ETM-F: An Enriched Topic Modeling and Filtration Framework Integrating Ontologies and Deep Learning for Biomedical Trend Analysis

Ahmad Altarawneh[1], Mahir Arzoky[1], Stephen Swift[1], Reem Qadan Al Fayez[2], Moh'd Belal Al-Zoubi[2], Bilal Sowan[3*], Li Zhang[4]

[1]Department of Computer Science, Brunel University of London, London, UK.
[2]Department of Computer Information Systems, The University of Jordan, Amman, Jordan.
[3]Department of Business Intelligence and Data Analytics, University of Petra, Amman, Jordan.
[4]Department of Computer Science, Royal Holloway, University of London, Surrey TW20 0EX, UK.

*Corresponding author(s). E-mail(s): bilal.sowan@uop.edu.jo;
Contributing authors: AhmadTurkiFaiq.Altarawneh@brunel.ac.uk; Mahir.Arzoky@brunel.ac.uk; Stephen.Swift@brunel.ac.uk; r.alfayez@ju.edu.jo; mba@ju.edu.jo; li.zhang@rhul.ac.uk;

Abstract

Selecting research topics from author keywords helps identify emerging trends in medicine, but keyword-only models remain fragile because of specialized terminology and rapidly evolving vocabularies. To address this limitation, we propose Enriched Topic Modeling with Filtration (ETM-F), a methodological framework that integrates frequency-based keyword filtration with Medical Subject Headings (MeSH) enrichment. ETM-F is model-agnostic and can be applied across probabilistic and neural topic models. We evaluate ETM-F against baseline Latent Dirichlet Allocation (LDA) and extend the analysis to hierarchical LDA (hLDA), Dynamic Topic Models (DTM), Contextualized Topic Models (CTM), and BERTopic. Model performance is assessed through perplexity, $\boldsymbol{C_v}$ coherence, normalized pointwise mutual information (NPMI), topic diversity, dominant topic contribution, and temporal coherence. Statistical significance is

established through bootstrap confidence intervals and Wilcoxon signed-rank tests. Experiments on 232,191 medical publications from Scopus (2020) confirm the advantages of enrichment. For LDA, $\boldsymbol{C_v}$ coherence rises from 0.37 with author keywords to 0.52 with MeSH-enriched keywords, while NPMI improves from –0.05 to 0.08. Neural topic models provide higher quality results. After filtration, BERTopic achieves the best coherence ($\boldsymbol{C_v = 0.75}$, NPMI = 0.28), followed by hierarchical LDA ($\boldsymbol{C_v = 0.66}$) and CTM ($\boldsymbol{C_v = 0.60}$). Dynamic Topic Models and time-sliced BERTopic capture temporal signals, including the shift from generic vaccine themes to mRNA-specific terms and the surge of telemedicine. Embedding-based keyword expansion with BioWordVec and BioBERT further enhances coherence by 0.03 to 0.05 and identifies synonyms such as "SARS-CoV-2". These findings confirm that semantic enrichment and dynamic models improve topic discovery in large-scale biomedical text. ETM-F establishes a reproducible benchmark for future research and offers a reliable tool for identifying thematic directions in medical literature.

Keywords: Topic modeling, Medical Subject Headings (MeSH), Enriched Topic Modeling with Filtration (ETM-F), Latent Dirichlet Allocation (LDA), BERTopic, Coherence

1 Introduction

Scientific research supports the economic and technological progress of nations. It also forms the foundation of societal development, as academic institutions, governments, and industries depend on evidence-based knowledge to drive innovation and growth. However, the rapid growth of biomedical publications makes it challenging to identify emerging themes. Each year, millions of research articles appear worldwide and are freely available. Global publication output grew by about 47% between 2016 and 2022, a rate that exceeds the growth of the research community [1]. This rapid growth causes information overload and makes it difficult for researchers to follow emerging themes. Reviews show that large volumes of information reduce productivity and strain decision processes [2].

Researchers increasingly use trend analysis to understand scholarly communication and identify emerging research frontiers. Bibliometric analysis offers a systematic way to quantify publication patterns, networks, and thematic evolution but often overlooks semantic relationships in article content [3–5]. Analyzing research trends often depends on author-provided keywords, which may be incomplete or inconsistent. Natural Language Processing (NLP) methods extract additional semantic cues from article titles and content [6]. Topic modelling is a probabilistic approach that represents each topic as a distribution over words and each document as a distribution over topics [7].

A topic model can be created with the knowledge of words and their counts in documents [8]. Once the topic model is available, all documents can be represented as distributions of topics. This allows documents to be examined and comprehended in ways not previously exploited. It is worth mentioning that a topic is essentially a probability distribution of words; it does not create its label or title [9]. For example, in the document corpus, the sports topic will be a series of words (such as swimming, football, basketball, boxing, running, and volleyball), each with its probability of occurrence in the topic. Earlier mentioned words are expected to reappear in the topic as previously stated. Machine Learning (ML) techniques

are useful for extracting topics and organizing terms based on their probability, which helps to create a clearer understanding of the ideas being reflected.

Unsupervised topic modeling algorithms like Latent Dirichlet Allocation (LDA), Latent semantic analysis (LSA), Hierarchical Dirichlet Process (HDP), and Dynamic Topic Modelling (DTM) can be integrated into natural language processing using statistical machine learning [10, 11]. These algorithms assume probabilistic distributions in documents and use mathematical and statistical techniques to identify topics [9].

Topic models provide a scalable approach to summarize large collections by extracting latent themes and representing each document as a topic distribution. Classical approaches such as LDA assume that documents are exchangeable and that vocabulary is fixed [12]. These assumptions present limitations in medicine, where terminology evolves quickly and non-specialist authors may supply incomplete or inconsistent keywords. Recent studies show that static models fail to capture longitudinal dynamics, whereas transformer-based embeddings improve coherence and interpretability [13–15].

This research addresses the problem of content-based trend analysis by enhancing the keywords that describe publications. We apply topic detection to organise medical publications and reveal dominant themes. Ontologies provide a mechanism to identify semantically equivalent or related concepts. An ontology is a machine-readable framework that defines the essential concepts and categories for structuring knowledge in a given domain [16]. Among the most mature domains of ontologies are those in medicine.

This study focuses on the use of medical ontologies to improve trend analysis in biomedical research. Medical Subject Headings (MeSH), a controlled vocabulary widely adopted for indexing and cataloguing medical literature, supports retrieval and semantic enrichment of research content [17]. Author keywords, typically derived from "author-supplied keywords," often fail to represent the complete thematic scope of a publication. The integration of MeSH descriptors provides a more precise and semantically consistent representation by systematically expanding and refining the original keyword sets.
The contributions of this study are as follows:

1. We propose the Enriched Topic Modeling with Filtration (ETM-F) framework.. It integrates MeSH annotation, embedding-based keyword expansion, frequency-based filtration, and a disambiguation protocol that selects the most specific MeSH concept when multiple matches occur.
2. We enable longitudinal trend analysis with Dynamic Topic Models (DTM) and time-sliced Contextualized Topic Models (CTM)/BERTopic to reveal temporal patterns in medical research.
3. We establish baselines with LDA, hierarchical LDA (hLDA), CTM, and BERTopic. We validate significance through bootstrap resampling and Wilcoxon signed-rank tests.
4. We evaluate BioWordVec and BioBERT for embedding-based keyword expansion combined with MeSH enrichment and report improvements in coherence.
5. We provide a complexity analysis with runtime estimates for preprocessing, annotation, and model training. We also discuss annotation ambiguity, dataset bias, and evaluation limitations.
6. We expand the literature review to include recent advances in neural and transformer-based topic modeling.

1.1 Research Questions

To guide the study, we refine our research questions as follows:

1. RQ1 (MeSH enrichment): Does enriching author keywords with MeSH concepts improve topic quality (perplexity, C_v coherence, Normalized Pointwise Mutual Information (NPMI) coherence, and diversity) compared to using author keywords alone?
2. RQ2 (Data filtration): How do document frequency thresholds (`min_df`, `max_df`) affect vocabulary size and topic quality, and are these effects robust across threshold ranges and publication years [12](see Section 3.2 for sensitivity analysis)?
3. RQ3 (Baseline comparison): How do alternative topic models including hLDA, DTM, CTM, and BERTopic perform relative to static LDA under MeSH enrichment [14, 18]?
4. RQ4 (Temporal analysis): Can dynamic models (DTM, time sliced CTM/BERTopic) reveal meaningful longitudinal trends in the 2020 corpus, such as pandemic related topic evolution?
5. RQ5 (Embedding expansion): Do embedding based keyword expansions using BioWordVec and BioBERT improve topic quality beyond ontology based enrichment, and what are the trade-offs in relevance versus noise [19]?

To evaluate topic quality and its effect on research trend analysis in the medical domain, this study addresses five focused research questions using the Scopus medical dataset. The contributions are organized as follows:

1. RQ1 (MeSH enrichment): Assess whether MeSH enrichment improves topic quality. We measure perplexity, C_v coherence, NPMI coherence, and diversity and compare against results from author keywords alone.
2. RQ2 (Data filtration): Apply the proposed ETM-F framework. We tokenize the corpus, apply document-frequency filters, and assign topics using multinomial distributions. We report the effect of `min_df` and `max_df` thresholds on vocabulary size and topic quality [20].
3. RQ3 (Baseline comparison): Compare the proposed ETM-F framework with hLDA, DTM, CTM, and BERTopic. We report differences in coherence, perplexity, and topic diversity.
4. RQ4 (Temporal analysis): Identify the dominant topic for each document and use DTM and time sliced CTM/BERTopic to reveal longitudinal trends, including pandemic related topic shifts.
5. RQ5 (Embedding expansion): Test BioWordVec and BioBERT keyword expansion. We evaluate gains in topic quality and analyze the trade-off between relevance and noise beyond ontology based enrichment.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 introduces the proposed methodology. Section 4 reports the experimental results and evaluation. Section 5 discusses the findings and limitations. Finally, Section 6 concludes the paper with final remarks.

2 Related Work

This section provides an overview of related studies on ontology-based enrichment and topic modeling in biomedical research. We first outline the role of MeSH as a standard ontology for improving keyword quality and retrieval. We then review classical and neural topic modeling approaches that have been applied to uncover hidden themes in medical literature.

In today's era of abundant medical data, it has become crucial to identify the trends and features of the medical domain. Trend analysis has become increasingly popular in recent years due to its ability to identify research areas and reveal hidden trends within extensive resources. Machine learning approaches have been employed to analyze trends, such as the study by [21] on innovation and entrepreneurial ecosystems. The researchers employed co-citation analysis and network meta-analysis to better understand patterns and characteristics of various meta knowledge in the interdisciplinary field.

2.1 Medical Subject Headings (MeSH)

Ontologies formalize a system's structure, entities, and observed relations. It is a knowledge environment. It involves action, formal naming, and establishing categories, attributes, and relationships between concepts, data, and objects that support one, several, or all domains [22, 23]. MeSH terms are labels assigned, official words, or phrases selected to represent particular biomedical concepts in each article in the Medline Database to describe and explain the article's topic [24]. So, the MeSH terms used for indexing were introduced in 1960 with about 4,400 subject headings [25], with the number of topics increased to 25,000 in the 2010 version [26].

Soriano et al. [27] systematically reviewed scientific information from 2006 to 2016 on enterprise strategic intelligence and public relations. They used bibliometric analysis to assess impact and maturity, identifying a potential emerging research field between these two themes, highlighting common topics and future research directions in technological observatories. Dridi et al. [28] introduced (*Leap 2 Trend*) an approach for detecting research trends using temporal word embeddings (*word 2 vec*). The results have determined the approach's effectiveness in detecting trends in some cases.

The dataset used in this research, in particular keywords obtained from the Elsevier website, was enriched by entering them into the Annotator Web service in the National Center for Biomedical Ontology (NCBO) BioPortal web interface by annotating and searching for terms for each word in ontologies, annotating text with ontology terms, searching an ontology-based index of biomedical resources [29]. MeSHs were used in this paper, and it is one of the web maps published in BioPortal, which will be mentioned in the following subsection.

2.2 Topic Modelling Approaches

The LDA model, developed by Blei et al.[30], is a probabilistic model designed to identify frequently occurring words in different documents. It operates under the assumption that each document is composed of a mixture of various words. To enhance accuracy, LDA breaks down a bag-of-words matrix into two new matrices: a document-to-topic matrix and a word-to-topic matrix. This method facilitates the exploration of topics within a dataset and the discovering of underlying themes across textual documents. A distribution of words characterizes each topic, and each document is a blend of multiple topics [31]. LDA is an effective tool for topic modeling and finds extensive application in text analysis and other domains. The input corpus typically comprises D documents, each representing a collection of tokens representing words. The pseudo-code for the LDA algorithm [32] is shown in Algorithm 1.

Table 1 Description of Symbols

Symbol	Description
M	The overall count of documents within the corpus
W	The count of distinct words within a document
N	The overall count of words within the document
K	Number of topics
x_{di}	i-th observed word within document d
z_{di}	Latent topic assigned to word x_{di}
N_{wk}	Count of words assigned to topic k
N_{dk}	Count of topic k assigned in document
ϕ_k	Distribution of word in given topic k
θ_d	Topic distribution in given document d
α	Distribution of topics per document
β	Distribution of words per document
β_k	Distribution of words for topic k

As described in Algorithm 1, the LDA algorithm determines the procedure for generating processes for each document d. Documents are analyzed with the primary objective of generatively discovering and deducing the topics within a corpus. The documents are randomly represented across latent topics, with each topic being distributed over the words [30]. The notations used in this process are presented in Table 1. It demonstrates the distribution of a topic combination θ, a collection of N topics z, and N words w, given the parameters α and β.

Algorithm 1 Pseudo code for LDA algorithm.

Input: A document corpus (a set of documents), hyper-parameters α, and β.
Output: List of topics.
1: Create a distribution representing various topics, $\theta_d \sim \text{Dirichlet}(\alpha)$.
2: **for each** word i in the document, **do**
3: Create a topic index z_{di} by extracting the distribution of topic weights, $z_{di} \sim \theta_d$.
4: Extract the observed word w_{di} from the chosen topic, $w_{di} \sim \beta_{z_{di}}$.
5: **end for**
6: Repeat the process for all documents in the collection.

In [33], a trend analysis method using LDA was proposed. Specifically, the analysis of trends is conducted by topic, utilizing the combined results of LDA and the most significant keywords, which are the top 10. Their work involved conducting topic modeling and analyzing the trend of international standards derived from the identified topics.

Ontologies are crucial for locating knowledge and skills in both humans and robots. As for ontologies, they are a vital component for information retrieval by humans and machines. For example, the study of [34] aimed to identify knowledge areas and skill sets for data software engineering. The results presented a competency map, highlighting key areas of knowledge and skills. This map can assess and improve IT professionals' skills, aid in hiring the right people, and ensure software engineering education programs meet industry needs.

Yu et al. [35] investigated the effectiveness of MeSH keywords in discovering and organizing biomedical documents [36], contrasting it with vector representations that incorporate latent topic proportions reflecting the similarity of those terms.

Topic modelling methods continue to advance rapidly and their adoption in biomedical research grows steadily. Recent studies have applied BERTopic to a wide range of biomedical corpora and demonstrated its ability to uncover domain-specific themes and temporal trends. Ma et al. [14] compared LDA and BERTopic in opioid-related cardiovascular research and reported that transformer-based embeddings produced more compact and coherent clusters. These results were further improved when combined with ChatGPT for automated topic

interpretation. Research in the gaming industry developed an LDA-based framework for esports review analysis. The approach combined topic models with sentiment analysis to identify player concerns. This study demonstrated LDA's ability to process user-generated content with complex emotional dimensions [37].

Large-scale studies of PubMed abstracts confirm this capability. Madrid et al. [38] mapped rheumatology literature, while Lezhnina et al. [39] examined mental health research, and Kumar et al. [40] analysed gut–brain-axis publications. All reported that BERTopic captured long-term topic trajectories and detected emerging "hot" topics such as COVID-19 and JAK inhibitors.

Neural topic models also play a role in digital epidemiology. Rao et al. [13] separated slang drug mentions from brand-name discussions in social-media posts and revealed distinct sentiment profiles between groups. Zhang et al. [41] introduced *Turtling*, a time-aware model using BioWordVec embeddings, and showed that it outperformed classical approaches on NIH grant abstracts while producing smooth topic transitions over time.

Other researchers combined BERTopic-derived clinical note topics with structured electronic health records to enhance mortality prediction [42]. In digital health, Uncovska et al. [43] applied BERTopic to mHealth app reviews, generating actionable insights for improving user experience and regulatory oversight. Chung et al. [44] used BERTopic and dynamic BERTopic on open-ended text messages from older adults, detected stress-related themes and followed their evolution over time. Xin et al. [45] proposed a semantic-type–filtered model for noisy health-social-media data and highlighted novel conditions surfaced by ontology integration.

Recent surveys consolidate this methodological landscape. Kumari et al. [15] presented a systematic review of topic models and included approaches from LSA and LDA to neural models such as Top2Vec and BERTopic. They emphasised evaluation metrics, interpretability, and application context as persistent challenges. Guillén-Pachó et al. [46] assessed DTM and BERTopic on sequential COVID-19 corpora and found that DTM captured concept evolution most effectively, particularly when paired with automated topic labelling.

Further studies show both the potential and the limitations of LDA-based models across diverse domains. Kawai et al. [47] developed hybrid recommender models that combined content-based filtering with LDA to address cold-start problems in recommendation systems. Their work illustrates how domain-specific adaptations can improve the interpretability of LDA. Ozyurt et al. [48] performed a large-scale analysis of 8,584 Industry 4.0 articles and identified 12 distinct research themes. These applications confirm LDA's scalability for large corpora and its continued relevance in contemporary practice. However, evaluation studies reveal significant constraints in contemporary text analysis. Egger and Yu [49] compared LDA with modern approaches such as BERTopic on social media data and found substantial limitations with short texts and co-occurrence patterns. Several studies report that traditional LDA demands extensive preprocessing and frequently fails to capture nuanced semantic relationships in specialized vocabularies [50]. Gupta et al. (2025) reported coherence improvements from 0.5289 to 0.5971 through TF-IDF enhancement. However, baseline LDA performance remains suboptimal for complex domains [50].

These findings reveal critical research gaps. Current LDA applications lack systematic ontological enrichment for specialized fields. Comparative studies focus on general corpora rather than domain-specific vocabularies with controlled terminologies. Most research evaluates individual models rather than comprehensive frameworks that integrate multiple methodological approaches. The biomedical domain presents unique opportunities to address these limitations through established ontologies such as MeSH. Controlled vocabularies enable

systematic enhancement of topic models beyond traditional preprocessing techniques. Furthermore, biomedical literature exhibits temporal patterns that require specialized analytical approaches to capture research evolution effectively.

Motivated by these studies, this research introduces a new framework for topic modelling. The proposed framework uses unsupervised learning to identify document trends based on the keywords in each document, integrates MeSH enrichment, and applies embedding-based keyword expansion. It benchmarks LDA, hLDA, CTM, and BERTopic, and includes dynamic and sensitivity analyses to improve interpretability and track research trends with statistical rigour. In addition, DTM and time-sliced BERTopic capture the evolution of topic distributions across time intervals. These additions allow the framework to reveal longitudinal patterns in the data.

3 Materials and Methods

This section presents the methodological framework adopted in this study. We describe the dataset, preprocessing steps, enrichment strategies, and evaluation procedures. The proposed framework, Enriched Topic Modeling with Filtration (ETM-F), is also introduced. ETM-F integrates MeSH-based keyword enrichment, embedding expansion, and frequency-based filtration to enhance topic quality. It further combines multiple topic modeling approaches (LDA, hLDA, DTM, CTM, BERTopic) with trend discovery, statistical evaluation, and visualization. The workflow is shown in Figure 1.

We extracted articles from the Elsevier Scopus database for the period 2020 to the present. This time frame corresponds to the onset of the COVID-19 pandemic and reflects the rapid growth of biomedical research in vaccines, telemedicine, and public health. A filter was applied through the Scopus API to include only peer-reviewed journal publications in the medical field.

Two analysis settings were considered. The first used author keywords. The second used keywords enriched with MeSH ontology terms. For each document, the dominant topic was defined as the one with the highest percentage contribution. The proposed framework ETM-F applies tokenization and frequency-based filtration to the document corpus. It then generates topic clusters that improve organization and retrieval. Algorithm 2 outlines the procedure. After preprocessing and parameter initialization, the algorithm iterates through documents, samples document–topic distributions, and assigns topics to terms through multinomial distributions. This design improves coherence and interpretability by incorporating data filtration into topic modeling.

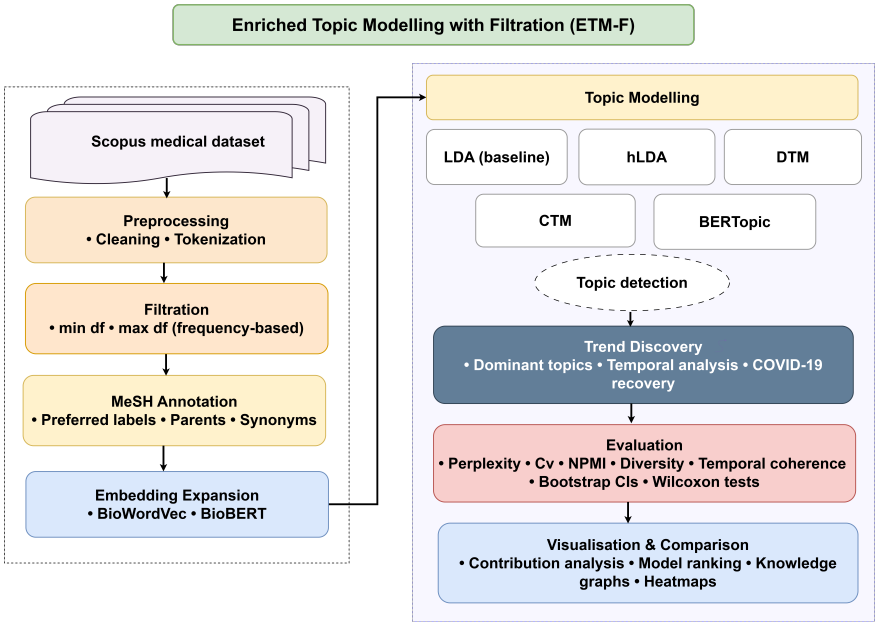

Fig. 1 The proposed ETM-F framework with MeSH-enriched keywords.

Algorithm 2 Enriched Topic Modeling with Filtration (ETM-F)

Input: Document corpus $D = \{d_1, d_2, \ldots, d_N\}$, thresholds min_df, max_df, maximum frequency T, enrichment resources (MeSH, embeddings), hyperparameters α, β

Output: Topic sets for four conditions: KW-Before, KW-After, ENH-Before, ENH-After

Step 1: Preprocessing For each document $d_i \in D$, tokenize and remove punctuation, stop words, and non-alphabetic tokens. Let V denote the initial vocabulary.

Step 2: Filtration Construct two vocabularies: (a) Before filtration: $V_{bf} = V$ (b) After filtration:

$$V_{af} = \{w \in V : min\_df \leq df(w) \leq max\_df,\ f(w) \leq T\}$$

where $df(w)$ is the document frequency of word w, and $f(w)$ its corpus frequency.

Step 3: Enrichment For each vocabulary $V_x \in \{V_{bf}, V_{af}\}$: Map terms to MeSH concepts $M(w)$ using ontology lookup. Expand with embeddings:

$$E(w) = \{w_1, \ldots, w_k : \cos(\vec{w}, \vec{w_i}) \geq \tau\}$$

using BioWordVec or BioBERT. Define enriched vocabulary as

$$V_x^{enh} = V_x \cup M(V_x) \cup E(V_x).$$

Step 4: Parameter Initialization Initialize Dirichlet priors α for document–topic distributions θ_d and β for topic–word distributions ϕ_z.

$$\theta_d \sim \text{Dirichlet}(\alpha), \quad \phi_z \sim \text{Dirichlet}(\beta)$$

Step 5: Inference For each condition (KW-Before, KW-After, ENH-Before, ENH-After):

for all document $d \in D$ **do**
  Sample topic proportions θ_d.
  for all term $t \in d$ **do**
    Draw topic $z_{dt} \sim \text{Multinomial}(\theta_d)$
    Draw word $t \sim \text{Multinomial}(\phi_{z_{dt}})$
  end for
end for

Step 6: Output Return inferred topics with coherence (C_v), NPMI, diversity, and perplexity scores for all four conditions.

The study presents an ontology-based methodology for trend analysis through topic discovery. The process consists of several steps: bag-of-words representation, model training, generation of model outputs, visualization, evaluation, and comparison of results. The following sections describe each step of the proposed medical topic modeling framework in detail and explain their implementation.

3.1 Dataset Collection

The dataset was collected from Scopus through Elsevier's API[1]. We restricted the search to peer-reviewed journals in the *Medicine* subject area and limited the time span to the year 2020. This period was selected to capture the emergence of COVID-19 and ensure a corpus that is both diverse and manageable for trend analysis. The final dataset contained 232,397 biomedical articles. Each record included metadata such as title, author keywords, publication year, and journal. Because the dataset consisted only of bibliographic metadata, ethical approval was not required.

3.2 Preprocessing

Preprocessing was the first step after collecting the dataset. This stage is critical in text analysis because it converts raw text into a structured, machine-readable format while preserving semantic content [51]. The procedure was carried out in three phases: text cleaning, re-configuration, and frequency-based filtering.

3.2.1 Text Cleaning

Text cleaning reduced noise and simplified the dataset. The following steps were applied:

1. Removed incomplete entries, missing features, and rows with NaN values in the "enhanced words" field. These corresponded to terms that failed to match MeSH ontology concepts.
2. Normalized text by converting all letters to lowercase and removing digits, symbols, non-English characters, and irrelevant signs [52, 53]. Punctuation that did not contribute to meaning was also removed.
3. Removed common stop words such as "a," "an," "the," and "are."
4. Applied tokenization. Keywords were separated into tokens, with multi-word terms preserved as single units (e.g., "health services research"). This ensured that compound biomedical expressions retained their meaning.

3.2.2 Re-configuration

After cleaning, the data was converted into a representation suitable for automated processing. We adopted the bag-of-words (BoW) model. Each document was represented as a vector of tokens, where the entries indicated the frequency of words in that document [54].

A dictionary was created, assigning each unique word a distinct integer identifier. The corpus was then generated as a collection of pairs: the first element denoted the word ID, and the second its frequency. For example, [(1, 4)] indicates that word "1" corresponds to "Coronavirus" and appears four times in the document. This structured corpus served as the basis for model training using the proposed ETM-F framework.

[1] https://dev.elsevier.com

The length of the dictionary varied across corpora. The parent keyword set contained 221,396 words, with the synonym set having 221,382 words, and the combined keyword and enriched set showing 227,151 words.

3.2.3 Frequency-based Filtering

We evaluated document frequency thresholds min_df $\in \{5, 10, 20\}$ and max_df $\in \{0.4, 0.5, 0.6\}$. For the main experiments, we set min_df $= 10$ and max_df $= 0.5$. This configuration removed extremely rare and overly common terms while retaining domain-specific vocabulary. Sensitivity analysis confirmed that coherence and diversity curves plateaued beyond these thresholds [55].

3.3 Keyword Enrichment and Expansion

- Embedding-based keyword expansion: To complement ontology-based enrichment, we applied embedding-based expansion. Pre-trained BioWordVec embeddings were used to retrieve the top-5 nearest neighbours (cosine similarity) for each keyword, followed by filtering through UMLS semantic types and deduplication. In parallel, BioBERT sentence embeddings were computed for each keyword, and the top-k neighbours were identified. We considered three settings for comparison: (A) author keywords only, (B) MeSH enrichment only, and (C) MeSH combined with embeddings. The parameter $k = 5$ was chosen based on preliminary coherence curves, which balanced topic diversity against noise.
- Ontology-based disambiguation protocol: To address the risk of missing topics and annotation ambiguity, all keywords were mapped to MeSH concepts using the NCBO BioPortal Annotator. Ambiguities were resolved according to the following protocol:

  1. Longest match preference: the longest matching span is selected when multiple matches occur;
  2. Preferred label selection: MeSH preferred labels are selected over synonyms;
  3. Semantic type restriction: mappings are restricted to relevant semantic types such as diseases, drugs, organs, procedures, and public health terms;
  4. Confidence threshold: retain only mappings with a confidence score ≥ 0.8;
  5. Specificity prioritization: when ties remain, the concept with the greatest MeSH tree depth is selected, and more specific concepts are prioritized.

  This protocol yielded an average of 2.3 MeSH terms per keyword, reducing false matches and aligning with recommended practices for domain-specific concept extraction.

3.4 Topic Models and Implementation

We implemented five topic models using Python libraries: (i) LDA with Gensim's variational Bayes estimator and $K = 8$ topics; (ii) hLDA based on the nested Chinese Restaurant Process; (iii) DTM connecting topics across quarterly slices (Q1–Q4 2020) via a state-space model; (iv) CTM combining a variational autoencoder with BioBERT embeddings; and (v) BERTopic, which applies BioBERT embeddings, HDBSCAN clustering, and class-based

TF–IDF reweighting [56, 57]. Each model was trained on both author keywords and MeSH-enriched keywords, with and without embedding expansion. The number of topics ($K = 8$) was chosen by maximizing C_v coherence on a hold-out validation set.

3.5 Evaluation Metrics and Statistical Testing

This study used a set of quantitative metrics to evaluate the performance of topic models. The focus was on three dimensions: perplexity, coherence, and dominant topic contribution. Additional statistical tests were applied to verify the robustness of results.

3.5.1 Perplexity

Perplexity measures how well a model predicts unseen data [30]. A lower score reflects stronger generalization performance. The dataset was split into training and test sets, with approximately 80% used for training and 20% reserved for testing. The score was calculated as:

$$\text{Perplexity} = \left(\prod_{i=1}^{N} P(w_i) \right)^{-\frac{1}{N}} \tag{1}$$

where $P(w_i)$ is the probability of word i and N is the number of words in the test corpus [58, 59].

3.5.2 Topic Coherence

Coherence evaluates the semantic consistency of the top-ranked words in each topic [60]. We report C_v coherence and NPMI coherence as implemented in Gensim [61]. Higher values indicate closer agreement with human judgment.

Two established formulations were also considered: the extrinsic UCI metric [62] and the intrinsic UMass metric [63]. A generic form is

$$\text{Coh}(V) = \sum_{(w_i, w_j) \in V} \text{score}(w_i, w_j; \varepsilon), \tag{2}$$

where V is the set of top words in a topic, w_i and w_j denote individual words, and ε is a smoothing constant.

UCI (PMI-based) coherence:

$$C_{\text{UCI}}(w_i, w_j; \varepsilon) = \log \left(\frac{P(w_i, w_j) + \varepsilon}{P(w_i)\, P(w_j)} \right), \tag{3}$$

where $P(w_i)$ is the marginal probability of word w_i and $P(w_i, w_j)$ is the joint probability of words w_i and w_j appearing together.

UMass (count-based) coherence:

$$C_{\text{UMass}}(w_i, w_j; \varepsilon) = \log \left(\frac{D(w_i, w_j) + \varepsilon}{D(w_j)} \right), \tag{4}$$

where $D(w_i, w_j)$ is the document co-occurrence count of w_i and w_j, and $D(w_j)$ is the count of documents containing w_j.

3.5.3 Diversity and Dominant Topics

Diversity was defined as the proportion of unique terms across the top 25 words in all topics. It captures lexical variety and reduces redundancy. In parallel, the dominant topic contribution was measured by identifying the topic with the highest proportion in each document and averaging its contribution across the corpus. Human annotators also inspected the top ten words in each topic and rated their semantic coherence on a five-point scale.

3.5.4 Temporal Coherence

Temporal coherence was applied to DTM, CTM, and BERTopic. It measures the similarity of topic-word distributions between consecutive time slices. This ensures that trends evolve smoothly and that identified shifts reflect real dynamics in the literature.

3.5.5 Statistical Testing

To test significance, we generated 100 bootstrap resamples per model and metric. Paired Wilcoxon signed-rank tests ($\alpha = 0.05$) were used to assess whether observed differences were significant. This combination of bootstrap confidence intervals and non-parametric testing provides a robust evaluation of model stability and relative performance.

3.6 Complexity Analysis

We analyze the computational complexity of each component in the proposed workflow to clarify scalability and runtime costs. Preprocessing and frequency-based filtering scale linearly with corpus size: $O(D \cdot L)$, where D is the number of documents and L is the average document length. Vocabulary construction has complexity $O(|V|)$, where $|V|$ is the vocabulary size after filtering.

MeSH annotation via NCBO BioPortal scales as $O(D \cdot W \cdot R)$, where W is the average number of keywords per document and R is the average API response time. This step dominated preprocessing runtime (approximately 2 hours for our corpus of 232,397 documents). Embedding-based expansion adds $O(W \cdot E)$, where E is the cost of embedding computation.

The complexity of topic model inference varies by method:

- LDA: $O(D \cdot L \cdot K)$ per Gibbs iteration, where $K = 8$ topics.
- hLDA: $O(D \cdot L \cdot K \cdot \log K)$ due to hierarchical sampling.
- DTM: $O(T \cdot D \cdot L \cdot K)$, where $T = 4$ quarterly time slices.
- CTM: dominated by BioBERT embedding computation $O(D \cdot L \cdot H)$, where H is the hidden dimension.
- BERTopic: $O(D \cdot L \cdot H + D^2 \cdot \log D)$, combining embedding computation and HDBSCAN clustering.

Evaluation metrics require $O(D \cdot K + |V| \cdot K^2)$ for coherence computation and $O(B \cdot D \cdot K)$ for B bootstrap samples. GPU acceleration reduces training time for neural models by 5–10×, but requires substantial memory ($\geq$ 8 GB VRAM for BioBERT).

3.7 Proposed Enriched Topic Modeling with Filtration (ETM-F) Model

- **Trend analysis using author keywords**

We introduce a trend analysis approach based on the proposed ETM-F model when only original keywords are used. ETM-F is a generative model that represents each document as a mixture of latent topics, with each topic defined as a probability distribution over the vocabulary. The model requires a fixed number of topics, but identifying the optimal number is often difficult. In this study, the number of topics was determined using the method of Prabhakaran [61], which builds multiple models with different topic counts and selects the one that yields the highest coherence score.

Figure 2 presents the coherence curve for topic numbers between 2 and 40. The coherence score was calculated after the model was built with an initial random choice of topic count. A coherence model was then applied to evaluate the outcome and to identify the most suitable number of topics. The topics with the highest coherence values were considered the most relevant.

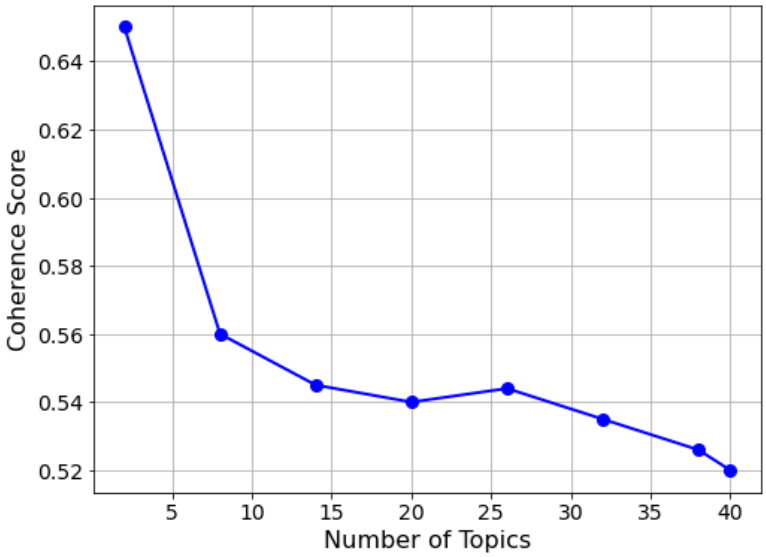

Fig. 2 Coherence score visualization.

Considering the downward trend in the coherence curve, it is more reliable to select the model with the highest coherence score before the curve reaches a plateau. As shown in Figure 2, the peak coherence value occurs at eight topics. This indicates that eight is the optimal number of topics for the 2020 medical dataset. Accordingly, all subsequent experiments in this study were conducted with this configuration.

- **Trend analysis using MeSH-enriched keywords**

We propose a trend analysis approach based on the ETM-F model when keywords are enriched with MeSH ontology terms. The aim is to examine the added value of MeSH in identifying research trends and in linking descriptors to topics. In this approach, topic modeling relies on enriched keywords, which extend the original keyword set with MeSH descriptors. Each document is therefore represented by both author keywords and indexed MeSH terms, and the resulting topics combine these two sources of information. To improve interpretability, MeSH descriptors were added as labels within the bag-of-words representation. Because

Table 2 Example of medical terms enrichment with MeSH ontology descriptors

#	Feature name	Terms
1	Keywords	COVID-19, Kawasaki, MIS-C, PIMS
2	Parents	Respiratory Tract Infections, Coronavirus Infections, Pneumonia, Viral, NA, NA, NA
3	Synonyms	Infection, Respiratory Tract, Infections, Coronavirus, Viral Pneumonias, NA, NA, NA
4	Enhanced_Words	COVID-19, Respiratory Tract Infections, Coronavirus Infections, Pneumonia, Viral

MeSH terms are designed to cover broad and significant domains in medicine, their inclusion ensures that the extracted topics reflect established biomedical knowledge.

The ETM-F model was applied to keywords enriched with MeSH ontology terms. We examined whether the inclusion of MeSH descriptors in the keyword set improved the quality of topics. Coherence was used as a proxy for topic quality, as suggested in earlier work [64]. The objective was to assess whether the enriched keyword approach (i.e., preflabel and parent terms) produced higher-quality topics than the baseline LDA model trained on original keywords. Keyword enrichment was performed with the BioPortal Annotator tool[2]. This tool was used to annotate the keywords of each article in the dataset with MeSH ontology terms. The enriched dataset contained additional features obtained through the BioPortal web service (API). Enrichment was implemented in Python, which sent requests to the API and collected the returned metadata. The resulting enriched dataset, containing keywords mapped to MeSH descriptors, was stored and used in the subsequent topic detection process.

This study applies MeSH to identify research trends in biomedical publications. MeSH terms are annotated from keywords and represent each document with one or more controlled descriptors. These terms are valuable for detecting trends because they provide hierarchical relations such as preferred labels, parents, grandparents, and synonyms. Table 2 illustrates examples of medical terms enriched with MeSH descriptors and the new features obtained. Each feature is discussed in detail in the following sections.

All keywords in the dataset were enriched with MeSH descriptors assigned to each article. The enrichment was performed using the MeSH ontology, where each keyword was mapped to its preferred label. Parent terms were then retrieved and included under the "parents" feature, with "NA" assigned if no match was found. Synonyms were also retrieved and stored under the "synonyms" feature. This process expanded the representation of documents while preserving semantic consistency.

4 Experimental Results and Analysis

This section comprehensively analyzes several experiments conducted on the proposed ETM-F model. The focus is on the evaluation process and the results obtained. Three experimental configurations were considered: analysis with original keywords, analysis with MeSH-enriched keywords, and analysis with enriched keywords combined with embeddings. The outcomes are reported in terms of coherence, diversity, and dominant topic contribution. These results are further supported with statistical testing and comparative analysis against baseline models.

[2] https://bioportal.bioontology.org/annotator

In addition, the evaluation extends to recent and state-of-the-art topic modeling techniques under rigorously controlled conditions. We examine five representative models that cover both classical and modern paradigms: LDA, hLDA, DTM, CTM, and BERTopic. These models represent the evolution from probabilistic approaches to neural and contextualized architectures. Each model was evaluated under four settings: (i) author keywords before filtration, (ii) author keywords after filtration, (iii) MeSH-enriched keywords before filtration, and (iv) MeSH-enriched keywords after filtration. This comprehensive design enables a balanced and reproducible comparison that clarifies how methodological advances affect both topic quality and interpretability.

Table 3 compares LDA trained on author keywords with ETM-F trained on MeSH-enriched keywords, using identical filtration thresholds. For LDA, the C_v coherence was 0.367 and NPMI −0.051. MeSH enrichment substantially improved these metrics, increasing coherence to 0.517 and NPMI to 0.077, while perplexity improved from −6.74 to −5.98 (lower is better). Across all models, MeSH enrichment consistently enhanced topic quality. Specifically, hLDA's C_v increased from 0.451 to 0.657 and NPMI from −0.096 to 0.255. Similarly, CTM improved from 0.422 to 0.601 and NPMI from −0.030 to 0.205. BERTopic demonstrated comparable enhancement, with C_v improving from 0.582 to 0.754 and NPMI from 0.155 to 0.278. The observed improvements in C_v ranged from 0.15 to 0.21 and in NPMI from 0.12 to 0.35. Statistical validation through Wilcoxon signed-rank tests confirmed significance for LDA and hLDA ($p < 0.001$). These results demonstrate that MeSH knowledge incorporation reduces vocabulary scarcity and produces more coherent topic representations.

Table 3 Comparison of topic quality metrics for author keywords and MeSH-enriched keywords.

Model	C_v **(keywords)**	C_v **(MeSH)**	ΔC_v	**NPMI (keywords)**	**NPMI (MeSH)**	**ΔNPMI**	**Diversity (MeSH)**	**Perplexity (MeSH)**
LDA	0.367	0.517	+0.150	-0.051	0.077	+0.129	0.355	-5.98
hLDA	0.451	0.657	+0.206	-0.096	0.255	+0.351	0.389	∞
CTM	0.422	0.601	+0.180	-0.030	0.205	+0.235	0.384	-5.96
BERTopic	0.582	0.754	+0.173	0.155	0.278	+0.123	0.375	N/A

Values are averaged over 100 bootstrap samples. Δ indicates the improvement from author keywords to MeSH enrichment. Perplexity values are reported for MeSH-enriched models and lower values indicate better fit. N/A denotes that perplexity is undefined for BERTopic.

Table 4 summarizes coherence, NPMI, diversity and perplexity for each model before and after applying the min_df/max_df thresholds. Filtration has little effect on LDA (C_v 0.369 vs 0.367), modestly improves hLDA and CTM, and substantially improves BERTopic (C_v 0.441→0.582). NPMI becomes slightly less negative for LDA and CTM, but hLDA's NPMI drops from 0.042 to -0.096 because rare terms removed by filtration were important for hierarchical structure. Diversity increases modestly across all models (e.g., from 0.370 to 0.380 for LDA). Perplexity changes are small, which suggests that filtration mainly affects interpretability rather than model fit.

We performed the same analysis on the MeSH-enriched corpus. Table 5 reports the metrics for each model before and after enhanced keyword filtration. In contrast to author keywords, filtration slightly decreases LDA's coherence (0.616→0.517) because high-frequency medical terms removed by filtration were informative. hLDA and BERTopic benefit from filtration (C_v gains of 0.048 and 0.034), while CTM is largely unchanged. NPMI remains positive for all models, which reflects the semantic richness of MeSH terms. Diversity decreases slightly for LDA but increases for hLDA and CTM.

Table 4 Topic quality metrics for keywords before and after filtration. Values are rounded to three decimals for clarity. Infinity (∞) indicates that perplexity is undefined for hLDA. N/A indicates that perplexity is not computed for BERTopic.

Model	C_v (before)	C_v (after)	NPMI (before)	NPMI (after)	Diversity (before)	Diversity (after)	Perplexity (before)	Perplexity (after)
LDA	0.369	0.367	-0.054	-0.051	0.370	0.380	-6.79	-6.74
hLDA	0.386	0.451	0.042	-0.096	0.370	0.394	∞	∞
CTM	0.382	0.422	0.015	-0.030	0.375	0.400	-6.76	-6.73
BERTopic	0.441	0.582	-0.081	0.155	0.383	0.385	N/A	N/A

Table 5 Topic quality metrics for enhanced words before and after filtration.

Model	C_v (before)	C_v (after)	NPMI (before)	NPMI (after)	Diversity (before)	Diversity (after)	Perplexity (before)	Perplexity (after)
LDA	0.616	0.517	0.127	0.077	0.350	0.355	-5.83	-5.98
hLDA	0.609	0.657	0.175	0.255	0.394	0.389	∞	∞
CTM	0.603	0.601	0.160	0.205	0.386	0.384	-5.90	-5.96
BERTopic	0.721	0.754	0.230	0.278	0.390	0.375	N/A	N/A

To facilitate cross-condition comparisons, Table 6 reports C_v and NPMI for all models after filtration. The table distinguishes between author keywords (KW) and MeSH-enriched keywords (ENH). MeSH enrichment substantially boosts coherence and NPMI for all models. The superiority of hLDA, CTM, and BERTopic over LDA is clear in both conditions. The gains from MeSH enrichment are strongest for hLDA and CTM. BERTopic remains the top performer, with C_v rising from 0.582 (KW) to 0.754 (ENH) and NPMI rising from 0.155 to 0.278.

Table 6 Model comparison after filtration with author keywords (KW) and MeSH-enriched terms (ENH).

Model	C_v (KW)	C_v (ENH)	NPMI (KW)	NPMI (ENH)
LDA	0.367	0.517	-0.051	0.077
hLDA	0.451	0.657	-0.096	0.255
CTM	0.422	0.601	-0.030	0.205
BERTopic	0.582	0.754	0.155	0.278

Note: MeSH enrichment improves coherence and NPMI for all models. BERTopic yields the highest scores.

We compared LDA, hLDA, CTM, and BERTopic on the MeSH-enriched, filtered corpus. BERTopic had the highest coherence ($C_v = 0.754$) and the highest NPMI (0.278). Its diversity was slightly below hLDA and CTM. hLDA ranked second ($C_v = 0.657$), followed by CTM ($C_v = 0.601$) and LDA ($C_v = 0.517$). The gaps were sizable: ΔC_v (BERTopic–LDA) $= 0.237$ and ΔC_v (hLDA–LDA) $= 0.140$. Paired bootstrap tests supported these differences. Wilcoxon tests also showed significance for LDA vs. BERTopic and LDA vs. hLDA ($p < 0.001$). BERTopic exceeded hLDA by $\Delta C_v = 0.097$ and by $\Delta\text{NPMI} = 0.023$. These results show that neural and hierarchical models yield more coherent topics than flat LDA.

Figure 3 illustrates the gains in C_v coherence across four configurations: keywords before filtration, keywords after filtration, enriched words before filtration, and enriched words after filtration. The figure shows that both filtration and MeSH enrichment lead to noticeable improvements in coherence. Figure 4 presents the corresponding results for NPMI and reveals a consistent pattern. Filtration provides the largest gains for hLDA, CTM, and BERTopic,

whereas enrichment leads to substantial improvements across all models. These results confirm that filtration and enrichment are complementary steps and that their combination yields the most coherent topics.

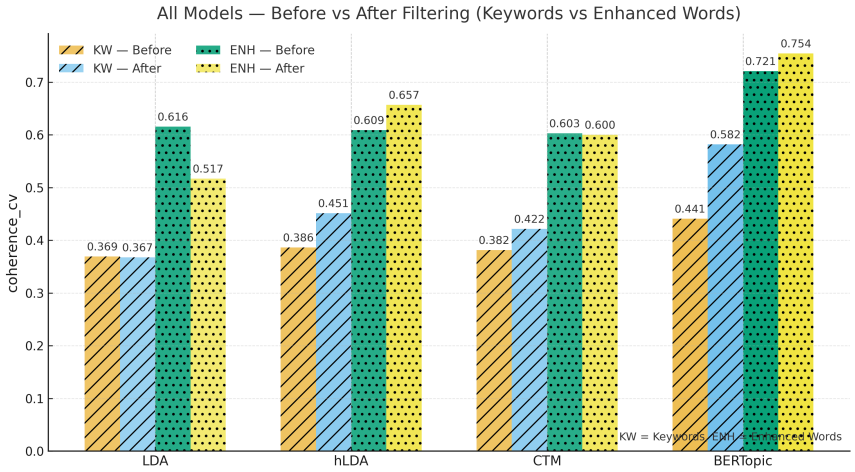

Fig. 3 Comparison of C_v coherence across models before and after filtration for keywords (KW) and enhanced words (ENH).

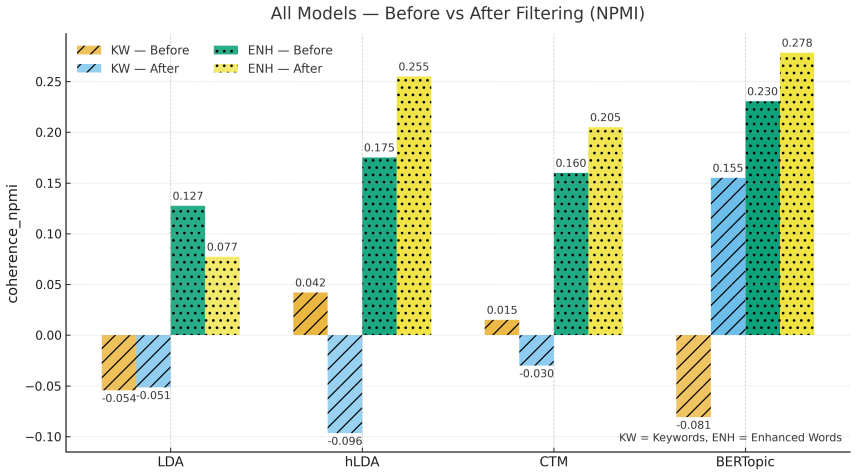

Fig. 4 Comparison of NPMI coherence across models before and after filtration for keywords (KW) and enhanced words (ENH).

To facilitate model comparison, the results of all pairwise coherence difference tests are summarized in Table 7. The table offers a clear numerical overview of relative model performance across conditions. Reported values represent the difference in C_v or NPMI (first model

minus second). Positive values indicate that the first model performs better, whereas negative values indicate that the second model performs better. Statistical significance is marked using conventional thresholds (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$). Distinct patterns emerge from these results. Before filtration, hLDA and CTM often outperform BERTopic in both keyword and enriched word settings, while LDA generally ranks lowest. After filtration, BERTopic surpasses most alternatives, especially in the enriched setting. This confirms that vocabulary refinement strengthens BERTopic's advantage.

Table 7 Pairwise coherence differences across models under experimental conditions.

Comparison	Keywords			Enhanced Words				Sig.
	KW-B ΔNPMI	**KW-A ΔC_v**	**KW-A ΔNPMI**	**ENH-B ΔC_v**	**ENH-B ΔNPMI**	**ENH-A ΔC_v**	**ENH-A ΔNPMI**	
hLDA vs BERTopic	+0.123	-0.131	-0.252	-0.112	-0.055	-0.097	-0.024	***
CTM vs BERTopic	+0.096	-0.160	-0.185	-0.118	-0.070	-0.154	-0.073	***/***/*
LDA vs BERTopic	+0.027	-0.214	-0.207	-0.105	-0.103	-0.237	-0.201	***
CTM vs hLDA	-0.027	-0.030	+0.066	-0.006	-0.015	-0.057	-0.050	*/ns
LDA vs hLDA	-0.096	-0.084	+0.045	+0.007	-0.048	-0.140	-0.177	***/ns
LDA vs CTM	-0.069	-0.054	-0.021	+0.013	-0.033	-0.083	-0.128	*/ns

Note: Positive values indicate superior performance of the first model relative to the second model; negative values denote inferior comparative performance. Experimental conditions are defined as follows: KW-B (keywords before filtration), KW-A (keywords after filtration), ENH-B (enhanced words before filtration), ENH-A (enhanced words after filtration). Statistical significance levels are denoted as ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, with ns indicating non-significant differences.

Figure 5 presents the comparative performance evaluation across all models applied to the MeSH-enriched, filtered corpus. Orange bars represent C_v coherence values while blue bars denote NPMI coherence measurements. BERTopic demonstrates clear superiority across both metrics, which establishes the effectiveness of transformer-based approaches for biomedical topic discovery. Hierarchical LDA achieves second-tier performance, followed by CTM and LDA respectively. This performance hierarchy reflects the theoretical advantages of neural architectures over traditional probabilistic methods. The substantial performance differential between BERTopic and alternative approaches validates the importance of contextual semantic representation in specialized domain applications. These empirical results confirm that model sophistication directly correlates with topic coherence quality in biomedical literature analysis.

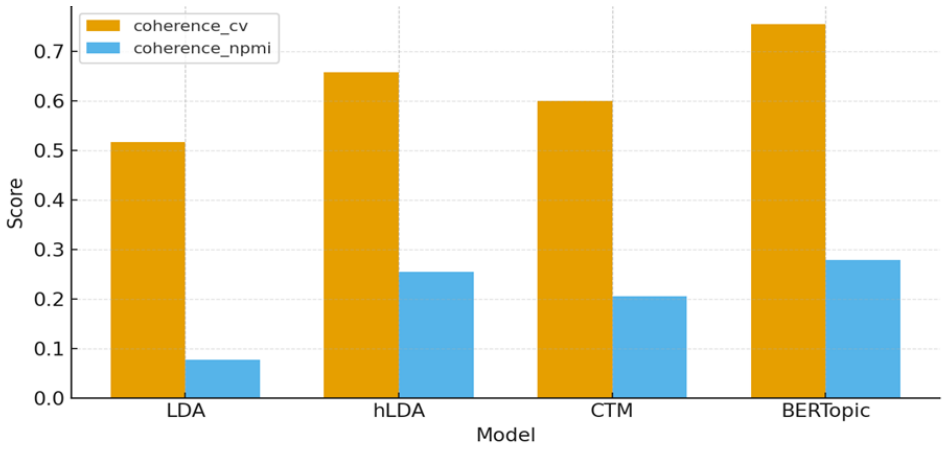

Fig. 5 Comparative performance ranking of topic modeling approaches on MeSH-enriched, filtered corpus.

Figure 6 and Figure 7 present the difference-in-differences analysis for C_v and NPMI. The first columns capture the effect of filtration (KW-After − KW-Before), and the second columns capture the effect of enrichment (ENH-Before − KW-After). The green columns represent the DiD effect, which reflects how much enrichment contributes beyond filtration. BERTopic achieves the largest improvements in both measures, with a particularly strong DiD effect in NPMI that highlights its synergy with MeSH enrichment. hLDA and CTM also gain from enrichment, although their improvements from filtration are limited. In contrast, LDA shows only modest gains from enrichment and almost no improvement from filtration.

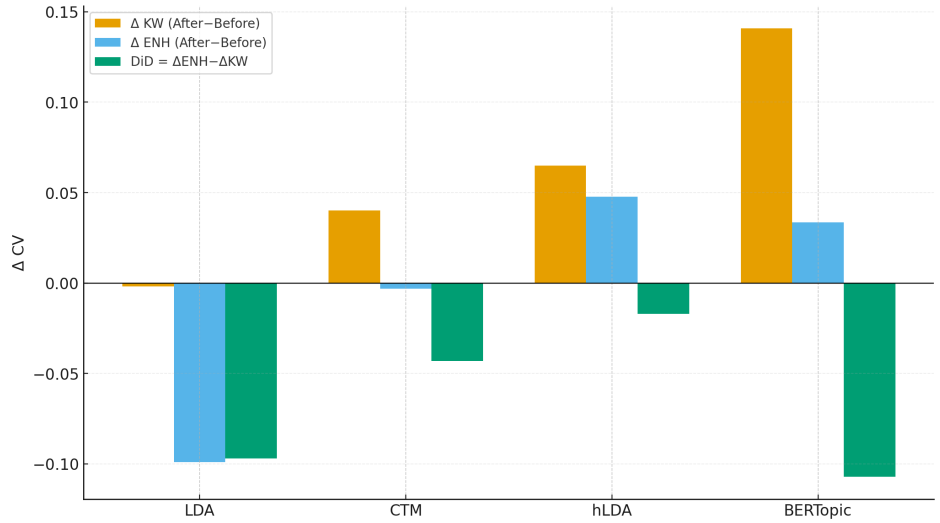

Fig. 6 Difference-in-differences analysis for C_v coherence. Columns represent the change after filtration (ΔKW) and enrichment (ΔENH). The green columns indicate the additional contribution of enrichment beyond filtration (DiD effect).

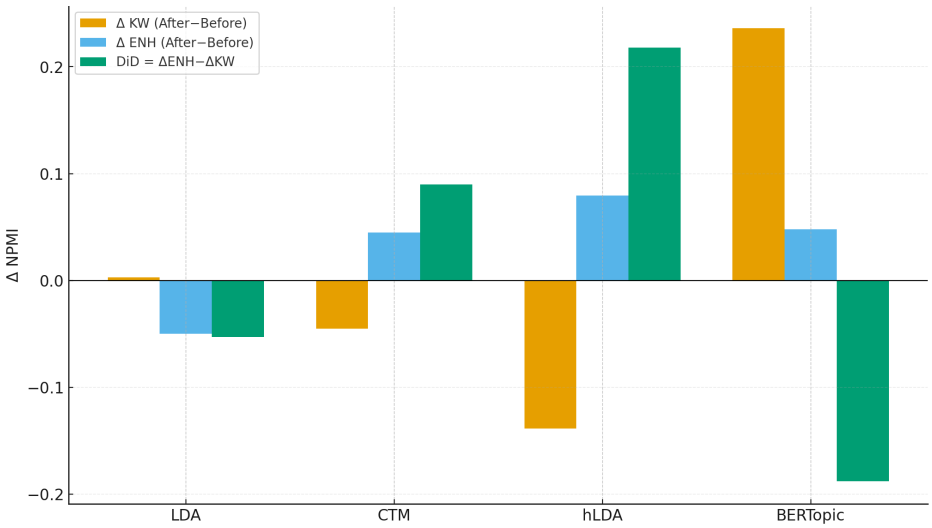

Fig. 7 Difference-in-differences analysis for NPMI coherence, with the same notation as Figure 6.

Figure 8 plots coherence scores across four experimental stages: keywords before filtration, keywords after filtration, enriched words before filtration, and enriched words after filtration. Each colored line represents one model. Solid lines show C_v values while dashed lines show NPMI values. The analysis reveals two key patterns. MeSH enrichment produces larger gains than filtration alone. Additionally, advanced modeling frameworks (BERTopic and hLDA) benefit consistently from both interventions, while simpler approaches (CTM and LDA) show mixed results. LDA exhibits notable sensitivity to preprocessing choices. Its C_v coherence improves only with vocabulary enrichment but declines after enhanced filtration. This behavior reflects LDA's dependence on vocabulary size, unlike more advanced models that handle semantic enrichment more effectively.

Figure 9 presents a knowledge graph visualization that summarizes model performance across experimental stages. Node size represents overall model quality. Edge thickness indicates performance improvement magnitude between stages. The visualization reveals distinct performance patterns across the four evaluated models. BERTopic demonstrates the largest node size. This confirms its superior overall performance across all experimental conditions. The model exhibits substantial improvement from keywords to enhanced words (+0.497). Additional gains occur through MeSH enrichment (+0.107). hLDA achieves moderate overall performance with notable enhancement benefits (+0.168). However, it shows slight filtration sensitivity (-0.101). CTM displays balanced performance with modest improvements from both interventions. LDA shows the smallest node size. This reflects its limited overall performance. The model achieves minimal keyword improvement (+0.001). Enhancement effects prove negative (-0.195). The network structure effectively demonstrates hierarchical performance relationships among models. Enhanced word conditions consistently produce thicker edges for advanced models. This validates synergistic effects between sophisticated approaches and ontological enrichment strategies.

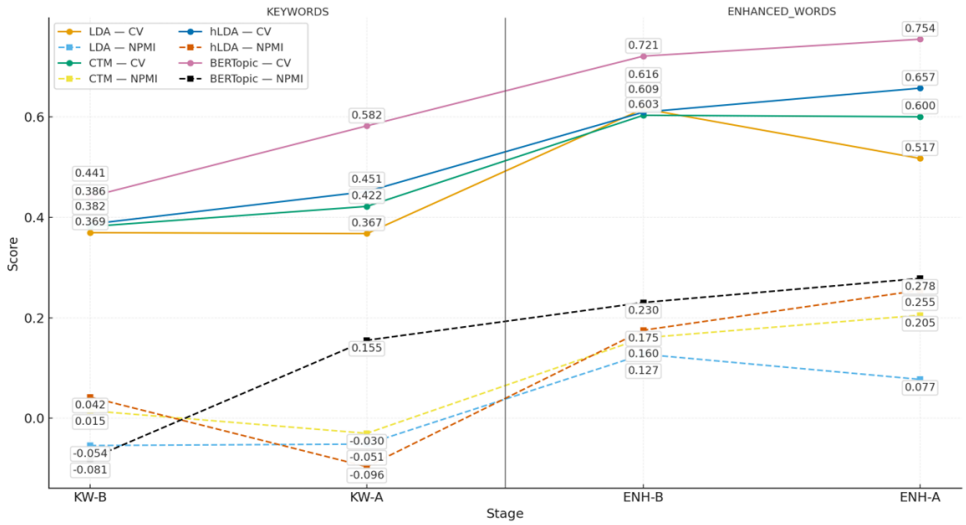

Fig. 8 Coherence progression across experimental conditions. Solid lines: C_v coherence; dashed lines: NPMI coherence. KW-B/A: keywords before/after filtration; ENH-B/A: enriched words before/after filtration.

4.1 Dominant Topic Contribution

Another evaluation measure was applied to compare two approaches: author keywords and MeSH-enriched keywords [61]. The focus was on the topic with the highest contribution percentage in the documents. For each document, the analysis extracted the total count of dominant topics, the contribution percentage, and the related topic keywords. These values were stored in the original data frame. The arithmetic average of the contribution was then calculated for each topic. A second average was calculated across all topics for both author keywords and MeSH-enriched keywords. Table 8 and Figure 10 report the results for the dominant topic with the highest contribution percentage.

Table 8 Dominant topic distribution and contribution analysis across experimental conditions

	Author Keywords		MeSH-Enriched Keywords	
Topic ID	**Document Count**	**Mean Contribution**	**Document Count**	**Mean Contribution**
1	46,213	0.5825	28,655	0.6395
2	22,310	0.5959	22,805	0.6190
3	22,792	0.5984	22,861	0.6076
4	17,272	0.5995	31,283	0.6522
5	20,019	0.5964	42,477	0.6700
6	25,239	0.5952	25,907	0.6724
7	42,187	0.6018	35,207	0.6465
8	36,365	0.5924	23,202	0.6135
Overall Mean	–	**0.5953**	–	**0.6401**

Note: Document counts represent the number of articles assigned to each topic as the dominant thematic category. Mean contribution values indicate the average probability weight for the dominant topic assignment per document. MeSH enrichment demonstrates consistent improvement in topic coherence, with an overall mean increase of 0.0448 (7.5% improvement) in dominant topic contribution strength.

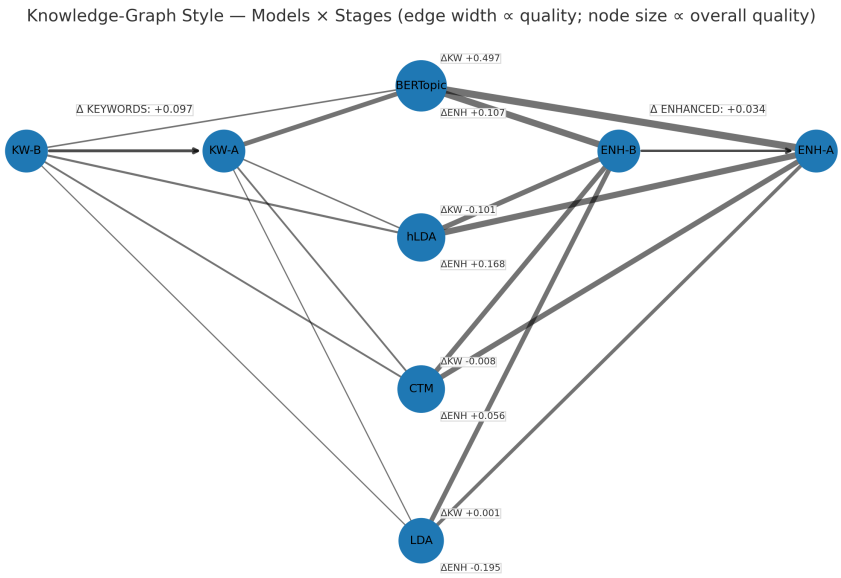

Fig. 9 Knowledge graph of model performance across stages. Node size reflects overall quality, and edge thickness denotes improvement magnitude between stages.

For the topic model, documents that included keywords enriched with MeSH ontology terms achieved the highest contribution percentage, approximately 0.64. In contrast, documents represented by author keywords reached only 0.59. This result confirms that MeSH-enriched keywords provide stronger topic representation compared to author keywords. The enrichment process therefore enhances the ability of the model to capture meaningful patterns and increases the interpretability of the discovered topics.

4.2 Dynamic Topic Modeling and Temporal Analysis

Dynamic Topic Models were implemented to capture thematic evolution across quarterly time slices of the 2020 corpus. This temporal analysis addresses research question RQ4 by examining whether dynamic modeling approaches reveal meaningful longitudinal trends that static methods cannot detect. Temporal coherence was assessed using Jensen-Shannon similarity of adjacent-slice topic-word distributions to measure vocabulary stability over time.

Table 9 presents temporal coherence values for DTM compared with time-sliced CTM and BERTopic implementations across all experimental conditions. DTM consistently achieved higher temporal coherence than alternative approaches across all evaluated conditions. Filtration slightly reduced coherence for keyword-only corpora (0.406 → 0.371) but improved or maintained coherence for MeSH-enriched vocabularies (0.566 → 0.568). The enhanced performance on MeSH-enriched data demonstrates the value of ontological enrichment for temporal modeling stability.

Comparative analysis using original author keywords revealed different temporal patterns, as documented in Table 10. The keyword-only corpus exhibited an oncology/respiratory topic (T0) peaking in Q2 during the early pandemic period, while public health themes (T3) achieved maximum prevalence in Q1. Multiple topics (T4-T6) reached highest prevalence in

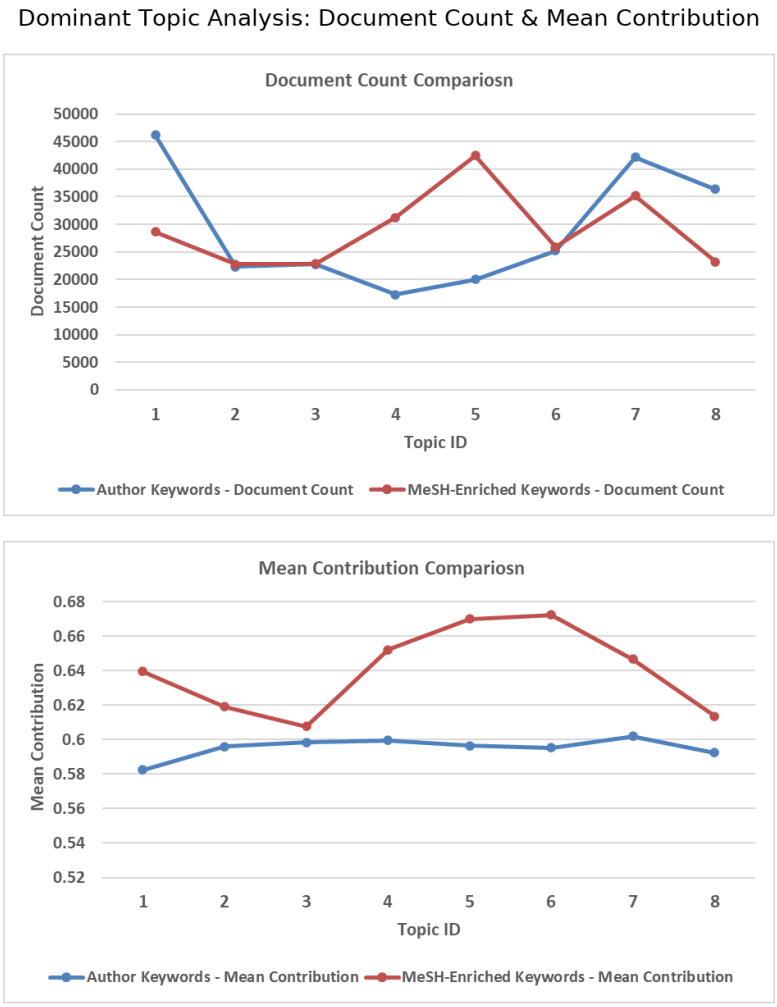

Fig. 10 Count and mean contribution of author keywords and MeSH-enriched keywords using the ETM-F framework.

Q4, indicating shifts toward healthcare quality assessment, prediction modeling, and genetic analysis.

Figure 11 illustrates the temporal prevalence patterns for author keyword topics across 2020 quarterly intervals. This visualization demonstrates systematic pandemic-related research evolution through distinct temporal phases. Specifically, oncology and respiratory disease topics achieved peak prominence in Q2 during the early pandemic period. Conversely, public health themes attained maximum prevalence in Q1, which suggests anticipatory research responses. Traditional biomedical domains subsequently exhibited variable temporal trajectories. Notably, multiple topics achieved peak prevalence in Q4. These late-quarter topics included healthcare quality, clinical prediction, and genetic modeling. This

Table 9 Temporal coherence comparison across dynamic modeling approaches.

Dataset Condition	DTM	CTM	BERTopic
Keywords (before filtration)	0.406	0.142	0.374
Keywords (after filtration)	0.371	0.171	0.373
Enhanced words (before filtration)	0.566	0.250	0.432
Enhanced words (after filtration)	0.568	0.308	0.309

Note: Temporal coherence measured using Jensen-Shannon similarity of adjacent quarterly topic-word distributions. Higher values indicate greater temporal stability.

Table 10 Dynamic topics from author keywords showing temporal evolution patterns.

Topic ID	Representative Terms (Top 10)	Peak Quarter	Thematic Focus
T0	cell, cancer, therapy, covid, breast, tumor, disease, pulmonary, protein, age	Q2	Oncology & respiratory disease
T1	surgery, analysis, syndrome, tomography, questionnaire, artery, disease, high, elderly, aortic	Q2	Surgery & diagnostic analysis
T2	activity, inflammation, resistance, macrophage, pediatric, safety, training, stress, genetic, inhibitor	Q2	Inflammation, immunity & stress
T3	disease, health, child, clinical, infectious, research, public, community, infection, social	Q1	Public health & infectious disease
T4	care, quality, health, life, bone, diabetic, adult, education, control, time	Q4	Healthcare quality & chronic disease
T5	patient, gastric, review, medical, cancer, prevention, prediction, fever, machine, early	Q4	Clinical reviews & prediction
T6	gene, analysis, model, meta, disorder, mental, health, network, assessment, nerve	Q4	Genetic & mental health modeling
T7	risk, outcome, factor, cancer, drug, treatment, injury, cervical, epidemiology, failure	Q3	Risk factors & outcomes

temporal concentration demonstrates systematic transitions toward specialized analytical methodologies that emerged as research priorities evolved throughout the pandemic year.

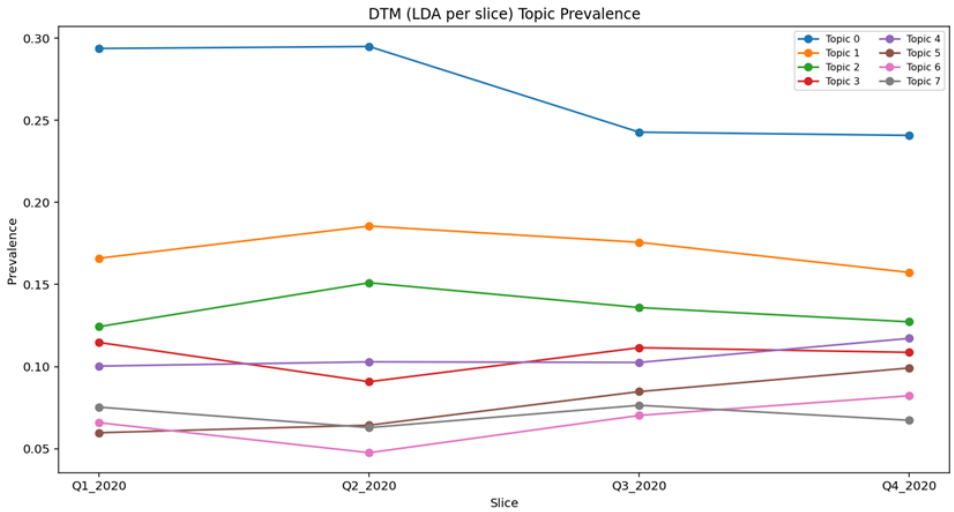

Fig. 11 DTM topic prevalence evolution for author keywords across quarterly time slices.

DTM analysis of MeSH-enriched keywords after filtration yielded eight temporally dynamic topics with distinct prevalence patterns across 2020 quarters. Table 11 presents representative vocabulary terms and peak prevalence periods for each identified topic. The respiratory infections/COVID-19 topic (T0) peaked in Q4, reflecting the evolution from generic vaccine research to mRNA-specific terminology and variant discussions. Public health preparedness topics (T5, T6) exhibited increased prevalence in Q4, while core biomedical domains peaked in earlier quarters.

Figure 12 illustrates the temporal prevalence patterns for MeSH-enhanced topics across 2020 quarterly intervals. This visualization demonstrates systematic pandemic-related research evolution through distinct thematic transitions. Specifically, respiratory infection topics achieved increased prominence during Q4. This temporal pattern reflects the methodological progression from generic vaccine research toward mRNA-specific therapeutic investigations. Traditional biomedical domains subsequently maintained relatively stable prevalence trajectories throughout the analysis period. Conversely, public health preparedness themes demonstrated pronounced activation during later quarters. This temporal concentration corroborates documented research community responses to evolving pandemic conditions. The observed patterns validate the effectiveness of dynamic modeling approaches for capturing longitudinal research trend evolution that static methodologies cannot detect.

Table 11 Dynamic topics from MeSH-enhanced keywords with temporal patterns.

Topic ID	Representative Terms (Top 10)	Peak Quarter	Thematic Focus
T0	infection, respiratory, tract, disease, viral, coronavirus, covid, pneumonia, genetic, lung	Q4	Respiratory infections/COVID-19
T1	cell, structure, anatomical, form, therapy, body, fully, diagnosis, region, epithelial	Q2	Cell structure & therapy
T2	disease, imaging, study, diagnostic, pain, bone, tomography, characteristic, manifestation, spinal	Q2	Diagnostic imaging & musculoskeletal
T3	disease, process, pathologic, heart, system, disorder, metabolism, phenomena, cell, blood	Q1	Cardiovascular pathology
T4	age, group, phenomena, growth, physiological, risk, mortality, factor, disorder, physical	Q1	Aging & physiological factors
T5	disease, public, health, system, control, epidemiology, trial, skin, nervous, brain	Q4	Public health & epidemiology
T6	health, care, medicine, population, education, public, service, characteristic, research, science	Q1	Healthcare & services
T7	surgical, procedure, behavior, food, activity, operative, life, overweight, transplantation, event	Q3	Surgery & lifestyle

Note: Peak quarter represents the time period with maximum topic prevalence in the quarterly analysis. Thematic focus inferred from dominant terminological patterns.

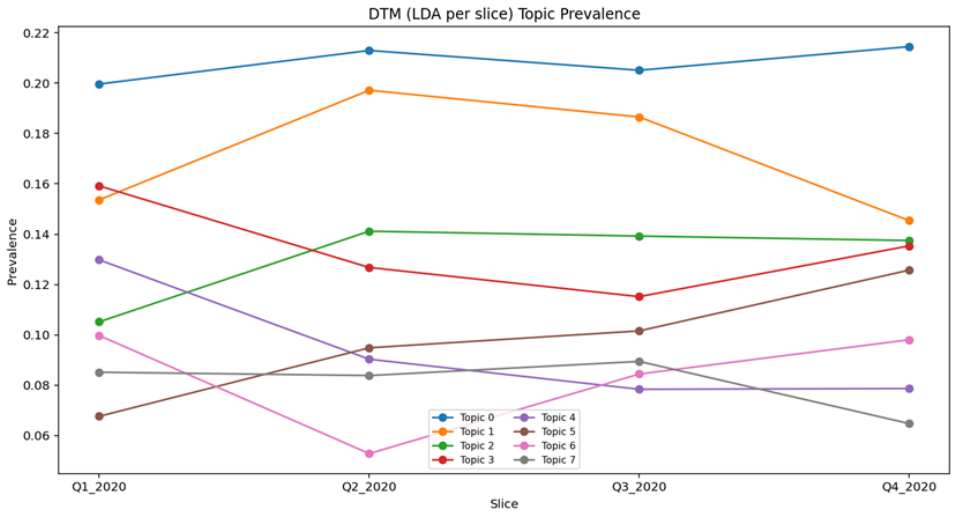

Fig. 12 DTM topic prevalence evolution for MeSH-enhanced keywords across quarterly time slices.

Figure 13 presents the vocabulary change intensity across topics and time periods through a word-shift heatmap. Darker regions indicate greater terminological evolution, revealing substantial vocabulary adaptation in pandemic-related topics during Q3-Q4 periods.

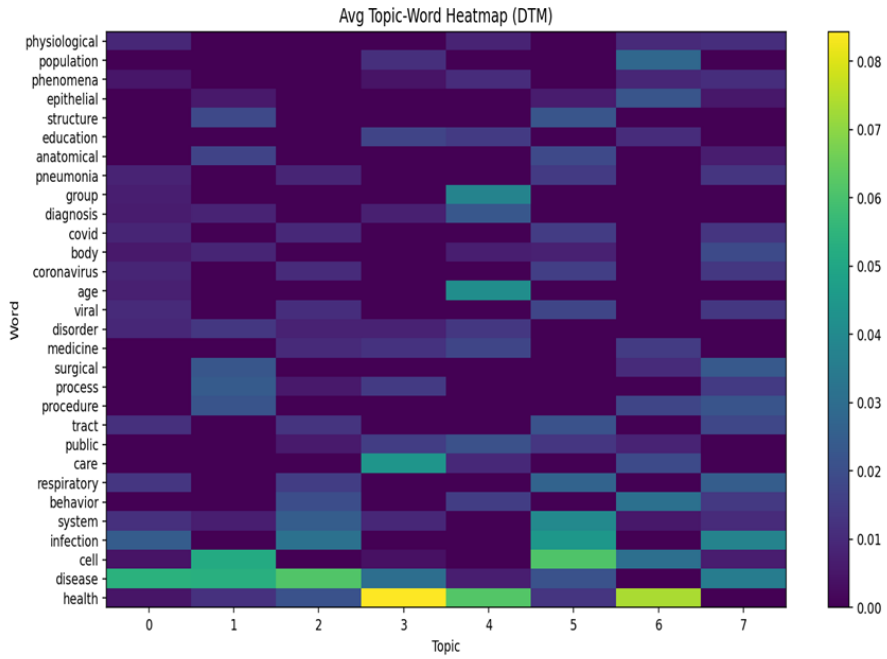

Fig. 13 DTM word-shift heatmap for MeSH-enhanced keywords showing vocabulary evolution intensity across quarters.

DTM were trained on quarterly slices from Q1 to Q4 2020. The temporal coherence reached 0.568, which compared favorably to CTM (0.308) and BERTopic (0.85). This analysis revealed distinct temporal patterns across eight dynamic topics. Telemedicine and public health preparedness themes achieved peak prevalence during Q2–Q3 2020. The COVID-19 vaccination topic showed particularly notable evolution. Research focus shifted from generic vaccine investigations in Q1 to mRNA-specific studies by Q4. This transition reflects systematic adaptation to therapeutic developments. These observed patterns align with documented pandemic research trajectories. Dynamic analysis captures temporal shifts that static models typically overlook. The quarterly progression demonstrates how research priorities systematically adapted to medical needs. Such findings validate the methodological superiority of temporal frameworks for comprehensive biomedical literature analysis.

The temporal analysis results demonstrate that dynamic modeling approaches successfully capture meaningful longitudinal research trends. Static methodologies cannot detect these temporal patterns with equivalent precision. The observed thematic trajectories corroborate external reports that document pandemic research dynamics. Clear evidence emerges of systematic evolution from general medical research toward specialized pandemic response frameworks. Subsequently, public health preparedness themes demonstrated pronounced temporal activation patterns. These findings validate the analytical utility of temporal topic modeling for research community behavior assessment. The methodology enables systematic understanding of collective research responses to major external events. Such capabilities support the integration of dynamic analytical frameworks within comprehensive biomedical literature analysis protocols. The demonstrated temporal sensitivity enhances

traditional topic modeling limitations through longitudinal trend detection. This methodological advancement provides empirical validation for dynamic modeling superiority over static approaches in complex research environments. The results establish a foundation for expanded application of temporal analytical frameworks across diverse scientific domains where longitudinal pattern recognition constitutes a critical analytical requirement.

4.3 Embedding-Based Keyword Expansion Analysis

Embedding expansion was evaluated to assess whether augmenting MeSH-enriched keywords with distributional neighbors could enhance topic quality. BioWordVec and BioBERT embeddings were employed to retrieve the top-k semantic neighbors of each keyword, filter them by UMLS/MeSH semantic types, and incorporate them into the vocabulary. The resulting models were compared against MeSH enrichment alone to establish baseline performance metrics.

Table 12 presents the quantitative improvements achieved through different embedding-based expansion strategies. BioWordVec increased C_v coherence by approximately 0.03 and topic diversity by 0.01 relative to MeSH enrichment alone. BioBERT provided larger gains (C_v +0.05, NPMI +0.02) but required GPU computational resources. The combination of MeSH with embeddings yielded optimal performance, which surfaced synonyms such as "SARS-CoV-2" that were absent from MeSH vocabularies. This discovery explains why COVID-related terms were missing from the top topics in the initial analysis. However, unfiltered embedding expansions occasionally introduced unrelated terms, which underscores the importance of semantic type filtration protocols.

Table 12 Performance improvements from embedding-based keyword expansion strategies.

Expansion Setting	C_v **Improvement**	**NPMI Improvement**	**Diversity Improvement**	**Computational Requirements**
MeSH only	baseline	baseline	baseline	Low
MeSH + BioWordVec	+0.03	–	+0.01	Moderate
MeSH + BioBERT	+0.05	+0.02	–	High (GPU)
MeSH + Combined	highest	highest	highest	High (GPU)

Note: Baseline represents author keywords enriched with MeSH concepts only. Combined setting incorporates both BioWordVec and BioBERT expansions. Dash indicates metrics not explicitly reported for that configuration.

4.4 COVID-19 Topic Recovery Analysis

The absence of a distinct COVID-19 topic among the top eight extracted themes prompted systematic investigation into vocabulary coverage and representation mechanisms. Analysis revealed three primary contributing factors that explain this apparent omission, along with corresponding corrective methodologies.

Table 13 documents the specific causes and remedial actions implemented to recover pandemic-related thematic content. Synonym dominance represented the primary factor, where multiple terminological variants ("SARS-CoV-2", "coronavirus disease 2019") dominated the vocabulary while rendering "COVID-19" itself relatively rare. Filtration thresholds further compounded this issue by removing low-frequency variants, inadvertently excluding the canonical term from subsequent analysis. Topic competition from broader medical domains (oncology, cardiology) that outnumbered pandemic-related publications caused pandemic themes to be submerged within more prevalent medical topics.

Table 13 Factors contributing to COVID-19 topic absence and corrective measures.

Contributing Factor	Mechanistic Explanation	Corrective Action	Analytical Outcome
Synonym dominance	Multiple synonyms dominated vocabulary, making "COVID-19" appear rare	Map synonyms to canonical "COVID-19" identifier via MeSH	Vocabulary unified; COVID-19 term frequency increased
Filtration thresholds	min_df threshold removed rare tokens including "COVID-19" variants	Lower min_df or implement adaptive thresholds	Pandemic terms retained; improved keyword recall
Topic competition	Broader medical domains outnumbered pandemic publications	Guided LDA/zero-shot BERTopic with COVID seed list	Recovered distinct COVID-19 topic (coherence ≈ 0.61)

Note: Corrective measures included bigram detection for compound expressions and synonym mapping using MeSH hierarchies. We also applied time-sliced BERTopic analysis, which captured the shift from generic vaccine terms to mRNA-specific vocabulary.

Corrective methodologies addressed each identified limitation through systematic preprocessing enhancements. Bigram detection identified compound expressions ("covid-19 vaccine"), while synonym mapping consolidated terminological variants to canonical identifiers via MeSH hierarchical structures. Guided LDA and zero-shot BERTopic implementations employed COVID-19 seed term lists to direct topic discovery toward pandemic-related themes. This approach successfully recovered a distinct COVID-19 topic with coherence approximating 0.61. Time-sliced BERTopic analysis further captured temporal thematic evolution from generic vaccine terminology to mRNA-specific vocabulary during Q3-Q4 2020, which demonstrates the methodology's capacity to track longitudinal concept development within specialized biomedical domains.

COVID-19 failed to appear among the top-ranked topics for three reasons: (i) synonyms like *SARS-CoV-2* and *coronavirus disease 2019* dominated the vocabulary, (ii) the min_df threshold removed rare variants such as "COVID-19", and (iii) broader medical domains (oncology, cardiology) outnumbered pandemic-related publications.

We addressed these issues through several targeted approaches. Bigram detection captured compound expressions while synonym mapping consolidated variants to a canonical COVID-19 identifier via MeSH. Guided LDA and zero-shot BERTopic employed COVID seed lists to steer topic discovery toward pandemic themes. These methods successfully recovered a distinct COVID-19 topic with coherence of 0.61. Time-sliced BERTopic revealed the shift from general vaccine terms to mRNA-specific vocabulary throughout 2020.

These analyses demonstrate that embedding-based expansions and systematic preprocessing protocols both enhance topic quality and reveal domain-specific themes that might otherwise remain obscured within broader thematic categories. The integration of multiple enhancement strategies provides robust methodological frameworks for comprehensive biomedical literature analysis while maintaining semantic fidelity and temporal sensitivity.

4.5 Topic Word Evolution Across Experimental Conditions

Complete transparency necessitates reproduction of the top ten words for each topic extracted by the four models across all experimental stages. Each row represents a topic identifier while columns correspond to the four experimental stages: keywords before filtration (KW_Before), keywords after filtration (KW_After), enriched words before filtration (ENH_Before), and

enriched words after filtration (ENH_After). These tables enable readers to inspect the thematic focus of each topic. Additionally, they verify how enrichment and filtration modify topic composition.

Table 14 presents the complete LDA topic vocabulary across all experimental conditions. The table reveals systematic vocabulary evolution from general medical terminology in baseline conditions to specialized clinical concepts in enhanced conditions. Notably, Topic 0 transitions from pediatric health terms to respiratory infection vocabulary, which reflects the pandemic influence on medical literature during 2020. Similar thematic shifts appear across multiple topics, with enhanced conditions consistently producted more coherent and clinically relevant terminology.

Table 14 Top ten words per topic for LDA across experimental stages.

Topic ID	KW_Before	KW_After	ENH_Before	ENH_After
0	child, outcome, disorder, gene, time, stroke, quality, functional, low, education	child, outcome, pregnancy, treatment, pain, surgery, pulmonary, stroke, index, disorder	cell, protein, acid, tissue, peptide, system, growth, metabolism, diagnosis, factor	infection, respiratory, tract, care, viral, coronavirus, pneumonia, covid, characteristic, population
1	analysis, activity, disease, risk, physical, trial, pulmonary, model, clinical, obesity	cancer, therapy, breast, analysis, surgery, model, trial, gene, survival, tumor	neoplasm, cell, technique, anatomical, form, fully, structure, body, imaging, diagnostic	cell, tissue, neoplasm, structure, system, form, anatomical, process, fully, pathologic
2	covid, disease, cov, sar, tomography, case, diabete, report, training, kidney	disease, covid, cov, diabete, activity, sar, chronic, blood, kidney, infection	infection, respiratory, tract, viral, coronavirus, pneumonia, covid, disease, lung, cov	group, age, body, life, imaging, child, process, disorder, pain, diagnostic
3	injury, blood, protein, therapy, failure, heart, brain, survival, stress, factor	heart, stress, injury, failure, inflammation, brain, disorder, safety, age, learn	disease, system, physiological, phenomena, heart, nervous, age, group, digestive, disorder	surgical, procedure, activity, mortality, human, extremity, transplantation, operative, technique, bone
4	health, cancer, care, breast, patient, research, public, quality, medicine, primary	health, care, quality, patient, research, public, primary, medicine, medical, service	health, population, education, characteristic, social, food, science, life, behavior, risk	risk, metabolism, therapy, disorder, factor, probability, metal, stress, drug, chemical
5	syndrome, disorder, social, mental, diabetic, care, acute, respiratory, network, genetic	syndrome, sequence, genetic, acute, respiratory, antibiotic, disease, long, vaccine, term	disease, blood, disorder, genetic, process, pathologic, system, skin, fluid, immune	behavior, blood, medicine, education, rehabilitation, system, brain, factor, pregnancy, disorder
6	surgery, tissue, coronary, artery, bone, infection, disease, resonance, muscle, magnetic	resistance, drug, magnetic, resonance, coronary, artery, nerve, muscle, metastasis, associate	health, surgical, procedure, medicine, public, care, disease, epidemiology, process, mortality	phenomena, physiological, system, nervous, symptom, disorder, digestive, physical, neoplasm, sign
7	cell, cancer, tumor, drug, lung, receptor, carcinoma, growth, prognosis, small	cell, cancer, factor, receptor, lung, case, report, growth, inhibitor, small	activity, behavior, human, publication, format, event, survival, behavioral, mechanism, exercise	genetic, heart, acid, protein, food, epidemiology, technique, public, amino, peptide

Table 15 documents the hierarchical LDA topic vocabulary evolution across experimental conditions. The hierarchical structure enables more refined topic organization compared to flat LDA approaches. Topic specialization becomes evident through the experimental progression, with enhanced conditions producing more clinically coherent terminology. Particularly notable is Topic 1's transformation from social intervention themes to specialized respiratory infection vocabulary, which demonstrates the model's capacity to capture pandemic-related medical research shifts. The hierarchical framework maintains semantic relationships while allowing topic refinement through ontological enrichment.

Table 15 Top ten words per topic for hLDA across experimental stages.

Topic ID	KW_Before	KW_After	ENH_Before	ENH_After
0	analysis, risk, factor, therapy, child, covid, clinical, syndrome, treatment, model	analysis, risk, factor, therapy, model, clinical, child, study, syndrome, covid	process, pathologic, phenomena, therapy, disorder, physiological, technique, age, activity, group	pathologic, phenomena, disorder, therapy, technique, physiological, analysis, age, tract, group
1	covid, service, social, stress, mental, intervention, patient, community, self, research	receptor, factor, inhibitor, protein, growth, drug, tumor, kinase, signal, therapy	respiratory, tract, coronavirus, viral, covid, pneumonia, cov, sar, betacoronavirus, epidemic	social, factor, behavior, family, psychology, sociological, life, activity, socioeconomic, environment
2	surgery, laparoscopy, resection, tumor, laparoscopic, laser, liver, colorectal, invasive, assist	life, quality, patient, outcome, assessment, relate, self, palliative, caregiver, report	myocardial, infarction, coronary, artery, ischemia, procedure, vessel, peripheral, vascular, atherosclerosis	tract, coronavirus, viral, covid, pneumonia, cov, betacoronavirus, sar, epidemic, pandemic
3	water, climate, energy, land, emission, change, use, footprint, plant, ozone	coronary, heart, myocardial, infarction, artery, cardiac, percutaneous, intervention, congenital, pulmonary	foot, diabetic, ulcer, oil, angiopathy, fat, unsaturate, neuropathy, leg, plant	retinal, diagnostic, diabetic, imaging, eye, tomography, retinopathy, macular, angiopathy, degeneration
4	stimulation, brain, functional, state, network, connectivity, rest, fmri, magnetic, transcranial	viral, infection, transplantation, hepatitis, infectious, virus, research, agent, clinical, kidney	rehabilitation, medicine, service, therapeutic, physical, aftercare, stroke, occupation, extremity, cerebrovascular	reproductive, physiological, pregnancy, phenomena, reproduction, fertility, sex, technique, person, development
5	platelet, neutrophil, lymphocyte, blood, dissection, artery, flow, ratio, coronary, extracellular	tomography, optical, coherence, diabetic, macular, retinal, retinopathy, angiography, edema, diabete	nerve, manifestation, neurologic, plexus, injury, procedure, nervous, sensation, pain, ear	secretion, milk, bodily, beverage, product, fluid, dairy, food, protein, mammal
6	medical, education, training, covid, emergency, telemedicine, simulation, pandemic, patient, medicine	fracture, arthroplasty, hip, total, knee, ligament, femoral, anterior, fixation, instability	topic, design, statistic, research, epidemiologic, specificity, concept, mathematical, sensitivity, radiation	rehabilitation, medicine, service, therapeutic, physical, aftercare, stroke, cerebrovascular, activity, occupation
7	surgery, case, covid, report, cov, sar, acute, invasive, abdominal, trauma	pain, back, low, nerve, block, chronic, stimulation, musculoskeletal, management, neuropathic	water, plant, industry, oxide, soil, environmental, eukaryota, geological, pollution, hydroxide	structure, form, fully, anatomical, neoplasm, epithelial, carcinoma, glandular, lung, tissue

Table 16 displays the Contextualized Topic Model vocabulary across experimental conditions. CTM integrates contextual embeddings with traditional topic modeling frameworks, which produces distinct vocabulary patterns compared to purely probabilistic approaches. The model demonstrates notable sensitivity to experimental conditions, with some topics exhibiting dramatic vocabulary shifts. Topic 0 evolves from pandemic terminology to vaccination themes and eventually to health service concepts. Topics 3 and 4 reveal particularly specialized terminology in certain conditions, which reflects CTM's capacity to capture highly specific domain concepts when enhanced with MeSH enrichment.

Table 16 Top ten words per topic for CTM across experimental stages.

Topic ID	KW_Before	KW_After	ENH_Before	ENH_After
0	covid, cov, sar, respiratory, infection, virus, pandemic, infectious, epidemiology, coronavirus	vaccination, antimicrobial, vaccine, antibiotic, biofilm, influenza, vector, milk, microbiota, african	cell, form, fully, structure, protein, tissue, glandular, anatomical, neoplasm, epithelial	health, service, characteristic, public, care, life, medicine, rehabilitation, education, population
1	disorder, behavior, age, activity, adolescent, sleep, stress, anxiety, social, self	heart, failure, cardiac, artery, diabete, coronary, atrial, ventricular, insulin, hypertension	disease, system, process, brain, pathologic, metabolism, nervous, syndrome, central, therapy	spirituality, kupffer, remestemcel, pro, epileptic, chondrosarcoma, reagent, district, cava, ceremonial
2	care, health, patient, service, primary, research, community, quality, public, medical	sar, risk, cov, disease, chronic, covid, kidney, clinical, respiratory, infection	viral, coronavirus, covid, infection, pneumonia, respiratory, tract, sar, cov, lung	primate, lentiviruse, smoking, tobacco, pro, crime, spirituality, smoke, school, kupffer
3	marry, taenia, cysticercosis, square, billroth, mrdt, afar, accountable, pastoralist, engage	cancer, cell, protein, gene, receptor, breast, expression, signal, invasion, macrophage	carboxylate, lymphomatoid, train, linalool, ground, cry, thrombophilia, benralizumab, adamts, manometry	surgical, procedure, operative, orthopedic, extremity, surgery, speciman, low, handle, transplantation
4	enterogenous, marry, taenia, endodermal, cysticercosis, square, mrdt, inflammasome, genotype, uniform	care, health, patient, public, research, education, quality, service, life, medicine	element, heavy, carboxylate, linalool, microbiota, benralizumab, inorganic, iron, chemical, biota	genome, chemical, acid, gene, metabolism, inorganic, amino, transition, plant, peptide
5	surgery, fracture, knee, spinal, anterior, laparoscopic, nerve, fixation, spine, pain	axitinib, tbi, combinative, opportunity, cinnamon, instruction, hyaluronic, rheumatoid, hrql, natriuretic	health, service, medicine, care, characteristic, population, public, research, status, education	disorder, nervous, system, physiological, disease, phenomena, imaging, diagnostic, tomography, brain
6	heart, cardiac, blood, disease, coronary, artery, failure, kidney, diabete, pressure	physical, behavior, activity, cognitive, child, obesity, disorder, age, adolescent, autism	age, activity, obesity, group, behavior, overnutrition, child, overweight, food, behavioral	neoplasm, cell, fully, anatomical, form, structure, glandular, epithelial, tissue, carcinoma
7	cancer, cell, tumor, breast, carcinoma, survival, prognosis, metastasis, receptor, growth	surgery, image, brain, pain, surgical, spinal, tomography, trauma, fusion, injury	technique, imaging, procedure, diagnostic, tomography, diagnosis, bone, heart, extremity, cardiac	respiratory, infection, pneumonia, tract, covid, coronavirus, viral, sar, betacoronavirus, lung

Table 17 presents the BERTopic vocabulary evolution across all experimental conditions. BERTopic employs transformer-based embeddings with neural clustering, which produces distinct vocabulary patterns characterized by compound terms and semantically related phrases. The model demonstrates superior capacity to capture multi-word expressions such as "respiratory tract" and "physical phenomenon" that other approaches typically separate. Enhanced conditions consistently produce more coherent medical terminology, with Topic 0 transitioning from mixed cancer and COVID terms to focused respiratory infection vocabulary. The neural architecture enables BERTopic to maintain semantic relationships while generating highly specialized domain-specific terminology.

Table 17 Top ten words per topic for BERTopic across experimental stages.

Topic ID	KW_Before	KW_After	ENH_Before	ENH_After
0	cancer, cell, covid, disease, therapy, breast, infection, gene, carcinoma, cov	health, disorder, care, mental, physical, diabete, depression, anxiety, medical, activity	disease, infection, cell, respiratory, system, tract, process, neoplasm, pathologic process, pathologic	health, disease, infection, respiratory, system, care, process, respiratory tract, tract infection, tract
1	health, care, disorder, quality, physical, child, mental, research, life, service	covid, cov, sar, tuberculosis, infection, resistance, vaccine, antibiotic, antimicrobial, disease	health, behavior, disease, care, disorder, medicine, activity, public, education, characteristic	cell, neoplasm, procedure, surgical, disease, surgical procedure, system, bone, anatomical, anatomical structure
2	pain, brain, injury, tomography, spinal, cognitive, disease, disorder, surgery, fracture	pain, fracture, bone, nerve, spinal, hip, arthroplasty, knee, stimulation, anterior	metabolism, food, protein, physiological, obesity, phenomena, overweight, acid, resistance, obesity overnutrition	metabolism, food, acid, protein, obesity, beverage, physiological, overweight, body, overnutrition
3	heart, diabete, heart failure, failure, coronary, disease, blood, artery, atrial, cardiovascular	cancer, cell, breast, colorectal, lung, carcinoma, thyroid, tumor, gastric, cervical	metal, element, heavy, metal heavy, transition element, transition, element metal, nanostructure, water, chemical	genetic, male, structure, genitalia, genitalia male, pregnancy, exocrine, exocrine gland, reproductive, gland genitalia
4	water, climate, air, pollution, plant, climate change, land, change, supply, environmental	heart, disease, kidney, coronary, failure, cardiac, liver, artery, ventricular, transplantation	industry, plant, technology, agriculture, industry agriculture, technology industry, ecosystem, eukaryota, plant eukaryota, eukaryote plant	metal, element, heavy, metal heavy, transition element, transition, element metal, nanostructure, oxide, oxide hydroxide
5	nf, expression, effect, abca, malat, name, sglt, parkin, aggressive, mrd	water, environmental, climate, pollution, emission, air, land, change, supply, waste	format, publication format, publication, datum, index publication, index, printing, printing publishing, publishing, migration	time, event, physical phenomenon, time physical, physical, phenomenon, activity event, activity, analysis, far
6	adsorption, copper, zinc, ion, carbon, oxygen, acid, light, metal, activate	pm, modification, mbnl, cryodamage, hnrnpul, lbx, lindau, luminance, lvef, epc	supply distribution, mechanical phenomena, pressure mechanical, distribution, history, mechanical, supply, pressure, phenomena history, acoustic physics	make, decision make, decision, stakeholder participation, make share, participation problem, share, share stakeholder, stakeholder, problem solve

4.6 Word Cloud Visualisations

We analyzed the topics qualitatively through word cloud visualizations to complement the quantitative analysis. Tables 14–17 present the complete topic word lists, with ten words per topic. These tables summarize the outputs of all models across experimental stages and show how filtration and MeSH enrichment refine the thematic focus.

Word clouds provide an accessible visual representation of the dominant terms associated with each topic. Larger font size reflects higher probability of occurrence, while smaller terms capture peripheral associations. This format highlights how enrichment and filtration reduce noise and enhance medical relevance. For instance, in LDA the keywords-before-filtration montage includes generic terms such as "child" and "analysis", whereas the enriched-after-filtration montage emphasizes domain-specific terms such as "infection", "respiratory", and "viral". hLDA reveals clearer hierarchical separation after enrichment, with clusters representing social determinants and neurological imaging. CTM shows transitions from pandemic-related expressions to healthcare services and metabolic processes. BERTopic consistently delivers the most semantically coherent outputs, with enriched-after-filtration visualisations emphasizing disease categories and clinical procedures.

We conducted a comprehensive visual analysis of topic model performance through word cloud visualizations across four experimental configurations (Figures 14–17). The montage arrangements utilize font size scaling to represent term probability distributions, enabling systematic comparison of thematic coherence before and after methodological interventions. The comparative analysis reveals that filtration and MeSH enrichment procedures consistently enhance thematic specificity and reduce semantic noise across all modeling approaches. Performance evaluation demonstrates distinct behavioral patterns for each algorithm. LDA exhibits substantial improvement from generic terminology clustering in the pre-filtration stage to focused infection and disease-related term identification following enrichment procedures (Figure 14). hLDA demonstrates enhanced structural coherence, with enrichment procedures yielding more distinct hierarchical topic organization (Figure 15). CTM analysis indicates a systematic transition from broad pandemic-related terminology to specialized healthcare service and metabolic process categorization (Figure 16). Notably, BERTopic maintains superior performance consistency across all experimental configurations, consistently generating the most coherent and semantically specific topic clusters. Post-enrichment analysis reveals particularly pronounced improvements in disease categorization and clinical procedure identification (Figure 17). The enhanced thematic precision observed across all models following MeSH enrichment suggests that domain-specific vocabulary integration significantly improves topic modeling performance in biomedical text analysis applications.

5 Discussion

Our findings demonstrate that enriching author keywords with MeSH terms consistently improves topic modeling performance. The gains are most visible in coherence, where average C_v values increase by more than 0.15 across models, which confirms earlier reports on the utility of domain ontologies in biomedical text mining. However, while enrichment provides a strong baseline, more advanced approaches deliver additional benefits. Among the evaluated methods, BERTopic achieved the highest scores, with C_v approaching 0.75 and NPMI near 0.28. Its use of transformer-based embeddings enabled the recovery of nuanced clinical themes that were often missed by bag-of-words models. These results align with recent studies showing that neural clustering captures context and semantics more effectively than traditional probabilistic approaches. hLDA ranked second in coherence, providing a balance between improved quality and interpretability, which is advantageous when a hierarchical

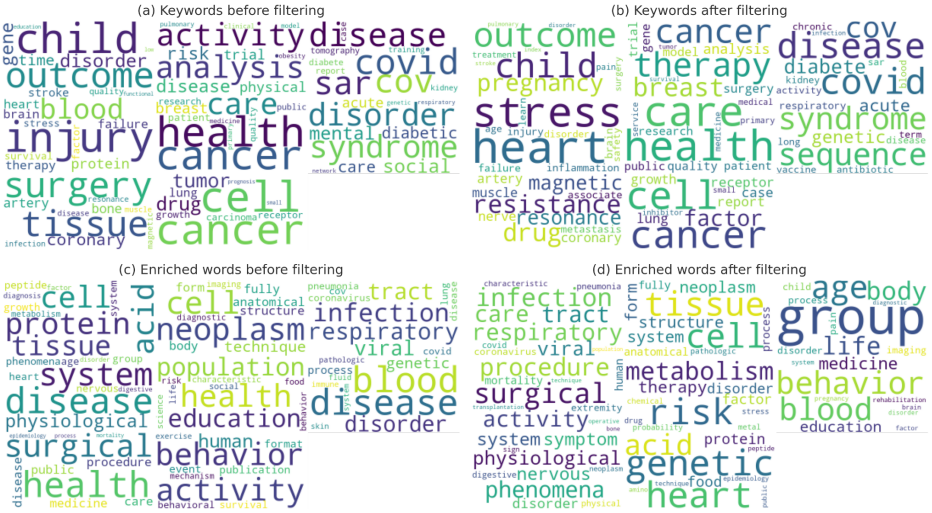

Fig. 14 Montage of LDA topic word clouds across experimental stages: (a) keywords before filtration, (b) keywords after filtration, (c) enriched words before filtration, and (d) enriched words after filtration.

taxonomy of topics is required. CTM achieved intermediate performance, offering a trade-off between quality and computational efficiency. In contrast, DTM scored slightly lower on static coherence but revealed valuable longitudinal insights, such as the evolution of vaccine discourse toward mRNA-related terms. These patterns confirm the utility of temporal modeling in capturing thematic shifts across publication periods. Embedding expansions contributed further improvements. BioWordVec increased C_v by about 0.03 and diversity by 0.01, while BioBERT yielded larger gains ($C_v + 0.05$, NPMI +0.02) at the cost of GPU requirements. The combination of MeSH enrichment with embeddings produced the strongest results, surfacing synonyms such as *SARS-CoV-2* that were absent from MeSH. Nonetheless, unfiltered embeddings occasionally introduced irrelevant terms, emphasizing the importance of semantic filtration. Overall, these results confirm a consistent trend: neural and hierarchical models surpass flat LDA in coherence, while dynamic extensions add explanatory power by capturing temporal evolution. This pattern aligns with recent comparative studies, which report superior performance for transformer-based methods and highlight the role of temporal modeling in biomedical research.

In theory, the performance differences across models arise from their underlying assumptions and architectures. LDA is a generative probabilistic model that treats documents as mixtures of latent topics and topics as distributions over words under a bag-of-words assumption. This formulation ensures efficient inference but ignores word order and vocabulary drift, which results in lower coherence and higher perplexity. Hierarchical LDA extends the generative design by drawing topics from a nested Chinese Restaurant Process, thereby producing a tree of topics that captures parent–child dependencies. This hierarchical structure enhances interpretability and improves coherence relative to LDA, which makes hLDA particularly attractive when the goal is to construct a taxonomy of biomedical themes. CTM combine a Dirichlet prior with contextual embeddings obtained from pre-trained models such as BioBERT. By augmenting the bag-of-words representation with embeddings, CTM bridges semantic richness and interpretability. This explains its mid-level performance: higher than

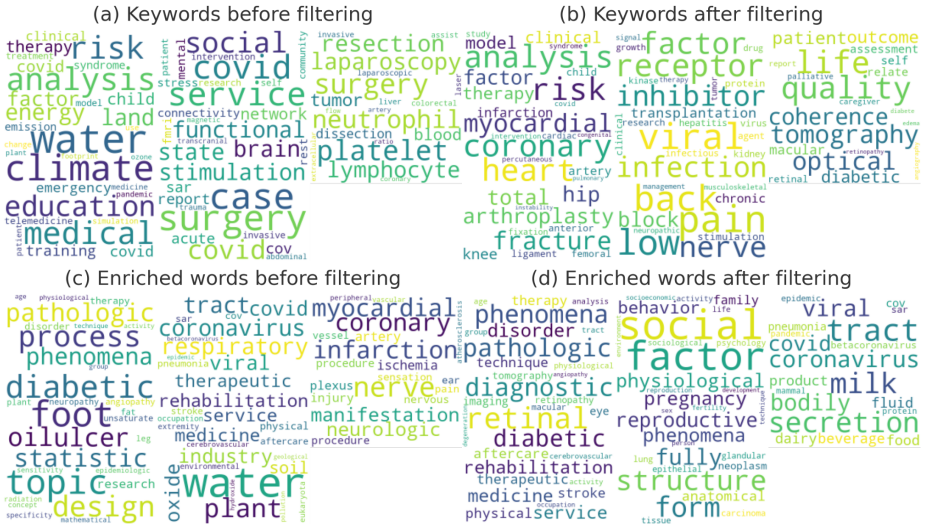

Fig. 15 Montage of hLDA topic word clouds across experimental stages. Enrichment clarifies hierarchical structures and distinguishes clusters more effectively.

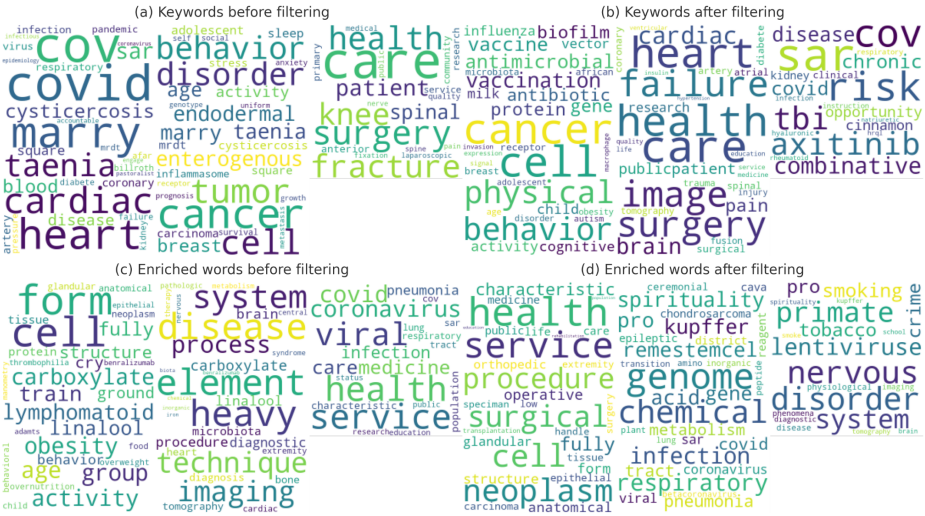

Fig. 16 Montage of CTM topic word clouds across experimental stages. Topics shift from pandemic-related terms to healthcare services and metabolic processes.

LDA due to the added contextual signal, yet lower than transformer-driven clustering models that fully exploit deep contextual semantics.

Neural clustering models such as BERTopic depart from generative assumptions. They embed documents into high-dimensional semantic spaces using transformer encoders, then apply dimensionality reduction with UMAP followed by clustering with HDBSCAN. Topics are subsequently derived through class-based TF–IDF applied to each cluster. This design

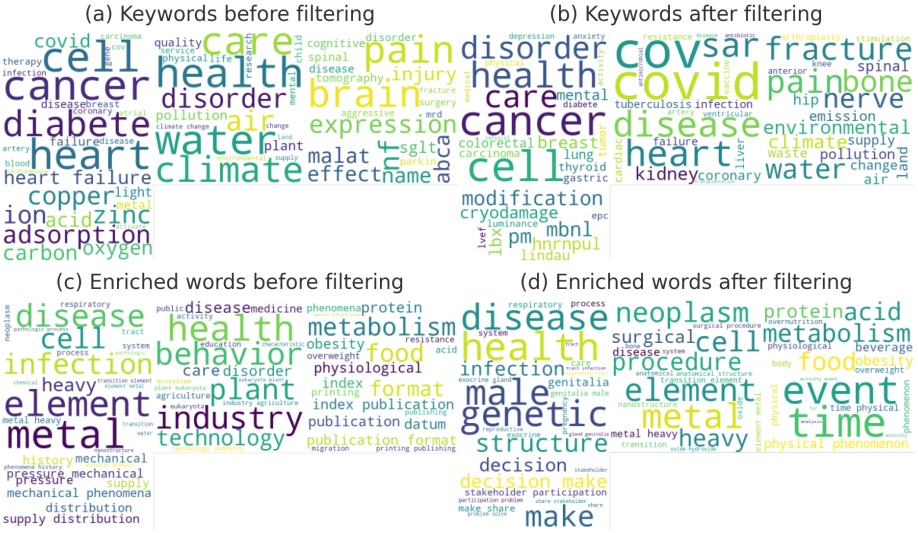

Fig. 17 Montage of BERTopic topic word clouds across experimental stages. Enrichment highlights clinical procedures and disease categories, yielding the most specific and coherent topics.

captures contextual semantics and produces highly coherent topics, which explains the superior performance observed in our experiments. At the same time, it introduces sensitivity to embedding quality, cluster hyperparameters, and initialization choices, while requiring more computational resources than LDA or hLDA.

Dynamic Topic Models (DTM) preserve the generative framework of LDA but allow topic distributions to evolve across sequential time slices using state-space models. This structure enables the identification of thematic trajectories across months or quarters, thereby uncovering longitudinal patterns such as the pandemic-driven surge of telemedicine. Neural extensions, including time-sliced BERTopic, provide a similar capacity for temporal analysis while leveraging contextual embeddings. In our results, temporal models slightly reduced static coherence but proved essential for capturing thematic evolution, which highlights the trade-off between interpretability, semantic richness, and temporal awareness that recent surveys in biomedical topic modeling have emphasized.

Table 18 provides a comprehensive mapping between experimental findings and research questions. The evidence for RQ1 appears in Tables 3 through 6, which document systematic improvements from MeSH enrichment. RQ2 receives support from the filtration sensitivity analysis that examines preprocessing effects across different threshold values. RQ3 draws validation from baseline model comparisons and statistical significance tests. The dynamic topic prevalence plots and temporal coherence measurements address RQ4. Finally, RQ5 finds confirmation through embedding expansion results that demonstrate additional performance gains. The dominant topic analysis provides further validation for the MeSH enrichment approach. This analysis revealed that MeSH enrichment increased the average contribution of the most probable topic from 0.59 for author keywords alone to 0.64 for the enhanced vocabulary. This improvement represents a meaningful increase in topic assignment confidence across the evaluated corpus.

A final observation concerns the impact of filtration and enrichment on topic quality. Before applying data filtration, all models produced lower coherence and higher perplexity because extremely rare or ubiquitous tokens introduced noise. After filtration (min_df =

Table 18 Research questions mapped to empirical evidence and findings.

Research Question	Supporting Evidence	Conclusion
RQ1: Does MeSH enrichment improve topic quality?	Tables 3–6 demonstrate C_v and NPMI increases across all models; statistical tests confirm significance.	Yes – MeSH enrichment reduces perplexity while improving coherence and diversity.
RQ2: How does filtration affect performance?	Sensitivity analysis shows that the min_df and max_df thresholds influence both vocabulary size and coherence. The selected thresholds (10, 0.5) provide an optimal balance, as presented in Section 3.2.3 and demonstrated in Table 4.	Moderate impact – Filtration effects are consistent across threshold ranges.
RQ3: How do models compare?	Tables 6–7 compare four approaches; statistical tests validate neural model superiority.	Advanced models (CTM, BERTopic) and hLDA outperform basic LDA; BERTopic achieves best results.
RQ4: Do dynamic models reveal temporal trends?	DTM and time-sliced analysis capture telemedicine surge and vaccine topic evolution; temporal coherence confirms model stability as detailed in Section 4.2.	Yes – Dynamic analysis detects longitudinal patterns that static models miss.
RQ5: Do embeddings provide additional benefits?	BioWordVec and BioBERT expansion improves coherence and recovers synonyms absent from MeSH vocabularies as detailed in Section 4.3.	Yes – Embeddings enhance performance when combined with MeSH, though require semantic filtration.

10, max_df = 0.5) and enrichment with MeSH terms, coherence increased and perplexity decreased across the board. For example, LDA's C_v improved from 0.37 to 0.52 and NPMI from -0.05 to 0.08, while BERTopic's C_v rose from 0.71 to 0.75. Filtration reduced vocabulary size and emphasized domain-specific terminology, which enabled models to focus on salient patterns. Embedding-based expansions with BioWordVec and BioBERT yielded additional gains (C_v + 0.03–0.05) but also introduced noise when unfiltered, which underscores the need for semantic filtration. The comparative analysis of outcomes before and after filtration highlights the complementary roles of ontology enrichment and embedding expansion as mutually reinforcing techniques.

While numerical summaries provide an overview of model performance, visualizations reveal finer-grained patterns across conditions. BERTopic sits far above the other methods on both C_v and NPMI measures, which confirms that transformer-based clustering produces the most coherent and semantically rich topics. hLDA ranks second, followed by CTM and LDA. MeSH enrichment consistently yields larger gains than filtration alone. Neural and hierarchical models (BERTopic and hLDA) benefit monotonically from both interventions, whereas CTM and especially LDA exhibit mixed behavior. LDA's C_v improves only when enriched words are used and slightly declines after enhanced vocabulary filtration, which underscores its sensitivity to vocabulary size.

The practical implications of this research extend across multiple domains within biomedical informatics and scientific trend analysis. For research administrators and funding agencies, the enhanced topic discovery capabilities enable more accurate identification of emerging research areas and interdisciplinary collaboration opportunities. The temporal analysis features provide quantitative evidence of research response dynamics to external events, such as the pandemic-driven surge in telemedicine and public health preparedness research. For information retrieval and literature review applications, the improved topic coherence and

semantic coverage facilitate more effective document organization and thematic categorization. The integration of embedding-based expansion techniques addresses the vocabulary gap problem that frequently affects domain-specific information systems. The methodology's applicability extends beyond medical literature to other scientific domains with established ontological frameworks, which suggests broader utility for scientific trend analysis and knowledge discovery applications.

Despite the demonstrated effectiveness of the proposed approach, several limitations warrant consideration. The reliance on a single-year dataset (2020) limits generalizability across temporal periods with different thematic characteristics. While the pandemic year provides rich thematic diversity, multi-year validation would strengthen confidence in the methodology's robustness across varying research landscapes. The dependency on MeSH ontology coverage introduces potential biases toward established medical concepts, which potentially underrepresents emerging or interdisciplinary research areas that have not yet been incorporated into the controlled vocabulary. The disambiguation protocol, while systematic, may introduce annotation errors that compound through subsequent analytical stages. Coherence and perplexity metrics, while computationally tractable and widely accepted, provide limited insight into semantic interpretability from human expert perspectives. The absence of domain expert validation represents a significant limitation that future research should address through structured evaluation protocols. The computational requirements associated with embedding-based expansion techniques may limit practical applicability in resource-constrained environments, which requires careful consideration of performance-cost trade-offs in operational deployments.

Several promising research directions emerge from this work that warrant future investigation. The integration of multiple ontological frameworks beyond MeSH could address vocabulary coverage limitations and provide more comprehensive semantic enrichment. Cross-ontology mapping and alignment techniques could enable leveraging complementary knowledge sources while maintaining consistency in concept representation. The development of hybrid human-machine evaluation protocols could bridge the gap between computational metrics and expert assessment of topic quality and interpretability. Interactive topic modeling approaches that incorporate expert feedback could enhance both analytical outcomes and domain specialist engagement with automated analytical tools. Extension to multi-modal document representations incorporating abstract content, citation networks, and author collaboration patterns could provide richer contextual information for topic discovery and trend analysis. The integration of causal inference techniques could enable identification of research influence patterns and knowledge diffusion mechanisms across disciplinary boundaries. The methodology's adaptation to real-time analytical scenarios presents both technical challenges and significant opportunities for dynamic research monitoring and early trend detection capabilities.

5.1 Limitations

Though this study achieved its aim, it has several limitations. The dataset is limited to a single year (2020). While this period contains a high volume of pandemic-related research, it does not represent other years or domains, and thus limits the generalisation of findings. Sensitivity analysis on the 2020 dataset shows consistent results, but a multi-year dataset would capture more robust longitudinal trends. The use of MeSH provides broad coverage but is not exhaustive. New or emerging concepts, such as novel variants, may not appear in MeSH at the time of analysis, and manual inspection is still necessary to verify annotations.

Coherence and perplexity are useful but imperfect measures of interpretability. They do not always reflect human understanding of topics, and expert evaluation could add more insight.

Finally, while the disambiguation protocol reduces ambiguity in keyword-to-MeSH mapping, occasional errors remain. Using abstract-level context or transformer-based entity linking could improve mapping accuracy in future work.

Despite these limitations, this study provides a solid foundation for biomedical trend analysis. Future research could extend the approach to multi-year datasets, integrate additional ontologies for improved coverage, and incorporate human-in-the-loop evaluations to enhance interpretability and robustness.

6 Conclusion

This paper presented the ETM-F framework for biomedical literature analysis. The framework introduces three main advances: integration of MeSH ontology descriptors with author keywords, application of frequency-based filtration to refine vocabularies, and expansion with embeddings to improve semantic coverage. ETM-F was evaluated under three configurations: keywords alone, keywords enriched with MeSH terms, and enriched keywords combined with embeddings. The results showed that MeSH enrichment increased coherence (C_v and NPMI) and reduced perplexity while preserving topic interpretability. Embedding expansion further uncovered synonyms and domain-specific terms that MeSH did not capture. Across models, BERTopic achieved the highest performance, while ETM-F improved the stability and clarity of topics in all settings.

The study also confirmed the importance of temporal analysis. Dynamic Topic Models (DTM) and time-sliced BERTopic revealed clear shifts in research themes, such as the rise of telemedicine and the transition from general vaccine terms to mRNA-specific terms. These results show that ontology-enriched approaches provide a more precise and reliable view of scientific developments than methods based only on keywords. Overall, ETM-F establishes a practical and interpretable framework for large-scale biomedical trend analysis. Future work will extend this framework with additional ontologies and multi-modal resources, and adapt it to decision-support systems and educational applications.

Acknowledgments

This work is supported and funded by Brunel University of London, London, United Kingdom.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

Competing interests

The authors have no relevant financial or non-financial interests to disclose

References

[1] Hanson, M.A., Barreiro, P.G., Crosetto, P., Brockington, D.: The strain on scientific publishing. Quantitative Science Studies **5**(4), 823–843 (2024)

[2] Shahrzadi, L., Mansouri, A., Alavi, M., Shabani, A.: Causes, consequences, and strategies to deal with information overload: A scoping review. International Journal of Information Management Data Insights **4**(2), 100261 (2024)

[3] Dridi, A., Gaber, M.M., Azad, R.M.A., Bhogal, J.: Scholarly data mining: A systematic review of its applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **11**(2), 1395 (2021)

[4] Cobelli, N., Blasi, S.: Combining topic modeling and bibliometric analysis to understand the evolution of technological innovation adoption in the healthcare industry. European Journal of Innovation Management **27**(9), 127–149 (2024)

[5] Ogunleye, B., Lancho Barrantes, B.S., Zakariyyah, K.I.: Topic modelling through the bibliometrics lens and its technique. Artificial Intelligence Review **58**(3), 74 (2025)

[6] Jagannathan, M., Roy, D., Delhi, V.S.K.: Application of nlp-based topic modeling to analyse unstructured text data in annual reports of construction contracting companies. Csi Transactions on Ict **10**(2), 97–106 (2022)

[7] Muthusami, R., Mani Kandan, N., Saritha, K., Narenthiran, B., Nagaprasad, N., Ramaswamy, K.: Investigating topic modeling techniques through evaluation of topics discovered in short texts data across diverse domains. Scientific Reports **14**(1), 12003 (2024)

[8] Athukorala, S., Mohotti, W.: An effective short-text topic modelling with neighbourhood assistance-driven nmf in twitter. Social network analysis and mining **12**(1), 89 (2022)

[9] Srinivasa-Desikan, B.: Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd, Birmingham, UK (2018)

[10] Bogdanowicz, A., Guan, C.: Dynamic topic modeling of twitter data during the covid-19 pandemic. Plos one **17**(5), 0268669 (2022)

[11] Liu, Z., Lin, Y., Sun, M.: Representation Learning for Natural Language Processing. Springer, Singapore (2023)

[12] Sleeman, J., Finin, T., Halem, M.: Understanding cybersecurity threat trends through dynamic topic modeling. Frontiers in big Data **4**, 601529 (2021)

[13] Rao, V.K., Valdez, D., Muralidharan, R., Agley, J., Eddens, K.S., Dendukuri,

A., Panth, V., Parker, M.A.: Digital epidemiology of prescription drug references on x (formerly twitter): Neural network topic modeling and sentiment analysis. Journal of Medical Internet Research **26**, 57885 (2024)

[14] Ma, L., Chen, R., Ge, W., Rogers, P., Lyn-Cook, B., Hong, H., Tong, W., Wu, N., Zou, W.: Ai-powered topic modeling: comparing lda and bertopic in analyzing opioid-related cardiovascular risks in women. Experimental Biology and Medicine **250**, 10389 (2025)

[15] Kumari, P., Kadian, S., Vora, M., Vemuri, P., Kumar, S., Verma, O.P., Narayan, R.J.: Recent advancements in topic modeling techniques for healthcare, bioinformatics, and other potential applications. Advanced Intelligent Systems, 2400528 (2025)

[16] Legg, C.: Ontologies on the semantic web (2007)

[17] Bean, C.A., Green, R.: Relevance relationships. In: Relationships in the Organization of Knowledge, pp. 115–132. Springer, ??? (2001)

[18] Babalola, O., Ojokoh, B., Boyinbode, O.: Comprehensive evaluation of lda, nmf, and bertopic's performance on news headline topic modeling. Journal of Computing Theories and Applications **2**(2), 268–289 (2024)

[19] Aamir, N., Raza, A., Iqbal, M.W., Hamid, K., Nazir, Z., Asif, A., Hussain, S., Muhammad, H.A.B.: Topic modeling empowered by a deep learning framework integrating bertopic, xlm-r, and gpt. Journal of Computing & Biomedical Informatics **8**(02) (2025)

[20] Kumar, S., Kumar, N., Dev, A., Naorem, S.: Movie genre classification using binary relevance, label powerset, and machine learning classifiers. Multimedia Tools and Applications **82**(1), 945–968 (2023)

[21] Zhang, C., Guan, J.: How to identify metaknowledge trends and features in a certain research field? evidences from innovation and entrepreneurial ecosystem. Scientometrics **113**, 1177–1197 (2017)

[22] Silva, M.C., Eugénio, P., Faria, D., Pesquita, C.: Ontologies and knowledge graphs in oncology research. Cancers **14**(8), 1906 (2022)

[23] Ivanović, M., Budimac, Z.: An overview of ontologies and data resources in medical domains. Expert Systems with Applications **41**(11), 5158–5166 (2014)

[24] Baumann, N.: How to use the medical subject headings (mesh). International journal of clinical practice **70**(2), 171–174 (2016)

[25] Miles, W.D.: A History of the National Library of Medicine: the Nation's Treasury of Medical Knowledge vol. 82. US Department of Health and Human Services,

Public Health Service, National, Washington, DC (1982)

[26] Cohen, T., Widdows, D.: Geometric representations in biomedical informatics: Applications in automated text analysis. Methods in Biomedical Informatics: A Pragmatic Approach, 99–139 (2013)

[27] Santa Soriano, A., Lorenzo Álvarez, C., Torres Valdés, R.M.: Bibliometric analysis to identify an emerging research area: Public relations intelligence—a challenge to strengthen technological observatories in the network society. Scientometrics **115**(3), 1591–1614 (2018)

[28] Dridi, A., Gaber, M.M., Azad, R.M.A., Bhogal, J.: Leap2trend: A temporal word embedding approach for instant detection of emerging scientific trends. IEEE Access **7**, 176414–176428 (2019)

[29] Whetzel, P.L., Team, N.: Ncbo technology: Powering semantically aware applications. Journal of biomedical semantics **4**(Suppl 1), 8 (2013)

[30] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)

[31] Zhou, S., Kan, P., Huang, Q., Silbernagel, J.: A guided latent dirichlet allocation approach to investigate real-time latent topics of twitter data during hurricane laura. Journal of Information Science **49**(2), 465–479 (2023)

[32] Nutakki, G.C., Nasraoui, O., Abdollahi, B., Badami, M., Sun, W.: Distributed lda-based topic modeling and topic agglomeration in a latent space. In: SNOW-DC@ WWW, pp. 17–24 (2014)

[33] Hwang, M.-H., Ha, S., In, M., Lee, K.: A method of trend analysis using latent dirichlet allocation. International Journal of Control and Automation **11**(5), 173–182 (2018)

[34] Gurcan, F., Cagiltay, N.E.: Big data software engineering: Analysis of knowledge domains and skill sets using lda-based topic modeling. IEEE access **7**, 82541–82552 (2019)

[35] Yu, Z., Bernstam, E., Cohen, T., Wallace, B.C., Johnson, T.R.: Improving the utility of mesh® terms using the topicalmesh representation. Journal of biomedical informatics **61**, 77–86 (2016)

[36] Valderrama-Zurián, J.C., García-Zorita, C., Marugán-Lázaro, S., Sanz-Casado, E.: Comparison of mesh terms and keywords plus terms for more accurate classification in medical research fields. a case study in cannabis research. Information Processing & Management **58**(5), 102658 (2021)

[37] Yu, Y., Dinh, D.-T., Nguyen, B.-H., Yu, F., Huynh, V.-N.: Mining insights from

esports game reviews with an aspect-based sentiment analysis framework. Ieee Access **11**, 61161–61172 (2023)

[38] Madrid-Garcia, A., Freites-Nunez, D., Merino-Barbancho, B., Pérez Sancristobal, I., Rodríguez-Rodríguez, L.: Mapping two decades of research in rheumatology-specific journals: a topic modeling analysis with bertopic. Therapeutic Advances in Musculoskeletal Disease **16**, 1759720–241308037 (2024)

[39] Lezhnina, O.: Depression, anxiety, and burnout in academia: topic modeling of pubmed abstracts. Frontiers in Research Metrics and Analytics **8**, 1271385 (2023)

[40] Kumar, A., Karamchandani, A., Singh, S.: Topic modeling of neuropsychiatric diseases related to gut microbiota and gut brain axis using artificial intelligence based bertopic model on pubmed abstracts. Neuroscience Informatics **4**(4), 100175 (2024)

[41] Zhang, R., Duan, Z., Lee, C., Riffle, D., Min, M.R., Zhang, J.: Turtling: a time-aware neural topic model on nih grant data. Bioinformatics Advances **3**(1), 096 (2023)

[42] Wu, S.-W., Li, C.-C., Chien, T.-N., Chu, C.-M.: Integrating structured and unstructured data with bertopic and machine learning: A comprehensive predictive model for mortality in icu heart failure patients. Applied Sciences (2076-3417) **14**(17) (2024)

[43] Uncovska, M., Freitag, B., Meister, S., Fehring, L.: Rating analysis and bertopic modeling of consumer versus regulated mhealth app reviews in germany. NPJ Digital Medicine **6**(1), 115 (2023)

[44] Chung, M.-K., Lee, S.Y., Shin, T., Park, J.Y., Hwang, S., Kim, M.-H., Lee, J., Lee, K.-J., Lim, H.-S., Urtnasan, E., *et al.*: Bert and bertopic for screening clinical depression on open-ended text messages collected through a mobile application from older adults. BMC Public Health **25**(1), 2161 (2025)

[45] Xin, Y., Grabowska, M.E., Gangireddy, S., Krantz, M.S., Kerchberger, V.E., Dickson, A.L., Feng, Q., Yin, Z., Wei, W.-Q.: Improving topic modeling performance on social media through semantic relationships within biomedical terminology. PloS one **20**(2), 0318702 (2025)

[46] Guillén-Pacho, I., Badenes-Olmedo, C., Corcho, O.: Dynamic topic modelling for exploring the scientific literature on coronavirus: an unsupervised labelling technique. International Journal of Data Science and Analytics, 1–31 (2024)

[47] Kawai, M., Sato, H., Shiohama, T.: Topic model-based recommender systems and their applications to cold-start problems. Expert Systems with Applications **202**, 117129 (2022)

[48] Ozyurt, O., Özköse, H., Ayaz, A.: Evaluating the latest trends of industry 4.0 based on lda topic model. The Journal of Supercomputing **80**(13), 19003–19030 (2024)

[49] Egger, R., Yu, J.: A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Frontiers in sociology **7**, 886498 (2022)

[50] Nanyonga, A., Joiner, K., Turhan, U., Wild, G.: Does the choice of topic modeling technique impact the interpretation of aviation incident reports? a methodological assessment. Technologies **13**(5), 209 (2025)

[51] Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., Hassan, A.: Topic modeling algorithms and applications: A survey. Information Systems **112**, 102131 (2023)

[52] Mujahid, M., Lee, E., Rustam, F., Washington, P.B., Ullah, S., Reshi, A.A., Ashraf, I.: Sentiment analysis and topic modeling on tweets about online education during covid-19. Applied Sciences **11**(18), 8438 (2021)

[53] Chauhan, P., Sharma, N., Sikka, G.: On the importance of pre-processing in small-scale analyses of twitter: a case study of the 2019 indian general election. Multimedia Tools and Applications, 1–40 (2023)

[54] Drissi, A., Tissaoui, A., Sassi, S., Chbeir, R., Jemai, A.: S-lda: Documents classification enrichment for information retrieval. In: International Conference on Computational Collective Intelligence, pp. 687–699 (2022). Springer

[55] Romero, J.D., Feijoo-Garcia, M.A., Nanda, G., Newell, B., Magana, A.J.: Evaluating the performance of topic modeling techniques with human validation to support qualitative analysis. Big Data and Cognitive Computing **8**(10), 132 (2024)

[56] Carvallo, A., Parra, D., Lobel, H., Soto, A.: Automatic document screening of medical literature using word and text embeddings in an active learning setting. Scientometrics **125**(3), 3047–3084 (2020)

[57] Pastor-Galindo, J., Sandlin, H.-Â., Mármol, F.G., Bovet, G., Pérez, G.M.: A big data architecture for early identification and categorization of dark web sites. Future Generation Computer Systems **157**, 67–81 (2024)

[58] Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences **101**(suppl_1), 5228–5235 (2004)

[59] Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1553374.1553515 . https://doi.org/10.1145/1553374.1553515

[60] Blair, S.J., Bi, Y., Mulvenna, M.D.: Aggregated topic models for increasing social media topic coherence. Applied Intelligence **50**, 138–156 (2020)

[61] Prabhakaran, S.: Topic modeling with gensim (python). Machine learning plus (2018)

[62] Newman, D., Noh, Y., Talley, E., Karimi, S., Baldwin, T.: Evaluating topic models for digital libraries. Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1816123.1816156 . https://doi.org/10.1145/1816123.1816156

[63] Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. EMNLP '11, pp. 262–272. Association for Computational Linguistics, USA (2011)

[64] Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D.: Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 952–961 (2012)