big data and cognitive computing

MDPI

Article

Evaluating the Performance of Topic Modeling Techniques with Human Validation to Support Qualitative Analysis

Julian D. Romero [1], Miguel A. Feijoo-Garcia [2], Gaurav Nanda [1], Brittany Newell [1] and Alejandra J. Magana [2,*]

[1] School of Engineering Technology, Purdue University, 401 N. Grant St., West Lafayette, IN 47907, USA; romerorj@purdue.edu (J.D.R.); gnanda@purdue.edu (G.N.); bnewell1@purdue.edu (B.N.)
[2] Department of Computer and Information Technology, Purdue University, 401 N. Grant St., West Lafayette, IN 47907, USA; mfeijoog@purdue.edu
* Correspondence: admagana@purdue.edu

Abstract: Examining the effectiveness of machine learning techniques in analyzing engineering students' decision-making processes through topic modeling during simulation-based design tasks is crucial for advancing educational methods and tools. Thus, this study presents a comparative analysis of different supervised and unsupervised machine learning techniques for topic modeling, along with human validation. Hence, this manuscript contributes by evaluating the effectiveness of these techniques in identifying nuanced topics within the argumentation framework and improving computational methods for assessing students' abilities and performance levels based on their informed decisions. This study examined the decision-making processes of engineering students as they participated in a simulation-based design challenge. During this task, students were prompted to use an argumentation framework to articulate their claims, evidence, and reasoning, by recording their informed design decisions in a design journal. This study combined qualitative and computational methods to analyze the students' design journals and ensured the accuracy of the findings through the researchers' review and interpretations of the results. Different machine learning models, including random forest, SVM, and K-nearest neighbors (KNNs), were tested for multilabel regression, using preprocessing techniques such as TF-IDF, GloVe, and BERT embeddings. Additionally, hyperparameter optimization and model interpretability were explored, along with models like RNNs with LSTM, XGBoost, and LightGBM. The results demonstrate that both supervised and unsupervised machine learning models effectively identified nuanced topics within the argumentation framework used during the design challenge of designing a zero-energy home for a Midwestern city using a CAD/CAE simulation platform. Notably, XGBoost exhibited superior predictive accuracy in estimating topic proportions, highlighting its potential for broader application in engineering education.

Keywords: argumentation framework; topic modeling; machine learning; qualitative analysis; natural language processing

Citation: Romero, J.D.; Feijoo-Garcia, M.A.; Nanda, G.; Newell, B.; Magana, A.J. Evaluating the Performance of Topic Modeling Techniques with Human Validation to Support Qualitative Analysis. *Big Data Cogn. Comput.* **2024**, *8*, 132. https://doi.org/10.3390/bdcc8100132

Academic Editor: Carson K. Leung

Received: 26 July 2024
Revised: 8 September 2024
Accepted: 10 September 2024
Published: 8 October 2024

1. Introduction

Integrating technology into engineering education offers substantial benefits that can significantly improve teaching and learning processes. Research consistently shows that when technology is effectively incorporated into educational practices, it can enhance student engagement, boost knowledge retention, and promote the development of higher-order thinking skills [1]. One important way that technology contributes is through its application in assessment, particularly in evaluating students' written responses. While computer-based assessments are widely used for straightforward formats such as multiple-choice or true/false questions, analyzing open-ended text responses poses a greater challenge. This difficulty arises from the complexity and volume of unstructured data inherent

in such responses. As a result, effective analysis requires advanced methods and tools to manage large amounts of unstructured text data. Therefore, this study contributes to analyzing and comparing analytical methods to help effectively computationally determine students' abilities and performance levels. That is, by improving the use of different computational tools to assess students' written responses (i.e., argumentation), this study aims to make the evaluation of their abilities and performance more accurate and efficient. This focus on analytical methods seeks to enhance the efficiency and accuracy of assessing student learning outcomes [1].

In social sciences, automated text analysis methods are increasingly leveraged to extract and categorize information from massive text collections. Methods related to natural language processing (NLP) have demonstrated the utility of topic models in extracting actionable insights from large volumes of text data in various contexts, including health education research and analyzing student responses [2,3]. These methods aid in mining words and identifying hidden semantic structures. Additionally, supervised learning schemes, which involve acquiring text datasets, extracting emotional features, and training classification models, highlight the importance of these methods in text classification tasks [4].

Previous studies on natural language processing (NLP) have implemented either supervised or unsupervised approaches. The existing literature lacks strong qualitative validation and has not systematically distinguished between these methods regarding their relative coverage and reliability. Thus, this manuscript compares and establishes reliability and validity by examining topic modeling techniques following unsupervised and supervised approaches. On the unsupervised lens, it was particularly considered latent Dirichlet allocation (LDA), which supports text analysis by providing nuanced categories based on natural groupings of topics, aiding in interpretation and refinement, thereby improving the overall evaluation process [3]. Additionally, leveraging qualitative analysis as training data in supervised machine learning models can automate and optimize topic modeling, improving efficiency. After quantitatively evaluating the predictions made by both the unsupervised and supervised approaches, the researchers systematically and semantically compared and interpreted the main topics predicted by each method focusing on assessing the accuracy and relevance of their results.

This study focuses on the context of engineering design by analyzing design journals from first-year engineering students. Documentation of their decision-making processes was recorded using an argumentation framework [5,6]. The students were tasked to design a zero-energy home using Aladdin, an integrated CAD/CAE platform that facilitates simulation-based learning. The aim was to help students apply their scientific knowledge to make informed trade-off decisions [7,8]. The argumentation framework helped students to articulate their claims, evidence, and reasoning supporting their design decisions in the form of an argument. Educationally speaking, the arguments describing design decisions need to be evaluated to identify if they are grounded in scientific principles. From a human perspective, it is difficult to evaluate hundreds of arguments at a time. Therefore, there is an opportunity to use computational methods to extend human capabilities in evaluating text data [9]. Thus, this study combines qualitative and computational methods to analyze student design journals and validate the findings through expert analysis. Hence, the guiding research questions for this study are as follows: How do supervised and unsupervised approaches compare in terms of coverage and computational efficiency for topic modeling tasks? How do supervised and unsupervised approaches compare in terms of interpretability for topic modeling tasks?

1.1. Topic Modeling as a Qualitative Research Approach

While quantitative research methods emphasize numerical data and statistical analysis to quantify relationships and patterns [10], qualitative research methods delve deeper into analyzing and understanding participants' experiences, perspectives, and the meanings they attribute to various aspects within the context of their social and cultural norms [11].

To date, different studies have discussed efforts to enhance transparency in reporting qualitative research synthesis, by developing and utilizing techniques such as meta-ethnography, thematic analysis, and critical interpretive synthesis [12]. For instance, researchers have emphasized refining meta-ethnography methods to enhance their relevance and effectiveness in qualitative synthesis, incorporating new insights to improve the quality and impact of such studies [13]. Furthermore, researchers have integrated qualitative techniques such as thematic analysis with topic modeling, to enhance identifying latent themes or topics within a large corpus of text data, improving the efficiency and effectiveness of qualitative analysis by automating the process of theme identification and extraction across various disciplines [14–16].

1.2. Machine Learning Models for Topic Modeling

Topic modeling serves as a method in natural language processing (NLP) and machine learning to identify prominent underlying topics within qualitative data. It involves automatically identifying thematic information in extensive text collections for data mining purposes, emphasizing co-occurring words [17–19]. In fact, topic modeling has been involved in different contexts, including text analysis, image annotation, and sentiment classification, among others [20,21]. Common techniques to address topic modeling include latent Dirichlet allocation (LDA), latent semantic analysis (LSA), and non-negative matrix factorization (NMF) [22]. For instance, LDA—a probabilistic generative model—represents documents as mixtures of topics and topics as mixtures of words [23], which has been widely used in education to extract topics from resources, uncovering latent themes and patterns in students' thinking and decision-making processes [24,25].

As supervised and unsupervised machine learning models have supported text mining and NLP tasks, they have been utilized to address topic modeling approaches. First, supervised models establish relationships between predictors and a target variable [26]. In particular, in topic modeling, supervised methods like convolutional neural networks (CNNs), long short-term memory (LSTM), support vector machines (SVM), and Naïve Bayes, have been developed using annotated data [27,28]. Despite the ability of supervised models to categorize topics within textual data, they require significant labeled data to avoid overfitting and perform well on "unseen" data, leading to the consideration of semi-supervised or unsupervised approaches [27,29–31]. In fact, previous studies have recognized Ensemble Learning as a powerful technique for improving text classification performance by combining the outputs of multiple classifiers, revealing the effectiveness of "ensemble methods" in enhancing accuracy and generalization ability in text classification tasks [32–34]. For example, XGBoost has been known for its high performance and efficiency in text analysis tasks [35], while light gradient-boosting machine (LGBM) for its efficiency and speed in training models on large-scale datasets [36].

In order to create the training data for the models of the given type, the process starts from the definition of how a raw text should be preprocessed into numerical feature vectors that can be used in the training of the model. One of them is the global vectors for word representation, also known as GloVe, which generates word vectors based on the occurrences of the words. These vectors are produced ahead as features in training data that support the model in executing better for any word's similarity, analogy, or even sentiment analysis [37–39]. Another model that can be used is the TF-IDF model which measures the necessity of specific words or word occurrences in a particular document in relation to all documents. However, there are some disadvantages of using TF-IDF such as it does not consider the distribution of the words under the classes and this can lead to poor performance of the model [40]. By transforming text into numerical vectors through these methods, the training data becomes suitable for the model to learn from and make accurate predictions.

Different studies have combined these models for different processing and analytical approaches to improve topic modeling performance by strengthening the capabilities of each algorithm. For instance, XGBoost has been combined with models like random forest,

SVM, and LSTM; moreover, random forest has been combined with SVM, LGBM, and LSTM to improve accuracy and robustness in topic modeling tasks [41]. Moreover, SVM has been integrated with deep learning models like CNN and LSTM, and LSTM has been combined with LDA, to enhance sentiment analysis and topic classification accuracy [42].

1.3. Topic Modeling in Computer Science and Engineering Education

Topic modeling has been a widely used tool in computer science and engineering education, aiding in various aspects of curriculum development, teaching methodologies, and research analysis. For instance, studies have utilized techniques to analyze hidden semantic structures within textual data, providing insights into it. The literature has demonstrated the use of various techniques to analyze knowledge domains and skill sets in specialized fields. For instance, latent Dirichlet allocation (LDA) has been used to explore Big Data software engineering [43], bibliometric analysis has been used to identify research directions for early career researchers in computer science [44], and unified modeling language (UML) has been used to support the understanding on bridging concepts from both computer science and other fields [45], among other applications. Therefore, in the field of computer science and engineering education, by leveraging topic modeling, instructors can create engaging learning experiences that align with industry demands and academic standards. Topic modeling has been instrumental in contextualized computing, where educators employ applications or multidisciplinary areas to teach topics in computer science [46].

By utilizing topic modeling techniques, instructors can identify relevant themes and concepts within educational texts, enabling them to tailor instructional materials to meet the specific needs and interests of students in computer science and engineering programs. Furthermore, utilizing topic modeling can support identifying intersections between disciplines, as well as components such as the stages of the argumentation framework, all looking to prepare students for the complexities of the modern world tailoring curricula to meet the evolving needs of the industry and academia [47].

On the other hand, topic modeling has been instrumental in identifying prevalent themes within user reviews, involving then sentiment analysis and user feedback on educational applications [48]. Also, in the context of higher education research, topic modeling has been utilized to analyze and categorize research publications, unveiling prevalent themes and trends within academic literature [49]. Therefore, the multifaceted application of topic modeling in these fields streamlines curriculum development, students' performance, thinking and learning processes, feedback analysis, as well as research endeavors, enhancing the quality and effectiveness of computer science and engineering education.

However, while topic modeling offers significant benefits, it also presents challenges that should be acknowledged. The short length of user feedback can make it difficult to accurately capture underlying themes, leading to less coherent and interpretable topics [50,51]. Additionally, topic models may generate topics that are statistically valid but semantically incoherent, making it challenging to derive meaningful insights [52]. Therefore, topic modeling should be seen as a tool to support, rather than replace, traditional topic analysis.

2. Methods

2.1. Context and Participants

This study was conducted on the data collected from a lesson design posed that involved the argumentation framework in a design challenge that prompted the design of a zero-energy home for a Midwestern city in the United States, using an integrated CAD/CAE platform to enable simulation-based learning [7,8]. It consisted of design journals for informed trade-off decisions of first-year undergraduate engineering technology students (N = 248) within an introductory course at a Midwestern university in the United States, and focused on the fundamental concepts of building designs and the development of engineering design projects. The course took place in the fall of 2020 under hybrid

teaching conditions due to the COVID-19 pandemic. The activity spanned four weeks within the course.

The design challenge prompted students to apply their knowledge of energy-related concepts to construct an energy-efficient home using Aladdin software, which serves as an integrated CAD/CAE platform for simulating design trade-offs [53]. The students were prompted to consider constraints on costs and size, aiming to achieve net-zero energy consumption.

Each student submitted a design journal in PDF format reporting their informed trade-off decisions on the design challenge. Figure 1 illustrates an example of the format the design journals followed. The design journals consisted of four different parts:

- Factor: Description of intended modifications or decisions toward the design.
- Argue prediction: Justification for the modification or decision on the design, with informed anticipated outcomes.
- Observation: Informed evidence (i.e., accounts of tangible elements) revealed and observed in the CAD/CAE software over the factor of analysis.
- Justification: In the justification section, students provide a reasoned justification for their proposed solutions.

	Factor	Argue Prediction	Observation	Justification
	Which factor of the design are you going to change and how are you going to change it?	BEFORE YOU MAKE THE CHANGE: Provide reasoning for what *you think* will happen due to the change? (Construct an argument justifying why you are changing the factor and the outcome you except. *Consider also* what evidence and reasoning supports your prediction)	AFTER THE CHANGE: What *actually happened* due to the change?	Explain *why you think* it happened. (Construct an argument justifying your explanation by providing all relevant reasons. *Consider also* what evidence and reasoning supports your explanation. Link science or other concepts you *think* are relevant with observations or other evidence. Are there alternative explanations?)
EXAMPLE	Solar panel tilt the orientation upward	Solar panels tilted upward generate more electrical energy [*CLAIM*] because they receive a greater energy from solar radiation when facing the sun directly [*EVIDENCE*] and convert the energy from solar radiation to electrical energy [*REASONING*]	Solar panels facing the south side caused the annual energy cost of the home to decrease.	Solar panels generated more energy [*CLAIM*] because when facing the sun directly, they receive more energy from the sunlight [*EVIDENCE*], and converted the energy from sunlight to electrical energy [*REASONING*]

Figure 1. Example of the prompted design journal.

First, the researchers conducted a preliminary hand-codification process over a random sample of 50 observations who completed the design challenge of creating energy-efficient homes [7]. The hand-codification process consisted of an initial hand-coding resulting in thirty-two different codes, a first axial coding resulting in six different grouping codes, and a second axial coding resulting in two different grouping codes. For each of the 50 random samples, the researchers annotated the frequency of each topic per record. The aim was to investigate how students informed their trade-off decision-making processes in design for energy-efficient homes, within the argumentation framework, and the recurring trends in students' designs related to economic decision-making and energy science [7,8].

To generate the training data for the supervised models, the researchers followed a multi-step process. In the first step, the researchers manually reviewed the responses from the 50 observations and identified individual pieces of information relevant to the

design challenge. This resulted in thirty-two distinct codes, each representing a specific aspect/theme of the student responses, such as considerations for energy efficiency, cost, or sustainability. Next, these thirty-two initial codes were grouped into broader categories through axial coding. The first level of axial coding consolidated the codes into six grouping codes, which grouped related topics under broader themes, such as "economic factors." Afterward, a second axial coding process further refined these six grouping codes into two overarching grouping codes, providing a higher-level categorization of the data, such as "energy science." For each of the 50 observations, the frequency of each identified topic was recorded, involving counting how often each of the thirty-two codes appeared in the responses of each sample. This detailed frequency count helped to quantify the prominence of different topics within the students' designs and decision-making processes. The annotated frequencies from each observation were compiled into a structured dataset, in which each record included a detailed breakdown of how often each topic was mentioned, and categorized according to the two final grouping codes. This structured dataset was then used to generate training data for the supervised models. The models were trained using the frequencies of topics to learn and identify patterns related to energy efficiency and economic considerations.

Moreover, the researchers reflected on the broader challenge of integrating new evidence into scientific explanations, focusing on how incorporating detailed, annotated data can support trade-off decisions by addressing systemic thinking and problem-solving [54]. This information was used in this study to generate training data comprising the fifty instances for the supervised models analyzed in this manuscript. Refer to Figure 2 depicting the general analysis method.

2.2. Data Collection and Processing

The data collection and processing were performed following a two-step procedure. First, the researchers extracted the data from the design journals presented in PDF format, of all students who completed the design challenge (N = 248). These design journals were organized in a table provided by the instructor, comprising four columns representing the stages of the argumentation framework. From these design journals, the data were extracted using Tabula [55]—a library in Python designed for extracting tables automatically. By removing the headers to avoid non-predictive elements, and compiling the data into a dataset, we ensured data quality for the model and reduced variability in the classifiers avoiding data noise. Each row in the dataset utilized in this study represented a student's design journal following the argumentation framework.

Subsequently, the researchers followed by performing pre-processing procedures over the resulting data frame: handling empty entries, filtering out short responses, removing stop words, lemmatizing, and tokenizing the text. Also, the researchers considered that techniques such as feature selection, word processing, and punctuation handling play an important role in enhancing text classification tasks [56]. These models were divided into two groups: (1) unsupervised, and (2) supervised.

Regarding the unsupervised approach, the foremost model implemented was latent Dirichlet allocation (LDA). This model learns by itself the prominent/emergent topics, given parameters that boost the performance. In contrast, the supervised approaches require a human coding of the texts where the classification tasks involve learning the association between words among the observations, utilizing labeled datasets as training sets to help predict outcomes and recognize further patterns over emergent topics. For the supervised approach, six different models were utilized for analysis and comparison, including K-nearest neighbor, support vector machine (SVM), extreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), recurrent neural network with long short-term memory (RNN—LSTM), and random forest. Hence, this study aims to compare supervised and unsupervised approaches for topic modeling tasks.

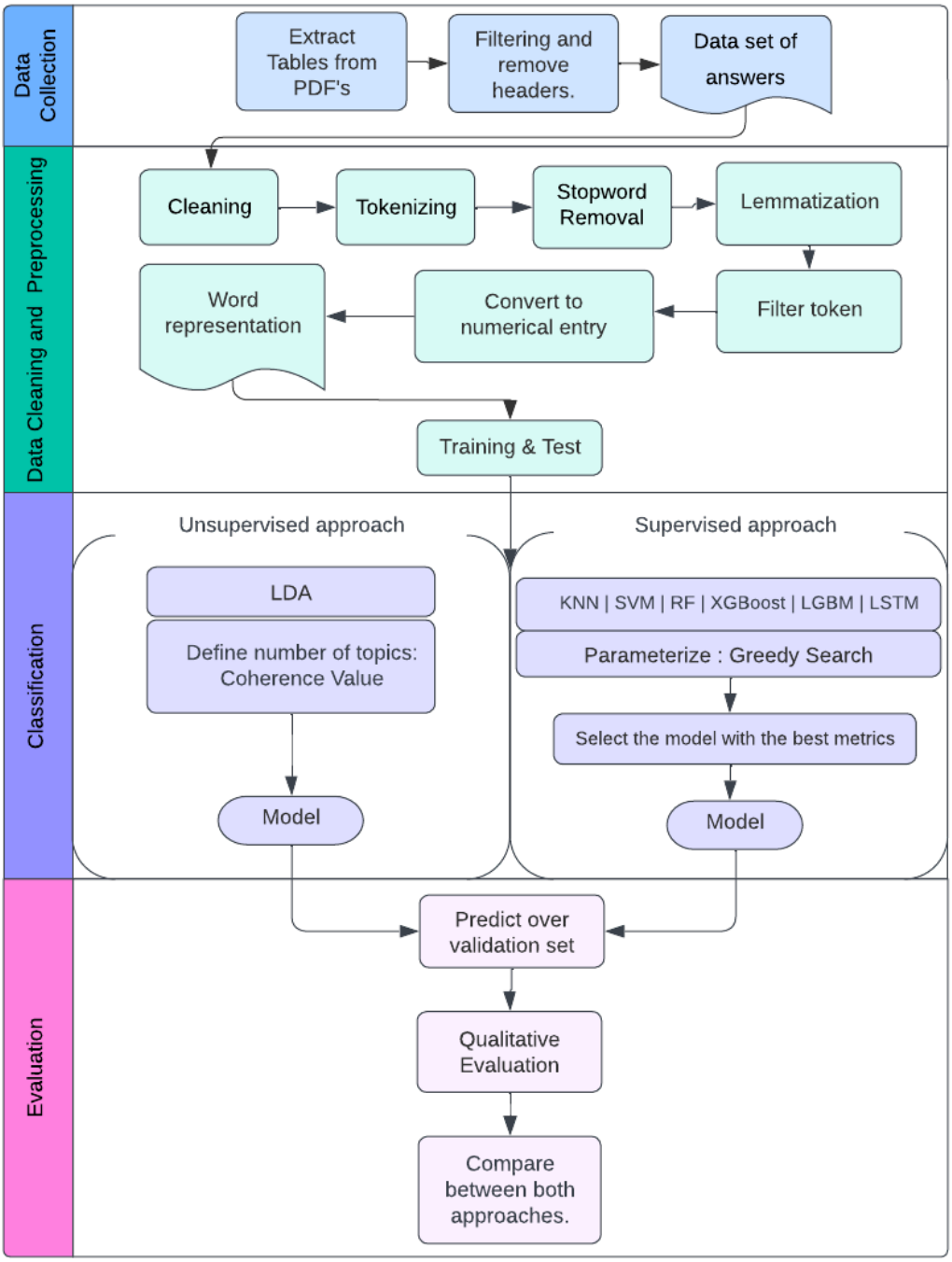

Figure 2. Overview of the comparative analysis of the supervised and unsupervised approaches for topic modeling with qualitative validation.

To ensure the validity and reliability of the data collection process, the design journals filled out by the students were standardized, in the sense that all were provided with the same type and format of journals to fill out. Moreover, when using Tabula for data extraction, it was ensured that there was no error or omission in extracting tables.In addition, during preprocessing stages such as removing non-predictive features and managing empty cells or values, measures were taken to maintain the bias and reliability of the data. Moreover, all the supervised models applied in the analysis were cross-validated using the cross-validation procedures to check the reliability of the results obtained from various partitions in the data and, therefore, improve the reliability of the results yielded.

2.3. Differences between Supervised and Unsupervised Approaches

2.3.1. Unsupervised Approach

For this approach, the researchers opted to utilize the latent Dirichlet allocation model (LDA) as the cornerstone for unsupervised topic modeling. This model is a probabilistic generative model that represents documents as mixtures of topics and topics as mixtures of words, leveraging the co-occurrences of words to form a topic. Also, it is one of the most common procedures for determining the underlying topics and has shown reliable results in similar contexts [3,16]. In addition, it has evidenced its ability to automatically detect topics from large text collections [16].

Considering that determining the optimal number of topics is required to train the LDA model, through a systematic approach, the researchers optimized its hyperparameters, paying particular attention to the number of topics. Since literature has highlighted that perplexity may not always correlate with human annotations and may diverge due to the similarity between topics, especially using a large number of topics, the metric "coherence" was considered as has been one of the most practical ways to determine the optimal number of topics [57]. That is, perplexity measures how well a model predicts the probability distribution of a sample, with lower values indicating a better fit [58]. Nevertheless, while this metric primarily evaluates how well a model predicts data, it does not assess the interpretability of the emerging topics. This can lead to it not accurately reflecting the semantic coherence of the topics [15]. In contrast, "coherence" focuses on the degree of semantic similarity among the top words in a topic, providing a more nuanced understanding of topic quality [59]. Also, this metric has evidence of high reliability in understanding and optimally choosing the number of topics in related tasks, showing a good correlation with human understanding [60,61]. To set out the best number of topics for the model, this study opted to leverage a greedy search algorithm assessing the coherence value [62]. Looking forward to finding the first local maxima, the researchers began with a low number of topics and increased by one topic at a time.

The model was trained using MALLET, a Java-based toolkit for statistical natural language processing (NLP) [63]. Right after the topics were extracted using MALLET, the metric for coherence was computed using Gensim [64], and the coherence score versus the number of topics, to define the ideal number to work within this study accordingly. Figure 3 depicts the first local maxima, corresponding to six emerging topics, which is consistent with the qualitative approach carried out in the previous related study [7]. The first local maximum occurs at six topics because the coherence value at this point is higher than the values at five topics and seven topics, indicating that the model with six topics achieves a good balance, offering meaningful structure and avoiding overfitting. That is, the six-topic point is the first point where the coherence value rises above its neighboring points, making it a logical choice for model selection. In contrast, the three-topic point is not chosen as its coherence value is not higher than the points around it, as represented by the downward trend in this point.

MALLET provided a list of the top 20 words that constituted each topic. For instance, for topic T_i, the list of the top k words, $W_i = \{w_{1i}, w_{2i}, \ldots, w_{ki}\}$, was outputted, in descending order of probability. That is, the first word is the most likely word associated with the topic, followed by the second most likely.

Afterward, the composition of each document (i.e., open-ended responses) was the output in terms of topics and associated weights using the Gibbs Sampling approach [65]. For example, for a given topic model with n topics $\{T_1, T_2, \ldots, T_n\}$, the composition of a response R_i can be represented as follows:

$$C(R_i) = p_{1i}T_1 + p_{2i}T_2 + p_{3i}T_3 + \cdots + p_{ni}T_n$$

where p_{ji} represents the relative weight associated with topic T_j and the sum of all topic weights for a document is 1, i.e., $\sum_{j=1}^{n} p_{ji} = 1$. Therefore, documents composed of multiple topics were expected to be assigned smaller weights for multiple topics, and documents

composed of a single topic were expected to have a high weight associated with that topic [16]. Thus, understanding the precise nuances of these subjects requires background information and expertise in the relevant field [66]. This output of MALLET for the LDA topic model was then analyzed qualitatively to interpret the theme (i.e., semantic) of the topic, with an optimal number of topics defined.

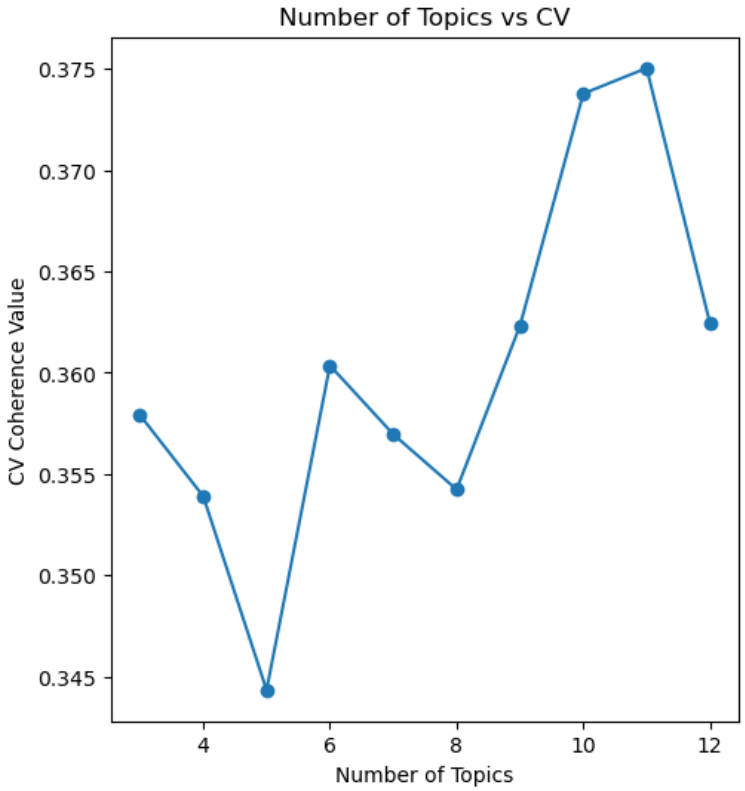

Figure 3. Coherence scores for the optimal number of topics.

The researchers independently labeled the topics and provided brief explanations by reviewing the top 20 words and the top 10 representative documents for each topic. They then reached a consensus on the final topic interpretations. This process ensured that the topic labels were well-informed and agreed upon by researchers.

2.3.2. Supervised Approaches

The researchers leveraged the data from previous work [7] using the proportion of each topic in each document to treat the task as a "multilabel regression problem." Thus, the goal was to learn how different topics in each document use words at different rates and predict the proportion of topics in unseen documents. That is, despite supervised algorithms that have been utilized to classify text based on human-coded discrete categories, certain drawbacks assume assigning "discrete labels", which means that a passage belongs or not to a topic. Nevertheless, a passage might represent a mixture of topics with different proportions, and the data often suffer from class imbalance, which is even more pronounced when the proportions are not taken into account, leading to potential bias in the model's performance.

For this analysis approach using supervised methods, data were selected consisting of 50 hand-coded instances from the previous study [7], as the core of the *training set*. On the other hand, the *test set* consisted of 20 random instances that were not previously evaluated, and which were manually coded by one of the researchers, ensuring objectivity. Afterward, these coding results were further validated by a second researcher, to ensure consistency and reliability.

The data structure for this stage included the text processed along with the labels representing the normalized frequencies of the topics identified in the previous study [7]. From this study, the resulting six topics are defined as follows: (1) 'energy efficiency and conservation', (2) 'insulation and thermal regulation', (3) 'material selection and construction', (4) 'environmental considerations', (5) 'design and space efficiency', and (6) 'economic considerations'.

For the following models, it was necessary to convert the data into numerical representations. GloVe (global vector for word representation) has been utilized for text preprocessing to construct a vector space of words [67]. This pre-trained word-embedding method has proven effective for text representation, aiding in classification tasks [39]. Also, this method can establish linguistic relationships due to training on extensive corpora, which enhances the analysis, utilizing the Common Crawl dataset, comprising 840 billion tokens.

Afterward, several methods for text classification and multilabel classification were considered, as no single model fits all problems. The framework presented in this study first utilized simple modeling techniques to provide insights into the task's complexity, including support vector machines (SVMs) and K-nearest neighbors (KNNs). In contrast, state-of-the-art models (SOAMs) aim to enhance performance and robustness, comprising ensemble learning techniques that combine multiple models to make predictions [68].

This approach leverages the diversity of various models to mitigate overfitting risks and enhance generalization. The ensemble-learning-based algorithms used in this framework included random forest (RF), extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and recurrent neural network (RNNs), specifically long short-term memory (LSTM) networks. In particular, LSTMs are particularly well-suited for text classification tasks due to their ability to capture sequential dependencies and handle long-range dependencies, mitigating the vanishing gradient problem.

Each model was hyper-parameterized using greedy search, which iteratively explores the search space by making the locally optimal choice at each step. Initially, we searched for broad steps within the search space. After identifying regions with better performance, the search was refined to focus on those areas, ensuring more precise optimization of the hyperparameters for each machine learning model. This approach helped avoid falling into local optima and allowed for testing of a wide range of possible parameters (refer to Table 1). While we acknowledge that greedy search could settle for local optima, we ensured to mitigate this limitation by combining the approach with five-fold cross-validation. That is, the technique assesses the performance of the greedy search and reduces the risk of overfitting. Thus, these issues were mitigated by evaluating the model's performance across different data folds to ensure reliability.

Table 1. Hyperparameter optimization.

Algorithm	Hyperparameter	Bounds
KNN	Number of neighbors	(3, 11)
	Weights	[uniform, distance]
	Algorithm	[auto, $ball_{tree}$, kd_{tree}, brute]
	Leaf Size	(10, 50)
	Distance	[Manhattan distance, Euclidean distance]
SVM	Kernel	[linear, poly, rbf, sigmoid]
	C	(0.01, 10)
	Gamma	[scale, auto]
	Degree	(2, 5)
Random forest	Number of estimators	(100, 600)
	Maximum depth	(3, 30)
	Minimum samples split	(3, 20)
	Minimum samples leaf	(3, 20)
	Maximum features	(0.5, 1.0)
	Maximum samples	(0.7, 1.0)

Table 1. *Cont.*

Algorithm	Hyperparameter	Bounds
LightGBM	Number of estimators	(50, 500)
	Maximum depth	(3, 50)
	Number of leaves	(10, 100)
	Learning rate	(0.001, 10)
	Minimum child weight	(1.0, 20.0)
XGBoost	Number of estimators	(50, 500)
	Maximum depth	(3, 50)
	Learning rate	(0.0001, 10.0)
	Regularization lambda	(0.0, 10.5)
	Gamma	(0.0, 100.0)
	Minimum child weight	(0.0, 200.0)
	Subsample	(0.5, 1.0)
	$\text{Colsample}_{\text{bytree}}$	(0.5, 1.0)
LSTM	Embedding dimension	(100, 300)
	LSTM units	(64, 128, 30)
	Learning rate	(0.0001 , 0.1)
	Batch size	(16, 32, 64)
	Epochs	(10, 300)

Afterward, the researchers performed a final validation using five-fold cross-validation, repeated ten times for each model, resulting in 50 estimations of MSE, for the models considered in this study. This enabled the comparison of the mean squared error (MSE) and variance between the models, ensuring assessing the models' ability to generalize across different subsets of the data.

The criteria for selecting the model included the use of MSE, making it suitable for gradient-based optimization algorithms. Although MSE is a common choice for regression problems, this study aimed to identify the main topics to understand how students make trade-off decisions in the context of design engineering. To address this, tailored metrics for these purposes were created, considering the following:

- Exact matches (EMs): The top three topics are the same and in the same order.
- Unordered matches (UMs): Top three topics are the same but in any order.
- Highest topic matches (HTMs): Whether the main topic appears in the top three predicted topics.
- Proportion of highest topic matches (PHTMs): Frequency of the main topic appearing in the predictions.
- Two of three matches (TTMs: At least two of the top three predicted topics match the actual labels.
- Main topic accuracy (MTA): The first topic prediction matches the actual topic.

As an illustration Table 2 of how the results and the evaluation for one instance:

Table 2. Model prediction example.

	Top 1	Top 2	Top 3
Test Set	Design and space efficiency	Energy efficiency and conservation	Environmental considerations
Model prediction	Design and space efficiency	Energy efficiency and conservation	Environmental considerations

From this evaluation, tailored metrics were used to select the best model. Then, an account of the predictions in the validation set was made for each component of the argumentation framework (i.e., claim, evidence, reasoning).

2.4. Qualitative Analysis: Supervised and Unsupervised Approaches in Topic Modeling

Combining analytical models with qualitative analysis is essential to evaluate the reliability of predictions over the validation set. Both the LDA and the supervised model predicted the same 100 instances—not previously seen by the supervised model—for the three sections of the argumentation framework: argument prediction (claim), observation (evidence), and justification (reasoning).

For each document, the predicted topic proportions were sorted in descending order to identify the top three topics. Researchers then manually validated each model and component of the argumentation framework, annotating the number of topics for the top three predicted topics for each answer. This process was repeated for both approaches. A second researcher conducted an additional evaluation to ensure objectivity, achieving substantial agreement with a kappa statistic of 0.80, according to the Landis and Koch benchmark scale [69]. Any minor discrepancies were resolved through discussion until a consensus was reached.

The topics were analyzed in detail to understand how the models distributed the top three topics for each column. This analysis aimed to determine alignment with previous studies on content and students' topics during argumentation. The comparison between the models provided insight into these alignments. Following this, the accuracy of human validation was checked to understand why one approach might be preferable and its benefits.

3. Trustworthiness, Validity, and Reliability

This study looked after the consistency and agreement between the two researchers who performed the analysis. A Ph.D. student with expertise in systems engineering, data analytics, and engineering education and a visiting scholar with expertise in industrial engineering and data analytics conducted together the data analysis process to ensure the inter-rater reliability of this study. The two researchers worked together through regular weekly meetings. Using a consensus approach, the two researchers independently analyzed the results for each model utilized in this study and then met to reflect on and discuss them, ensuring consistency in the analysis and resolving any discrepancies in the interpretations—for example, researchers manually evaluated the top three automated emerging topics for models XGBoost and LDA, each analyzing the alignment of each theme for 100 instances for each stage of the argumentation framework, resulting in analysis for both models of 600 instances for claim, evidence, and reasoning. The inter-rater reliability was assessed using Cohen's kappa [70]. With an overall kappa of 0.80, the researchers achieved a substantial agreement between them across all results, showing a high level of consistency and reliability in their ratings. Moreover, through the regular weekly meetings, the two researchers engaged in reflexivity to enhance the trustworthiness of this study by reflecting on potential biases in the resulting interpretations and further analyses.

4. Results

4.1. Differences between Supervised and Unsupervised Approaches

4.1.1. Supervised Models

The results obtained from the supervised models are summarized in Table 3. This table presents a comparison between the three actual main topics and the three main predicted topics for each model. The models demonstrate strong performance, with low mean squared error (MSE) values indicating accurate predictions. However, to fully understand each model's capabilities, we need to delve into more tailored metrics.

The highest topic match (HTM) metric evaluates how accurately the model identifies the most prominent topic. XGBoost achieved a perfect score with 100%, the highest topic matches, followed closely by random forest with 98.31%. LSTM attained 93.22%, and LGBM followed with 96.61%. This highlights the robustness of these models in capturing the dominant topics within the data.

Table 3. Supervised model performance comparison for the top three topics.

Model	Best Parameters	Metrics
Random forest	Number of estimators = 76, Maximum depth = 20, Maximum features = 'sqrt', Minimum samples leaf = 4	Mean squared error: 0.0037 Exact matches: 0.4576 Unordered matches: 0.7288 Highest topic matches: 58 out of 59 predictions Proportion of highest topic matches: 0.9831 Two of three matches: 1.0000 Main topic accuracy: 0.7119
LGBM	Number of trees = 210, Maximum depth = 3, Number of leaves = 31, Learning rate = 0.01, Estimator subsample: 0.7, $\text{Colsample}_{\text{bytree}}$ = 0.9	Mean squared error: 0.0070 Exact matches: 0.2373 Unordered matches: 0.4407 Highest Topic matches: 55 out of 59 predictions Proportion of highest topic matches: 0.9322 Two of three matches: 0.9661 Main topic accuracy: 0.5763
XGBoost	Number of estimators = 80, Maximum depth = 3, Learning rate = 0.1, Estimator subsample = 0.7, $\text{Colsample}_{\text{bytree}}$ = 0.8	Mean squared error: 0.0004 Exact matches: 0.5085 Unordered matches: 0.8475 Highest topic matches: 59 out of 59 predictions Proportion of highest topic matches: 1.0000 Two of three matches: 1.0000 Main topic accuracy: 0.8475
RNN: LSTM	LSTM units = 256, Learning rate = 0.001, Epochs = 10, Embedding dim = 300, Batch size = 64	Mean squared error: 0.0069 Exact matches: 0.1864 Unordered matches: 0.3559 Highest topic matches: 58 out of 59 predictions Proportion of highest topic matches: 0.9831 Two of three matches: 0.9831 Main topic accuracy: 0.5593
SVM	C = 1.0098, degree = 4, kernel = 'sigmoid'	Mean squared error: 0.0099 Exact matches: 0.1695 Unordered matches: 0.3051 Highest topic matches: 51 out of 59 predictions Proportion of highest topic matches: 0.8644 Two of three matches: 0.9661 Main topic accuracy: 0.5254
KNN	Number of neighbors = 17, Weights = uniform, Algorithm = 'auto' Leaf Size = 5	Mean squared error: 0.0147 Exact matches: 0.0508 Unordered matches: 0.1864 Highest topic matches: 39 out of 59 predictions Proportion of highest topic matches: 0.6610 Two of three matches: 0.8136 Main topic accuracy: 0.2881

The exact match rate (EMR) and unordered match rate (UMR) are critical for assessing how well models predict both the correct topics and their order. XGBoost performed best with an exact match rate of 50.85% and an unordered match rate of 84.75%. This indicates XGBoost's strong ability to accurately identify all three main topics in their correct order or identify the correct topics regardless of order. In contrast, LSTM using RNN showed lower performance, with an exact match rate of 18.64% and an unordered match rate of 35.59%, suggesting it struggled to predict all three topics and their order accurately, although it could identify individual topics reasonably well. Figure 4 illustrates a radar chart comparing the performance of the top three topics across the supervised models analyzed in this manuscript.

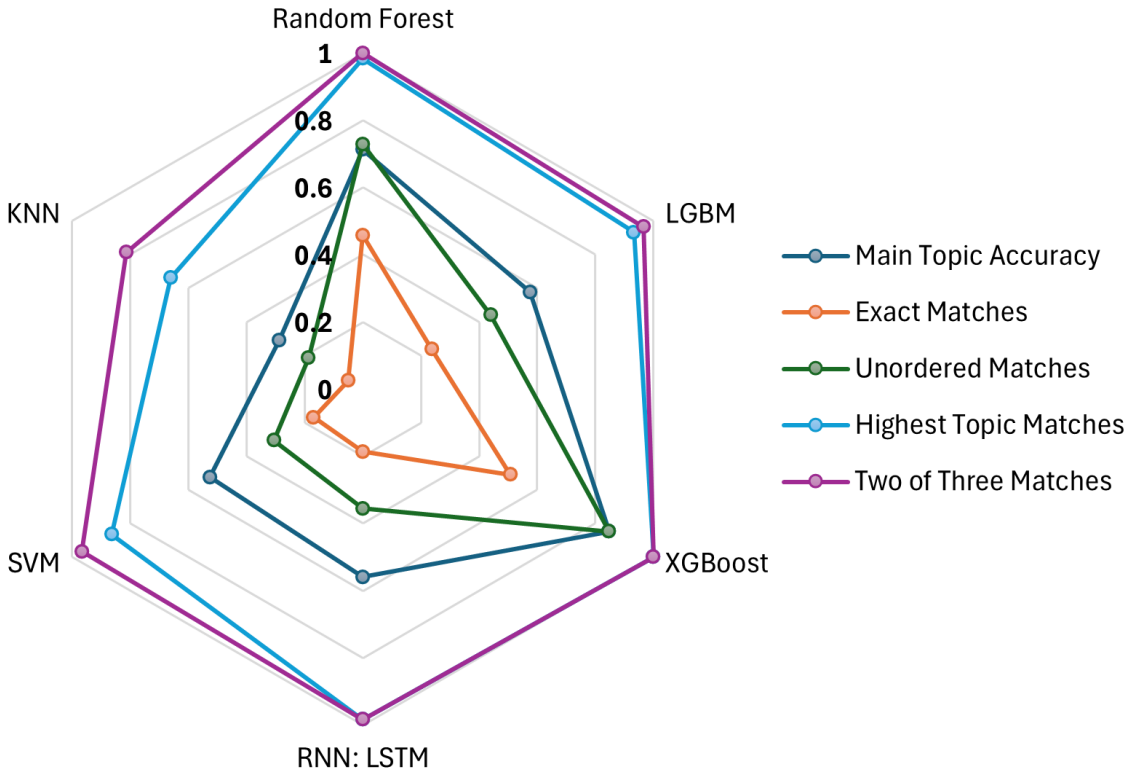

Figure 4. Radar chart of the performance comparison among the supervised models.

Apart from performance metrics, considerations such as computational efficiency and interpretability are important when choosing a model. XGBoost stands out for its high precision and minimal mean squared error (MSE), making it suitable for applications requiring precise predictions. In contrast, random forest demonstrates robust performance overall, emphasizing generalizability and accurate topic identification. Simpler models like SVM offer clear interpretative insights, which are valuable in contexts where understanding predictions is essential. Additionally, hybrid approaches and advanced ensemble techniques present promising avenues for improving predictive capabilities further.

4.1.2. Unsupervised Model: Latent Dirichlet allocation (LDA)

The LDA model identified six as the optimal number of topics for the argumentation framework, as depicted in Figure 3. This figure illustrates coherence scores for different topic numbers, with the highest score observed for six topics. Table 4 summarizes the themes identified by the LDA model across the entire corpus. For each topic, the table includes a prominent example, its corresponding 20 words, and its weight, providing a comprehensive overview of the main themes identified by the model.

Table 4. Labeling topics: interpretation.

Label	Topic Weight	Top 20 Words	Representative Quotes
Energy efficiency and conservation	0.12571	energy, solar, panel, home, decrease, house, net, window, increase, heat, cost, tree, amount, roof, increasing, annual, size, electrical, efficiency, consumption	"The addition of more trees around the home is beneficial and increases energy efficiency because they provide shade to the home when you need it. Because of this, this lowers A/C costs over the year."
Solar energy positioning and costs	0.26446	energy, south, house, side, roof, sun, window, winter, sunlight, solar, cost, facing, time, panel, area, reduce, block, summer, large, heating	"Hip roof is better because it has more area facing toward the sun compared to regular roof. Reducing the total area of the house can reduce the total cost of the house."

Table 4. *Cont.*

Label	Topic Weight	Top 20 Words	Representative Quotes
Solar energy generation	0.21765	solar, house, energy, sun, panel, sunlight, window, roof, tree, south, leaf, summer, light, radiation, angle, generate, heat, directly, facing, winter	"More solar panels will include more solar cells so more sunlight can hit the surface to create energy."
Economical considerations	0.07336	cost, house, solar, panel, make, budget, high, wall, side, sunlight, foundation, time, order, money, idea, energy, adding, space, expensive, made	"Lowering the walls will bring down the price because there is less material and thus make living in the house cheaper as there is less cost to cover."
Solar panel placement and solar heat gain	0.15969	solar, energy, house, sun, east, west, south, panel, sunlight, window, side, day, heating, heat, tree, receive, northern, direct, hemisphere, radiation	"Having solar panels on those sides of the house will produce more energy because when solar panels face the sun, they gain more energy from solar radiation and is converted into energy."
Insulation and thermal consideration in seasons	0.18272	house, energy, heat, winter, air, window, amount, summer, reduce, temperature, cool, heating, sunlight, insulation, wall, adding, net, side, cold, tree	"This should keep all hot and cold air in and shield from the opposite outside. That should reduce the amount of AC and heat the house uses. Shade from the windows should help with cooling in the summer. The amount of energy to heat and cool the house should be reduced because of the reduction of escape points. The insulation should work the same way as for the windows and walls."

Following a consensus on the preliminary findings, thematic interpretations emerged regarding the integration of solar energy and energy conservation in residential settings. These themes focus on optimizing energy efficiency while addressing economic and thermal comfort concerns. Specifically, the LDA model identified the following themes related to building energy-efficient homes:

- Energy efficiency and conservation: This theme focuses on energy efficiency in homes, including aspects such as solar panels, windows, heat, and costs. The frequent mention of terms like "decrease", "increase", "annual", "electrical", and "consumption" suggests a strong focus on improving energy efficiency and managing energy consumption in households.
- Solar energy positioning and costs: This theme centers around the positioning of elements in houses (e.g., "south", "side", "window", "roof") to maximize sunlight exposure, particularly in different seasons ("winter", "summer"). It also touches on the costs associated with solar energy.
- Solar energy generation: This theme emphasizes solar energy generation, focusing on aspects like sunlight, panels, roofs, and trees. It also considers seasonal changes ("summer", "winter") and the importance of angle and radiation for effective solar energy generation.
- Economical considerations: This theme revolves around the financial aspects of installing solar panels, such as cost, budget, and expenses. It also considers structural elements like walls and foundations.
- Solar panel placement and solar heat gain: This theme deals with the placement of solar panels on different sides of a house (e.g., "east", "west", "south") to maximize sunlight exposure throughout the day. It also mentions heating and direct sunlight, which are crucial for effective solar energy usage.
- Insulation and thermal consideration in seasons: This theme focuses on managing energy for heating and cooling throughout different seasons ("winter", "summer").

It mentions elements like insulation, windows, and temperature control to reduce energy consumption and maintain comfort in homes.

The results demonstrate how effectively LDA can uncover latent topics within student responses, revealing the underlying themes and patterns in their thinking. This outcome supports the understanding of how students approach tasks related to building energy-efficient homes and how they make informed trade-off decisions. The transitions between topics are nuanced, indicating the understanding and application of engineering design concepts integrating scientific and non-scientific considerations.

These findings underscore the need to combine quantitative and qualitative analyses to validate and interpret LDA results, ensuring a comprehensive understanding of themes in student responses. The comparison between the LDA model and the supervised machine learning model will further highlight the strengths and limitations of each approach. This comparison is essential for determining the most effective method consistent with the goals, for analyzing student responses in this educational context.

4.2. Qualitative Results: Supervised and Unsupervised Approaches in Topic Modeling

The researchers conducted a qualitative evaluation to assess the model's reliability in accurately identifying and predicting topics across new documents. This involved annotating each instance (i.e., design journal) to determine the alignment with the student's claimed text across the three stages of the argumentation framework. They categorized the analysis into agreement (three out of three), partial agreement (two out of three), and disagreement (one out of three) levels.

Inter-rater reliability was carefully considered to ensure consistency and agreement between two researchers conducting a qualitative evaluation. They followed a predefined protocol to independently assess the performance of LDA and XGBoost models across 100 instances within the argumentation framework. This involved evaluating how well the top three predicted topics aligned with the text for each instance, totaling 600 evaluations. Through regular discussions and reflection on discrepancies in their interpretations, the researchers maintained clarity in their qualitative evaluation process and criteria. The agreement between the two researchers was substantial, resulting in 83% agreement for LDA and 75% for XGBoost in the argue prediction stage, 75% for LDA and 73% for XGBoost in the observation stage, and 86% for LDA and 83% for XGBoost in the justification stage. These findings are detailed comprehensively in Table 5, which provides a comparative overview of how LDA and XGBoost performed in identifying topics across claim, evidence, and reasoning aspects of the argumentation framework.

Table 5. Topic identification results for LDA and XGBoost in the top three topics.

Model	Argue			Observation			Justification		
	3 of 3	**2 of 3**	**1 of 3**	**3 of 3**	**2 of 3**	**1 of 3**	**3 of 3**	**2 of 3**	**1 of 3**
LDA	71.0%	27.0%	2.0%	77.5%	19.3%	3.1%	59.3 %	37.50%	0.31%
XGBoost	65.6 %	32.3%	2.02%	77.5%	20.4%	2.04%	40.6%	55.2%	4.1%

From the results illustrated in Table 5, the following analysis is made for each component of the argumentation framework:

- **Argue category:**
  - LDA achieves higher percentages in identifying three out of three topics (71.0%) compared to XGBoost (65.0%), indicating better performance in comprehensive topic coverage.
  - XGBoost demonstrates a notable difference in identifying two out of three topics, showing better performance with 32.0% vs. 27.0% for LDA, and also performs slightly better in identifying one out of three topics (3.0% vs. 2.0% for LDA).

- **Observation category:**
  - Both models show similar performance through all three metrics, in identifying three out of three topics is identical with (77.5%).
  - XGBoost excels in identifying two out of three topics (67.7% vs. 46.5% for LDA) but slightly underperforms in identifying one out of three topics (8.1% vs. 5.1% for LDA).
  - XGBoost correctly identifies none of the topics (zero out of three) in this category, whereas LDA identifies 2.0%.
- **Justification category:**
  - LDA demonstrates superior performance in identifying three out of three topics (59.3%) compared to XGBoost (23.2%).
  - XGBoost excels in identifying two out of three topics (55.2% vs. 37.5% for LDA) but slightly underperforms in identifying one out of three topics (8.1% vs. 5.1% for LDA).

The findings underscore LDA's capability to encompass a diverse array of topics within argumentative texts, whereas XGBoost demonstrates proficiency in precisely identifying specific topics, especially in observational contexts. Visual representations through bubble charts (see Figures 5 and 6) effectively illustrate the distribution of the top three topics identified by each model. Topics are color-coded and grouped along the y-axis across the argumentation framework's three sections. The x-axis denotes proportions, with bubble size correlating to the frequency of each topic as the predominant focus. These visualizations provide clear insights into how LDA and XGBoost perform in topic identification across different stages of argumentation, enhancing understanding of their respective strengths in textual analysis.

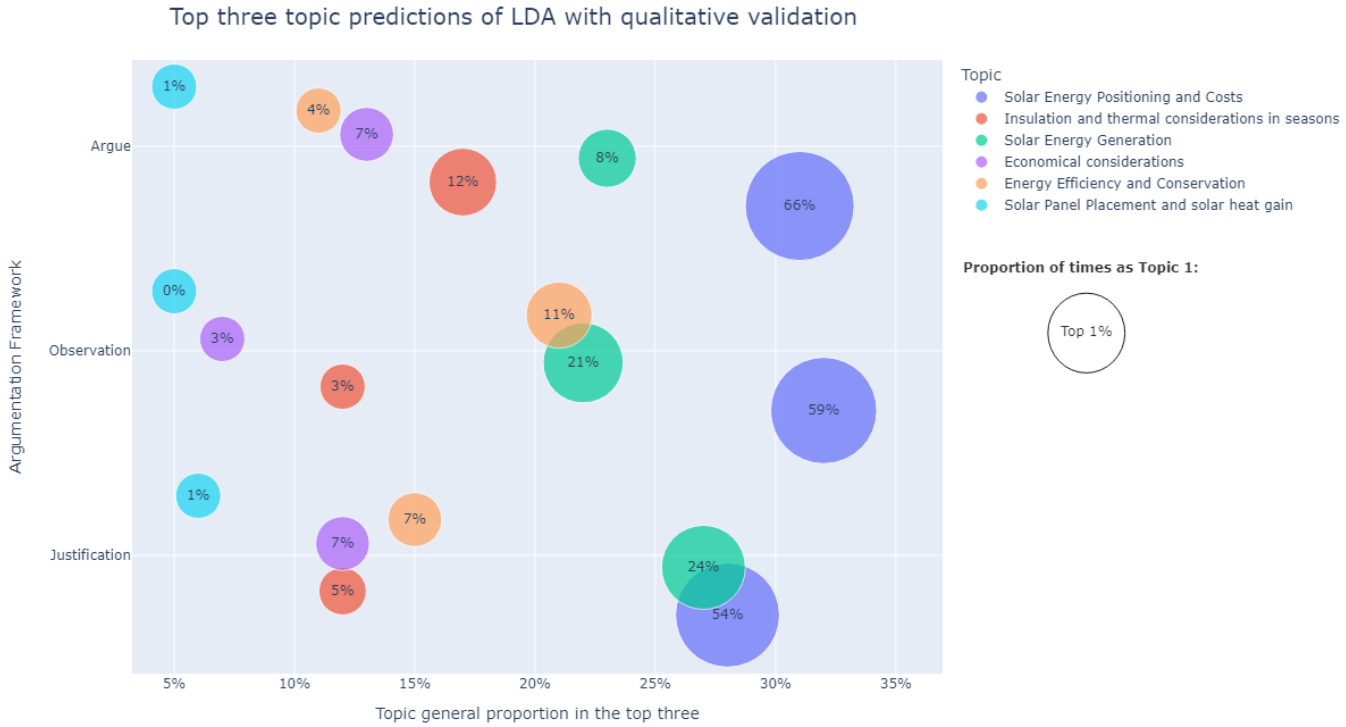

Figure 5. Top three topic predictions of LDA with qualitative validation.

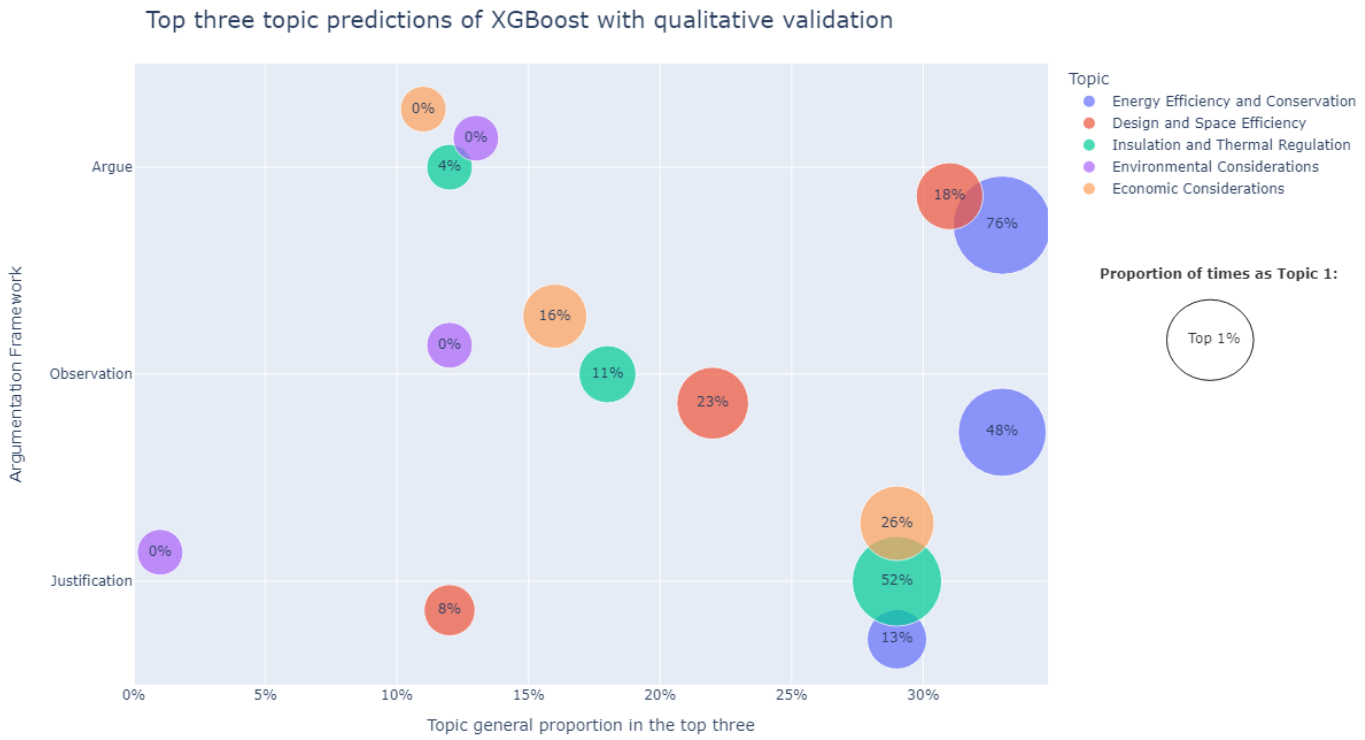

Figure 6. Top three topic predictions of XGBoost with qualitative validation.

4.2.1. LDA Argumentation Framework

Figure 5 illustrates the proportion of categories across different segments. The top three topics closely align with Mallet's output probabilities. Solar energy positioning and costs dominate, appearing 31% of the time and ranking highest in LDA at 26%. This category focuses on optimizing sunlight exposure through strategic placement. Solar energy generation follows at 23%. This pattern holds consistently across the observation and justification stages, with only a one-percentage point gap, highlighting their significance throughout the argumentation framework.

An example showcasing the prominence of solar energy positioning and generation, alongside insulation, is illustrated in the fragmented quote. It highlights strategic solar panel placement for energy capture and window positioning to manage sunlight, considering external factors like trees for cooling in summer and solar heating in winter. Table 6 depicts the representative quotes illustrated in the previous example.

Table 6. Representative quotes for LDA.

Topics Order	Quote
Solar energy positioning and costs	"Solar panels on the roof generate more energy because they are closer to the sun and thus have a higher chance to get more coverage."
Solar energy generation	"Solar panels facing the south will generate more electrical energy because they receive more energy from the sun."
Insulation and thermal considerations in seasons	"Trees outside the windows will make the summer times colder in the house due to lack of sunlight and make the house warmer in the winter due to the excess of sunlight coming into the house."

Patterns in the results are notable. Energy efficiency and conservation showed a significant increase from 11% in the argue prediction stage to 20% in the observation stage before declining, highlighting a strong focus on this topic during observational analysis. Economic considerations, addressing budget discussions and cost fluctuations, consistently

appeared at around 12% across all three components of the argumentation framework. While overlapping with cost discussions, economic considerations emphasize broader savings and perspectives rather than specific elements.

The dominant topic strongly correlated with overall proportions. Solar energy positioning and costs dominated over solar energy generation, particularly noticeable in the justification stage where Topic 1 held a substantial thirty percentage point lead despite a narrow 1% difference in general proportion. In the argue prediction stage, students predominantly focused on solar energy positioning and costs, with insulation as the second main topic, highlighting its significance in their arguments despite ranking third overall.

4.2.2. XGBoost Argumentation Framework

Figure 6 summarizes the validation results of XGBoost predictions, emphasizing the need for cautious interpretation due to small train and test set sizes (50 and 20 instances, respectively). Efficiency and conservation emerged as the primary theme, starting prominently at 76% in the argue prediction stage, decreasing to 48% in the observation stage, and further to 13% overall while remaining top-ranked despite the decline. Design and space efficiency closely followed, initially at 32% and gradually decreasing to 12% by the justification stage. Environmental considerations, initially significant, declined in prominence across the argumentation and observation stages, showing a decrease in justification.

These findings highlight the dynamic nature of the topic prominence across the different stages of the argumentation framework within the XGBoost predictions. While efficiency and conservation maintained a strong presence throughout, other themes like design and space efficiency and environmental considerations showed varied levels of prominence, reflecting shifts in focus as arguments progressed through their justification stages.

Figure 6 illustrates how topic proportions change across different stages of the argumentation framework. While design and space efficiency decrease notably, several other topics follow an opposite trend, indicating a strong connection between the argumentation stage and the importance of specific topics.

For example, economic considerations start at 12% in argue prediction and steadily increase in prominence, reaching 16% in the observation stage and notably increasing then to 26% in the justification stage. This rise positions economic considerations closely behind energy efficiency and conservation, representing nearly one-third of the topics identified as Topic 1. A similar trend is observed with insulation and thermal consideration, which also rises from 12% in the argue prediction stage to nearly one-third of the topic distribution, becoming the dominant Topic 1.

This increase in economic considerations and decrease in energy efficiency can be seen in the next randomly chosen instance within the justification stage where the model predicted "economic considerations." Although the predominance of economic factors in trade-off decision-making is apparent, this predominance is more evident in the justification stage, in contrast to the energy efficiency and conservation considerations, which were much less evident in the observation and justification stages than in the argue prediction stage.

The correlation with findings from the base study [7] is evident. Material selection and construction consistently failed to rank among the top three topics, likely due to its minimal representation in the training set at only 7%. Environmental consideration, while not dominant, remained significant in the argue prediction and observation stages. In the justification stage, themes like economic consideration, insulation and thermal regulation, and energy efficiency and conservation took prominence.

Comparing both models, LDA offered detailed insights with specific word outputs, while XGBoost benefited from supervised learning's interpretative advantages, potentially enhancing accuracy. Both models converged on topics such as economic consideration, energy efficiency and conservation, and insulation and thermal regulation. However, distinctions arose in topics like solar energy positioning and costs, closely linked to dis-

cussions on design and space efficiency, focusing on housing specifics. Table 7 depicts the representative quotes illustrated in the previous example.

Table 7. Representative quotes for XGBoost.

Topics Order	Quote
Energy efficiency and conservation, insulation and thermal regulation, economic considerations	"Adding a source of power to a house that lacks any will drastically lower the annual energy cost, because having more energy naturally provided will lead to less energy consumed. Raising the roof will help lower energy cost because heat rises, so raising the roof will help contain the heat and lower energy consumption. Increasing window size will decrease annual energy cost because more sunlight will get in, which will provide more heat in the winter months. Increasing the insulation R-value will decrease annual energy cost because the house will be better at maintaining its internal temperature, which will lead to less heat and A/C being used."
Insulation and thermal regulation, economic considerations, design and space efficiency	"Adding the solar panels drastically lowered the annual energy cost. Raising the roof did the opposite of what I expected and upped the energy consumption. Increasing the window size did decrease the annual energy cost, but only slightly. Increasing the insulation R value drastically lowered annual energy cost."
Economic considerations, insulation and thermal regulation, energy efficiency and conservation	"Solar panels generated a lot of energy, especially compared to no solar panels, because solar panels use energy from sunlight to make electricity that can be used to lower the energy cost. Raising the roof upped the energy consumption because it caused more space to heat, which will lead the heater to work harder to heat the same amount of living space. Increasing the window size did decrease the annual energy cost, but only slightly because the winter energy consumption dropped, but energy consumption increased in the summer which only led to marginal improvements. Increasing the insulation R value drastically lowered annual energy cost because the house better maintained its internal temperature, meaning that less heat and A/C were used. A/C and heat were the two sources taking up energy, and lowering them minimally over one day built up over the year."

Table 8 illustrates topic predictions across stages of the argumentation framework. Both models identify "solar energy positioning and costs" as primary, with XGBoost also associating it with "design and space efficiency." These included discussions on optimizing solar panels and window placement for energy efficiency, impacting conservation strategies like using sunlight for heating and maximizing electricity generation, even in winter conditions. This showcases the models' effectiveness in identifying and interpreting pertinent topics.

Table 8. Average differences in the proportions of the top three topics.

Model	Argue			Observation			Justification		
	1-2	2-3	1-3	1-2	2-3	1-3	1-2	2-3	1-3
LDA	15.03%	8.45%	23.48%	24.28%	10.79%	35.07%	14.08%	8.71%	22.80%
XGBoost	5.33%	6.20%	11.53%	5.75%	5.28%	11.04%	5.14%	4.75%	9.89%

Another important aspect to highlight is the proportion of topics identified by each model. Table 8 provides a summary of the average differences in the proportions of the top three topics, which are important for distinguishing between the models. These differences show statistical significance, indicating that the models vary significantly in how they prioritize and allocate importance to different topics. There is statistical significance,

determined by a p-value lower than 0.05 (p-value < 0.05) from a t-test of mean differences using t-Student, the assumptions over normality are assumed by the central limit theorem, underscores the distinct approaches and outputs of each model in topic identification and prioritization.

In the observation stage, LDA consistently highlighted "solar energy positioning and costs", while XGBoost shifted the focus to "energy efficiency and conservation." XGBoost introduced "economic consideration", citing cost reduction benefits from energy strategies. In the justification stage, both models emphasized "economic considerations", with LDA correctly noting "solar energy generation", supported by quotes on increased solar panel energy production. Table 9 depicts the representative quotes supporting this claim.

Table 9. Representative quotes for LDA and XGBoost.

Model	Topic 1	Topic 2	Topic 3	Quote
			Claim	
LDA	Solar energy positioning and costs	Solar energy generation	Energy efficiency and conservation	"Solar panels on the south side of the roof generate more electricity because they are exposed to the sun more throughout the day and convert the sunlight to electrical energy. Making the windows on the south side of the house larger will allow more sunlight in during the winter and let the sunlight warm the house more instead of using power on the heater."
XGBoost	Energy efficiency and conservation	Design and space efficiency	Environmental considerations	
			Observation	
LDA	Solar energy positioning and costs	Energy efficiency and conservation	Solar energy generation	"Solar panels on the south side caused the annual energy cost of the house to decrease. Larger windows on the south side of the house caused the energy cost of the house to decrease."
XGBoost	Energy efficiency and conservation	Economic considerations	Insulation and thermal regulation	
			Justification	
LDA	Solar energy generation	Economic considerations	Energy efficiency and conservation	"The solar panels generated more energy because when they are exposed to the sun more, they received more sunlight and converted more energy to electrical energy. The larger windows saved energy because the house was able to be warmed more from the sun in the winter, which saved energy from being spent on the heater."
XGBoost	Economic considerations	Energy efficiency and conservation	Insulation and thermal regulation	

The findings underscore the dynamic evolution of the topic's importance across argumentation stages, starting with a strong emphasis on design and space efficiency and transitioning towards economic considerations in final justifications. The bubble charts in Figures 5 and 6 illustrate these shifts, with XGBoost's insights from the training set offering clearer distinctions. Particularly, supervised models provided a nuanced understanding of how students navigated between topics, enriching qualitative insights and highlighting the evolving focus throughout the argumentation process.

5. Discussion

5.1. Comparison of Supervised and Unsupervised Approaches in Terms of Coverage and Computational Efficiency

The results suggest that both unsupervised and supervised approaches effectively identify nuanced topics within the argumentation framework. Specifically, latent Dirichlet allocation (LDA) reveals six distinct themes that elucidate aspects of building energy-efficient homes. These themes offer valuable insights into students' understanding and application of scientific knowledge and other non-scientific considerations (e.g., economic

considerations) in engineering design. This finding highlights the importance of employing both unsupervised and supervised methods in educational research, particularly in the context of engineering design. The six distinct themes identified by LDA not only reflect students' grasp of technical concepts related to energy-efficient homes but also underscore their ability to integrate economic factors into their design processes. By recognizing these themes, educators can tailor their instructional strategies to address both the scientific and practical aspects of engineering, fostering a more holistic understanding among students [71]. Furthermore, the insights gained from this analysis can inform curriculum development, ensuring that students are equipped with the necessary skills to navigate the complexities of real-world engineering challenges, where both technical knowledge and economic viability are crucial for successful outcomes. This dual approach can ultimately enhance students' critical thinking and problem-solving abilities, preparing them for future roles in sustainable design and engineering practices [72]. Moreover, this dual-method approach helps uncover nuanced insights that might be missed when relying on a single method. For instance, while unsupervised methods like LDA reveal emerging themes and patterns, supervised methods can validate these findings and explore their practical implications more deeply.

A compelling finding of this study is that it achieves high performance in classifying documents as mixtures of topics, surpassing previous single-topic or binary classification studies. Qualitative evaluation confirms reliability, with both models correctly identifying two out of three topics in about 87% of instances and all three topics in approximately 70% of documents. Another notable finding is XGBoost's accurate prediction of proportions across argumentation framework segments, providing insights into topic fluctuations and overall behavior. Surprisingly, LDA does not depict this transition as clearly as XGBoost [73]. However, given the small sample size, caution must be applied, as these findings might not be reproducible on a larger scale or across unseen applications.

XGBoost's detailed prediction of topic proportions across different segments of the argumentation framework offers valuable insights into the dynamic nature of topic distribution, shedding light on how topics evolve and interact within documents [35]. In contrast, while LDA provides a broad overview of topic distribution, its less detailed depiction of transitions between topics suggests limitations in capturing the finer aspects of topic changes [73].

These findings underscore the effectiveness of ensemble methods like XGBoost and random forest in predicting topic proportions and identifying representative topics, crucial for integrating quantitative and qualitative methods in teaching applications [74]. While these methods offer superior predictive accuracy and robustness, the suitability of different models should align with specific application needs. Moreover, the effectiveness of ensemble methods such as XGBoost and random forest highlights their potential as powerful tools for educators seeking to analyze student responses more effectively. By leveraging these advanced algorithms, educators can gain deeper insights into the nuances of student arguments, which can inform instructional design and curriculum development. Future research could explore hybrid and advanced ensemble techniques to enhance predictive performance while preserving interpretability.

5.2. Comparison of Supervised and Unsupervised Approaches in Terms of Interpretability

The findings from both models offer valuable insights for educators. The clear transition of topics identified by XGBoost can help educators understand how students' focus shifts throughout the argumentation process. This understanding can inform the design of instructional strategies that support students in making more coherent and logically structured arguments. Additionally, the emphasis on economic considerations highlights the importance of integrating cost-benefit analysis into the curriculum. Thus, instructors can use these insights to design activities that encourage students to consider both technical and economic aspects of their design decisions [75].

The combination of LDA and XGBoost provides a comprehensive view of student argumentation, revealing both the prevalent topics and the transitions between them. These insights can be used to enhance educational practices and better support students in developing robust and well-rounded arguments [76]. In fact, since supervised models are trained with predefined classes or results, they help reveal how different topics interact with student outcomes. On the other hand, unsupervised approaches, such as LDA, do not rely on predefined categories. Thus, despite LDA can reveal different topics that students discuss as part of their argumentation process, it may not clearly show how these topics are connected or how students shift their focus. For instance, while LDA can show that "sustainability" is a prevalent theme, it might not clarify whether discussions transition to "economic factors" [77]. Furthermore, although LDA identifies various topics, the lack of predefined categories can result in less precise insights [25], as it can make hard to explain how important or semantically relevant each theme is as well as a supervised model would.

Hence, when comparing supervised and unsupervised approaches for topic modeling, supervised methods generally offer clearer and more precise insights. For instance, XGBoost as a supervised method can track how students shift from one topic to another during their argumentation, such as moving from technical design considerations to economic implications [78]. While unsupervised models are good at identifying a range of topics, they often lack the precision in quantifying the prevalence or importance of these themes compared to supervised models. Therefore, the results support the idea that while unsupervised models are useful for discovering main themes, a deeper qualitative analysis or a combination of both supervised and unsupervised methods with manual qualitative analysis is needed. Consequently, this approach can help provide a more comprehensive and interpretable understanding of how student argumentation progresses across the argumentation framework [75].

6. Conclusions

This research demonstrates the potential of using LDA to uncover latent themes in student responses and compares it with the effectiveness of supervised learning methods. The insights gained from this analysis can inform educational strategies and support the development of more effective and sustainable building practices. By exploring interpretability techniques in future work, researchers can further enhance the utility and transparency of these models in practical applications. Nevertheless, the inter-rater reliability achieved by the researchers, with a Cohen's Kappa of 0.80, indicates consistency and reliability in the analysis and interpretation, reinforcing the trustworthiness of the findings and suggesting that the methodologies employed can be replicated in future studies.

In addition, this study provides a comprehensive analysis of students' responses to building energy-efficient homes using both latent Dirichlet allocation (LDA) and supervised learning methods. The six themes identified through LDA offer valuable insights into the key aspects of sustainable residential construction, including energy efficiency, solar energy utilization, and economic considerations. The comparison between LDA and supervised learning methods highlights the complementary strengths of unsupervised and supervised approaches in understanding and categorizing student responses. Therefore, by utilizing both unsupervised techniques like LDA and supervised machine learning models, the analysis can reveal nuanced/hidden themes and patterns that traditional qualitative methods might overlook. This approach not only helps identify emerging topics but also provides a structured framework for interpreting qualitative data, significantly improving the reliability and accuracy of the findings. However, it is important to acknowledge the challenge of semantically incoherent topics that may arise, particularly when text responses are brief or lack sufficient depth and context for effective thematic analysis. In these cases, models may struggle to identify meaningful patterns, leading to less coherent and interpretable topics. Moreover, the performance of machine learning models can vary significantly across different datasets. Hence, a hybrid approach combining manual qualitative review with computational methods offers a more nuanced understanding of

the data. This approach enhances the overall quality and interpretability of the findings by ensuring that the identified topics are coherent and meaningful, while also validating and enriching the themes identified through computational techniques.

From a computational perspective, supervised and unsupervised approaches differ in coverage, efficiency, and interpretability. Supervised methods use labeled data, which restricts coverage to predefined topics and can be computationally intensive. In contrast, unsupervised methods are more flexible, covering a broader range of topics without needing labels and generally requiring less computation. However, in terms of interpretability, supervised methods are often clearer because topics are associated with specific labels, making their meaning easier to understand. Unsupervised methods, on the other hand, rely on data patterns without predefined labels, which makes it harder to interpret the meaning of emerging topics.

While this study provides valuable insights into topic interpretation using key terms and expert knowledge, some questions remain about the interpretability of supervised models. Nevertheless, future research could enhance interpretability using techniques like SHAP (SHapley Additive explanations) and LIME (local interpretable model-agnostic explanations). These methods can provide deeper insights into how supervised models make decisions for each instance, increasing transparency and trustworthiness. This approach has established a procedure to predict topic mixtures and reliably evaluate documents, supporting teaching processes. Further studies are needed to identify the specific contributions of predictors to target variables.

Author Contributions: Conceptualization, J.D.R., M.A.F.-G., A.J.M. and G.N.; methodology, G.N.; software, J.D.R.; validation, J.D.R., M.A.F.-G. and G.N.; formal analysis, J.D.R. and M.A.F.-G.; investigation, J.D.R., M.A.F.-G., G.N., A.J.M. and B.N.; resources, A.J.M. and B.N.; data curation, B.N.; writing—original draft preparation, J.D.R. and M.A.F.-G.; writing—review and editing, A.J.M. and G.N.; visualization, J.D.R.; supervision, G.N.; project administration, A.J.M. and B.N.; funding acquisition, A.J.M. and B.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the US National Science Foundation under award nos. 2406698 and 2219271 and the Purdue Polytechnic Research Impact Areas post-doctoral researcher awards. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or Purdue University.

Institutional Review Board Statement: This study was reviewed and approved by the Institutional Review Board, IRB-2020-1294, and was deemed exempt. Researchers not involved in the collection of the data were provided with de-identified data for analysis.

Informed Consent Statement: This study was approved as exempt by the institution's IRB, as it was conducted within established or commonly accepted educational settings and involved normal educational practices. These practices were unlikely to adversely affect students' opportunities to learn required educational content or the assessment of educators who provide instruction.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to privacy and confidentiality considerations.

Acknowledgments: The authors would like to acknowledge the Institute for Future Intelligence for their continuous support and free access to Aladdin software, and the Undergraduate Research Experience Purdue-Colombia (UREP-C), for the research mobility of visiting scholars.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Akintayo, O.T.; Eden, C.A.; Ayeni, O.O.; Onyebuchi, N.C. Evaluating the impact of educational technology on learning outcomes in the higher education sector: A systematic review. *Open Access Res. J. Multidiscip. Stud.* **2024**, *7*, 52–72. [CrossRef]
2. Valdez, D.; Picket, A.C.; Young, B.; Golden, S.D. On mining words: The utility of topic models in health education research and practice. *Health Promot. Pract.* **2021**, *22*, 309–312. [CrossRef] [PubMed]
3. Nanda, G.; Jaiswal, A.; Castellanos, H.; Zhou, Y.; Choi, A.; Magana, A.J. Evaluating the Coverage and Depth of Latent Dirichlet Allocation Topic Model in Comparison with Human Coding of Qualitative Data: The Case of Education Research. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 473–490. [CrossRef]
4. Wang, Y.; Sohn, S.; Liu, S.; Shen, F.; Wang, L.; Atkinson, E.J.; Amin, S.; Liu, H. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 1. [CrossRef]
5. Moore, B.A.; Wright, J. Constructing written scientific explanations: A conceptual analysis supporting diverse and exceptional middle-and high-school students in developing science disciplinary literacy. *Front. Educ.* **2023**, *8*, 1305464. [CrossRef]
6. McNeill, K.L.; Martin, D.M. Claims, evidence, and reasoning. *Sci. Child.* **2011**, *48*, 52.
7. Feijoo-Garcia, M.A.; Holstrom, M.S.; Magana, A.J.; Newell, B.A. Simulation-Based Learning and Argumentation to Promote Informed Design Decision-Making Processes within a First-Year Engineering Technology Course. *Sustainability* **2024**, *16*, 2633. [CrossRef]
8. Feijoo-Garcia, M.A.; Newell, B.; Magana, A.J.; Holstrom, M. Argumentation Framework as an Educational Approach for Supporting Critical Design Thinking in Engineering Education. In Proceedings of the 2024 ASEE Annual Conference & Exposition, Portland, OR, USA, 23–26 June 2024.
9. Vieira, C.; Ortega-Alvarez, J.D.; Magana, A.J.; Boutin, M. Beyond analytics: Using computer-aided methods in educational research to extend qualitative data analysis. *Comput. Appl. Eng. Educ.* **2024**, *32*, e22749. [CrossRef]
10. Bloomfield, J.; Fisher, M.J. Quantitative research design. *J. Australas. Rehabil. Nurses Assoc.* **2019**, *22*, 27–30. [CrossRef]
11. Roni, S.M.; Merga, M.K.; Morris, J.E. *Conducting Quantitative Research in Education*; Springer: Berlin/Heidelberg, Germany, 2020.
12. Tong, A.; Flemming, K.; McInnes, E.; Oliver, S.; Craig, J. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. *BMC Med. Res. Methodol.* **2012**, *12*, 181. [CrossRef]
13. France, E.F.; Cunningham, M.; Ring, N.; Uny, I.; Duncan, E.A.; Jepson, R.G.; Maxwell, M.; Roberts, R.J.; Turley, R.L.; Booth, A.; et al. Improving reporting of meta-ethnography: The eMERGe reporting guidance. *BMC Med. Res. Methodol.* **2019**, *19*, 25. [CrossRef] [PubMed]
14. Gauthier, R.P.; Wallace, J.R. The computational thematic analysis toolkit. *Proc. ACM Hum.-Comput. Interact.* **2022**, *6*, 1–15. [CrossRef]
15. Kherwa, P.; Bansal, P. Topic modeling: A comprehensive review. *EAI Endorsed Trans. Scalable Inf. Syst.* **2019**, *7*, e2. [CrossRef]
16. Nanda, G.; Douglas, K.A.; Waller, D.R.; Merzdorf, H.E.; Goldwasser, D. Analyzing Large Collections of Open-Ended Feedback From MOOC Learners Using LDA Topic Modeling and Qualitative Analysis. *IEEE Trans. Learn. Technol.* **2021**, *14*, 146–160. [CrossRef]
17. Zhao, W.; Zou, W.; Chen, J.J. Topic Modeling for Cluster Analysis of Large Biological and Medical Datasets. *BMC Bioinform.* **2014**, *15*, S11 [CrossRef]
18. Mohammadiha, N.; Smaragdis, P.; Leijon, A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 2140–2151. [CrossRef]
19. Wu, X.; Feng, C.; Li, Q.; Zhu, J. Keyword Pool Generation for Web Text Collecting: A Framework Integrating Sample and Semantic Information. *Mathematics* **2024**, *12*, 405. [CrossRef]
20. Çatir, O. UNDERSTANDING EMPLOYEE VOICE USING MACHINE LEARNING METHOD: EXAMPLE OF HOTEL BUSINESSES. *Geoj. Tour. Geosites* **2022**, *43*, 955–963. [CrossRef]
21. George, L.; Sumathy, P. An integrated clustering and BERT framework for improved topic modeling. *Int. J. Inf. Technol.* **2023**, *15*, 2187–2195. [CrossRef]
22. Grün, B.; Hornik, K. topicmodels: An R package for fitting topic models. *J. Stat. Softw.* **2011**, *40*, 1–30. [CrossRef]
23. Ning, X.; Yim, D.; Khuntia, J. Online sustainability reporting and firm performance: Lessons learned from text mining. *Sustainability* **2021**, *13*, 1069. [CrossRef]
24. Muchene, L.; Safari, W. Two-stage topic modelling of scientific publications: A case study of University of Nairobi, Kenya. *PLoS ONE* **2021**, *16*, e0243208. [CrossRef] [PubMed]
25. Rahmi, N.A.; Rudiman, R. Latent Dirichlet Allocation Utilization as a Text Mining Method to Elaborate Learning Effectiveness. *JSE J. Sci. Eng.* **2023**, *1*, 23–29.
26. Wang, W.; Guo, B.; Shen, Y.; Yang, H.; Chen, Y.; Suo, X. Neural labeled LDA: A topic model for semi-supervised document classification. *Soft Comput.* **2021**, *25*, 14561–14571. [CrossRef]
27. Zhou, S.; Zhao, Y.; Bian, J.; Haynos, A.F.; Zhang, R. Exploring eating disorder topics on Twitter: Machine learning approach. *JMIR Med. Inform.* **2020**, *8*, e18273. [CrossRef]
28. Gou, Z.; Huo, Z.; Liu, Y.; Yang, Y. A method for constructing supervised topic model based on term frequency-inverse topic frequency. *Symmetry* **2019**, *11*, 1486. [CrossRef]
29. Hou, Y.Y.; Li, J.; Chen, X.B.; Ye, C.Q. Variational quantum semi-supervised classifier based on label propagation. *Chin. Phys. B* **2023**, *32*, 070309. [CrossRef]

30. Kimura, M.; Izawa, R. Density-Fixing: Simple yet Effective Regularization Method based on the Class Priors. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
31. Engelen, J.E.v.; Hoos, H.H. A Survey on Semi-Supervised Learning. *Mach. Learn.* **2019**, *109*, 373–440. [CrossRef]
32. Hudon, A.; Phraxayavong, K.; Potvin, S.; Dumais, A. Ensemble methods to optimize automated text classification in avatar therapy. *BioMedInformatics* **2024**, *4*, 423–436. [CrossRef]
33. Onan, A. Hybrid supervised clustering based ensemble scheme for text classification. *Kybernetes* **2017**, *46*, 330–348. [CrossRef]
34. Li, H.; Ma, Z.; Zhu, H.; Ma, Y.; Chang, Z. An ensemble classification algorithm of micro-blog sentiment based on feature selection and differential evolution. *IEEE Access* **2022**, *10*, 70467–70475. [CrossRef]
35. Das, M.; Banerjee, S.; Saha, P. Abusive and threatening language detection in urdu using boosting based and bert based models: A comparative approach. *arXiv* **2021**, arXiv:2111.14830.
36. Osman, M.; He, J.; Mokbal, F.M.M.; Zhu, N.; Qureshi, S. ML-LGBM: A machine learning model based on light gradient boosting machine for the detection of version number attacks in RPL-based networks. *IEEE Access* **2021**, *9*, 83654–83665. [CrossRef]
37. Çano, E.; Morisio, M. Quality of word embeddings on sentiment analysis tasks. In *Natural Language Processing and Information Systems*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 332–338. [CrossRef]
38. Wang, K.J. Making hong kong film. In *Hong Kong Popular Culture*; Hong Kong Studies Reader Series; Springer: Berlin/Heidelberg, Germany, 2020; pp. 33–116. [CrossRef]
39. Gatto, J.; Seegmiller, P.; Johnston, G.; Preum, S.M. Identifying the perceived severity of patient-generated telemedical queries regarding covid: Developing and evaluating a transfer learning–based solution. *JMIR Med. Inform.* **2022**, *10*, e37770. [CrossRef] [PubMed]
40. Lin, H.; Bu, N. A cnn-based framework for predicting public emotion and multi-level behaviors based on network public opinion. *Front. Psychol.* **2022**, *13*, 909439. [CrossRef] [PubMed]
41. Razali, M.N.; Mustapha, A.; Mostafa, S.A.; Gunasekaran, S.S. Football matches outcomes prediction based on gradient boosting algorithms and football rating system. *Hum. Factors Softw. Syst. Eng.* **2022**, *61*, 57.
42. Al Hanai, T.; Ghassemi, M.M.; Glass, J.R. Detecting Depression with Audio/Text Sequence Modeling of Interviews. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 1716–1720.
43. Gurcan, F.; Cagiltay, N.E. Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. *IEEE Access* **2019**, *7*, 82541–82552. [CrossRef]
44. Sydorenko, S.; Kuzminska, O.; Mazorchuk, M.; Barna, O. Bibliometric analysis in determining the research directions of early career researchers. *Inf. Technol. Learn. Tools* **2022**, *5*, 113–129.
45. Sanfilippo, F.; Austreng, K. Enhancing teaching methods on embedded systems with project-based learning. In Proceedings of the 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), Wollongong, Australia, 4–7 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 169–176.
46. Ariza, J.A.; Baez, H. Understanding the role of single-board computers in engineering and computer science education: A systematic literature review. *Comput. Appl. Eng. Educ.* **2022**, *30*, 304–329. [CrossRef]
47. Dolgopolovas, V.; Dagienė, V. Computational thinking: Enhancing STEAM and engineering education, from theory to practice. *Comput. Appl. Eng. Educ.* **2021**, *29*, 5–11. [CrossRef]
48. Shaik, T.; Tao, X.; Li, Y.; Dann, C.; McDonald, J.; Redmond, P.; Galligan, L. A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *IEEE Access* **2022**, *10*, 56720–56739. [CrossRef]
49. Fahlevvi, M.R. Sentiment Analysis And Topic Modeling on User Reviews of Online Tutoring Applications Using Support Vector Machine and Latent Dirichlet Allocation. *Knowbase Int. J. Knowl. Database* **2022**, *2*, 142–155. [CrossRef]
50. Gao, C.; Zeng, J.; Wen, Z.; Lo, D.; Xia, X.; King, I.; Lyu, M.R. Emerging app issue identification via online joint sentiment-topic tracing. *IEEE Trans. Softw. Eng.* **2021**, *48*, 3025–3043. [CrossRef]
51. Wang, Z. Extracting latent topics from user reviews using online LDA. In Proceedings of the 2018 International Conference on Information Technology and Management Engineering (ICITME 2018), Beijing, China, 26–27 August 2018; Atlantis Press: Amsterdam, The Netherlands, 2018; pp. 204–208.
52. Qiang, J.; Qian, Z.; Li, Y.; Yuan, Y.; Wu, X. Short text topic modeling techniques, applications, and performance: A survey. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 1427–1445. [CrossRef]
53. Xie, C.; Ding, X.; Jiang, R. Using Computer Graphics to Make Science Visible in Engineering Education. *IEEE Comput. Graph. Appl.* **2023**, *43*, 99–106. [CrossRef] [PubMed]
54. Feijóo-García, M.A.; Ramírez-Arévalo, H.H.; García, P.G.F. Collaborative Strategy for Software Engineering Courses at a South American University. In Proceedings of the CSEDU (2), Online, 23–25 April 2021; pp. 266–273.
55. Tabula. 2023. Available online: https://tabula.technology/ (accessed on 11 June 2024).
56. HaCohen-Kerner, Y.; Miller, D.C.; Yigal, Y. The influence of preprocessing on text classification using a bag-of-words representation. *PLoS ONE* **2020**, *15*, e0232525. [CrossRef]
57. Selection of the Optimal Number of Topics for LDA Topic Model-Taking Patent Policy Analysis as an Example. *Entropy* **2023**, *23*, 1301. [CrossRef]

58. Hagg, L.J.; Merkouris, S.S.; O'Dea, G.A.; Francis, L.M.; Greenwood, C.J.; Fuller-Tyszkiewicz, M.; Westrupp, E.M.; Macdonald, J.A.; Youssef, G.J. Examining analytic practices in latent dirichlet allocation within psychological science: Scoping review. *J. Med. Internet Res.* **2022**, *24*, e33166. [CrossRef] [PubMed]
59. Campagnolo, J.M.; Duarte, D.; Dal Bianco, G. Topic coherence metrics: How sensitive are they? *J. Inf. Data Manag.* **2022**, *13*. [CrossRef]
60. Röder, M.; Both, A.; Hinneburg, A. Exploring the space of topic coherence measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China, 2–6 February 2015; pp. 399–408.
61. Zhou, K.; Wang, J.; Ashuri, B.; Chen, J. Discovering the Research Topics on Construction Safety and Health Using Semi-Supervised Topic Modeling. *Buildings* **2023**, *13*, 1169. [CrossRef]
62. Jensen, F.B.; Kuperman, W.A.; Porter, M.B.; Schmidt, H. *Computational Ocean Acoustics*; Springer: Berlin/Heidelberg, Germany, 1995; Volume 121. [CrossRef]
63. Mimno, D. Mallet: MAchine Learning for LanguagE Toolkit. Available online: http://mallet.cs.umass.edu (accessed on 11 June 2024).
64. Murshed, B.A.H.; Mallappa, S.; Abawajy, J.; Saif, M.A.N.; Al-Ariki, H.D.E.; Abdulwahab, H.M. Short text topic modelling approaches in the context of big data: Taxonomy, survey, and analysis. *Artif. Intell. Rev.* **2023**, *56*, 5133–5260. [CrossRef] [PubMed]
65. Martino, L.; Elvira, V.; Camps-Valls, G. The recycling Gibbs sampler for efficient learning. *Digit. Signal Process.* **2018**, *74*, 1–13. [CrossRef]
66. Bisgin, H.; Liu, Z.; Fang, H.; Xu, X.; Xu, X.; Tong, W. Mining FDA drug labels using an unsupervised learning technique—Topic modeling. *BMC Bioinform.* **2011**, *12*, S11. [CrossRef]
67. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
68. Sagi, O.; Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
69. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
70. Warrens, M.J. Five ways to look at Cohen's kappa. *J. Psychol. Psychother.* **2015**, *5*, e197. [CrossRef]
71. Buch, A. Ideas of holistic engineering meet engineering work practices. In *Engineering Professionalism*; Brill: Leiden, The Netherlands, 2016; pp. 145–169.
72. Wan, X.; Wang, T. Automatic labeling of topic models using text summaries. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 2297–2305.
73. Tan, Y.; Ou, Z. Topic-weak-correlated latent dirichlet allocation. In Proceedings of the 2010 7th International Symposium on Chinese Spoken Language Processing, Tainan, Taiwan, 29 November–3 December 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 224–228.
74. Wang, Y.; Pan, Z.; Zheng, J.; Qian, L.; Li, M. A hybrid ensemble method for pulsar candidate classification. *Astrophys. Space Sci.* **2019**, *364*, 139. [CrossRef]
75. Mathis, C.A.; Siverling, E.A.; Glancy, A.W.; Moore, T.J. Teachers' incorporation of argumentation to support engineering learning in STEM integration curricula. *J. Pre-Coll. Eng. Educ. Res. (J-PEER)* **2017**, *7*, 6. [CrossRef]
76. Liu, Y.; Wang, H.; Fei, Y.; Liu, Y.; Shen, L.; Zhuang, Z.; Zhang, X. Research on the prediction of green plum acidity based on improved XGBoost. *Sensors* **2021**, *21*, 930. [CrossRef]
77. Meisert, A.; Böttcher, F. Towards a discourse-based understanding of sustainability education and decision making. *Sustainability* **2019**, *11*, 5902. [CrossRef]
78. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.