INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Semantic Topic Discovery for Curriculum Analysis of NCERT Chemistry Textbooks Using Transformer-Based Language Models

Saifullah Rahimi[1]*, **Hiren Joshi**[1], **Hardik Joshi**[1]

1 Department of Computer Science , Gujarat University, Ahmedabad-380009, Gujarat, India

OPEN ACCESS

Received: 25/08/2026

Accepted: 01/10/2026

Published: 25/10/2026

Citation: Rahimi S, Joshi H, Joshi H (2026) Semantic Topic Discovery for Curriculum Analysis of NCERT Chemistry Textbooks Using Transformer-Based Language Models. Indian Journal of Science and Technology 19(38): 2612-2623. https://doi.org/10.17485/IJST/v19i38.1220

* **Corresponding author**.

saifullah.r@gujaratuniversity.ac.in

Funding: None

Competing Interests: None

Copyright: © 2026 Rahimi et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: This study aims to develop and compare an AI-based approach for analyzing the conceptual structure of NCERT Chemistry curricula through BERTopic and Latent Dirichlet Allocation (LDA). **Method:** The textbooks from Classes XI and XII (19 chapters) were gathered from the official NCERT website repository and converted to machine-readable text through OCR. Texts were normalized, filtered, tokenized, lemmatized, and classified into approximately 500-token chunks. Further BERTopic used Sentence Transformer embeddings, HDBSCAN, UMAP, and c-TF-IDF. LDA served as a comparative model. Models were evaluated using C_v topic coherence and qualitative evaluation of topic interpretability and alignment between curriculum and domains. **Findings:** BERTopic obtained a C_v coherence score of 0.64, contrasted with 0.54 for LDA, indicating an approximately 19% relative optimization. BERTopic also generated more distinct and domain-aligned topics, reflecting major areas including “organic”, “inorganic”, and “physical chemistry”, as well as concepts such as “chemical bonding”, “equilibrium”, “thermodynamics”, “electrochemistry”, and “atomic structure”. Alternatively, LDA produced overlapping and lower interpretable topics, including combinations of “thermodynamics”, “equilibrium,” and “reaction kinetics”. The outcomes are consistent with the expected strength of contextual embeddings over bag-of-words interpretation for capturing semantic relationships. Conversely evaluation was based on coherence and qualitative analysis rather than statistical significance testing. **Novelty:** The analysis benchmarks BERTopic against LDA on school based NCERT Chemistry textbooks and validates its potential for fine-grained, domain-aligned curriculum assessment.

Keywords: Educational Data Mining; Curriculum Analysis; Transformer Models; BERTopic; Latent Dirichlet Allocation

1 Introduction

Recent advances in natural language processing have expanded the capacity to analyze large educational text collections beyond traditional frequency-based approaches. While BoW-based models represent documents primarily through patterns of word occurrence, contextualized and embedding-based approaches incorporate richer semantic representations that can improve the coherence and interpretability of discovered topics [1]. Recent comparative research further demonstrates the continuing importance of evaluating classical and embedding-based topic models across different corpora rather than assuming that one approach is universally superior [2]. More recent embedding-based approaches, including BERTopic, leverage contextual representations to identify semantically related groups of documents and terms, offering an alternative for capturing contextual relationships that conventional bag-of-words models may overlook [3]. Consequently, topic modeling is increasingly being considered not only as a tool for topic discovery but also as a means of examining the structure and evolution of knowledge within large textual corpora. In addition to that, recent studies have failed to develop any domain-specific assessment frameworks for analyzing such curriculum features as conceptual density, curriculum alignment, and knowledge progression [4],[5]. A recent analysis of 793 publications spanning 2000–2024 demonstrates the growing methodological development of educational data mining and predictive modeling in the era of AI, with increasing attention to emerging analytical techniques and their applications in educational contexts [6]. Romero and Ventura [7] emphasized the importance of content analysis automation for gaining insights into the educational environment.

Most existing research is limited to higher education datasets, learning analytics and scholarly publications, while school level textbooks have been neglected, particularly in developing educational contexts [4]. Additionally, traditional topic modeling methods like Latent Dirichlet Allocation (LDA) depend on bag-of-words representations that are unable to capture the contextual relationships between words. This limitation often leads to less coherence and clarity of issues. The recent studies have demonstrated that the use of contextual embeddings has a strong positive impact on the topic coherence and semantic quality [1]. Moreover, frameworks like OCTIS highlight the importance of systematic evaluation and comparison of topic models through established metrics and benchmarking methodologies [8]. This application is particularly relevant to curriculum analysis in India, where the National Education Policy (NEP 2020) emphasizes conceptual understanding and interdisciplinary learning.

1.1 Transformer-Based Topic Modeling

Transformer based models have achieved significant progress on semantic text representation. BERT made contextual embedding of text possible [9]. Comparative studies [10],[11] have demonstrated that embedding-based models are better to traditional methods in terms of coherence, semantic richness and interpretability. Recent studies illustrate the use of BERTopic in identifying complex and semantically coherent topics in large corpora, especially in sustainability and interdisciplinary research [12].

1.2 Comparative Analysis of Topic Modeling Techniques

The recent work shows that there is always a trade-off involved in topic modeling methods . The empirical investigation into the integration of clustering, LDA, and co-word analysis has revealed the potential of hybrid approaches to increase the consistency of classification and topic validation, provided that the embedded domain knowledge is taken into account [13]. In addition, many bibliometric analyses have confirmed the efficacy of the integration of topic modeling and network analysis for increasing the precision of thematic extraction in artificial intelligence [14]. Yet, the comprehensive benchmarking based on school-level dataset is scarce.

1.3 Research Gaps

Accordingly, the central research problem addressed in this study is whether contextual embedding-based topic modeling can provide a more informative and interpretable representation of curricular knowledge than a conventional probabilistic topic model when applied to school-level science textbooks. The models are evaluated using quantitative topic-coherence measures together with qualitative examination of topic interpretability and curricular relevance. The study therefore connects recent advances in contextual NLP with computational curriculum analysis, rather than treating topic modeling solely as a generic text-mining task. The study makes the following contributions:

- Methodological comparison: It provides a systematic comparison of LDA and BERTopic for identifying latent thematic structures in a school-level textbook corpus.
- Curricular application: It applies both approaches to NCERT Chemistry textbooks for Classes XI and XII, providing a computational analysis of the organization of curricular content.

- Multi-dimensional evaluation: It combines quantitative topic-coherence assessment with qualitative evaluation of topic interpretability and curricular relevance.

2 Methodology

In this research paper, we present a comparative artificial Intelligence based framework for curriculum analysis with topic modeling techniques. Figure 1, displays the multi-stage pipeline is depicted from data preprocessing to comparative evaluation. The system combines transformer-based semantic modeling (BERTopic) and traditional probabilistic methods (LDA) to evaluate their capacity to extract meaningful structures from instructional material. The full pipeline consists of five stages as follows:

1. Data acquisition and preprocessing
2. Text segmentation (chunking)
3. Topic modeling (BERTopic and LDA)
4. Topic representation and extraction
5. Quantitative and qualitative evaluation

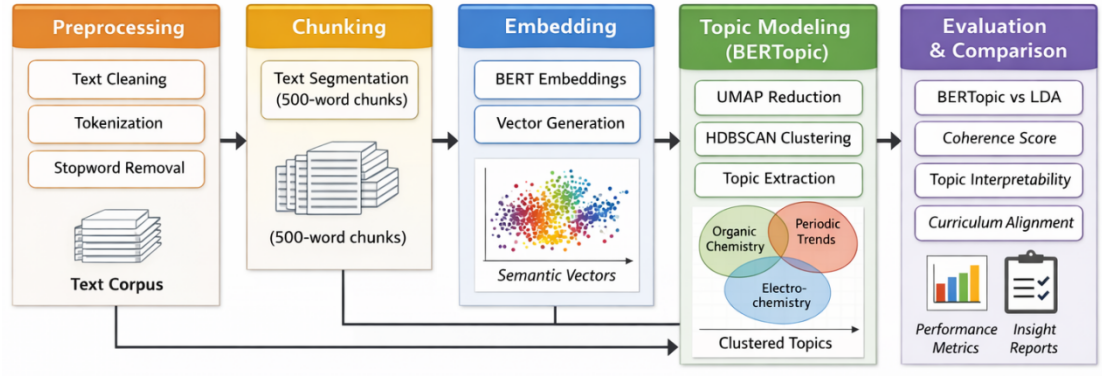


Fig 1. Proposed Methodology Framework

2.1 Dataset and Preprocessing

This data set consists of NCERT Chemistry books for class XI and XII containing all chapters in order to cover the whole syllabus. All textbooks were collected from the NCERT official repository (https://ncert.nic.in/). In total, the samples are 19 chapters, the exact dataset supporting the findings of this study are available on (https://github.com/saifujasoor/NCERT_Chemistry.git). Further the data were converted chapter by chapter to the machine-readable form through Optical Character Recognition (OCR). A robust preprocessing pipeline was developed to prepare the data to have the best possible input for topic modeling. This pipeline includes linguistic normalization and domain-specific filtering by means of NLTK, spaCy, and regex. The models were implemented using Python in a Google Colab environment. Key libraries include:

- BERTopic for transformer-based topic modeling
- Scikit-learn for LDA implementation
- Gensim for coherence evaluation
- spaCy and NLTK for preprocessing

Hyperparameters were tuned to optimize topic granularity:

- Chunk size: 500 tokens
- HDBSCAN min cluster size: 3
- UMAP neighbors: 5

These are the following preprocessing steps:

1. Text Normalization: lowercasing, removing punctuation/special characters/OCR artifacts.
2. Tokenization: Sentence and word tokenization by spaCy linguistic models.
3. Stopword Removal: Removing common English stopwords and domain-specific non-informative keywords (e.g., "chapter", "figure", "example").
4. Lemmatization: The step when words are converted to their root forms to keep semantic consistency.
5. Chunking: The step of splitting the text to fixed-length semantic chunks for further modeling.

These preprocessing steps make the dataset clean and ready to model topics with the maximum consistency.

Formally, given a corpus of documents:

$$\mathrm{D} = \{\mathrm{d}_1, \mathrm{d}_2, \ldots, \mathrm{d}_\mathrm{n}\} \tag{1}$$

Each document d_i is transformed into a cleaned representation:

$$d_i' = preprocess(d_i) \tag{2}$$

2.2 Text Segmentation (Chunking)

To enhance topic granularity and semantic coherence, each page is segmented into smaller portions of around 500 chunks. Formally, a document d_i is transformed into a set of chunks:

$$d_i \rightarrow \{ci1, ci2, ..., c_{ij}\} \tag{3}$$

where c_{ij} denotes a segment of around 500 chunks. This phase guarantees the preservation of local semantic structures and enhances clustering performance.

2.3 Topic Modeling Approaches

2.3.1. Latent Dirichlet Allocation (LDA)

LDA is a probabilistic generative model positing that documents are mixes of latent themes, with topics represented as distributions over words. LDA represents each document as a combination of latent themes, with each topic characterized by a distribution of words.

The generative process is characterized as follows:

$$\theta_d \sim Dirichlet(\alpha) \tag{4}$$

$$\phi_k \sim Dirichlet(\beta) \tag{5}$$

$$z_{d,n} \sim Multinomial(\theta_d) \tag{6}$$

$$w_{d,n} \sim Multinomial(\theta_{z_{d,n}}) \tag{7}$$

Where:

θ_d: topic distribution for document d
ϕ_k: word distribution for topic k
$z_{d,n}$: topic assignment
$w_{d,n}$: observed word

2.3.2. BERTopic (Transformer-Based Topic Modeling)

BERTopic blends semantic embeddings with clustering and statistical topic representation. BERTopic utilizes contextual embeddings and clustering methodologies to provide semantically profound topics. The execution comprises:

1. Embedding Generation: SentenceTransformer model (all-MiniLM-L6-v2)
2. Dimensionality Reduction: UMAP for preserving semantic structure
3. Clustering: HDBSCAN for density-based grouping
4. Representation of Topics: Class-based TF-IDF (c-TF-IDF)

This method captures contextual associations among words, rendering it especially appropriate for domain-specific teaching material.

1. ***Embedding Generation***

   At this stage, each text segment is converted into a dense vector representation via a transformer-based model (BERT). Every document segment is converted into a dense vector via a transformer model:

$$E = f_{BERT}(C) \tag{8}$$

   where:

   C: set of document chunks

   $E \in R^{nxd}$: embedding matrix

   These embeddings encapsulate semantic significance and constitute the basis for grouping.

2. ***Dimensionality Reduction***

   To preserve local structure and reduce computational complexity:

$$E' = UMAP(E) \tag{9}$$

3. ***Clustering***

   Density-based clustering groups semantically similar documents:

$$C = HDBSCAN(E') \tag{10}$$

4. ***Topic Representation (c-TF-IDF)***

   This approach categorizes semantically analogous segments into clusters, yielding comprehensible topics aligned with curriculum subjects such as organic chemistry, thermodynamics, and atomic structure.

   Topics are represented using class-based TF-IDF:

$$cTFIDF_{t,w} = \frac{f_{t,w}}{\Sigma_w f_{t,w}} \log\left(\frac{N}{n_w}\right) \tag{11}$$

   Where:

   $f_{t,w}$: Frequency of word www in topic t

   N: Total number of documents

   n_w: Number of documents containing word w

2.4 Evaluation Metric

To ensure rigorous comparison, Topic Coherence (C_v) evaluation metric is used. Topic coherence measures semantic similarity between words:

$$C_v = \frac{1}{N}\sum_{i,j}\log\frac{D(w_i,w_j)+\epsilon}{D(w_j)} \tag{12}$$

Where:
$D(w_i,w_j)$: co-occurrence frequency
$D(w_j)$: frequency of word w_j

The final stage evaluates the extracted topics using both quantitative and qualitative measures. In addition, a comparative analysis with LDA is performed. The performance of BERTopic and LDA is evaluated based on:

- Quantitative metrics (Coherence Score and Topic Diversity)
- Topic interpretability
- Domain alignment with curriculum structure
- Noise robustness

The results enable identification of the most effective model for educational text analysis. Which ensures that the extracted topics are not only statistically valid but also educationally meaningful.

3 Results and Discussion

BERTopic generated semantically related topic groups through contextual embeddings, dimensionality reduction, density-based clustering, and class-based TF-IDF representation, whereas LDA represented topics through probabilistic word distributions. In Figure 2, presents the BERTopic pipeline, highlighting the transformation from raw text to semantically coherent topics. Unlike generic topic modeling visualizations, the presented relevance and interpretability. The BERTopic model, applied to the NCERT Chemistry corpus, refer to Table 1, generated 4 topics in initial stage, after tuning the model in stage Tuned v1 we got 5 generated topics and in next round model tuning finally the tuned BERTopic model successfully identified fine-grained conceptual structures, producing 19 distinct topics.

topics are derived directly from the NCERT Chemistry corpus, ensuring domain-specific

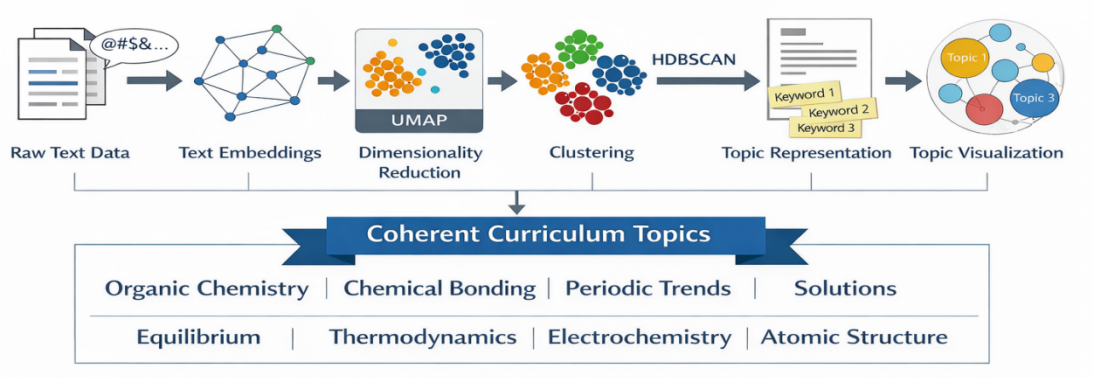


Fig 2. BERTopic Pipeline from raw text to coherent topics

Table 1. Fine-grained BERTopic results

Stage	Topics	Meaning
Initial	4	Over generalized
Tuned v1	5	Slight Improvement
Tuned v2(Final)	19	Fine-grained conceptual extraction

3.1 Quantitative Evaluation

The primary quantitative evaluation used the C_v topic-coherence measure. BERTopic achieved a coherence score of 0.6443, whereas LDA achieved 0.5414. The absolute difference was 0.1029, corresponding to an approximately 19.0% higher coherence score for BERTopic relative to LDA:

$$\Delta C_v = \frac{C_v^{\text{BERTopic}} - C_v^{\text{LDA}}}{C_v^{\text{LDA}}} \text{x}100 \tag{13}$$

$$\Delta C_v = \frac{0.6443 - 0.5414}{0.5414} x100 \approx 19.0\% \tag{14}$$

A higher value of coherence implies that the topic words extracted by BERTopic are more semantically related than the ones by LDA for this corpus. This is consistent with the difference in their methodologies. LDA being more of a word frequency based algorithm, and BERTopic using contextual embeddings to represent text segments, and consequently grouping together expressions that are semantically related, irrespective of their superficial differences. A second metric, Topic Diversity (TD) was also calculated, to compare the lexical diversity of topics identified by the two algorithms. BERTopic reported a TD score of 0.7059, while LDA reported a score of 0.7000. The TD scores for both the models were very close to each other, implying that both were able to differentiate the topics produced by them reasonably well. And so for this corpus, the main advantage of BERTopic seems to be an improvement in the semantic coherence of generated topics, instead of a significant increase in lexical diversity.

3.1.1. Comparison at a common Topic Number

A possible limitation in comparing the two models in their default configurations is that they determine topic structures differently. LDA requires a predefined number of topics, whereas BERTopic uses embedding-based clustering to determine the number of clusters it finds. To reduce the influence of the topic granularity, an additional comparison was made with a common topic number. Obtained results represented in Table 2. The BERTopic model was reduced to a target of ten topics and resulted in nine non-outlier topics. LDA was subsequently trained with the same number of topics (K=9). Both models were evaluated with the same top-10-word representation and C_v coherence procedure.

Table 2. Comparison at a Common Topic Number (K = 9)

Metric	BERTopic	LDA
Number of topics	9	9
C_v coherence	0.6760	0.5777
Topic Diversity	0.7778	0.7444

BERTopic achieved a C_v score of 0.6760 compared to 0.5777 for LDA, with a relative difference of approximately 17.0%:

$$\frac{0.6760 - 0.5777}{0.5777} \times 100 \approx 17.0\% \tag{15}$$

Topic diversity was also higher for BERTopic (0.7778) than for LDA (0.7444), representing an approximate 4.5% relative improvement. The experiment thus strengthens the primary finding. The higher coherence of BERTopic cannot be explained only by the difference in the number of topics because of the same topic granularity. The result suggests that contextual semantic representation is advantageous in organizing the concepts present in the NCERT Chemistry corpus.

3.1.2. Robustness Analysis

In order to establish if the difference in performance was dependent on the particular random seed, both models were evaluated on 5 different random seeds (42, 52, 62, 72 and 82). The C_v scores obtained for each run are represented in Table 3.

BERTopic showed a range of values between 0.6418 and 0.7021, with a mean of 0.6808 $\pm$ 0.0214. The results for LDA ranged from 0.5823 to 0.6373, with a mean of 0.6086 $\pm$ 0.0217. Overall, BERTopic demonstrated a significantly better performance with a mean coherence value that was approximately 11.9% higher than the one demonstrated by LDA across the robustness experiment. Additionally, both algorithms showed a very similar variance, with the standard deviation across the runs being 0.0214 for BERTopic and 0.0217 for LDA. This implies that, on average, the results obtained for the two models were stable

Table 3. Five-seed robustness analysis

Model	MinC_v	MaxC_v	MeanC_v	SD
BERTopic	0.6418	0.7021	0.6808	0.0214
LDA	0.5823	0.6373	0.6086	0.0217

across different random seeds. The total number of topics identified by BERTopic ranged between 10 and 19, with a mean of 15.8 ± 3.06. The range of the number of topics is explained by the nature of the algorithm, which relies on the density of points in the embedding space to form clusters, as opposed to a fixed number of them as in LDA . As such, the deviation in the number of topics for BERTopic should be viewed as a deviation in the granularity of topics rather than a failure to capture semantic similarity. Lastly, while the robustness experiment was conducted on a corpus of documents related to chemistry textbooks, the results should not be extrapolated to other subjects, textbooks, languages or educational systems.

Table 4. Summary of BERTopic and LDA Evaluation Results

Evaluation	BERTopic	LDA
Primary C_v	0.6443	0.5414
Primary Topic Diversity	0.7059	0.7000
K = 9 C_v	0.6760	0.5777
K = 9 Diversity	0.7778	0.7444
Five-seed mean C_v	0.6808	0.6086
Five-seed SD C_v	0.0214	0.0217

The combined results of the experiment are displayed in Table 4. Three findings from this study are worth consideration: First, BERTopic's C_v consistently outperformed LDA's in the primary experiment, the common topic-number experiment, and the five-seed robustness analysis. Second, Topic Diversity was generally comparable in the primary experiment and was moderately better in the common topic-number experiment for BERTopic. Thus, Topic Diversity does not appear to be a distinguishing factor that drives BERTopic's advantage over LDA in terms of generating dramatically more diverse sets of lexically distinct topics. Third, the similar standard deviations found in the robustness analysis suggest that BERTopic's higher mean was not achieved by having dramatically more variable coherence scores across the five seeds. As a result, the primary conclusion from this study is that BERTopic provides stronger semantic coherence and better differentiated contextual groups than LDA on average, not that it consistently achieves dramatically more lexical diversity.

3.2 Qualitative Analysis

Qualitative evaluation further supports the benefits of BERTopic. As seen in Figure 3, the highest weight is given to the Concept of Topic 4 (Equilibrium), indicating that the concepts of Equilibrium have high lexical concentration and conceptual density in the curriculum. The model demonstrates high semantic coherence by identifying clusters of organic, inorganic and physical chemistry. The results confirm that the transformer-based topic modeling can effectively capture hierarchical curriculum frameworks. The BERTopic model is able to successfully capture the natural disciplinary structure of NCERT Chemistry curriculum by identifying distinct clusters corresponding to the categories of physical, organic and inorganic chemistry.

The model effectively recognized domain-specific clusters related to:

- Organic Chemistry (carbon, alcohol, alkyl groups)
- Chemical Bonding & Molecular Structure (bond, orbital, molecule, electron, overlap)
- Periodic Trends (element, transition, metal)
- Solutions (Physical Chemistry) (solution, pressure, solvent, vapour)
- Chemical equilibrium (equilibrium constant, concentration, pressure)
- Thermodynamics (system, heat, entropy)
- Electrochemistry (cell, electrode, conductivity)
- Atomic Structure & Quantum Theory (electron, wavelength, radiation, quantum)

On the other hand, the LDA model identifies ten topics. In this case, several topics were semantically overlapping and less interpretable. Many of the problems with LDA combined unrelated concepts such as thermodynamics, equilibrium and reaction kinetics. This suggested the limitations of the method in explaining contextual relations in the dataset. We also performed a

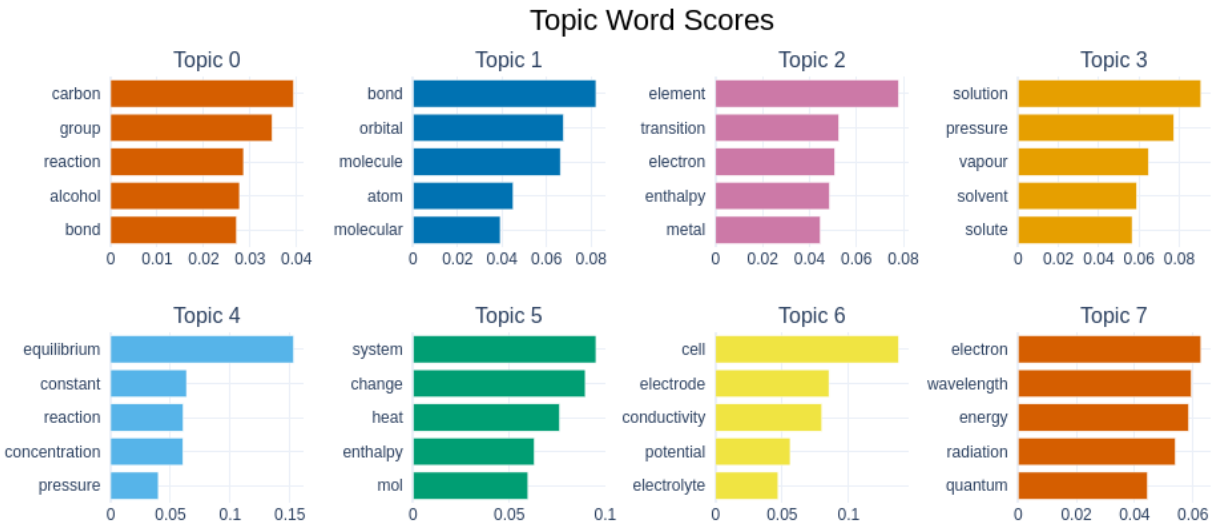


Fig 3. Topic Word Scores obtained by the BERTopic model

Comparative Evaluation of Topic Modeling Approaches as seen in Table 5, The LDA model yielded overlapping and semantically ambiguous topics. The BERTopic model produced clearly distinguishable and domain-specific clusters, demonstrating an improved contextual understanding and interpretability.

Table 5. Comparative Evaluation of Topic Modeling Approaches

Feature	BERTopic	LDA
Topic clarity	High	Low
Domain separation	Clear	Moderate
Noise Handling	Strong	Weak
Interpretability	High	Moderate
Context awareness	Yes (Transformer-based)	No (Bag-of-Words)

The quantitative evaluation of the BERTopic clustering outcome was followed by the qualitative analysis of topic representations. Specifically, the model was able to identify groups of domain-specific terms related to particular aspects of Chemistry. Indeed, the topics incorporated terms related to atomic structure and properties, chemical bonds, the chemical equilibrium, thermodynamics, electrochemistry, solutions, and organic chemistry (Figure 3).

Additionally, Figure 4 illustrates the topic-similarity matrix, which shows the degree of semantic similarity between the topics identified by BERTopic. Darker blue cells indicate higher similarity, whereas lighter cells indicate lower similarity. The matrix reveals stronger relationships between certain chemistry topics, reflecting conceptual connections among related subject areas. For example, the solution and equilibrium topics show a relatively strong association, while some topics, such as wavelength and radiation energy, exhibit weaker similarity with several other topics. This analysis provides a complementary view of the relationships among the discovered topics and helps assess their semantic differentiation. Overall, the qualitative analysis helps to validate the results of clustering by evaluating the topics' semantic representation. In combination with the coherence and diversity scores, the inspection of topics' words facilitates the interpretation of the clustering outcomes, thus validating the method for further use.

3.3 Comparison with Related Topic Modeling Studies

In order to locate the present results within the scope of previous research, the current results were compared with the results of representative studies in Table 6, considering conventional and contextualized topic-modeling approaches.

The comparison demonstrates a broadly consistent advantage of contextualized topic modeling across several datasets and application domains. Given that different corpora, preprocessing pipelines, topic numbers, and evaluation criteria were

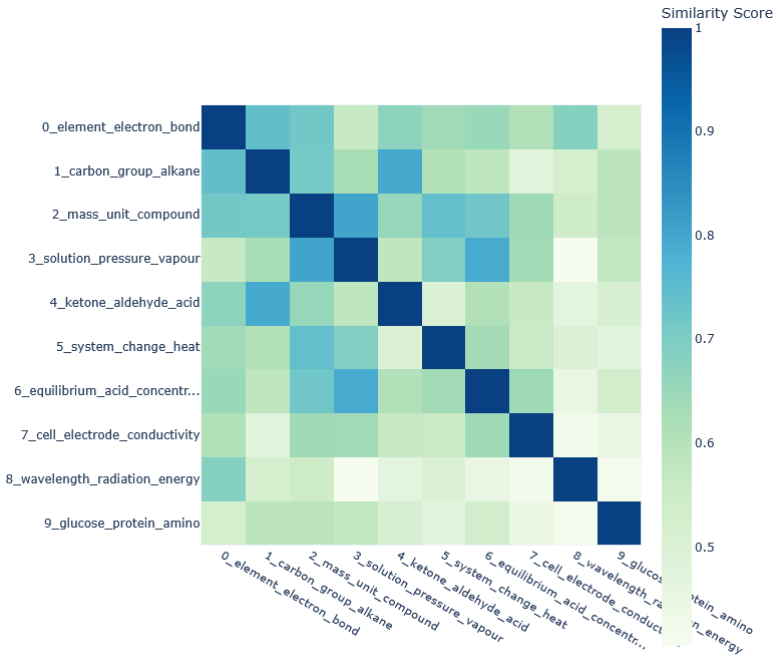


Fig 4. Topic-similarity heatmap of BERTopic-derived topics

considered in these papers, absolute coherence values should not be directly compared across studies . Therefore, the focus was placed on models' behaviors, evaluation methodologies, and reported findings with the current results.

Table 6. Comparison with existing topic-modeling studies

Study	Models	Evaluation	Main Finding	Present Study
Dieng et al.[10]	LDA, ETM	Coherence, Diversity	Embedding-space modeling improved topic quality in several settings.	Supports evaluating both semantic coherence and topic diversity.
Bianchi et al.[1]	LDA, Contextual models	Coherence, Diversity	Contextualized document representations produced more coherent topics than conventional representations.	Consistent with higher BERTopic coherence.
Terragniet al.[8]	Multiple models	Multiple topic quality metrics	Emphasized controlled and systematic evaluation.	Supports K= 9 comparison and multi metric evaluation.
Grootendorst[3]	BERTopic, LDA, others	Coherence, Diversity	BERTopic generally achieved stronger topic coherence than LDA.	Directly agrees with the higher BERtopic coherence observed here.
Yu et al.[15]	LDA	Topic analysis	LDA identified broad thematic structures in AI in education.	Demonstrates usefulness of topic modeling for educational corpora.
Oprea & Bâra[16]	LDA, BERT, BERT Clustering	Comparative thematic analysis	Contextual approaches produced more domain specific themes.	Consistent with present study BERTopic topic differentiation.
Present study	BERTopic, LDA	C_v, Diversity, Robustness	BERTopic consistently has higher coherence.	Extends comparison to NCERT school textbooks.

4 Conclusion

The present study introduces a novel approach towards curriculum intelligence via applying an AI-based topic modeling methodology to the domain of school education. Specifically, the research was conducted on the pages of Indian school NCERT Chemistry textbooks for classes XI and XII, comparing the performance of a transformer-based algorithm called BERTopic and a traditional probabilistic Latent Dirichlet Allocation (LDA). The study revealed that BERTopic exhibited better performance in terms of topic coherence, providing more differentiated and chemistry-specific topics than LDA. In the initial experiment, BERTopic demonstrated a higher coherence value (C_v) of 0.6443 versus 0.5414 for LDA, thus outperforming the traditional algorithm by almost 19%. Another benefit of using a contextualized encoder such as a BERT transformer lies in generating more diversified topics related to major fields within chemistry. The additional experiments conducted under the K=9 -based methodology confirmed the initial findings, with BERTopic providing a higher cohesion value (0.6760) compared to LDA (0.5777). Furthermore, the Topic Diversity metric was also better for BERTopic (0.7778) versus LDA (0.7444). Additionally, when experimenting with different random seeds, BERTopic again showed better performance with a higher mean of the C_v (0.6808) compared to LDA (0.6086). These results suggest that the utilization of contextual embeddings may be beneficial for conducting curriculum analysis in the field of chemistry.

In conclusion, the current study demonstrated the effectiveness of using a transformer-based encoder for conducting a more accurate and fine-grained analysis of the school curriculum. The findings of the study can be instrumental in developing AI-driven curriculum intelligence systems, supporting the analysis of curricula at multiple levels and helping align educational content with the latest educational frameworks and policies, including NEP 2020 and SDG 4. The next step in this research is to apply the methodology to larger and more diverse data sets, including multi-disciplinary curricula, and validate the results with experts in the field. The alternative research design could involve combining topic modeling with knowledge graphs to build a more comprehensive understanding of curriculum content and structure.

5 Acknowledgments

The authors have no acknowledgements to declare. This study didn't receive any specific funds.

Author details: Dr. Hiren Joshi serves as a Professor and Dr. Hardik Joshi as an Associate Professor.

Data availability: The data supporting the findings of this study are publicly available from the official National Council of Educational Research and Training (NCERT) website (https://ncert.nic.in/). No proprietary or restricted datasets were used.

References

1) Bianchi F, Terragni S, Hovy D. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. Online. Association for Computational Linguistics. 2021. https://doi.org/10.18653/v1/2021.acl-short.96.
2) Lamirel JC, Lareau F, Prouteau T, Malaterre C. Bag-of-words or embedding-based topic models? Comparing LDA, BERTopic, and six other methods. *Scientometrics*. 2026;131(9):6213–6237. https://doi.org/10.1007/s11192-026-05745-4.
3) Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*. 2022. https://doi.org/10.48550/arXiv.2203.05794.
4) Bond M, et al. A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*. 2024;21(1):4. https://doi.org/10.1186/s41239-023-00436-z.
5) Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*. 2019;16(1):39. https://doi.org/10.1186/s41239-019-0171-0.
6) López-Meneses E, Mellado-Moreno PC, Herrerías CG, Pelícano-Piris N. Educational Data Mining and Predictive Modeling in the Age of Artificial Intelligence: An In-Depth Analysis of Research Dynamics. *Computers*. 2025;14(2):68. https://doi.org/10.3390/computers14020068.
7) Romero C, Ventura S. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*. 2020;10(3):e1355. https://doi.org/10.1002/widm.1355.
8) Terragni S, Fersini E, Galuzzi BG, Tropeano P, Candelieri A. OCTIS: Comparing and Optimizing Topic models is Simple! Online. Association for Computational Linguistics. 2021. https://doi.org/10.18653/v1/2021.eacl-demos.31.
9) Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*. 2019;p. 1810.04805. https://doi.org/10.48550/arXiv.1810.04805.
10) Dieng AB, Ruiz FJR, Blei DM. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*. 2020;8:439–453. https://doi.org/10.1162/tacl_a_00325.
11) Egger R, Yu J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*. 2022;7:886498. https://doi.org/10.3389/fsoc.2022.886498.
12) Raman R, Pattnaik D, Lathabai HH, Kumar C, Govindan K, Nedungadi P. Green and sustainable AI research: an integrated thematic and topic modeling analysis. *Journal of Big Data*. 2024;11(1):55. https://doi.org/10.1186/s40537-024-00920-x.
13) Chang IC, Yu TK, Chang YJ, Yu TY. Applying Text Mining, Clustering Analysis, and Latent Dirichlet Allocation Techniques for Topic Classification of Environmental Education Journals. *Sustainability*. 2021;13(19):10856. https://doi.org/10.3390/su131910856.
14) Omarsaib M, Mitha SB, Vahed A, Mohamed GM. Mapping the AI Surge in Higher Education: A Bibliometric Study Spanning a Decade (2015–2025). *Informatics*. 2025;12(4):137. https://doi.org/10.3390/informatics12040137.

15) Yu JH, Chauhan D, Iqbal RA, Yeoh E. Mapping academic perspectives on AI in education: trends, challenges, and sentiments in educational research (2018–2024). *Educational Technology Research and Development*. 2025;73(1):199–227. https://doi.org/10.1007/s11423-024-10425-2.
16) Oprea SV, Bâra A. Transforming Education With Large Language Models: Trends, Themes, and Untapped Potential. *IEEE Access*. 2025;13:87292–87312. https://doi.org/10.1109/ACCESS.2025.3570649.