A Comparative Study of ChatGPT-Generated and Scholar-Written Literature Abstracts: A Case Study in the Field of Information Resource Management

Authors: Zhang Qiang, Wang Xiaoran, Gao Ying, Zhou Hong

Date: 2023-08-28T00:00:00+00:00

Abstract

Purpose/Significance To investigate the similarities and differences between ChatGPT-generated and scholar-written Chinese academic paper abstracts, and analyze the disparities in their content characteristics, thereby providing insights for AI-generated academic paper detection and related research. Method/Process First, taking the field of information resource management as an example, we extracted 500 highly cited papers from each of library science, information science, and archival science from the past three years. Based on the obtained paper titles, we generated corresponding abstract texts using prompt-based methods with the ChatGPT tool to construct a dataset. Second, we employed nine machine learning and deep learning algorithms to classify and detect ChatGPT-generated versus scholar-written abstracts. Finally, we conducted a multi-angle analysis of their similarities and differences from the perspectives of text features, topic models, and ROUGE evaluation to reveal their distinct characteristics. Results/Conclusion Mainstream machine learning and deep learning algorithms trained on the dataset can effectively distinguish whether an abstract is AI-generated or scholar-written, with BERT and ERNIE achieving the best performance, while RF and Xgboost performed best among the machine learning algorithms. ChatGPT-generated abstracts contain more characters and sentences than scholar-written ones, with keywords primarily being templated transitional terms. While the text topics are largely similar between the two, differences exist in themes such as “disciplinary system” and “digital humanities.” Quantitative analysis using ROUGE and cosine similarity demonstrates that ChatGPT-generated abstracts exhibit an obvious phenomenon of being “similar in form” rather than “similar in spirit” compared to scholar-written abstracts.

Full Text

A Comparative Study of ChatGPT-Generated and Scholar-Written Literature Abstracts: A Case Study in Information Resource Management

Zhang Qiang[1], **Wang Xiaoran**[2], **Gao Ying**[1], **Zhou Hong**[3],[4] [1]School of Information Management, Central China Normal University, Wuhan 430079 [2]School of Computer and Information, Anhui Polytechnic University, Wuhu 241000 [3]School of Economics and Management, University of Chinese Academy of Sciences, Beijing 101190 [4]Wuhan Library and Intelligence Center, Chinese Academy of Sciences, Wuhan 430071

Abstract

[**Purpose/Significance**] This study investigates the similarities and differences between ChatGPT-generated abstracts and those written by scholars in Chinese academic papers, analyzing disparities in content characteristics to provide insights for AI-generated academic text detection and related research.

[**Method/Process**] Taking the field of information resource management as a case study, we first extracted 500 highly cited papers from each of three sub-disciplines—library science, information science, and archival science—published over the past three years. Using the paper titles obtained, we applied the ChatGPT tool through prompt engineering to generate corresponding abstract texts, thereby constructing a dataset. Second, we employed nine machine learning and deep learning algorithms to classify and detect whether abstracts were ChatGPT-generated or scholar-written. Finally, we conducted a multi-angle analysis of their similarities and differences from the perspectives of textual features, topic modeling, and ROUGE evaluation to reveal distinctions between the two types of abstracts.

[**Result/Conclusion**] The mainstream machine learning and deep learning algorithms trained on our dataset can effectively distinguish between AI-generated and scholar-written abstracts. ERNIE and BERT achieved the best performance, while among machine learning algorithms, RF and Xgboost performed optimally. ChatGPT-generated abstracts contain more characters and sentences than scholar-written ones, with keywords typically being templated transitional terms. While the textual themes largely overlap, differences exist in themes such as “disciplinary system” and “digital humanities.” Quantitative analysis using ROUGE and cosine similarity metrics indicates that ChatGPT-generated abstracts exhibit clear “resemblance in form” rather than “resemblance in substance” when compared to scholar-written abstracts.

Keywords: ChatGPT; Text Classification; Text Features; Paper Abstracts; Classification Number: G353

Introduction

At the end of 2022, ChatGPT emerged as the fastest-growing consumer application in history, marking generative artificial intelligence (Artificial Intelligence Generated Content, AIGC) as a new research hotspot in both academia and industry [1]. Generative AI represents a new generation of artificial intelligence that learns from large-scale corpora to generate new data, text, images, and other content. It has broad application prospects in natural language processing, image generation, machine translation, speech generation, artistic creation, and other fields, potentially triggering a new round of technological revolution and industrial restructuring [2].

As a conversational interactive application, ChatGPT represents the highest level of industrialization in AIGC, possessing strong natural language understanding and generation capabilities. It can comprehend user intentions and generate corresponding responses based on user prompts [3]. In the education sector, surveys indicate that 89% of university students in the United States are using ChatGPT to complete academic assignments, demonstrating that ChatGPT has already achieved the level of a junior researcher and can produce academic papers with complete formatting and fluent logic [4]. In the academic domain, some scholars have listed ChatGPT as a co-author, triggering disputes over the ownership of copyright for AI-generated content. In response to ChatGPT being listed as an author, *Nature* has added two new principles to its submission guidelines: large language models cannot be listed as paper authors, and papers using large language models must clearly state this in the methods or acknowledgments section [5]. Domestic academic journals, represented by *Library and Information Service*, have also explicitly stated in their submission policies that they do not accept academic papers with AI tools listed as authors [6]. These developments demonstrate that academic journals attach great importance to the academic ethics issues triggered by ChatGPT, and there is an urgent need for methods and standards to determine whether academic papers are AI-generated.

Based on the above analysis, it is particularly important to identify whether academic paper content is generated by AI tools like ChatGPT and to analyze the characteristic differences between AI-generated text and manually written content. Specifically, this study takes Chinese academic paper abstracts from the information resource management discipline—encompassing library science, information science, and archival science—as the research object, focusing on two main questions: (1) Can statistical machine learning and deep learning methods distinguish whether Chinese academic paper abstracts are written by scholars or generated by AI? (2) What are the similarities and differences in textual features between scholar-written academic abstracts and AI-generated ones? This research can provide references for evaluating the quality of AI-generated academic paper texts and assist journals in assessing the originality of academic papers. Additionally, by analyzing the content characteristics of human-written versus AI-generated Chinese academic paper abstracts, this

study explores the features, quality, and comparative advantages and disadvantages of AI-generated content, thereby promoting the rational use of AI tools in academic paper writing and academic publishing ethics.

1 Related Research

The official OpenAI website introduces the methodology behind the ChatGPT model, whose core technologies include Transformer-based pre-trained models, Reinforcement Learning from Human Feedback (RLHF), supervised fine-tuning, and reward models [7]. In essence, ChatGPT is optimized through supervised fine-tuning, reward models, and reinforcement learning after pre-training to generate reasonable and fluent conversational information, enabling it to possess human common sense and values while appropriately avoiding sensitive issues. Current research on ChatGPT mainly includes two aspects:

The first aspect concerns the impact and influence of ChatGPT on specific disciplines or fields. Chris and Richard [8] argue that ChatGPT-like generative AI technologies can accelerate scientific research and generate innovative hypotheses, thereby advancing knowledge development, but they express concerns regarding data bias, text ethics, and research reproducibility. Pawan et al. [9] propose a development path for generative AI's involvement in human resource management research, connecting it with various aspects of HR management processes, practices, relationships, and outcomes to explore future research directions in the field. Dai Ling et al. [10] contend that ChatGPT-like AI technologies break through barriers of time, space, and individuals, connecting various fields throughout history in learning networks, which is conducive to the digital transformation of the education industry and the reform of educational ecosystems, but also poses challenges to educational ethics and data security. In the field of information resource management, Lu Wei et al. [11] explore the impact of large language models like ChatGPT on information resource management from six aspects: supporting algorithms and technologies, information resource construction, information organization and retrieval, content security and evaluation, and human-computer intelligent interaction and collaboration. Zhang Zhixiong et al. [12] summarize the development history of generative AI and analyze its impact on literature and intelligence work from perspectives including data organization methods, knowledge service models, intelligence analysis methods, literature usage patterns, and workforce development. They argue that the essence of rapid generative AI development lies in enhanced knowledge acquisition capabilities, that high-value corpora form the foundation of generative AI, and that the literature and intelligence field, which manages domains containing high-value human knowledge, needs to actively adapt and develop in the era of generative AI. Brady D. et al. [13] outline the technical principles behind ChatGPT as a chatbot and discuss through interviews how ChatGPT shows great promise in library areas such as search and discovery, reference and information services, cataloging and metadata generation, and content creation, while still requiring vigilance regarding ethical issues like privacy and

bias. Cao Shujin et al. [14] explore the impact of generative AI on intelligence research from three perspectives—research questions, research data, and research paradigms—and analyze changes in intelligence practice work from four service levels, arguing that intelligence science should actively embrace the new generation of AI while maintaining an objective and 审视 attitude. Zhou Wenhuan [15] analyzes the current state of digitalization and intelligentization research in the archival field, concluding that ChatGPT has broad application prospects in archival text summarization, archival classification, intelligent archival information retrieval, archival knowledge Q&A, and archival preservation and security, which can improve the efficiency, accuracy, and intelligence level of archival management.

The second aspect concerns the performance and evaluation of ChatGPT in various text generation tasks. In 2023, OpenAI released performance data for GPT-4-based ChatGPT on various examination tasks, showing that it scored above 90% of humans on the U.S. bar exam. Zheng et al. [16], considering that ChatGPT's training dataset only extends to before 2021, evaluated its performance by repeatedly questioning ChatGPT about an academic paper not in its database, with results indicating current limitations. Fredricton [17] argues that before using ChatGPT for writing, one must know that it fabricates nonexistent citations, and that authors should not use ChatGPT-generated content in scientific writing. He also contends that journals need not require authors to declare ChatGPT's role in scientific writing, as ChatGPT, like thesauruses and grammar checkers, is merely a writing tool, and authors themselves must be responsible for their decisions. In the Chinese domain, Zhang Huaping et al. [18] compared ChatGPT with several existing pre-trained models, finding that ChatGPT already achieves high accuracy in Chinese sentiment analysis tasks but frequently makes factual errors in closed-book question answering. Bao Tong et al. [19] compared ChatGPT with multiple pre-trained models across several Chinese public datasets to analyze its performance in entity extraction, relation extraction, and event extraction, with results showing that ChatGPT performs better on event extraction tasks than on the other two. Shi Yilong et al. [20] explored the comparative advantages and disadvantages of ChatGPT versus highly-upvoted human answers on Zhihu regarding the same questions, finding that ChatGPT enables more convenient access to desired information and that its response text features are close to highly-upvoted human answers, but that answer quality varies significantly across different themes and is accompanied by false information.

In summary, beyond discussing the opportunities and challenges that ChatGPT-like generative AI tools bring to disciplines or fields, empirical studies evaluating their performance have begun to emerge. However, relevant research on Chinese corpora remains relatively insufficient, and studies analyzing the similarities and differences between text content generated by ChatGPT-like tools and manually written academic content in Chinese journals are far from adequate. This paper selects abstract texts from academic literature in library science, information science, and archival science under the information resource management

domain as the basic dataset, uses ChatGPT to generate corresponding abstracts based on literature titles, and explores ChatGPT's performance in generating Chinese academic papers while analyzing the differences between the two types of abstracts.

2 Research Design

To analyze the similarities and differences between scholar-written academic paper abstracts and ChatGPT-generated abstracts, this study designed a research framework as shown in Figure 1 [Figure 1: see original paper].

2.1 Data Sources and Processing

Figure 1 Research Framework

This study takes journal papers in the information resource management field as research objects, dividing them into three categories according to sub-disciplines: library science, information science, and archival science. Considering the representativeness of academic literature, we selected core journals as the data source. Additionally, considering that ChatGPT's training data was last updated in September 2021 (currently, GPT-4's training data cutoff is also September 2021), we ultimately determined to select 500 highly cited papers from each of the three sub-disciplines published between September 2018 and August 2021, totaling 1,500 papers as the basic research sample. For interdisciplinary journals such as *Library and Information Service*, we manually intervened in the classification and screening process. Specific source journal names and paper quantities are shown in Table 1 .

Table 1 Source Journal Names and Paper Quantities

After obtaining the relevant paper titles, we required ChatGPT to generate corresponding abstract texts. In the use of generative AI tools, the importance of Prompt is self-evident. Prompt is a language containing guiding information that enables the model to better understand and generate content. The quality of Prompt directly affects the model's output results. This study referenced the CRISPE framework to write Prompts, which breaks down the prompt creation process into clear, structured steps [21]. CR (Capacity and Role) represents the capability and role that the questioner wants ChatGPT to play. I (Insight) provides ChatGPT with background information and context to fully understand the background and requirements. S (Statement) represents the user's clearly defined task objectives so that ChatGPT can meet the user's response needs. P (Personality) represents the style in which the user wants ChatGPT to respond, which helps ChatGPT generate personalized content. E (Experiment) represents the user's request for ChatGPT to generate multiple examples and answers under rough search conditions, allowing users to compare and evaluate among diverse options.

The final Prompt we determined for requiring ChatGPT to generate academic

paper abstracts is: "Assume you are a researcher in library science/information science/archival science. You have conceived a new academic paper with the title 'XXX'. Please write an abstract for this title in the style of a Chinese researcher and according to the requirements of Chinese academic journals."

Due to current usage limits on GPT-4, this study used self-written Python code to call the GPT-3.5 API to batch obtain ChatGPT-generated abstract content. The calling code is shown in Figure 2 [Figure 2: see original paper].

Figure 2 Code for Calling ChatGPT to Generate Academic Text Abstracts

Finally, the 1,500 scholar-written abstracts and 1,500 ChatGPT-generated abstracts were saved as local Excel files. The data preprocessing procedure is described below:

1. **Domain Lexicon Construction**: We used the keywords from the 1,500 scholar-written papers as the initial domain lexicon, supplemented with conventional terms from the library, information, and archival fields. After manual screening, 1,376 words were finalized as the domain lexicon for subsequent word segmentation.
2. **Stopword List Construction**: To capture the textual features of ChatGPT-generated abstracts as comprehensively as possible, we only considered adding punctuation marks and meaningless function words to the stopword list.
3. **Text Segmentation**: We used self-written Python code to call the LTP natural language processing package, loading the domain lexicon and stopword list to segment the abstract texts.
4. **Classification Labeling**: ChatGPT-generated abstracts and scholar-written abstracts were labeled as 0 and 1, respectively.

2.2 Text Classification Annotation

The research objective of this study is to explore whether current mainstream machine learning and deep learning algorithms can distinguish between ChatGPT-generated and scholar-written academic paper abstracts and identify their differences. From a coarse-grained perspective, this problem can be transformed into a classic binary classification problem. After vectorizing the text representation using the TF-IDF method, we conducted classification experiments using seven common machine learning classification algorithms (SVM, NB, K-Nearest Neighbors, Decision Tree, Logistic Regression, Random Forest, Xgboost) and two deep pre-trained language models (BERT and ERNIE). We selected Accuracy, Precision, Recall, and F1-Score as evaluation metrics. Following conventional dataset partitioning standards in machine learning, 70% of the text dataset was used as the training set to train classification models, while the remaining 30% was used as the test set to evaluate

model performance.

2.3 Text Content Analysis

In addition to classifying and identifying ChatGPT-generated versus scholar-written academic paper abstracts, this study also employed text feature detection, topic model consistency detection, and ROUGE evaluation to interpret the differences between the two from the content level.

Text feature detection primarily includes character-level, word-level, and sentence-level feature detection. Specifically, character-level analysis examines differences in abstract length between ChatGPT-generated and scholar-written abstracts; word-level analysis compares high-frequency keywords; and sentence-level analysis examines differences in the number of sentences in abstracts. Textual features typically reflect the core characteristics of a text, and analyzing them from the perspectives of characters, words, and sentences facilitates comparison of differences in key concepts, academic terminology, and language expression.

Topic model consistency detection primarily uses the LDA topic model to compare topic distributions between the two types of abstracts. The LDA (Latent Dirichlet Allocation) model, proposed by Blei et al. in 2003, is a three-layer Bayesian network model comprising words, topics, and documents, used to discover latent topics in text data and assign documents to these topics. Using the LDA topic model can identify thematic differences between ChatGPT-generated and scholar-written academic paper abstracts, thereby revealing their content-level similarities and differences. The document generation process in the LDA topic model is shown in Figure 3 [Figure 3: see original paper].

Figure 3 LDA Topic Model Diagram

Step 1: Select a document according to prior probability
Step 2: Sample to generate document topics from Dirichlet distribution α
Step 3: Sample to generate document topic distribution from topic distribution
Step 4: Sample to generate the j-th word's topic from the topic distribution of document i
Step 5: Sample to generate word j from the word distribution corresponding to topic z_{ij}

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of automatic evaluation metrics used to measure the similarity between generated text summaries or machine translation results and reference summaries. It is widely used in evaluating summary generation and machine translation tasks in natural language processing. ROUGE primarily focuses on recall rate, treating generated summaries and reference summaries as bag-of-words models and measuring their similarity by calculating the degree of word overlap—that is, determining how much content from the reference summary is included in the generated summary. Common ROUGE metrics include:

ROUGE-N: This metric calculates the recall rate of N-grams (continuous N words) between generated summaries and reference summaries. The calculation method is shown in Formula (1).

```
ROUGE-N = Count_{match}(N-gram) / Count_{reference}(N-gram)
```

Formula (1)

Where the denominator is the number of N-grams in scholar-written abstracts, and the numerator is the number of co-occurring N-grams between scholar-written and ChatGPT-generated abstracts. ROUGE-N is concise and has word order features, but its value drops sharply as N increases. Generally, ROUGE-1 and ROUGE-2 are used as evaluation metrics.

ROUGE-L: This metric calculates the recall rate of the Longest Common Subsequence (LCS). It measures long-distance dependencies and sequential consistency between generated summaries and reference summaries. The calculation method is shown in Formulas (2)-(4).

```
LCS_R = LCS(X,Y) / m
LCS_P = LCS(X,Y) / n
ROUGE-L = (1 + $\beta$$^{2}$) * LCS_R * LCS_P / (LCS_R + $\beta$$^{2}$ * LCS_P)
```

Formula (2) Formula (3) Formula (4)

Where LCS(X,Y) is the length of the longest common subsequence between X and Y, m and n represent the lengths of scholar-written abstracts and ChatGPT-generated abstracts respectively, LCS_R and LCS_P represent recall and precision rates, and β is used to balance their importance. ROUGE-L's characteristic is that it does not require specifying N-gram length like ROUGE-N, but it only considers the length of the longest subsequence, making it more suitable for evaluating short abstract extraction.

Additionally, this study employed cosine similarity to detect the similarity between ChatGPT-generated and scholar-written abstracts, thereby forming a comparison with ROUGE metric results.

3.1 Abstract Classification Results

Based on the binary classification procedure and nine classification models described above, this study conducted classification tests on ChatGPT-generated academic paper abstracts and scholar-written abstracts. The results are shown in Table 2 .

Table 2 Comparison of Classification Effects for ChatGPT-Generated and Scholar-Written Abstracts Across Different Models

As shown in Table 2, among the nine classification models, the deep learning-based ERNIE achieved the best classification effect, followed by the BERT model. The reason is that this study's research object is Chinese academic paper abstracts, and ERNIE is Baidu's further optimization based on the BERT

model specifically for Chinese NLP tasks. Additionally, among the seven machine learning classification models, except for NB and KNN, the overall F1-Scores of the remaining five machine learning classification models exceeded 90%, indicating they all achieved good classification effects, while NB and KNN performed poorly on this task. From the perspective of the three sub-disciplines, classification algorithms achieved lower F1-Scores in the archival science domain compared to library science and information science, suggesting that from a text classification perspective, ChatGPT-generated and scholar-written abstracts are more similar in the archival science field.

In text classification experiments, feature words are crucial for classification determination, and feature selection directly affects model classification effectiveness and capability. This study selected the machine learning classification models RF and Xgboost, which achieved F1-Scores exceeding 90% across all three sub-disciplinary domains, and analyzed the top 10 feature words for each algorithm. The results are shown in Table 3 .

Table 3 Top 10 Feature Words for RF and Xgboost Algorithms

As shown in Table 3, although the keywords and rankings differ between the two algorithms across the three domains, words such as “this paper,” “propose,” “discuss,” and “finally” appear in both algorithms across domains, indicating these words can effectively distinguish whether abstracts are ChatGPT-generated or scholar-written. More specifically, RF and Xgboost achieve relatively close F1-Scores in library science, while Xgboost outperforms RF in information science, and RF outperforms Xgboost in archival science. Examining the feature words reveals that in information science, words like “context,” “graph,” and “compute” have more important classification features, while in archival science, words like “significance,” “digitalization,” “importance,” and “preservation” have more important classification features.

3.2 Text Content Analysis

Beyond verifying whether machine learning models can distinguish between ChatGPT-generated and scholar-written academic abstracts, it is necessary to examine their similarities and differences from the text content level to understand their internal textual differences. This section explores the differences between the two from three aspects: text features, topic models, and ROUGE evaluation.

3.2.1 Text Feature Analysis

As a typical short text, academic abstracts reflect textual characteristics at the character, word, and sentence levels. Abstract length refers to the number of Chinese characters in an academic abstract. We statistically analyzed the abstract lengths of both types across the three sub-disciplines and plotted histograms, as shown in Figure 4 [Figure 4: see original paper].

Figure 4 Normal Distribution Fitted Histogram of Abstract Lengths

To intuitively display the statistical information of abstract lengths for ChatGPT-generated and scholar-written abstracts, we used mean, standard deviation, skewness, and kurtosis as statistical indicators, as shown in Table 4 .

Table 4 Statistical Information Table of Abstract Lengths

As shown in Figure 4 and Table 4, overall, the mean character count of ChatGPT-generated abstracts is 336.55, while that of scholar-written abstracts is 271.84, showing a relatively large difference. However, the difference in standard deviation is small, indicating similar data dispersion. The kurtosis gap is large, indicating a significant difference in peak sharpness, with scholar-written abstract character count distribution having a sharper peak. Looking at specific sub-disciplines, information science shows the smallest gap in mean values, while archival science shows the largest. For standard deviation, information science shows the smallest gap, archival science the largest. For skewness, information science shows the smallest gap, library science the largest. For kurtosis, information science shows the smallest gap, library science the largest.

Beyond examining differences at the character level, writing habits reflected through word usage can also indicate writing style differences. We extracted keywords from both types using TF-IDF and TextRank algorithms, with results shown in Table 5 .

Table 5 Keywords in ChatGPT-Generated and Scholar-Written Abstracts (TF-IDF and TextRank)

As shown in Table 5, overall, the keywords of the two types are relatively similar. However, ChatGPT-generated abstracts often contain transitional and structural words like “this paper,” “propose,” and “finally,” while scholar-written abstracts mostly contain substantive nouns or verbs. Examining the three subdisciplines, the keywords in the information science domain are more similar between the two types, while library science and archival science show more differences.

In addition to considering character and word perspectives, we also analyzed differences in sentence quantity between ChatGPT-generated and scholar-written abstracts, striving to dissect their similarities and differences from multiple angles. We statistically analyzed the sentence counts of both types across the three sub-disciplines and plotted histograms, as shown in Figure 5 [Figure 5: see original paper].

Figure 5 Normal Distribution Fitted Histogram of Abstract Sentence Counts

Similar to the analysis of abstract length statistics, to more intuitively display the statistical information of sentence counts, we continued using mean, standard deviation, skewness, and kurtosis as statistical indicators, as shown in Table 6 .

Table 6 Statistical Information Table of Abstract Sentence Counts

As shown in Figure 5 and Table 6, overall, ChatGPT-generated abstracts have nearly twice as many sentences as scholar-written abstracts, with higher standard deviation, indicating that sentence counts in ChatGPT-generated abstracts are more dispersed overall. The skewness of the two types is similar, indicating consistent distribution patterns. Scholar-written abstracts have greater kurtosis, indicating a sharper peak in their distribution. Looking at specific sub-disciplines, ChatGPT-generated abstracts exceed scholar-written abstracts in sentence count across all domains. For standard deviation, library science shows similar values, while information science and archival science show lower standard deviations for scholar-written abstracts. For skewness, library science and information science are relatively close, while archival science shows lower skewness for scholar-written abstracts. For kurtosis, library science and archival science show larger gaps between ChatGPT-generated and scholar-written abstracts, with scholar-written abstract sentence count kurtosis values exceeding 3, indicating sharper peaks in these two domains.

3.2.2 Topic Model Analysis

We conducted LDA topic model analysis on ChatGPT-generated and scholar-written academic abstracts to grasp their differences in textual themes. We used perplexity to measure the model, as shown in Figure 6 [Figure 6: see original paper]. Perplexity describes topic similarity; generally, lower perplexity values are better, but when the number of topics is too large, the model tends to be overfitted.

Based on the plotted perplexity-topic number line graphs for both types of abstracts and considering the optimal number of topics, we ultimately determined 9 as the optimal topic number.

Figure 6 Perplexity-Topic Number Line Graph for ChatGPT-Generated and Scholar-Written Abstracts

After determining the optimal topic number, we imported the preprocessed texts into the LDA topic model for training, obtained the "topic-word" distribution, and summarized the top 5 words by probability for each of the 9 topics, as shown in Table 7 .

Table 7 "Topic-Word" Distribution

As shown in Table 7, the topic distributions of ChatGPT-generated and scholar-written abstracts are relatively consistent, with themes such as "smart library," "data governance," "influencing factors," and "information literacy" appearing in both. The differing themes are mainly reflected in "disciplinary system," "emergency events," and "digital humanities," indicating greater stylistic differences in these themes.

3.2.3 ROUGE Evaluation

To quantitatively evaluate the similarity between ChatGPT-generated and scholar-written abstracts, we employed the most commonly used ROUGE-1, ROUGE-2, and ROUGE-L metrics for evaluating automatic text summarization, plus cosine similarity to detect their similarity. ROUGE primarily counts the number of overlapping basic units between the two, while cosine similarity mainly measures their similarity in overall direction. The results are shown in Table 8 .

Table 8 ROUGE and Cosine Similarity Evaluation Results

Due to current ChatGPT input character limitations and because ChatGPT itself is a generative AI tool whose main function is to generate responses based on user input, we could not compare it with current benchmark algorithms in ROUGE evaluation, as those algorithms generate summaries based on full texts. By comparing with the performance of current mainstream benchmark algorithms on public evaluation datasets [22,23], although ChatGPT's summary evaluation scores remain relatively low, its cosine similarity is high. This indicates that the generated content exhibits a "resemblance in form" phenomenon—that is, it appears similar on the surface but has low overlapping unit data. Among the three sub-disciplinary domains, information science shows the highest similarity between the two types, while archival science shows the lowest, indicating that compared to information science, ChatGPT-generated content differs more from scholar-written abstracts in the archival science domain.

Conclusion

This study takes highly cited papers from three sub-disciplines of the information resource management field in recent years as research objects. Based on obtained paper titles, we designed prompts to call the ChatGPT API to generate AI-written paper abstracts. Using nine machine learning and deep learning algorithms, we classified and identified the two types of abstracts and analyzed their differences from multiple textual content perspectives.

In terms of classification identification, mainstream machine learning or deep learning classification models can effectively identify whether abstract texts are ChatGPT-generated or scholar-written. Among the nine selected classification models, the two deep learning models ERNIE and BERT achieved the best classification results. Among machine learning algorithms, except for NB and KNN, the overall F1-Scores of the remaining five machine learning algorithms exceeded 90%. From the sub-disciplinary perspective, classification algorithms achieved lower F1-Scores in archival science abstract classification compared to library and information science.

In text feature analysis, at the character level, ChatGPT-generated abstracts have longer average lengths than scholar-written abstracts, with similar data dispersion but large kurtosis gaps. Specifically, information science shows the

smallest gaps across all metrics, while archival science shows the largest gaps in mean and standard deviation, library science shows the largest gap in skewness, and library science shows the largest gap in kurtosis. At the word level, the overall keywords are relatively similar, but ChatGPT-generated abstracts often accompany transitional words like “this paper,” “propose,” and “finally.” At the sentence level, ChatGPT-generated abstracts have nearly twice as many sentences as scholar-written abstracts, with higher standard deviation, similar skewness, and scholar-written abstracts showing greater kurtosis. Specifically, archival science shows the largest gap in mean values, archival science shows the largest gap in standard deviation, archival science shows the largest gap in skewness, and library science shows the largest gap in kurtosis.

In topic model analysis, the topic distributions of ChatGPT-generated and scholar-written abstracts are relatively consistent, with major differences in “disciplinary system,” “emergency events,” and “digital humanities.” In ROUGE evaluation, current ChatGPT-generated abstract evaluation scores are low across ROUGE-1, ROUGE-2, and ROUGE-L metrics, but cosine similarity is high, indicating a “resemblance in form but not in substance” phenomenon. Among the three sub-disciplinary domains, information science shows the highest scores, while archival science shows the lowest, indicating that ChatGPT-generated abstracts differ more from scholar-written abstracts in the archival science domain.

This study has several limitations. Due to current ChatGPT API constraints, we used GPT-3.5-based ChatGPT to generate abstracts rather than the latest GPT-4 model. Additionally, current ChatGPT has many limitations on input character count. Future research will examine introduction, main text, and conclusion sections to more comprehensively analyze differences between the two types. Furthermore, this study only used Chinese academic papers from the information resource management field as research objects; future research will consider comparative analysis across different disciplinary domains.

References

[1] Wang F Y, Li J, Qin R, et al. ChatGPT for Computational Social Systems: From Conversational Applications to Human-Oriented Operating Systems[J]. IEEE Transactions on Computational Social Systems, 2023, 10(2): 414-425.

[2] Mondal S, Das S, Vrana V G. How to Bell the Cat? A Theoretical Review of Generative Artificial Intelligence towards Digital Disruption in All Walks of Life[J]. Technologies, 2023, 11(2): 44.

[3] Cheung K S. Real Estate Insights Unleashing the potential of ChatGPT in property valuation reports: the “Red Book” compliance Chain-of-thought (CoT) prompt engineering[J]. Journal of Property Investment & Finance, 2023, ahead-of-print(ahead-of-print).

[4] Productive Teaching Tool or Innovative Cheating?[EB/OL]. [2023-08-28].

https://study.com/resources/perceptions-of-ChatGPT-in-schools.

[5] Initial submission | Nature[EB/OL]. [2023-08-28]. https://www.nature.com/nature/for-authors/initial-submission.

[6] Library and Information Service AI Policy Statement[EB/OL]. [2023-08-28]. https://www.lis.ac.cn/CN/column/column27.shtml.

[7] Introducing ChatGPT[EB/OL]. [2023-03-30]. https://openai.com/blog/ChatGPT.

[8] Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science[J]. Nature, 2023, 614(7947): 214-216.

[9] Budhwar P, Chowdhury S, Wood G, et al. Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT[J]. Human Resource Management Journal, 2023, 33(3): 606-659.

[10] Dai Ling, Hu Jiao, Zhu Zhiting. New Strategies for Empowering Educational Digital Transformation with ChatGPT[J]. Open Education Research, 2023, 29(4): 41-48.

[11] Lu Wei, Liu Jiawei, Ma Yongqiang, et al. The Impact of Large Models Represented by ChatGPT on Information Resource Management[J]. Library and Intelligence Knowledge, 2023, 40(2): 6-9+70.

[12] Zhang Zhixiong, Yu Gaihong, Liu Yi, et al. The Impact of ChatGPT on Literature and Intelligence Work[J]. Data Analysis and Knowledge Discovery, 2023, 7(3): 36-42.

[13] Lund B D, Wang T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries?[J]. Library Hi Tech News, 2023, 40(3): 26-29.

[14] Cao Shujin, Cao Ruye. The Impact of Generative AI on Intelligence Science Research and Practice from the Perspective of ChatGPT[J]. Modern Intelligence, 2023, 43(4): 3-10.

[15] Zhou Wenhuan. The Application and Significance of ChatGPT in the Archival Field[J]. China Archives, 2023(3): 62-63.

[16] Zheng H, Zhan H. ChatGPT in Scientific Writing: A Cautionary Tale[J]. The American Journal of Medicine, 2023, 136(8): 725-726.e6.

[17] ScientistSeesSquirrel. How to use ChatGPT in scientific writing[EB/OL]. (2023-06-20) [2023-08-11]. https://scientistseessquirrel.wordpress.com/2023/06/20/how-to-use-ChatGPT-in-scientific-writing/.

[18] Zhang Huaping, Li Linhan, Li Chunjin. ChatGPT Chinese Performance Evaluation and Risk Response[J]. Data Analysis and Knowledge Discovery, 2023, 7(3): 16-25.

[19] Bao Tong, Zhang Chengzhi. Evaluation of ChatGPT's Chinese Information Extraction Capability—Taking Three Typical Extraction Tasks as Examples[J]. Data Analysis and Knowledge Discovery: 1-16.

[20] Shi Yilong, Xu Xin. A Comparison of ChatGPT Machine Answers and Zhihu Human Answers[J]. Library Tribune: 1-10.

[21] Nigh M. ChatGPT3 Prompt Engineering[CP/OL]. (2023-08-12)[2023-08-12]. https://github.com/mattnigh/ChatGPT3-Free-Prompt-List.

[22] Wang Hongbin, Jin Ziling, Mao Cunli. Extractive Automatic Summarization of News Text Combining Hierarchical Attention[J]. Computer Science and Exploration, 2022, 16(4): 877-887.

[23] Zhao Jiangjiang, Wang Yang, Xu Yingying, et al. An Extractive Automatic Summarization Model Based on Knowledge Distillation[J]. Computer Science, 2023, 50(S1): 214-220.

Author Contributions

Zhang Qiang: Designed the research framework, conducted experimental analysis and processing, wrote and revised the initial draft.
Wang Xiaoran: Conducted experiments and analyzed results.
Gao Ying: Revised the paper and finalized the version.
Zhou Hong: Provided paper revision suggestions.

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