification speed, not only for large multi-classification tasks, but also for incremental training, and not easy to be affected by missing data. However, the disadvantage is that the classification performance is not guaranteed because the assumption of conditional independence about features is not always satisfied; it cannot learn the interactions between features. It exhibits high sensitivity to the representation of the input data, such as discrete, continuous and very small values, which directly affects the classification results:

$$precision = TP / (TP + FP) \tag{9}$$

$$recall = TP / (TP + FN) \tag{10}$$

$$F1 = 2PR / (P + R) \tag{11}$$

The constructed plain Bayesian classifier can be evaluated by checking the full rate (9), checking the accuracy rate (10), and F1 value (11). Where the check accuracy rate indicates the proportion of correctly predicted samples to the total samples; the check accuracy rate indicates how many of the samples predicted as correct samples are really correct samples; and the F1 value is the reconciled average of the check accuracy rate and the check rate, which is the weighted average of the check

accuracy rate and the check rate, where both the check accuracy rate and the check rate have the same weights. Selecting a larger F1 value as the threshold can improve the performance of the classifier.

3. Regulatory Mechanisms for Text-Mining Based Computer-Assisted Publishing

3.1. Computer-aided publishing process under convergent publishing

Through the previous computer-aided publishing and traditional publishing and a large number of comparative analysis between the process, proposed as shown in Figure 3 based on comparative analysis of the computer-aided publishing total process. Figure in the solid line box in the link is a computer-aided publishing unit must complete their own work, the dotted line box link can be entrusted to another service unit or individual to complete the work. The following to refine each step in the process.

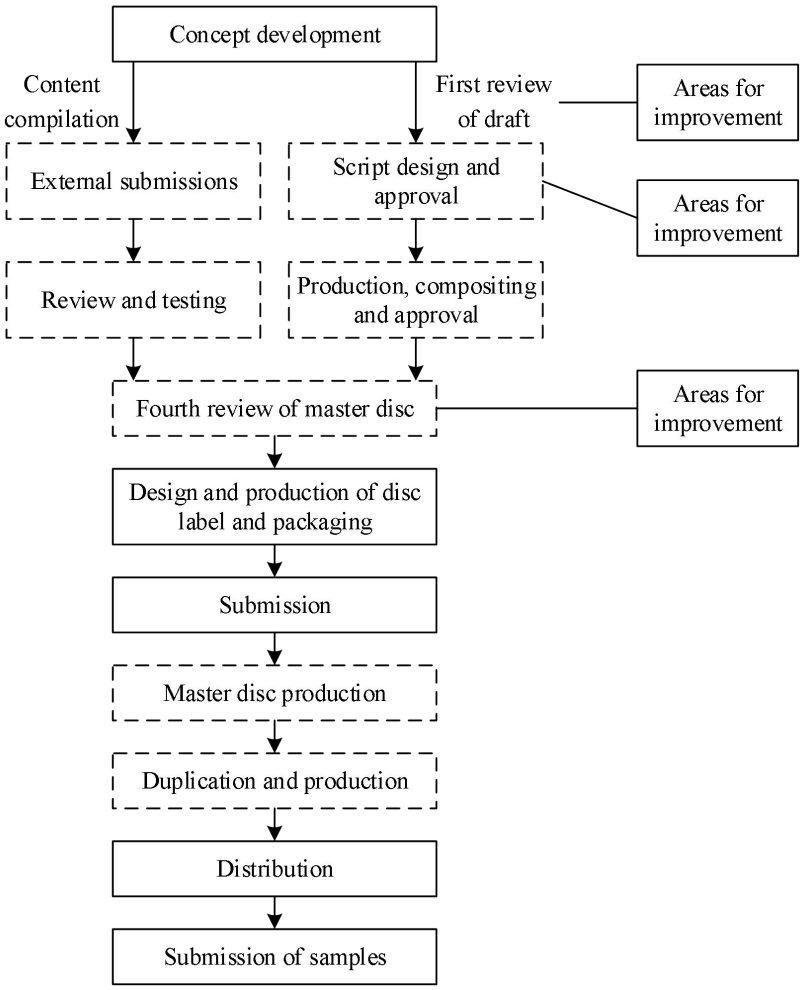

Figure 3. Electronic Publishing Process Flow Based on Comparative Analysis Method

3.2. Lapses in publishing ethics

This development from cultural units to enterprise restructuring has posed a new challenge to the survival and development of the publishing industry. In the past, the publishing industry as a single cultural unit has been unable to meet the requirements of the new market environment, and in order to adapt to the new market and cater to the needs of readers, the change of the publishing industry is bound to bring about a lot of problems and contradictions, among which, the failure of the publishing ethics is a serious problem. For example, many practitioners in the publishing industry have abused their rights, disregarded social morality, pursued personal fame and fortune, and even caused a very bad social impact on the society, and also hindered the healthy and orderly development of the publishing industry.

Ethical and moral misconduct means that the moral values and the system of ethical principles, which are the meaning of existence and the norms of life in social life, are either missing or ineffective, and that they cannot play a normal role in regulating and guiding social life and people's personal life, thus manifesting themselves in uncontrollable, disordered and chaotic social life and personal life.

4. Convergent Publishing's Ethical Mining of Computer-Assisted Publishing Regulation

4.1. Similarity distribution of computer-aided publishing terms

Traditional text similarity calculation models: models such as Word2vec do not fully consider semantic information, and lexical and positional information are not referenced. In order to get the similarity between texts more accurately, according to the word vectors trained by the BERT model, the cosine distance is used to set the threshold, the focused word pairs and non-focused word pairs are separated, and the similarity between them is calculated separately, and finally the text similarity is calculated by linear weighting. The composite lexical distribution and lexical weights are shown in Table 1, the average of the TF-IDF weights in the set of focused word pairs and the set of non-focused word pairs in the corresponding text, of which the most lexical in computer-assisted publishing is the verb, which accounts for 73.62%.

Table 1. Distribution of Compound Parts of Speech

Part of Speech	Percentage (%)	Characteristics	Weight
Verb	73.62	Verb	0.72
Adjective	13.57		
Noun	8.36	Noun	0.24
Adverb	2.25		
Preposition	1.04	Adverb	0.1
Conjunction	0.71		
Exclamatory	0.45	Adjective	0.22

The BERT model as a whole contains a large scale of parameters, and if the represented text is too long, it is difficult to train the model effectively. This paper, however, mainly applies word vectors to short texts such as evaluations, so the parameter size of the model can be effectively ensured. Because of this, when training the BERT model, while combining the open-source training model, the pre-training model parameters open-sourced by other researchers are used, which effectively reduces the consumption of resources and the difficulty of the experiment. The results of similarity calculation are shown in Table 2, which demonstrates the example training results of word vectors of the experimental corpus, in which the average value of similarity of near-synonyms under the publication keyword is 0.864.

Table 2. Similarity calculation results

Key words	Similarity			
Publishing	Publish a book	Printing	Distribution	Editor
	0.879	0.864	0.836	0.877
Media	Media	Radio	Newspapers	Magazines
	0.854	0.942	0.821	0.814
Books	Communication	Manuscript	Reading	Bookishness
	0.905	0.869	0.894	0.864

4.2. Computer-Aided Publishing Feature Clustering

In the field of computer-aided publishing, the keywords were cleaned and a total of 2495 keywords were obtained, and when calculating the threshold of high-frequency keywords, it was found that after data preprocessing, this paper selects the keywords with a word frequency greater than or equal to 5 as high-frequency keywords. The computer-aided publishing keywords are shown in Table 3. Among them, the first 13 high-frequency words are calculated to have a cumulative percentage of 31.01%, which can represent to a large extent the hot topic research of librarianship for the citation of journalism and communication.

Table 3. Computer-assisted publishing keywords

Key fields	Occurrence frequency	Percentage (%)	Cumulative percentage
Library (W1)	89	3.57	3.57
Open access (W2)	79	3.17	6.74
Reading promotion (W3)	73	2.93	9.67
University libraries (W4)	68	2.73	12.4
Public libraries (W5)	63	2.53	14.93
Digital libraries (W6)	62	2.48	17.41
Digital reading (W7)	56	2.24	19.65
Copyright (W8)	53	2.12	21.77
Information behavior (W9)	53	2.12	23.89
Information resources Journal (W10)	51	2.04	25.93
Mass reading (W11)	48	1.92	27.85
Open access journals (W12)	46	1.84	29.69
Knowledge services (W13)	33	1.32	31.01

In the correlation matrix, too many values of 0 are easy to cause too much error in the statistics, so 1 is used to subtract with all the values on the correlation matrix to get the dissimilarity matrix that indicates the degree of dissimilarity between the two words, which can be used to carry out the systematic clustering in SPSS, and the dissimilarity matrix of the computer-aided publication of the previous 13 keywords, for example, is shown in Fig. 4. Among them, the dissimilarity similarity of the same keywords takes the value of 0, and the average dissimilarity similarity of the 13 keywords is 0.73>0.7, the dissimilarity similarity is relatively high, and it is suitable for keyword clustering analysis.

	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13
W1	0.00												
W2	1.54	0.00											
W3	0.07	1.30	0.00										
W4	1.02	1.46	0.59	0.00									
W5	1.63	0.55	1.34	1.58	0.00								
W6	1.16	1.37	0.33	1.43	1.15	0.00							
W7	1.78	0.64	0.65	0.27	0.33	0.62	0.00						
W8	0.45	0.07	0.43	0.28	0.51	1.51	1.02	0.00					
W9	1.54	0.21	1.70	0.89	0.95	0.62	1.12	0.69	0.00				
W10	0.53	0.95	0.74	0.22	0.15	1.77	0.34	1.47	1.21	0.00			
W11	0.10	0.51	0.73	0.71	0.51	0.68	1.57	1.25	1.03	0.17	0.00		
W12	1.72	0.43	0.75	1.43	1.26	0.80	0.89	0.10	0.10	0.79	1.01	0.00	
W13	1.40	0.52	0.73	1.20	1.77	0.56	0.23	0.86	0.41	1.35	1.42	0.08	0.00

Figure 4. The dissimilarity matrix of computer-assisted publishing

In this paper, we choose inter-group linkage with Euclidean square distance for systematic clustering analysis, and then determine the central keyword of each category according to the adhesion, analyze the semantic relationship of keyword combination, and name each class group. The keyword clustering tree of computer-aided publishing is shown in Figure 5. The maximum distance of the dendrogram is 1.15 and approximates the cohesion process on the graph.

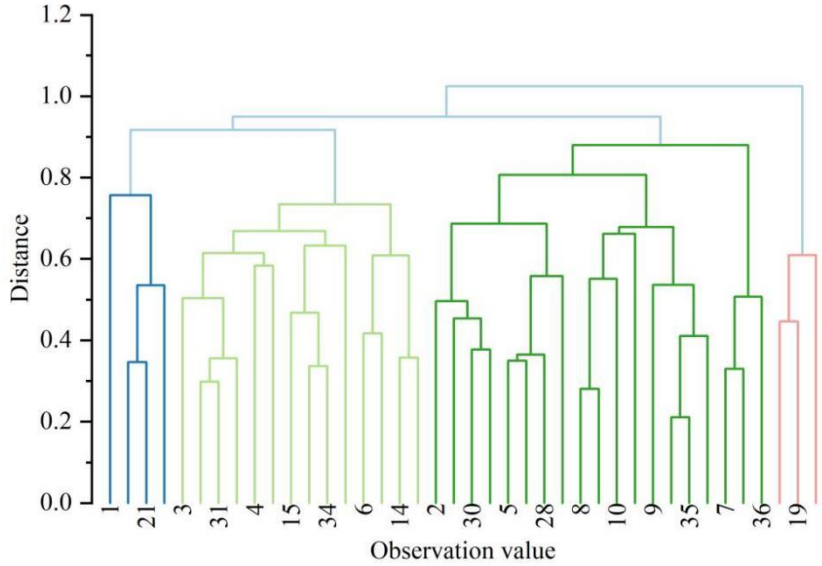

Figure 5. Keyword clustering tree of computer-aided publishing

4.3. Keyword LDA Topic Extraction

The general trend of confusion is to decrease with the increase of the number of topics, and the optimal number of topics for computer-aided publishing is determined to be 8 by the confusion analysis.The probability distribution of topic-word items for computer-aided publishing is shown in Table 4. Each column represents the probability that the topic-word belongs to this topic, and there are several rows for generating several topics. each row in theta file represents the document, and each column represents the probability that this document belongs to the topic, and there are several columns for generating several topics. In the field of computer-aided publishing, the probability distribution of all topic-word items reaches 0.01675 on average, and the extracted topic features are more reasonable.
The ethics of computer-aided publishing concentrates on the essential features of computer-aided publishing ethics, which are the fundamental principles and standards to be followed in adjusting all kinds of interpersonal relationships in computer-aided publishing activities, throughout the computer-aided publishing activities, with the significance and role of overarching and leading. The establishment of computer-aided publishing ethical and moral principles based on three things: first, to be conducive to the sound and perfect socialist moral system. Second, to ensure the health, order and prosperity of the network publishing industry. Finally, it should be conducive to the formation of a new type of computer-aided publishing ethics.

Table 4. Computer-assisted publishing topic - term probability distribution

Topic	1	2	3	4	5	6	7	8
W1	0.0157	0.0262	0.0232	0.026	0.0096	0.0172	0.0073	0.0289
W2	0.0080	0.0056	0.0243	0.0012	0.0184	0.0014	0.0260	0.0082
W3	0.0014	0.0192	0.0119	0.0162	0.026	0.0141	0.0207	0.0285
W4	0.0166	0.0222	0.0117	0.0166	0.0135	0.0144	0.0223	0.0112
W5	0.0011	0.0278	0.016	0.0287	0.0263	0.0114	0.0147	0.0073
W6	0.0116	0.0298	0.0129	0.0246	0.0292	0.0186	0.0092	0.0157
W7	0.0270	0.0032	0.0266	0.0262	0.0294	0.0071	0.0254	0.0228
W8	0.0214	0.0276	0.0063	0.0038	0.0157	0.0035	0.0258	0.0279
W9	0.0128	0.0263	0.0158	0.0170	0.0229	0.0271	0.0276	0.0013
W10	0.0222	0.0015	0.0135	0.0193	0.0044	0.0189	0.0277	0.0144
W11	0.0171	0.0046	0.0036	0.0219	0.021	0.0141	0.0112	0.0165
W12	0.0168	0.028	0.0062	0.0095	0.0251	0.0205	0.0071	0.0284
W13	0.0104	0.023	0.0059	0.0142	0.0014	0.0200	0.0275	0.0166

5. Conclusion

In this paper, text mining methods such as text similarity calculation, K-Means clustering, LDA model and plain Bayes are comprehensively applied to deeply analyze the ethical regulation of computer-assisted publishing under the convergent publishing model, and relevant ethical texts are

collected for analysis. In the distribution of computer-assisted publishing word attributes, the proportion of verbs reaches 73.62%, and the average similarity of near-synonyms under the publishing domain is 0.864. High-frequency keywords are extracted, of which the cumulative proportion of the top 13 high-frequency words is 31.01%, which can effectively study the hot topics. The average dissimilarity similarity of the 13 key high-frequency words is 0.73, which is more than 0.7, which can effectively regulate the Computer-Assisted Publishing Ethics.

Funding

Open Fund Project of Shanghai Jiao Tong University - Deguit Press Joint Laboratory: Full-Chain Analysis of Retraction Factors in Academic Ethics and Dynamic Path Planning for Academic Misconduct Prevention (Project Number: STTU-DG-2024-003); 2024 Editor · Renhe Fund: Research on Strategies for Journal Editorial Departments to Assist Deceived Authors in 'Anti-Fraud' under the New Situation (Project Number: XBRH2024-002-022); Research Project of Chongqing Science and Technology Journal Editors Association in 2024: Research on the Path of Empowering Equal Dialogue between Editors and Authors of Medical Comprehensive Journals with New Quality Productivity; Project of the "Publishing Integration and Innovation Fund" of the China University Science and Technology Journal Research Association in 2025: Research on the Construction and Implementation Mechanism of the Collaborative Training Model for "AI + Editing" Compound Talents (Project Number: CUJS2025-CBRH-030).

References

1. Wong, R. (2023). Role of generative artificial intelligence in publishing. What is acceptable, what is not. The Journal of ExtraCorporeal Technology, 55(3), 103-104.
2. Ugwu, N. F., Igbinlade, A. S., Ochiaka, R. E., Ezeani, U. D., Okorie, N. C., Opele, J. K., ... & Ojobola, F. B. (2024). Clarifying Ethical Dilemmas in Using Artificial Intelligence in Research Writing: A Rapid Review. Higher Learning Research Communications, 14(2), 29-47.
3. Morton, B., Vercueil, A., Masekela, R., Heinz, E., Reimer, L., Saleh, S., ... & Oriyo, N. (2022). Consensus statement on measures to promote equitable authorship in the publication of research from international partnerships. Anaesthesia, 77(3), 264-276.
4. Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., ... & de Oliveira, N. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. Patterns, 4(10).
5. Zhou, J., Müller, H., Holzinger, A., & Chen, F. (2024). Ethical ChatGPT: Concerns, challenges, and commandments. Electronics, 13(17), 3417.
6. Mijwil, M. M., Hiran, K. K., Doshi, R., Dadhich, M., Al-Mistarehi, A. H., & Bala, I. (2023). ChatGPT and the future of academic integrity in the artificial intelligence era: A new frontier. Al-Salam Journal for Engineering and Technology, 2(2), 116-127.
7. Adams, C., Pente, P., Lemermeyer, G., & Rockwell, G. (2023). Ethical principles for artificial intelligence in K-12 education. Computers and Education: Artificial Intelligence, 4, 100131.
8. Chen, J., & Sun, Q. (2025). Review of the Current Situation and Future Pathways of the Integration of Publishing Theory and Practice. Science-Technology & Publication, 44(11), 116-124.
9. Leminen, S., Huhtala, J. P., Rajahonka, M., & Westerlund, M. (2016). Business model convergence and divergence in publishing industries. In Media Convergence Handbook-Vol. 1: Journalism, Broadcasting, and Social Media Aspects of Convergence (pp. 187-200). Berlin, Heidelberg: Springer Berlin Heidelberg.
10. Besancenot, D., & Vranceanu, R. (2017). A model of scholarly publishing with hybrid academic journals. Theory and Decision, 82(1), 131-150.
11. Tondi, F. (2017). Alternative Publishing Models in a Changing Cultural Landscape. LOGOS: The Journal of the World Book Community, 28(4).
12. Silva, A. C., & Borges, M. M. (2015, October). Hybrid publishing design methods for technical books. In Proceedings of the 3rd International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 411-417).

13. Pieterse, H. (2012). A Hybrid Model for Scholarly Publishing in South Africa: The Challenge to Engage with an Open Access Environment within a Global Arena. Information, Medium and Society, 9(4), 1.
14. Kasani, P. H., Cho, K. H., Jang, J. W., & Yun, C. H. (2024). Influence of artificial intelligence and chatbots on research integrity and publication ethics. Science Editing, 11(1), 12-25.
15. Qadhi, S. M., Alduais, A., Chaaban, Y., & Khraisheh, M. (2024). Generative AI, research ethics, and higher education research: Insights from a scientometric analysis. Information, 15(6), 325.
16. Rahimi, F., & Abadi, A. T. B. (2023). ChatGPT and publication ethics. Archives of medical research, 54(3), 272-274.
17. Hosseini, M., Resnik, D. B., & Holmes, K. (2023). The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. Research Ethics, 19(4), 449-465.
18. Kocak, Z. (2024). Publication ethics in the era of artificial intelligence. Journal of Korean Medical Science, 39(33).
19. da Veiga, A. (2025). Ethical guidelines for the use of generative artificial intelligence and artificial intelligence-assisted tools in scholarly publishing: a thematic analysis. Science Editing, 12(1), 28-34.
20. Moy, L. (2023). Guidelines for use of large language models by authors, reviewers, and editors: considerations for imaging journals. Radiology, 309(1), e239024.
21. Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences, 622, 178-210.
22. Yu, D., & Xiang, B. (2023). Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling. Expert systems with applications, 225, 120114.
23. Peretz, O., Koren, M., & Koren, O. (2024). Naive Bayes classifier–An ensemble procedure for recall and precision enrichment. Engineering Applications of Artificial Intelligence, 136, 108972.

About the Author

Yongguo Hu, Graduated from the Chongqing Medical University in 2012.Employed at Editorial Department of Health Medicine Research and Practice, Southwest University. Her research interests include Ethics of Scientific and Technological Publishing.

Li Dai, Graduated from the Zunyi Medical University in 2016.Employed at Editorial Department of Health Medicine Research and Practice, Southwest University. Her research interests include Ethics of Scientific and Technological Publishing.

Zonghui Wu, Graduated from the Chongqing Medical University in 2011.Employed at Editorial Department of Health Medicine Research and Practice of Southwest University, and Southwest University Hospital. Her research interests include Ethics of Scientific and Technological Publishing.

Wenjie Huang, Graduated from the Chongqing Medical University in 2016.Employed at Editorial Department of Health Medicine Research and Practice, Southwest University. Her research interests include Ethics of Scientific and Technological Publishing.

Jie Liu, Graduated from the Southwest University in 2022.Employed at Editorial Department of Health Medicine Research and Practice, Southwest University. Her research interests include Ethics of Scientific and Technological Publishing.

Yahui Li, Graduated from the Chongqing University of Posts and Telecommunications in 2014.Employed at Editorial Department of Computer Science. Her research interests include Ethics of Scientific and Technological Publishing.

Yongsong Yan, Graduated from the Chongqing University in 2017.Employed at Editorial Department of Nano Materials Science, Chongqing University. Her research interests include Ethics of Scientific and Technological Publishing.