https://doi.org/10.51408/issi2025_208

Attempts to Enable Generative AI for Topic Recognition: A Case Study of ChatGPT

Wenting Tang[1], Wen Lou[2]

[1]*twt2543535795@163.com,*[2]*wlou@infor.ecnu.edu.cn*
East China Normal University, 3663 Zhongshan North Road, 200062, Shanghai (China)

Introduction

In recent years, the application of generative AI has seen growing use in NLP tasks like keyword extraction, entity recognition, and translation (Lu et al., 2024), yet its role in topic recognition remains underexplored.

Traditional topic models like LDA and PLSA build thematic spaces via word co-occurrence matrices, often causing semantic ambiguity and feature sparsity in theme inference. In contrast, generative AI develops deep contextual semantic representations through massive corpus pre-training, enabling accurate identification of implicit themes and effective mitigation of theme recognition bias. This study aims to explore the application of generative AI, specifically ChatGPT, in topic recognition in scientific literature. The study attempts to achieve efficient topic recognition through two different strategies and compares two strategies with machine learning methods. The study evaluates the advantages and limitations of generative AI in topic recognition, providing richer offering empirical insights for its practical use in literature analysis.

Methodology

Strategy One: Topic Recognition Based on Excel Files

This strategy enables ChatGPT to process large metadata from Excel files. Using PubMed as the source, the study filters medical literature published between 2000 and 2020, with article types including Clinical Trial, Meta-Analysis, and Randomized Controlled Trial. Using web scraping, key data like titles, abstracts, keywords, and publication dates are extracted and formatted into an Excel file with 17,000 records.

For topic recognition, this strategy attempts to use ChatGPT to perform topic recognition based on the BERT model (Sawant et al., 2022). The strategy provides ChatGPT with a basic explanation of the BERT framework and uses specific instructions to guide ChatGPT in performing BERT-based clustering. Specifically, ChatGPT is instructed not to directly provide the BERT model code but to encode each piece of metadata using the BERT model, extract its semantic features, and apply a clustering algorithm to group similar literature into categories for topic aggregation and recognition.

Strategy Two: Topic Recognition Based on Abstract Content

This strategy involves directly inputting the literature titles and abstracts into ChatGPT in the form of a dialogue to perform topic recognition. Specifically, this strategy guides ChatGPT to follow the steps of the DBSCAN model for topic clustering (Luchi & Rodrigues, 2019).

The strategy first instructs ChatGPT to removestopwords and numbers, normalize word forms, and construct a document vocabulary list from the abstracts. It then calculates TF-IDF and cosine similarity to assess topic similarity. With defined ε and MinPts, it classifies metadata into core, border, and noise points, further organizing the data into topic categories to provide insights into potential research topics.

Results and Discussion

Discussion of Strategy One: Topic Recognition Based on Excel Files

Table 1. BERT Keywords VS ChatGPT Keywords.

	Topic Feature Keywords Identified by ChatGPT	Topic Feature Keywords Identified by BERT
Topic 1	hypertension, treatment, risk	hypertension, amlodipine, antihypertensive
Topic 2	cancer, lung, factors	acupuncture, rehabilitation, stroke
Topic 3	infection, H. pylori, gastric	nutrition, parenteral, enteral
Topic 4	community, effectiveness, intervention	propofol, anesthesia, dexmedetomidin e
Topic 5	clinical, randomized, controlled	rectal, laparoscopic, anastomosis

Both ChatGPT and the BERT model identified topics related to hypertension treatment and antihypertensive drugs. However, ChatGPT emphasized a broader evaluation of "effects" and "risks," while BERT concentrated on specific medications like "amlodipine" and their impact on blood pressure control. BERT's topics were more detailed, exploring specific treatments, whereas ChatGPT identified overarching themes about hypertension treatment effectiveness.

For other topics, there was minimal overlap between ChatGPT and BERT. ChatGPT's themes were broader, suitable for detecting trends in large datasets, while BERT excelled in semantic accuracy and context, particularly in recognizing technical terms and treatment methods.

According to Bougioukas 's literature review (Bougioukas et al., 2021), keywords like "systematic review" and "study" appear most often, aligning with ChatGPT's Topic 1. Medical terms such as "acupuncture," "cancer," and "effectiveness" match BERT's Topic 2 and ChatGPT's Topics 1, 2, and 4. This suggests bibliometric methods produce research topics semantically and topically similar to those from generative AI.

Discussion of Strategy Two: Topic Recognition Based on Abstract Content

In topic recognition based on abstract content, although effective topic clustering was achieved by following the steps of the DBSCAN model, practical challenges still arose.

First, the issue of selecting ε and MinPts. The key to DBSCAN lies in selecting ε and MinPts. Manual tuning often requires multiple trials to optimize clustering, during which ChatGPT may produce memory errors—like fabricating cosine similarities between fictional documents—causing result deviations and reducing topic recognition accuracy.

Second, the issue of accurately comparing numerical values. Since DBSCAN's reliance on cosine similarity involves comparing small decimals. ChatGPT may misjudge values with varying decimal places (e.g., seeing 0.3 as smaller than 0.11), leading to misclassification of core points and distorted clustering.

Third, there is the issue of input and output word count limits. While batch processing helps mitigate word count restrictions, merging data from different batches may exceed the system's capacity, reducing efficiency and impacting the stability of the results.

Conclusion

This study explores the application of generative artificial intelligence in topic recognition of medical literature through two strategies: Excel files and abstract content. In the Excel-based approach, only one ChatGPT topic aligned with BERT's; BERT captured finer details, while ChatGPT identified broader themes but missed semantic nuance. The abstract-based strategy enabled effective clustering but faced issues with parameter tuning, numerical precision, and word count limits.

Overall, generative AI holds promise for topic recognition but requires further optimization for large-scale data and semantic precision. Future work will integrate traditional methods with generative AI to enhance efficiency and accuracy.

Acknowledgments

This work was supported by the Independent Research Project of Key Laboratory of Frontier Theory and Application of Statistics and Data Science of Ministry of Education, Identification and Application of Key Technologies in Medical Research Empowered by Digital Intelligence [KLATASDS2406], and by the Shanghai Planning Office of Philosophy and Social Science Project, Research on the Theoretical Framework and Implementation Path of Intelligent Sharing of Health Science Data[2024BJC005].

References

Bougioukas, K. I., Vounzoulaki, E., Mantsiou, C. D., et al. (2021). Global mapping of overviews of systematic reviews in healthcare published between 2000 and 2020: a bibliometric analysis. Journal of Clinical Epidemiology, 137, 58–72.

Luchi, D., & Rodrigues, A. L. (2019). Sampling approaches for applying DBSCAN to large datasets. Pattern Recognition Letters, 117, 90-96.

Lu, W., Liu, Y. P., Shi, X., et al. (2024). Academic text mining driven by large models: Construction of inference-end instruction strategies and capability evaluation. Journal of the China Society for Scientific and Technical Information, 43(08), 946-959.

Sawant, S., Yu, J., Pandya, K., et al. (2022). An enhanced BERTopic framework and algorithm for improving topic coherence and diversity. IEEE 24th International Conference on High Performance Computing & Communications.