

Text mining in biomedical literature: identifying chemical and disease relationships

Nurul Mutiah, Dian Prawira

Department of Information Systems, Faculty of Mathematics and Natural Science, Tanjungpura University, Pontianak, Indonesia

Article Info

Article history:

Received May 7, 2025

Revised Jul 14, 2026

Accepted Aug 12, 2026

Keywords:

Biomedical text mining
Chemical disease relationship
Knowledge graph
Named entity recognition
Relation extraction

ABSTRACT

Text mining is an essential method for extracting valuable information from biomedical literature, especially concerning chemical-disease relationships (CDR). This study integrates named entity recognition (NER), relation extraction (RE), classification, and topic modeling to analyze kratom-related research articles from PubMed. Using the biocreative V chemical-disease relation (BC5CDR) dataset, NER achieved high recall and accuracy, demonstrating its ability to identify diverse entities. However, low precision revealed challenges with false positives due to ambiguous contexts and term overgeneralization. Network analysis of extracted relationships identified “Neuropathic” as the most central entity, underscoring its importance in linking chemicals like “mitragynine” to clinical conditions. Classification with bidirectional encoder representations from transformers for biomedical text mining (BioBERT) yielded moderate results, with an F1-score of 0.48, precision of 0.43, and recall of 0.38, highlighting the need for larger, curated datasets to improve accuracy. Bidirectional encoder representations topic modeling (BERTopic) uncovered five main topics with moderate coherence ($C_V = 0.5567$), emphasizing themes related to kratom’s toxicity and clinical implications, although capturing semantic relationships remained challenging. This research demonstrates the potential of text mining in biomedical knowledge discovery, revealing key insights and thematic patterns in scientific literature. Improved data preprocessing, model fine-tuning, and manual validation are recommended to enhance the reliability and application of these methods in future studies.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Nurul Mutiah

Department of Information Systems, Faculty of Mathematics and Natural Science, Tanjungpura University

Prof. H. Hadari Nawawi Street, Pontianak, West Kalimantan, Indonesia

Email: nurul@sisfo.untan.ac.id

1. INTRODUCTION

The growing use of technology in biomedical research has increased the number of discoveries and experimental data. This trend is closely related to the rising quantity of scientific publications in biomedical each year and the increasing number of citations in the PubMed database, which, in 2023, contains 36 million citations and abstracts of biomedical literature in electronic text format. However, biomedical literature is primarily available in electronic form in unstructured formats [1] and is not optimized for effective information extraction [2]. Literature searches that rely only on titles, abstracts, and keywords often have limited sensitivity because scientific publication databases typically restrict searches to these elements, thereby reducing the specificity of the search results [3]. One of the main goals of large-scale biomedical literature processing is to extract relevant biomedical information and uncover new knowledge or the latest research findings in the biomedical field [4], [5]. It can also improve the quality of research and discovery by

integrating the results of information extraction into new findings in the biomedical field [6]. Collecting and updating biological data from millions of articles is a time-consuming and challenging process because simple term-based searches often yield irrelevant documents or overlook important information due to homonyms, synonyms, abbreviations, and term mismatches [7], [8]. The vast amount and complexity of biomedical literature make it time-consuming for researchers to read data and uncover new knowledge manually [9]. The use of computational technologies, such as artificial intelligence (AI), in the extraction and retrieval of large amounts of biomedical literature can greatly reduce search time and lead to the discovery of new knowledge [7]–[9].

Biomedical information extraction plays an important role in automatically identifying semantic structures, such as entities and relations, from large volumes of unstructured biomedical text articles [10]. Extracting information from large collections of scientific text is an effective approach for knowledge acquisition [11]. The abundance of diverse named entity classes in the biomedical field results in a combinatorial explosion of relationships between entities, making it costly and time-consuming to identify the most significant relationships through biological experiments [10]. Text mining techniques can uncover hidden information from large text corpora such as trends, patterns, and data duplication, as well as improve the conceptual understanding of unstructured data through classification, clustering, and summarization techniques [12], [13]. Text mining can support the process of knowledge extraction and discovery in biomedical scientific articles. Biomedical text mining aims to extract, analyze, and summarize meaningful information from structured and unstructured biomedical texts, enabling automatic text mining and generation techniques to help researchers gain valuable insights from the vast amount of available biomedical literature data [14]. Text mining techniques, such as named entity recognition (NER) [10], [15] and relation extraction (RE) [5], [16], can be used to extract information from scientific literature articles. NER is a part of information extraction that aims to identify and categorize entities in unstructured text documents and can be divided into general NER and domain-specific NER [10], [17]. The NER model forms the foundation for the relationship extraction (RE) process by identifying entities that are subsequently used as input for the RE model, these identified entities play a crucial role in establishing relationships between them [18].

Multiple studies have been conducted on RE, specifically extracting relationships between biomedical entities using generative pre-trained transformer (GPT) models [19] and utilize stacking-based ensemble learning and attention mechanisms by combining the advantages of multiple pre-trained models [20]. Research on NER using bidirectional encoder representations from transformers (BERT) with prompt contrastive learning [21], neural network-based joint learning approach [22], machine reading comprehension framework [23], and hybrid graph-based neural network [24] to identify disease and chemical entities within document texts. Previous studies indicate that NER and RE methods can effectively extract information from biomedical texts.

This study aims to identify entities and extract relationships between them, with a focus on the connection between chemicals and diseases in biomedical scientific articles using NER and RE text mining techniques. The model used for NER and RE is the biocreative V chemical-disease relation (BC5CDR) model from spaCy. The connection between chemicals and diseases is crucial in biomedical research, while entity RE plays a role in extracting information from biomedical literature [25]. The relationship between chemicals and diseases is represented in the form of a knowledge graph. The knowledge graph is a semantic network that represents information using various nodes (objects) and edges (relational interactions) to provide clear knowledge regarding entity attributes and their connections [26]. This study also performs topic modeling with Bidirectional encoder representations topic modeling (BERTopic). This aims to link information extraction results with the main topics in biomedical articles.

The dataset used in this study is a scientific article about *Mitragyna speciosa*, commonly known as kratom, which was obtained from the PubMed database using the National Center for Biotechnology Information (NCBI) Entrez application programming interface (API). Kratom has many benefits [27], but its use can cause negative impacts [28] and has been classified as a new psychoactive substance [29]. This study utilizes a dataset of scientific articles on kratom to demonstrate the effectiveness of text mining techniques. Specifically, it highlights the value of NER, RE, and topic modeling in uncovering hidden information and knowledge within biomedical literature. This method links biomedical entity extraction, relationships, and topic modeling to better understand kratom's biological aspects, benefits, risks, and regulations.

2. METHOD

This study focuses on the automated extraction of biomedical information from scientific literature. This research uses a computational approach based on text mining and natural language processing (NLP) to extract, model, and analyze the relationship between chemical compounds and diseases from scientific literature. The research design includes several main stages: data collection, text preprocessing, entity and RE, knowledge graph construction, and embedding-based topic modeling, as illustrated in Figure 1.

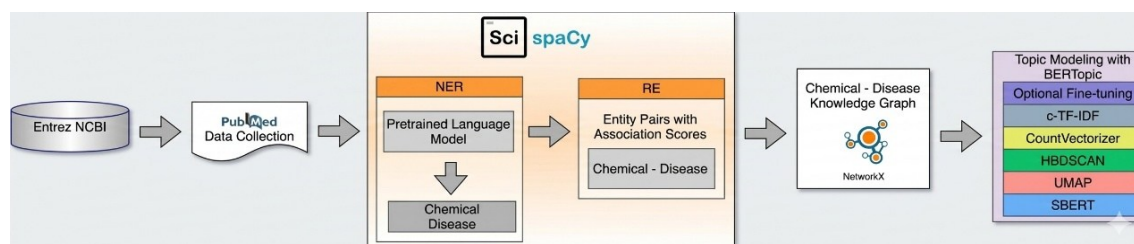


Figure 1. Research method

2.1. Data collection

The data for this study were collected from PubMed using the NCBI Entrez API, which provides access to metadata and abstracts from kratom articles. The data collection process was performed by executing a search query using specific keywords: “*Mitragyna speciosa*”, “kratom benefits”, “kratom effects”, and “kratom regulation”. These keywords were selected to ensure coverage of scientific articles relevant to the research topic, including information on kratom’s benefits, risks, regulations, and potential in a biomedical context. The collected dataset consists of 612 articles with important metadata, including title, abstract, author, and keywords.

2.2. Named entity recognition

This NER stage utilizes the SpaCy model enhanced with the BC5CDR, a pre-trained model specifically tailored for biomedical literature. The primary focus of this NER implementation is to identify two types of biomedical entities: chemical and disease, which are particularly relevant to kratom research. The NER process starts by using the tokenization and named entity tagging capabilities of SpaCy, optimized by ScispaCy’s `en_ner_bc5cdr_md` model. This model was selected due to its proven performance in handling complex biomedical terminology and specialized linguistic patterns, ensuring higher reliability in entity detection.

2.3. Relation extraction: knowledge graph construction

The RE process utilizes the BC5CDR model-based pipeline, specifically designed for RE in biomedical literature. This model is trained on the BC5CDR dataset, which includes explicit relationship annotations between chemical and disease entities, such as “Chemical A is used to treat Disease B”. The extracted relationships between chemicals and diseases are then represented in a knowledge graph. In this graph, nodes represent chemical and disease entities, while edges represent relationships between entities, weighted based on association scores. The knowledge graph is constructed and analyzed using the NetworkX library, which allows for systematic exploration of network structures, identification of relationship patterns, and topological analysis.

2.4. Topic modeling

To identify key themes in a collection of scientific literature, this study applies embedding-based topic modeling using the BERTopic framework. BERTopic is a topic modeling technique that combines class-based clustering with vector-based representation to produce more intuitive and meaningful topic representations. Semantic representations of documents are generated using sentence-BERT (SBERT), which are then dimensionally reduced using uniform manifold approximation and projection (UMAP). Next, the clustering process is performed using hierarchical density-based spatial clustering of applications with noise (HDBSCAN) to group documents based on semantic similarity. Keyword representations for each topic are obtained using a class-based term frequency-inverse document frequency (TF-IDF) class-based term frequency-inverse document frequency (c-TF-IDF) approach, with the option of fine-tuning the embedding model to improve the suitability for the biomedical domain. In this study, topic modeling seeks to identify the primary themes in scientific articles about kratom, including medical benefits, side effects, and legal regulations.

3. RESULTS AND DISCUSSION

3.1. Named entity recognition

Biomedical entities extracted from the dataset include disease and chemical entities, which were identified using SpaCy and SciSpaCy libraries. SpaCy is a comprehensive NLP library that facilitates essential tasks such as tokenization, part-of-speech tagging, and NER. In addition, the SciSpaCy extension provides domain-specific models that are fine-tuned to meet the unique requirements of biomedical text mining, allowing for more precise entity classification.

SciSpaCy offers pre-trained models that are designed to handle terminology and language commonly used in biomedical research. The model utilized in this research is `en_ner_bc5cdr_md`, specifically designed to identify two categories of biomedical entities: disease and chemical. Examples of biomedical entity extraction results with this model are shown in Table 1. The BC5CDR NER model was tested on the dataset, categorizing 332 “chemical” entities and 265 “disease” entities.

The evaluation results show that the model has a precision of 0.17, which indicates a relatively low proportion of correct positive predictions. However, the model demonstrates a very high Recall rate of 0.99, indicating that almost all relevant entities were successfully identified. The F1-score of 0.29 indicates an imbalance between precision and recall, while the overall accuracy is 0.99, suggesting most predictions, including those of the negative class, are correct. According to the confusion matrix heatmap in Figure 2, the model achieved 330 correct predictions for the label “chemical” with no errors. For the label “disease,” there were 259 correct predictions along with just 2 incorrect ones. This indicates the model’s excellent ability to detect entities with high accuracy, although low precision indicates the model’s tendency to produce many false positive predictions. Overall, the precision of the BC5CDR NER model needs improvement to achieve a better balance between prediction accuracy and entity detection capability.

The observed performance, characterized by exceptionally high recall (0.99) but lower precision (0.17), is consistent with the inherent challenges of biomedical NER systems. Entity recognition in specialized medical domains often encounters a ‘precision-recall trade-off’ due to the linguistic ambiguity of chemical nomenclatures that overlap with common biological terms [10]. This high sensitivity is crucial in clinical discovery tools to ensure no relevant chemical-disease associations are overlooked, although it necessitates subsequent human-in-the-loop validation or more stringent rule-based filtering, as suggested in comprehensive surveys of deep learning for NER [30].

Table 1. Results of chemical and disease entity extraction using the BC5CDR model

Entity	Label
7-hydroxymitragynine	Chemical
9-hydroxylated oxindole alkaloids speciofoline	Chemical
9,10-phenanthrenequinone	Chemical
Abnormal behavior of blood concentration	Disease
Abnormal liver	Disease

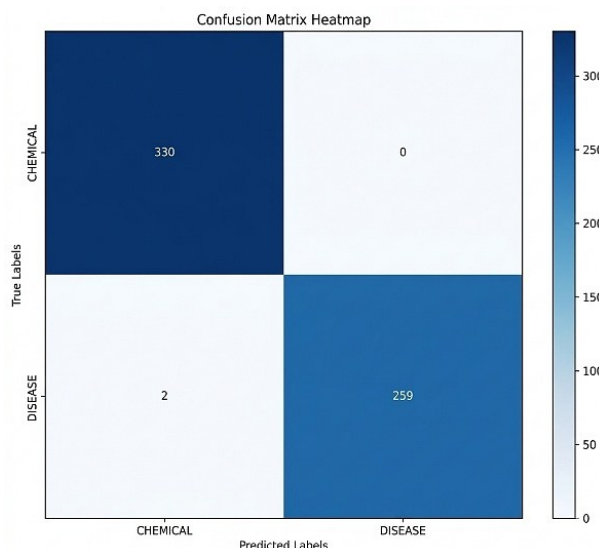


Figure 2. NER confusion matrix heatmap

3.2. Relation extraction: biomedical application on kratom

In biomedical research, the relationship between chemical compounds and diseases is essential for developing new therapies, understanding drug side effects, and uncovering disease mechanisms. This section discusses the implementation of NLP models to extract chemical and disease entities and their relationships from scientific texts. The connection between chemical and disease entities was determined by analyzing sentences that included both. Key verbs (such as “cause,” “lead,” and “associated”) were recognized as the links between the chemical and disease entities.

This study employs the BC5CDR model to extract the relationship between chemical compounds and diseases from scientific articles discussing the kratom plant (*Mitragyna speciosa*). Kratom has long been used traditionally as a medicine, but scientific literature shows a complex relationship with various medical conditions. Understanding the relationships between the chemicals in kratom and various diseases is crucial for evaluating their biomedical effects. This process aids in identifying potential therapeutic benefits, adverse effects, and interactions with other substances, offering a clearer insight into their pharmacological properties. By mapping these relationships, researchers can reveal important patterns that guide medical usage, regulatory decisions, and future research directions. Table 2 shows the results of kratom plant extraction, illustrating the relationship between chemicals and disease.

Table 2. Results of chemical and disease entity extraction using the BC5CDR model

Chemical	Disease	Relation
7-hydroxymitragynine	Chronic pain	Used
Mitragynine	Cholestatic hepatitis	Demonstrated
Mitragynine	Drug abuse	Required
Mitragynine	Psychotic symptoms	Associated
Mitragynine	Liver injury	Associated
Mitragynine	Neuropathic pain	Reduced

According to the results regarding the relationship between chemicals and diseases, three categories describe this relationship:

- i) Positive relationship: the use of kratom for medicine
 - 7-Hydroxymitragynine and chronic pain: this relationship suggests that the compound 7-hydroxymitragynine is “used” to treat chronic pain. This compound is known to have a stronger analgesic potential than morphine [31].
 - Mitragynine and opiate addiction: this relationship shows that mitragynine can “reduce” opioid addiction, supporting its role as a substitution therapy for opioid use disorder [32].
- ii) Negative relationships: risks and side effects
 - Mitragynine and liver injury, toxicity: this relationship indicates that mitragynine use is “associated” with liver injury and toxicity. These effects emphasize the possible risks associated with long-term or high-dose use of mitragynine [33].
 - Mitragynine and hemorrhagic stroke: this relationship suggests a possible connection between Mitragynine and hemorrhagic stroke [34]. These findings emphasize the necessity of closely monitoring the use of kratom.
- iii) Controversial relationship: potential for abuse: mitragynine and drug abuse, psychotic symptoms: this association indicates that mitragynine is associated with drug abuse and psychotic symptoms. These findings emphasize concerns regarding the potential for dependency and neuropsychological effects associated with kratom use [35].

The data visualization illustrating the relationship between chemicals and diseases in the kratom scientific article dataset is presented as a knowledge graph, specifically focusing on the mitragynine chemical type. The visualization of the knowledge graph shows the relationship between mitragynine and various medical conditions, offering a clear overview of its therapeutic benefits, side effects, and clinical potential. A knowledge graph utilizes a weighted graph structure, which is comprised of nodes and edges. Nodes represent chemical entities (such as mitragynine) and diseases (such as pain, depression, and seizure). Edges connect two nodes to show the relationship between them, with labels being verbs that indicate the nature of the relationship (such as alleviating, causes, and associated). The knowledge graph in Figure 3 clearly represents the relationship between mitragynine and various medical conditions. By combining both positive and negative relationships, this graph provides a useful tool for understanding the potential of mitragynine in greater depth. Visual information is essential for assessing the balance between the benefits and risks of mitragynine. The patterns observed in the graph guide further toxicology studies, molecular mechanisms, and clinical trials. Centrality metric calculations are used to assess the quality of relationships visualized in the knowledge graph. There are three types of centrality metrics used in network analysis. Degree centrality measures the extent to which a node is directly connected to other nodes. Betweenness centrality assesses how often a node acts as a link in the shortest path between other nodes. Closeness centrality evaluates how easily a node can access other nodes based on the shortest paths. These metrics help to understand the roles and importance of different nodes within a network. Table 3 presents a sample of the results from measuring centrality metrics on the extracted chemical disease relations.

The higher the degree centrality value of a node, the more direct relationships it has. According to the centrality metrics measurement results, neuropathic has the highest degree centrality value (34, 965, 034,

965, 034, 900). This indicates that neuropathic is connected to more nodes in the network, making it the entity with the most connections. Mitragynine has a significantly lower value (944, 055, 944, 055, 944), but it is still important as a representative of the key chemical in kratom. Pain (12, 937, 062, 937, 062, 900) and neuropathic pain (17, 482, 517, 482, 517, 400) play significant roles in the network, representing clinical conditions frequently associated with kratom use. In this result, every node has a betweenness centrality value of 0.0. This indicates that no node acts as a key connector in the network, all relationships are direct and do not require intermediary nodes. Closeness centrality indicates that higher values suggest faster access or closer relationships within a network. Neuropathic recorded the highest closeness centrality value of 34, 965, 034, 965, 034, 900, demonstrating that this node has relatively quick access to other nodes. This highlights its role as a major hub in the network. Pain and neuropathic pain play a crucial role in linking chemicals, such as mitragynine, to disease.

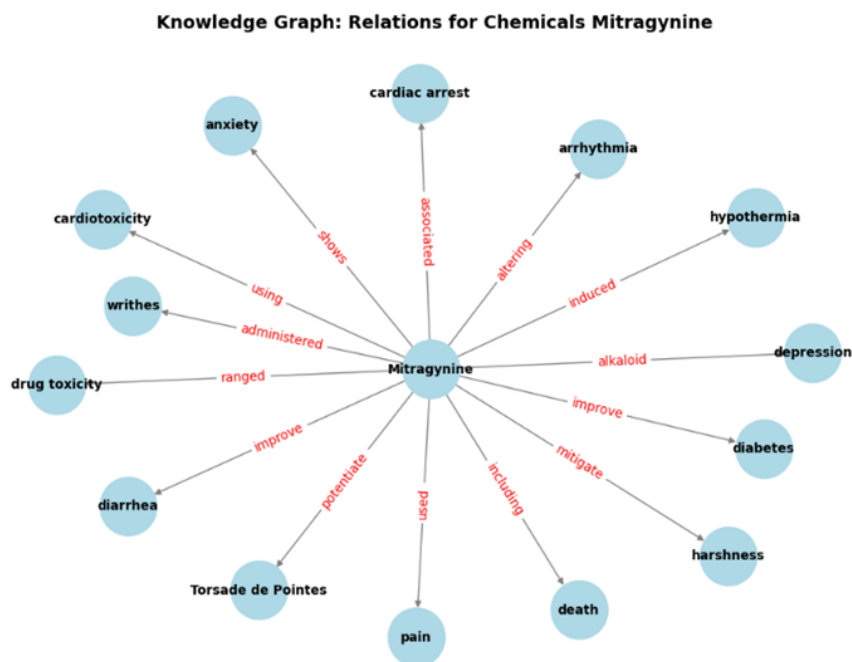


Figure 3. Knowledge graph illustrates the relationship between the chemical “mitragynine” and disease

Table 3. Results of centrality metrics for CDR

Node	Degree centrality	Betweenness centrality	Closeness centrality
Mitragynine	944.055.944.055.944	0.0	0.0
Pain	12.937.062.937.062.900	0.0	12.971.175.166.297.100
Chronic pain	2.097.902.097.902.090	0.0	2.097.902.097.902.090
Neuropathic	34.965.034.965.034.900	0.0	34.965.034.965.034.900
Hypersensitivity	6.993.006.993.006.990	0.0	6.993.006.993.006.990
Neuropathic pain	17.482.517.482.517.400	0.0	17.482.517.482.517.400

3.3. Topic modeling

The stages of topic modeling with BERTopic begin with document embedding, where document vectors are represented using sentence transformers. Next, HDBSCAN is used for document clustering, followed by dimension reduction using UMAP. Once each document in the corpus has been assigned to a cluster, a topic representation is generated using c-TF-IDF (class-based TF-IDF). The next step is to analyze the topic using a model that identifies five specific topics. This analysis produces two main outputs: the topics themselves and their corresponding probabilities. The topics represent the themes identified within a document, while the probabilities indicate the extent to which a document aligns with each given topic. The BERTopic model was executed three times to ensure consistency in the results. The findings indicate that the main topics generated remain consistent with each model run, as illustrated in Figure 4. Figure 3 illustrates the frequency distribution of topics across all analyzed documents. A total of 149 datasets belong to topic 1, 111 datasets to topic 2, 20 datasets each to topics 3 and 4, and 13 datasets to topic 5. Additionally, 82 datasets are not classified into any specific topic (referred to as topic -1). This lack of classification may be attributed to the limited number of specified topics, which are five in total. This restriction on the number of topics is

implemented to maintain a high coherence score, which makes the resulting topics easier for people to understand. A higher coherence score indicates that the topics are more relevant and meaningful within the context of human language.

Figure 5 presents a bar chart visualization of the most significant words associated with each topic, based on their c-TF-IDF scores. The words at the top of the chart indicate a higher frequency of occurrence within that topic cluster, while those at the bottom represent lower frequencies. This diagram also facilitates comparisons among sets of words within each topic. For instance, the most relevant words from topic number 2 include opioid, report, pain, withdrawal, and user. This highlights the relationship between the benefits of kratom, particularly in alleviating symptoms associated with opioid use.

The coherence score of 0.5567 indicates a moderate but acceptable level of thematic consistency for a specialized biomedical corpus. As demonstrated by Egger and Yu [36] in their comparative study of transformer-based topic models, BERTopic typically yields more interpretable clusters than traditional LDA in technical domains, even if the coherence metrics remain in the 0.5–0.6 range. Our findings confirm that c-TF-IDF effectively captures the distinct nuances between ‘opioid withdrawal’ and ‘toxicity’ topics, which are essential for understanding the multifaceted nature of kratom research [37].

Topic	Count	Name	Representation	Representative_Docs
0	-1	82 -1_effect_mitragynin_use_user	[effect, mitragynin, use, user, report, extrac...	[indigen known tradit addict potenti asia year...
1	0	149 0_mitragynin_alkaloid_effect_opioid	[mitragynin, alkaloid, effect, opioid, recepto...	[mitragynin indol alkaloid isol previous repor...
2	1	111 1_opioid_report_pain_withdraw	[opioid, report, pain, withdraw, user, effect,...	[opioidreceptor mediat effect emerg potenti su...
3	2	20 2_exposur_poison_report Associ	[exposur, poison, report, associ, effect, amon...	[interest remedi increas western countri recen...
4	3	20 3_mitragynin_concentr_case_report	[mitragynin, concentr, case, report, detect, d...	[substanc major chemic compon mitragynin eluci...
5	4	13 4_liver_injuri_dili_report	[liver, injuri, dili, report, hepatotox, herba...	[druginduc liver injuri dili among common caus...

Figure 4. Topics generated using BERTopic

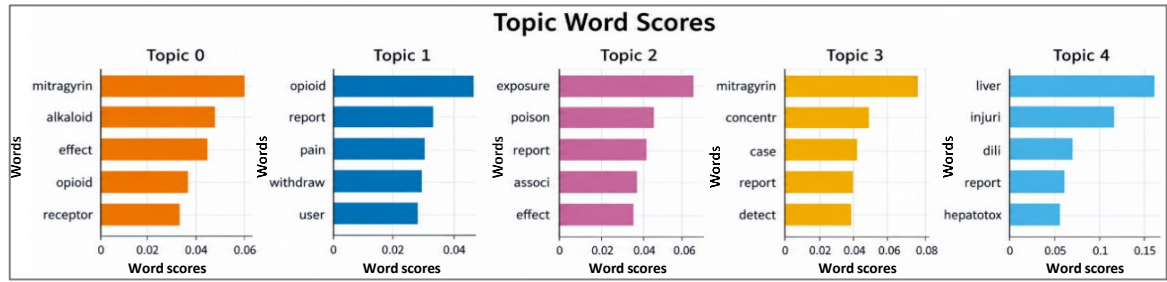


Figure 5. Visualization of topics based on topic word scores

4. CONCLUSION

This study shows that text mining methods, such as NER, RE, and topic modeling, effectively extract and analyze biomedical knowledge from scientific literature. The findings show that these methods effectively identify important entities, relationships, and thematic patterns in kratom research, promoting biomedical knowledge discovery and literature analysis. Additionally, network-based analysis highlights the key role of clinically relevant concepts in organizing CDR. Meanwhile, topic modeling emphasizes the main themes related to toxicity and clinical implications. This study has some limitations related to linguistic variability and semantic ambiguity in biomedical texts, which may affect the consistency of information extraction and thematic representation. Additionally, the quality of topics generated through topic modeling is influenced by the size of the corpus and its semantic diversity, which may limit the coverage of less prominent concepts. Future work should focus on expanding domain-specific corpora, integrating biomedical ontologies to improve semantic disambiguation, and incorporating expert validation to enhance the robustness and interpretability of text mining results.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to the Faculty of Mathematics and Natural Sciences at Tanjungpura University for providing the computing facilities that supported the completion of this research.

FUNDING INFORMATION

This work was supported by the Faculty of Mathematics and Natural Sciences, Tanjungpura University, Indonesia, through the Budget Execution Document (DIPA) for the 2023 fiscal year.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nurul Mutiah	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	
Dian Prawira		✓		✓		✓		✓	✓	✓				

- C : Conceptualization
- M : Methodology
- So : Software
- Va : Validation
- Fo : Formal analysis
- I : Investigation
- R : Resources
- D : Data Curation
- O : Writing - Original Draft
- E : Writing - Review & Editing
- Vi : Visualization
- Su : Supervision
- P : Project administration
- Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that supports the findings of this study are available from the corresponding author, [NM], upon reasonable request.

REFERENCES

[1] I. C. Wiest *et al.*, “Privacy-preserving large language models for structured medical information retrieval,” *npj Digital Medicine*, vol. 7, no. 1, Sep. 2024, doi: 10.1038/s41746-024-01233-2.

[2] N. R.-Ferreira *et al.*, “The treasury chest of text mining: piling available resources for powerful biomedical text mining,” *BioChem*, vol. 1, no. 2, pp. 60–80, Jul. 2021, doi: 10.3390/biochem1020007.

[3] B. B. L. P. de Vries, M. van Smeden, F. R. Rosendaal, and R. H. H. Groenwold, “Title, abstract, and keyword searching resulted in poor recovery of articles in systematic reviews of epidemiologic practice,” *Journal of Clinical Epidemiology*, vol. 121, pp. 55–61, 2020, doi: 10.1016/j.jclinepi.2020.01.009.

[4] S. Zhao, C. Su, Z. Lu, and F. Wang, “Recent advances in biomedical literature mining,” *Briefings in Bioinformatics*, vol. 22, no. 3, pp. 1–19, 2021, doi: 10.1093/bib/bbaa057.

[5] A. R. Hsieh and C. Y. Tsai, “Biomedical literature mining: graph kernel-based learning for gene-gene interaction extraction,” *European Journal of Medical Research*, vol. 29, no. 1, 2024, doi: 10.1186/s40001-024-01983-5.

[6] R. K. Pathak, D. B. Singh, and R. Singh, “Introduction to basics of bioinformatics,” in *Bioinformatics: Methods and Applications*, Elsevier, 2021, pp. 1–15, doi: 10.1016/B978-0-323-89775-4.00006-7.

[7] A. Krithara, J. G. Mork, A. Nentidis, and G. Paliouras, “The road from manual to automatic semantic indexing of biomedical literature: a 10-year journey,” *Frontiers in Research Metrics and Analytics*, vol. 8, 2023, doi: 10.3389/frma.2023.1250930.

[8] A. Lamurias, S. Jesus, V. Neveu, R. M. Salek, and F. M. Couto, “Information retrieval using machine learning for biomarker curation in the exposome-explorer,” *Frontiers in Research Metrics and Analytics*, vol. 6, 2021, doi: 10.3389/frma.2021.689264.

[9] H. Zong *et al.*, “Advancing Chinese biomedical text mining with community challenges,” *Journal of Biomedical Informatics*, vol. 157, 2024, doi: 10.1016/j.jbi.2024.104716.

[10] N. Perera, M. Dehmer, and F. E.-Streib, “Named entity recognition and relation detection for biomedical information extraction,” *Frontiers in Cell and Developmental Biology*, vol. 8, Aug. 2020, doi: 10.3389/fcell.2020.00673.

[11] S. Jing *et al.*, “Correlation analysis and text classification of chemical accident cases based on word embedding,” *Process Safety and Environmental Protection*, vol. 158, pp. 698–710, Feb. 2022, doi: 10.1016/j.psep.2021.12.038.

[12] A. Udgave and P. Kulkarni, “Text mining and text analytics of research articles,” *PalArch's Journal of Archaeology of Egypt / Egyptology*, vol. 17, no. 6, pp. 4483–4489, 2020.

[13] K. Thakur and V. Kumar, “Application of text mining techniques on scholarly research articles: methods and tools,” *New Review of Academic Librarianship*, vol. 28, no. 3, pp. 279–302, Jul. 2022, doi: 10.1080/13614533.2021.1918190.

[14] Q. Chen *et al.*, “An extensive benchmark study on biomedical text generation and mining with ChatGPT,” *Bioinformatics*, vol. 39, no. 9, 2023, doi: 10.1093/bioinformatics/btad557.

[15] B. D. Tho, M. T. Nguyen, D. T. Le, L. L. Ying, S. Inoue, and T. T. Nguyen, “Improving biomedical named entity recognition with additional external contexts,” *Journal of Biomedical Informatics*, vol. 156, 2024, doi: 10.1016/j.jbi.2024.104674.

[16] W. Zheng, W. Hou, and J. C. W. Lin, “A deep learning-based feature entity relationship extraction method for telemedicine sensing big data,” *Mobile Networks and Applications*, vol. 28, no. 6, pp. 1987–1996, 2023, doi: 10.1007/s11036-022-02024-3.

[17] Y. J. Park, M. A. Lee, G. J. Yang, S. J. Park, and C. B. Sohn, “Web interface of NER and RE with BERT for biomedical text mining,” *Applied Sciences*, vol. 13, no. 8, 2023, doi: 10.3390/app13085163.

[18] H. U. Haq, V. Kocaman, and D. Talby, “Connecting the dots in clinical document understanding with relation extraction at scale,” *Software Impacts*, vol. 12, 2022, doi: 10.1016/j.simpa.2022.100294.

- [19] J. Zhang, Y. Wang, H. Xu, and L. Li, "A study of biomedical relation extraction using GPT models," in *AMIA Summits on Translational Science Proceedings*, 2024.
- [20] Y. Jia, H. Wang, Z. Yuan, L. Zhu, and Z. L. Xiang, "Biomedical relation extraction method based on ensemble learning and attention mechanism," *BMC Bioinformatics*, vol. 25, no. 1, 2024, doi: 10.1186/s12859-024-05951-y.
- [21] Z. Cai, Y. Fan, and J. Xin, "Research on medical text named entity recognition model based on prompt contrastive learning," in *ACM International Conference Proceeding Series*, 2024, pp. 139–144, doi: 10.1145/3705677.3705701.
- [22] L. Luo, Z. Yang, M. Cao, L. Wang, Y. Zhang, and H. Lin, "A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature," *Journal of Biomedical Informatics*, vol. 103, 2020, doi: 10.1016/j.jbi.2020.103384.
- [23] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang, "MRC4BioER: joint extraction of biomedical entities and relations in the machine reading comprehension framework," *Journal of Biomedical Informatics*, vol. 125, 2022, doi: 10.1016/j.jbi.2021.103956.
- [24] H. Lu, L. Li, Z. Li, and S. Zhao, "Extracting chemical-induced disease relation by integrating a hierarchical concentrative attention and a hybrid graph-based neural network," *Journal of Biomedical Informatics*, vol. 121, 2021, doi: 10.1016/j.jbi.2021.103874.
- [25] J. Chen, B. Hu, W. Peng, Q. Chen, and B. Tang, "Biomedical relation extraction via knowledge-enhanced reading comprehension," *BMC Bioinformatics*, vol. 23, no. 1, Dec. 2022, doi: 10.1186/s12859-021-04534-5.
- [26] G. Agrawal, Y. Deng, J. Park, H. Liu, and Y. C. Chen, "Building knowledge graphs from unstructured texts: applications and impact analyses in cybersecurity education," *Information*, vol. 13, no. 11, pp. 1–22, 2022, doi: 10.3390/info13110526.
- [27] T. Limcharoen *et al.*, "Inhibition of α -glucosidase and pancreatic lipase properties of *Mitragyna speciosa* (korth.) havil. (kratom) leaves," *Nutrients*, vol. 14, no. 19, Sep. 2022, doi: 10.3390/nu14193909.
- [28] S. C. Eastlack, E. M. Cornett, and A. D. Kaye, "Kratom—pharmacology, clinical implications, and outlook: a comprehensive review," *Pain and Therapy*, vol. 9, no. 1, pp. 55–69, Jun. 2020, doi: 10.1007/s40122-020-00151-x.
- [29] S. C. P. N and A. Adhari, "Policy on the criminalization of kratom plants as narcotics category I," in *Proceedings of the 3rd Tarumanagara International Conference on the Applications of Social Sciences and Humanities (TICASH 2021)*, 2022, vol. 655, pp. 1332–1335, doi: 10.2991/assehr.k.220404.213.
- [30] Z. L. Chen, "Research and application of clustering algorithm for text big data," *Computational Intelligence and Neuroscience*, pp. 1–8, Jun. 2022, doi: 10.1155/2022/7042778.
- [31] A. B.-Rod *et al.*, "Evaluation of the rewarding effects of mitragynine and 7-hydroxymitragynine in an intracranial self-stimulation procedure in male and female rats," *Drug and Alcohol Dependence*, vol. 215, no. 5, Oct. 2020, doi: 10.1016/j.drugalcdep.2020.108235.
- [32] A. S. Alford, H. L. Moreno, M. M. Benjamin, C. F. Dickinson, and M. T. Hamann, "Exploring the therapeutic potential of mitragynine and corynoxine: kratom-derived indole and oxindole alkaloids for pain management," *Pharmaceuticals*, vol. 18, no. 2, Feb. 2025, doi: 10.3390/ph18020222.
- [33] K. Roma, S. Mohammed, B. Sieck, K. Naik, and S. Wahid, "Kratom-induced acute liver injury: a case study and the importance of herbal supplement regulation," *Journal of Hepatology*, vol. 79, no. 2, pp. 581–584, 2023, doi: 10.1016/j.jhep.2023.04.026.
- [34] N. Nacca *et al.*, "Kratom adulterated with phenylethylamine and associated intracerebral hemorrhage: linking toxicologists and public health officials to identify dangerous adulterants," *Journal of Medical Toxicology*, vol. 16, no. 1, pp. 71–74, Jan. 2020, doi: 10.1007/s13181-019-00741-y.
- [35] M. Awad, H. H. Burke, and S. A. Oakman, "Kratom-induced psychiatric decompensation and paranoid delusions," *Cureus*, vol. 16, no. 2, pp. 2–5, Feb. 2024, doi: 10.7759/cureus.54626.
- [36] R. Egger and J. Yu, "A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts," *Frontiers in Sociology*, vol. 7, pp. 1–16, 2022, doi: 10.3389/fsoc.2022.886498.
- [37] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

BIOGRAPHIES OF AUTHORS



Nurul Mutiah    holds a Master of Informatics degree from Bandung Institute of Technology, Indonesia. She is a lecturer at the Department of Information Systems at Tanjungpura University, Indonesia since 2015. In addition, she is serving as head of a research group on IT governance. Her research interests are text mining, enterprise architecture, IT governance and audit, systems dynamics, and information systems management. In this research, she contributed significantly to the data collection process via NCBI Entrez API, the technical implementation of named entity recognition (NER) and relation extraction (RE) models using SciSpacy, and the development of the knowledge graph visualization. She has published over 40 papers in international journals and conferences, as well as one book on text mining. She can be contacted at email: nurul@sisfo.untan.ac.id.



Dian Prawira    is a lecturer in the Department of Information Systems at Tanjungpura University, Indonesia. He completed his postgraduate education, earning a Master of Information Technology from the Department of Electrical Engineering and Information Technology at Gadjah Mada University in 2011. His primary research concentration and interests lie in the fields of text mining, natural language processing (NLP), and web and mobile application development. He has authored a notable book titled 'Text mining.' In this research, he contributed significantly to the data collection process via NCBI Entrez API and the technical implementation of named entity recognition (NER). He is committed to advancing information systems through data-driven research and practical technological innovation. He can be contacted at email: dianprawira@sisfo.untan.ac.id.