

ORIGINAL ARTICLE

Research on a data mining algorithm based on BERTopic for medication rules in Traditional Chinese Medicine prescriptions

Hongchen Li¹  | Xinyi Lu¹ | Yujia Wu² | Jie Luo³

¹School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, Hangzhou, China

²The First School of Clinical Medicine, Zhejiang Chinese Medical University, Hangzhou, China

³Institute of Innovation and Entrepreneurship, Zhejiang Chinese Medical University, Hangzhou, China

Correspondence

Jie Luo, Institute of Innovation and Entrepreneurship, Zhejiang Chinese Medical University, Hangzhou, China.
Email: roger@zcmu.edu.cn

Abstract

Background: A data mining algorithm is proposed based on BERTopic to provide new insights into the analysis of medication rules in Traditional Chinese Medicine (TCM) prescriptions.

Methods: Using the BERTopic algorithm, collected TCM prescriptions for corneal diseases are converted to embeddings through a transformer based on the Bidirectional Encoder Representations from Transformers pre-trained model. Then, Uniform Manifold Approximation and Projection is applied to perform dimensionality reduction in prescription embeddings. Subsequently, Hierarchical Density-Based Spatial Clustering of Applications with Noise is used for clustering. Finally, class-based term frequency-inverse document frequency is used to generate several main drug combinations from the clustered results.

Results: The highest frequency of drugs used included *Buddleja officinalis*, *Bidens pilosa*, *Angelica sinensis*, *Eriocaulon buergerianum*, and *Raw Rehmannia glutinosa*. The most frequent drug combinations were “*Eriocaulon buergerianum*, *Raw Rehmannia glutinosa*, *Prunella vulgaris*, *Notopterygium incisum*” “*Lycii Fructus*, *Bidens pilosa*, *Buddleja officinalis*” and “*Kochiae Fructus*, *Cortex Dictamni*.”

Conclusions: The proposed data mining algorithm based on BERTopic demonstrated promising outcomes in the analysis of TCM prescription medication rules. This method exhibited simplicity and efficiency, thereby offering a novel avenue for analysis.

KEYWORDS

BERTopic, data mining, medication rules, traditional Chinese medicine

Abbreviations: BERT, bidirectional encoder representations from transformers; c-TF-IDF, class-based term frequency-inverse document frequency; DNN, deep neural network; HDBSCAN, hierarchical density-based spatial clustering of applications with noise; NLP, natural language processing; TCM, traditional Chinese medicine; TF-IDF, term frequency-inverse document frequency; UMAP, uniform manifold approximation and projection.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Medicine Advances* published by John Wiley & Sons Ltd on behalf of Tsinghua University Press.

1 | INTRODUCTION

Data mining is a method that can extract potentially valuable information from a large amount of random and fuzzy data, and is commonly used to analyze the medication rules for Traditional Chinese Medicine (TCM) prescriptions. The mining methods mainly used in the field of TCM include association analysis, clustering analysis, factor analysis, the genetic algorithm, and deep neural networks (DNNs) [1–3]. Bidirectional Encoder Representations from Transformers (BERT) is an advanced deep learning pre-trained language model that has achieved remarkable results in a wide range of natural language processing (NLP) tasks [4]. In this study, a topic modeling algorithm called BERTopic was used to analyze the medication rules in 1276 TCM prescriptions for corneal diseases from the Zhejiang Hospital of TCM from 2016 to 2021.

2 | METHODS

Clinical TCM mainly uses TCM prescriptions to treat diseases. Data mining methods can effectively mine prescription data deeply to provide the basis for the analysis of medication rules [5]. The theory of TCM properties contains ambiguity and conceptuality, which makes the research of TCM a great challenge as many words and phrases have been used through extension, metonymy, or other methods, resulting in the phenomenon of polysemy [6]. Conceptual and ambiguous challenges also exist while translating sentence and word meanings with NLP. To enable computers to understand and process natural language, people have trained many excellent models using DNN technology [7].

In this paper, we adopt the BERTopic algorithm to obtain the core prescription through topic extraction and then analyze the correlation between drugs according to the normalized c-TF-IDF score.

2.1 | BERTopic

BERTopic is a neural topic model with a class-based (c) Term Frequency–Inverse Document Frequency (TF-IDF) algorithm that mainly uses BERT, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), Uniform Manifold Approximation and Projection (UMAP), and c-TF-IDF [8]. It can understand the input text using a neural network and extract several topics discussed in the text, each containing several keywords. The technical flow chart is shown in Figure 1.

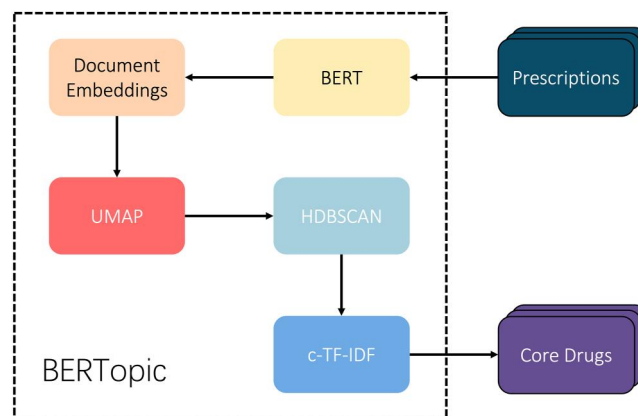


FIGURE 1 Technical flow chart of BERTopic. First, prescriptions in the form of documents are embedded by the bidirectional encoder representations from transformers pre-trained model. Then, UMAP is used for dimensionality reduction. Next, embeddings are clustered through HDBSCAN. Finally, core drugs are generated by c-TF-IDF as several topic words from the document.

2.2 | Document embeddings

BERTopic assumes that documents that contain the same topic are semantically similar. The documents are embedded through Sentence-BERT (SBERT) [9]. This is a framework that can convert sentences and paragraphs to dense vector representations using the BERT pre-trained model.

In this study, for TCM prescriptions, the Chinese-supported *paraphrase-multilingual-MiniLM-L12-v2* pre-trained model is used to perform the embedding step for each prescription, which results in prescription embedding.

2.3 | Dimension reduction

The embeddings have hundreds of dimensions, which makes differences in distance measurement unclear. To improve the clustering effect, the dimensionality of embeddings needs to be reduced. BERTopic uses UMAP for dimension reduction to reduce the dimensionality of document embeddings.

UMAP is an advanced dimension reduction algorithm based on the theoretical framework of Riemannian geometry and algebraic topology [10]. It assumes that the data satisfy the following three conditions: (1) data should be uniformly distributed on an existing Riemannian manifold; (2) the underlying manifold of interest is locally connected; and (3) the topology of the manifold is constant.

The algorithm performs excellently and has been shown to retain more global and local features of high-dimensional data in the lower dimension, as in the research sentence vector with hundreds of dimensions are able to reduce to no more than 10 vector groups or even less; hence, it works well with real-world data and has been widely applied in various fields [11]. Through manifold modeling, the equivalent fuzzy topology can be used to obtain the projection of prescription embeddings in lower dimensions.

2.4 | Document clustering

The dimensionality reduced embeddings are clustered by HDBSCAN. HDBSCAN is a spatial clustering algorithm based on hierarchical density with noise, which can solve clustering problems with various densities through parameter control [12]. It mainly includes five steps: spatial transform, building the minimum spanning tree, constructing the hierarchical structure of cluster groups, compressing the cluster tree, and extracting clusters.

Compared with traditional clustering algorithms, the HDBSCAN algorithm effectively improves the clustering effect of multiple topics.

2.5 | Topic representation

Term Frequency-Inverse Document Frequency is a statistical method that can be used to assess the importance of a word in a document and extract its keywords:

$$TF-IDF_{t,d} = TF_{t,d} * IDF_t \quad (1)$$

where the size of the TF-IDF product reflects the importance of term t in document d . Term frequency (TF) refers to the frequency with which term t appears in document d :

$$TF_{t,d} = \frac{T}{D} \quad (2)$$

where T represents the number of occurrences of term t in document d and D represents the total number of words in document d . The Inverse Document Frequency (IDF) reflects the frequency with which term t appears in the corpus:

$$IDF_t = \log\left(\frac{N}{c}\right) \quad (3)$$

where N represents the total number of documents in the corpus and c represents the number of documents in the corpus that contain term t .

For TCM prescriptions, the value of TF is the frequency with which a single TCM appears in a series of prescriptions and the value of IDF reflects the frequency with which the TCM appears in the corpus. The product of TF and IDF indicates the importance of the drug for a series of prescriptions.

Extracting too many documents directly through TF-IDF results in too many unclassified keywords, which is the main disadvantage. BERTopic has improved the TF-IDF algorithm and proposed a class-based TF-IDF algorithm, with only document d making some changes on concatenating all documents in a cluster (namely a class) as a new single document d :

$$W_{t,c} = TF_{t,c} * \log\left(1 + \frac{A}{TF_t}\right) \quad (4)$$

where c represents a class, A represents the average number of words in class c , $TF_{t,c}$ represents the frequency of term t in class c and TF_t represents the frequency of term t in all classes.

According to the clustering results, c-TF-IDF splices documents in each class and calculates TF-IDF, thereby effectively mining keywords in each topic class.

3 | APPLICATION

3.1 | Dataset

The prescription data used in this study was sourced from the Department of Ophthalmology at Zhejiang Hospital of TCM, and covered the period from January 2016 to December 2021. After the data were filtered, a total of 1276 valid prescriptions for the treatment of corneal disorders were selected for analysis in this study.

Microsoft Excel was used as the storage format for the data, which recorded the drugs and their dosages in each prescription. Among the 1276 prescriptions used for treating corneal disorders, a total of 127 Chinese herbal medicines were included. Through frequency analysis, it was found that *Buddleja officinalis* and *Bidens pilosa* appeared the most frequently in the prescriptions (Figure 2).

3.2 | Experimental setup

The process of extracting topic keywords from the documents using BERTopic is similar to the process of

extracting core prescriptions. Traditional Chinese Medicine can be regarded as words, whereas TCM prescriptions are sentences. A series of prescriptions for treating specific diseases can be considered as a paragraph in the document with a clear theme. Therefore, the collected TCM prescriptions were organized into paragraphs so that BERTopic could obtain their topics and information, which were used to analyze the medication rules in prescriptions.

The collected clinical TCM prescriptions were formatted as an Excel table. The *openpyxl* library in Python 3.9 easily reads the table into arrays. Then, the drugs in each prescription were concatenated, separated by spaces, and finally constructed into strings to form a document that BERTopic could understand. All the prescription strings were stored in an array, thereby forming paragraphs that could be analyzed.

The *paraphrase-multilingual-MiniLM-L12-v2* model was used, for which each prescription is embedded in a 384-dimensional sentence vector through a *Sentence Transformer*. To improve the clustering effect, the 384-dimensional sentence vector group was reduced to a five-dimensional vector group through UMAP. Then, the vector group was clustered through HDBSCAN to obtain a set of several compatible prescriptions for the disease. Finally, c-TF-IDF extracted topic words from prescriptions in each classification to obtain potential core drug combinations for the disease.

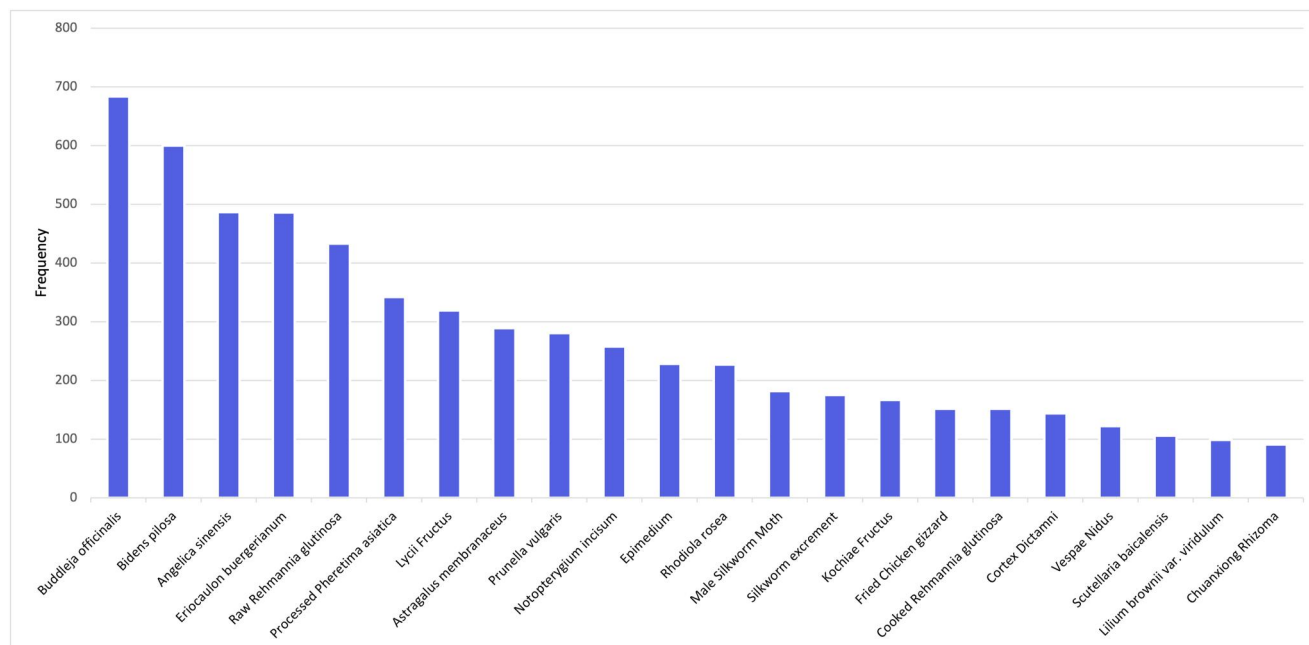


FIGURE 2 Drug frequency distribution. Frequency represents the number of times a drug appeared in all prescriptions. The drugs are sorted in order of frequency, from highest to lowest. The ordinate is the frequency of each drug and the abscissa is the drug. All the drug names contained in the prescriptions included in the experiment were regulated in accordance with the “Chinese Materia Medica” and “Pharmacopoeia of the People’s Republic of China”.

4 | RESULTS

4.1 | Cluster analysis

After the prescriptions were embedded using BERT, dimensionality was reduced through UMAP to improve the clustering effect. Figure 3 shows the clustering results

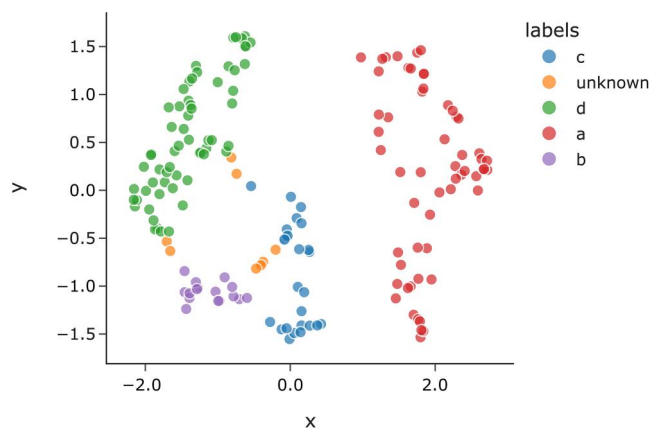


FIGURE 3 Results of prescription clustering. Prescriptions were divided into four categories, where a, b, c and d represent the 4 recognized clusters, unknown represents the noise data. Discrete values are marked as unknown. During clustering, the data were five-dimensional, however, for visualization purposes, the data in the figure was reduced to two dimensions.

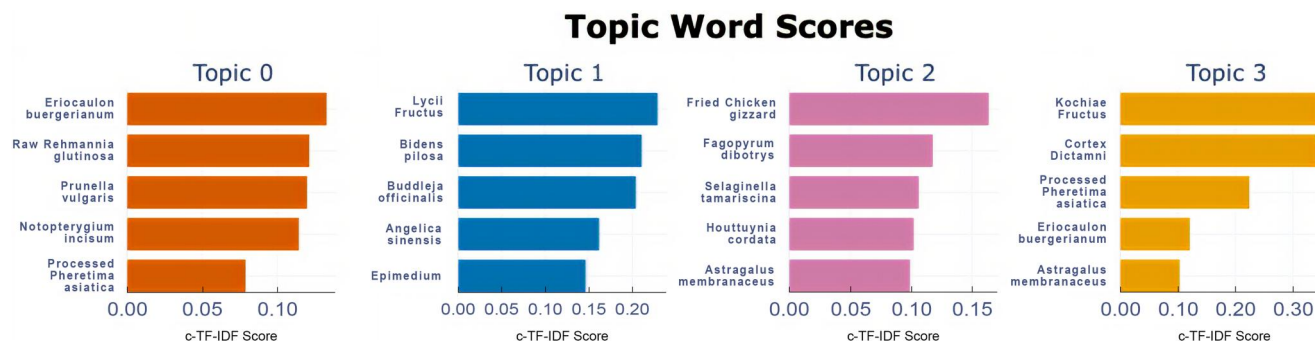


FIGURE 4 Potential core drug combinations. Topic 0–3 are topic word scores of prescriptions as documents. The results are sorted in order of frequency, from highest to lowest. The ordinate is the core drug and the abscissa is the frequency of each drug. The frequencies of Topic 0 and Topic 1 are close, and the results are most likely to be the main drug combination for the disease.

after dimensionality reduction. All prescriptions were classified into four categories and noise was ignored. The prescriptions were used to treat corneal diseases and each category was considered as a different direction for treatment.

4.2 | Core drug combinations

Through the analysis of BERTopic, four topics were extracted from the prescriptions, which identified four potential directions for medication for the treatment of the disease. The results are presented in Figure 4.

4.3 | Analysis of drug association

The normalization results for the Topic 0 frequency data are shown in Table 1, with Eriocaulon buergerianum, Raw Rehmannia glutinosa, Prunella vulgaris, and Notopterygium incisum as the main drug combinations.

The normalization results for the Topic 1 frequency data are shown in Table 2, with Lycii Fructus, Bidens pilosa, and Buddleja officinalis as the main drug combinations.

The normalization results for the Topic 2 frequency data are shown in Table 3.

The normalization results for the Topic 3 frequency data are shown in Table 4, with Kochiae Fructus and Cortex Dictamni as the main drug combinations.

4.4 | Drug frequency analysis

Among the 1276 TCM prescriptions for the treatment of corneal diseases, 127 drugs appeared, with a cumulative frequency of 8521 times. There were 22 high-frequency drugs with a frequency greater than or equal to 90,

TABLE 1 Normalization results for Topic 0.

Number	Name	Rate
1	Eriocaulon buergerianum	1
2	Raw rehmannia glutinosa	0.823964997
3	Prunella vulgaris	0.800917872
4	Notopterygium incisum	0.717323639
5	Processed pheretima asiatica	0.176097512
6	Angelica sinensis	0.153616430
7	Cooked rehmannia glutinosa	0.094778020
8	Chuanxiong rhizoma	0.086225339
9	Scutellaria baicalensis	0.031561448
10	Buddleja officinalis	0

TABLE 2 Normalization results for Topic 1.

Number	Name	Rate
1	Lycii fructus	1
2	Bidens pilosa	0.896727280
3	Buddleja officinalis	0.859092171
4	Angelica sinensis	0.622881305
5	Epimedium	0.535057132
6	Male silkworm moth	0.440211308
7	Rhodiola rosea	0.220790726
8	Raw rehmannia glutinosa	0.179506823
9	Silkworm excrement	0.015697210
10	Lotus leaf	0

which accounted for 70.70% of the total drug frequency. Table 5 shows the frequency of prescriptions for corneal disorders.

TABLE 3 Normalization results for Topic 2.

Number	Name	Rate
1	Fried chicken gizzard	1
2	Fagopyrum dibotrys	0.482333610
3	Selaginella tamariscina	0.353147508
4	Houttuynia cordata	0.305662613
5	Astragalus membranaceus	0.271872005
6	Actinidia valvata dunn	0.217213023
7	Lysimachia christinae	0.181239659
8	Actinidia chinensis planch radix	0.160029286
9	Vespae nidus	0.008234572
10	Processed pheretima asiatica	0

TABLE 4 Normalization results for Topic 3.

Number	Name	Rate
1	Kochiae fructus	1
2	Cortex dictamni	0.987173983
3	Processed pheretima asiatica	0.563268572
4	Eriocaulon buergerianum	0.193992237
5	Astragalus membranaceus	0.131285012
6	Vespae nidus	0.107117904
7	Fried vespae nidus	0.096110332
8	Prunella vulgaris	0.072933842
9	Silkworm excrement	0.064182601
10	Buddleja officinalis	0

TABLE 5 Frequency analysis of high-frequency drugs (frequency ≥ 90).

Number	Name	Frequency	Rate
1	Buddleja officinalis	684	8.03
2	Bidens pilosa	600	7.04
3	Angelica sinensis	487	5.72
4	Eriocaulon buergerianum	486	5.70
5	Raw rehmannia glutinosa	433	5.08
6	Processed pheretima asiatica	342	4.01
7	Lycii fructus	319	3.74
8	Astragalus membranaceus	289	3.39
9	Prunella vulgaris	281	3.30
10	Notopterygium incisum	258	3.03
11	Epimedium	228	2.68
12	Rhodiola rosea	227	2.66
13	Male silkworm moth	182	2.14
14	Silkworm excrement	175	2.05
15	Kochiae fructus	167	1.96
16	Fried chicken gizzard	152	1.78
17	Cooked rehmannia glutinosa	152	1.78
18	Cortex dictamni	144	1.69
19	Vespae nidus	122	1.43
20	Scutellaria baicalensis	106	1.24
21	Lilium brownii var. viridulum	99	1.16
22	Chuanxiong rhizoma	91	1.07

5 | RELATED WORK

5.1 | Cluster analysis

Currently, in cluster analysis, experts are often used to manually specify the vector weights of each TCM, which is an error prone and time-consuming process. The K -means algorithm is used for clustering [13]. This method is simple and easy to implement; however, the clustering effect depends on the choice of the K value and initial point [14] and is easily disturbed by noisy data.

5.2 | Term Frequency–Inverse Document Frequency

TF-IDF is also a TCM prescription data mining method [15]. Its goal is to calculate the TF-IDF value of each TCM. Table 6 shows the results of TF-IDF ≥ 5 . The results

showed that drugs commonly used in the treatment of corneal diseases included Buddleja officinalis, Bidens pilosa, Angelica sinensis, Eriocaulon buergerianum, Lycii Fructus, Raw Rehmannia glutinosa, Processed Pheretima asiatica, Epimedium, and Prunella vulgaris.

The results of TF-IDF only represent the types of core drugs in the batch of prescriptions, which is a macro analysis of prescriptions without local information. It is difficult to specify the combination of several drugs and hence the compatibility relationship cannot be explained well.

6 | DISCUSSION

At present, data mining technology is widely used in the field of TCM. In this paper, we proposed a data mining algorithm based on BERTopic. In the analysis of TCM prescription data used to treat corneal diseases, the results obtained good evaluations by clinical doctors.

TABLE 6 Results of term frequency-inverse document frequency analysis.

Number	Name	Rate
1	Buddleja officinalis	23.55
2	Bidens pilosa	22.53
3	Angelica sinensis	18.95
4	Eriocaulon buergerianum	18.60
5	Lycii fructus	16.18
6	Raw rehmannia glutinosa	16.13
7	Processed pheretima asiatica	14.13
8	Epimedium	13.09
9	Prunella vulgaris	12.58
10	Notopterygium incisum	12.54
11	Male silkworm moth	12.11
12	Astragalus membranaceus	11.44
13	Rhodiola rosea	10.89
14	Silkworm excrement	9.78
15	Kochiae fructus	8.45
16	Cooked rehmannia glutinosa	7.60
17	Cortex dictamni	7.55
18	Vespae nidus	6.68
19	Fried chicken gizzard	6.67
20	Scutellaria baicalensis	5.88
21	Lilium brownii var. viridulum	5.55
22	Lotus leaf	5.53

In this study, based on the topic modeling ability of BERTopic, first, prescriptions were clustered and then core drugs were extracted for each category of prescription. This demonstrated the potential for local compatibility. The HDBSCAN algorithm was used for clustering, which reduced the interference of noise, to some extent, and achieved good clustering results. Simultaneously, BERTopic used the powerful BERT pre-trained model for semantic analysis, which has great potential in the application of TCM prescription data mining.

As a neural topic model, BERTopic has shown a good comprehensive effect with its high efficiency and satisfactory clustering results, which brings about more thinking in the application field of language models. The compatibility relationship of deep-level TCM prescriptions will be further explored using a language model, which will introduce new ideas to the study of medication rules and provide a more comprehensive analysis.

Finally, as a new method to explore TCM prescription, our research still has the limitation of sample size and data

support. Future research is required to explore more specific correlations between TCM prescriptions using language models, and more clinical research and evaluation are also needed in order to enrich our understanding of data mining and TCM treatment of diseases.

AUTHOR CONTRIBUTIONS

Hongchen Li: Conceptualization (supporting); data curation (equal); formal analysis (lead); investigation (equal); methodology (lead); project administration (lead); software (lead); validation (equal); visualization (lead); writing—original draft (lead). **Xinyi Lu:** Data curation (equal); formal analysis (supporting); investigation (equal); methodology (supporting); software (supporting); validation (equal); visualization (supporting). **Yujia Wu:** Data curation (equal), formal analysis (supporting), investigation (equal), methodology (supporting), resources (supporting), validation (equal). **Jie Luo:** Conceptualization (lead); data curation (equal); funding acquisition (lead); investigation (equal); project administration (supporting); resources (lead); supervision (lead); validation (equal); writing—review and editing (lead).

ACKNOWLEDGMENTS

None.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ETHICS STATEMENT

This article is a practice-oriented case study description that made extensive use of secondary information sources and also drew upon the professional knowledge of the co-authors. As such, the creation of this case study article did not involve any formal study, nor did it involve human participation in a study. As such, IRB review was not required for this article.

CONSENT TO PARTICIPATE

Not applicable.

ORCID

Hongchen Li  <https://orcid.org/0009-0006-2317-8305>

REFERENCES

- [1] Feng M, Wang Y, Bai D, Wang JH. Research progress of TCM prescription analysis methods based on data mining. World Chin Med. 2022;17(23):3411–6.

- [2] Chen XT, Ruan CT, Zhang YC, Chen HJ. Heterogeneous information network based clustering for precision traditional Chinese medicine. *BMC Med Inf Decis Making*. 2019;19(Suppl 6):264. <https://doi.org/10.1186/s12911-019-0963-0>
- [3] Xu DT, Peng KS, Huang ZJ, Guo XQ. Data mining of acupuncture prescription patterns for early-onset ovarian insufficiency based on association rules and cluster analysis. *Glob Tradit Chin Med*. 2022;15(12):2381–7.
- [4] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding; 2018. arXiv: 1810.04805.
- [5] Tang SH, Yang HJ. Review of study on traditional Chinese medicine medication regularity. *Chin J Exp Tradit Med Formulae*. 2013;19(5):359–63.
- [6] Wang T, Xu YX, Li WL, Xu JF. The application of cluster analysis in the study of TCM prescription. *Clin J Tradit Chin Med*. 2017;29(12):2035–7.
- [7] Natural language understanding: instructions for (present and future) use R navigli. Available from: <https://www.ijcai.org/Proceedings/2018/812>. Accessed on 25 June 2023.
- [8] Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure; 2022. arXiv: 2203.05794.
- [9] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks; 2019. arXiv: 1908.10084.
- [10] McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction; 2018. arXiv: 1802.03426.
- [11] Image and signal processing. Available from: <https://link.springer.com/book/10.1007/978-3-030-51935-3>. Accessed on 25 June 2023.
- [12] McInnes L, Healy J. Accelerated hierarchical density based clustering. In: 2017 IEEE international conference on data mining workshops (ICDMW). New Orleans: IEEE; 2017. p. 33–42. <https://doi.org/10.1109/ICDMW.2017.12>
- [13] Ma JH, Zhu Y, Lu M, Yu YX. Efficacy and prescription of Chinese medicine intervention on post-abortion based on K-means clustering algorithm. *World Chin Med*. 2022;17(4): 537–42.
- [14] Zhu HL, Zhao YY, Wang XY, Xu YL. Research on data analysis of traditional Chinese medicine with improved differential evolution clustering algorithm. *J Healthc Eng*. 2021; 2021:1–10. <https://doi.org/10.1155/2021/4468741>
- [15] Qu DD, Yang T, Hu KF. Application of NLP in automatic extraction of symptom information from Chinese medical cases. *Softw Guide*. 2021;20(2):44–8.

How to cite this article: Li HC, Lu XY, Wu YJ, Luo J. Research on a data mining algorithm based on BERTopic for medication rules in Traditional Chinese Medicine prescriptions. *Med Adv*. 2023;1(4):353–60. <https://doi.org/10.1002/med4.39>