AI-driven analysis: optimizing tertiary education policy through machine learning insights

Christian Y. Sy [a,1,*], Lany L. Maceda [a,2], Mideth B. Abisado [b,3]

[a,b] Computer Science and Information and Technology Department, Bicol University, Legazpi City, Philippines
[b] College of Computing and Information Technology, National University, Metro Manila, Philippines
[1] cysy@bicol-u.edu.ph; [2] llmaceda@bicol-u.edu.ph; [3] mbabisado@national-u.edu.ph
* corresponding author

ARTICLE INFO

Article history
Received March 3, 2023
Revised March 19, 2024
Accepted March 29, 2024
Available online May 31, 2024

Keywords
Multiclass text classification
Topic modeling
BERT
BERTopic
UAQTE program

ABSTRACT

Tertiary education is universally recognized as crucial for equipping individuals with the essential knowledge and skills, advancing efforts towards equitable access to quality higher education worldwide. The Philippines' Universal Access to Quality Tertiary Education (UAQTE) Act exemplifies this commitment by providing eligible Filipino students with free tertiary education. This study evaluates the UAQTE program's implementation from the perspectives of student beneficiaries, employing a combined approach of qualitative analysis and machine learning techniques. Supervised and unsupervised machine learning methods, including multiclass text classification using BERT and topic modeling with BERTopic, are utilized to analyze student responses, revealing insights into their experiences and perceptions of the program. While BERT proves effective in certain categories, challenges such as overfitting and the delicate balance of sequence length versus model performance are identified. BERTopic highlights the importance of capturing two-word combinations for enhancing topic coherence, with key themes identified including "Educational Opportunity," "Program Implementation," "Financial Support," and "Appreciation and Gratitude," underscoring their significance within the UAQTE program. The convergence of themes between BERT and BERTopic provides a nuanced perspective on students' experiences, emphasizing the program's commitment to addressing financial barriers. Recommendations for program enhancement encompass refining focus areas, strengthening support systems, and fostering continuous monitoring and interdisciplinary collaboration to tackle emerging challenges effectively. Additionally, technical optimization recommendations include refining model configurations, exploring advanced techniques to mitigate overfitting, and conducting further research to enhance transformer-based models' effectiveness in analyzing student experiences. This study underscores the importance of ongoing evaluation and improvement of the UAQTE program to ensure its efficacy in providing quality tertiary education and meeting the diverse needs of Filipino students.

1. Introduction

Tertiary education equips individuals with the competence, ability, and proficiency needed for success across various fields. It provides professional training, empowering individuals to impact society positively. Recognizing its significance, there is a global effort to promote equal access to tertiary education, emphasizing its role in economic development, poverty reduction, and sustainable growth.

https://doi.org/10.26555/ijain.v10i2.1525

Under the 4th Sustainable Development Goal (SDG), which is "Quality Education," the United Nations (UN) stresses the significance of inclusive and high-quality education, particularly in nations with low incomes. This aims to ensure equitable access to high-quality tertiary education, eliminate socio-economic barriers, and position education as a driving force for empowerment, societal progress, and sustainable development.

The current condition of tertiary education in the Philippines reflects a landscape undergoing significant transformations driven by initiatives such as the Universal Access to Quality Tertiary Education (UAQTE) program. Despite progress in enhancing access to higher education, challenges persist in maintaining quality, relevance, and inclusivity across the tertiary education sector. The UAQTE program, implemented to address longstanding economic barriers to tertiary education, has undoubtedly expanded opportunities for Filipino youth by covering tuition fees and related expenses in public State Universities and Colleges (SUCs). Its impact is evidenced by a notable increase in enrollment rates, benefitting approximately 2 million students as of 2022. However, while the UAQTE program has improved access, evaluations and studies reveal ongoing concerns regarding the quality of education provided and the program's long-term sustainability. Questions arise about whether the increased enrollment rates have strained resources and infrastructure, potentially affecting the overall quality of education. Additionally, there are discussions surrounding the alignment of curriculum and programs with industry needs and technological advancements, ensuring graduates are equipped with relevant skills for the workforce. Moreover, socio-economic disparities persist despite the UAQTE program's efforts, with marginalized communities still facing barriers to accessing quality tertiary education.

Given these understandings, it is imperative to do an in-depth assessment of the execution of the UAQTE program to ascertain its efficacy and pinpoint potential areas for improvement. This approach allows for a thorough examination of the program's impact and provides an opportunity to gather insights from those directly affected. Nevertheless, prior research has frequently neglected the qualitative features of the initiative, leading to a limited comprehension of the viewpoints of student recipients. Although qualitative data in its raw form provides useful insights, it may not comprehensively reflect students' varied experiences and views. Therefore, it is evident that a more distinct qualitative investigation is required, with a specific emphasis on the viewpoints of student recipients. The absence of research employing sophisticated machine learning methodologies for analysis constrains the capacity to acquire a more profound understanding of the actualization of the UAQTE program. It is crucial to address this gap to provide valuable information for future policy actions and developments to enhance the accessibility, quality, and sustainability of higher education in the Philippines. This will help bridge the gap between policy aims and stakeholders' experiences. This novel approach contributes to the existing body of knowledge by providing insights that traditional qualitative analysis alone may not reveal. Student responses are categorized into predefined themes through supervised learning, enabling targeted analysis. Conversely, unsupervised learning reveals latent themes within the data, providing a holistic view of student experiences without predefined assumptions. This dual approach identifies expected and unexpected insights, enriching the understanding of the UAQTE program's effectiveness and challenges.

The present study employs the "Boses Ko" or "My Voice" participatory toolkit, facilitating the digital sharing of student beneficiaries' experiences on the program's implementation. The toolkit presented here is the outcome of a joint project between Bicol University (BU) and National University (NU), with financial support from the CHED-LAKAS (Commission on Higher Education - Leading the Advancement of Knowledge in Agriculture and Sciences) program. This program emphasizes research and development initiatives in science and technology. Using the pre-processed dataset, qualitative modeling involves multiclass text classification with BERT and topic modeling with BERTopic. BERT's advanced language processing capabilities capture semantic nuances, improving the accuracy of identifying underlying topics. The models produced undergo assessment through automated metrics and expert manual review to guarantee their reliability. The rigorous evaluation of machine learning approaches undertaken in this study not only contributes to the existing body of knowledge by validating the accuracy and efficiency of these techniques in analyzing qualitative data but also demonstrates their efficacy in uncovering valuable insights from student narratives, opening new avenues for future studies in educational research and policy evaluation. Additionally, this inclusive approach fosters collaboration and participation, enabling stakeholders to shape the program effectively while uncovering alignments in student experiences. Through supervised and unsupervised machine learning, the objective is to provide valuable insights that enhance the UAQTE program's effectiveness and ensure its sustainable success in serving Filipino students, thereby contributing to a more comprehensive understanding of the program's impact and guiding future policy decisions and reforms in tertiary education.

2. Literature Review

This section provides context on the Universal Access to Quality Tertiary Education (UAQTE) program in the Philippines, supervised and unsupervised machine learning techniques, the role of domain experts, and the evaluation metrics. UAQTE aims to enhance accessibility to higher education, while machine learning methods help analyze its impact through student responses. Domain experts and evaluation metrics are crucial in refining and validating generated models. This background sheds light on the implementation and efficacy of the UAQTE program, emphasizing the interdisciplinary approach required for well-informed educational policy intervention.

2.1. UAQTE Program

The UAQTE program, established under Republic Act No. 10931 in the Philippines, marks a significant initiative to address disparities in higher education accessibility. Enacted on August 13, 2017, this program is designed to ensure equitable access to quality tertiary education, foster academic excellence, and provide comprehensive support to eligible Filipino students [1]. By alleviating financial barriers and promoting quality education standards, UAQTE aspires to bolster opportunities for social mobility and contribute to national development. The UAQTE program takes a holistic approach, utilizing financial aid, resource distribution, and outreauch initiatives. Financial assistance is crucial, covering tuition and other costs to alleviate financial pressure on students and families [2], [3]. Concurrently, strategies for resource allocation are directed towards reinforcing the capacity of higher education institutions (HEIs) and technical-vocational institutions (TVIs) to deliver high-quality education and support services [4]. Investments in infrastructure, faculty training, and curriculum improvement synergistically contribute to elevating educational standards and enhancing student achievements [5], [6]. Furthermore, organized initiatives are implemented to educate qualified students about the advantages and qualification requirements of UAQTE, guaranteeing that disadvantaged communities are informed and empowered to pursue higher education opportunities.

The anticipated outcomes of the UAQTE program encompass a spectrum of advancements, including increased enrollment, heightened retention and completion rates, and an overall enhancement in the quality and relevance of tertiary education [7]. By mitigating financial obstacles and providing comprehensive support services, UAQTE is poised to elevate enrollment rates, particularly among disadvantaged groups. Concurrent efforts to bolster retention and completion rates aim to address the underlying factors contributing to student attrition and foster higher graduation rates [8], [9]. Moreover, through strategic investments in institutional capacity and curriculum development, UAQTE seeks to furnish students with the requisite skills and knowledge vital for success in the contemporary workforce. In principle, the UAQTE program signifies a significant stride to increase the availability of tertiary education. By tackling financial impediments, fortifying institutional capacity, and providing holistic support to students, UAQTE endeavors to cultivate equitable access, enhance educational outcomes, and drive national progress. Continuous evaluation and monitoring will be indispensable in gauging the program's impact [10], determining areas for enhancement, and ensuring its efficacy in catering to the needs of Filipino students and society as a whole.

2.2. Supervised and Unsupervised Machine Learning

Assessing the impact of the UAQTE program relies significantly on supervised and unsupervised machine learning techniques. Supervised learning, which depends on labeled data [11]–[13], facilitates systematic analysis of qualitative responses from student beneficiaries, offering comprehensive insights into program effectiveness. Through methods like multiclass text classification, responses are categorized systematically, efficiently extracting meaningful insights [14], [15] into various dimensions of the UAQTE program's impact. The integration of models like BERT (Bidirectional Encoder Representations from Transformers) in multiclass text classification enhances accuracy by capturing nuances and contexts within student responses [16]–[18]. Through supervised learning, researchers uncover nuanced insights [19], [20] into the UAQTE program's impact, highlighting areas for improvement and potential challenges.

On the other hand, unsupervised machine learning, particularly topic modeling, uncovers latent themes and patterns [21]–[23] within student responses without predefined categories. Researchers can use models like BERTopic to identify hidden themes [24], [25], providing a deeper understanding of their viewpoints. This data-driven approach enhances the comprehensiveness of the assessment by uncovering insights that may not be apparent through supervised methods alone [26], [27]. Nonetheless, unsupervised learning may encounter challenges, such as the emergence of ambiguous or irrelevant topics, necessitating manual validation to ensure accuracy and relevance. The combination of supervised and unsupervised machine learning techniques offers a holistic understanding [28], [29] of the UAQTE program's impact. By categorizing student responses and uncovering latent topics, researchers can identify patterns, trends, and areas of concern related to the program's implementation and effectiveness. These insights provide invaluable guidance for policymakers, educators, and stakeholders, informing decision-making processes to enhance the accessibility of tertiary education. By adopting this comprehensive approach, the study offers an in-depth understanding of the UAQTE program's impact, thereby enabling evidence-based strategies for its enhancement and sustainability in supporting Filipino students.

2.3. Role of Domain Experts

In crafting predefined categories for multiclass text classification [30], [31], domain experts play a pivotal role in the development of machine learning models for text analysis. Their active engagement ensures that these categories effectively encompass the diverse dimensions of qualitative responses [26], [32], [33]. By leveraging their specialized knowledge, domain experts align the classification model with the intricacies of the specific research domain, thereby enhancing the accuracy and contextual relevance of class labels [34], [35]. Furthermore, domain experts play a vital part in validating machine learning models by assessing the accuracy of model predictions against the true labels they provide [36]. This ongoing validation process contributes to refining the classification system, bolstering its reliability, and reinforcing the credibility of the analysis [37]. In topic modeling, domain experts undertake the crucial responsibility of identifying and labeling generated topic models with appropriate themes [38]–[40]. They draw upon their expertise to label the generated models, reflecting observed word similarities and accurately capturing each topic's essence [41], [42]. This thorough labeling improves the interpretability of the generated topic models, aiding in informing policy outcomes. Furthermore, domain experts validate the topics to ensure alignment with the research goals [21], [43], [44]. Their active participation maximizes the effectiveness of topic modeling methodologies, enriching the analysis and contributing to an inclusive knowledge of the UAQTE program's impact on tertiary education in the Philippines.

2.4. Model Evaluations Metrics

Evaluation metrics serve as critical benchmarks for evaluating the effectiveness of multiclass text classification models and topic modeling techniques, offering valuable insights into their performance and generalization capabilities [45], [46]. These metrics encompass various factors, including training accuracy, validation accuracy, and test accuracy, which are essential in gauging the accuracy and reliability of multiclass text classification models [47], [48]. Additionally, precision, recall, and F1-score provide nuanced assessments of the model's ability to accurately classify instances across multiple classes while maintaining a balance between false positives and false negatives [49], [50]. The confusion matrix further facilitates a detailed analysis of the model's performance across different classes [51], [52]. Furthermore, domain experts play a key role in the evaluation process by providing true labels and validating the model's predictions against ground truth [53]. Their involvement significantly enhances the reliability and credibility of multiclass text classification, ensuring the precision of classifications and reinforcing the trustworthiness of the analysis [54]. Through collaborative efforts between machine learning algorithms and domain experts, the accuracy of model predictions is verified, contributing to a deeper understanding of the UAQTE program.

In evaluating topic modeling, supplementary metrics such as silhouette scores, coherence scores, and manual examination of generated topics are employed [55], [56]. Silhouette scores quantify the cohesion and separation of clusters formed by the model, providing a quantitative assessment of the quality of topic clusters [57]. Coherence scores evaluate the semantic coherence of topics by assessing the semantic similarity between words within each topic [58]. Domain experts conduct a manual examination to refine the model further and validate its coherence, ensuring it aligns with the research objectives and the intricacies of the UAQTE program. The collaborative efforts between machine learning algorithms and domain experts in evaluating multiclass text classification models and topic models play a crucial role in determining the intricacies of student experiences and perceptions associated with the UAQTE program. Through integrating evaluation metrics and expert validation, multiclass text classification and topic modeling techniques offer a comprehensive approach to uncovering valuable insights from qualitative data.

Analyzing the findings from the various studies in the literature review reveals several recurrent themes, disparities, and gaps in the existing literature regarding evaluating the UAQTE program and utilizing machine learning techniques. Common themes across studies include improving accessibility to higher education, the need for comprehensive evaluation methodologies, and the importance of informed policy decisions to enhance educational outcomes. However, discrepancies arise in the methods employed, with some studies focusing solely on supervised learning techniques while others explore both supervised and unsupervised approaches. Additionally, there is a gap in the literature regarding the long-term impact of the UAQTE program on student outcomes and national development goals, suggesting a need for longitudinal studies to assess program sustainability and effectiveness over time. Moreover, while some studies highlight the role of domain experts in refining machine learning models, others may overlook the significance of expert validation, indicating a potential gap in methodological objectivity.

3. Method

The methodology employed is outlined in this section. Fig. 1 presents the information processing phases covering data preparation, tokenization and embeddings, model training, model evaluation and interpretation, and hyperparameter tuning and inference.

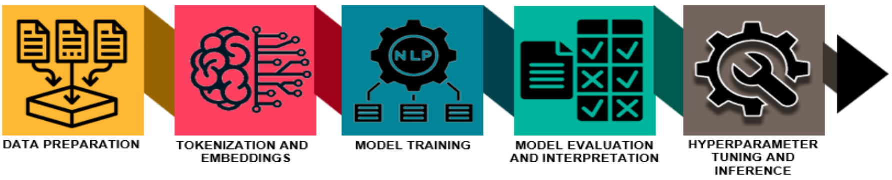
NLP
DATA PREPARATION
TOKENIZATION AND EMBEDDINGS
MODEL TRAINING
MODEL EVALUATION AND INTERPRETATION
HYPERPARAMETER TUNING AND INFERENCE

Fig. 1. Information Processing Phase

3.1. Data Preparation

The utilization of the "Boses Ko" or "My Voice" toolkit for data collection from student beneficiaries of the UAQTE program stems from its grassroots approach, emphasizing the importance of prioritizing perspectives directly from those involved in the program. This approach ensures the authenticity and relevance of the collected data, as it directly captures the experiences and insights of the beneficiaries themselves. The qualitative question guiding the study, "Write your experiences as one of the beneficiaries of the UAQTE program," further underscores the focus on understanding student experiences within the program, aligning closely with the research objectives. It facilitates the direct sharing of experiences by student beneficiaries, providing a platform for them to express their thoughts, concerns, and observations regarding the implementation of the UAQTE program. This participatory approach empowers students to voice their opinions and contribute valuable insights that may not have been captured through traditional survey methods alone. By leveraging digital tools for data collection, the toolkit enables efficient and accessible participation, allowing a diverse range of students to share their perspectives regardless of geographical location or logistical constraints. Moreover, the toolkit's development, a collective effort between BU and NU with support from the CHED-LAKAS program, highlights its credibility and relevance in science and technology research and development. This institutional support further validates the toolkit's effectiveness in capturing meaningful data pertinent to the UAQTE program's implementation and impact.

Convenience sampling, a non-probabilistic method, was utilized to select 3,325 student responses from State Universities and Colleges (SUCs) to evaluate the UAQTE program. Despite its departure from random selection principles, convenience sampling was chosen due to its practicality and efficiency in accessing a large pool of participants within the study's constraints. Given the study's logistical limitations and the urgency to collect data across various SUCs within a limited timeframe, convenience sampling proved to be a feasible approach for efficiently gathering a substantial number of student responses. Although convenience sampling may introduce biases and limit the generalizability of findings, its implementation allowed researchers to swiftly gather diverse perspectives from students enrolled in different SUCs across the Philippines. While not as rigorous as probabilistic sampling methods, the large sample size of 3,325 responses helps mitigate some of the limitations of convenience sampling, providing a more comprehensive dataset for analysis. Despite its shortcomings, convenience sampling enabled researchers to capture a broad spectrum of student experiences and perceptions, contributing valuable insights into its impact on tertiary education in the country.

Data cleaning procedures are essential for ensuring the quality and consistency of the dataset used in analyzing the impact of the UAQTE program. By removing non-English, identical, and empty responses, researchers can eliminate irrelevant data points, thus enhancing the reliability of subsequent analysis. Additionally, text standardization techniques are crucial in refining the text representation within the dataset. The dataset is transformed into lowercase, and any special characters, punctuation marks, and digits are removed. These steps help minimize noise and interference with the modeling process, ensuring that the data accurately reflects the responses of student beneficiaries. Tokenization and stopwords removal are vital pre-processing steps that further refine the dataset for analysis. Implemented through tools such as the Natural Language Toolkit (NLTK) library, tokenization breaks down responses into individual tokens, simplifying the analysis of the dataset. By removing common stopwords, researchers can enhance the interpretability of the generated models that convey core themes, allowing for more meaningful insights to be extracted from the dataset.

Domain experts play a crucial responsibility in evaluating multiclass text classification and topic modeling techniques by providing invaluable insights, validating model predictions, and enhancing the reliability of the analysis. One of their primary roles is to provide true labels for the data, especially in supervised learning scenarios. These true labels serve as benchmarks against which the model's predictions are compared, allowing for the assessment of its accuracy and performance. Through their expertise, domain experts can accurately identify and assign appropriate labels to the data, ensuring that the classification or clustering process aligns with the nuances of the research domain. They contribute to validating model predictions, particularly in cases where human judgment and interpretation are necessary. By reviewing and assessing the model's outputs against their expertise and knowledge, domain experts can verify the accuracy and relevance of the generated classifications or topics. Their validation helps reinforce the credibility of the analysis, assuring that the model's outputs are consistent with domain-specific knowledge and expectations. Table 1 presents the identified categories for multiclass text classification.

An 80-20 train-validation split is applied to partition the data, allocating 80% of the entire dataset for training. Within this training subset, 80% is dedicated to actual model training, with the remaining 20% reserved for validation purposes. This partitioning approach ensures that the model is trained on a considerable percentage of the dataset while validation is utilized to monitor model performance and mitigate overfitting risks. The remaining 20% of the entire dataset is reserved for testing the model's performance, providing an independent assessment of its generalization ability.

Table 1. Domain-Experts Identified Categories for Multiclass Text Classification

Categories	Description
Financial Support	Responses mentioning the financial aid offered, relief from financial pressures, and assistance with tuition fees, allowances, and other expenditures.
Educational Opportunity	Responses expressing students' appreciation for the program, which enables them to pursue their desired courses, continue their education, and access high-quality schooling.
Family Support	Responses express families' gratitude for the support provided by the program and the ease it brings to their lives, as it relieves financial burdens, allowing them to save money and allocate resources to other expenses.
Academic Focus and Personal Development	Responses describe students being more focused on studying, becoming more responsible, and having more chances to invest in school projects due to the financial support received. They also attribute personal growth, increased enthusiasm, and improved class standings to being part of the program.
Program Implementation	Responses cover various viewpoints on the program's execution, showcasing favorable and unfavorable perspectives.

3.2. Tokenization and Embeddings

Tokenization is essential in preparing text data for transformer-based machine learning models such as BERT, particularly in multiclass text classification with the UAQTE student responses dataset. This process entails breaking down the text into smaller units, usually words or subwords, facilitated by specialized tokenizers available in Hugging Face's transformers library. For instance, the BERT tokenizer utilizes a WordPiece tokenizer to transform words into subwords based on a predetermined vocabulary. Once tokenization is complete, the tokenized text data requires formatting into an appropriate input structure for transformer models. This includes adding special tokens like [CLS] (classification token) at the beginning of each sentence and [SEP] (separator token) between sentences. Additionally, sequences are padded to a fixed length, and attention masks are generated to differentiate actual words from padding tokens. These formatting steps ensure uniform input lengths and aid the model in focusing on relevant tokens during the training and inference phases.

Document embedding is critical in topic modeling, transforming textual data into numerical interpretations, and capturing the essence of documents. This study employs the BERTopic and TF-IDF algorithms to analyze the dataset, ensuring the capture of contextual, semantic, and class-specific information aligning with research goals. BERTopic stands out for its use of pre-trained BERT models, starting with tokenization to break down documents into single sub-tokens mapped to word vectors from BERT. By considering surrounding words, it captures contextual and semantic nuances. Techniques like mean or max pooling yield condensed vector representations encoding rich information about word meanings and context.

In contrast, TF-IDF emphasizes term importance based on term frequencies and inverse document frequencies. TF-IDF vectors assign a unique dimension to every term in the document collection, where the values are computed using term frequencies and inverse document frequencies. This approach ranks terms according to their importance within documents and their rarity across the entire corpus, providing clear and understandable rankings of term significance.

$$tf - idf(t) \; = \; tf(t, d) \; x \; idf(t) \tag{1}$$

Equation (1) evaluates the weight 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 distinctiveness in the entire document collection ('idf').

3.3. Model Training

Training models with BERT for multiclass text classification tasks using the UAQTE student responses dataset involves several essential steps. The pre-trained BERT model is initially loaded from the Hugging Face's transformers library, leveraging its extensive contextual understanding of the language for diverse natural language processing tasks, including multiclass text classification. Following model initialization, defining an optimizer and a loss function is crucial for training efficiency. Commonly used optimizers such as Adam or SGD, paired with loss functions like Cross Entropy Loss, contribute to parameter adjustment and loss minimization during training, ensuring convergence toward accurate predictions. Data loaders are then constructed to handle pre-processed text data effectively, converting it into PyTorch or TensorFlow datasets. These loaders manage tasks such as shuffling, batching, and loading data onto the GPU, optimizing resource utilization and streamlining the training process. During the training loop, batches of data from the training set are iterated through. The input data is fed through the BERT model to generate predictions in each iteration. Subsequently, the loss is computed by relating the model's predictions with the true labels, followed by a backward pass to compute gradients and update model parameters using the chosen optimizer. This iterative process continues for multiple epochs, with the model's weights adjusted iteratively to enhance performance.

After obtaining document embeddings, BERTopic employs Uniform Manifold Approximation and Projection (UMAP) to transform the high-dimensional embeddings into a lower-dimensional space. This process preserves the data's structure, facilitating enhanced visualization and clustering effectiveness by simplifying the data representation. Subsequently, BERTopic applies the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm to the lower-dimensional UMAP space to extract meaningful topics from the dataset. HDBSCAN distinguishes dense clusters of data points representing topics or subtopics, showcasing its versatility in handling datasets with varying shapes and sizes. One notable feature of BERTopic is its autonomous topic number detection capability. By investigating data-driven insights in the concentration and distribution of document vectors, BERTopic identifies accepted cluster boundaries without requiring researchers to specify the number of topics beforehand. This feature streamlines the topic modeling process, enhancing efficiency. Fig. 2 illustrates the process from document embeddings using BERT modeling to topic representation.

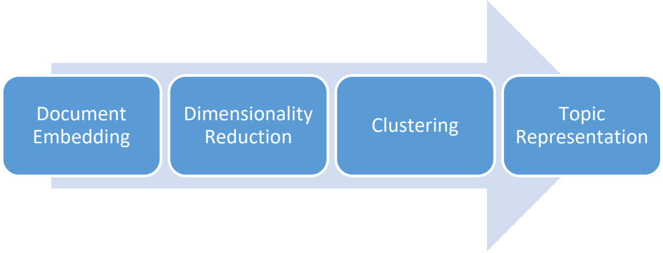

Fig. 2. Topic Extraction using BERTopic

3.4. Model Evaluation and Interpretation

Evaluation metrics are crucial in assessing the effectiveness of multiclass text classification models and topic modeling techniques. Precision, recall, and F1-score are commonly used metrics for evaluating classification models. Equation (2) precision measures the proportion of true positive predictions among all positive predictions made by the model. It indicates the model's ability to avoid false positives, providing insights into its precision in classifying instances correctly.

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives} \tag{2}$$

Equation (3) recall, on the other hand, calculates the proportion of true positive predictions among all actual positive instances. It assesses the model's ability to capture all relevant instances within a class, indicating its sensitivity.

$$Recall = \frac{True\ Positives}{True\ Positives+False\ Negatives} \tag{3}$$

Equation (4) F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's overall performance across all classes. It helps assess the model's accuracy while considering false positives and negatives, making it a robust metric for evaluating classification performance.

$$F1 - score = 2\ x\ \frac{Precision\ x\ Recall}{Precision\ + Recall} \tag{4}$$

The performance of trained multiclass text classification models is comprehensively assessed during the model evaluation phase. This evaluation examines various metrics to gauge the models' effectiveness in handling the UAQTE student responses dataset. Initially, the assessment considers training accuracy, reflecting how well the models have learned from the training data by measuring the proportion of correctly classified instances within this dataset. Subsequently, validation accuracy is assessed to understand the models' generalization performance on unseen data, providing insights into their ability to perform accurately on examples beyond the training set. Finally, test accuracy serves as a vital metric in evaluating the overall performance of the models on entirely novel and unobserved instances, offering a practical indication of their efficacy. The evaluation procedure incorporates precision, recall, and F1-score metrics to comprehensively assess the models' performance concerning classification accuracy by class. The confusion matrix, a tabular representation, provides a detailed breakdown of the model's predictions compared to the actual class labels. It helps visualize the model's performance across different classes, highlighting areas of correct classification and misclassification. The confusion matrix allows for calculating metrics such as accuracy, precision, recall, and F1-score for each class, providing a comprehensive assessment of the model's performance across multiple categories.

In topic modeling, domain experts' silhouette and coherence scores and manual assessment were used to evaluate the topic models. The Silhouette score ranges from -1 to +1 and is calculated based on the average distance of data points within clusters and the distance to the nearest neighboring cluster. A high positive score indicates well-separated clusters, a near-zero score suggests boundary cases, and a negative score indicates potential clustering errors. In equation (5), 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.

$$s = \frac{b-a}{max(a,b)} \tag{5}$$

Coherence scores, on the other hand, range from 0 to 1 and reflect the semantic coherence of topics generated by the model. These scores are computed based on the similarity between words within each topic. Higher coherence scores signify more interpretable topics. In equation (6), 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.

$$C\_V(T)\ =\ 2\ /\ (|T|(|T|\ -\ 1))\ *\ \text{∑\_}\{i = 1\}\text{\^{}}\{|T|\}\ \text{∑\_}\{j\ \neq i\}\ sim(Word\_i, Word\_j) \tag{6}$$

Domain experts are critical in evaluating topic models, providing contextual insights, and ensuring that the topics align with the research domain. They validate the coherence and relevance of the topics generated. Combining quantitative metrics like silhouette and coherence scores with qualitative assessments from domain experts offers a comprehensive evaluation of topic model quality, ensuring statistically sound and semantically meaningful results.

The aim of interpreting the generated topic model was to designate suitable labels based on the observed word similarities. This procedure greatly improved understanding of topics and facilitated efficient communication, allowing practical implementations in the UAQTE program. Moreover, it was pivotal in contributing to a more informed evaluation of the program's effects and guiding future policy decisions concerning tertiary education. An essential aspect of this interpretive process was the active involvement of domain experts. Their specialized knowledge and expertise were instrumental in identifying appropriate labels, refining topics, and ensuring their validity and interpretability. This collaborative effort ensured that the topics aligned closely with the research objectives and provided valuable insights into the UAQTE program. Table 2 represents the finalized labels selected by the domain experts.

Table 2. Domain-Experts Identified Labels for Topic Modeling

Categories	Description
Educational Opportunity	Responses expressing students' appreciation for the program, which enables them to pursue their desired courses, continue their education, and access high-quality schooling.
Program Implementation	Responses cover various viewpoints on the program's execution, showcasing both favorable and unfavorable perspectives.
Financial Support	Responses mentioning the financial aid offered, relief from financial pressures, and assistance with tuition fees, allowances, and other expenditures.
Gratitude and Appreciation	Responses exhibiting a range of expressions of appreciation for being chosen as scholarship program recipients.

3.5. Hyperparameter Tuning and Inference

Fine-tuning hyperparameters is critical in optimizing a multiclass text classification model's efficiency through experiments involving parameters like learning rate, sample size, and number of epochs, which aim to achieve peak performance. Additionally, monitoring the model's performance on a validation set during training facilitates the adjustment of hyperparameters to ensure convergence toward accurate predictions. Here are the hyperparameters utilized:

- Batch Size: Experimenting with batch sizes ranging from 16 to 64 allowed observation of their impact on training dynamics and model convergence. While a larger batch size could accelerate training, it might lead to memory constraints, whereas a smaller batch size could introduce more noise during optimization;
- Epochs: Altering the number of epochs, ranging from 1 to 15, affects the training duration and model evaluation. Increasing epochs enabled the model to extract more insights from the data, yet excessive epochs could result in overfitting on the training set;
- Learning Rates: Adjusting learning rates from 1e-5 to 5e-5 allowed evaluating the model's performance sensitivity. A higher learning rate might expedite convergence but risk overshooting the optimal solution, whereas a lower rate could lead to slower convergence but more stable training;
- Epsilon: This small value added to the denominator of the AdamW optimizer prevents division by zero. While a default value of 1e-8 ensures numerical stability during training, exploring adjustments to 8 allows observation of any changes in training dynamics or model performance;
- Max-length: Varying the max-length between 128 and 256 enabled investigations into the effect of considering different amounts of context on model performance. While a larger max-length allowed the model to capture more contextual information, it might have required more computational resources and memory;

Hyperparameter tuning encompasses systematically adjusting these parameters and assessing their influence on the model's performance metrics, including accuracy, precision, recall, and F1-score. This iterative process aims to identify the optimal hyperparameter configuration, thereby improving the model's ability to generalize and perform well on unseen data.

For topic modeling, the following are the hyper-parameters used:

- min_topic_size: Adjusting this parameter from 10 to 30 allowed for exploring the impact of document count on topic validity. As the minimum document count increased, topics may have become more refined and granular, while lowering it could potentially result in broader topics encompassing fewer documents;

- top_n_words. Setting this parameter from 5 to 10 determined the number of top words showcased per topic, typically representing the most significant terms. Selecting a particular value aided in comprehending the key terms associated with each topic, facilitating a more insightful interpretation of the topics;
- num_topics. This parameter's selection depended on the dataset's nature and the anticipated number of topics. Adjusting num_topics allowed for tailoring the topic modeling process to suit the specific characteristics and complexities of the dataset;
- ngram_range. This parameter determined the range of n-grams used for document term frequency representation, commonly expressed as (min_n, max_n). For example, when ngram_range is set to (1, 2), individual words and two-word phrases are considered, influencing the granularity of topics. The specified ngram ranges included (1,1), (1,2), and (2,2), allowing for variations in the level of detail captured in the topic modeling process.

It is essential to adjust these parameters appropriately to ensure that the topic modeling process matches the distinctive attributes of the dataset, thereby impacting the quality and comprehensibility of the extracted topics. Ultimately, the interpretability and relevance of the topics to the particular research objectives determine their quality.

During the final inference phase in multiclass text classification, domain experts play a central role in validating predictions made by trained models and providing the true labels for the dataset. This validation process is essential for ensuring the accuracy and reliability of the model's predictions, as it confirms their alignment with the ground truth provided by experts. Valuable insights and feedback are exchanged through a feedback loop between domain experts and machine learning models based on domain knowledge and expertise. This collaborative effort guides the iterative optimization of the models' performance, enhancing their predictive capabilities. Furthermore, the inference phase evaluates the models' performance on new data, validating their generalization capabilities. This ensures that the insights derived from the predictions are accurate and trustworthy, contributing to informed decision-making in education policy analysis. Overall, the involvement of domain experts in validating predictions and providing true labels, along with fine-tuned models, advances natural language processing techniques and facilitates informed decision-making in education policy analysis.

4. Results and Discussion

The multiclass text classification task utilizing BERT architecture offered valuable insights into its performance across a range of hyperparameter configurations, as seen in Table 3. Accuracy scores, batch sizes, epochs, learning rates, and max-length parameters are all critical factors influencing the performance of machine learning models. Analyzing the results of the experiments reveals several trends and insights regarding the impact of different hyperparameter configurations on model performance.

Generally, increasing the batch size from 16 to 32 did not consistently improve model performance, as seen in the varied training, validation, and test accuracy scores across different configurations. However, models trained with a larger batch size tended to converge faster, as evidenced by the lower number of epochs required to reach similar accuracy levels compared to smaller batch sizes. This suggests that larger batch sizes may accelerate training but do not necessarily lead to better performance. Across various batch sizes and epochs, models trained with a learning rate of 1e-5 consistently demonstrated higher accuracy than those trained with higher learning rates (3e-5 and 5e-5). This indicates that lower learning rates facilitated more stable and effective learning, allowing the model to converge to better solutions. The performance difference between learning rates of 1e-5 and 3e-5 was not substantial, suggesting that the choice of learning rate within this range may not significantly impact model performance. Increasing the maximum sequence length from 128 to 256 tokens improved accuracy, particularly in the training phase. However, this improvement was not consistently reflected in validation and test accuracy scores. Models trained with a max-length of 256 tokens tended to exhibit slightly higher training accuracy but comparable validation and test accuracy than those trained with a max-length of 128 tokens. This suggests that while longer sequences facilitate better learning during training, they may not necessarily lead to significant improvements in generalization performance. Notably, larger maximum sequence lengths did not consistently enhance accuracy, revealing complexities in balancing sequence length and model performance. This suggests that accommodating longer sequences may capture more context but also introduces challenges such as increased computational overhead and potential information reduction.

Table 3. BERT Hyperparameters and Accuracy Scores

Batch Size	Epoch	Learning rate	Max-length	Training Accuracy	Validation Accuracy	Test Accuracy
16	3	1e-5	128	73%	72%	66%
16	3	1e-5	256	77%	73%	68%
32	3	1e-5	128	65%	65%	65%
32	3	1e-5	256	70%	70%	67%
16	5	1e-5	128	84%	73%	71%
16	5	1e-5	256	85%	75%	70%
32	5	1e-5	128	76%	72%	68%
32	5	1e-5	256	79%	72%	69%
16	3	3e-5	128	85%	76%	73%
16	3	3e-5	256	87%	75%	71%
32	3	3e-5	128	82%	75%	71%
32	3	3e-5	256	81%	75%	70%
16	5	3e-5	128	95%	75%	73%
16	5	3e-5	256	96%	75%	73%
32	5	3e-5	128	93%	74%	72%
32	5	3e-5	256	92%	74%	72%
16	3	5e-5	128	89%	76%	73%
16	3	5e-5	256	88%	76%	72%
32	3	5e-5	128	85%	76%	72%
32	3	5e-5	256	85%	76%	73%
16	5	5e-5	128	93%	74%	72%
16	5	5e-5	256	97%	76%	73%
32	5	5e-5	128	96%	75%	73%
32	5	5e-5	256	96%	74%	72%

The observed trends in accuracy metrics and performance across training, validation, and test datasets offer valuable results regarding the model's functionality and alignment with the study's objectives. These accuracy trends provide a nuanced understanding of the model's capacity to fulfill its intended purpose. Notably, higher accuracy scores, especially in the validation and test datasets, signify the model's adeptness in generalizing to unseen data, suggesting its potential applicability in real-world scenarios. Furthermore, the consistency between the observed accuracy trends and the anticipated outcomes underscores the model's reliability and stability. For instance, configurations with smaller batch sizes and lower learning rates consistently yielded superior accuracy scores, supporting that incremental updates during training and stable learning contribute to enhanced model performance. Similarly, the noted enhancements in accuracy with longer maximum sequence lengths validate the hypothesis that capturing more context could bolster classification accuracy. Additionally, scrutinizing the trends across training, validation, and test accuracies provides deeper insights into the model's behavior throughout different training and evaluation stages. The unity in performance across these metrics indicates the model's ability to effectively learn from the training data and generalize well to unseen instances. This consistency across phases further fortifies confidence in the model's reliability and robustness for real-world applications.

During the experimentation process, several challenges were encountered that impacted the performance and generalization capabilities of the model. One significant challenge was the complexity of balancing sequence length and model performance. Increasing the maximum sequence length could capture more context and introduce challenges such as increased computational overhead and potential information reduction. This complexity required careful consideration to find the optimal sequence length that maximized performance without compromising efficiency. Another challenge observed was instances of overfitting, particularly starting from the fifth epoch of training. Overfitting occurs when the model learns the training data too well, resulting in poor performance on unseen data. To mitigate overfitting, early stopping or regularization techniques were employed. Early stopping involved halting the training process when the model's performance on the validation set stopped improving, preventing it from memorizing the training data and enhancing its generalization capabilities. The choice of hyperparameters, such as batch sizes, epochs, and learning rates, presented challenges in finding the right balance between convergence speed and model performance. Larger batch sizes tended to converge faster but did not consistently lead to better performance. Lower learning rates facilitated more stable learning but required more epochs to converge. Finding the optimal combination of these hyperparameters required extensive experimentation and fine-tuning.

A systematic approach was adopted to address these challenges, including thorough experimentation, careful monitoring of performance metrics, and adjusting hyperparameters based on observed trends. Regularization techniques like dropout were also employed to prevent overfitting by randomly dropping units during training. Mitigating these challenges had implications for the model's generalization capabilities. By finding the optimal sequence length, avoiding overfitting through early stopping and regularization, and fine-tuning hyperparameters, the model's ability to generalize to unseen data was improved. This ensured that the model could effectively capture the underlying patterns in the data without memorizing noise, enhancing its reliability and applicability in real-world scenarios.

As detailed in Table 4, the performance metrics of BERT exhibited strong precision, recall, and F1-score for "Family Support," indicating its robust performance in this category. "Financial Support" also showed promising results, although there is potential for precision enhancement. Similarly, "Academic Focus & Personal Development" displayed balanced precision-recall metrics, suggesting reliable classification performance. However, categories such as "Educational Opportunity" and "Program Implementation" exhibited lower scores, particularly in the recall, highlighting areas where further model refinement may be beneficial.

Table 4. BERT Performance Metrics by Category

Categories (BERT)	Precision	Recall	F1-Score
Academic Focus & Personal Development	70%	78%	74%
Educational Opportunity	54%	60%	57%
Family Support	95%	96%	96%
Financial Support	73%	77%	75%
Program Implementation	80%	55%	65%
Weighted Average	**74%**	**73%**	**73%**

Overall, while BERT model demonstrated effectiveness in certain categories, there remains room for improvement, particularly in those with lower recall scores. Analyzing misclassified instances and fine-tuning model parameters could enhance performance across all categories. Moreover, addressing challenges related to the classification of similar terms into multiple categories is crucial to improving overall accuracy and reducing confusion. By refining the model's ability to differentiate between subtle distinctions in meaning, it can better capture the nuances within the text data and yield more accurate classifications across diverse categories. The heatmap visualization of the confusion matrix, as illustrated in Fig. 3, offered a nuanced representation of BERT's classification performance, complementing the precision, recall, and F1-score metrics. For instance, in the "Academic Focus & Personal Development" (AF&PD) category, BERT accurately classified 103 instances (true positives) but misclassified 17 instances (false negatives), aligning with its recall of 78%. Similar observations were made across other categories such as "Educational Opportunity" (EO), "Family Support" (FaS), "Financial Support" (FiS), and "Program Implementation" (PI).

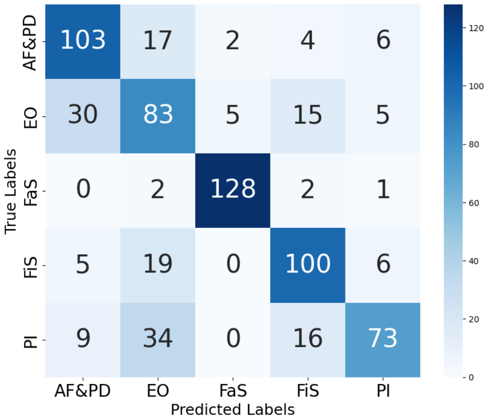

Fig. 3. BERT Confusion Matrix Heatmap

Notably, categories with lower recall scores, such as PI, displayed more false negatives, suggesting areas with potential for improvement. Conversely, categories with high precision and recall, like FaS, demonstrated fewer misclassifications. This detailed analysis, coupled with precision, recall, and F1-score metrics, offered a comprehensive understanding of BERT's classification performance across diverse categories, thus guiding optimization strategies to enhance accuracy. For topic modeling using the BERTopic approach, the results in Table 5 showcased configurations that achieved acceptable silhouette and coherence scores from numerous experiments with varying hyperparameters. These findings unveiled several intriguing insights. Notably, the selection of the n-gram range significantly impacted the quality of generated topics. Models employing a range of (1, 2), which incorporated bigrams, consistently outperformed other configurations in terms of coherence scores. This suggested capturing two-word combinations in topics enhanced their coherence and semantic meaning.

Table 5. BERTopic Hyperparameters and Evaluation Scores

Exp	# of Topics	Top n Words	Min Topic Size	n-gram range	Silhouette Scores	Coherence Scores
1	9	10	13	(1,1)	0.765	0.849
2	8	10	13	(1,2)	0.814	0.863
3	6	10	13	(2,2)	0.802	0.859
4	6	10	15	(1,1)	0.781	0.853
5	8	10	15	(1,2)	0.768	0.850
6	6	10	15	(2,2)	0.786	0.845
7	6	10	20	(1,1)	0.768	0.862
8	7	10	20	(1,2)	0.783	0.840
9	6	10	20	(2,2)	0.752	0.834
10	4	10	25	(1,1)	0.795	0.822
11	3	10	25	(1,2)	0.805	0.832
12	4	10	25	(2,2)	0.808	0.830

Nonetheless, the optimal number of topics ranged from 3 to 9, indicating that the most suitable number of topics depended on the specific characteristics of the dataset. Moreover, the experiment noted that a smaller minimum topic size of 13 yielded better results in both silhouette and coherence scores. This suggested that involving fewer documents per topic resulted in distinct clusters. Overall, these findings underscored the position of meticulously choosing hyperparameters and adjusting them based on dataset features to attain optimal outcomes. Implementing both the silhouette and coherence methods provided a robust framework for evaluating topic quality and guiding the selection of the most fitting model configuration.

Several key themes were discovered, "Educational Opportunity" was the most frequently assigned label, indicating its pervasive and central nature within the dataset. "Program Implementation" emerged as another prominent theme that was consistently recognized, highlighting responses related to the execution of the UAQTE program. "Financial Support" also appeared as a recurring and significant theme, likely associated with financial assistance or funding. Lastly, "Appreciation and Gratitude" implied expressions of thanks and acknowledgment. Table 6 showcases the labeled model, incorporating meticulously selected hyperparameters, silhouette, and coherence scores meeting stringent criteria. It delineates the varied thematic landscape, including words such as "education," "grateful," and "opportunity" under Educational Opportunity, and descriptors like "helpful," "good," and "experience" associated with Program Implementation. Furthermore, financial support encompasses terms like "cost," "expenses," and "financial," while expressions such as "continue," "cost," and "money" align with their context.

Table 6. BERTopic Domain Experts' Labeled Model

Topic	Words	Label
0	college, education, free, funded, grateful, help, opportunity, pursue, state, study	Educational Opportunity
1	beneficial, best experience, convenient, experience, good, helpful, improve, lacking, need improved, quality system, quite	Program Implementation
2	amazing, awesome, excellent, fun, good, great, great help, interesting, job, nice	Program Implementation
3	advantage, college, diminished, education, expenses, financial, free, helped, parents, student	Financial Support
4	accessible, beneficial, beneficial, quite lacking, utilized, outstanding, helpful, good overall, helpful, quite lacking	Educational Opportunity
5	best, complaints, enough, enough though, experience, good, good experience, hope improve, questions, really	Program Implementation
6	comment, education, extremely, free, grateful, helpful, nice, opportunity, really helpful, super, thank	Educational Opportunity
7	continue, cost, education, family, free, guarantee, money, program, quality, receive	Financial Support

Fig. 4 presents the intertopic distance map, which highlights clusters sharing common labels with those designated by domain experts. This placement between clusters and domain expert assigned labels strongly indicates the quality and relevance of these topics. Notably, topics 0, 4, and 6 are classified under "Educational Opportunity," while topics 1, 2, and 5 are grouped as "Program Implementation," and topics 3 and 7 form a cluster labeled "Financial Support." This correlation between metric scores and expert evaluations decisively validates the effectiveness of quantitative analysis in identifying central themes within the dataset.

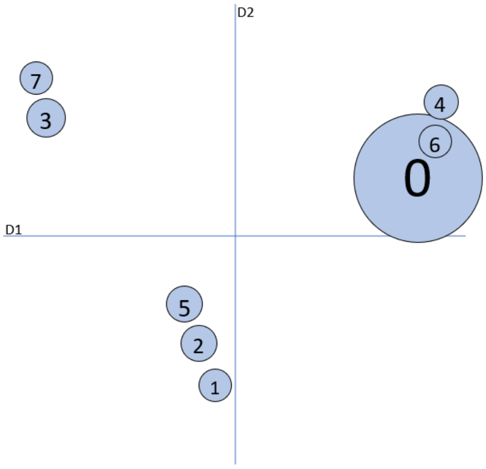

Fig. 4. BERTopic Intertopic Distance Map

The findings derived from the BERT model's predictions and BERTopic have significant implications for real-world applications, particularly in education policy analysis. By accurately classifying and analyzing student responses to the UAQTE program, these models provide policymakers with invaluable comprehension of the program's effectiveness and areas for improvement. Firstly, the BERT model's robust performance in categorizing responses related to family support, financial support, academic focus, and personal development offers policymakers a detailed understanding of how the UAQTE program impacts students in these key areas. This allows policymakers to identify specific aspects of the program that are working well and areas that may require attention or enhancement. For example, suppose the model indicates high student satisfaction regarding family support but identifies gaps in financial assistance. In that case, policymakers can allocate resources more effectively to address these issues and ensure that the program meets the diverse needs of students. Secondly, BERTopic's ability to identify key themes and topics within the dataset further enhances policymakers' understanding of student experiences and perceptions of the UAQTE program. By uncovering recurring themes such as educational opportunities, program implementation, financial support, and appreciation and gratitude, BERTopic provides policymakers with qualitative insights that complement the quantitative analysis provided by the BERT model. These can help policymakers identify predominant trends, challenges, and areas of success within the program, guiding the development of targeted interventions and policy adjustments. Combining the BERT model's predictions and BERTopic's thematic analysis equips policymakers with a comprehensive understanding of the UAQTE program's impact on students and the factors influencing its effectiveness. With these, policymakers can make informed decisions, allocate resources strategically, and design evidence-based interventions to enhance the program's effectiveness and ultimately improve the quality of tertiary education in the Philippines.

The analysis of multiclass text classification using BERT and topic modeling with BERTopic provided a comprehensive understanding of the capabilities and challenges associated with transformer-based models in natural language processing tasks. In multiclass text classification, BERT consistently demonstrated improved accuracy with smaller batch sizes, higher numbers of epochs, and optimal learning rates, highlighting its efficacy in accurately classifying student responses. However, challenges like overfitting and the complexity of balancing sequence length versus model performance were noted, emphasizing the need for careful model tuning. Examining the classification outcomes presented notable insights into students' experiences within the UAQTE program. Categories "Family Support" and "Financial Support" exhibited high precision, recall, and F1 scores, indicating the program's efficacy in addressing student needs in these domains. Conversely, categories like "Educational Opportunity" and "Program Implementation" displayed lower performance metrics, signaling potential areas for enhancement. This underscores the importance of further refining the program to better cater to the identified needs and optimize outcomes.

In topic modeling implementing BERTopic, models integrating bigrams consistently outperformed others, highlighting the significance of capturing two-word combinations for enhancing topic coherence and understanding. By combining silhouette and coherence scores, researchers can effectively assess the topic quality and determine the most suitable model configuration. Moreover, the alignment between hyperparameter configurations and key themes identified by domain experts emphasizes the importance of integrating technical insights with domain-specific knowledge, thereby enhancing the overall effectiveness and relevance of the topic modeling process. The identification of key themes within the dataset, including "Educational Opportunity," "Program Implementation," "Financial Support," and "Appreciation and Gratitude," underscores their pervasive presence and significance in the UAQTE program.

The convergence of themes between BERT and BERTopic provided a nuanced perspective on students' experiences in the UAQTE program. Both models consistently identified themes like "Financial Support," underscoring the program's commitment to addressing financial barriers. This alignment reinforces the significance of financial assistance in facilitating students' educational endeavors. Similarly, if "Educational Opportunity" emerges as a common theme across both analyses, it emphasizes the importance of providing equitable educational experiences. This overlap signals the need for interventions to broaden access to quality education within the program, highlighting areas where improvements can be made to enhance academic outcomes. By leveraging insights from multiclass text classification and topic modeling, program administrators can comprehensively understand the multifaceted aspects of students' experiences, enabling informed decision-making to optimize program effectiveness and address underlying issues effectively. This integrated approach empowers administrators to tailor interventions and resources to support students better, ultimately improving the overall impact and success of the UAQTE program.

Additionally, assessing the long-term effects of the UAQTE program on student outcomes beyond the duration of their participation could provide valuable insights into the program's sustained impact. Longitudinal studies tracking participants over an extended period could help identify any lasting benefits or challenges associated with program participation. Considering the intersectionality of student experiences within the UAQTE program, future research could explore how socio-economic status, race, ethnicity, and gender influence program outcomes. This intersectional approach could help ensure that the program effectively addresses diverse student populations' unique needs and challenges. Lastly, exploring innovative uses of natural language processing techniques beyond multiclass text classification and topic modeling could offer new perspectives on understanding student experiences and informing program improvements. For example, sentiment analysis could be employed to analyze the emotional tone of student responses, providing insights into their overall satisfaction and well-being within the program.

5. Conclusion

Based on the results of the analysis, several actionable recommendations can be proposed to optimize the impact of the UAQTE program on student experiences. Firstly, it is essential to address categories with lower performance metrics, particularly focusing on "Educational Opportunity" and "Program Implementation." This could involve revising program structures, enhancing resources, or providing additional support to ensure students' needs are adequately met in these domains. Secondly, strengthening support systems related to specific themes like "Educational Opportunity" and "Program Implementation" is crucial. This may entail providing students with additional mentorship, guidance, or resources to ensure equitable access to educational opportunities and effective program implementations. Leveraging advanced analytics techniques such as BERTopic can offer deeper insights into student experiences. Specifically, utilizing bigrams and other linguistic features can provide a more nuanced understanding, enabling the identification of hidden patterns within student responses for targeted interventions and improvements. Continuous monitoring and evaluation are critical, necessitating the implementation of a system for regular assessment of the program's effectiveness. This involves collecting participant feedback, assessing performance metrics, and making iterative adjustments to address emerging challenges and optimize outcomes effectively. Lastly, fostering interdisciplinary collaboration between technical experts, educators, policymakers, and other stakeholders involved in the UAQTE program is essential. This collaborative approach enables a holistic optimization strategy, leveraging diverse expertise and perspectives to effectively tailor interventions and support systems. By prioritizing these recommendations, the UAQTE program can enhance its impact on student experiences, ensuring that students receive the support and resources necessary for academic and personal success. Additionally, technical optimization recommendations include refining model configurations, exploring advanced techniques to mitigate overfitting, and conducting further research to enhance transformer-based models' effectiveness in analyzing student experiences. Several potential avenues can be explored for future work to enhance program effectiveness in addressing student needs within the UAQTE program. Firstly, investigating the root causes of lower performance metrics in categories such as "Educational Opportunity" and "Program Implementation" could provide valuable insights. This research could involve qualitative methods such as interviews or focus groups with program participants to uncover specific challenges or gaps in service delivery. Furthermore, exploring the effectiveness of tailored interventions targeting these identified areas for improvement could be beneficial. This could involve implementing targeted support systems or interventions to enhance educational opportunities and improve program implementation processes. The impact of these interventions could be evaluated using a rigorous experimental design, measuring outcomes such as academic achievement, program satisfaction, and overall well-being among participants.

Acknowledgment

The authors extend our heartfelt thanks to the Philippine Commission on Higher Education (CHED) and the Leading the Advancement of Knowledge in Agriculture and Science (LAKAS) Project No. 2021-007, particularly the eParticipation 2.1 initiative titled "Harnessing Natural Language Processing (NLP) for Community Participation," for their generous funding and support, which made this research possible.

Declarations

Author contribution. Dr. Maceda and Dr. Abisado headed the development of the "Boses Ko" toolkit, a pivotal tool in aggregating student responses for our qualitative research analysis. Their leadership extended beyond the toolkit's creation, guiding diverse research domains such as methodology, model validation, and the refinement of findings. Concurrently, Christian conducted meticulous experiments and wrote the research paper. All authors had approved the final version.

Funding statement. This research received funding from the Philippine Commission on Higher Education (CHED) and the Leading the Advancement of Knowledge in Agriculture and Science (LAKAS) Project No. 2021-007, specifically, the eParticipation 2.1 initiative, titled "Harnessing Natural Language Processing (NLP) for Community Participation.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] M. K. P. Ortiz, K. A. M. Melad, N. V. V. Araos, A. C. Orbeta, and C. M. Reyes, “Process Evaluation of the Universal Access to Quality Tertiary Education Act (RA 10931): Status and Prospects for Improved Implementation,” Quezon City, 2019–36, 2019. [Online]. Available at: https://www.pids.gov.ph.

[2] K. Saguin, “The Politics of De-Privatisation: Philippine Higher Education in Transition,” *J. Contemp. Asia*, vol. 53, no. 3, pp. 471–493, May 2023, doi: 10.1080/00472336.2022.2035424.

[3] J. N. Emmanuel, “Affordability In College Access: Improving Equitable Value for Low-Income, First-Generation, and Students of Color,” *Vermont Connect.*, vol. 44, no. April, pp. 1–16, 2023, [Online]. Available at: https://scholarworks.uvm.edu/tvc.

[4] D. Edralin and R. Pastrana, “Technical and vocational education and training in the Philippines: In retrospect and its future directions,” *Bedan Res. J.*, vol. 8, no. 1, pp. 138–172, Apr. 2023, doi: 10.58870/berj.v8i1.50.

[5] A. K. Maiya and P. S. Aithal, “A Review based Research Topic Identification on How to Improve the Quality Services of Higher Education Institutions in Academic, Administrative, and Research Areas,” *Int. J. Manag. Technol. Soc. Sci.*, vol. 8, no. 3, pp. 103–153, Aug. 2023, doi: 10.47992/IJMTS.2581.6012.0292.

[6] M. Kayyali, “The Relationship between Rankings and Academic Quality Manager of Higher Education Quality and Assessment Council HEQAC,” *Sci. Innov. Technol. IJMSIT Rev. Pap.*, vol. 4, no. 3, pp. 1–11, 2023, [Online]. Available at: https://ijmsit.com/volume-4-issue-3/.

[7] T. J. Phillips and A. Dissertation, “Culturally Responsive College Student Retention Theory & Practice,” The Purdue University Graduate School Statement Of Committee Approval, pp. 1-116, 2023. [Online]. Available at: https://hammer.purdue.edu/articles/thesis/Culturally_Responsive.

[8] P. P. Xie, Z. Li, C. Ma, and J. Zhao, “Education Management Intervention in Managing School Problems,” *J. World Englishes Educ. Pract.*, vol. 6, no. 1, pp. 35–87, Jan. 2024, doi: 10.32996/jweep.2024.6.1.3.

[9] F. A. Orji and J. Vassileva, “Modeling the Impact of Motivation Factors on Students’ Study Strategies and Performance Using Machine Learning,” *J. Educ. Technol. Syst.*, vol. 52, no. 2, pp. 274–296, Dec. 2023, doi: 10.1177/00472395231191139.

[10] S. Romlah, A. Imron, Maisyaroh, A. Sunandar, and Z. A. Dami, "A free education policy in Indonesia for equitable access and improvement of the quality of learning," *Cogent Educ.*, vol. 10, no. 2, pp. 1–27, Dec. 2023, doi: 10.1080/2331186X.2023.2245734.

[11] N. A. A. Khleel and K. Nehéz, "Detection of code smells using machine learning techniques combined with data-balancing methods," *Int. J. Adv. Intell. Informatics*, vol. 9, no. 3, p. 402, Nov. 2023, doi: 10.26555/ijain.v9i3.981.

[12] J. Xiao, L. Wang, J. Zhao, and A. Fu, "Research on Adaptive Learning Prediction Based on XAPI," *Int. J. Inf. Educ. Technol.*, vol. 10, no. 9, pp. 679–684, Sep. 2020, doi: 10.18178/ijiet.2020.10.9.1443.

[13] E. Barbierato and A. Gatti, "The Challenges of Machine Learning: A Critical Review," *Electronics*, vol. 13, no. 2, p. 416, Jan. 2024, doi: 10.3390/electronics13020416.

[14] V. K. Rao and M. A. Sowjanya, "Integrated Intelligent Framework for Sensor Data Analysis," *World Acad. J. Eng. Sci.*, vol. 7, no. 3, pp. 52–59, 2020, [Online]. Available at: https://www.isroset.org/pdf_paper_view.php?paper_id=2083&.

[15] A. A. Kurniawan, S. Madenda, S. Wirawan, and R. J. Suhatril, "Multidisciplinary classification for Indonesian scientific articles abstract using pre-trained BERT model," *Int. J. Adv. Intell. Informatics*, vol. 9, no. 2, p. 331, Jul. 2023, doi: 10.26555/ijain.v9i2.1051.

[16] A. H. Oliaee, S. Das, J. Liu, and M. A. Rahman, "Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types," *Nat. Lang. Process. J.*, vol. 3, p. 100007, Jun. 2023, doi: 10.1016/j.nlp.2023.100007.

[17] M. Tezgider, B. Yildiz, and G. Aydin, "Text classification using improved bidirectional transformer," *Concurr. Comput. Pract. Exp.*, vol. 34, no. 9, p. e6486, Apr. 2022, doi: 10.1002/cpe.6486.

[18] R. Ghnemat, A. Shaout, and A. M. Al-Sowi, "Higher Education Transformation for Artificial Intelligence Revolution: Transformation Framework," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 19, pp. 224–241, Oct. 2022, doi: 10.3991/ijet.v17i19.33309.

[19] M. Mukasheva, A. Mukhiyadin, U. Makhazhanova, and S. Serikbayeva, "The Behaviour of the Ensemble Learning Model in Analysing Educational Data on COVID-19," *Int. J. Inf. Educ. Technol.*, vol. 13, no. 12, pp. 1868–1878, Dec. 2023, doi: 10.18178/ijiet.2023.13.12.2000.

[20] A. Ülkü, "Artificial intelligence-based large language models and integrity of exams and assignments in higher education: the case of tourism courses," *Tour. Manag. Stud.*, vol. 19, no. 4, pp. 21–34, Oct. 2023, doi: 10.18089/tms.2023.190402.

[21] A. A. Saeed and N. G. M. Jameel, "Intelligent feature selection using particle swarm optimization algorithm with a decision tree for DDoS attack detection," *Int. J. Adv. Intell. Informatics*, vol. 7, no. 1, p. 37, Mar. 2021, doi: 10.26555/ijain.v7i1.553.

[22] R. S. Concepcion II *et al.*, "Lettuce growth stage identification based on phytomorphological variations using coupled color superpixels and multifold watershed transformation," *Int. J. Adv. Intell. Informatics*, vol. 6, no. 3, p. 261, Nov. 2020, doi: 10.26555/ijain.v6i3.435.

[23] L. Jiang, T. Zhang, and T. Huang, "Empirical Research of Hot Topic Recognition and its Evolution Path Method for Scientific and Technological Literature," *J. Adv. Comput. Intell. Intell. Informatics*, vol. 26, no. 3, pp. 299–308, May 2022, doi: 10.20965/jaciii.2022.p0299.

[24] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Inf. Syst.*, vol. 112, p. 102131, Feb. 2023, doi: 10.1016/j.is.2022.102131.

[25] F. Alhaj, A. Al-Haj, A. Sharieh, and R. Jabri, "Improving Arabic Cognitive Distortion Classification in Twitter using BERTopic," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 1, pp. 854–860, 2022, doi: 10.14569/IJACSA.2022.0130199.

[26] B. Gencoglu, M. Helms-Lorenz, R. Maulana, E. P. W. A. Jansen, and O. Gencoglu, "Machine and expert judgments of student perceptions of teaching behavior in secondary education: Added value of topic modeling with big data," *Comput. Educ.*, vol. 193, p. 104682, Feb. 2023, doi: 10.1016/j.compedu.2022.104682.

[27] 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 &amp; Education,” *Comput. Educ.*, vol. 151, p. 103855, Jul. 2020, doi: 10.1016/j.compedu.2020.103855.

[28] M. H. Mobarak *et al.*, “Scope of machine learning in materials research—A review,” *Appl. Surf. Sci. Adv.*, vol. 18, p. 100523, Dec. 2023, doi: 10.1016/j.apsadv.2023.100523.

[29] M. M. Cencer, J. S. Moore, and R. S. Assary, “Machine learning for polymeric materials: an introduction,” *Polym. Int.*, vol. 71, no. 5, pp. 537–542, May 2022, doi: 10.1002/pi.6345.

[30] L. Zhou, F. Zhang, S. Zhang, and M. Xu, “Study on the Personalized Learning Model of Learner-Learning Resource Matching,” *Int. J. Inf. Educ. Technol.*, vol. 11, no. 3, pp. 143–147, Jan. 2021, doi: 10.18178/ijiet.2021.11.3.1503.

[31] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, “A Survey on Text Classification Algorithms: From Text to Predictions,” *Information*, vol. 13, no. 2, p. 83, Feb. 2022, doi: 10.3390/info13020083.

[32] V. Dogra, S. Verma, A. Singh, M. N. Talib, and M. Humayun, “Banking news-events representation and classification with a novel hybrid model using DistilBERT and rule-based features,” *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 3039–3054, 2021, [Online]. Available at: https://turcomat.org/index.php/turkbilmat/article/view/4954/4152.

[33] D. Te’eni *et al.*, “Reciprocal Human-Machine Learning: A Theory and an Instantiation for the Case of Message Classification,” *Manage. Sci.*, pp. 1–26, Nov. 2023, doi: 10.1287/mnsc.2022.03518.

[34] E. Hassan, T. Abd El-Hafeez, and M. Y. Shams, “Optimizing classification of diseases through language model analysis of symptoms,” *Sci. Rep.*, vol. 14, no. 1, p. 1507, Jan. 2024, doi: 10.1038/s41598-024-51615-5.

[35] Y. Ge *et al.*, “OpenAGI: When LLM Meets Domain Experts,” *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 1–30, Apr. 2023. [Online]. Available at: https://arxiv.org/abs/2304.04370v6.

[36] D. Kerrigan, J. Hullman, and E. Bertini, “A Survey of Domain Knowledge Elicitation in Applied Machine Learning,” *Multimodal Technol. Interact.*, vol. 5, no. 12, p. 73, Nov. 2021, doi: 10.3390/mti5120073.

[37] X. Lan, C. Gao, D. Jin, and Y. Li, “Stance Detection with Collaborative Role-Infused LLM-Based Agents,” *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 18, pp. 891–903, May 2024, doi: 10.1609/icwsm.v18i1.31360.

[38] E. (Olivia) Park, B. (Kevin) Chae, J. Kwon, and W.-H. Kim, “The Effects of Green Restaurant Attributes on Customer Satisfaction Using the Structural Topic Model on Online Customer Reviews,” *Sustainability*, vol. 12, no. 7, p. 2843, Apr. 2020, doi: 10.3390/su12072843.

[39] W. Chen, F. Rabhi, W. Liao, and I. Al-Qudah, “Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study,” *Electronics*, vol. 12, no. 12, p. 2605, Jun. 2023, doi: 10.3390/electronics12122605.

[40] I. Guillén-Pacho, C. Badenes-Olmedo, and O. Corcho, “Dynamic Topic Modelling for Exploring the Scientific Literature on Coronavirus: An Unsupervised Labelling Technique,” *Res. Sq.*, pp. 1–43, May 2023, doi: 10.21203/rs.3.rs-2872880/v1.

[41] P. Akbarighatar, I. Pappas, and P. Vassilakopoulou, “A sociotechnical perspective for responsible AI maturity models: Findings from a mixed-method literature review,” *Int. J. Inf. Manag. Data Insights*, vol. 3, no. 2, p. 100193, Nov. 2023, doi: 10.1016/j.jjimei.2023.100193.

[42] A. Alamsyah and N. D. Girawan, “Improving Clothing Product Quality and Reducing Waste Based on Consumer Review Using RoBERTa and BERTopic Language Model,” *Big Data Cogn. Comput.*, vol. 7, no. 4, p. 168, Oct. 2023, doi: 10.3390/bdcc7040168.

[43] P. Gupta, B. Ding, C. Guan, and D. Ding, “Generative AI: A systematic review using topic modelling techniques,” *Data Inf. Manag.*, vol. 8, no. 2, p. 100066, Jun. 2024, doi: 10.1016/j.dim.2024.100066.

[44] Z. Li *et al.*, “Improving the TENOR of Labeling: Re-evaluating Topic Models for Content Analysis,” *arXiv Comput. Languag*, vol. 1, pp. 1–20, 2024, [Online]. Available at: https://arxiv.org/pdf/2401.16348.

[45] S. Haque, Z. Eberhart, A. Bansal, and C. McMillan, “Semantic similarity metrics for evaluating source code summarization,” in *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*, May 2022, vol. 12, pp. 36–47, doi: 10.1145/3524610.3527909.

[46] D. Kici, G. Malik, M. Cevik, D. Parikh, and A. Başar, “A BERT-based transfer learning approach to text classification on software requirements specifications,” in *Proceedings of the Canadian Conference on Artificial Intelligence*, Jun. 2021, pp. 1–13, doi: 10.21428/594757db.a4880a62.

[47] T. Jagrič and A. Herman, “AI Model for Industry Classification Based on Website Data,” *Information*, vol. 15, no. 2, p. 89, Feb. 2024, doi: 10.3390/info15020089.

[48] J. Van Landeghem, M. Blaschko, B. Anckaert, and M.-F. Moens, “Benchmarking Scalable Predictive Uncertainty in Text Classification,” *IEEE Access*, vol. 10, pp. 43703–43737, 2022, doi: 10.1109/ACCESS.2022.3168734.

[49] J. Savla, D. Mehta DJSCE Aruna Gawade DJSCE Ramchandra Mangrulkar DJSCE, V. Vora, D. Mehta, A. Gawade, and R. Mangrulkar, “Classification of Diverse AI Generated Content: An In-Depth Exploration using Machine Learning and Knowledge Graphs,” *Res. Sq.*, pp. 1–23, Oct. 2023, doi: 10.21203/RS.3.RS-3500331/V1.

[50] Y. Mu *et al.*, “A BERT model generates diagnostically relevant semantic embeddings from pathology synopses with active learning,” *Commun. Med.*, vol. 1, no. 1, p. 11, Jul. 2021, doi: 10.1038/s43856-021-00008-0.

[51] S. Naaz, Z. Abedin, and D. Rizvi, “Sequence Classification of Tweets with Transfer Learning via BERT in the Field of Disaster Management,” *ICST Trans. Scalable Inf. Syst.*, vol. 8, no. 31, p. 169071, Jul. 2018, doi: 10.4108/eai.23-3-2021.169071.

[52] R. K. Kaliyar, A. Goswami, and P. Narang, “FakeBERT: Fake news detection in social media with a BERT-based deep learning approach,” *Multimed. Tools Appl.*, vol. 80, no. 8, pp. 11765–11788, Mar. 2021, doi: 10.1007/s11042-020-10183-2.

[53] S. Lebovitz, N. Levina, and H. Lifshitz-Assa, “Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts’ Know-What,” *MIS Q.*, vol. 45, no. 3, pp. 1501–1526, Sep. 2021, doi: 10.25300/MISQ/2021/16564.

[54] C. Lozano-Murcia, F. P. Romero, J. Serrano-Guerrero, and J. A. Olivas, “A Comparison between Explainable Machine Learning Methods for Classification and Regression Problems in the Actuarial Context,” *Mathematics*, vol. 11, no. 14, p. 3088, Jul. 2023, doi: 10.3390/math11143088.

[55] Y. Zhang, P. Calyam, T. Joshi, S. Nair, and D. Xu, “Domain-specific Topic Model for Knowledge Discovery through Conversational Agents in Data Intensive Scientific Communities,” in *2018 IEEE International Conference on Big Data (Big Data)*, Dec. 2018, pp. 4886–4895, doi: 10.1109/BigData.2018.8622309.

[56] T. Xiang, S. Chen, Y. Zhang, and R. Zhu, “TrendFlow: A Machine Learning Framework for Research Trend Analysis,” *Appl. Sci.*, vol. 13, no. 12, p. 7029, Jun. 2023, doi: 10.3390/app13127029.

[57] N. R. Mohammed and M. Mohammed, “Assessment of Twitter Data Clusters with Cosine-Based Validation Metrics Using Hybrid Topic Models,” *Ingénierie des systèmes d Inf.*, vol. 25, no. 6, pp. 755–769, Dec. 2020, doi: 10.18280/isi.250606.

[58] V. K. Garbhapu, “A comparative analysis of Latent Semantic analysis and Latent Dirichlet allocation topic modeling methods using Bible data,” *Indian J. Sci. Technol.*, vol. 13, no. 44, pp. 4474–4482, Nov. 2020, doi: 10.17485/IJST/v13i44.1479.