https://doi.org/10.65091/icicset.v2i1.15

Query Refinement using Latent Dirichlet Allocation

Dhiraj Poudel
Nepal College of Information Technology
Lalitpur, Nepal
dhiraj.201612@ncit.edu.np

Rishav Dahal
Nepal College of Information Technology
Lalitpur, Nepal
rishav.201629@ncit.edu.np

Sandhya Gotame
Nepal College of Information Technology
Lalitpur, Nepal
sandhya.201637@ncit.edu.np

Binit Bikram KC
Nepal College of Information Technology
Lalitpur, Nepal
binit.201610@ncit.edu.np

Ashim Khadka*
Nepal College of Information Technology
Lalitpur, Nepal
ashim.khadka@ncit.edu.np

*Corresponding author

***Abstract*—This study introduces a query refinement technique that leverages Latent Dirichlect Algorithm (LDA) and coherence techniques. The proposed method falls under unsupervised learning, firstly preprocessing large text corpus, applying topic modelling via LDA and finding the optimal number of topics using coherence scores eventually leading to the topics which are meaningful and relevant to the text corpus. We made use of coherence measures to make sure the topics extracted from the corpus are relevant to each other and understandable. The results demonstrated that the method was successful in topic extraction and modelling of large text corpus. However, the results significantly vary on the quality of text corpus on which the method is being applied.**

***Index Terms*—Semantic Query Refinement, Latent Dirichlet Allocation (LDA), Topic Modeling, Unsupervised Learning**

I. Introduction

As the amount of information keeps increasing in this technology-driven era, users find it increasingly difficult to retrieve genuinely relevant information [1]. Studies show that too much digital information can overwhelm users, reduce productivity and make search tasks less efficient [2].

Traditional search engine systems often rely on keyword-based approaches and perform lexical search, where queries are matched exactly with the document terms using techniques like inverted indexing and term frequency-inverse document frequency (TF-IDF) [3]. Such lexical searching is limited, meaning it fails to understand context, meaning, synonyms, the relationship between words or the user's true intent [4]. This discrepancy between user expectations and search results is frequently caused by imprecise or ambiguous inquiries that fail to sufficiently convey the user's emotional undertone or contextual needs. To address these shortcomings, semantic query refinement techniques have emerged aiming to interpret the user's true intent and meaning.

This study proposes a NLP-based query refinement technique which, instead of relying on literal keyword matching, actually interprets semantic meaning to increase search results. The system functions as an intermediary API which enhances user queries before they are sent to the appropriate knowledge base or the search engine. Since the system is developed as a flexible API, it can be easily adopted across a variety of platforms and is applicable in a wide range of scenarios, hence improving search accuracy.

The study explored and contrasted the effectiveness of various language representation and topic modelling models, such as Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI), and Bidirectional Encoder Representations from Transformers (BERT), in order to maximize keyword generation and enhance the relevancy of search results. The results generated by each model were evaluated to determine which approach was most suitable in the domain-specific queries for capturing semantic meaning.

Latent Dirichlet Allocation is a topic modelling technique which identifies the primary topics and how they are distributed throughout a collection of documents. It is a probabilistic technique for finding themes in massive text corpora and a Bayesian approach to topic modelling [5].

Latent Semantic Indexing (LSI) uses Singular Value Decomposition to reduce a term document matrix's dimensionality, reveal latent semantic structure and capture relationships between conceptually similar words even if they don't appear together explicitly [6].

BERT is a transformer-based language model that has already been trained. It processes text bidirectionally, capturing word relationships in both directions using methods like Next Sentence Prediction and Masked Language Modelling [7].

II. Literature Review

Almost all traditional search systems use lexical search, which is keyword-based and retrieves results by matching the query terms exactly to the document terms. It does so often using inverted indexing and TF-IDF scoring. While this method is precise and efficient, since the results you'll get will exactly match your query term, it cannot handle variations such as synonyms, misspellings or different variants of words. It fails to understand the context and treats each term separately, which makes it bad for vague or complicated questions. It also fails to capture user intent, reducing the relevancy of the retrieved information [1], [8], [9].

By adding related keywords, such as synonyms, word variations, or any other contextually relevant terms, query

expansion improves and refines user input, increasing retrieval accuracy and resolving issues with lexical search. Traditional Query Expansion techniques include synonym expansion, stemming and pseudo-relevance feedback alongside classic relevance feedback processes like Rocchio Algorithm [10]–[13]. Advanced semantic enrichment methods apply ontologies and knowledge bases together with word embeddings to understand the semantic meaning of words while generating contextually suitable expansions [14], [15].

Topic modelling identifies underlying semantic patterns in text by analyzing the frequent co-occurrence of words across different topics. LDA represents generative frameworks which deliver interpretable models to users. Conventional methods, including LDA demonstrate poor performance when dealing with complex meanings and large vocabularies. Embedded topic models (ETM) resolve these problems by integrating pretrained word embeddings into a probabilistic framework to achieve interpretability alongside semantic richness while maintaining effective performance with large vocabularies and uncommon terms [16], [17].

The field of neural topic modelling has progressed through the implementation of transformer-based embeddings along with clustering techniques through the models BERTopic and Top2Vec. BERTopic produces meaningful topic representations which maintain interpretability and flexibility for short or multilingual texts after applying BERT embeddings, followed by UMAP dimensionality reduction and HDBSCAN clustering and class-based TF-IDF cluster summarization [18], [19]. Top2Vec generates joint semantic topic vectors by combining document and word embeddings in a process which resembles LDA.

The approach determines topic numbers automatically while minimizing preprocessing needs and generates topic clusters that demonstrate excellent semantic cohesion [20], [21]. The integration of individual semantic context and topic structures through embeddings enables improved topic modelling while enhancing semantic query refinement systems.

The study evaluated SVD/LSA and LDA unsupervised topic extraction methods on large-scale text corpora. Both methods effectively discovered meaningful topics as LDA generated probabilistic topic distributions from text data, while LSA detected latent patterns [22]. The research indicates that both models rely strongly on the quality and representativeness of text corpora for better performance. Automated coherence measurement requires further development because manual interpretation continues to be necessary for connecting generated themes to human understanding.

The study demonstrates a hierarchical deep learning system to classify queries in QA systems, but this study does not focus on it as the main point. Their research applies embeddings to enhance semantic understanding but uses supervised learning for question categorization instead of unsupervised topic or query refinement methods. The study investigates NLP semantic modelling yet remains outside the scope of unsupervised, coherence-driven topic modelling frameworks that form the core of this research [23].

III. Methodology

The research applies multiple stages of unsupervised learning through Latent Dirichlet Allocation (LDA), which uses coherence metrics to transform medical question–answer pairs into enriched keyword-based queries.

A. Data Collection

We utilize the MedQuAD dataset, a comprehensive collection of 47,457 medical question–answer pairs compiled from twelve U.S. National Institutes of Health (NIH) websites, covering a wide range of medical topics such as treatments, diagnoses, and side effects [24]. This domain-specific data provides a rich and diverse linguistic foundation for developing and evaluating our semantic query refinement framework.

B. Text Preprocessing

The answer texts undergo transformation from their original full paragraph form into a structured question $\rightarrow$ keyword(s) format. The preprocessing stages include:

1) Cleaning and normalizing text (e.g., removing punctuation, lowercasing).
2) Tokenizing text and applying lemmatization.
3) Filtering tokens by part-of-speech, retaining only nouns, adjectives, and verbs to emphasize semantic content while discarding less meaningful tokens.

The advanced representation system emphasizes linguistically meaningful components which establishes an optimal foundation for semantic modeling purposes.

C. Coherence Guided Topic Modeling

LDA topic modeling operates at the core of this methodology where topic coherence metrics guide the selection of the best number of topics.

1) Multiple LDA models are trained between 5 and 50 topics to determine the optimal number of topics.
2) Each model undergoes coherence scoring with c_ev coherence as the primary metric which evaluates topic interpretability and consistency through frameworks like Gensim's CoherenceModel [25].

The model that achieves the highest coherence score stands as the optimal configuration because it validates topics statistically and makes them understandable to humans.

D. Implementation Details and Reproducibility

The experiments were conducted using two separate computing environments to support different stages of the proposed framework. Latent Dirichlet Allocation (LDA) topic modeling was performed on Google Colab (Free Tier), while BERT-based clustering experiments were executed on the Kaggle platform.

The Google Colab environment provided TPU-based runtime resources with approximately 47.0 GB of RAM and 225.3 GB of available disk space. The BERT clustering experiments were conducted on Kaggle using dual NVIDIA T4 GPUs, with an average runtime of 3 minutes and 28 seconds per execution.

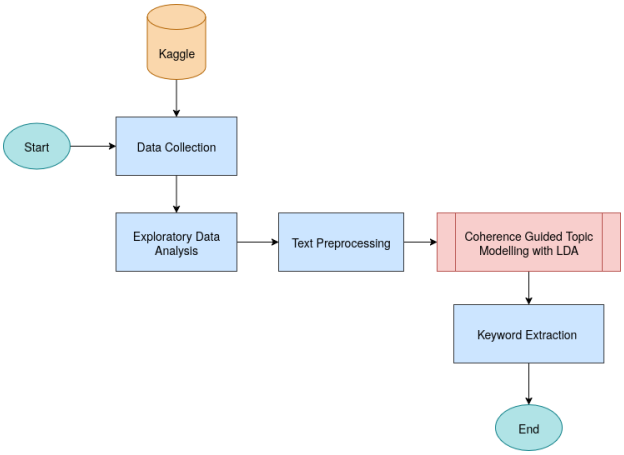

Fig. 1. Research Steps.

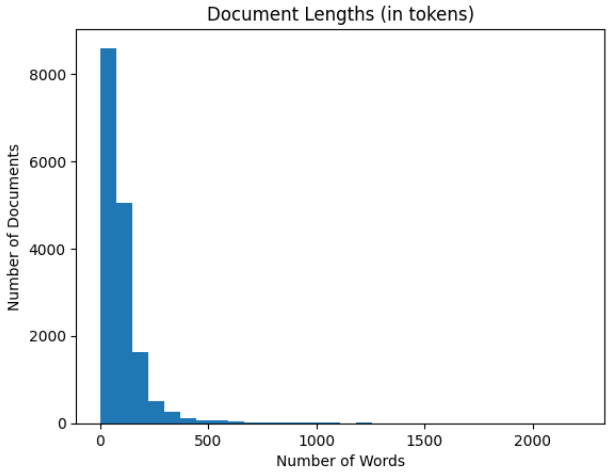

Fig. 2. Document Lengths

Text preprocessing was carried out using spaCy's *en_core_web_sm* model. The preprocessing pipeline included text normalization, tokenization, lemmatization, and part-of-speech filtering, where only nouns, verbs, and adjectives were retained to emphasize semantically meaningful terms.

LDA topic modeling was implemented using *gensim* version 4.3.2. Multiple models were trained by varying the number of topics from 5 to 50. Each model was trained for 15 iterations with a fixed random state of 100. Topic coherence was evaluated using the *c_v* coherence metric, and the model with the highest coherence score was selected as the optimal configuration.

For semantic clustering, sentence embeddings were generated using the *all-MiniLM-L6-v2* model from the Sentence-Transformers framework. The resulting embeddings were clustered using the k-means algorithm implemented via *scikit-learn* version 1.7.1. Cluster-level keywords were extracted using Term Frequency–Inverse Document Frequency (TF–IDF) weighting.

All trained models, dictionaries, intermediate results, plots, and generated labels were stored to support reproducibility and potential reuse. A complete reproduction of the experiments requires the same software versions and comparable computational resources.

IV. Results and Discussion

This section presents the key findings of our study and interprets their significance in the context of semantic query refinement.

A. Exploratory Data Analysis

We conducted exploratory data analysis of the MedQuAD dataset before topic modeling to evaluate both its structural and lexical characteristics. The initial analysis helped establish proper guidelines for preprocessing choices and topic modeling configuration.

The majority of answer texts in the dataset contain fewer than 300 tokens as illustrated in Figure 2. A smaller proportion extends beyond 500 tokens, and only a handful exceed 1,000 tokens. The dataset comprises both brief factual answers and extensive explanatory texts which create this long-tail distribution pattern. Topic model development requires knowledge about document length variability because both very brief and very lengthy documents affect topic coherence and interpretability.

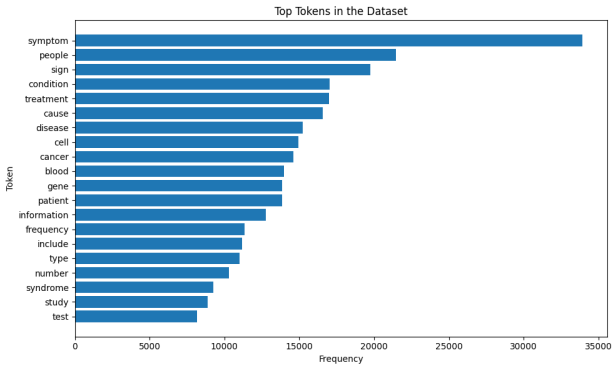

Fig. 3. Top Tokens

The dataset's most common tokens appear in Figure 3 after the preprocessing step. The dataset contains numerous medical terms including "symptom" and "condition" along with "treatment" and "disease" and "cancer" and "gene" which demonstrate its medical focus. The high frequency of patient-related words including "people" and "patient" in the dataset shows it presents information from clinical and non-clinical viewpoints. The analysis demonstrates that the dataset contains a substantial number of medical terms which makes LDA appropriate for discovering meaningful topics.

B. Topic Modeling

We used coherence scores as an evaluation metric to determine the best number of topics for our analysis.

The evaluation of topic models through coherence scores represents a popular method to measure topic interpretability

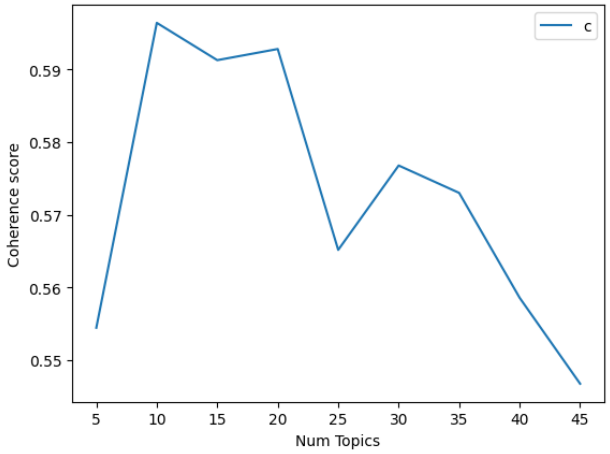

Fig. 4. Coherence score per topic.

because it evaluates semantic similarity between key topic words. The evaluation process involved computing coherence scores for LDA models which had topic counts ranging from 5 to 50 with 5 as the increment.

The coherence scores from different models showed a distinct maximum at 10 topics as displayed in Figure 4. The model with 10 topics achieved the highest semantic consistency of keywords which balanced between too few topics and too many topics. The evaluation results identified the 10-topic LDA model as the best configuration which would be used for keyword extraction and semantic query refinement.

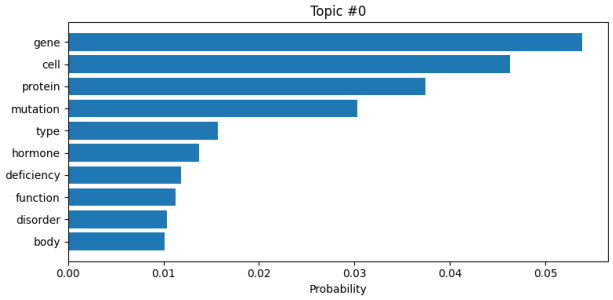

Fig. 5. Top keywords for topic 0.

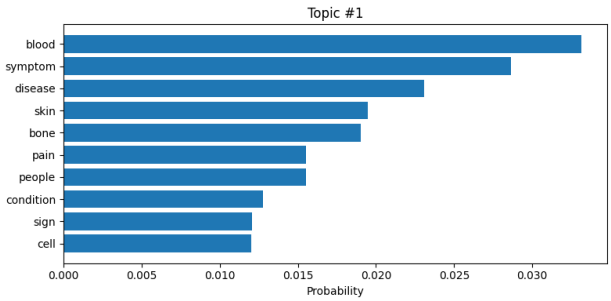

Fig. 6. Top keywords for topic 1.

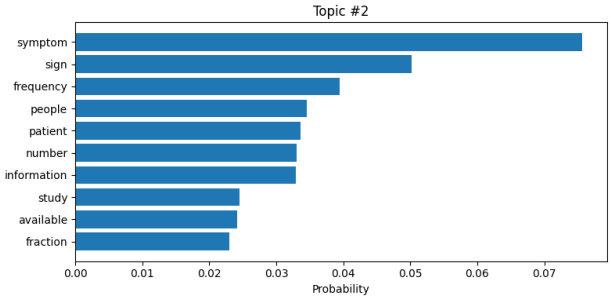

Fig. 7. Top keywords for topic 2.

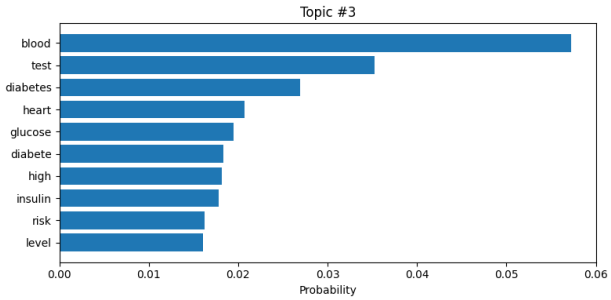

Fig. 8. Top keywords for topic 3.

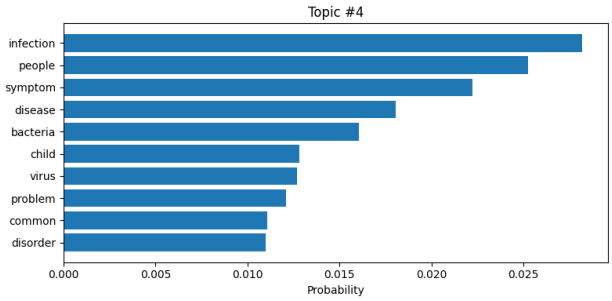

Fig. 9. Top keywords for topic 4.

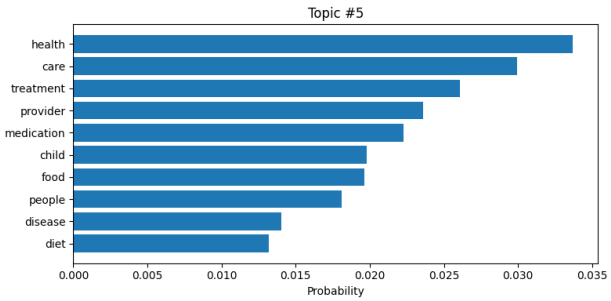

Fig. 10. Top keywords for topic 5.

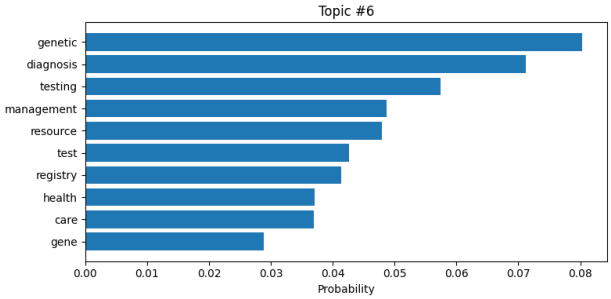

Fig. 11. Top keywords for topic 6.

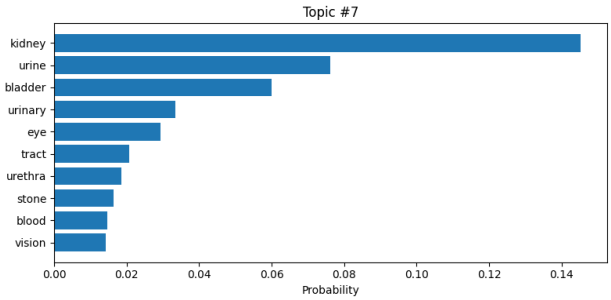

Fig. 12. Top keywords for topic 7.

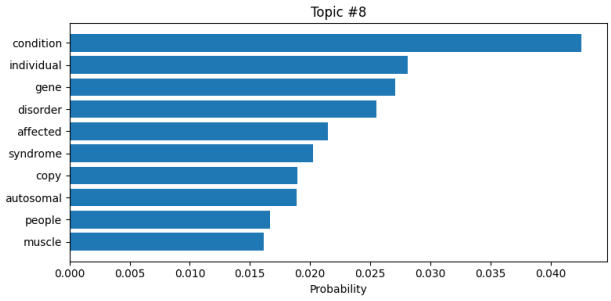

Fig. 13. Top keywords for topic 8.

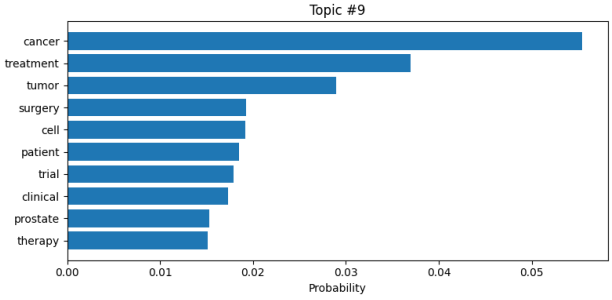

Fig. 14. Top keywords for topic 9.

The top keywords from each topic served as the basis for further interpreting the selected 10-topic model. Figures 5 through 14 presents representative words grouped by topic which demonstrate semantically coherent clusters such as gene–cell–protein in Topic 0 and health–care–treatment in Topic 5. The groupings demonstrate that the model effectively identifies concepts which hold medical meaning.

Fig. 15. Topic Distance Map

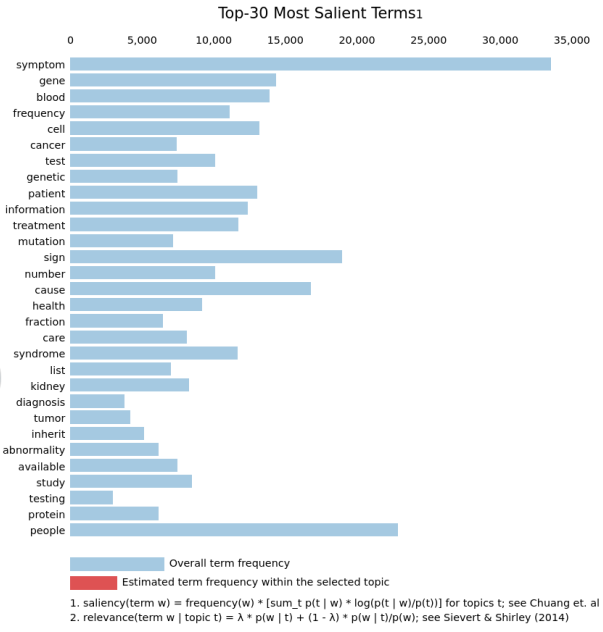

Fig. 16. Salient Terms

The visual representation produced through PyLDAVis, shown in Figure 15, displays topic similarity and word saliency

in two dimensions. Each topic is represented as a circle, and its size corresponds to its overall frequency. The distance between the circles shows how different the topics are from each other. In addition to this, Figure **??** displays the most salient words for each topic. The two visualizations together demonstrate that the topics maintain a good level of separation while having little intersection, which supports their interpretability.

C. Alternative Unsupervised Approaches

This research primarily studied Latent Dirichlet Allocation (LDA) but also tested Latent Semantic Indexing (LSI) and BERT embeddings to determine their performance in discovering semantic patterns.

1) Latent Semantic Indexing (LSI): The LSI model trained on the identical preprocessed corpus produced topics which lacked the coherence that LDA exhibited. The model identified some meaningful connections between topics yet showed increased topic blending which weakened the clarity of results.

2) BERT Embeddings: BERT sentence-level embeddings of answers were clustered through k-means as part of the BERT-based analysis. The processing method created groups of documents which matched semantic similarity better than other clustering techniques. The clusters proved difficult to interpret through explicit keyword topics since BERT embeddings operate through contextualized semantics instead of sparse word distributions.

D. Case Study: Query Results Across Models

Following the comparative analysis of topic modeling and embedding-based approaches, this section presents a case study evaluating how LDA, LSI, and BERT respond to real-world medical queries. Three representative queries were selected, ranging from vague symptom descriptions to disease-oriented questions, in order to assess semantic relevance and precision across models.

1) Query 1: Pain in Urinary Tract: To evaluate symptom-level semantic matching, the query "pain in urinary tract" was tested against the trained models. Table I summarizes the top keywords generated by each approach.

TABLE I
COMPARISON OF KEYWORDS GENERATED FOR "PAIN IN URINARY TRACT"

LDA (with Coherence)	**LSI**	**BERT**
symptom, cause, infection, child, disease, pain, people, diarrhea, intestine, tract	cancer, resource, blood, management, trial, clinical, diagnosis, registry, tumor, heart	kidney, urine, bladder, blood, urinary, stone, cause, disease, tract, dialysis

2) Query 2: I Have Pain in My Knee: The second query represents a vague symptom-based input frequently encountered in medical search scenarios. The extracted keywords are shown in Table II.

LDA again captured broad symptom-related terms but lacked anatomical specificity. LSI returned research-oriented and unrelated medical terminology, demonstrating weak semantic alignment. BERT provided the most coherent representation by focusing on symptom-level descriptors closely related to pain and medical conditions.

TABLE II
COMPARISON OF KEYWORDS GENERATED FOR "I HAVE PAIN IN MY KNEE"

LDA (with Coherence)	**LSI**	**BERT**
symptom, cause, infection, child, disease, pain, people, diarrhea, intestine, tract	cancer, resource, blood, management, trial, clinical, diagnosis, registry, tumor, heart	symptom, pain, sign, blood, cause, skin, disease, people, condition

3) Query 3: What Causes Diabetes?: The third query represents a disease-centric information need focused on causation. The resulting keywords are presented in Table III.

TABLE III
COMPARISON OF KEYWORDS GENERATED FOR "WHAT CAUSES DIABETES?"

LDA (with Coherence)	**LSI**	**BERT**
blood, kidney, diabetes, people, disease, heart, high, risk, level	cancer, resource, blood, management, trial, clinical, diagnosis, registry, tumor, heart	blood, diabetes, risk, high, cholesterol, food, glucose, insulin

For this query, LDA identified frequently co-occurring disease-related terms but did not clearly express causal relationships. LSI again produced unrelated clinical research terms. In contrast, BERT successfully highlighted medically meaningful causal factors such as *insulin*, *glucose*, *cholesterol*, and *dietary components*, demonstrating superior semantic understanding.

4) Overall Observation: Across all three queries, the BERT-based method consistently achieved the most accurate semantic matching. While LDA provided acceptable results through statistically derived topic distributions, it often included generic terms that reduced precision. LSI showed consistently poor performance for short medical queries. These results indicate that contextual embedding-based models are better suited for real-world medical query understanding and semantic search applications.

REFERENCES

[1] M. Arnold, M. Goldschmitt, and T. Rigotti, "Dealing with information overload: A comprehensive review," *Frontiers in Psychology*, vol. 14, Art. no. 1122200, 2023. [Online]. Available: https://doi.org/10.3389/fpsyg.2023.1122200

[2] P. Brey, "Foreword," in *Information Overload: An International Challenge for Professional Engineers and Technical Communicators*. Leiden, Netherlands: Brill, 2012. [Online]. Available: https://brill.com/downloadpdf/display/book/9789004549906/BP000016.pdf

[3] A. Jalilifard, V. F. Caridá, A. F. Mansano, R. S. Cristo, and F. P. C. da Fonseca, "Semantic sensitive TF-IDF to determine word relevance in documents," in *Advances in Computing and Network Communications*, S. M. Thampi *et al.*, Eds. Singapore: Springer, 2021, pp. 265–277. [Online]. Available: https://doi.org/10.1007/978-981-33-6987-0_27

[4] C. Dean, "SEO's evolution: Lexical search vs. semantic search revealed," *SurgeGraph Blog*, Mar. 19, 2025. [Online]. Available: https://surgegraph.io/seo/lexical-search-vs-semantic-search

[5] J. Murel and E. Kavlakoglu, "What is latent Dirichlet allocation?," *IBM Think*, Apr. 22, 2024. [Online]. Available: https://www.ibm.com/think/topics/latent-dirichlet-allocation

[6] A. Mirzal, "Clustering and latent semantic indexing aspects of the singular value decomposition," arXiv:1011.4104, Nov. 2010. [Online]. Available: https://arxiv.org/abs/1011.4104

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805, Oct. 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[8] S. Kuzi, M. Zhang, C. Li, M. Bendersky, and M. Najork, "Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach," arXiv:2010.01195, Oct. 2020. [Online]. Available: https://arxiv.org/abs/2010.01195

[9] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Communications of the ACM*, vol. 30, no. 11, pp. 964–971, Nov. 1987. [Online]. Available: https://doi.org/10.1145/32206.32212

[10] H. K. Azad and A. Deepak, "Query expansion techniques for information retrieval: A survey," *Information Processing & Management*, vol. 56, no. 5, pp. 1698–1735, Sep. 2019. [Online]. Available: https://doi.org/10.1016/j.ipm.2019.04.007

[11] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2009. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/relevance-feedback-and-query-expansion-1.html

[12] N. Goharian, "Query expansion techniques (relevance feedback, thesaurus, semantic network)," Georgetown Univ., COSC 488 Lecture Slides. [Online]. Available: https://people.cs.georgetown.edu/ñazli/classes/ir-Slides/QueryExpansion-RF-13.pdf

[13] D. Sharma, R. Pamula, and D. S. Chauhan, "Semantic approaches for query expansion," *Evolutionary Intelligence*, vol. 14, pp. 1101–1116, Mar. 2021. [Online]. Available: https://doi.org/10.1007/s12065-020-00554-x

[14] Q. Liu, H. Chen, L. Zhang, J. Wang, and S. Li, "Query expansion based on word embeddings and ontologies for efficient information retrieval," *The Scientific World Journal*, 2017. [Online]. Available: https://thesai.org/Downloads/Volume12No11/Paper_42-Query_Expansion_based_on_Word_Embeddings_and_Ontologies.pdf

[15] Y. Slimani, A. Jarir, and M. O. Bensedik, "Query expansion based on NLP and word embeddings," in *Proc. TREC Common Core Track*, 2018. [Online]. Available: https://trec.nist.gov/pubs/trec27/papers/JARIR-CC.pdf

[16] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," arXiv:1907.04907, Jul. 2019. [Online]. Available: https://arxiv.org/abs/1907.04907

[17] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020. [Online]. Available: https://doi.org/10.1162/tacl_a_00325

[18] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv:2203.05794, Mar. 2022. [Online]. Available: https://arxiv.org/abs/2203.05794

[19] Xenoss, "Topic modeling techniques comparison," Xenoss Blog, 2025. [Online]. Available: https://xenoss.io/blog/topic-modeling-techniques-comparison

[20] D. Angelov, "Top2Vec: Distributed representations of topics," arXiv:2008.09470, Aug. 2020. [Online]. Available: https://arxiv.org/abs/2008.09470

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv:1301.3781, Jan. 2013. [Online]. Available: https://arxiv.org/abs/1301.3781

[22] H. Henderi, B. H. Hayadi, S. Sofiana, P. Padeli, and D. Setiyadi, "Unsupervised learning methods for topic extraction and modeling in large-scale text corpora using LSA and LDA," *Journal of Applied Data Sciences*, vol. 4, no. 3, pp. 103–118, Sep. 2023. [Online]. Available: https://doi.org/10.47738/jads.v4i3.102

[23] K. P. Moholkar, S. H. Patil, and J. Naveenkumar, "Question classification for efficient QA system," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 2, pp. 1876–1884, Apr. 2021. [Online]. Available: http://dx.doi.org/10.17762/turcomat.v12i2.1526

[24] PythonAfroz, "MedQuAD: Medical question-answer dataset for AI research," Kaggle Datasets. [Online]. Available: https://www.kaggle.com/datasets/pythonafroz/medquad-medical-question-answer-for-ai-research

[25] One-Off Coder, "Topic modeling with Gensim," 2025. [Online]. Available: https://datascience.oneoffcoder.com/topic-modeling-gensim.html