This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/ICCIAA68481.2026.11544024, 2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIAA)

Ontology-Enriched Abstract Topic Modelling for Biomedical Trend Discovery: A MeSH-based LDA Framework with BI Visual Analytics

Ahmad Altarawneh
Dept. of Computer Science
Brunel University of London
London, UK
ahmadturkifaiq.altarawneh@brunel.ac.uk

Stephen Swift
Dept. of Computer Science
Brunel University of London
London, UK
stephen.swift@brunel.ac.uk

Allan Tucker
Dept. of Computer Science
Brunel University of London
London, UK
allan.tucker@brunel.ac.uk

Mahir Arzoky
Dept. of Computer Science
Brunel University of London
London, UK
mahir.arzoky@brunel.ac.uk

***Abstract*—The explosive growth of biomedical literature increasingly makes it difficult to identify incipient topics and long-term developments. Citation-based indicators are slow to capture developments and often fail to represent influence. Whilst Latent Dirichlet Allocation (LDA) supports content-based trend discovery, its bag-of-words assumption can overlook synonymy, polysemy, and hierarchical relations that are central to biomedical terminology. Moreover, recent work provides limited direct evidence on how far MeSH-enriched abstract representations improve LDA-based trend analysis when compared against the original abstract text. To fill this gap, an ontology-driven methodology for the pre-topic modelling of article abstracts by enriching them with hierarchical Medical Subject Headings (MeSH) terms is introduced in this study. We evaluate four representations, namely original abstracts, MeSH parent categories, MeSH-matched terms with parents, and a fully enriched representation, using a decade-long Scopus corpus of 624,486 biomedical journal articles (2015–2024). Model performance is assessed using coherence, distinctiveness, perplexity, and stability. The fully enriched representation achieves the strongest overall results (coherence 0.606, distinctiveness 0.996, stability 0.725, perplexity 1782.040) relative to the original abstract baseline. Temporal topic analysis further reveals interpretable patterns, including the COVID-19 surge and longer-term structural shifts such as the rise of machine-learning diagnostics. These findings indicate that MeSH-guided abstract enrichment can strengthen topic quality and support more effective biomedical trend discovery through business-intelligence visual analytics.**

***Index Terms*—Biomedical research trends, MeSH enrichment, Abstract-based analysis, Topic modelling, Latent Dirichlet Allocation, Semantic ontology, Business intelligence, Trend detection.**

I. Introduction

The biomedical sciences are experiencing a deluge of publications; millions of articles are published every year across thousands of different journals, and they overwhelm individual researchers and make it hard to recognise emerging topics, as Web of Science indexes more than three million articles in 2023 alone [1]. By 2025, PubMed contains over 38 million indexed and up-to-date articles and adds over one million records annually [2]. Health practitioners, along with researchers, struggle to extract valuable information from these large published research databases due to their extensive size and massive data quantity [3].

Conventional bibliometric metrics, including citation numbers, h-indexes and impact factors, provide a backward-looking view of academic impact; yet, they are ill-suited to identify trends in time since they have a time lag and hidden citation phenomena [4]. Conversely, content-based methods often have a lack of semantic depth, despite the direct analysis of textual data using natural language processing (NLP) to extract additional content [5], [6]. Besides, techniques such as LDA are regularly used to derive topical structures of datasets [7]. The ambiguous connotation of keywords, the domain-specific vocabulary, and the abundance of synonyms usually hinder the issue of determining the real thematic framework of the literature, disrupting a comprehensive understanding of existing trends in research by the LDA technique [8], [6]. For example, the separate articles that discuss "heart attack" and "myocardial infarction" lack a direct connection through text analysis unless we understand these terms as equivalent expressions of each other and referring to the same concept [8].

A further limitation seldom discussed is the keyword gap. Author keywords offer a succinct glimpse of a paper's themes but are usually limited to three to five generic terms and often suffer from inconsistent terminology [9]. As a result, keyword-only analyses may overlook nuanced or emerging concepts. By contrast, the full abstract summarises objectives, methods, results and conclusions in several sentences, offering a much richer lexical context. Topic modelling on complete abstracts, therefore, uncovers subtler topics that might be invisible in keyword lists.

Recent studies suggest that combining topic modelling with structured ontologies can provide deeper semantic insights [10], [6]. Innovative approaches integrate LDA with ontologies such as MeSH, but most either post-process LDA topics with MeSH labels or use MeSH terms as features. Enriching biomedical research texts with MeSH terms facilitates uncovering research trends [11] because the ontology offers a structured interpretation of medical concepts, enabling researchers to identify hidden patterns [12].

By embedding MeSH terms directly within document abstracts, researchers would benefit from both semantic connection and improved domain-specific analysis capabilities and speedier emerging research area recognition [10]. Research on the performance difference between original abstract text and MeSH-enriched text in LDA topic modelling applications is still

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/ICCIAA68481.2026.11544024, 2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIAA)

limited, and no recent studies have directly investigated this approach to the best of our knowledge. In this study, we aim to investigate the use of MeSH ontology term annotations to enrich the vocabulary of research paper abstracts and assess whether this enriched vocabulary can improve the discovery of research trends in the biomedical field using the LDA algorithm. This study makes three main contributions:

- We propose an ontology-enriched abstract representation that inserts MeSH parent categories inline within each abstract, preserving narrative flow while embedding hierarchical knowledge.
- Evaluate and compare the performance of the LDA algorithm on enriched versus non-enriched vocabularies through comprehensive experimental analysis.
- We provide a suite of business-intelligence visualisations to aid decision-making in trend analysis. This is accomplished using Microsoft Power BI that enable interactive exploration of topic structures and temporal dynamics.

Our research question: How can the enrichment of research paper abstracts with MeSH ontology term annotations improve the effectiveness of trend analysis in the medical field using LDA and visualisations? This overarching question encompasses the following aspects:

- **RQ1:** The process of enriching abstracts with MeSH terms.
- **RQ2:** The effectiveness of using these enriched abstracts in trend analysis.
- **RQ3:** The use of LDA for topic discovery on enriched and non-enriched abstracts.
- **RQ4:** The application of visualisations to aid decision-making in trend analysis.

The paper is organised as follows: Section 2 reviews NLP, Machine learning (ML), and MeSH ontology. Section 3 covers dataset, preprocessing, enrichment, and topic modelling. Section 4 presents results and analysis. Section 5 discusses findings and BI contributions. Section 6 concludes with limitations and future work.

II. LITERATURE REVIEW

A. Natural Language Processing (NLP) in Biomedical Research

NLP stands as a subfield of artificial intelligence that analyses human language text while deriving its underlying meaning from both speech and written text [13]. NLP techniques make it feasible to convert this unstructured text into structured knowledge [14]. The biomedical field recognises NLP technology as essential because this system creates methods to examine complete document collections which human researchers cannot handle [15].

Research on NLP applications demonstrates practical use in both cancer hallmark extraction [16], and identifying novel concepts during the pandemic [17]. In the context of the COVID-19 pandemic, this pandemic generated an astonishingly large number of scientific papers, and text-mining systems with NLP techniques helped researchers and clinicians understand them via named entity identification and abstract condensation methods [13]. Similarly, research insights using NLP methods established multiple thematic clusters that exist within antibiotic-resistant bacteria publications while demonstrating that new treatment approaches have begun gaining popularity recently [18]. Scientists normalise scientific text which comes from original papers as a prerequisite step before processing begins [14]. Text processing requires splitting words (tokenisation), then converting everything to lowercase, followed by the removal of punctuation marks along with stop words, and the reduction of words to their basic forms using lemmatisation ("patients" becomes "patient") [15].

Several common NLP approaches have been tailored for biomedical text mining. The primary characteristic of biomedical NLP includes extensive reliance on controlled vocabularies together with ontologies, including Medical Subject Headings (MeSH) for articles [19], [6].

B. Machine Learning (ML) and Topic Modelling

ML techniques permeate biomedical research, powering diagnostic imaging, genomics, drug discovery and public health surveillance [20]. Several alternative topic modelling techniques have been explored in the literature, including Latent Semantic Analysis (LSA) and probabilistic latent semantic analysis (pLSA), which laid the groundwork for uncovering word co-occurrence patterns but suffered from overfitting and limited generalisation [21].

The introduction of Latent Dirichlet Allocation (LDA) addressed these shortcomings by imposing Dirichlet priors on document–topic and topic–word distributions, thereby regularising the model and enabling prediction on unseen documents [7], [22]. While neural topic models, such as embedded topic models (ETM), contextualised topic models (CTM) and BERTopic, leverage transformer embeddings for improved coherence [23], [24], they demand substantial computational resources and often obscure the contribution of individual words [25]. LDA thus remains a robust baseline for real-time trend analysis: it scales to millions of documents, yields interpretable mixtures of topics and can be seamlessly integrated with domain ontologies [23]. Fig. 1 illustrates the generative process of LDA: Each document is modelled as a mixture of latent topics; each topic is a distribution over words, and Dirichlet priors govern the allocations [7].

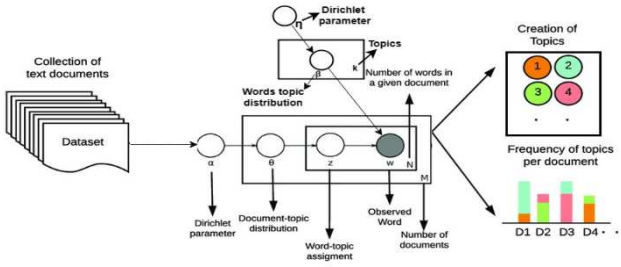

Fig. 1. LDA Model [7].

Among the corpus documents, LDA assumes they consist of latent topics (θ), while each topic contains a word distribution (w). M refers to the total number of documents. The Dirichlet

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/ICCIAA68481.2026.11544024, 2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIAA)

prior parameter (α) is the Dirichlet prior that controls the number of topics in each document, while z identifies topic assignments, and N represents total document words. The probability distribution of the top words for a particular given topic K is denoted by the symbol βk. The topic hyper-parameter is represented by Eta (η). Different topics are denoted by K. The right side of this schematic shows how topics are distributed among example documents [7], [22].

C. Medical Subject Headings (MeSH) Ontology and Semantic Enrichment

Ontology provides structured vocabularies and hierarchical relationships that codify domain knowledge [10]. In medicine, the MeSH thesaurus, developed by the U.S. National Library of Medicine, serves as the cornerstone for indexing MEDLINE and PubMed records and is updated annually to reflect evolving terminology [26]. It organises more than 30 000 descriptors into a multi-level tree: broad headings at the top (e.g., Viral Diseases) branch into progressively more specific concepts (e.g., COVID-19) [27]. Each descriptor is accompanied by a unique identifier, a definition and a set of entry terms or synonyms [28]. This hierarchical structure supports parent–child relationships, enabling semantic generalisation and finer-grained retrieval, and makes MeSH invaluable for unifying variant expressions across the biomedical literature [27].

Annotating text with MeSH descriptors therefore enhances semantic interpretation, resolves synonyms and facilitates consistent indexing [6]. For example, the colloquial phrase "heart attack" or "cardiac infarction" can be normalised to the MeSH descriptor Myocardial Infarction, ensuring that literature using different wording is aggregated under a common concept [8]. Though MeSH was shown to work well for summarising biomedical articles, this approach alone cannot identify hot topics and recent research [29], [6]. Consequently, statistics-based NLP analysis, known as topic modelling through LDA, gains a clear standard classification by associating {"blood", "pressure", "hypertension", "risk"} with the term "hypertension" in the MeSH system [30].

III. PROPOSED FRAMEWORK AND METHODOLOGY

A. Dataset Collection and Scope

Fig. 2 visualises the proposed workflow, which begins with the Scopus dataset, applies MeSH annotation, performs text preprocessing, generates bags of words for both original, and MeSH-enriched abstracts. Topic detection via LDA is followed by trend discovery, visualisation and evaluation.

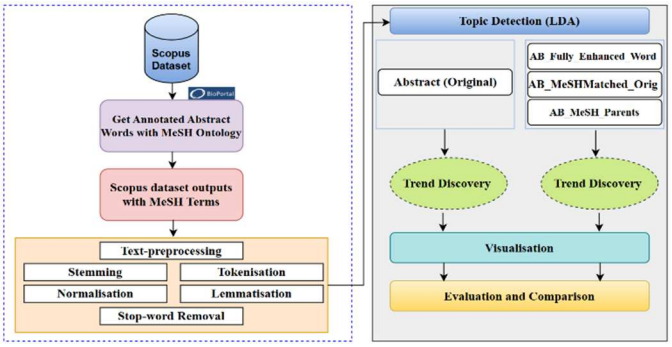

Fig. 2. MeSH-Enriched Topic Modelling Framework.

We collected metadata and abstracts for 624,486 biomedical articles from Scopus via Elsevier's API, covering the period from 2015 to 2024. To ensure broad coverage, we queried Scopus for journal articles whose subject areas included medicine, health sciences and life sciences. For each article, we collected the following fields: title, abstract, Yr (publication year), affiliation, aggregation type (e.g., journal) and subtype description (e.g., article, review).

In this study, the terms AB_Fully_Enhanced_Word and Fully Enhanced, as used in figures and tables, refer to the same representation denoted in the text as fully enriched abstract terms.

B. MeSH-Based Abstract Enrichment

After collecting the dataset from Scopus, we augmented the abstracts using the MeSH ontology. The enrichment pipeline proceeded as follows. First, we matched tokens in the abstract to MeSH descriptors via annotation service. The matched descriptors were appended to the original abstract to form the MeSH matched terms field (matched terms with parent categories appended). Next, we traversed the MeSH tree upwards to extract the immediate parent terms of each matched descriptor; these parent terms were added to create the MeSH parents field (only parent categories). Finally, we combined the original abstract, the matched descriptors and their parents to produce the fully enriched terms field (original abstract words captured in MeSH and their corresponding parent term, as well as any additional abstract words not captured in MeSH). For comparison purposes, we retained the unmodified abstract (original). TABLE I summarises the four representations with a simple example, illustrating how MeSH descriptors and their parents enrich the narrative. Where the notation $((\rightarrow))$ means that the term on the left is categorised under the MeSH term on the right.

TABLE I.
FEATURES ADDED AFTER MESH ONTOLOGY ENRICHMENT

Representation & Description	Example
***Abstract* (*Original*)** This field contains the original abstract words provided by the author.	The lockdown due to the COVID-19 pandemic imposed in many countries led to social isolation and the interruption of activities that were useful in stimulating cognition, etc.
AB_MeSH_Parents: This field includes only parent terms, with each represented by a unique MeSH identifier or prefLabel.	Coronavirus Infections, Pneumonia, Viral, Social Behaviour, Sociological Factors, Mental Processes.
AB_MeSHMatched_Orig: This field contains the original abstract words captured in MeSH, its corresponding parent term.	COVID-19 pandemic →Coronavirus Infections, Pneumonia, Viral; social isolation →Social Behaviour, Sociological Factors; cognition →Mental Processes.
AB_Fully_Enriched_Word: This field contains the original abstract words captured in MeSH, its corresponding parent term,	The lockdown due to the COVID-19 pandemic (Coronavirus Infections, Pneumonia, Viral), imposed in many countries led to social isolation (Social Behaviour, Sociological Factors) and the interruption of activities that were useful in

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/ICCIAA68481.2026.11544024, 2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIAA)

Representation & Description	Example
as well as any additional abstract words not captured in MeSH.	stimulating cognition (Mental Processes), etc.

C. Pre-processing Pipeline

To ensure that the representation of all representations is comparable, all the four keyword representations were processed by a unified pre-processing pipeline. The procedures encompassed:

- **Temporal filtering and deduplication:** Only records published from 2015 to 2024 were retained. Duplicate entries were removed by de-duplicating on the unique article identifier (EID).
- **Language filtering:** Retained only English abstracts to ensure consistency in lexical processing.
- **Text cleaning:** All textual fields were lowercased; brackets, punctuation marks, numeric characters and missing values were removed; whitespace was collapsed; and non-alphabetic tokens were discarded. This step produced clean token sequences suitable for NLP.
- **Tokenisation:** MeSH fields are broken down based on delimiters (i.e., comma, semicolon, pipe, etc.). Multiword descriptors are joined by underscores (e.g., "drug resistance" →"drug_resistance".
- **Stop-word removal and lexical filtering:** Common english stop words and a domain-specific stop list (e.g., "methods", "results", "study") were removed. A stricter lexical filtering was applied to retain primarily content-bearing terms.
- **Lemmatisation and stemming:** Reduced inflected forms to their lemmas or stems to conflate morphological variants.
- **Vocabulary control and weighting:** A CountVectoriser was used with a minimum document frequency threshold (min_df = 10), a maximum document frequency ratio (max_df = 0.40) and a cap of 1500 on term frequency.

D. Topic Modelling and Evaluation with LDA

We trained separate LDA models on each of the four representations. For each representation, we performed a grid search over topic numbers K and trained the models using collapsed Gibbs sampling implemented in the Gensim library. The Dirichlet hyperparameters α (document–topic prior) and β (topic–word prior) were set to symmetric values of $50/K$. We evaluated model performance using the metrics described below and selected the K that maximised coherence while minimising perplexity and maximising distinctiveness. For each representation, the model was repeatedly run ten times with different random seeds to assess stability.

1) Topic Coherence (C_V)

Coherence measures the semantic relatedness of the top words within a topic [22]. We adopted the C_V coherence metric, which combines a sliding window, word segmentation, one-gram NPMI and cosine similarity. For a topic with top-T words $W = \{w_1, w_2, \dots, w_T\}$, the NPMI between two words w_i and w_j is defined as:

$$C_{\mathrm{NPMI}}(w_i, w_j) = \frac{\log \frac{p(w_i, w_j)+\varepsilon}{p(w_i)p(w_j)}}{-\log(p(w_i, w_j)+\varepsilon)}, \quad (1)$$

Where

- $p(w_i, w_j)$ is the probability that the two words co-occur within a sliding window,
- $p(w_i)$ and $p(w_j)$ are their marginal probabilities,
- and ε is a small constant for numerical stability.

The C_V score is then computed as the average cosine similarity between the vectors of NPMI scores for each pair of words. In our experiments we set $T = 10$ and estimated probabilities using the training corpus.

2) Perplexity

Perplexity measures how well a probabilistic model predicts unseen data [22]. We computed log likelihoods on a 20% held out set and exponentiated the average negative log likelihood to obtain perplexity. Lower perplexity indicates better generalisation [7]. The perplexity is defined as:

$$\mathrm{Perplexity}(D_{\mathrm{test}}) = \exp\left(-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right) \quad (2)$$

Where

- w_d: sequence of words in document d.
- N_d: number of words in document d.
- $p(w_d)$: probability of w_d under the LDA model.

3) Jaccard similarity & Distinctiveness

For each pair of topics i and j, we consider their sets of top words T_i and T_j, and compute the Jaccard similarity [31]:

$$J(T_i, T_j) = \frac{|T_i \cap T_j|}{|T_i \cup T_j|} \quad (3)$$

which measures the proportion of shared words. We then define distinctiveness as one minus the average Jaccard similarity across all $\binom{K}{2}$ pairs of topics [31]:

$$\mathrm{D} = 1 - \frac{1}{\binom{K}{2}} \sum_{i<j} J(T_i, T_j). \quad (4)$$

Where

- T_i: set of the top-T words of topic i.
- $|\cdot|$: set cardinality.
- K: number of topics.

Higher distinctiveness indicates less overlap between topics.

To assess the reproducibility of topics across runs, we trained each representation ten times with different seeds. For

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/ICCIAA68481.2026.11544024, 2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIAA)

each pair of runs, we matched topics by maximum overlap and computed the average Jaccard similarity of their top T word sets. Stability was computed as the mean of these similarities across all run pairs.

4) Bootstrap Significance Tests

To compare models statistically, we performed paired bootstrap tests on the coherence scores [32]. For each pair of representations we resampled documents with replacement 1000 times, trained the corresponding LDA models and computed the difference in coherence. We then derived 95% confidence intervals and p values for the mean difference using the bootstrap distribution.

IV. RESULTS AND ANALYSIS

A. Optimal Topic Selection

The best choice of topics requires finding a balance between interpretability (coherence and distinctiveness) against model fit (perplexity). Fig. 3 presents a three-dimensional landscape of the performance for candidate counts of topics for each representation. Enriched abstracts reach peak coherence and distinctiveness at relatively low topic counts: both fully enriched terms and MeSH matched terms maximise coherence at $K = 16$. A similar pattern is observed for the MeSH matched representation, whereas the MeSH parent only representation peaks at K =12, and the original abstract requires a larger K (around 34) to achieve its highest coherence.

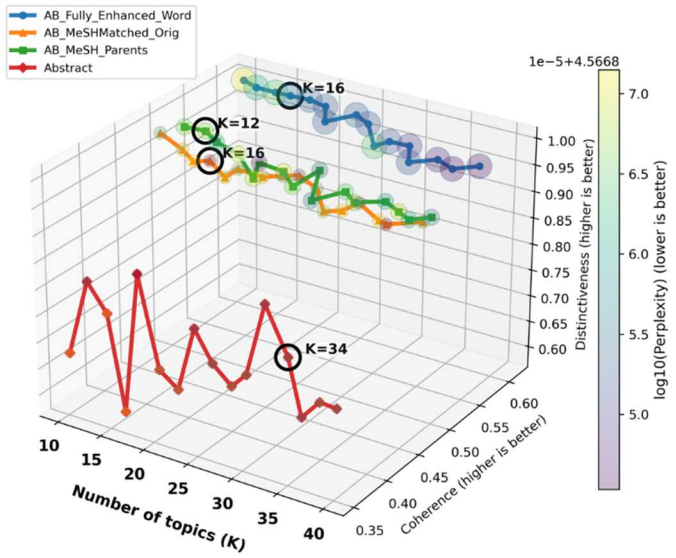

Fig. 3. Optimal Number of Topics (K) for Each Representation.

B. Model Evaluation for Each Representation

TABLE II reports the coherence, distinctiveness and perplexity values at the optimal K. The fully enriched representation exhibits the highest coherence (0.606) and distinctiveness (0.996) and the lowest perplexity (1782.040), which corresponds to a 65.8% improvement when compared to the original abstract. MeSH matched terms achieves moderate improvement with coherence 0.447, distinctiveness 0.909 and perplexity 23349.537, a 22.1% increase over the original abstract. MeSH parents yields coherence 0.424, which is 15.8%, and distinctiveness 0.864 and perplexity 13608.789. The original abstract retains the lowest coherence (0.366), distinctiveness (0.693) and the highest perplexity (36885.640). The enriched representation with MeSH annotations outperforms the representation of the original abstract words in all three metrics.

TABLE II.
AVERAGE LDA PERFORMANCE ACROSS ALL FIELDS (10 RUNS)

Field	Coherence (↑)	Gain vs. Abstract	DST	PPL (↓)
Abstract (original)	0.366	–	0.693	36885.640
AB-Parents	0.424	15.8%	0.864	13608.789
AB-MeSHMatched	0.447	22.2%	0.909	23349.537
AB-Fully Enriched	**0.606**	**65.8%**	**0.996**	**1782.040**

Fig. 4 compares the normalised performance metrics at the selected topic counts using a radar chart. The fully enriched representation achieves the highest coherence (0.606), highest distinctiveness (0.996) and lowest perplexity (1782.040). The partially enriched representation (MeSH matched) also performs well, while MeSH parents and the original abstract occupy progressively smaller areas, reflecting lower coherence and higher perplexity. This radar view succinctly illustrates the trade-offs between semantic enrichment and model complexity.

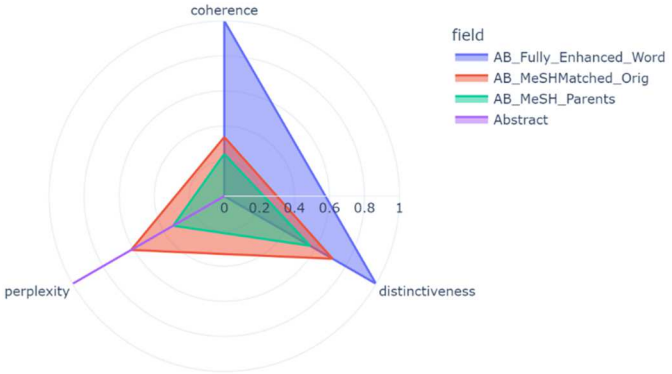

Fig. 4. Power BI Radar (Spider) Chart of Normalised Metrics.

C. Bootstrap Significance Testing

To determine whether the observed coherence differences between representations reflect genuine improvements rather than random variation, we conducted bootstrap-based pairwise comparisons of mean scores. Positive differences indicate that the first representation outperforms the second. The results of TABLE III show that fully enriched terms provide the greatest and the most stable improvement, outperforming the original Abstract by 0.229 (CI: 0.219–0.241), followed by gains over MeSH parents (0.108, CI: 0.095–0.120) and MeSH matched (0.079, CI: 0.065–0.093). Additionally, smaller improvements are observed for MeSH matched over MeSH parents (0.029, CI: 0.019–0.038) and for both enriched fields over the original Abstract (0.150 and 0.122, respectively). All intervals exclude zero, and all contrasts are statistically significant ($p < 0.001$), confirming that MeSH enrichment yields robust improvements in coherence.

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/ICCIAA68481.2026.11544024, 2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIAA)

TABLE III.
BOOTSTRAP SIGNIFICANCE TESTS

A	B	Mean difference (A–B)	CI [Low: High]	P (bootstrap)
Fully Enriched	***MeSH Matched***	**0.079**	**[0.065–0.093]**	< 0.001
Fully Enriched	***Parents***	**0.108**	**[0.095–0.120]**	< 0.001
Fully Enriched	***Abstract***	**0.229**	**[0.219–0.241]**	< 0.001
MeSH Matched	***Parents***	0.029	[0.019–0.038]	< 0.001
MeSH Matched	***Abstract***	0.150	[0.142–0.158]	< 0.001
Parents	***Abstract***	0.122	[0.116–0.126]	< 0.001

D. Topic Stability and Jaccard Similarity

The consistency of the topics in the ten independent runs of the LDA was assessed by computing the mean of the average pairwise Jaccard similarity for each of four representations. High values of mean Jaccard similarity indicate that similar semantic structures are re-created when the model is re-trained using independent random seeds. As shown by TABLE IV, the representation of fully enriched words is most stable (0.725), with several pairs of runs achieving Jaccard indices close to or equal to 1.00, indicating strong convergence and reproducibility of word sets with each other in the extracted topics. The MeSH Parents representation and the MeSH matched representation were getting values of 0.451 and 0.359, respectively. The original abstract has a stability of 0.421, showing the instability of topics that are extracted from un-enriched text.

TABLE IV.
TOPIC STABILITY AND JACCARD SIMILARITY

Group	Mean Jaccard	Min Jaccard	Max Jaccard	SD Jaccard
Abstract (original)	0.421	0.305	0.627	0.076
AB-Parents	0.451	0.340	0.602	0.062
AB-MeSHMatched	0.359	0.254	0.492	0.064
AB-Fully Enriched	**0.725**	0.553	1.000	0.206
Average	**0.489**	**0.363**	**0.680**	**0.102**

E. Topic Labelling and Coherence Improvements

TABLE V presents exemplar topics based on the original abstract, as well as on the fully enriched representation. In the original abstract topics, abstract words are dominated by generic academic terms (e.g., "study", "using", "data"), which results in low coherence. Enrichment adds domain-specific descriptors such as "Neoplasms", "Chemotherapy", "Cytodiagnosis", "Viral", "Vaccines", "Pollutants", "Emissions", "Mellitus" and "Insulin", which is an ontological anchoring that reduces the variances of lexicon and enhances the conceptual consistency between documents.

TABLE V.
EXEMPLAR TOPICS (10 KEYWORDS EACH) COMPARING ABSTRACT (ORIGINAL) VS AB-FULLY-ENRICHED-WORD

Topic	Abstract (Original) vs. AB-Fully Enriched Word (Top 10 keywords)
T1	**Original** → Clinical; Disease; Treatment; Research; Patients; Review; Health; Diseases; Studies; Cancer. **Enriched**→ Neoplasms; Tumor; Intestinal; Chemotherapy; Neoplastic; Immunotherapy; Glandular; Cytodiagnosis; Colonic; Liver. **Original =0.359 → Enr=0.613 (Gain=+0.254)**
T2	**Original** → Death People; Risk Kidney; Failure Death; Kidney Failure; Semaglutide; People Diabetes; Corrects; Vol; Permanent Molars; Compound. **Enriched**→ Diabetes; Overweight; Obesity; Dietary; Anthropometry; Mellitus; Expectancy; Attributable; Ethiopia; Hexoses. **Original =0.353 → Enr=0.605 (Gain=+0.252)**
T3	**Original** → Health; Public; Public Health; Disease; Using; Study; Human; Infection; Infectious; Used. **Enriched**→ Virus; Viral; Animals; Mosquito-borne; Fever; Vaccine; Zoonotic; Vaccines; Protozoan; Influenza. **Original =0.389 → Enr=0.603 (Gain=+0.214)**
T4	**Original** → Health; Public; Public Health; Using; Used; Water; Study; Potential; High; Different. **Enriched**→ Pollution; Water; Geological; Gases; Toxic; Pollutants; Metal; Emissions; Monitoring; Summer. **Original =0.357 → Enr=0.638 (Gain=+0. 281)**
T5	**Original** → Images; Image; Medical; Imaging; Using; Data; Method; Proposed; Model; Used. **Enriched**→ Network; Networks; Images; Computer; Image; Accuracy; Algorithms; Machine; Neural; Segmentation. **Original =0.373 → Enr=0.612 (Gain=+0.239)**
T6	**Original** → Cells; Cell; Potential; Treatment; Effects; Used; Using; Activity; Study; Cancer. **Enriched**→ Protein; Expression; Targets; Discovery; Pathway; Signaling; Genes; Binding; Connective; Biochemistry. **Original =0.362 → Enr=0.605 (Gain=+0.244)**
T7	**Original** → Health; Public; Public Health; Research; Data; Study; Care; Social; COVID-19; Clinical. **Enriched**→ Coronavirus; Viral; COVID-19; Sars-cov-2; Syndrome-related; Virus; Attire; Vaccines; Vaccine; Masks. **Original =0.365 → Enr=0.605 (Gain=+0.240)**

F. Topic Dominance and Mixture Patterns

TABLE VI shows topic dominance statistics. The fully enriched representation has a mean dominance of 0.994, and 98% of documents have dominance $\geqslant$0.90, indicating almost deterministic assignments. On the other hand, the original abstract yields a diffuse topic mixture, as indicated by an average dominance of 0.725, with only 35.3% of documents having dominance $\geqslant$ 0.90. The dominance values of the MeSH-matched and parent representations are also high (> 0.97), while the dominance of the original abstract is small (0.725), and only 35% of documents reach the 0.90 threshold. The difference between the highest and second-highest topic probabilities (top-2 gap) follows a similar pattern. All these findings suggest that the process of enrichment yields stronger topic assignments and eliminates ambiguity in document classification.

TABLE VI.
TOPIC DOMINANCE METRICS FOR EACH REPRESENTATION

Field	Mean Dominance	Top-2 Gap	Proportion $\geqslant$ 0.90
Abstract (original)	0.725	0.597	0.353
AB-Parents	0.978	0.958	0.933
AB-MeSH Matched	0.986	0.972	0.956
AB-Fully Enriched	**0.994**	**0.987**	**0.980**

G. Temporal Topic Dynamics

Modern radial network mapping techniques of the extracted topics, using intelligent business intelligence (BI) visualisation, help illustrate how word clusters evolve over time. Fig. 5 plots the most common topic keywords around a circle, in colours

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/ICCIAA68481.2026.11544024, 2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIAA)

indicating the year in which the documents associated with each keyword peaked. Nodes representing keywords are connected by edges whose thickness denotes co-occurrence strength.

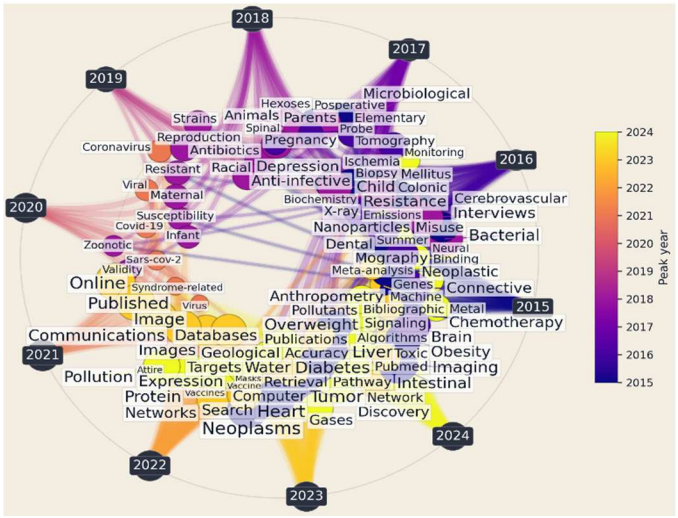

Fig. 5. Topic Trends over Time Using Power BI Radial Networks

Due to the dominance of the pandemic, the difference is that almost no information on topics related to the COVID-19 pandemic is present before 2020, whereas there is a surge during 2020–2022 and a decline by 2024, but it is still above the previous levels. Machine-learning diagnostics grows steadily and peaks in 2023–2024, reflecting the increasing adoption of AI in healthcare.

H. *Advanced Business Intelligence (BI) Visualisation for Topic Modelling Evaluation*

1) Benchmarking Topic Trend Clarity Using Line Charts

Fig. 6 uses Power BI small-multiples line charts to track dominant-topic prevalence from 2015 to 2024, contrasting fully enriched terms with the original abstract. The enriched representation has higher and better-separated trajectories with a significant increase into the early 2020s and higher around 2021-2022 with better representation of the temporal signal and better topic dominance. In comparison, the original abstracts generate compressed and lower counts of lines with weak separation.

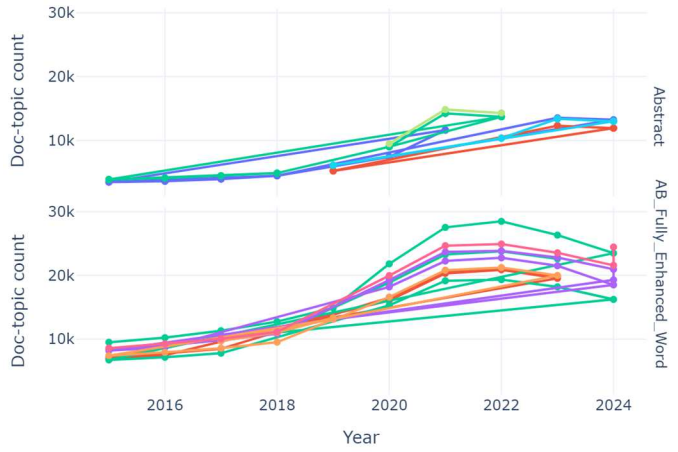

Fig. 6. Power BI Small-Multiples Line Chart of Top Topics over Time

V. DISCUSSION

A. *Impact of MeSH Enrichment on Research Trend Insights*

Across all metrics the fully enriched abstracts significantly outperformed all the other representations. Their coherence score improved to 0.606 and distinctiveness to 0.996, and perplexity reduced to 1782, where the semantically coherent and highly predictive topics were indicated. The stability and dominance tests revealed a mean stability of 0.725 and almost 98% of the documents getting high confidence, compared with markedly lower values of the matched (0.447 coherence, 23349 perplexity) as well as parent (0.424, 13609) variants. Bootstrap tests established these gains to be statistically significant, and qualitative inspection established that enriched topics were very much consistent with domain knowledge. Collectively, the findings confirm that the incorporation of both matched descriptors as well as the addition of their hierarchical parents in abstracts is a significant and evident benefit in trend analysis.

B. *Economic and Business-Intelligence Implications*

The contemporary Business Intelligence (BI) visualisations transformed our evaluation findings into clear decision-support tools for the scientists and administrators. The radar chart equalised coherence, distinctiveness and perplexity at the selected topic counts, which directly indicated the high level of balance as accomplished by the fully enriched representation. A three-dimensional performance landscape mapped candidate topic counts against these metrics, which was used to select the optimal number of topics for each representation. Radial network maps and small-multiple line charts were used to monitor the temporal dynamics of topics to make predictions about the emergent topics, and use this information to align funding and strategic investments, respectively. By integrating these visual analytics, our framework demonstrates how BI can underpin both scientific insight and economic decision-making.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

The paper describes a comprehensive framework for detecting trends in biomedical research using abstract enrichment and MeSH terms in combination with LDA topic modelling. The fully enriched abstract representation markedly improved topic coherence, distinctiveness, stability and dominance while reducing perplexity. The fully enriched representation achieved a coherence of 0.606, a distinctiveness of 0.996 and a perplexity of 1,782, which is way beyond the original abstract. Enriched topics captured both transient phenomena and long-term structural shifts, revealing a more nuanced view of the biomedical landscape than either citation-based metrics or raw text analyses. Compared with the original abstract baseline, our method provides a more reliable and more accurate representation of the research activity. By integrating MeSH enrichment with topic modelling and business-intelligence visualisation, we provide a scalable framework for real-time research trend detection and decision support.

Our current approach is based on traditional LDA, which assumes the possibility of forming a set of words that represent the topics in a statistical way but ignores contextual semantics. Future work will examine recent neural topic models, including BERTopic, Contextualised Topic Models (CTM), Embedding

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/ICCIAA68481.2026.11544024, 2026 2nd International Conference on Computational Intelligence Approaches and Applications (ICCIAA)

Topic Models (ETM) and Top2Vec to determine whether semantic enrichment has additional advantages to more sophisticated models.

ACKNOWLEDGMENT

The authors thank Middle East University, Amman, Jordan, for supporting this research. The primary affiliation for this work is Brunel University of London.

REFERENCES

[1] A. M. Pezzullo and S. Boccia, 'Growth and challenges in biomedical scientific publishing: the phenomenon of mega-journals', *Eur J Public Health*, vol. 34, no. Supplement_3, Nov. 2024, doi: 10.1093/eurpub/ckae144.677.

[2] Q. Jin, R. Leaman, and Z. Lu, 'PubMed and beyond: biomedical literature search in the age of artificial intelligence', *EBioMedicine*, vol. 100, no. 4, p. 104988, 2024, doi: 10.1016/j.ebiom.2024.104988.

[3] I. Guillén-Pacho, C. Badenes-Olmedo, and O. Corcho, 'Dynamic topic modelling for exploring the scientific literature on coronavirus: an unsupervised labelling technique', *Int J Data Sci Anal*, 2024, doi: 10.1007/s41060-024-00610-0.

[4] Manoj Kumar L, R. J. George, and A. P.S, 'Bibliometric Analysis for Medical Research', *Indian J Psychol Med*, vol. 45, no. 3, pp. 277–282, May 2023, doi: 10.1177/02537176221103617.

[5] W. Han, X. Han, S. Zhou, and Q. Zhu, 'The Development History and Research Tendency of Medical Informatics: Topic Evolution Analysis', *JMIR Med Inform*, vol. 10, no. 1, pp. 1–16, 2022, doi: 10.2196/31918.

[6] A. Altarawneh et al., "ETM-F: an enriched topic modeling and filtration framework integrating ontologies and deep learning for biomedical trend analysis," Knowl. Inf. Syst., vol. 68, art. 25, 2026, doi:10.1007/s10115-025-02633-w.

[7] E. F. Dinsa, M. Das, and T. U. Abebe, 'A topic modeling approach for analyzing and categorizing electronic healthcare documents in Afaan Oromo without label information', *Sci Rep*, vol. 14, no. 1, p. 32051, Dec. 2024, doi: 10.1038/s41598-024-83743-3.

[8] Y. Cao *et al.*, 'X-LDA: An interpretable and knowledge-informed heterogeneous graph learning framework for LncRNA-disease association prediction', *Comput Biol Med*, vol. 167, Dec. 2023, doi: 10.1016/j.compbiomed.2023.107634.

[9] A. Dridi, M. M. Gaber, R. M. A. Azad, and J. Bhogal, 'Scholarly data mining: A systematic review of its applications', *Wiley Interdiscip Rev Data Min Knowl Discov*, no. October, pp. 1–23, 2020, doi: 10.1002/widm.1395.

[10] H. Turki *et al.*, 'MeSH2Matrix: combining MeSH keywords and machine learning for biomedical relation classification based on PubMed', *J Biomed Semantics*, vol. 15, no. 1, p. 18, Oct. 2024, doi: 10.1186/s13326-024-00319-w.

[11] X. Wang, R. E. Mercer, and F. Rudzicz, 'MeSHup: A Corpus for Full Text Biomedical Document Indexing', Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.13604

[12] F. Taglino *et al.*, 'An ontology-based approach for modelling and querying Alzheimer's disease data', *BMC Med Inform Decis Mak*, vol. 23, no. 1, p. 153, Aug. 2023, doi: 10.1186/s12911-023-02211-6.

[13] A. Gupta, S. Aeron, A. Agrawal, and H. Gupta, 'Trends in COVID-19 Publications: Streamlining Research Using NLP and LDA', *Front Digit Health*, vol. 3, Jul. 2021, doi: 10.3389/fdgth.2021.686720.

[14] Sateesh Kumar Rongali, 'Natural Language Processing (NLP) in Artificial Intelligence', *World Journal of Advanced Research and Reviews*, vol. 25, no. 1, pp. 1931–1935, Jan. 2025, doi: 10.30574/wjarr.2025.25.1.0275.

[15] Z. Qiang, K. Taylor, and W. Wang, 'How Does A Text Preprocessing Pipeline Affect Ontology Syntactic Matching?', Nov. 2024, [Online]. Available: http://arxiv.org/abs/2411.03962

[16] E. Batbaatar, V. H. Pham, and K. H. Ryu, 'Multi-task topic analysis framework for hallmarks of cancer withweak supervision', *Applied Sciences (Switzerland)*, vol. 10, no. 3, Feb. 2020, doi: 10.3390/app10030834.

[17] R. Prabhakar Kaila, 'Informational Flow on Twitter-Corona Virus Outbreak-Topic Modelling Approach', *International Journal of Advanced Research in Engineering and Technology (IJARET)*, vol. 11, no. 3, pp. 128–134, 2020, [Online]. Available: http://www.iaeme.com/IJARET/index.asp128http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=11&IType=3JournalImpact Factor

[18] C. F. Méndez-Cruz, J. Rodríguez-Herrera, A. Varela-Vega, V. Mateo-Estrada, and S. Castillo-Ramírez, 'Unsupervised learning and natural language processing highlight research trends in a superbug', *Front Artif Intell*, vol. 7, 2024, doi: 10.3389/frai.2024.1336071.

[19] A. Arbaaeen and A. Shah, 'Ontology-based approach to semantically enhanced question answering for closed domain: A review', *Information (Switzerland)*, vol. 12, no. 5, 2021, doi: 10.3390/info12050200.

[20] G. Yazıcı and T. Ozansoy Çadırcı, 'Creating meaningful insights from customer reviews: a methodological comparison of topic modeling algorithms and their use in marketing research', *Journal of Marketing Analytics*, vol. 12, no. 4, pp. 865–887, Dec. 2024, doi: 10.1057/s41270-023-00256-0.

[21] U. Patel, P. Kathiria, C. S. Mansuri, S. Madhvani, and V. Parikh, 'Use of Topic Analysis for Enhancing Healthcare Technologies', *Scalable Computing: Practice and Experience*, vol. 25, no. 3, pp. 1350–1360, Apr. 2024, doi: 10.12694/scpe.v25i3.2360.

[22] X. Wu, T. Nguyen, and A. T. Luu, 'A survey on neural topic models: methods, applications, and challenges', *Artif Intell Rev*, vol. 57, no. 2, p. 18, Jan. 2024, doi: 10.1007/s10462-023-10661-7.

[23] L. Ma *et al.*, 'AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women', *Exp Biol Med*, vol. 250, 2025, doi: 10.3389/ebm.2025.10389.

[24] M. AlShaikh-Hasan and G. Ghinea, 'Predicting Student Performance Using Deep Learning and Machine Learning Techniques', in *2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA)*, IEEE, Apr. 2025, pp. 1–6. doi: 10.1109/ICCIAA65327.2025.11013240.

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

[26] F. S. Tonin, V. Gmünder, A. F. Bonetti, A. M. Mendes, and F. Fernandez-Llimos, 'Use of "Pharmaceutical services" Medical Subject Headings (MeSH) in articles assessing pharmacists' interventions', *Exploratory Research in Clinical and Social Pharmacy*, vol. 7, p. 100172, Sep. 2022, doi: 10.1016/j.rcsop.2022.100172.

[27] Y. K. Agarwal, D. Pandey, and L. S. Umrao, 'Hybrid query refinement based approach for enhanced biomedical image retrieval', *Multimed Tools Appl*, vol. 83, no. 16, pp. 49515–49536, Nov. 2023, doi: 10.1007/s11042-023-17469-1.

[28] Z.-H. Guo *et al.*, 'MeSHHeading2vec: a new method for representing MeSH headings as vectors based on graph embedding algorithm', *Brief Bioinform*, vol. 22, no. 2, pp. 2085–2095, Mar. 2021, doi: 10.1093/bib/bbaa037.

[29] S. Vasudevan, 'Searching of Literature Through Medical Subject Headings (MeSH): An Update', *International Journal of Medical Sciences and Nursing Research*, vol. 3, no. 3, pp. 13–15, Sep. 2023, doi: 10.55349/ijmsnr.2023331315.

[30] S. S. Kim *et al.*, 'A Compendium of Age-Related PheWAS and GWAS Traits for Human Genetic Association Studies, Their Networks and Genetic Correlations', *Front Genet*, vol. 12, Jun. 2021, doi: 10.3389/fgene.2021.680560.

[31] L. Sun, J. Chen, L. J. Li, and L. Li, 'Similarity-based metric analysis approach for predicting osteogenic differentiation correlation coefficients and discovering the novel osteogenic-related gene FOXA1 in BMSCs', *PeerJ*, vol. 12, p. e18068, Sep. 2024, doi: 10.7717/peerj.18068.

[32] R. Maitra, V. Melnykov, and S. N. Lahiri, 'Bootstrapping for Significance of Compact Clusters in Multidimensional Datasets', *J Am Stat Assoc*, vol. 107, no. 497, pp. 378–392, Mar. 2012, doi: 10.1080/01621459.2011.646935.