REGULAR ARTICLE

ANSABE?! The Role of Topic Modeling in Exploring Student Engagement and Curriculum Relevance During the Pandemic

Paula Joy Martinez*, Jason Catacutan and Warren May de la Cruz

*Correspondence:
pmartinez.msds2024@aim.edu
Aboitiz School of Innovation, Technology, and Entrepreneurship, Paseo de Roxas, Legazpi Village, 1229 Makati, Philippines
Full list of author information is available at the end of the article

Abstract

Topic modeling is a powerful tool for extracting themes from large text corpora, and this study explores the application of two prominent methods: Latent Dirichlet Allocation (LDA) and BERTopic. Using an academic learning platform as a case study, our analysis reveals that LDA effectively captures broad and overlapping topics, reflecting the diverse subjects registered within the platform. However, this method also exhibits limitations, such as the tendency for topics to overlap due to dimensionality reduction distortions. In contrast, BERTopic demonstrates its strength in capturing nuanced themes through its integration with contextual embeddings and clustering techniques, allowing for better interpretability and coherence. Despite BERTopic's versatility, LDA remains a viable option for simpler tasks, particularly when ease of use and established methods are prioritized. Our findings highlight the importance of thorough text preprocessing and underscore the need for hyperparameter tuning in BERTopic and grid search optimization for LDA. Future directions include enhancing BERTopic's performance through hyperparameter adjustments and optimizing LDA's Dirichlet priors to improve topic granularity and interpretability.

Keywords: topic modeling; latent dirichlet allocation; bertopic

1 Introduction and Objectives

1.1 Introduction

The COVID-19 pandemic has profoundly disrupted educational systems worldwide, forcing institutions to quickly adapt to remote and online learning environments. This sudden shift has raised important questions regarding the effectiveness and relevance of the curriculum delivered through digital platforms, as well as the level of student engagement in these new learning contexts. Understanding how students interact with academic content and the relevance of the topics discussed during this period is essential for assessing the success of remote education and guiding future educational strategies.

Topic modeling, a key technique in natural language processing (NLP), offers a powerful method for analyzing large volumes of textual data, such as student discussions and online forum posts. By identifying underlying themes and tracking their evolution over time, topic modeling can provide valuable insights into student engagement and curriculum relevance. This study applies two prominent topic modeling techniques—Latent Dirichlet Allocation (LDA) and BERTopic—to analyze student interactions on academic platforms during the pandemic. The primary

goal is to assess whether the topics introduced by teachers aligned with student needs and concerns, and to examine how student participation trends reflect the usability and effectiveness of these platforms.

The rationale behind this study lies in the need to understand the pandemic's impact on educational engagement and curriculum effectiveness. As the transition to online learning occurred abruptly, there is an urgent need to evaluate how well the educational content delivered during this period resonated with students and addressed the unique challenges posed by remote learning. Through a data-driven approach, topic modeling allows us to uncover valuable insights by analyzing the actual content of student discussions.

Additionally, comparing the effectiveness of two topic modeling techniques—LDA and BERTopic—enables a deeper understanding of how to best capture and interpret thematic structures within educational data. This comparison is particularly important given the different strengths and limitations of these methods: LDA offers simplicity and computational efficiency, while BERTopic provides a more dynamic and detailed analysis. The findings of this study have the potential to inform educators, policymakers, and institutions about the effectiveness of remote learning strategies, while contributing to the development of more responsive and relevant educational practices through data science.

1.2 Objectives

The primary objective of this research is to assess student engagement and the relevance of the curriculum on academic platforms during the COVID-19 pandemic, utilizing advanced topic modeling techniques. This study specifically aims to identify the dominant themes discussed among students in online academic forums during the pandemic by applying Latent Dirichlet Allocation (LDA) and BERTopic methods. Additionally, the study seeks to analyze the evolution of these topics over time, providing insights into how students' engagement with various themes fluctuated throughout the pandemic.

A key part of the study involves comparing the granularity and effectiveness of LDA and BERTopic in capturing the thematic structures within student discussions. This includes evaluating how LDA, with its static topic modeling, compares to BERTopic's dynamic capability to reflect changing topics over time. Moreover, the study intends to evaluate student participation trends, examining how effectively academic platforms facilitated student engagement during the pandemic. This analysis will highlight any potential gaps in the remote-learning infrastructure based on observed trends in student interaction.

The structure of this paper is as follows: Section 2 outlines the dataset and details the preprocessing techniques applied. Section 3 explains the methodology and the natural language processing pipeline used. Section 4 presents the findings and offers an in-depth analysis of the results. Section 5 provides concluding remarks and suggests potential directions for future research. Finally, Section 6 outlines the assumptions underlying this study.

2 Data Description and Preparation

2.1 Data Source

The data comes from an educational setting, potentially an online learning platform. The dataset consists of three primary dataframes:

1 `df_Entries`
Contains information about student entries or submissions. This may include details such as timestamps, entry content, and user IDs.
2 `df_grades`
Stores grade information, including user IDs, scores, and possibly the total points for each assignment.
3 `df_Rep`
Contains textual data, likely representing student responses or feedback. This dataframe also includes user IDs and the associated text.

Since this project focuses on topic modeling, only the `df_Rep` dataframe was used, as it contains the relevant text data. The columns of this dataframe, along with their descriptions, are listed in Table 1.

Table 1 Column Descriptions of df_Rep.

Column Name	Description
collected_at	The timestamp indicating when the data was logged or recorded by the system. This might differ from when the actual event occurred.
metadata_event_time	The exact time when the event (e.g., discussion entry creation) occurred, potentially before the data was logged.
metadata_event_name	The name of the event, such as discussion entry creation.
metadata_context_id	Identifier for the specific context of the event (e.g., course or discussion).
metadata_context_role	The role of the user in the context, such as student or teacher.
metadata_user_id	Unique identifier for the user associated with the event.
body_assignment_id	Identifier for the assignment associated with the event.
body_discussion_topic_id	Identifier for the discussion topic related to the event.
body_discussion_entry_id	Identifier for the specific discussion entry created.
body_submission_id	Identifier for the submission made by the user (if applicable).
body_user_id	Identifier for the user who made the entry or submission.
body_parent_discussion_entry_id	Identifier for the parent discussion entry (if it's a reply).
body_text	The content of the discussion entry or submission.

2.2 Data Preprocessing

This project involves extensive preprocessing steps to ensure the integrity of the dataset used for topic modeling.

2.2.1 Data Filtering

Data filtering primary involves the removal of invalid and irrelevant rows:

1 **Rows with No Text.** These rows are irrelevant to the topic modeling task. In the initial phase of data cleaning, rows devoid of textual content were identified and removed. Since topic modeling relies on analyzing text to discern patterns and themes, the absence of text renders these rows non-contributory. Retaining such rows would not only dilute the analytical process but also increase computational overhead without providing any insights.
2 **Duplicates.** Duplicates can distort the analysis by giving disproportionate weight to repeated data, leading to biased outcomes. For example, if a document is duplicated multiple times, the topic modeling might overemphasize

the themes within that document, skewing the overall topic distribution. To address this issue, all identified duplicates were removed to ensure that each data point contributed equally to the analysis. This process was applied both at the dataframe level (by removing rows that are exact copies) and at the column level (by eliminating texts that are exact copies).

3 **Text Preprocessing.** After removing invalid rows and duplicates, text preprocessing was applied to refine the data further. This step removed the following:

   (a) **Stopwords, Punctuation, and Other Tokens.** Stopwords, punctuation, emails, numbers, and URLs were removed from the text to prevent irrelevant tokens from skewing the analysis.

   (b) **Alphanumeric Strings.** Alphanumeric strings (e.g., "test123") can introduce noise into the data. Such tokens were filtered out to ensure a cleaner text corpus.

   (c) **Consecutive Duplicate Tokens.** Repeated adjacent tokens within a text (e.g., "hello hello") were identified and removed to further refine the quality of the input data.

4 **Non-English Texts.** The topic modeling analysis is focused exclusively on English texts, as texts in other languages may not be relevant and could introduce noise into the analysis. However, a potential future direction could involve isolating Filipino texts to provide additional context for the project.

5 **Obtain Top 10% of the Topics.** This reduces the noise from less prevalent topics. Focusing on the top 10% of topics allows the analysis to concentrate on the most dominant themes within the dataset. This mitigates the impact of noise that could be introduced by less prevalent topics, which may not be as relevant or insightful. By honing in on these dominant topics, the results are more likely to reflect the core patterns and discussions present in the data, thereby enhancing the overall accuracy and relevance of the findings. A more data-driven method, however, involves applying the Pareto principle: identifying the 20% of topics that account for 80% of the total counts.

   (a) **Improve Interpretability.** Interpreting every identified topic can be overwhelming, especially when the number of topics is large. Focusing on the top 10% of topics makes the analysis more manageable and comprehensible. This decision to filter down to the top 10% was driven by the need to enhance the interpretability of the results. When too many topics are considered, the analysis and interpretation can become convoluted and less meaningful, particularly when the topics are only marginally relevant. In this case, many topics contained only a single entry.

   (b) **Optimize Resources.** Filtering the data in this way also allows for more efficient use of computational resources. Analyzing only the top 10% of topics reduces the amount of processing power and time required. Large-scale topic modeling can be computationally intensive, and narrowing the scope of the analysis helps to optimize resource utilization. This approach not only enhances computational efficiency but also ensures that resources are directed towards analyzing the most critical aspects of the dataset, rather than expending effort on less impactful topics.

2.2.2 Heuristic Filtering

The purpose of heuristic filtering is to identify and exclude outlier documents that could potentially skew the analysis based on text length. This is because extreme variations in text length can significantly affect the performance and results of topic modeling algorithms.

For instance, exceptionally short texts may lack sufficient context and content, leading to inadequate representation of the underlying themes (See Table 2). Conversely, excessively long texts might dominate the analysis, overshadowing shorter but equally important contributions. By filtering out documents that fall outside a predefined length range, we aim to achieve a more balanced dataset where each document contributes meaningfully to the overall topic distribution.

Table 2 Sample of non-meaningful texts identified during data cleaning. These examples include short, repetitive, or otherwise trivial texts that were excluded from the analysis to ensure a more meaningful dataset for topic modeling.

Sample Texts
Good work!
well said.
True False
Awwwww! ¡3
Tiring ;-;
Thanks Ana
Thanks Ana
True False
Well said!
False True

The heuristic filtering process was conducted in two key stages: before and after the application of topic modeling.

1 **Stage 1: Token Space Outlier Detection.** Before applying any topic modeling techniques, we employed the Kullback-Leibler (KL) divergence, which measures how one probability distribution diverges from a second, reference distribution. In this context, the reference distribution is derived from the entire corpus of text lengths, representing the expected distribution of text lengths across the dataset. By applying this metric, we can identify documents with text lengths that significantly deviate from the norm established by the entire corpus, which could indicate anomalies or noise within the data. The removal of outliers was specifically carried out using the interquartile range (IQR) method. Documents with scores falling below the lower quartile or above the upper quartile were flagged as outliers and considered for exclusion from further analysis (Figure 1).

2 **Stage 2: Topic Space Outlier Detection.** Following the implementation of Latent Dirichlet Allocation (LDA) for topic modeling, we conducted a second round of heuristic filtering in the topic space. Here, KL divergence was again utilized, this time to measure the divergence of topic distributions within individual documents compared to the overall topic distribution in the corpus. To achieve this, we first computed the topic distribution vectors for each document and identified their nearest neighbors based on these distributions. We

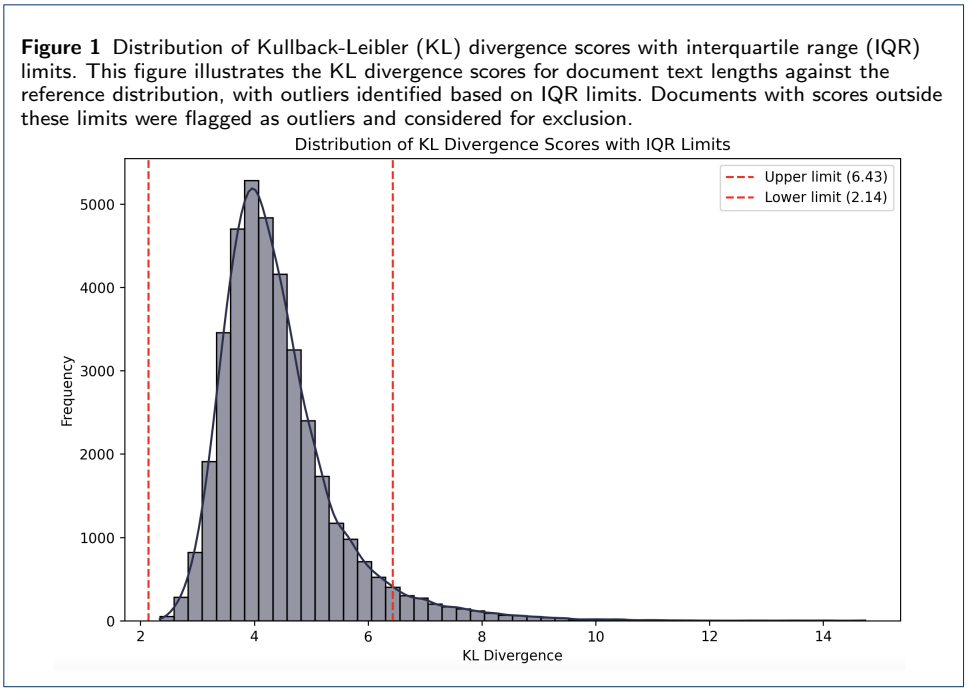


Figure 1 Distribution of Kullback-Leibler (KL) divergence scores with interquartile range (IQR) limits. This figure illustrates the KL divergence scores for document text lengths against the reference distribution, with outliers identified based on IQR limits. Documents with scores outside these limits were flagged as outliers and considered for exclusion.

then calculated the KL divergence between each document's topic distribution and the average topic distribution of its neighbors. This measure helped identify documents with extreme divergence scores, which were flagged as outliers. These outliers, determined by thresholds set through the interquartile range, were removed to ensure a more homogeneous dataset. This approach enhances the robustness and reliability of the topic modeling results by focusing on documents that align closely with the overall topic distribution, thereby improving the accuracy of the analysis.

3 Methodology

3.1 Latent Dirichlet Allocation

There were three classic topic models to choose from: Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA). Ultimately, LDA was chosen for its generative probabilistic framework, which allowed each text to be represented as a mixture of topics. This flexibility was particularly valuable given the complex nature of real-world documents. For instance, student essays on data science often touch on various subfields, such as natural language processing, data governance, and machine learning, which makes LDA's capability to model multiple topics per document a strong fit.

One of LDA's core strengths is that it models documents as a distribution over topics, and topics as distributions over words. This means it captures the nuanced overlap between topics and how they share terms, reflecting the complex relationships that exist in natural text. Additionally, LDA assigns a probability score to each topic per document, which allows for the identification of dominant themes as well as secondary ones. This probabilistic nature makes it adaptable to documents that aren't strictly about one subject but weave in multiple themes.

However, while LDA's probabilistic model provides significant flexibility, it also introduces challenges. Tuning the hyperparameters, such as the number of topics and the Dirichlet priors, can be difficult and greatly impacts the quality of the output. In this project, an asymmetric apriori strategy on the document-topic distribution was employed to capture the varying importance of topics across documents. This approach assigns a higher prior probability to dominant topics while allowing for less frequent topics to still be represented, better reflecting the natural imbalance found in real-world text corpora.

The same apriori strategy was applied to the topic-word distribution to account for the uneven contribution of words across topics. By assigning higher prior probabilities to certain words that are more likely to occur within a dominant topic, this method improves the model's ability to generate coherent topics. At the same time, less frequent but still meaningful words were preserved, ensuring that niche or specialized topics could still emerge. Domain expertise can further improve these strategies by setting actual values for the priors.

Finally, the number of predefined topics was determined using the coherence score, which evaluates the semantic similarity between the words within a topic. The coherence score helps assess the interpretability of the topics, with higher scores indicating more coherent and meaningful groupings of words. In this project, 50 iterations were evaluated for both LDA and BERTopic, each with a different number of topics. A good rule of thumb is to select the iteration with the highest coherence score. However, in this case, the highest coherence score only resulted in two topics (Figure 2) for LDA, which were too broad to capture the underlying themes of the dataset's conversations. Table 3 presents the keywords associated with one of these topics, illustrating the lack of granularity that makes it difficult to differentiate between distinct themes. To address this issue, the number of topics with the third highest coherence score, which was 32, was chosen. This adjustment led to a more nuanced and detailed representation of the topics, allowing for better differentiation between the various themes present in the dataset.

3.2 BERTopic

One significant drawback of LDA is its inability to capture temporal trends, which poses challenges in dynamic topic modeling. This limitation led to the adoption of BERTopic, which excels at tracking topic evolution over time and provides better topic representations. As will be discussed next, BERTopic not only addresses the temporal aspects but also outperforms LDA in terms of producing clearer and more interpretable topic distributions through its modular pipeline.

BERTopic was a recently released topic modeling library that leverages transformers and count-TF-IDF to create clusters for topic identification. Aside from ease of use, BERTopic's biggest strength lies in its modularity. It's best to think of the package not as an individual algorithm, but a pipeline that can be fully customized depending on the specific topic modeling use case. However, given the lack of domain expertise with the provided dataset, this study will utilize the default recommendations of the package author (see Figure 3), with exception to exploring additional representation techniques at the last step.

Figure 2 Coherence scores for different numbers of topics in LDA. The highest coherence score, which corresponds to only two topics, was too broad for meaningful analysis. Consequently, the number of topics with the third highest coherence score, which is 32, was selected for a more detailed and granular topic representation.

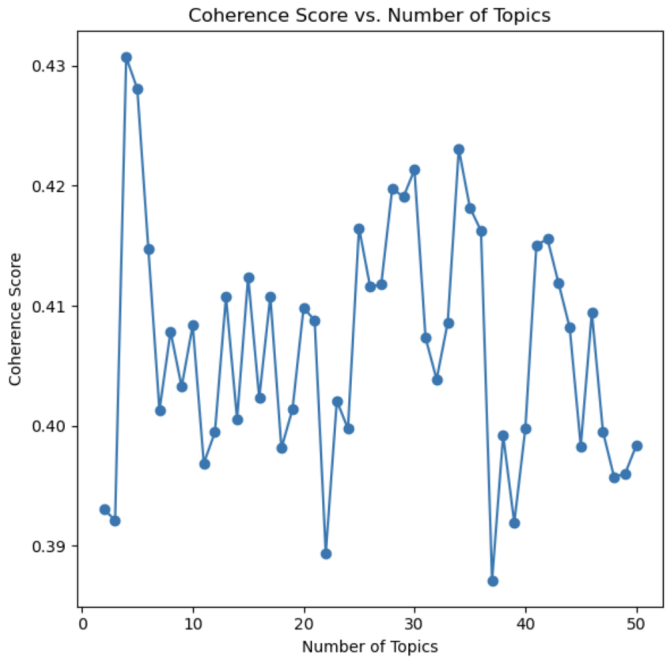


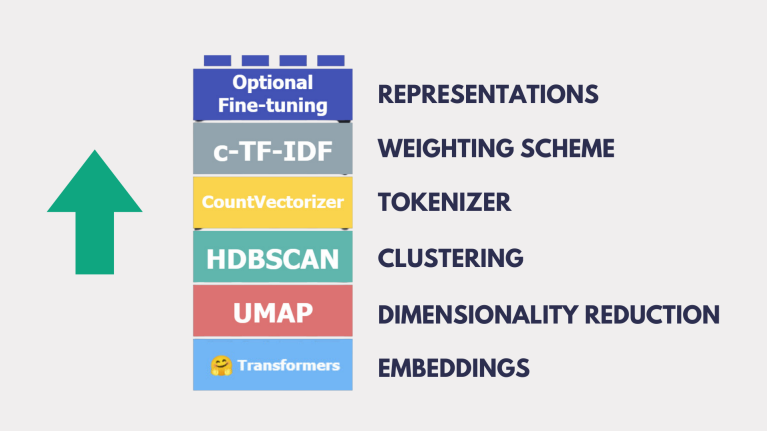
Table 3 Keywords associated with one of the topics, illustrating the lack of granularity that makes it difficult to differentiate between distinct themes.

Keyword	Weight
think	0.009
people	0.008
like	0.008
time	0.007
life	0.007
way	0.007
thing	0.006
love	0.005
experience	0.005
learn	0.005
find	0.005
different	0.005
know	0.005
feel	0.004
story	0.004

3.1.1 Text Representation - Embedding

The first step in BERTopic involves converting textual data into numerical representations that can be processed by machine learning algorithms. To achieve this, BERTopic typically uses sentence transformers, with the `all-MiniLM-L6-v2` model being the default choice. This model is well-regarded for its ability to capture se-

Figure 3 Overview of the BERTopic pipeline, illustrating its modular components and customization options. The pipeline integrates transformers and count-TF-IDF for topic identification. For this study, the default settings were used, with additional representation techniques explored in the final step.



mantic relationships between documents, enabling it to generate embeddings that reflect the meaning of the text rather than just the individual words.

3.1.2 Dimensionality Reduction

Once the text data is converted into high-dimensional embeddings, the next challenge is to manage the "curse of dimensionality," which can make clustering difficult and less effective. To address this, BERTopic employs Uniform Manifold Approximation and Projection (UMAP), a dimensionality reduction technique. UMAP helps to preserve both the local and global structures of the data, ensuring that the relationships between data points are maintained while reducing the number of dimensions.

3.1.3 Clustering

With the reduced-dimensional embeddings in place, BERTopic applies a clustering algorithm to group similar documents together. HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is typically used for this purpose. HDBSCAN is particularly adept at identifying clusters of varying densities, which is essential in NLP tasks where topics may not be uniformly distributed across documents. However, the choice of clustering algorithm can vary depending on the specific use case, and adjustments may be made to optimize the performance.

3.1.4 Tokenizer

Once the clusters are formed, BERTopic focuses on generating topic representations that are both understandable and interpretable. This step often involves the use of n-grams and the removal of stop words to refine the topic descriptions. BERTopic relies on a modified TF-IDF (Term Frequency-Inverse Document Frequency) approach, known as c-TF-IDF, to calculate word importance within each cluster. Unlike traditional TF-IDF, which focuses on document-level frequencies, c-TF-IDF

emphasizes the differences between clusters by treating each cluster as a single document. This method ensures that the most distinctive words for each topic are highlighted, leading to clearer and more meaningful topic representations.

3.1.5 Representations

After the initial topic representation, BERTopic offers multiple ways to refine the keywords and improve topic coherence. The primary method used is a combination of the Bag-of-Words model and class-based TF-IDF (c-TF-IDF), which helps emphasize distinctive terms for each topic. However, BERTopic also allows for the incorporation of advanced techniques such as KeyBERT-inspired keyword extraction, Part-of-Speech (POS) tagging using Spacy, and Maximal Marginal Relevance (MMR). These techniques enhance the quality of the keywords by ensuring they better capture the essence of the topics, thus offering a more precise understanding of the themes.

Unlike LDA, where keyword extraction is more rigid, these additional techniques in BERTopic provide multiple perspectives, allowing for more nuanced interpretations of topics. We also explored the integration of large language models (LLMs) to further enhance the topic refinement process. Specifically, the Llama 2 family of models was utilized to assess its ability to generate more contextually rich and accurate topic representations. This integration represents a significant step forward in leveraging LLMs to improve the interpretability and quality of topic modeling results.

4 Analysis and Findings

4.1 Topic Results of Latent Dirichlet Allocation

Even though the selected number of topics for LDA (32) provided topic distributions that are more interpretable and balanced in terms of both diversity and generality, compared to the suggestion of the coherence score (2), further analysis of these topics shows that there are significant overlaps and even topic misrepresentations. Figure 4 shows an instance of poor representation. The topic has apparent overlaps between topics encompassing vaccine development, the spread of the COVID-19 virus, and history. We surmise that this could be because the dataset only covers one month, so separate discourses revolving around the pandemic and history could really be limited, given that the nature of academic learning platforms is unpredictable and largely dictated by professors of certain classes. That is, if professors do not desire to talk about the pandemic or history, then those topics would not arise in the first place.

On the other hand, Figure 5 shows a more coherent and sensible topic generated by LDA, which is mainly about national governance in the Philippines covering different sectors such as the economy, migration, and education.

The top 6 topics based on `pyLDAvis` are shown in Table 4. The overall interpretation is that these topics are general subjects taken by students in academic institutions. However, factoring in the milieu of the dataset based on the covered timeframe (the year was 2021), certain topics related to the pandemic inevitably came about. The associated keywords for each topic are also provided in the Table 4.

Figure 4 Instance of poor topic representation in LDA. The figure illustrates a topic with significant overlap between discussions on vaccine development, the spread of the COVID-19 virus, and historical context. Such overlaps highlight challenges in achieving distinct and meaningful topic separation, potentially influenced by the limited temporal scope of the dataset and the specific academic focus of the platform.

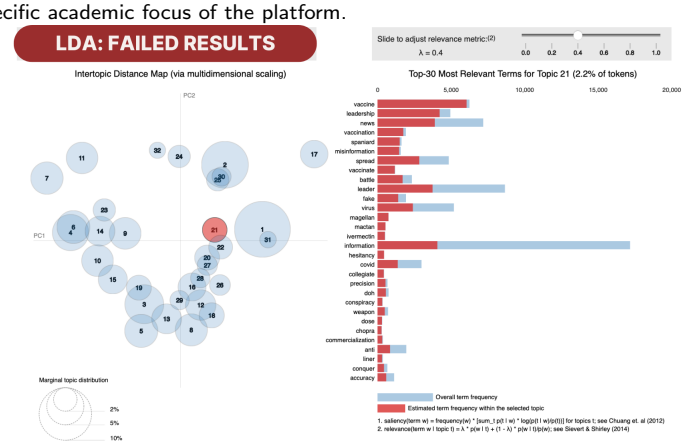
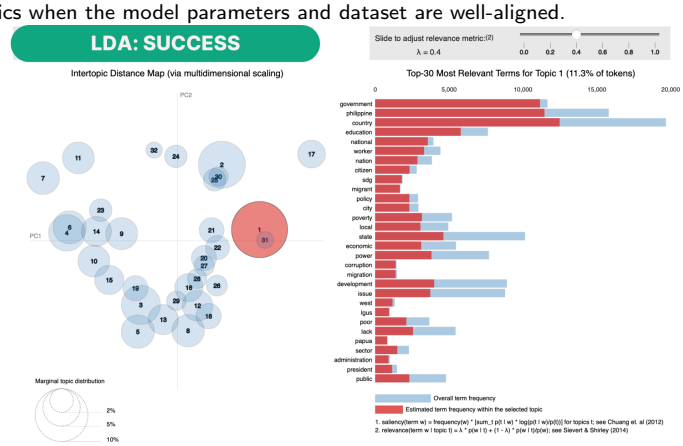


Figure 5 Example of a coherent and meaningful topic generated by LDA. The topic, primarily focused on national governance in the Philippines, encompasses various sectors such as the economy, migration, and education. This demonstrates LDA's ability to produce insightful and distinct topics when the model parameters and dataset are well-aligned.



4.2 Topic Results of BERTopic

4.2.1 Topics across the Entire Timeframe

For BERTopic, while the coherence scores suggest that 43 topics was optimal (Figure 6), further examination using Intertopic Distance Maps and Hierarchical Clustering revealed that 9 topics provided a better balance between coherence and interpretability (Figure 7). This decision led to more meaningful visualizations and overall improved topic modeling results.

Hence, we selected a more concise set of 9 topics that still exhibited strong coherence, and the outcomes from both visualizations demonstrated enhanced segmentation. Figures 8 and 9 shows the Hierarchical Clustering and Intertopic Distance Map of this adjustment, respectively.

As previously mentioned, additional representation techniques were utilized to enhance cluster representation and intepretability (see Figure 10). This includes

Table 4 Top keywords for each identified topic in LDA.

Topic	Keywords and Weights
Philosophy & Ethics	0.032*"philosophy" + 0.011*"think" + 0.011*"question" + 0.009*"life" + 0.009*"idea" + 0.009*"society" + 0.009*"labor" + 0.008*"truth" + 0.007*"understand" + 0.007*"system" + 0.007*"world" + 0.007*"answer" + 0.006*"believe" + 0.006*"way" + 0.006*"people"
Vaccine Development	0.019*"vaccine" + 0.014*"people" + 0.012*"news" + 0.012*"leadership" + 0.011*"leader" + 0.010*"information" + 0.006*"battle" + 0.005*"spread" + 0.005*"spaniard" + 0.005*"think" + 0.005*"vaccination" + 0.005*"team" + 0.005*"philippine" + 0.004*"misinformation" + 0.004*"medium"
PH Gov't and Current Events	0.019*"country" + 0.016*"philippine" + 0.016*"government" + 0.008*"people" + 0.008*"education" + 0.007*"development" + 0.007*"state" + 0.006*"poverty" + 0.006*"economic" + 0.006*"right" + 0.005*"need" + 0.005*"power" + 0.005*"national" + 0.005*"worker" + 0.005*"issue"
Arts and Literature	0.039*"art" + 0.013*"like" + 0.012*"song" + 0.012*"painting" + 0.010*"piece" + 0.010*"music" + 0.009*"artist" + 0.008*"love" + 0.008*"time" + 0.008*"artwork" + 0.007*"look" + 0.007*"work" + 0.007*"see" + 0.007*"think" + 0.007*"feel"
Biology	0.034*"cell" + 0.017*"animal" + 0.012*"plant" + 0.010*"organism" + 0.008*"function" + 0.006*"water" + 0.006*"food" + 0.005*"different" + 0.005*"form" + 0.005*"life" + 0.005*"body" + 0.005*"sep" + 0.005*"human" + 0.004*"access" + 0.004*"membrane"
Filipino and Pop Culture	0.029*"culture" + 0.012*"energy" + 0.012*"like" + 0.012*"food" + 0.010*"pop" + 0.008*"jollibee" + 0.008*"high" + 0.007*"filipino" + 0.007*"eat" + 0.006*"popular" + 0.006*"k" + 0.006*"mcdonald" + 0.006*"people" + 0.006*"consumer" + 0.005*"taste"

KeyBERT, MMR, and POS from Spacy. Additionally, Llama 2 - 7b was leveraged to automate topic labelling (see Table 5). However, while automated topic labeling with LLMs can expedite the process, it is not a panacea. Topic 8, for example, is better described as encompassing broader physical activity or education rather than solely street dance. This underscores the importance of familiarity with the data, additional representations, and fine-tuning prompts.

4.2.2 Dynamic Topic Modeling

BERTopic's dynamic topic modeling feature enables the exploration of how topics evolve over time. As illustrated in Figures 11 and 12, many topics have declined

Figure 6 Coherence scores and visualizations for different numbers of topics under BERTopic. Although an initial analysis suggested 43 topics, subsequent evaluation using Intertopic Distance Maps and Hierarchical Clustering indicated that 9 topics (Figure 6) offered a better balance between coherence and interpretability. This refinement led to more meaningful visualizations and enhanced topic modeling results.

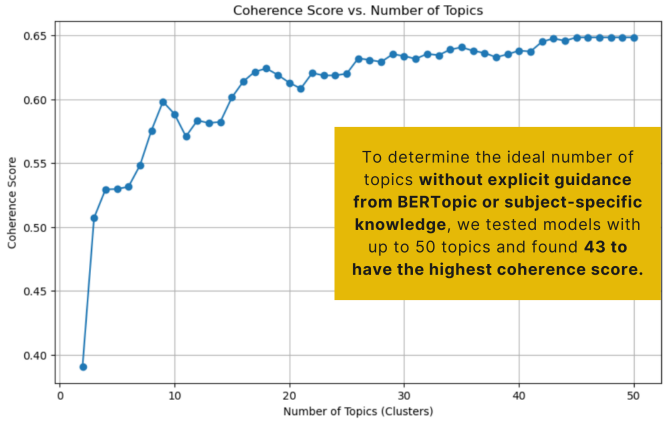


Figure 7 Our intertopic distance map revealed that many topics were clustered closely together. This observation was confirmed by hierarchical clustering, which demonstrated the formation of distinct topic groups.

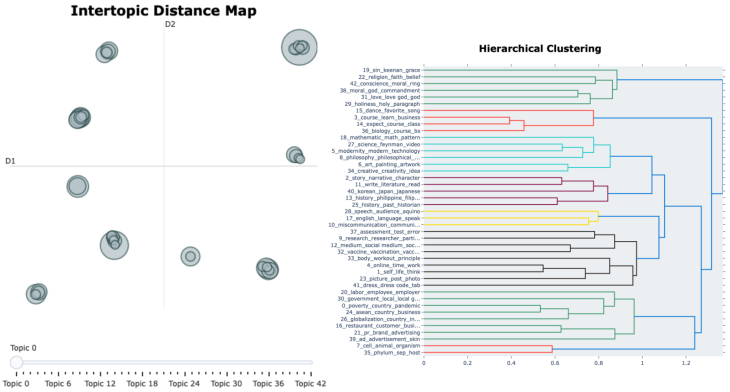


Table 5 Refined topic labels resulting from enhanced cluster representation techniques, including KeyBERT, Maximal Marginal Relevance (MMR), Part-of-Speech tagging (POS) from Spacy, and automated topic labeling using Llama 2 - 7b.

Topic Number	Topic Label
1	Creative Thinking and Problem Solving
2	Philippine Local Government: Balancing Autonomy and National Interest
3	Pandemic Vaccine Hesitancy
4	Online Learning Challenges
5	God and Love
6	Cell Biology
7	Labor Law and Obligations
8	Street Dance
9	Fitness Principles and Individualization

Figure 8 Hierarchical Clustering visualization of the refined 9-topic model. This adjusted clustering demonstrates improved segmentation and clearer topic differentiation compared to the initial model.

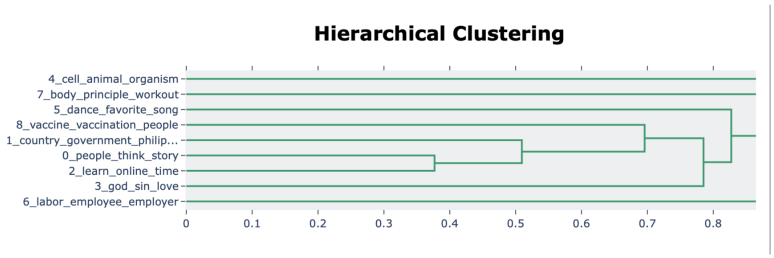


Figure 9 Intertopic Distance Map for the refined 9-topic model. The map illustrates the spatial distribution of topics, showcasing enhanced separation and coherence among the topics after the adjustment.

in prominence throughout the dataset, potentially indicating a loss of interest or engagement during this period.

4.2.3 Comparison of Topics between LDA and BERTopic

Building on the earlier comparison, Latent Dirichlet Allocation (LDA) proved effective in identifying the common classroom subjects typically discussed in academic learning platforms. Its strength lies in capturing the core, expected topics that dominate such environments, providing a reliable baseline for understanding the primary themes. However, when applied to more nuanced or emerging discussions, LDA tends to show its limitations. It often focuses on broad, overlapping topics that, while relevant, might not fully encapsulate the diversity of conversations happening within a dynamic student body.

On the other hand, BERTopic goes beyond this foundational level by offering a more granular and diverse clustering of texts. Unlike LDA, BERTopic is capable of uncovering less obvious, emerging topics that reflect the evolving interests

Figure 10 BERTopic's modularity enables the simultaneous use of various representation models within its pipeline, offering a broader understanding of the underlying topics. Additionally, we leveraged LLMs (Llama 2 - 7b) to enhance the topic labeling process.

	Topic	Count	Name	CustomName	Representation	KeyBERT	MMR	Llama2	POS	Representative_Docs
0	-1	15551	-1_people_think_time_like	people - think - time - like - life - way - co...	[people, think, time, like, life, way, country...	[pandemic, government, society, health, philip...	[think, time, like, life, country, experience,...	[Philippines Pandemic Response, , , , , , , ,]	[people, time, life, way, country, work, able,...	[window peer philippine governance politic pan...
1	0	10739	0_think_people_story_like	think - people - story - like - way - life - p...	[think, people, story, like, way, life, philos...	[philosophy, concept, social medium, thinking,...	[life, philosophy, art, experience, social, me...	[Creative Thinking and Problem Solving, , , , ...	[think, people, story, way, life, philosophy, ...	[creative thinking process involve science acc...
2	1	2985	1_country_philippine_government_filipino	country - philippine - government - filipino -...	[country, philippine, government, filipino, po...	[local government, government, national goverm...	[philippine, filipino, poverty, education, dev...	[Philippine Local Government: Balancing Autono...	[country, philippine, government, filipino, po...	[thing come mind maximum word hear word philip...
3	2	2977	2_learn_online_time_course	learn - online - time - course - class - math ...	[learn, online, time, course, class, math, mat...	[online learning, study, assessment, student, ...	[online, math, expect, work, student, study, a...	[Online Learning Challenges, , , , , , , ,]	[learn, online, time, course, class, math, mat...	[look mathematic regard real world function in...
4	3	1522	3_god_sin_love_religion	god - sin - love - religion - moral - love god...	[god, sin, love, religion, moral, love god, li...	[relationship god, theology, god love, christi...	[sin, religion, moral, love god, conscience, h...	[God and Love, , , , , , , ,]	[sin, love, religion, moral, life, keenan, goo...	[thought and/or question keenan article sin ke...
5	4	847	4_cell_animal_organism_plant	cell - animal - organism - plant - energy - me...	[cell, animal, organism, plant, energy, membra...	[animal cell, cell, cell wall, cell membrane, ...	[cell, organism, membrane, animal cell, access...	[Cell Biology, , , , , , , ,]	[cell, animal, organism, plant, energy, membra...	[main difference include animal cell have cent...
6	5	687	5_pandemic_vaccine_people_god	pandemic - vaccine - people - god - government...	[pandemic, vaccine, people, god, government, t...	[vaccine, vaccination, vaccinate, pandemic, qu...	[pandemic, vaccine, government, virus, vaccina...	[Pandemic Vaccine Hesitancy, , , , , , , ,]	[pandemic, vaccine, people, government, time, ...	[speed current vaccine develop see unprecedent...
7	6	528	6_labor_obligation_employee_employer	labor - obligation - employee - employer - law...	[labor, obligation, employee, employer, law, b...	[labor law, labor code, article labor, employm...	[obligation, employee, employer, beer, labor l...	[Labor Law and Obligations, , , , , , , ,]	[labor, obligation, employee, employer, law, b...	[article section cover aspect come labor inclu...
8	7	448	7_dance_favorite_watch_pronoun	dance - favorite - watch - pronoun - dancing -...	[dance, favorite, watch, pronoun, dancing, son...	[street dance, dance style, dance class, dance...	[dance, favorite, dancing, street dance, music...	[Street Dance, , , , , , , ,]	[dance, favorite, watch, pronoun, dancing, son...	[know background street dance think street dan...
9	8	209	8_body_principle_workout_exercise	body - principle - workout - exercise - fitnes...	[body, principle, workout, exercise, fitness, ...	[principle training, fitness goal, fitness, wo...	[workout, exercise, fitness, routine, principl...	[Fitness Principles and Individualization, , , ...	[body, principle, workout, exercise, fitness, ...	[principle training stand principle individual...

Figure 11 BERTopic's dynamic topic modeling feature enables us to track the evolution of topics over time.

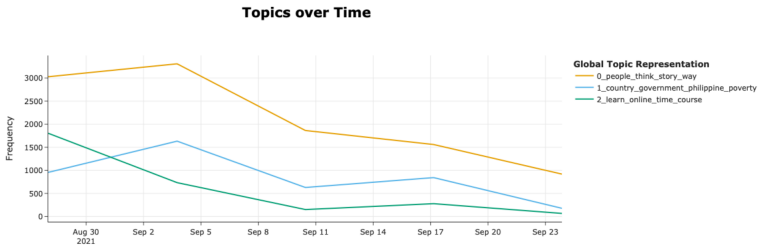
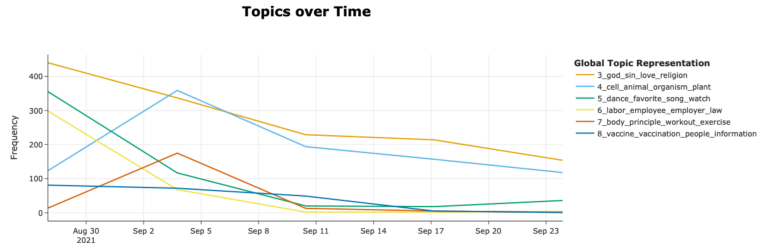


Figure 12 BERTopic's dynamic topic modeling feature enables us to track the evolution of topics over time.



and concerns of students. This model does not just stick to the traditional subject boundaries; it digs deeper into the context, revealing subtopics and new themes that might not be immediately apparent with LDA. This makes BERTopic particularly valuable in environments where understanding the full spectrum of student discourse is crucial, as it provides richer insights into the shifting landscape of student engagement and thought.

Furthermore, when it comes to temporal topic modeling, LDA falls short. While it's technically possible to use LDA to analyze topics over time by manually segmenting data into specific timeframes, this approach is labor-intensive and unsustainable for long-term analysis. LDA's static nature means it does not adapt easily to changes over time, making it less suitable for ongoing, dynamic environments where topics are continually evolving. In contrast, BERTopic's flexibility and ability to model topics temporally without extensive manual intervention make it a far superior choice for capturing how discussions and themes develop over time. This is particularly important in academic settings, where the relevance of topics can shift rapidly in response to new information, events, or trends.

4.2.3 Comparison Beyond Topics

Beyond the topics themselves, there are additional aspects in which LDA and BERTopic can be compared. These are summarized in Table 6.

5 Conclusion and Recommendations

5.1 Topic-specific Conclusions

1. LDA generates broad and overlapping topics Based on the visualization provided by `pyLDAvis`, LDA was able to capture the natural dynamics of an academic learning platform, i.e., its dependence on the subjects registered within the platform. While we anticipated this outcome, one notable observation is that the topics consist of a few distinct themes, while the rest tend to overlap. For example, there are two topics that seem to focus on ethics and religion. However, one appears to be a more specific subset of the other, with the former concentrating on religion and the latter covering more general literature. This overlap, observed across other topics as well, highlights some limitations of `pyLDAvis`. It is important to remember that the visualizations are projected onto a lower-dimensional space, which can introduce distortions. As a result, the perceived overlaps might be misleading and may not exist in higher-dimensional representations. Thus, there is merit to examining the top keywords of each topic rather than relying solely on the intertopic distance map.

2. BERTopic is perfect for experts

The BERTopic model effectively captures the nuances of the academic learning platform, accurately identifying key themes aligned with registered subjects. While some topics were well-defined, others exhibited overlap, highlighting the importance of user input and domain expertise. As demonstrated, adjusting the number of topics can significantly improve coherence and interpretability, a feature designed into BERTopic to facilitate customization for specific use cases.

Table 6 Comparison of Latent Dirichlet Allocation (LDA) and BERTopic across various aspects such as ease of use, topic granularity, interpretability, and computational efficiency. This comparison highlights the strengths and limitations of each method.

Aspect	**Latent Dirichlet Allocation (LDA)**	**BERTopic**
Ease of Use and Versatility	- **Implementation:** Well-established, widely used, many libraries available (e.g., Gensim, Scikit-learn). - **Learning Curve:** Moderate, requires understanding of probabilistic models and parameter tuning. - **Customization Opportunities:** Limited to parameters like number of topics, alpha, and beta.	- **Implementation:** Can be more complex due to dependencies on BERT-based embeddings. - **Learning Curve:** Steeper, involves understanding of contextual embeddings and clustering. - **Customization Opportunities:** High, allows for tuning embedding models, clustering parameters, and additional text preprocessing such as dimensionality reduction.
Topic Granularity and Interpretability	- **Granularity:** Fixed number of topics, which can be adjusted through the parameters. - **Interpretability:** Moderately interpretable, but can be challenging to decipher topic meanings, especially with large datasets.	- **Granularity:** Flexible, as the number of topics can adapt based on clustering results. - **Interpretability:** Generally better due to the use of contextual embeddings, which capture nuanced meanings, and availability of other topic representations such as KeyBERT, Maximal Marginal Relevance (MMR), and Part-of-Speech (POS).
Computational Efficiency	Generally efficient for smaller datasets but can become slow with larger datasets or many iterations. Computational load depends on the number of topics and corpus size.	Can be computationally intensive due to BERT embeddings and clustering. Efficiency can be lower with large corpora and complex models.

5.2 Implementation-Specific Conclusions

1. If it ain't broke, don't fix it

While BERTopic is highly effective and versatile, LDA is often more than adequate for simpler, less complex topic modeling tasks. LDA's integration with tools like `pyLDAvis` makes it easy to visualize and interpret topics, which is especially valuable for teams or projects with limited experience in topic modeling. For static, non-high-impact projects where the primary goal is straightforward topic discovery, we recommend sticking with LDA. It's tried and tested, and its ease of use makes it a solid choice when there's little familiarity with more advanced models like BERTopic.

2. Preprocessing is Key

Regardless of the model used, proper text preprocessing is crucial for meaningful results. LDA, in particular, is sensitive to the quality of input data, so traditional steps like tokenization, stopword removal, and lemmatization are essential. BERTopic, on the other hand, benefits from the contextual understanding provided by transformer-based embeddings but still performs better with clean data. We emphasize the importance of thorough preprocessing to ensure that noise does not dilute the coherence and quality of topics.

5.3 Future Directions

1. Hyperparameter Tuning

Future work should focus on hyperparameter tuning for BERTopic modules to refine topic extraction and improve model performance. This includes experimenting with different embedding models and clustering algorithms to enhance topic coherence and relevance.

2. Grid Search for LDA

Performing a grid search to optimize the Dirichlet priors in LDA is another promising direction. By systematically varying the Dirichlet parameters and evaluating their impact on topic coherence, more precise and interpretable topics can be achieved, potentially improving overall model performance.

6 Appendix: Assumptions of the Study

Several assumptions underpin this study, categorized into three key areas: pre-processing and post-processing, topic modeling, and evaluation and analysis.

1. **Pre-processing and Post-processing:** This study assumes that invalid or irrelevant data will be identified and removed during pre-processing, ensuring the dataset's integrity and relevance. The dataset is expected to contain both common and rare words, with the understanding that rare words may contribute less significantly to topic modeling. Given limited compute resources, the complexity and scale of the modeling approach may be affected. It is assumed that students are the primary actors in the dataset, with their responses and feedback forming the core of the analysis.
2. **Topic Modeling:** In terms of topic modeling, it is assumed that each document may contain multiple topics, reflecting the multifaceted nature of student responses. Furthermore, topic distributions are likely to be asymmetric, with some topics potentially being more prevalent than others, mirroring natural variations in topic frequency within the dataset.
3. **Evaluation and Analysis:** For evaluation and analysis, the study assumes that the data reflects the unique experiences of students during the COVID-19 pandemic, contextualizing the findings within this specific timeframe. Consequently, patterns and topics identified may be influenced by pandemic-related circumstances, requiring careful interpretation within this context.

These assumptions provide a framework for the study's methodology and guide the interpretation of results, acknowledging both the strengths and limitations of the approach.