International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Unsupervised Machine Learning Approaches in NLP: A Comparative Study of Topic Modeling with BERTopic and LDA

Christian Y. Sy *[1], Lany L. Maceda[2], Nancy M. Flores[3], Mideth B. Abisado[4]

Submitted: 05/02/2024 **Revised**: 13/03/2024 **Accepted**: 19/03/2024

Abstract*:* This research aimed to understand the issues and challenges encountered by beneficiaries of the Philippines' Universal Access to Quality Tertiary Education (UAQTE) program, using a comparative analysis of BERTopic and Latent Dirichlet Allocation (LDA) topic modeling techniques. The "Boses Ko" or "My Voice" toolkit was utilized to gather student responses from the ground up. The study found that BERTopic excelled in semantic relevance and coherence, while LDA effectively formed distinct clusters. The evaluation combined automatic metrics, such as silhouette and coherence scores, with domain experts' insights. Key themes identified included "Academic Difficulties," "Financial Difficulties," "Grant Disbursement," "Pandemic-Related Challenges," and "Program Implementation." The research concluded with actionable recommendations for the UAQTE program, advocating for enhanced academic support, improved financial assistance, flexible grant disbursement, strategies to tackle pandemic-related challenges, and establishing a structured feedback mechanism. These suggestions guide policy reforms, encouraging continuous evaluation to ensure long-term effectiveness in the educational sector. Overall, this study provides valuable insights into the application of topic modeling in educational policy analysis and emphasizes the need for nuanced model selection and interpretation for impactful policy development.

Keywords: *unsupervised machine learning, natural language processing (NLP), topic modeling, BERTopic, Latent Dirichlet Allocation (LDA), UAQTE program.*

1. Introduction

In a global landscape where the pursuit of fair educational opportunities transcends borders, initiatives aimed at offering free tertiary education are tailored to ensure unimpeded access to higher learning while dismantling the financial barriers that have long hindered students, particularly in developing nations, from pursuing tertiary education studies [1], [2]. It enables individuals to escape the recurring cycle of poverty and propels progress at the national level, thereby playing a crucial role in socioeconomic development. Tertiary education equips students with the knowledge, skills, and critical thinking abilities needed to overcome the intricate challenges of today's world. This, in turn, broadens their horizons, offering a wide array of career opportunities and empowering graduates to make meaningful contributions to their communities and nation [3].

Philippines, a nation marked by enduring economic and social disparities, tertiary education takes on a transformative role. It serves as a means of social upward mobility, affording individuals from economically disadvantaged backgrounds the same educational opportunities as their more privileged counterparts [4], [5]. Subsequently, it mitigates disparities and cultivates an environment characterized by inclusivity within the higher education domain, aligning with the broader goal of creating a fair and inclusive society [6].

The significance of tertiary education aligns with the broader global agenda, notably encapsulated within the fourth Sustainable Development Goal (SDG) [7], [8]. Recognized by the United Nations, "Quality Education" is an indispensable foundation that facilitates the achievement of sustainable development [9], [10]. With its intrinsic influence to effect significant personal and societal reforms, tertiary education plays a crucial role in achieving this goal [11]. Realizing this, the Universal Access to Quality Tertiary Education (UAQTE) program, also known as Republic Act No. 10931, was enacted on August 13, 2017, requiring all public higher education institutions (HEIs) and government-run technical-vocational institutions (TVIs) to provide free quality tertiary education to eligible Filipino students. This program aligns to provide every Filipino with the opportunity to pursue higher education by removing financial obstacles that have traditionally discouraged students from enrolling in colleges and universities. This praiseworthy effort symbolizes the government's dedication to creating a society where education is considered a right rather than a privilege.

As educational opportunities continue to broaden, it becomes increasingly critical to examine the multifaceted issues and challenges encountered by the beneficiaries of these programs [12]. This understanding significantly influences initiatives like the UAQTE program, directly enriching the educational journey of its recipients. It is pivotal in determining the program's success and effectiveness, ensuring a more fulfilling educational experience. This research aims to comprehensively understand the diverse issues and challenges encountered by beneficiaries of the UAQTE program. Addressing these complexities is essential for improving the program's effectiveness, enriching the educational experience of student beneficiaries, and positively impacting the broader educational landscape. The study implements topic modeling techniques, specifically BERTopic and Latent Dirichlet Allocation (LDA), to identify patterns and correlations in the data. This approach offers an in-depth insight into the implementation of the UAQTE program and its effects on stakeholders.

[1,2] *Computer Science and Information and Technology Department, Bicol University, Legazpi City, Philippines*
[3] *University of the Cordilleras, Baguio City, Philippines*
[4] *College of Computing and Information Technology, National University, Metro Manila, Philippines*
** Corresponding Author Email: cysy@bicol-u.edu.ph*

Topic modeling, within the domain of natural language processing and data analysis, plays a significant role in understanding and extracting meaningful insights from textual data [13] – [16]. It is particularly valuable in contexts where large volumes of unstructured text data need to be organized, categorized, and summarized [17], [19]. This investigation seeks to offer evidence-based insights aimed at facilitating policy reforms, ultimately improving and optimizing the UAQTE program. This study utilizes the "Boses Ko" or “My Voice” participatory toolkit, a digital platform resulting from a collaboration between Bicol University (BU) and National University (NU), funded by the Commission on Higher Education - Leading the Advancement of Knowledge in Agriculture and Sciences (CHED-LAKAS) program, emphasizing research and development efforts in science and technology. Through the "Boses Ko" toolkit, student beneficiaries can express their perspectives and articulate the issues and challenges encountered within the UAQTE program.

Building upon the pre-processed dataset, this study leverages two distinct qualitative modeling techniques: BERTopic and LDA (Latent Dirichlet Allocation), each possessing unique strengths in revealing hidden patterns and groupings within stakeholders' viewpoints. Combining BERTopic and LDA for topic modeling merges the strengths of BERT's semantic depth and LDA's interpretability, offering a more comprehensive analysis of stakeholders' perspectives on UAQTE implementation. BERTopic utilizes the state-of-the-art BERT (Bidirectional Encoder Representations from Transformers) language model [20]–[23] to explore semantic intricacies within textual data. LDA, a well-established methodology, takes a more conventional approach to achieve similar objectives [24]–[27].

BERTopic's use of BERT embeddings offers a comprehensive view of the semantic context of the text [28], while LDA's term-based analysis brings a traditional yet interpretable perspective [29]. This balances the feature representation, providing semantic context and term frequency information.

The evaluation of the generated models adopts a comprehensive approach that combines automated metrics, including the Silhouette Score and Coherence Score, with manual assessment of topics, incorporating insights from domain experts. Automated metrics offer efficiency, objectivity, and a quantitative foundation for assessing topics, facilitating rapid model refinement and baseline assessment. Domain expert views introduce vital qualitative perspectives, considering factors like relevance, interpretability, and real-world applicability, which automated metrics alone cannot capture [30], [31]. This approach ensures a more holistic evaluation, enhancing the model's reliability and relevance to specific domains and applications while thoroughly exploring the varied viewpoints of stakeholders regarding the UAQTE implementation, revealing valuable insights that form the basis for evidence-based policy reforms.

Silhouette Score is a metric used to evaluate the quality of clustering. It measures how similar each data point in one cluster is to the other data points in the same cluster compared to the nearest neighboring cluster [32]. On the other hand, Coherence Score is used to evaluate the quality and interpretability of topics in topic modeling. It measures how semantically related the top words within a topic are and how well the topic forms a coherent and meaningful theme [33], [34]. Drawing from their specialized knowledge and insightful judgment, domain experts actively contribute to the interpretation and assessment of models. Their contribution ensures that the topics generated are relevant, coherent, and aligned with the specific nuances of the domain [35]. This facilitates subjective assessment, empowering evaluators to consider factors like interpretability, coherence, and relevance, which hold significance in real-world applications. Furthermore, it supplements model refinement and incorporates domain-specific expertise, ultimately enriching contextual comprehension.

Ultimately, this research aims to generate data-driven insights for policy reforms, leading to the enhancement and optimization of the UAQTE program to better serve students and the wider educational sector. Central to this approach is the emphasis on collaboration and active participation, which empowers stakeholders to significantly contribute to shaping the UAQTE program. Their involvement is crucial in refining current strategies and influencing the direction of future policies in tertiary education, ensuring that they are more aligned with the actual needs and challenges faced by students.

2. Methodology

The methodology employed is outlined in this section. Fig. 1, representing the information processing phases, delineates the steps, encompassing data collection and dataset pre-processing, feature extraction, topic modeling, topic interpretation, labeling, and model evaluation.

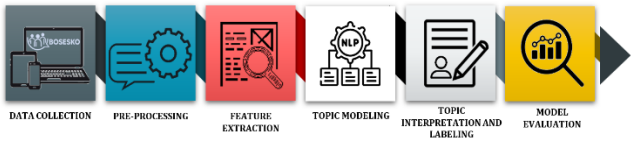

Fig 1. Information Processing Phase

2.1. Data Collection

The "Boses Ko" or "My Voice" toolkit has been instrumental in adopting a grassroots approach to gathering data from student beneficiaries, prioritizing the perspectives of those directly involved in the UAQTE program. Specifically fitted to assess the viewpoints of student beneficiaries, the qualitative question guiding this study is, "What are the issues and challenges encountered as one of the beneficiaries of the UAQTE program?" A sample size of 2,800 student beneficiaries, selected from State Universities and Colleges (SUCs) in the Bicol region, was incorporated into the study. This diverse representation facilitated a comprehensive examination of the UAQTE program's implementation across various institutional contexts and frameworks.

2.2. Data Pre-processing

Preparing the collected responses before feature extraction involved vital data pre-processing steps. These crucial procedures included the removal of non-English, duplicate, non-grantee, and empty responses. Subsequently, the cleaned dataset was subjected to text standardization, encompassing procedures such as converting the text to lowercase and removing special characters, punctuation marks, and digits. This streamlining process aimed to create a more coherent and refined textual representation, reducing noise and potential interference in the modeling tasks, thereby enhancing its suitability for subsequent analysis.

Tokenization and removing stopwords were essential pre-processing tasks accomplished by employing the Natural Language Toolkit (NLTK) library. Tokenization divided the responses into individual words or tokens, facilitating easier analysis and manipulation of the text data. Within the responses, common stopwords like "the," "is," "this," "and," "it," "for," "of," and "in" were frequently found, yet individually carried limited semantic value. Eliminating these stopwords contributed to the

quality, interpretability, and efficiency of the generated topics within the UAQTE framework. This process optimized the topic modeling outcomes by reducing noise and highlighting content words that conveyed the core themes.

Similarly, incorporating domain-specific words as additional stopwords like "UAQTE," "issues," "challenges," and "beneficiaries" provided several benefits in text analysis. It reduces noise by eliminating specialized terms that may not be pertinent, resulting in more focused topics. Furthermore, highlighting essential terms over common ones produced more interpretable and meaningful results. On the other hand, the study opted not to utilize stemming and lemmatization techniques, as these approaches can oversimplify words by reducing them to their most basic forms. This oversimplification could risk the loss of essential meaning. For instance, if words like "synchronous" and "arrangements" were stemmed to "synchron" and "arrang," respectively, it could render them unrecognizable and potentially introduce confusion and a loss of textual clarity. Similarly, the lemmatization of "better" to "good" alters the comparative aspect and may impact the overall message.

2.3. Feature Extraction

BERTopic leverages BERT embeddings, which are dense vector representations capturing contextual information about words. This process involves acquiring BERT embeddings for each document using pre-trained BERT models [36], [37]. These embeddings, generated through this approach, are the foundational features for document clustering based on their similarity. BERTopic, stands out due to its reliance on pre-trained BERT models. What truly distinguishes BERTopic is its remarkable ability to grasp the subtle nuances of context and semantics. The process begins with tokenization, breaking each document into individual sub-tokens, which are then associated with BERT word vectors. This is accomplished by the BERT models, which consider the words around them to create contextual embeddings [38]. BERTopic utilizes pooling techniques like mean and max pooling to produce standardized vectors for every document. This creates a rich repository of dense vector representations that comprehensively capture information regarding word meanings and context [39].

Whereas in the feature extraction process of LDA, two of the most commonly employed techniques are the Bag of Words (BoW) representation and the Term Frequency-Inverse Document Frequency (TF-IDF) transformation. These methods are the foundation for revealing hidden structures and gaining deeper insights from text data. The Bag of Words model represents each document as a vector, where each element corresponds to the frequency of specific words. LDA employs this model to convert text into numerical features, allowing it to uncover latent topics and understand their distribution across the entire document collection. TF-IDF further enriches the quality of feature representation; after implementing the Bag of Words (BoW), TF-IDF is applied to enhance the effectiveness of topic modeling. In this process, each word is assigned a weight that represents its importance within individual documents and across the entire corpus. Integrating TF-IDF after the Bag of Words leads to a more informative, differentiating, and resource-efficient feature matrix.

This approach enables LDA to assign greater significance to words that are distinctive to particular documents and topics, while diminishing the importance of common words. Implementing TF-IDF after Bag of Words is crucial as it refines the feature representation, reduces noise, and improves the distinction of topics in textual data. It ensures that the topic modeling process is more attuned to the unique characteristics of the dataset, resulting in more meaningful and accurate topic assignments.

$$tf - idf(t) = tf(t,d)\, x\, idf(t) \quad (1)$$

Equation 1 evaluates the significance of the term 't' within a document 'd' across a document collection. It considers two key elements: the term's frequency within the document ('tf') and its rarity or uniqueness in the entire document collection ('idf').

2.4. Topic Modeling

Following the implementation of document embeddings, BERTopic utilizes Uniform Manifold Approximation and Projection (UMAP), a dimensionality reduction technique. UMAP transforms the high-dimensional embeddings into a lower-dimensional space while preserving the inherent structure and relationships within the data. This reduction is instrumental in enhancing the effectiveness of data visualization and clustering [40]. To extract meaningful topics from the dataset, BERTopic leverages the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) clustering algorithm applied to the lower-dimensional UMAP space. HDBSCAN adeptly identifies dense clusters of data points, effectively representing topics or subtopics in the dataset [41], [42].

HDBSCAN's adaptability in identifying clusters of various shapes and sizes makes it well-suited for diverse datasets. Nevertheless, what distinguishes BERTopic is its remarkable ability to autonomously detect the number of topics, a feature that frees researchers from the need to specify the topic count in advance. By analyzing data-driven insights into the density and distribution of document vectors, BERTopic distinguishes natural cluster boundaries, streamlining the topic modeling process for enhanced efficiency. Fig. 2 illustrates the workflow from document embeddings using BERT modeling to topic representation.

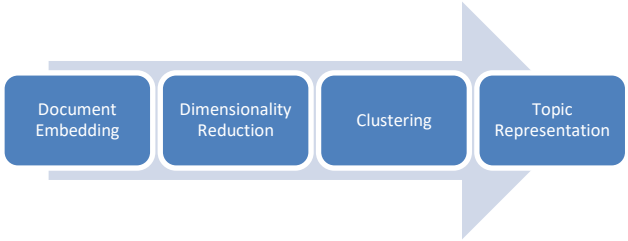

Fig 2. Topic Extraction using BERTopic

For LDA, following the initiation of the feature extraction process with BoW, where each document is represented as a vector of word frequencies, LDA steps in to uncover latent topics within the corpus. LDA operates on the document-term matrix derived from BoW, treating each document as a mixture of topics and each word's occurrence as attributable to one of these topics. The algorithm iteratively refines its estimates, adjusting the topic assignments for each word based on the co-occurrence patterns observed across the entire corpus.

Incorporating TF-IDF as a refinement of the BoW representation further enriches the feature vectors. TF-IDF assigns weights to each term, considering its frequency within individual documents and its importance across the entire corpus. This TF-IDF matrix becomes the input for LDA, which then analyzes the significance

and distribution of terms to identify underlying topics. The algorithm distinguishes topics based on the prevalence of specific terms across documents, aiming to capture the semantic relationships among words.

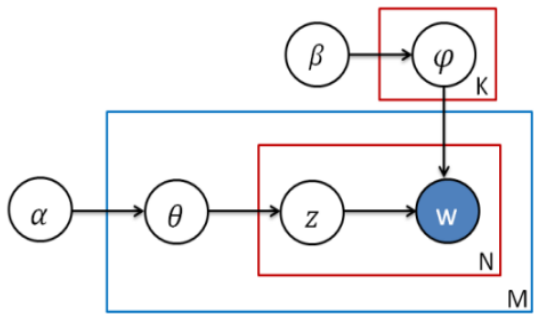

Fig 3. Topic Extraction using LDA

Fig. 3 illustrates the generative process within the Latent Dirichlet Allocation (LDA) model. It presents a structured view of how documents are formed, encompassing the selection of topics, the distribution of words within these topics, and the influence of hyperparameters α and β on topic and word distributions.

- α (alpha). Represents the parameter influencing the Dirichlet prior to the per-document topic distributions (θ). It determines the distribution of topics across individual documents.
- β (beta). Serves as the parameter controlling the Dirichlet prior for the per-topic word distributions (φ). It defines the distribution of words within topics.
- θ(theta)M. Represents the topic distribution specific to document M. It illustrates the mixture of topics present within a particular document.
- φ(phi)K. Symbolizes the word distribution for topic K. It showcases the likelihood of each word occurring within the specific topic K.
- Zmn. Denotes the assignment of topics to individual words, specifically the topic assigned to the nth word in document m.
- Wmn. Stands for the individual words themselves within the corpus. Each Wmn represents a distinct word in the document, and in conjunction with Zmn, it illustrates the specific word assigned to a particular topic within the document.

2.5. Hyper-Parameters

The following are the hyper-parameters used for BERTopic:

- min_topic_size. Defines the minimum document count for valid topics. Adjusting it affects topic size and granularity, potentially merging or discarding topics with too few documents.
- top_n_words. Sets the number of top words displayed per topic, typically the most representative terms. Choosing a specific top_n_words value helps understand each topic's key terms for better interpretation.
- num_topics. Defines the desired number of topics extracted from your dataset. Choose it based on your dataset's nature and expected topic count.

Table 1 presents the range of hyperparameters employed to generate BERTopic models, reflecting the numerous experiments conducted to ascertain the optimal settings for producing coherent and representative topics from the given corpus.

Table 1. Hyper-Parameters Used for BERTopic

Number of topics	*Top n words*	*Min topic size*
5 to 20	5 to 10	10 to 30

For LDA, the following are the hyper-parameters used:

- Number of Topics (K). Determines the granularity and diversity of themes or subjects extracted from the corpus.
- Number of Words. Refers to the vocabulary size or the maximum number of unique words considered in the analysis.
- Passes. Specifies the number of times the algorithm passes through the entire corpus during training. Each pass involves updating the topic distributions for documents and word distributions for topics.
- Iterations. Refer to the number of iterations within each pass through the corpus. It controls the number of times the model iterates over the entire dataset to refine topic assignments.
- Alpha (α). Influences the distribution of topics across documents. A higher alpha result in documents containing more topics, while a lower alpha leads to sparser topic distributions within documents.
- Eta (η). Influences the sparsity of per-topic word distributions. Higher values encourage broader word distributions for topics, whereas lower values lead to more focused word distributions.

Table 2. Hyper-Parameters Used for LDA

Number of topics	*Number of words*	*Passes*	*Iterations*	*alpha*	*eta*
5 to 20	5 to 10	10 to 100	100 to 5000	0.01	0.01

Table 2 presents the variations across multiple experiments to identify the most suitable hyperparameter configurations that ensure the generation of coherent, interpretable, and contextually relevant topics from the underlying dataset using the LDA model. The systematic investigation of diverse hyperparameter combinations within these experiments was pivotal in unveiling the optimal settings.

Properly tuning these is vital to ensure that the topic modeling process aligns with the unique characteristics of the dataset, influencing the quality and interpretability of the topics extracted. Such settings are essential as they enable the extraction of informative and meaningful topics, crucial for enhancing comprehension and knowledge extraction from the dataset. The quality of topics is assessed by their interpretability and relevance to the specific research goals, particularly when employing topic modeling techniques such as BERTopic and LDA.

2.6. Topic Interpretation and Labeling

The objective of interpreting a topic model is to assign relevant labels based on observed word similarities within the generated models. This approach significantly improved topic comprehension and facilitated effective communication, thus enabling practical applications in shaping the UAQTE program. Furthermore, it played a crucial role in providing a more informed assessment of the program's impact and guiding future policy decisions pertaining to tertiary education.

Collaborating closely with experts across diverse fields, including Commission of Higher Education (CHED) administrators, social scientists, data scientists, and UAQTE recipients, facilitated the selection of more comprehensible and research-aligned topic labels. Table 3 exhibits the curated labels derived through this collaborative effort.

Table 3. Domain-Experts Identified Labels

Categories	*Description*
Financial Difficulties	Refers to the challenges faced due to limited financial resources, including difficulties in budgeting, paying for miscellaneous fees, boarding, daily expenses, and additional educational costs not covered by the program or scholarship.
Grant Disbursement	Refers to issues of scholarships or financial aid being delayed in their disbursement, causing financial strain and difficulties in meeting educational expenses.
Academic Difficulties	Refers to the challenges encountered in fulfilling various academic obligations, such as preparing and submitting requirements, managing coursework, and striving to achieve satisfactory academic performance. Other factors like lack of or poor facilities and infrastructures are likewise part of this category.
Pandemic-Related Challenges	Refers to the obstacles and difficulties arising from the COVID-19 pandemic, including the transition to online learning, lack of access to reliable internet connectivity and necessary devices, disruptions in academic schedules, and the impact on mental health, stress, anxiety, and feelings of isolation.
Program Implementation	Refers to the diverse viewpoints regarding the program's implementation, incorporating various positive and negative perspectives.

2.7. Model Evaluation

The Silhouette score is calculated as follows:

$$s = \frac{b - a}{max(a, b)} \quad (2)$$

The silhouette score (s) of a data point is determined by comparing its average distance (a) to other points within the same cluster and its average distance (b) to points in the nearest neighboring cluster. When a silhouette score for an object approaches a high positive value, nearly +1, it signifies a strong alignment of the object with its designated cluster and a considerable dissimilarity from neighboring clusters. This scenario suggests well-defined and distinct clusters that are appropriately separated. Conversely, a silhouette scores close to zero indicates that the object resides near or precisely on the boundary between two neighboring clusters, implying a certain level of ambiguity in cluster assignments. Lastly, when the silhouette score approaches a negative value, nearly -1, it signals that the object might be erroneously assigned to a neighboring cluster rather than its own. This implies potential issues such as cluster overlap or poorly defined cluster boundaries, impacting the accuracy and cohesion of the clustering structure. Overall, interpreting silhouette scores aids in understanding clusters' separation and clarity levels, guiding assessments regarding cluster quality and potential overlaps within the dataset. The Coherence score is calculated as follows:

$$C\_V(T) = 2 / (|T|(|T| - 1)) * \sum\_\{i = 1\}\^\{|T|\} \sum\_\{j \neq i\} sim(Word\_i, Word\_j) \quad (3)$$

The coherence score (CV (T)) for a given topic T is calculated based on the number of words in the topic (|T|), the representation of two distinct words (Wordi and Wordj) within that topic, and the similarity measure (sim (Wordi, Wordj)) between these word pairs. The coherence score provides valuable insights into the meaningfulness and connectedness of words within a topic. A coherence score near 0 suggests the topic lacks substantial connections among words, resulting in a challenging interpretation. When the coherence score ranges from approximately 0.2 to 0.4, it indicates some coherence within the topic, but the overall interpretability remains constrained. A score between 0.4 and 0.6 implies reasonably coherent topics, enhancing their interpretability. As the coherence score reaches approximately 0.6 to 0.8, it signifies well-defined topics with high coherence, rendering them easily interpretable. A coherence score nearing 0.8 to 1 denotes exceptional topics, showcasing closely related words that significantly contribute to high interpretability within the given topic.

Domain experts are crucial in the evaluation process, contributing invaluable contextual knowledge and subject matter expertise. Their feedback is essential due to the multidimensional criteria that topic models need to fulfil: statistically robust, semantically meaningful, and contextually relevant within the specific domain. By leveraging their expertise, domain experts serve as crucial validators, ensuring that the generated topics align with the context of the research objectives, thereby enhancing the credibility and applicability of the generated insights. When paired with coherence scores and the examination of topics by domain experts, Silhouette scores constitute a robust evaluation approach. Good silhouette scores, coherence scores, and interpretable topics collectively signify the commendable quality of generated topics. This combination of quantitative metrics and the insightful judgment of domain experts forms a comprehensive assessment framework, ensuring a nuanced and reliable evaluation of the overall quality and effectiveness of the topic model.

3. Results and Discussion

This section presents key results and findings obtained through topic modeling the UAQTE dataset using the BERTopic and LDA approaches. The following results in Tables 4 and 5 highlight the configurations that achieved acceptable silhouette and coherence scores from the numerous topic modeling experiments. These scores serve as quantitative assessments, gauging the efficacy and relevance of outputs generated by the algorithms. Higher scores in both silhouette and coherence metrics typically signify excellent results. Higher silhouette scores point to well-defined clusters, while higher coherence scores indicate more easily interpretable topics within the data.

Table 4. BERTopic Hyperparameters and Evaluation Scores

Num of topics	*Min topic Size*	*Top n words*	*Silhouette Score*	*Coherence Score*
5	30	10	0.711	0.881
6	30	10	0.742	0.882
7	15	10	0.610	0.879
7	30	10	0.696	0.889
7	25	10	0.746	0.872
8	25	10	0.764	0.882
9	25	10	0.744	0.877
10	20	10	0.640	0.878
10	20	10	0.681	0.876
11	15	10	0.643	0.877
13	20	10	0.645	0.875
15	15	10	0.684	0.876

Table 4 presents the analysis conducted using BERTopic to assess the impact of variations in the number of topics, minimum topic size, and top words on topic quality, evaluating silhouette and coherence scores. Changes in the number of topics and minimum topic size led to fluctuations in silhouette scores, reaching a peak at eight topics. However, coherence scores remained consistently higher, approximately between 0.870-0.890, indicating stable and relevant topic connections. Conversely, alterations in the selection

of top words did not influence the scores.

Table 5. LDA Hyperparameters and Evaluation Scores

Num of topics	*Number of Words*	*Passes*	*Iterations*	*alpha*	*eta*	*Silhouette Score*	*Coherence Score*
8	10	20	1000	0.01	0.01	0.771	0.618
8	10	50	1000	0.01	0.01	0.761	0.594
8	10	10	3000	0.01	0.01	0.769	0.617
8	10	50	3000	0.01	0.01	0.789	0.623
8	10	30	5000	0.01	0.01	0.761	0.591
8	10	100	5000	0.01	0.01	0.770	0.627
10	10	20	1000	0.01	0.01	0.728	0.603
10	10	50	1000	0.01	0.01	0.746	0.596
10	10	10	3000	0.01	0.01	0.740	0.605
10	10	50	3000	0.01	0.01	0.739	0.604
10	10	10	5000	0.01	0.01	0.735	0.645
10	10	50	5000	0.01	0.01	0.771	0.634

On the other hand, the LDA analysis explored different configurations of topics, words, passes, and iterations, observing fluctuations in silhouette and coherence scores across diverse parameter settings, as seen in Table 5. While BERTopic exhibited higher coherence scores overall, LDA displayed higher silhouette scores in certain settings, such as eight topics with 50 passes and 3000 iterations. This suggests that while BERTopic tended to maintain greater semantic relevance across topics, LDA excelled in defining more distinct clusters within the data, as evidenced by higher silhouette scores in specific setups.

Although LDA generally yielded coherence scores ranging from 0.591 to 0.645, signifying occasional challenges in capturing nuanced semantic relationships due to its probabilistic nature, it demonstrated a similar level of effectiveness as BERTopic in modeling associations between topics and words. Notably, when LDA coherence scores reached the range of 0.6 to 0.80, they indicated the generation of particularly coherent and meaningful topics. This range denotes well-defined topics with high coherence, rendering them easily interpretable and underscoring LDA's capacity to provide insightful dataset representations. This aspect further complements BERTopic's consistent ability to maintain semantic relevance.

The observed results showcase the inherent strengths of each method: BERTopic excels in semantic relevance and contextual coherence, while LDA might outperform in delineating more distinct and separated clusters. The differences in results highlight each model's nuanced strengths and preferences, emphasizing the need to consider the specific goals and nuances of the dataset when selecting a topic modeling approach despite the general state-of-the-art status of BERTopic. Moreover, the findings highlight the importance of meticulous hyperparameter selection and fine-tuning aligned with dataset characteristics for optimal topic modeling outcomes. By leveraging both BERTopic and LDA, researchers can gain a more comprehensive insight into the dataset, using the semantic richness of BERTopic alongside the cluster delineation expertise of LDA.

Table 6 showcases a BERTopic model meticulously labeled by domain experts and configured with specific hyperparameters: 8 topics, 10 top words per topic, and a minimum topic size of 25. Notably, the model demonstrates balanced performance, with a silhouette score of 0.764, indicating reasonably well-separated clusters. Additionally, it achieves a significant coherence score of 0.882, suggesting high semantic coherence and interpretability among the identified topics. These metrics affirm the model's effectiveness in extracting meaningful insights from the dataset.

Table 6. BERTopic Labeled Model

Topic	*Words*	*Label*
0	conducive, learning, limited, access, materials, facility, equipment, classroom, infrastructure, resources	Academic Difficulties
1	work, time, applied, sustain, needs, jobs, ill, try, look, much, possible, job, tried, night, study	Financial Difficulties
2	lack, facilities, equipment, experience, quality, amenities, available, institution, services, facility	Academic Difficulties
3	financial, crisis, tried, best, save, money, situations, management, difficulty, expenses	Financial Difficulties
4	late, releasing, fund, okay, waiting, updates, delays, payment, issue, subsidy	Grant Disbursement
5	online, classes, access, internet, pandemic, struggle, lack, gadgets, reliable, schooling	Pandemic-Related Challenges
6	study, good, scholar, academic, pressure, completing, education, maintaining, level, performance	Academic Difficulties
7	drop, maintaining, grade, biggest, good, continuously, excelling, academically, consistency, high	Academic Difficulties

Table 7 presents the domain experts labeled LDA model, representing the highest scores in Silhouette and Coherence among the LDA experiments. This model showcases the following hyperparameters: 8 topics, 10 top words, 50 passes, 3000 iterations, alpha value of 0.01, and eta value of 0.01, achieving Silhouette Scores of 0.789 and Coherence Scores of 0.623.

Table 7. LDA Labeled Model

Topic	*Words*	*Label*
0	hard, expectations, scholarship, support, pressure, help, high, study, passing, excel	Academic Difficulties
1	lack, facilities, equipment, information, encounter, poor, resources, school, infrastructures, classrooms	Academic Difficulties
2	school, expenses, financial, problems, pay, need, requirements, family, enough, fee	Financial Difficulties
3	education, quality, access, lack, sustain, limited, free, learning, financial, expenses	Financial Difficulties
4	pressure, passing, academic, expenses, perform, requirements, work, studying, school	Academic Difficulties
5	grades, pandemic, online, good, maintaining, experience, pressure, academic, learning, classes	Pandemic-Related Challenges
6	school, free, money, tuition, allowance, delayed, expensive, release, bills, patience	Grant Disbursement
7	encounter, education, program, free, study, things, tuition, regarding, facilities, process	Program Implementation

The comprehensive analysis of domain experts' top labels derived from BERTopic revealed key themes including "Academic Difficulties," "Financial Difficulties," "Grant Disbursement," "Pandemic-Related Challenges," and "Program Implementation." These themes capture significant aspects of the dataset, highlighting academic hurdles, financial constraints, the allocation of grants, challenges related to the pandemic, and insights into the implementation of programs.

In comparison, the top labels derived from the extensive LDA experiments also depicted recurring themes resembling BERTopic, featuring "Academic Difficulties," "Financial Difficulties," "Pandemic-Related Challenges," "Grant Disbursement," and "Program Implementation." While these shared themes between BERTopic and LDA underscore the significance of core topics such as academic challenges, financial constraints, grant allocation, and program implementation, notable differences were observed.

Specifically, the sequence or emphasis between "Pandemic-Related Challenges" and "Grant Disbursement" differed between BERTopic and LDA. This discrepancy indicates variations in the prioritization or representation of these specific themes within the models' outcomes, highlighting nuanced differences in thematic recognition despite the overall convergence of identified topics across both methodologies. The alignment between the top labels from BERTopic and LDA signifies the robustness and consistency of these identified themes within the dataset. This convergence emphasizes the significance and relevance of these topics, corroborating their substantial presence and recurring nature as determined by domain experts across both topic modeling methodologies.

4. Conclusion and Recommendation

Analyzing BERTopic and LDA methodologies with varied parameters and evaluation metrics revealed critical insights for effective topic modeling. Silhouette and coherence scores were essential in evaluating model performance, showcasing the nuanced impacts of fine-tuning parameters. BERTopic excelled in semantic relevance and coherence, while LDA demonstrated proficiency in defining distinct clusters. Understanding these model disparities underscores the importance of aligning modeling choices with specific research aims and dataset nuances. Researchers can comprehensively understand complex datasets by leveraging BERTopic's semantic richness and LDA's cluster delineation capabilities.

Moreover, comparing domain expert-labeled themes from both models revealed commonalities in key topics such as academic and financial challenges, yet found nuanced variations in their representations. These findings stress the importance of a discerning approach in selecting, integrating, and interpreting models, considering their distinct strengths and the intricate nuances of the dataset. This holistic strategy promises to deliver precise, comprehensive, and meaningful insights applicable across diverse domains within topic modeling research.

Based on the perceived key themes such as "Academic Difficulties," "Financial Difficulties," "Grant Disbursement," "Pandemic-Related Challenges," and "Program Implementation," several recommendations are proposed to improve the UAQTE program further. First, there is a clear need to strengthen academic support initiatives. Implementing tailored programs or mentorship schemes can aid students in coping with academic challenges and subsequently enhance their academic performance. Second, the program should consider enhancing financial assistance packages by revising or expanding the aid to better meet students' specific financial needs, as indicated by the theme of "Financial Difficulties." Third, flexible policies to address "Grant Disbursement" should be considered, including the alignment of grant allocation strategies considering the actual needs of students. Addressing "Pandemic-Related Challenges" remains crucial, given that this study covered students who experienced these difficulties during the pandemic. Implementing support mechanisms, such as remote learning assistance or flexible funding options during health crises, and integrating mental health support and guidance programs is equally important to aid students in coping with the psychological impacts and associated challenges. Lastly, the utilization of the BOSES KO toolkit as a tailored feedback tool within UAQTE, empowering beneficiaries to express their concerns within a structured feedback system, has the potential to provide invaluable insights. This aims to consistently enhance the program's effectiveness and impact by enabling UAQTE beneficiaries to communicate their concerns effectively.

These recommendations offer tangible areas where policy reforms and improvements within the UAQTE program can be implemented to support students better and enhance the educational sector's overall effectiveness. However, it is important to consider that policy reforms are iterative and require continuous monitoring and adjustments. Therefore, while the analysis provides a solid foundation for policy recommendations, evaluations and adaptability to changing circumstances will be essential for sustained improvement in the UAQTE program's impact on students and the broader educational landscape.

References

[1] P. G. Altbach, L. Reisberg, and L. E. Rumbley, "Trends in Global Higher Education: Tracking an Academic Revolution A Report Prepared for the UNESCO 2009 World Conference on Higher Education Published with support from SIDA/SAREC," 2009.

[2] L. Mishra, T. Gupta, and A. Shree, "Online teaching-learning in higher education during lockdown period of COVID-19 pandemic," International Journal of Educational Research Open, vol. 1, Jan. 2020, doi: 10.1016/j.ijedro.2020.100012.

[3] V. Erdoğan, "Integrating 4C Skills of 21st Century into 4 Language Skills in EFL Classes," International Journal of Education and Research, 2019, [Online]. Available: www.ijern.com

[4] A. S. R. Manstead, "The psychology of social class: How socioeconomic status impacts thought, feelings, and behaviour," British Journal of Social Psychology, vol. 57, no. 2, pp. 267–291, Apr. 2018, doi: 10.1111/bjso.12251.

[5] T. Kromydas, "Rethinking higher education and its relationship with social inequalities: Past knowledge, present state, and future potential," Palgrave Commun, vol. 3, no. 1, Dec. 2017, doi: 10.1057/s41599-017.

[6] T. Kestin, van den Belt, L. Denby, K., T. Ross, and M. Hawkes, "Getting started with the SDGs in universities: A guide for universities, higher education institutions, and the academic sector.," Andrew Wilks, 2017.

[7] G. Nhamo and V. Mjimba, "Sustainable Development Goals Series Quality Education Sustainable Development Goals and Institutions of Higher Education," 2020.

[8] V. Vaccari and M. P. Gardinier, "Toward one world or many? A comparative analysis of OECD and UNESCO global education policy documents," International Journal of Development Education and Global Learning, vol. 11, no. 1, Jun. 2019, doi: 10.18546/ijdegl.11.1.05.

[9] R. J. Didham and P. Ofei-Manu, "Adaptive capacity as an educational goal to advance policy for integrating DRR into quality education for sustainable development," International Journal of Disaster Risk Reduction, vol. 47, Aug. 2020, doi: 10.1016/j.ijdrr.2020.101631.

[10] V. Odell, P. Molthan-Hill, S. Martin, and S. Sterling, "Transformative Education to Address All Sustainable Development Goals," 2020, pp. 1–12. doi: 10.1007/978-3-319-69902-8_106-1.

[11] K. Kohl et al., "A whole-institution approach towards sustainability: a crucial aspect of higher education's individual and collective engagement with the SDGs and beyond," International Journal of Sustainability in Higher Education, vol. 23, no. 2. Emerald Group Holdings Ltd., pp. 218–236, February 21, 2022.

[12] D. G. Smith, "Diversity's Promise for Higher Education:

Making It Work," JHU Press, 2020.

[13] T. Shaik et al., "A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis," IEEE Access, vol. 10, pp. 56720–56739, 2022, doi: 10.1109/ACCESS.2022.3177752.

[14] O. Umidjon, "Unlocking the Power of Natural Language Processing (NLP) for Text Analysis," *World scientific research journal* 17, no. 1, 2023.

[15] K. R. Prasad, M. Mohammed, and R. M. Noorullah, "Hybrid topic cluster models for social healthcare data," International Journal of Advanced Computer Science and Applications, vol. 10, no. 11, pp. 490–506, 2019, doi: 10.14569/IJACSA.2019.0101168.

[16] S. Likhitha, B. S., and H. M., "A Detailed Survey on Topic Modeling for Document and Short Text Data," Int J Comput Appl, vol. 178, no. 39, pp. 1–9, Aug. 2019, doi: 10.5120/ijca2019919265.

[17] B. A. H. Murshed, S. Mallappa, J. Abawajy, M. A. N. Saif, H. D. E. Al-ariki, and H. M. Abdulwahab, "Short text topic modeling approaches in the context of big data: taxonomy, survey, and analysis," Artif Intell Rev, vol. 56, no. 6, pp. 5133–5260, Jun. 2023, doi: 10.1007/s10462-022-10254-w.

[18] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," Inf Syst, vol. 94, Dec. 2020, doi: 10.1016/j.is.2020.101582.

[19] W. Luis Roldan-Baluis, N. Alcas Zapata, and M. Soledad Mañaccasa Vásquez, "The Effect of Natural Language Processing on the Analysis of Unstructured Text: A Systematic Review," International Journal of Advanced Computer Science and Applications, 13(5), 2022.

[20] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, and Most. M. J. Mim, "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," IEEE access, 2023.

[21] E. Qais, V. Mn, and P. E. S. Mca, "Short Text Analytics based on BERT by using Multivariate Filter Methods for Feature Selection," 2023, doi: 10.21203/rs.3.rs-3336617/v1.

[22] B. V Pranay Kumar and M. Sadanandam, "A Fusion Architecture of BERT and RoBERTa for Enhanced Performance of Sentiment Analysis of Social Media Platforms," International Journal of Computing and Digital Systems, 15(1), 51-66, 2023.

[23] F. Alhaj, A. Al-Haj, A. Sharieh, and R. Jabri, "Improving Arabic Cognitive Distortion Classification in Twitter using BERTopic," International Journal of Advanced Computer Science and Applications, 13(1), 854-860, 2022.

[24] D. Maier et al., "LDA Topic Modeling in Communication Research Applying LDA topic modeling in communication research: Toward a valid and reliable methodology Communication Methods and Measures: Special Issue on Computational Methods," 2021.

[25] Zoya, S. Latif, F. Shafait, and R. Latif, "Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling," IEEE Access, vol. 9, pp. 127531–127547, 2021, doi: 10.1109/ACCESS.2021.3112620.

[26] D. Choi and B. Song, "Exploring technological trends in logistics: Topic modeling-based patent analysis," Sustainability (Switzerland), vol. 10, no. 8, Aug. 2018, doi: 10.3390/su10082810.

[27] T. Wang and C. Y. Liu, "JSEA: A Program Comprehension Tool Adopting LDA-based Topic Modeling," International Journal of Advanced Computer Science and Applications, 8(3), 2017. [Online]. Available: https://github.com/jseaTool/JSEA

[28] R. Surbakti Saragih, S. Subagio, R. Aditya, and R. Watrianthos, "Jurnal Media Informatika Budidarma BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory," 2023, doi: 10.30865/mib.v7i3.6426.

[29] W. Zha, Q. Ye, J. Li, and K. Ozbay, "A social media Data-Driven analysis for transport policy response to the COVID-19 pandemic outbreak in Wuhan, China," Transp Res Part A Policy Pract, vol. 172, Jun. 2023, doi: 10.1016/j.tra.2023.103669.

[30] J. A. Da Silva Amaral and F. B. De Lima Neto, "A Model for Selecting Relevant Topics in Documents Aimed at Compliance Processes," in 2021 IEEE Latin American Conference on Computational Intelligence, LA-CCI 2021, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/LA-CCI48322.2021.9769786.

[31] C. Cheng and B. Morkos, "Exploring topic modeling for generalising design requirements in complex design," Journal of Engineering Design, 34(11), 922-940, 2023.

[32] T. Saheb and M. Dehghani, "Artificial intelligence for Sustainability in Energy Industry: A Contextual Topic Modeling and Content Analys," Sustainable Computing: Informatics and Systems, 35, 100699, 2022.

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

[34] S. H. Mohammed and S. Al-Augby, "LSA & LDA Topic Modeling Classification: Comparison study on E-books," Indonesian Journal of Electrical Engineering and Computer Science, vol. 19, no. 1, 2020, doi: 10.11591/ijeecs.v19.i1.pp%25p.

[35] X. Chen, D. Zou, G. Cheng, and H. Xie, "Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education," Comput Educ, vol. 151, Jul. 2020, doi: 10.1016/j.compedu.2020.103855.

[36] A. Analytics, D. Fontes, and H. Silvestre Da Silva, "MMAA Mestrado em Métodos Analíticos Avançados MAPINTEL: Enhancing Competitive Intelligence Acquisition Through Embedding and Visual Analytics," 2021.

[37] S. Sarkar, A. Alhamadani, L. Alkulaib, and C. T. Lu, "Predicting Depression and Anxiety on Reddit: a Multi-task Learning Approach," in Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 427–435. doi: 10.1109/ASONAM55673.2022.10068655.

[38] W. Qi, "Beyond Sentiment: Leveraging Topic Metrics for Political Stance Classification," arXiv preprint arXiv:2310.15429, Oct. 2023.

[39] E. Aytaç and M. Khayet, "A Topic Modeling Approach to Discover the Global and Local Subjects in Membrane Distillation Separation Process," Separations, vol. 10, no. 9, p. 482, Sep. 2023, doi: 10.3390/separations10090482.

[40] M. Rujas, B. Merino-Barbancho, P. Arroyo, and G. Fico, "Development of a Natural Language Processing-Based System for Characterizing Eating Disorders," 2023.

[41] M. H. Weng, S. Wu, and M. Dyer, "Identification and Visualization of Key Topics in Scientific Publications with Transformer-Based Language Models and Document Clustering Methods," Applied Sciences (Switzerland), vol. 12, no. 21, Nov. 2022, doi: 10.3390/app122111220.

[42] J. Yang, H. Jang, and K. Yu, "Analyzing Geographic Questions Using Embedding-based Topic Modeling," ISPRS Int J Geoinf, vol. 12, no. 2, Feb. 2023, doi: 10.3390/ijgi12020052.