Research on Lexical Characteristics and Provenance Methods of AI-Generated Paper Abstracts: A Case Study of Information Resource Management (Postprint)

Authors: Wang Yibo, Wang Yihu

Date: 2026-04-01T17:39:51+00:00

Abstract

[Purpose/Significance] By comparatively analyzing the lexical characteristics and provenance methods of abstracts written by scholars versus those generated by artificial intelligence, this study provides a reference for the provenance detection of generative artificial intelligence in research paper abstracts.

[Method/Process] An experimental dataset was constructed by selecting 2,000 scholar-written abstracts published between 2018 and 2022 in 20 core journals within the field of Information Resource Management, along with 18,000 abstracts generated by 9 mainstream large language models from both domestic and international sources. On this basis, the lexical features of the two types of abstracts, such as unique words and common phrases, were comparatively analyzed. A multi-classification model was constructed to perform provenance detection, ultimately leading to the development of an AI-generated abstract detection assistant.

[Result/Conclusion] The research indicates that the lexical selection style of scholar-written abstracts is professional, rigorous, and closely aligned with the research topics. In contrast, AI-generated abstracts frequently utilize templated unique words and phrases. Compared to sentence-level detection, large models perform better at the text level, with the fine-tuned BERT model showing outstanding performance. The AI-generated abstract detection assistant developed based on the experimental data provides technical support for maintaining academic integrity.

Full Text

Research on the Lexical Characteristics and Provenance Detection of AI-Generated Research Abstracts: A Case Study in the Field of Information Resource Management

Wang Yibo[1], **Wang Yihu**[2] (1. Peking University Library, Beijing 100871; 2. Department of Information Management, Peking University, Beijing 100871)

Abstract

[**Purpose / Significance**] This study provides a reference for the provenance detection of generative artificial intelligence (AI) in the context of academic paper abstracts. By comparing and analyzing the lexical characteristics and provenance methods of abstracts authored by human scholars versus those generated by Large Language Models (LLMs), the study aims to provide a theoretical and empirical basis for identifying AI-generated content in academic publishing.

[**Methods / Process**] An experimental dataset was constructed by selecting 2,000 abstracts written by scholars from 20 core journals in the field of Information Resource Management published between 2018 and 2022. These were supplemented by 18,000 abstracts generated by nine mainstream LLMs from both domestic and international sources. Based on this dataset, the study conducted a comparative analysis of lexical features—such as unique words and common phrases—between the two types of abstracts. Subsequently, a multi-classification model was constructed to perform provenance detection, culminating in the development of an "AI-Generated Paper Abstract Detection Assistant."

[**Results / Conclusion**] The research indicates that the lexical style of scholar-written abstracts is professional, rigorous, and closely aligned with specific research topics. In contrast, AI-generated abstracts frequently employ templated unique words and phrases. Compared to sentence-level detection, large models perform better at the text-level, with the fine-tuned BERT model demonstrating outstanding performance. The AI-Generated Paper Abstract Detection Assistant provides technical support for maintaining academic integrity in the digital intelligence era.

Keywords: Large Language Models, Lexical Feature Analysis, Attribution Detection, Information Resource Management, Deep Learning **Classification Number:** G353.1 **DOI:** 10.31193/SSAP.J.ISSN.2096-6695.2026.01.04

---

1. Introduction

The writing and publication of academic papers represent a critical component of scientific research for faculty and students in higher education. With the rapid advancement of Large Language Models (LLMs) and generative AI tech-

nologies, the academic community is facing unprecedented challenges regarding the authenticity and integrity of scholarly outputs. While AI tools can enhance writing efficiency, their potential for misuse in generating entire sections of research papers—particularly abstracts—necessitates robust detection and provenance methods.

Currently, publishing institutions have imposed restrictions on Artificial Intelligence Generated Content (AIGC). For instance, in 2023, the journal *Science* banned the use of text or graphics generated by AI tools and stipulated that AI cannot be credited as an author [**?**]. Similarly, Elsevier and various Chinese journals such as *Library and Information Service* have issued policies requiring formal declarations if generative AI is used [**?**, **?**]. Consequently, AIGC detection has become an indispensable process to ensure the originality of academic achievements. This paper explores the linguistic nuances that distinguish human writing from machine-generated text, focusing specifically on lexical diversity, density, and structural patterns within the field of Information Resource Management.

2. Data Sources and Preprocessing

2.1 Data Source Construction

The dataset consists of two primary components: 1. **Scholar-Authored Abstracts:** 2,000 abstracts were selected from 20 core journals in Information Resource Management (e.g., *Journal of Library Science in China*, *Journal of Intelligence*) indexed in the CSSCI between 2018 and 2022. To eliminate influence from ChatGPT, papers were selected based on high citation frequency prior to 2023. 2. **AI-Generated Abstracts:** 18,000 abstracts were generated using nine prominent LLMs (including ChatGPT, Claude, Baidu Ernie Bot, Kimi, and Zhipu GLM). Using the CO-STAR framework [**?**], prompts were designed to simulate high-quality academic writing based on the original titles and keywords of the scholar-authored papers.

2.2 Data Preprocessing

The preprocessing pipeline included: - **Sensitive Information Filtering:** Removing samples where models refused to generate content due to political sensitivity and replacing them with alternative papers from the same authors. - **Domain Lexicon Construction:** Building a lexicon of 2,172 high-frequency keywords from 5,000 bibliographic records to assist in precise segmentation. - **Word Segmentation:** Utilizing the `jieba` tool with the custom domain lexicon and a stop-word list of 1,419 terms. - **Category Labeling:** Marking abstracts with labels 0 (scholar) and 1–9 (specific AI models).

Figure 1

Figure 1: Figure 1

3. Comparative Analysis of Lexical Features

3.1 Unique Word Features

Unique words are terms appearing exclusively in one category. Analysis shows that scholar-authored abstracts contain the highest number of unique words (3,509), significantly exceeding all AI models (e.g., Tongyi Qianwen: 837; Doubao: 58). A chi-square test ($p < 0.03$) confirmed significant differences in category distribution.

Scholars' unique vocabulary is characterized by: - **Professional Terminology (26%):** Mathematical and statistical terms like "mean" or "power exponent." - **Institutional References (17%):** Specific university names and administrative terms. - **Contextual Specificity:** Vocabulary closely tied to the actual research problem.

AI-generated abstracts, conversely, rely on: - **Cultural/General Terms:** Domestic models often use "tea culture" or "social trends." - **Functional Fillers (21%):** Terms like "aimed at," "based on this," and "conversely," which create a "structurally standardized but semantically hollow" style.

3.2 Common Phrase Features (N-gram Analysis)

Using N-gram analysis ($2 \leq N \leq 10$), we identified the top 1,000 frequent phrases. While both groups share technical terms like "blockchain technology," scholars use specific phrases related to national policies ("high-quality development"). AI models rely on formulaic templates such as "against the background of the information explosion era" or "this study adopts the literature analysis method."

4. Provenance Detection Methods

The detection problem was modeled as a multi-class classification task. We compared eight models across document-level and sentence-level granularities.

4.1 Model Performance

As shown in , document-level detection outperformed sentence-level detection due to richer contextual information. - **Fine-tuned BERT:** Achieved the highest performance (Accuracy: 0.9410, F1-Score: 0.9404). - **LightGBM:** Followed closely with an F1-Score of 0.9177. - **Traditional Models:** SVM, Logistic Regression, and Random Forest maintained scores above 80%. - **Naive Bayes:** Performed poorly (F1: 0.4661) due to its inability to capture complex semantic features.

Figure 2

Figure 2: Figure 2

4.2 Feature Importance

All models identified eight key feature terms—"aims to," "research," "provide," "explore," "through," "optimize," "enhance," and "adopt"—as critical for distinguishing provenance.

4.3 Comparison with Commercial Systems and LLM Self-Detection

Tests on 100 samples showed that the CNKI Personal AIGC Detection System identified 31.2% as suspected AIGC, while standard plagiarism checks failed entirely. Interestingly, using Zhipu GLM as a detector proved ineffective; while it identified 78.8% of scholar-written texts correctly, it misidentified 84.6% of AI-generated abstracts as scholar-written.

5. Implementation: Detection Assistant

We developed the "AI-Generated Research Paper Abstract Detection Assistant" using a B/S architecture with a Streamlit front-end and a LightGBM back-end. In practical tests, the assistant correctly identified the probability of an abstract being generated by a specific model (e.g., identifying a Baidu Ernie Bot abstract with 47.3% probability).

6. Conclusion

This study demonstrates that AI-generated abstracts exhibit distinct "fingerprints" characterized by templated language and lower lexical diversity. While deep learning models like BERT provide robust provenance tracing at the document level, sentence-level detection remains a challenge. Future research will expand this framework to other disciplines (Science, Medicine) and explore cross-disciplinary domain-adaptive pre-training to enhance generalization.