

Ontology-based Keyword Enriched Topic Modelling for Biomedical Research Trends: A Business Intelligence Perspective

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

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

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

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

Abstract—Over the past ten years, the biomedical literature has increased exponentially, and it is becoming increasingly cumbersome for the researchers and policymakers to identify emerging directions in research and allocate resources appropriately. Bibliometric indicators that are traditional and topic models based on keywords have weaknesses of synonym ambiguity, lag times and shallow semantics. The paper presents a new pipeline, which adds hierarchical concepts of the Medical Subject Headings (MeSH) ontology to author keywords and then uses Latent Dirichlet Allocation (LDA) to identify topics. A data set of 624,555 journal articles from 2015 to 2024 has been collected from the Scopus API in medicine, health professions and molecular biology. Four different systematic representations, raw keywords, MeSH parent categories, direct MeSH matches, and fully enriched keywords are pre-processed. The results showed that using fully enriched keywords with MeSH ontology annotations significantly enhanced topic coherence from 0.338 (baseline keywords) to 0.574, producing more interpretable and semantically consistent topics. The spatial analysis demonstrates another evidence of changing research priorities at the geographical level. The suggested framework illustrates the strength of ontology-driven modelling and visual analytics to assist in making business-intelligence decisions in rapidly evolving biomedical settings.

Index Terms—Topic Modelling, MeSH Ontology, Biomedical literature, Research Trends, LDA, Semantic Enrichment, Business Intelligence, Biomedical Informatics.

I. INTRODUCTION

The volume of published literature in the scientific field, especially in the area of biomedicine, is growing at an unprecedented rate. This information overload makes it increasingly more difficult for health practitioners, researchers and funding agencies to extract valuable information and identify emerging topics [1]. Researchers and policymakers need sophisticated tools to identify new research directions and to properly target direct funding and reward applications [2]. Traditional techniques for detection of research trends have important deficiencies, due to traditional methods that depend mainly on textual content and bibliometric measurement methods [3].

The process of obtaining and detecting new research directions from the extensive scholarly text content remains both incomplete and difficult to execute [2]. Consequently,

discovering hidden and hot patterns within the vast number of articles remains challenging and complex that creates difficulties for researchers attempting to identify precise new research directions. Methods based on textual content (e.g. raw author keywords) are often lacking in semantic depth [4], [5] Domain specific vocabulary, the ambiguous meaning of keywords (polysemy) and synonyms conceal actual trends among research topics, hindering their proper understanding [6]. Furthermore, the widely used bibliometric measurement methodology, which is based on citation counting and co-citation networks, is inherently slow [2]. New scholarly topics take months or even years to receive sufficient citations, and in the process these tracking methods lag behind modern research discoveries [3], [5]. Moreover, a high citation count does not always characterise real significance, especially if authors cite a paper for its tool rather than its content which is a cardinal limitation of bibliometrics [7].

Topic modelling, particularly Latent Dirichlet Allocation (LDA) model, is a powerful statistical technique for clustering large text collections into meaningful topics [8]. However, once applied to unprocessed biomedical text, LDA models have difficulty with domain related features [9]. The vocabulary contained in biomedical literature is extensive and includes many terms and abbreviations that represent a single concept [10]. LDA treats lexical items as independent tokens, and therefore, does not have an inherent mechanism to understand that "COVID-19" and "Coronavirus" might refer to the same concept [11]. Moreover, traditional NLP methods based on the Term Frequency–Inverse Document Frequency (TF-IDF) processing and word co-occurrence metrics lack the ability to detect concept semantic meanings as well as linkages between concepts [12]. This limitation prevents the model from capturing the full semantic meaning and linkages between concepts.

To overcome this ambiguity and complexity in the concepts of medical literature, we use the Medical Subject Headings (MeSH) ontology, a richly structured controlled vocabulary created by the U.S. National Library of Medicine [13]. MeSH is used as a basic framework for biomedical research literature because MeSH supports the semantic enrichment of ambiguous concepts, which helps to enhance the identification of research trends [14], [13]. MeSH terms also enhance the retrieval of studies exploring similar concepts regardless of the terminology used by authors [15]. Consequently, using MeSH keywords to

enrich texts with rich concepts can be more effective than using user-generated bibliometric metadata and raw texts of scholarly publications [14], [5].

In this study, we propose an ontology-enriched topic modelling pipeline that enriches authors' keywords with MeSH annotations before applying LDA to extract topics and discover research trends. By enriching each keywords in published medical research with matched MeSH annotations, four complementary keyword representations (raw keywords, MeSH parents, direct MeSH matches, and fully enriched keywords) are created in the pipeline, which are then further analysed using LDA analysis. The aim of the study is to bridge domain knowledge and unsupervised learning to enhance research trend detection. The contributions of this paper are three-fold:

- A MeSH-based keyword enrichment pipeline that provides a semantic cohesion for topics.
- A comparative evaluation of topic models trained on raw and MeSH enriched keywords based on a set of different metrics.
- The use of the best performing models to identify temporal and geographic trends in biomedical research, translating topic models into business intelligence insights that can be used for strategic funding and policy decisions.

Our objectives are to:

- Collect a large dataset of biomedical articles covering ten years from Scopus and extract author keywords.
- Semantic enrichment of the author keywords in the corpus using MeSH (Medical Subject Headings) annotations and parent terms of the MeSH terms, resulting in multiple representational aspects.
- The implementation of a unified preprocessing pipeline and LDA model training across the multiple representational aspects.
- Evaluate the models using coherence, distinctiveness, perplexity, Jaccard similarity, bootstrap significance tests and dominance metrics.

From these objectives arise our research questions:

- **RQ1:** How does MeSH-based enrichment affect topic coherence compared to raw keywords?
- **RQ2:** Does the inclusion of parent categories increase topic stability and distinctiveness?
- **RQ3:** How does semantic enrichment affect topic dominance and the detection of geographical trends?
- **RQ4:** Can the resulting insights guide strategic planning and resource allocation in biomedical research and industry?

The remainder of this paper is organised as follows. Section 2 reviews the literature on topic modelling, LDA and MeSH based semantic enrichment. Section 3 presents the data set, preprocessing pipeline and the LDA modelling strategy using MeSH enrichment. Section 4 gives the experimental results,

visualisations, and business intelligence implications. Section 5 provides some concluding remarks and suggests directions for future research.

II. LITERATURE REVIEW

A. Topic Modelling in Biomedical Research

Topic modelling (also called topic discovery, topic extraction, or topic analysis) is a statistical method that uses a set of algorithms to find, show, extract, and label the hidden structure of data in huge amounts of document data [16]. A topic is a group of words that typically appears together and is focused on a certain set of terms, similar to the manner in which individuals construct their written works [17]. To put it simply, each document in the corpus includes a percentage of various topics that are mentioned based on the words used in the document [18]. A document has a given likelihood of generating several topics, and topic modelling is made up of certain terms that constitute the topic [17]. Consequently, the goal of topic modelling is to identify a group of terms and topics in different publications.

Using the topic modelling approach based on the original text, this technique can be employed in various tasks such as document retrieval, summarisation, or discovering hidden patterns in the textual data [19]. A multitude of modelling techniques have been used to explore research trends, classification of scholarly publications, and clinical narratives (e.g. electronic health records) and identify shared patterns in patient care [20]. Scarpino et al. used topic-modelling techniques to distinguish recurring themes within texts describing post-acute sequelae of COVID-19 (PASC) and compare them with accounts written by infected healthcare workers who did not exhibit PASC symptoms [17]. Martinis et al. further took this line of inquiry and used various analysing techniques such as VADER sentiment extraction and LDA on a data set from the "Sindrome Post COVID-19" blog which expanded the interpretative possibilities through co-existing thematic analysis [21]. A further refinement of such approaches are given by Muthusami et al., where a comparative evaluation of probabilistic models such as LDA, to non-probabilistic approaches such as a non-negative matrix factorisation (NMF), shows the unique strengths and limitations of each one in extracting topics from short textual data [19].

Lately, BERT-based topic modelling using deep learning models has gained considerable success in the science of revealing research trend detection [22]. Interestingly, the results of a study showed that although BERTopic has an edge in the visual separation of topics, the actual substantive topics that resulted are highly similar to LDA, and the expert evaluators rated the performance of both models to be similar [23]. This means that LDA is in fact competitive against even more modern neural approaches, especially when carefully tuned, and helps reinforce the argument that LDA is still heavily utilised in literature analysis [18].

B. LDA-Driven Topic Discovery

Latent Dirichlet Allocation (LDA) is an unsupervised technique that is one of the most widely used topic modelling algorithms, presented by Blei et al. [24]. Documents with large sizes, such as academic articles, clinical reports, and healthcare

literature, benefit from LDA analysis to generate hidden thematic patterns that normally stay undetected [8]. The LDA model initially generates the topic-word distribution for K topics randomly from the previous distribution (Dirichlet distribution) [25].

LDA stands as a core technique in NLP for analysing unstructured text since it has been extensively used in biomedical text mining research [12]. Using standard NLP principles, the LDA model begins analysing a significant amount of unstructured data [8]. Before training the model to extract topics, the procedure involves preparing tokens, the corpus and a dictionary and then cleansing the data [26]. Although LDA is popularly used for topic modelling in the field, it shows clear limitations in biomedical research. LDA can extract general themes from large text collections [8], but it cannot understand semantic meaning or recognise different terms referring to the same concept [11].

C. Novel MeSH Ontology Enrichment and Semantic for LDA

MeSH is an extensive and controlled vocabulary that has been created by the National Library of Medicine (NLM) for the indexing of biomedical literature [13]. It is one of the most widely used knowledge organisation systems in the biomedical field and thus can be a useful tool to give additional content to terminology with great semantic depth [15]. Fig. 1 shows the hierarchy of the term COVID-19 in MeSH ontology.

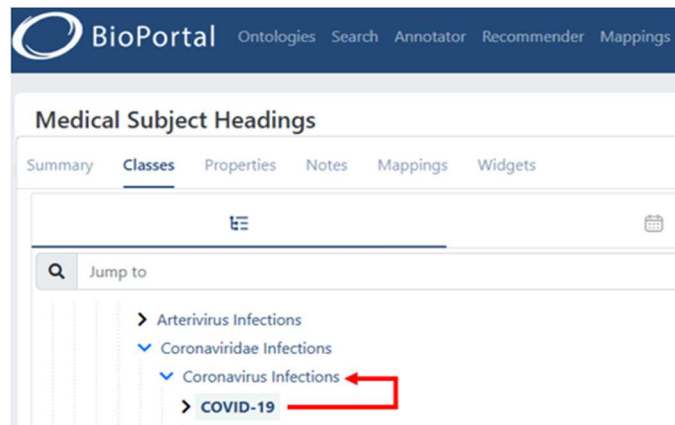


Fig. 1. MeSH Ontology Hierarchy for the Term COVID-19.

The use of MeSH enrichment has been shown to lead to better performance for classification [27], [10], increased performance during information retrieval [27], and relation extraction [14]. Several research studies have combined MeSH with topic modelling [5]. Serrano Nájera et al. used LDA to identify trends of drug discovery after the rise of documents with MeSH synonyms [28], while Valderrama-Zurians et al. compared MeSH indexing with keywords Plus in the construction of models of topics of cannabis research [29]. Most previous studies performed MeSH mapping either post-hoc to label LDA topics or as additional features; our approach goes further by enriching the text before modelling. The combination of ontologies and topic models could improve both the interpretability and the coherence. Recent studies highlight the value of integrating the MeSH-enriched content into LDA that leads to notable improvements in topic coherence and interpretability [30], [5].

III. PROPOSED FRAMEWORK AND METHODOLOGY

A. Dataset Collection and Scope

Metadata for 624,555 journal articles were extracted from Scopus through Elsevier’s API [31], using an advanced query that limited subject areas to Medicine (MEDI), Health Professions (HEAL) and Biochemistry, Genetics & Molecular Biology (BIOC) and high-impact topics, such as clinical research, public health, pharmacology, medical imaging, oncology, and infectious diseases. Only peer-reviewed journal publications between 2015 and 2024 were retained to ensure quality and comparability. For each article, we extracted the following fields: EID, authors, title, keywords, abstract, year, DOI, affiliation, aggregate type, subtype, and subtype description. Fig. 2 illustrates the framework.

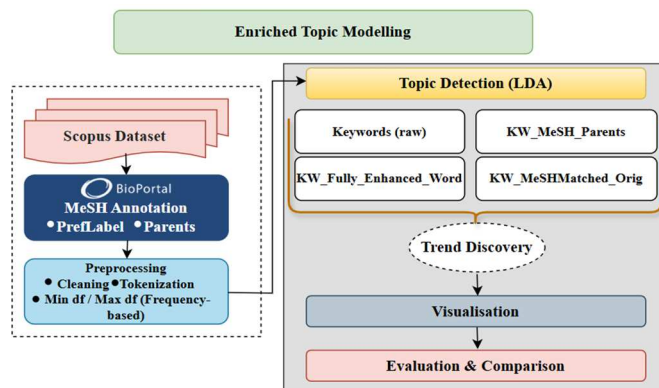


Fig. 2. MeSH-Enhanced Topic Modelling Framework.

In this study, the terms KW_Fully_Enhanced_Word and Fully Enriched, as used in figures and tables, refer to the same representation denoted in the text as Fully Enriched Word.

B. MeSH-Based Keyword Enrichment

The proposed pipeline enriches author-provided keywords by utilising the MeSH ontology. For each article, we mapped each author keyword to its MeSH descriptor and parent categories. Keyword matching was carried out against the MeSH thesaurus using the NCBO BioPortal API; when a correspondence was obtained between a MeSH descriptor, the parent terms (prefLabel) were taken. TABLE I illustrates the enrichment procedure. Where the notation (“→”) 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
Keywords (raw): This field contains the original keywords provided by the author and exported from Scopus, listed as comma-separated values.	Alzheimers disease, caregivers, cognitive impairment, lockdown, COVID-19, dementia, SARS-CoV-2, etc.
KW_MeSH_Parents: This field includes only parent terms, with each represented by a unique MeSH identifier or prefLabel.	Pathologic Processes, Cognition Disorders, Coronavirus Infections, Pneumonia, Viral, Brain Diseases, Neurocognitive Disorders, Severe acute respiratory syndrome-related coronavirus, etc.

Representation & Description	Example
<p><i>KW_MeSHMatched_Orig:</i></p> <p>This field contains the original keyword captured in MeSH, its corresponding parent term.</p>	<p>Alzheimers disease→Pathologic Processes, cognitive impairment→Cognition Disorders, COVID-19→Coronavirus Infections, Pneumonia, Viral; dementia→Brain Diseases, Neurocognitive Disorders; SARS-CoV-2→Severe acute respiratory syndrome-related coronavirus.</p>
<p><i>KW_Fully_Enriched_Word:</i></p> <p>This field contains the original keyword captured in MeSH, its corresponding parent term, as well as any additional keywords not captured in MeSH.</p>	<p>Alzheimers disease (Pathologic Processes), caregivers, cognitive impairment (Cognition Disorders), COVID-19 (Coronavirus Infections, Pneumonia, Viral), dementia (Brain Diseases, Neurocognitive Disorders), lockdown, SARS-CoV-2 (Severe acute respiratory syndrome-related coronavirus), etc.</p>

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:

- **Cleaning:** Brackets, punctuation, and non-alphanumeric characters are removed, extra spaces are removed, and the cleaned text is normalised to lowercase.
- **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"). Free text keywords are tokenised using a regular expression scheme that retains hyphens and underscores.
- **Filtering:** Only tokens that are purely numeric and one character long or empty of alphabets are removed, and then the English stopwords are also filtered out to reduce the noise.
- **Vectorisation:** The bag-of-words representation was built on a consistent vocabulary across the fields of all documents.

D. Topic Modelling and Evaluation with LDA

We used the following evaluations on the resulting models:

1) Pointwise Mutual Information (PMI) and Topic Coherence

Coherence is a measure of how well the words in a topic relate to each other [32]. It correlates strongly with human interpretability of the topics extracted from the model [33]. The PMI calculation formula is shown in Eq. 1. [32].

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)} \quad (1)$$

Where

- $P(w)$: denotes the probability of finding the word w in the sliding window,
- $P(w_i, w_j)$: represents the probability of both words w_i and w_j occurring together in the sliding window,

- Epsilon (ϵ) = 1 is added to avoid the logarithm of zero, to ensure the coherence score returns real numbers.

Coherence is then the sum of PMI scores over all pairs of top words w_i, w_j within a topic set TS [9]:

$$Coherence(TS) = \sum_{w_i, w_j \in TS} PMI(w_i, w_j) \quad (2)$$

Higher levels of coherence scores indicate a higher level of clarity and interpretability of the topics that are derived [23].

2) Perplexity

Perplexity is a classic way to check how well the model predicts a sample of text [34]. Lower perplexity indicates a better model performance, signifying that the model predicts the next word in the corpus with a higher probability [35]. The calculation formula for the perplexity value is shown in Eq. 3.

$$Perplexity(D) = \exp\left(-\frac{\sum_{d=1}^M \log P(W_d)}{\sum_{d=1}^M N_d}\right) \quad (3)$$

Where

- D : represents the set of all words in the document,
- M : denotes the number of documents,
- N_d : denotes the number of words in each document d ,
- $P(W_d)$: represents the probability of each word W_d appearing in the document.

3) Jaccard similarity

To measure the stability of topics across multiple runs, we computed the Jaccard similarity between sets of top keywords. For two sets A and B , the Jaccard index is [36]:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

where A, B are sets of words.

4) Topic Distinctiveness

Topic distinctiveness quantifies how unique each topic is relative to others by measuring the dissimilarity between their keyword distributions. The Kullback–Leibler (KL) divergence is defined as [37]:

$$D_{KL}(h_i || h_j) = \sum_k h_i(k) \log \frac{h_i(k)}{h_j(k)} \quad (5)$$

Where

- $h_i(k)$: Probability of topic k in distribution h_i ,
- $h_j(k)$: Probability of topic k in distribution h_j .

Higher distinctiveness indicates that the topics are well separated and have little lexical overlap, while low values indicate a higher level of topical overlap [37].

5) Bootstrap Significance Tests and Dominance Metrics

We calculated the mean difference in coherence between the two representations of a pair, 95% confidence intervals and p-values. Statistical differences in coherence between representations were analysed using a bootstrap test to test for

statistical significance [38]. Higher dominance means that there are documents having strong association with single topic; lower dominance means they are a mix of topics.

IV. RESULTS AND ANALYSIS

A. Selection of the Optimal Number of Topics

In the first experiment, we aimed to determine the optimal number of K (number of topics) before training the model on the representations by plotting coherence curves. The model selection scores for the four datasets, including the raw keywords, MeSH parents, direct MeSH matches, and fully enriched keywords, are shown in Fig. 3.

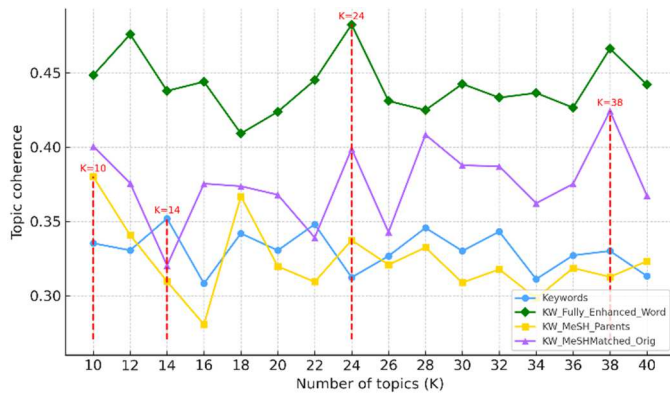


Fig. 3. Optimal Number of Topics (K) for Each Representation.

Coherence was highest at K=14, using the raw representation of the keyword. When mapping of keywords to the associated MeSH parent terms, the optimum was found at K=10. Direct matching of the MeSH terms captured in the field achieved a K=38 value for maximum coherence, suggesting the richer and more specific vocabulary that was involved. The fully enriched representation achieved the highest overall coherence at K=24, achieving a balance between granularity with respect to topics and interpretability.

B. Topic Quality and Performance Metrics

TABLE II summarises the average coherence, distinctiveness (DST) and perplexity (PPL) obtained from ten LDA runs on each representation. For the raw keyword, the coherence showed a relatively low value (0.338) and the highest level of distinctiveness, which would be characteristic of very fragmented and isolated topics. When one associates the keywords with their MeSH parent, coherence increases to 0.473, which is 39.8%, and the perplexity is reduced significantly. Direct alignment with MeSH terms additionally improves coherence to 0.501, a 48.1% increase over baseline. Fully enriched keywords achieve the maximum coherence value (0.574), which corresponds to a 69.8% improvement when compared to the baseline.

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

Field	Coherence	Gain vs. Keywords	DST	PPL
Keywords	0.338	–	0.967	5754.130
Parents	0.473	+39.8 %	0.965	405.300
MeSHMatched	0.501	+48.1 %	0.858	2371.260
Fully Enriched	0.574	+69.8 %	0.828	8585.170

C. Bootstrap Significance Testing

The results presented in TABLE III show that the fully enriched MeSH-based keyword field results in the most coherent topics with a large effect size of 0.109, which compares well to that of the raw Keywords field ($p < .001$; mean improvement of 0.109, 95% confidence interval "CI": 0.097–0.122). Furthermore, it outperforms MeSH Matched and MeSH Parents by 0.067 and 0.118, respectively, which are supported by narrow confidence intervals as well as highly significant p-values. Conversely, MeSH Matched gives a slight improvement of 0.042, while MeSH parent terms offer no statistically significant benefit (mean difference -0.010, $p = 0.225$). Collectively, these results provide evidence that full enrichment provides a much clearer and more semantically coherent topic structure than the raw keywords field.

TABLE III.
BOOTSTRAP SIGNIFICANCE TESTS

A	B	Mean difference (A–B)	95 % CI	P (bootstrap)
Fully Enriched	MeSH Matched	0.067	[0.051, 0.082]	< 0.001
Fully Enriched	Parents	0.118	[0.103, 0.133]	< 0.001
Fully Enriched	Keywords	0.109	[0.097, 0.122]	< 0.001
MeSH Matched	Parents	0.052	[0.034, 0.068]	< 0.001
MeSH Matched	Keywords	0.042	[0.027, 0.057]	< 0.001
Parents	Keywords	-0.010	[-0.024, 0.005]	0.225

D. Topic Stability and Jaccard Similarity

The raw keywords reveal a moderate degree of internal similarity (mean Jaccard similarity (J) = 0.387, Standard Deviation (SD) = 0.118), indicating considerable variation in the sets of keywords generated across different runs, as summarised in TABLE IV. By contrast, MeSH parents have very high internal similarity (mean J = 0.810, range 0.727–1.000), suggesting that use of MeSH parent terms is providing a much more stable and consistent representation of topics. MeSH matched is also better than the raw keywords (mean J = 0.478, SD = 0.043), which has a balance between stability and diversity. Although a fully enriched word is less similar on average (0.320), it does so with a very narrow spread (SD = 0.024), which is an indication of controlled diversity, rather than random noise.

TABLE IV.
TOPIC STABILITY AND JACCARD SIMILARITY

Group	Mean Jaccard	Min Jaccard	Max Jaccard	SD Jaccard
Keywords	0.387	0.269	0.683	0.118
Parents	0.810	0.727	1.000	0.075
MeSHMatched	0.478	0.383	0.595	0.043
Fully Enriched	0.320	0.276	0.369	0.024
Average	0.499	0.414	0.662	0.065

E. Topic Labelling and Coherence Improvements

The enriched keyword representation has a significant impact on topic coherence and can increase the coherence score, which passed in 0.329–0.349 in the raw keyword configuration,

to 0.537-0.607. This improvement has been translated into gains of between 0.192-0.272 for the seven topics, as shown in TABLE V. The antimicrobial resistance theme showed the greatest increase, from 0.335 to 0.607 (+0.272), and topics related to viral infections, obesity and nutrition and lifestyle correlated with chronic diseases all reached improvements of more than 0.262. Wider areas (more domains), such as public health systems and mental and behavioural disorders, show similar increases of about 0.208-0.259. Collectively, these findings confirm that enrichment produces clearer, more coherent, and more interpretable topic structures than the raw keyword representation.

TABLE V.
TOPICS OBTAINED AFTER APPLYING LDA

Topic	Raw keywords vs. Enriched Keywords
T1	Raw → novel, susceptibility, foodborne, multidrug, resistant, soil, probiotics, poultry, antifungal, essential. Enriched → resistance, antimicrobial, microbiological phenomena, E. coli, Staphylococcus, drug resistance, nanoparticles, heavy metals. Raw=0.335 → Enr=0.607 (Gain=+0.272)
T2	Raw → Ebola, TB, prophylaxis, zoonosis, meningitis, remote, transmitted, outbreak. Enriched → virus, viral infections, vaccine, influenza, immunity, biological products, genetic, immune response. Raw=0.336 → Enr=0.598 (Gain=+0.262)
T3	Raw → radiomics, radiotherapy, head, tracking, imaging, dosimetry, exposure. Enriched → cancer, neoplasms, tumour cells, stem cells, breast neoplasms, therapy, physiological phenomena. Raw=0.329 → Enr=0.537 (Gain=+0.208)
T4	Raw → aids, overweight, attitude, workforce, malnutrition, morbidity, pregnant, home, inequality, nurses. Enriched → obesity, overweight, nutritional status, diet, physical activity, public health, age groups, food consumption. Raw=0.343 → Enr=0.602 (Gain=+0.259)
T5	Raw → complications, impairment, cognition, epilepsy, Parkinson's, depressive, schizophrenia, prenatal, obstructive. Enriched → mental disorders, behavioral disorders, social problems, emotions, violence, psychological health, symptoms, age groups. Raw=0.338 → Enr=0.567 (Gain=+0.229)
T6	Raw → tropical, zoonosis, leishmaniasis, plasmodium, outbreaks, mosquito, aedes, Bangladesh, Nigeria. Enriched → mosquito-borne diseases, malaria, protozoan infections, flavivirus, Africa, tropical diseases. Raw=0.337 → Enr=0.552 (Gain=+0.215)
T7	Raw → urban, smoking, inactivity, pollution, diabetes, hypertension, cardiovascular, city, environment. Enriched → chronic disease, cardiovascular diseases, diabetes mellitus, smoking, physical inactivity, risk factors, urban population. Raw=0.330 → Enr=0.576 (Gain=+0.246)

F. Topic Dominance and Mixture Patterns

Dominance metrics reveal how decisively documents align with their most probable topic. TABLE VI shows the mean dominant weight topics for all representations. Raw keywords assign about 81% probability to the dominant topic. MeSH parent and MeSH matched fields lower dominance, yielding more balanced mixtures. The fully enriched field produces nearly perfect dominance (99.7%), indicating that MeSH terms act as strong categorical anchors, effectively reducing ambiguity. While this limits the mixture of topics per document, it is advantageous for specific trend detection where clear separation between research topics is required, minimising noise

often found in the raw keywords field. This mirrors the high coherence observed earlier.

TABLE VI.
TOPIC DOMINANCE

Field	Mean Dominant Weight	Mean Top-Two Gap
Keywords	0.809	0.692
Parents	0.758	0.618
MeSH Matched	0.731	0.558
Fully Enriched	0.997	0.994

G. Semantic Relationship Graphs

Fig. 4 shows a comparative study of the comparison of semantic graphs produced from raw and enriched keyword sets.

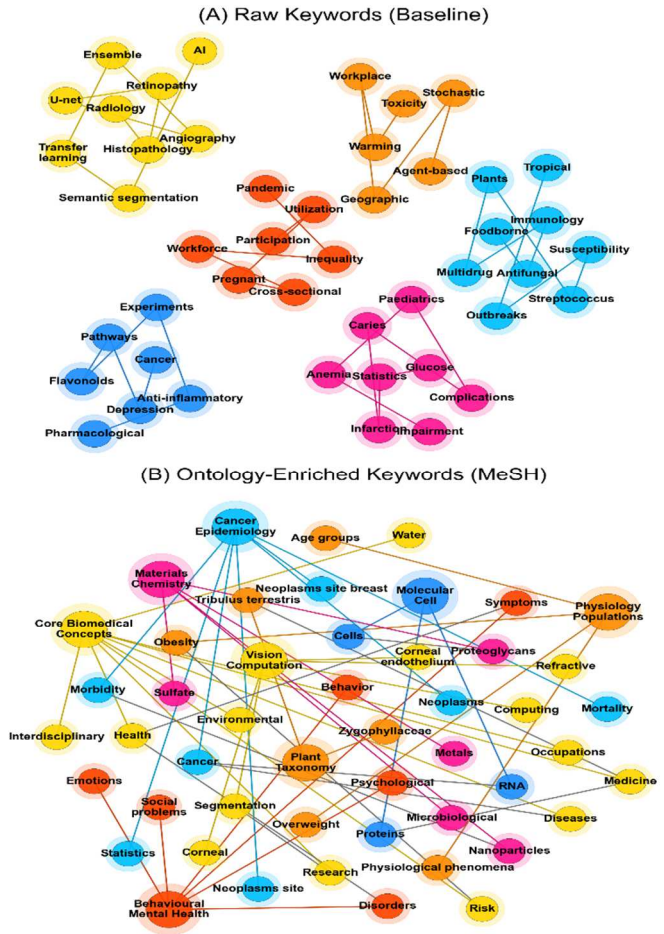


Fig. 4. Semantic Relationship Graphs.

At the top of the graph, the raw keywords group into fragmented ensembles with little interconnections, offering only a limited representation of the conceptual landscape. Meanwhile, at the bottom, the enriched graph shows not only an increase in density but also the appearance of hubs that connect different, disparate subfields, hence developing latent trends and cross-disciplinary links. This contrast underscores the ability of MeSH enrichment to expose hidden structures.

H. Geographic Distribution of Topics

Using Microsoft Power BI, an animated choropleth world map was created to represent the spatial distribution of topic-

specific research output by countries and to show the change in this distribution over time. Spatial patterns shown in the animated maps (Fig. 5) show that in the beginning of the second decade of the twenty-first century, the United States and several Western European countries (United Kingdom, Germany, Italy) dominated in most of the areas of research. By the middle of the 2020s, China, India, and Brazil have come to the fore, signalling a transition to a multipolar research arena. The use of MeSH enrichment has clearly highlighted the emerging research centres in East Asia and Latin America.

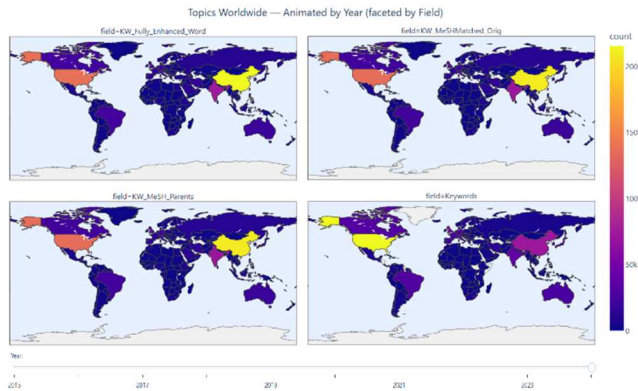


Fig. 5. Geographical Distributions.

V. DISCUSSION

Compared to all the metrics evaluated, ontology-based enrichment significantly improves the quality of the topic. Enrichment of keywords with MeSH parent terms was found to increase coherence by about 40% as compared with raw keywords. Incorporation of MeSH match descriptors brought the coherence even further up to 0.501, at the cost of lower distinctiveness. The fully enriched words, where raw keywords, matched descriptors and parent categories are integrated, had the best coherence (0.574) and were a ~70% improvement on baseline keywords. Although the fully enriched words demonstrated higher perplexity (8585), this is an expected consequence of the significantly expanded vocabulary introduced by the MeSH ontology terms. While perplexity measures predictive likelihood, it does not always correlate with human interpretability. These gains were statistically significant and robust across different runs and thus show that domain knowledge contributes significantly to the topic interpretability and stability.

In practice, diffuse amalgamations of unrelated keywords coalesce into sensible topics; for example, a raw topic with food, soil and resistance transforms into a focused antimicrobial resistance topic. This helps in easy topic labelling, and the highly dominant topics generated are consistent with MeSH enrichment, producing representations that are in close alignment with document semantics. Geographic patterns point to a shift from a U.S./Europe centred landscape to a more multipolar sphere where China, India, and Brazil will contribute more. These insights provide an argument for tools which can detect long-term as well as short-term trends.

The resulting framework serves as a powerful business intelligence tool, which will enable scholars, funding agencies and policymakers to go beyond the limitations that are inherent

in retrospective bibliometric analyses. By providing a dynamic, early-stage view of emerging research trends, this system facilitates data-driven decision-making, ensuring that resources are allocated to the most promising and rapidly evolving areas of biomedical science. The geographic analysis shows a shift from Western dominance to increasing contributions from Asia and Latin America, signalling opportunities for international collaboration and investment. By converting unstructured metadata into actionable knowledge, the methodology supports evidence-based decision-making.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

This study shows that it is significantly greater for augmenting author keywords with the MeSH ontology to improve the topic modelling to detect trends in biomedical literature. The Scopus dataset was processed in order to generate four representations in terms of keywords. Following standardised preprocessing and several runs of LDA modelling, this fully enriched representation showed significantly higher coherence for the topics, and the average score for this was 0.57. This performance surpasses that of the raw keywords by nearly 70%, yielding more stable and highly interpretable topics. Several evaluations and visualisations were used to verify the superiority of the enriched model. The pipeline that is developed not only contributes to a better understanding on the scientific side, but it also provides an instrument for policymakers and industry actors to track the research developments and invest in resources more appropriately.

Our methodology is domain agnostic, which means that this approach can be applied to any domain where a hierarchical ontology applies (i.e., any field where this type of injection of parent concepts into keyword metadata can be used prior to topic modelling). Nevertheless, there are a number of limitations which are worth discussing. Firstly, the total reliance on keywords ignores the rich information present in abstracts; the future work will increase the enrichment pipeline to include abstracts and full text. Secondly, the LDA framework is based on static topics, while topics in dynamic topic models might better reflect evolving topics. Finally, the approach proposed, while improving topic coherence, neural topic models like BERTopic can provide a higher semantic fidelity; comparative tasks between ontology-enriched LDA and embedding-based models can be explored in the future.

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] 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.
- [2] R. Senthil, T. Anand, C. S. Somala, and K. M. Saravanan, "Bibliometric analysis of artificial intelligence in healthcare research: Trends and future directions," *Future Healthc J*, vol. 11, no. 3, p. 100182, 2024, doi: 10.1016/j.fhj.2024.100182.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] R. Costa, T. Bragato, and R. Fileto, “Semantic Similarity: A Domain Analysis,” *Brazilian Journal of Information Science: research trends*, vol. 18, p. (18), 24., 2024, doi: 10.36311/1981.
- [7] 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.
- [8] J. Walsh, J. Cave, and F. Griffiths, “Combining Topic Modeling, Sentiment Analysis, and Corpus Linguistics to Analyze Unstructured Web-Based Patient Experience Data: Case Study of Modafinil Experiences,” *J Med Internet Res*, vol. 26, 2024, doi: 10.2196/54321.
- [9] A. Krishnan, “Exploring the power of topic modeling techniques in analyzing customer reviews: a comparative analysis,” Aug. 2023, doi: 10.48550.
- [10] X. Pan, P. Huang, S. Li, and L. Cui, “MCRWR: a new method to measure the similarity of documents based on semantic network,” *BMC Bioinformatics*, vol. 23, no. 1, Dec. 2022, doi: 10.1186/s12859-022-04578-1.
- [11] 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.
- [12] A. Benlahbib, A. Boumhidi, A. Fahfouh, and H. Alami, “Comparative Analysis of Traditional and Modern NLP Techniques on the CoLa Dataset: From POS Tagging to Large Language Models,” *IEEE Open Journal of the Computer Society*, 2025, doi: 10.1109/OJCS.2025.3526712.
- [13] 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.
- [14] 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.
- [15] F. Fernandez-Llimos, L. G. Negrão, C. Bond, and D. Stewart, “Influence of automated indexing in Medical Subject Headings (MeSH) selection for pharmacy practice journals,” *Research in Social and Administrative Pharmacy*, vol. 20, no. 9, pp. 911–917, Sep. 2024, doi: 10.1016/j.sapharm.2024.06.003.
- [16] M. Yousef and D. Voskergian, “TextNetTopics: Text Classification Based Word Grouping as Topics and Topics’ Scoring,” *Front Genet*, vol. 13, Jun. 2022, doi: 10.3389/fgene.2022.893378.
- [17] I. Scarpino, C. Zucco, R. Vallelunga, F. Luzza, and M. Cannataro, “Investigating Topic Modeling Techniques to Extract Meaningful Insights in Italian Long COVID Narration,” *BioTech*, vol. 11, no. 3, p. 41, Sep. 2022, doi: 10.3390/biotech11030041.
- [18] 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.
- [19] R. Muthusami, N. Mani Kandan, K. Saritha, B. Narenthiran, N. Nagaprasad, and K. Ramaswamy, “Investigating topic modeling techniques through evaluation of topics discovered in short texts data across diverse domains,” *Sci Rep*, vol. 14, no. 1, p. 12003, May 2024, doi: 10.1038/s41598-024-61738-4.
- [20] T. Porturas and R. A. Taylor, “Forty years of emergency medicine research: Uncovering research themes and trends through topic modeling,” *Am J Emerg Med*, vol. 45, pp. 213–220, Jul. 2021, doi: 10.1016/j.ajem.2020.08.036.
- [21] M. C. Martinis, I. Scarpino, C. Zucco, and M. Cannataro, “Investigating the Sentiment in Italian Long-COVID Narrations,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, Jul. 2023, pp. 678–690. doi: 10.1007/978-3-031-36021-3_65.
- [22] B. B. Ozkara, M. Karabacak, K. Margetis, W. Smith, M. Wintermark, and V. S. Yedavalli, “Trends in stroke-related journals: Examination of publication patterns using topic modeling,” *Journal of Stroke and Cerebrovascular Diseases*, vol. 33, no. 6, p. 107665, Jun. 2024, doi: 10.1016/j.jstrokecerebrovasdis.2024.107665.
- [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] Blei, D. M., A. Y. Ng, and Michael I., “Latent dirichlet allocation,” *Journal of Machine Learning Research* 3, pp. 993–1022, 2003.
- [25] H. Maennel, “Exact marginal inference in Latent Dirichlet Allocation,” Mar. 2020, [Online]. Available: <http://arxiv.org/abs/2004.00115>
- [26] P.-J. M. Noble, C. Appleton, A. D. Radford, and G. Nenadic, “Using topic modelling for unsupervised annotation of electronic health records to identify an outbreak of disease in UK dogs,” *PLoS One*, vol. 16, no. 12, p. e0260402, Dec. 2021, doi: 10.1371/journal.pone.0260402.
- [27] H. Kammoun, I. Gabsi, and I. Amous, “MeSH-Based Semantic Indexing Approach to Enhance Biomedical Information Retrieval,” *Computer Journal*, vol. 65, no. 3, pp. 516–536, Mar. 2022, doi: 10.1093/comjnl/bxaa073.
- [28] G. Serrano Nájera, D. Narganes Carlón, and D. J. Crowther, “TrendyGenes, a computational pipeline for the detection of literature trends in academia and drug discovery,” *Sci Rep*, vol. 11, no. 1, p. 15747, Aug. 2021, doi: 10.1038/s41598-021-94897-9.
- [29] J. C. Valderrama-Zurián, C. García-Zorita, S. Marugán-Lázaro, and E. Sanz-Casado, “Comparison of MeSH terms and KeyWords Plus terms for more accurate classification in medical research fields. A case study in cannabis research,” *Inf Process Manag*, vol. 58, no. 5, p. 102658, Sep. 2021, doi: 10.1016/j.ipm.2021.102658.
- [30] You and Sukjin, “The MeSH-Term Query Expansion Models using LDA Topic Models in Health Information Retrieval,” *Journal of Korean Library and Information Science Society*, vol. 52, no. 1, pp. 79–108, 2021, doi: 10.16981/KLISS.52.1.202103.79.
- [31] Elsevier, “Elsevier Research Products APIs.” Accessed: Dec. 11, 2025. [Online]. Available: <https://dev.elsevier.com/>
- [32] J. M. Campagnolo, D. Duarte, and G. Dal Bianco, “Topic Coherence Metrics: How Sensitive Are They?,” *Journal of Information and Data Management*, vol. 13, no. 4, Oct. 2022, doi: 10.5753/jidm.2022.2181.
- [33] J. D. Romero, M. A. Fejoo-García, G. Nanda, B. Newell, and A. J. Magana, “Evaluating the Performance of Topic Modeling Techniques with Human Validation to Support Qualitative Analysis,” *Big Data and Cognitive Computing*, vol. 8, no. 10, p. 132, Oct. 2024, doi: 10.3390/bdcc8100132.
- [34] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, “Topic modeling algorithms and applications: A survey,” *Inf Syst*, vol. 112, p. 102131, Feb. 2023, doi: 10.1016/j.is.2022.102131.
- [35] T. Zhang, Z. Li, and Y. Qi, “Research on digital literacy literature: Topic mining and evolutionary analysis,” *Journal of Librarianship and Information Science*, Jan. 2025, doi: 10.1177/09610006241309645.
- [36] J. Rieger, C. Jentsch, and J. Rahnenführer, “LDAPrototype: a model selection algorithm to improve reliability of latent Dirichlet allocation,” *PeerJ Comput Sci*, vol. 10, p. e2279, Sep. 2024, doi: 10.7717/peerj-cs.2279.
- [37] S. Pan, M. X. Zhou, Y. Song, W. Qian, F. Wang, and S. Liu, “Optimizing temporal topic segmentation for intelligent text visualization,” in *Proceedings of the 2013 international conference on Intelligent user interfaces*, New York, NY, USA: ACM, Mar. 2013, pp. 339–350. doi: 10.1145/2449396.2449441.
- [38] 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.