International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I3P125

Volume 72 Issue 3, 277-287, March 2024

Original Article

Leveraging Transformer-based BERTopic Model on Stakeholder Insights Towards Philippine UAQTE

Christian Y. Sy[1], Mary Joy P. Canon[2], Lany L. Maceda[3], Nancy M. Flores[4], Thelma D. Palaoag[5], Mideth B. Abisado[6]

[1,2,3]*Bicol University, Legazpi City, Philippines.*
[4,5]*University of the Cordilleras, Baguio City, Philippines.*
[6]*National University, Manila, Philippines.*

[1]*Corresponding Author : cysy@bicol-u.edu.ph*

Received: 28 November 2023 Revised: 15 February 2024 Accepted: 23 February 2024 Published: 17 March 2024

Abstract *- Free tertiary education expanded the scope of opportunities, enabling individuals to transform their aspirations into tangible achievements and empowering the nation's brightest minds to pave the way for economic and social development. This study aims to analyze and investigate student beneficiaries' perceptions within the Philippines' Universal Access to Quality Tertiary Education (UAQTE) framework, contributing to a more informed assessment of its impact on pursuing inclusive educational policies and reforms. The "Boses Ko" toolkit was utilized to collect responses, adopting a ground-up approach to capture insights directly from student beneficiaries across various Higher Education Institutions (HEIs). Through unsupervised machine learning using BERTopic modeling, latent topics and themes were identified within the qualitative data, enabling a holistic understanding of stakeholders' views. Silhouette and coherence scores and manual assessment by domain experts were used to evaluate the models. Key themes like "Educational Opportunity," "Program Implementation," and "Financial Support" were identified. Recommendations for policy reforms include enhancing educational opportunities, streamlining program implementation, and sustaining financial support within the UAQTE program.*

Keywords *- Free tertiary education, Stakeholder perceptions, Unsupervised machine learning, Topic modeling, BERTopic.*

1. Introduction

Tertiary education equips individuals with the essential knowledge, skills, and expertise required to excel in their respective fields [1]. It provides them with specialized education and training, enabling them to make meaningful contributions to society while pursuing their academic and professional goals. Through tertiary education, individuals gain a deep understanding of their respective disciplines, develop critical thinking abilities, and acquire practical skills that are highly valuable in their professional aspirations [2]. In developing nations, tertiary education is vital for cultural preservation, human capital development, poverty reduction, capacity building, and economic growth [3], [4]. Ensuring equitable access to quality tertiary higher education is a top priority globally, given its role in economic development and poverty reduction, particularly in developing countries [5]. Investing in this fosters sustainable development, unlocks human potential, and shapes a more prosperous and equitable future [6]. The United Nations (UN) recognizes the significance of the 4th Sustainable Development Goal (SDG), "Quality Education," for higher education in low-income nations. Emphasizing inclusive, quality education and lifelong learning opportunities, they aim to ensure equitable access to high-quality education, going beyond socioeconomic barriers [7], [8], [9]. This commitment highlights education as a catalyst for empowerment, social progress, and sustainable development [10]. Acknowledging this need, the Philippine government passed the Universal Access to Quality Tertiary Education Act (UAQTE), also referred to as Republic Act No. 10931, on August 13, 2017. This legislation mandates that all public higher education institutions (HEIs) and government-operated technical-vocational institutions (TVIs) offer free quality tertiary education to eligible Filipino students. RA10931 aims to enhance Philippine tertiary education through crucial provisions, including funding for broader participation, equal access to quality education, prioritizing academically capable disadvantaged students, efficient resource use, guidance for career choices, and recognizing public-private collaboration in the system [11]. The UAQTE program's implementation employs diverse strategies encompassing financial subsidies, resource allocation, and outreach initiatives. Financial subsidies provide crucial support to students by covering various educational expenses, including tuition fees and miscellaneous costs, alleviating financial burdens for students and families [12], [13]. Resource allocation strategies aim to strengthen the capacity

of higher education institutions (HEIs) and technical-vocational institutions (TVIs) by investing in infrastructure, faculty development, and curriculum enhancement, thereby enhancing educational standards and student outcomes [14], [15], [16].

Furthermore, targeted outreach and awareness campaigns are conducted to inform eligible students about UAQTE benefits and eligibility criteria, ensuring equitable access to higher education opportunities, particularly for marginalized and underserved populations. The Commission on Higher Education (CHED) reported that as of 2022, approximately 1.97 million students from various higher education institutions (HEIs) in the Philippines benefited greatly from the UAQTE program. Moreover, in the same period, the government's commitment to supporting tertiary education was further exemplified by the substantial number of beneficiaries receiving assistance. A total of 364,168 deserving students were fortunate recipients of the tertiary education subsidies (TES) program [17].

Considering these facts, evaluating the UAQTE program's implementation from a holistic perspective is crucial. This approach enables a more thorough assessment of the program's efficacy and facilitates the identification of challenges or areas requiring improvement. Moreover, it allows for gathering insights from individuals directly affected by the UAQTE initiative. However, previous research has often overlooked the qualitative dimensions of the UAQTE program's implementation, resulting in an inadequate representation of student beneficiaries' voices and viewpoints. While raw qualitative data offers valuable insights, it may not fully capture their nuanced experiences and perceptions. Thus, there is an apparent necessity for a more nuanced qualitative analysis of the UAQTE program, specifically focusing on the perspectives and experiences of student beneficiaries.

Additionally, the scarcity of research employing advanced analytical techniques like topic modeling and natural language processing hinders efforts to gain deeper insights into the UAQTE program's implementation and impact, particularly regarding the perspectives of student beneficiaries. Finally, a disconnect persists between policy objectives and the experiences of stakeholders, especially student beneficiaries, highlighting the necessity of bridging this gap to inform future policy decisions and reforms aimed at enhancing the accessibility, quality, and sustainability of tertiary education in the Philippines. This research aims to leverage unsupervised machine learning, implementing topic modeling techniques, natural language processing tools, and algorithms to uncover crucial insights, patterns, and correlations within the data, providing a more inclusive understanding of the UAQTE program's implementation and its impact on stakeholders. Topic modeling is a statistical technique that discovers latent topics or themes within a collection of documents without prior knowledge of the specific categories or labels [19], [20]. This method facilitates unveiling the underlying structure in textual data by grouping words that demonstrate frequent co-occurrence within similar contexts. Utilizing models such as BERTopic, researchers can uncover latent themes within student responses, offering a more profound insight into their experiences and perspectives concerning the UAQTE program [23], [24]. This unsupervised methodology enables a data-driven comprehension of student experiences, thereby enriching the comprehensiveness of the assessment [25], [26].

The study adopts the "Boses Ko" or "My Voice" participatory toolkit, which serves as a digital mechanism through which student beneficiaries share their experiences related to the program's implementation. The toolkit is a collaborative project between Bicol University (BU) and National University (NU), funded by the CHED-LAKAS (Leading the Advancement of Knowledge in Agriculture and Sciences) program, which focuses on research and development efforts in science and technology. Based on the pre-processed dataset, qualitative modeling using BERTopic harnesses the power of the BERT language model to capture the semantic nuances within the text, resulting in a more accurate representation of the underlying topics [27], [24]. The generated models are then evaluated using automated metrics and manual assessment of topics, including the silhouette score, the coherence score, and input from domain experts.

Domain experts play a critical role in the identification and categorization of the generated topic models with relevant themes [29], [30], [31]. Leveraging their expertise, experts assign labels based on observed word similarities within the models, ensuring precise representation of each topic [32], [33]. This meticulous labeling enhances the interpretability of the topic models, facilitating their practical application in informing policy decisions, such as those related to initiatives like the UAQTE program. Moreover, domain experts refine and validate the generated topics to ensure alignment with research objectives and program intricacies [34], [35], [36]. This approach provides a deeper understanding of the various perspectives that exist among stakeholders regarding the UAQTE implementation, revealing insights for informed decision-making.

Silhouette scoring is a metric used to evaluate the quality of clusters or topics generated by unsupervised learning algorithms, including topic modeling [37]. It measures the likeness of an object to its specific cluster, focusing on cohesion rather than emphasizing separation from other clusters. Coherence scoring evaluates the semantic relevance and coherence of the words within a given topic, ensuring meaningful connections and a cohesive thematic representation [38]. Domain experts play a vital role in topic modeling by contributing their specialized knowledge and judgment in model interpretation and evaluation [39], [40]. They help ensure that the generated topics are relevant,

coherent, and aligned with the domain's nuances [41], [42]. Their involvement is essential for meaningful and accurate topic modeling within a specific domain. Using domain experts to evaluate the generated topic models is valuable due to the qualitative insights they offer, which automated metrics may not capture. It allows for subjective assessment, enabling evaluators to consider interpretability, coherence, and relevance, which is crucial in practical applications [43], [33], also supports model refinement and considers domain-specific knowledge, enhancing contextual understanding [45], [46].

The study aims to better understand the stakeholders' perspectives on the UAQTE implementation by incorporating topic modeling as an unsupervised technique. It helps identify the key themes, discover the relationships between topics, and facilitates a comprehensive analysis of the textual data without relying on predefined categories or labels. Furthermore, taking into account the experiences and viewpoints of the students, a deeper understanding can be gained of the program's benefits, issues, challenges, potential areas for improvement, and sustainability. This inclusive approach fosters collaboration and participation in shaping the UAQTE program, better assessing its impact, and guiding future tertiary education policy decisions and reforms.

2. Methodology

This section details the methodology employed in our study, as presented in Figure 1 – the information processing phases. The process encompasses systematic data collection, pre-processing, and the implementation of document embeddings to enable advanced analysis. Central to our approach is the application of topic modeling techniques, which uncover hidden patterns and themes within the dataset. Subsequent steps included the interpretation and labeling of identified topics, adding contextual depth. The methodology culminates in model evaluation, ensuring the effectiveness of our techniques in aligning with the research objectives.

2.1. Data Collection

The "Boses Ko" toolkit was pivotal in collecting data from the student beneficiaries employing a ground-up approach. This means that the data collection starts from the grassroots level, emphasizing the perspectives of individuals involved in the UAQTE program.

The qualitative question for this study, "In 3-4 sentences, write your experiences as one of the beneficiaries of the UAQTE program." is directed at assessing student beneficiaries' perspectives. The study included a sample size of 2,800 student beneficiaries, selected from various State Universities and Colleges (SUCs) throughout the Bicol region, and this representation enabled a comprehensive analysis of the UAQTE program's implementation across different institutional contexts and frameworks.

2.2. Data Pre-processing

Data pre-processing involved critical steps that prepared the gathered responses before BERT embeddings [47], [48]. These essential steps encompassed tasks such as data cleaning and filtering by which non-English, duplicate, non-grantee, and blank responses were removed. This was followed by text standardization by converting the cleaned dataset to lowercase, eliminating special characters, punctuation marks, and digits to reduce noise and potential interference with the modeling process. This simplification aimed to create a more streamlined and coherent text representation, enhancing its suitability for subsequent modeling tasks.

Tokenization and stopwords removal by implementing the Natural Language Toolkit (NLTK) library were necessary pre-processing steps. Responses were tokenized into individual words or tokens, making analyzing and processing text data easier. Stopwords like "as," "one," "of," "the," "it," "me," "a," and "in" are common in the responses but typically lack significant meaning alone. Eliminating these enhanced the quality, interpretability, and efficiency of the generated topics within the UAQTE framework by reducing noise and emphasizing content words that conveyed the core themes.

Conversely, the study refrained from employing stemming and lemmatization techniques due to their potential to oversimplify words, reducing them to their most basic forms, which could risk the loss of significant meaning [49], [50]. For example, if the words "tuition" and "finances" were stemmed to "tuit" and "financ," respectively, it could result in unrecognizable words, potentially leading to confusion and a loss of clarity in the text. Similarly, "better" might be lemmatized to "good," which changes the comparison and could impact the overall message.

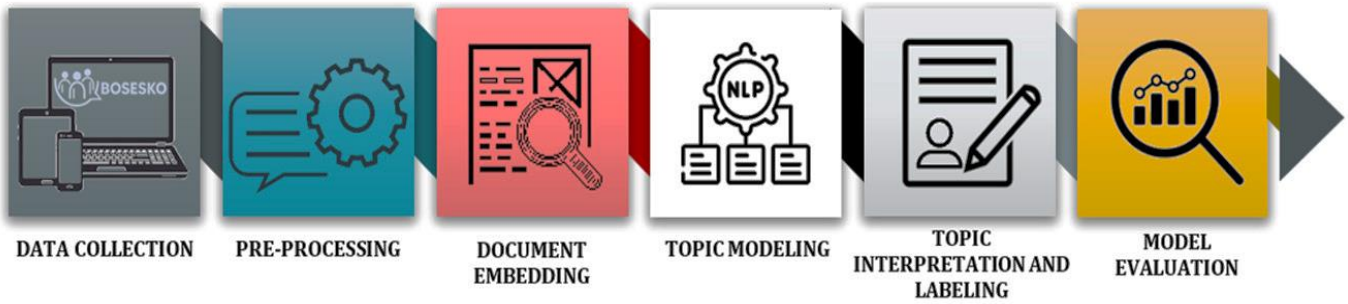

Fig. 1 Information processing phases

2.3. Document Embeddings

Document embedding is a crucial step in topic modeling as it transforms textual data into numerical representations that capture the essence of the documents. This study analyzed the dataset using a combination of the BERTopic and TF-IDF algorithms. This approach ensures that all contextual, semantic, and class-specific information is captured, leading to a complete understanding of topics and insights correlating with the research goals. As a modern and cutting-edge method, BERTopic or Bidirectional Encoder Representations from Transformers for Topic Modeling stands out for using pre-trained BERT models. BERTopic begins with tokenization, breaking down each document into individual sub-tokens and mapping them to word vectors from BERT [21], [22]. What sets BERTopic apart is its ability to capture contextual and semantic nuances, as BERT models consider the surrounding words to create contextual embeddings. Pooling techniques like mean or max pooling are applied to obtain a fixed-size vector for each document, resulting in dense vector representations that encode rich information about word meanings and context.

On the contrary, the term frequency-inverse document frequency (TF-IDF) utilizes a distinct method centered on term frequencies and inverse document frequencies to prioritize term significance. In this methodology, each dimension in these vectors corresponds to a unique term across the entire document corpus [18]. The values within these vectors are determined by two pivotal factors: term frequencies (TF), indicating the frequency of a term's occurrence within a document, and inverse document frequencies (IDF), gauging the scarcity of a term across the complete corpus. TF-IDF assigns greater weights to terms that appear frequently within a document but are rare across the entire corpus, thereby underscoring their significance.

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

Equation 1 calculates the importance of the term 't' in a document 'd' within a collection of documents. It combines two factors: the frequency of the term within the document ('tf') and the uniqueness or rarity of the term in the entire collection ('idf'). TF-IDF's strength lies in its ability to prioritize terms based on their significance within a document and their uncommonness across the entire dataset, resulting in straightforward and interpretable term importance rankings. This quality makes it well-suited for specific applications where a clear understanding of term importance is predominant. While TF-IDF provides simplicity and interpretability, it does not explicitly capture contextual information or semantics, making it appropriate for scenarios requiring only a high-level overview of themes. In specific scenarios, extending TF-IDF with Class-Based Term Frequency-Inverse Document Frequency (c-TF-IDF) can be advantageous, particularly when addressing some of its limitations and incorporating class-specific information into your document embeddings. However, this study opted not to utilize c-TF-IDF because the dataset inherently belongs to a specific domain, focusing on implementing Universal Access to Quality Tertiary Education (UAQTE) in the Philippines.

2.4. Topic Modeling/Extraction

After implementing document embeddings, BERTopic employs Uniform Manifold Approximation and Projection (UMAP), a dimensionality reduction technique. UMAP reduces the high-dimensional embeddings into a lower-dimensional space while preserving the data's structure and relationships [28]. This reduction helps in visualizing and clustering the data more effectively. Next is to extract meaningful topics from the dataset. BERTopic uses a Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) clustering algorithm to the lower-dimensional UMAP space. HDBSCAN identifies dense clusters of data points representing topics or subtopics in the data [32]. HDBSCAN's versatility in discovering clusters of varying shapes and sizes makes it particularly suitable for a wide range of datasets. However, what sets BERTopic apart is its remarkable capability for autonomous topic number detection, a feature that liberates researchers from the task of specifying the number of topics beforehand. By analyzing data-driven insights in the density and distribution of document vectors, BERTopic identifies natural cluster boundaries, streamlining the topic modeling process for efficiency and adaptability. Incorporating TF-IDF into the topic modeling process alongside BERTopic is a valuable and effective approach. BERTopic's strength lies in capturing semantic context, and when combined with TF-IDF, which emphasizes term importance and frequency, it enhances the overall quality of the modeling process. Initially, BERTopic generates topics based on semantic embeddings, while concurrently, TF-IDF is employed to represent documents traditionally, considering term frequencies. Figure 2 illustrates the workflow from document embeddings using BERT modeling to topic representation utilizing TF-IDF. Overall, the integration of TF-IDF with BERTopic combines the advantages of context-rich embeddings with traditional term frequency-based representations, providing a more comprehensive and interpretable approach to topic modeling and text analysis.

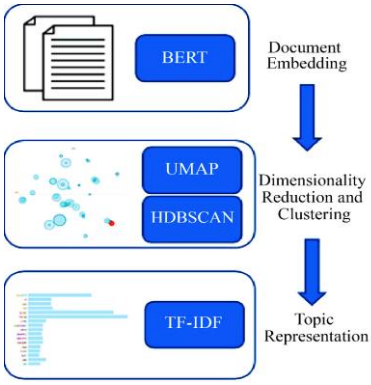

Fig. 2 Integration of BERTopic and TF-IDF

2.5. Hyper-Parameters

The following are the hyper-parameters used:

- min_topic_size. This parameter 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. This parameter 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. This parameter defines the desired number of topics extracted from your dataset. Choose it based on your dataset's nature and expected topic count.
- ngram_range. This parameter specifies the n-gram range for document term frequency representation, usually given as (min_n, max_n). For instance, ngram_range = (1, 2) considers both single words and two-word phrases, affecting topic granularity.

The topic modeling parameters selected for this study included a range of values to explore the optimal configuration. The number of topics varied from 5 to 20, allowing for a comprehensive investigation into different thematic clusters within the dataset. The top n words considered for each topic ranged from 5 to 10, offering flexibility in capturing the most significant terms associated with each identified theme. A range of 10 to 30 was set as the minimum topic size criterion to ensure meaningful topic sizes.

Lastly, n-gram ranges, specifying the sequence of word combinations, were considered with values (1,1), (1,2), and (2,2), allowing for the examination of unigram and bigram structures. 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. Ultimately, the quality of topics is assessed by their interpretability and relevance to the specific research goals.

2.6. Identification of Appropriate Themes

The primary objective in naming a topic model was to assign appropriate labels based on the word similarities observed within the generated models. This enhanced topic comprehension and effective communication, enabling their practical use in shaping the UAQTE program and contributing to a more informed assessment of its impact and informed future policy decisions related to tertiary education.

Critical to identifying labels was the involvement of domain experts, including CHED administrators, social scientists, data scientists, and UAQTE recipients. Their specialized knowledge contributed valuable contextual understanding, refined and validated topics, and enhanced their interpretability, ensuring alignment with the research objectives. Out of the many labels considered, Table 1 represents the finalized labels selected by the domain experts.

Table 1. Identified labels

Categories	Description
Financial Support	Responses that refer to the financial assistance provided, alleviation of financial burdens, and support with tuition fees, allowances, and expenses.
Educational Opportunity	Responses that describe students' gratitude towards the program, enabling them to pursue their preferred courses, continue their studies, and access quality education.
Program Implementation	Responses encompass a range of perspectives regarding the program's implementation, reflecting both positive and negative viewpoints.
Gratitude and Appreciation	Responses vary in their expressions of gratitude for being selected as one of the beneficiaries of the scholarship program.
No Label	Responses where a particular text or document does not have a predefined category or class assigned to it.

2.7. Model Evaluation

The Silhouette score is calculated as follows:

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

Where *s* is the silhouette score for the data point, *a* is the average distance from the data point to other points in the same cluster, and *b* is the average distance from the data point to points in the nearest neighboring cluster. The silhouette score can range from -1 to +1, where:

- A high positive value (close to +1) indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters, suggesting that the clusters are well-separated.
- A value near zero suggests that the object is on or very close to the boundary between two neighboring clusters.
- A negative value (close to -1) indicates that the object is incorrectly assigned to the neighboring cluster rather than its own, suggesting that the clusters may be overlapping or poorly defined.

Coherence scores range from 0 to 1, but the values may vary based on the coherence measure used. Some coherence measures can produce scores outside this range. The Coherence score is calculated as follows:

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

Where C_V(T) is the coherence score for the topic T, |T| is the number of words in the topic T, Word*i* and Word*j* represent two different words in the topic, the double summation ∑_{i=1} ^{|T|} ∑_ {j ≠ i} iterates over all pairs of distinct words in the topic ($i \neq j$) and sim (Word*i*, Word*j*) is the similarity measure between the two words.

The coherence score can range from 0 to 1, where:

- Coherence Score ≈ 0: Indicates that the topics lack meaningful connections, making them difficult to interpret.
- Coherence Score ≈ 0.2-0.4: Suggests that the topics exhibit some coherence, but their interpretability remains limited.
- Coherence Score ≈ 0.4-0.6: Implies that the topics are reasonably coherent and relatively interpretable.
- Coherence Score ≈ 0.6-0.8: This signifies that the topics are well-defined and highly coherent, making them easily interpretable.
- Coherence Score ≈ 0.8-1: Reflects excellent topics with closely related words, offering high interpretability.

Domain experts bring valuable contextual knowledge and subject matter expertise to the evaluation process. Their feedback is crucial because topic models should be statistically sound, semantically meaningful, and relevant to the specific domain. Domain experts can validate whether the topics generated make sense in the context of the research objectives.

Silhouette scores can be used in conjunction with coherence scores and manual inspection of topics by domain experts. A high silhouette score, coherence scores, and interpretable topics are a good sign of the quality of topics. Combining quantitative metrics and expert qualitative judgment provides a comprehensive assessment of topic model quality.

3. Results and Discussion

This section presents key results and findings from employing the BERTopic approach to model topics within the UAQTE dataset. Table 2 presents the configurations that yielded acceptable silhouette and coherence scores across various topic modeling experiments with varying hyperparameters.

Table 2. 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

These results offer insights into the effectiveness of specific hyperparameter settings in achieving coherent and meaningful topic structures, contributing to a comprehensive understanding of the dataset's inherent complexity.

The results obtained from the experimentation with various hyperparameter configurations offer valuable insights into the effectiveness of different settings in generating coherent and meaningful topic structures within the UAQTE dataset. Across the range of experiments conducted, notable variations in silhouette and coherence scores were observed, indicating the impact of hyperparameters on the quality of the generated topics. Specifically, configurations employing a unigram range (1,1), such as Experiment 1 and Experiment 4, yielded silhouette scores ranging from 0.765 to 0.781 and coherence scores ranging from 0.849 to 0.853, respectively.

These results suggest that while unigrams effectively capture individual words, they may overlook nuanced semantic relationships in multi-word expressions. In contrast, experiments utilizing a bigram range (1,2) consistently outperformed other configurations regarding coherence scores. For instance, Experiment 2 achieved a silhouette score of 0.814 and a coherence score of 0.863, indicating a significant improvement in capturing meaningful topic structures by considering two-word combinations.

Variations in the number of topics and minimum topic size also influenced the quality of the generated topics. Experiments with a smaller minimum topic size of 13, such as Experiments 1 to 3, generally exhibited higher silhouette and coherence scores than those with larger minimum topic sizes. This observation suggests that requiring fewer documents per topic leads to more well-defined clusters and enhances the interpretability of the resulting topics. Interestingly, the optimal number of topics varied across experiments, ranging from 3 in Experiment 11 to 9 in Experiment 1.

This variability underscores the importance of considering the dataset's specific characteristics and thematic diversity when determining the appropriate number of topics. Additionally, experiments with a larger n-gram range (2,2) generally yielded lower silhouette and coherence scores than their counterparts with a bigram range (1,2), indicating that while capturing longer sequences may enhance topic granularity, it may also introduce noise and reduce topic coherence. These emphasize the importance of carefully selecting and fine-tuning hyperparameters based on dataset characteristics to achieve the best results in topic modeling. By leveraging a combination of silhouette and coherence scores, researchers can effectively evaluate the quality of topics and guide the selection of the most suitable model configuration for analyzing textual data within the UAQTE dataset. Domain experts have identified several key themes in the dataset, "Educational Opportunity" emerges as the most prominent and pervasive theme within the dataset.

Table 3. Labeled model

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

This label encompasses discussions about access to education, opportunities for academic advancement, and the importance of educational support systems. The presence of terms such as “funded,” “study,” “free,” “pursue,” and “college” within Topic 0 of the labeled model in Table 3 underscores the centrality of educational opportunities captured by the UAQTE dataset. “Program Implementation” is another crucial theme consistently recognized by domain experts. This theme encompasses discussions surrounding the execution and operation of the UAQTE program itself. Terms like “helpful,” “convenient,” “complaints,” “need to be improved,” and “quality system” within Topic 1, Topic 2, and Topic 5 of the labeled model reflect respondents’ perceptions and experiences related to the implementation of the UAQTE program.

“Financial Support” emerges as a recurring and significant theme, reflecting discussions related to financial assistance or funding opportunities. Terms such as “financial,” “expenses,” and “cost” within Topic 3 and Topic 7 of the labeled model highlight the importance of financial support in facilitating access to education and alleviating economic barriers. Lastly, expressions of “Appreciation and Gratitude” are evident throughout the dataset, indicating respondents’ sentiments of thanks and acknowledgement towards various aspects of the UAQTE program. While not explicitly captured as a standalone theme within the presented labeled model, expressions of gratitude are implicitly woven into discussions across multiple topics, reflecting respondents’ positive experiences and appreciation for the opportunities provided by the UAQTE program.

By integrating domain expert insights with topic modeling results, a more comprehensive understanding of the thematic landscape within the UAQTE dataset is achieved. The alignment between expert-identified themes and the generated topics further validates the relevance and interpretability of the topic modeling outcomes, enhancing the ability to extract meaningful insights from the textual data. These insights contribute to a deeper understanding of the issues and perspectives prevalent within the UAQTE dataset, informing future research directions and programmatic initiatives within the domain.

Figure 3, the intertopic distance map, provides a visual representation of the relationships between the generated clusters, offering deeper insights into the thematic structure of the UAQTE dataset.

Examining the map makes it apparent that clusters exhibit a high degree of alignment with the labels assigned by domain experts. This observation serves as a robust indicator of the quality and relevance of the generated topics.

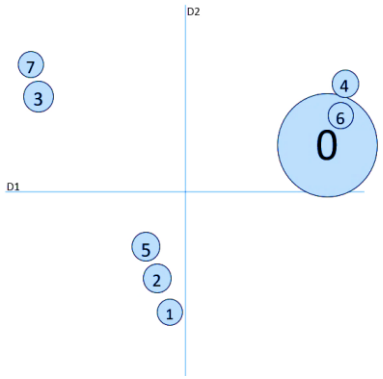

Fig. 3 Intertopic distance map

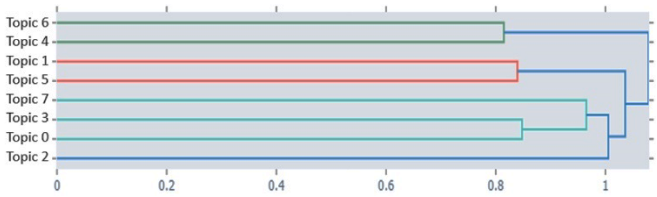

Fig. 4 Hierarchical clustering

For example, clusters containing topics 0, 4, and 6 are closely grouped, forming a cohesive cluster labeled "Educational Opportunity." Similarly, topics 1, 2, and 5 are clustered, corresponding to the "Program Implementation" label. Additionally, topics 3 and 7 form a distinct cluster identified as "Financial Support." This clustering pattern mirrors the thematic categories identified by domain experts, underscoring the consistency and effectiveness of the topic modeling approach in capturing central themes within the dataset. Furthermore, the alignment between the clustering patterns and expert evaluations is reinforced by the quantitative metric scores obtained during the topic modeling process.

The strong correspondence between metric scores such as silhouette and coherence scores and expert-assigned labels provides further validation of the effectiveness of quantitative analysis in identifying central themes within the UAQTE dataset. This alignment between qualitative expert insights and quantitative metrics enhances confidence in the validity and interpretability of the generated topics, facilitating a more nuanced understanding of the dataset's thematic landscape. Overall, the intertopic distance map is a powerful analytical tool for visualizing the thematic relationships within the UAQTE dataset.

The alignment between generated clusters and expert-assigned labels, coupled with the validation provided by quantitative metrics, underscores the robustness and effectiveness of the topic modeling approach in uncovering meaningful insights from textual data. This analytical framework enables researchers to identify and explore central themes within the dataset, facilitating deeper exploration and interpretation of the underlying issues and perspectives present within the UAQTE dataset.

The hierarchical clustering results confirm and reinforce the thematic relationships observed in the intertopic distance map. Notably, topics 6 and 4 and 1 and 5 exhibit clustering in both methods, suggesting shared thematic elements between these pairs. This consistent clustering pattern across different analytical approaches highlights the robustness and reliability of the identified thematic clusters. Moreover, while topic 2 appears to stand alone in the hierarchical clustering analysis, its connections to topics 3, 0, and subsequently to topic 7 and its eventual linkage with topics 1 and 5 align closely with the findings from the intertopic distance map. This alignment underscores the coherence and relevance of the identified thematic clusters, further validating the consistency of the results obtained from both analyses.

The agreement between the hierarchical clustering results and the intertopic distance map strongly indicates the dataset's thematic structure. By corroborating thematic relationships across multiple analytical methods, this agreement reinforces the reliability of the identified themes. It provides researchers with a robust framework for understanding the underlying thematic landscape within the UAQTE dataset.

4. Conclusion and Recommendations

Implementing BERTopic modeling has demonstrated significant advantages in uncovering the nuanced thematic landscape within the UAQTE dataset. This can be attributed to several key factors. Firstly, BERTopic leverages advanced transformer-based language models, such as BERT (Bidirectional Encoder Representations from Transformers), renowned for capturing intricate semantic relationships within textual data. By harnessing the contextual understanding provided by BERT, BERTopic effectively extracts distinct topics while preserving their coherence and semantic meaning, thus surpassing traditional topic modeling approaches.

Moreover, the strategic incorporation of n-gram ranges, particularly the preference for models incorporating bigrams, played a pivotal role in enhancing topic quality. Bigrams, capturing two-word combinations, inherently enrich topic coherence by incorporating nuanced semantic relationships that unigrams may miss. This emphasis on bigrams underscores the significance of considering multi-word expressions in topic modeling, thereby elevating the quality and interpretability of the generated topics.

Furthermore, adopting a dual evaluation methodology, comprising both automatic metrics and manual evaluation by domain experts, contributed significantly to the robustness of the topic modeling process. Integrating automated metrics such as silhouette and coherence scores facilitated quantitative assessment, providing objective topic quality and coherence measures. Concurrently, manual evaluation by domain experts offered qualitative insights, validating the relevance and interpretability of the generated topics.

This dual evaluation approach ensured a comprehensive assessment of the topic models, mitigating biases and enhancing confidence in the reliability of the results. Integrating BERTopic modeling with dual evaluation methodologies has enabled a comprehensive understanding of the thematic priorities within the UAQTE dataset. By leveraging analytical techniques alongside expert insights, policymakers can make informed decisions to optimize program effectiveness and ensure equitable access to quality tertiary education for all beneficiaries. This multifaceted approach underscores the significance of employing cutting-edge methodologies and rigorous evaluation methods in tackling complex research questions and guiding policy interventions effectively.

Given the predominantly positive perspective of the beneficiaries, the researchers propose the following targeted policy recommendations to further align and enhance the

UAQTE program with the themes identified in the dataset: First, to enhance Educational Opportunities, the UAQTE program should develop policies that broaden access to quality tertiary education and prioritize addressing the educational disparities faced by the disadvantaged sector.

These measures should involve increasing scholarship availability and facilitating diverse educational pathways. Second, in streamlining program implementation, recognizing the critical role of program implementation, it is essential to consider policies focusing on streamlining processes to address areas that need improvement, thereby rectifying any issues lacking and addressing any complaints.

Clear and transparent communication channels must be established to ensure effective information dissemination among program administrators, educational institutions, and beneficiaries. Lastly, to sustain financial support, the program should formulate policies to guarantee its long-term sustainability, emphasizing efficient procedures for providing free higher education and alleviating the financial burden on students and parents.This sustainability is maintained through continuous funding allocation, facilitating students' access to tertiary education without cost.

Funding Statement

Philippine Commission on Higher Education (CHED) Leading the Advancement of Knowledge in Agriculture and Science (LAKAS) Project No. 2021-007, eParticipation 2.1: Harnessing Natural Language Processing (NLP) for Community Participation.

Acknowledgements

The researchers express gratitude to the Philippine Commission on Higher Education (CHED) and the Leading the Advancement of Knowledge in Agriculture and Science (LAKAS). The generous financial support provided by this project has been instrumental in facilitating our research endeavors. The researchers deeply appreciate the invaluable contribution of CHED and the LAKAS Project, recognizing their pivotal role in advancing the pursuit of knowledge.

References

[1] Antonietta Di Giulio, and Rico Defila, "Enabling University Educators to Equip Students with Inter-and Transdisciplinary Competencies," *International Journal of Sustainability in Higher Education*, vol. 18, no. 5, pp. 630-647, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[2] S. Tomy, and E. Pardede, "An Entrepreneurial Intention Model Focusing on Higher Education," *International Journal of Entrepreneurial Behaviour and Research*, vol. 26, no. 7, pp. 1423-1447, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Mark Stafford-Smith et al., "Integration: The Key to Implementing the Sustainable Development Goals," *Sustainability Science*, vol. 12, no. 6, pp. 911-919, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[4] Anekwe Rita Ifeoma et al., "Effect of Entrepreneurship Development on Poverty Alleviation in Nigeria," *IOSR Journal of Business and Management (IOSR-JBM)*, vol. 20, no. 2, ver. 10, pp. 80-87, 2018. [Google Scholar] [Publisher Link]

[5] Savo Heleta, and Tohiera Bagus, "Sustainable Development Goals and Higher Education: Leaving Many Behind," *Higher Education*, vol. 81, no. 1, pp. 163-177, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Peter Messerli et al., "*Global Sustainable Development Report 2019: The Future is Now-Science for Achieving Sustainable Development*," Independent Group of Scientists, pp. 1-216, 2019. [Google Scholar] [Publisher Link]

[7] Raima Nazar et al., "Role of Quality Education for Sustainable Development Goals (SDGS)," *People: International Journal of Social Sciences*, vol. 4, no. 2, pp. 486-501, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] William Rosa et al., *A New Era in Global Health Nursing and the United Nations 2030 Agenda for Sustainable Development*, Springer Publishing Company, pp. 1-624, 2017. [Google Scholar] [Publisher Link]

[9] Serena Clark et al., "Including Digital Connection in the United Nations Sustainable Development Goals: A Systems Thinking Approach for Achieving the SDGs," *Sustainability*, vol. 14, no. 3, pp. 1-13, 2022. [Crossref] [Google Scholar] [Publisher Link]

[10] W. Leal Filho et al., "Using the Sustainable Development Goals towards a Better Understanding of Sustainability Challenges," *International Journal of Sustainable Development and World Ecology*, vol. 26, no. 2, pp. 179-190, 2019. [Crossref] [Google Scholar] [Publisher Link]

[11] Congress of the Philippines, Republic Act No. 10931, Seventeenth Congress, Republic of the Philippines, 2017. [Online]. Available: https://lawphil.net/statutes/repacts/ra2017/ra_10931_2017.html

[12] Kidjie Saguin, "The Politics of De-Privatisation: Philippine Higher Education in Transition," *Journal of Contemporary Asia*, vol. 53, no. 3, pp. 471-493, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Joy N. Emmanuel, "Affordability in College Access: Improving Equitable Value for Low-Income, First-Generation, and Students of Color," *The Vermont Connection*, vol. 44, no. 1, pp. 1-16, 2023. [Google Scholar] [Publisher Link]

[14] Divina Edralin, and Ronald Pastrana, "Technical and Vocational Education and Training in the Philippines: In Retrospect and its Future Directions," *Bedan Research Journal*, vol. 8, no. 1, pp. 138-172, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Adithya Kumar 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," *International Journal of Management, Technology, and Social Sciences (IJMTS)*, vol. 8, no. 3, pp. 103-153, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Mustafa Kayyali, "The Relationship between Rankings and Academic Quality," *International Journal of Management, Sciences, Innovation, and Technology IJMSIT*, vol. 4, no. 3, pp. 1-11, 2023. [Google Scholar] [Publisher Link]

[17] Stephanie Sevillano, 2.46M Filipinos Get Free College Education under the Duterte Admin, Philippine News Agency, 2022. [Online]. Available: https://www.pna.gov.ph/articles/1175587

[18] Firas Saidi, Zouheir Trabelsi, and Eswari Thangaraj, "A Novel Framework for Semantic Classification of Cyber Terrorist Communities on Twitter," *Engineering Applications of Artificial Intelligence*, vol. 115, 2022. [Crossref] [Google Scholar] [Publisher Link]

[19] Karoliina Isoaho, Daria Gritsenko, and Eetu Mäkelä, "Topic Modeling and Text Analysis for Qualitative Policy Research," *Policy Studies Journal*, vol. 49, no. 1, pp. 300-324, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] Minkyoung Song et al., "Comparative Analysis of National Cyber Security Strategies Using Topic Modelling," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 12, pp. 62-69, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Yu Meng et al., "Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations," *WWW 2022 - Proceedings of the ACM Web Conference*, New York, United States, pp. 3143-3152, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Raquel Silveira et al., "Topic Modelling of Legal Documents via LEGAL-BERT 1," *Proceedings of the First International Workshop RELATED - Relations in the Legal Domain, in Conjunction with ICAIL 22021*, pp. 1-9, 2021. [Google Scholar] [Publisher Link]

[23] Aly Abdelrazek et al., "Topic Modeling Algorithms and Applications: A Survey," *Information Systems*, vol. 112, 2023.[CrossRef] [Google Scholar] [Publisher Link]

[24] Fatima Alhaj et al., "Improving Arabic Cognitive Distortion Classification in Twitter using BERTopic," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 13, no. 1, pp. 854-860, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[25] Bilge Gencoglu et al., "Machine and Expert Judgments of Student Perceptions of Teaching Behavior in Secondary Education: Added Value of Topic Modeling with Big Data," *Computers & Education*, vol. 193, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Xieling Chen et al., "Detecting Latent Topics and Trends in Educational Technologies over Four Decades Using Structural Topic Modeling: A Retrospective of all Volumes of Computers & Education," *Computers & Education*, vol. 151, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[27] Wenqi Fan et al., "Recommender Systems in the Era of Large Language Models (LLMs)," *arXiv*, 2023 [CrossRef] [Google Scholar] [Publisher Link]

[28] Sahil Sawant et al., "An Enhanced BERTopic Framework and Algorithm for Improving Topic Coherence and Diversity," *2022 IEEE 24th Int Conf on High Performance Computing & Communications*; *8th Int Conf on Data Science & Systems*; *20th Int Conf on Smart City*; *8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, Hainan, China, pp. 2251-2257, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[29] Eunhye (Olivia) Park et al., "The Effects of Green Restaurant Attributes on Customer Satisfaction Using the Structural Topic Model on Online Customer Reviews," *Sustainability*, vol. 12, no. 7, pp. 1-20, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[30] Weisi Chen et al., "Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study," *Electronics*, vol. 12, no. 12, pp. 1-22, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[31] Ibai Guillén-Pacho, Carlos Badenes-Olmedo, and Oscar Corcho, "Dynamic Topic Modelling for Exploring the Scientiic Literature on Coronavirus: An Unsupervised Labelling Technique," *Research Square*, 2023, [CrossRef] [Google Scholar] [Publisher Link]

[32] Pouria Akbarighatar, Ilias Pappas, and Polyxeni Vassilakopoulou, "A Sociotechnical Perspective for Responsible AI Maturity Models: Findings from A Mixed-Method Literature Review," *International Journal of Information Management Data Insights*, vol. 3, no. 2, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[33] Andry Alamsyah, and Nadhif Ditertian Girawan, "Improving Clothing Product Quality and Reducing Waste Based on Consumer Review Using RoBERTa and BERTopic Language Model," *Big Data and Cognitive Computing*, vol. 7, no. 4, pp. 1-21, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[34] Priyanka Gupta et al., "Generative AI: A Systematic Review Using Topic Modelling Techniques," *Data and Information Management*, pp. 1-66, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[35] Zongxia Li et al., "Beyond Automated Evaluation Metrics: Evaluating Topic Models on Practical Social Science Content Analysis Tasks," *arXiv*, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[36] Chau Minh Pham et al., "TopicGPT: A Prompt-based Topic Modeling Framework," *arXiv*, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] Lijimol George, and P. Sumathy, "An Integrated Clustering and BERT Framework for Improved Topic Modeling," *International Journal of Information Technology*, vol. 15, pp. 2187-2195, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[38] Maarten Grootendorst, "BERTopic: Neural Topic Modeling with A Class-Based TF-IDF Procedure," *arXiv*, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[39] Soya Park et al., "Facilitating Knowledge Sharing from Domain Experts to Data Scientists for Building NLP Models," *26th International Conference on Intelligent User Interfaces*, TX USA, pp. 585-596, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[40] Ryan J. Gallagher et al., "Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 529-542, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[41] Yueshen Xu et al., "Hierarchical Topic Modeling with Automatic Knowledge Mining," *Expert Systems with Applications*, vol. 103, pp. 106-117, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[42] Giuseppe Carenini et al., "Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis," *Journal of Medical Internet Research*, vol. 23, no. 2, pp. 1-12, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[43] Angela Fan, Finale Doshi-Velez, and Luke Miratrix, "Prior Matters: Simple and General Methods for Evaluating and Improving Topic Quality in Topic Modeling," *arXiv*, pp. 1-18, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[44] Alexander Hoyle et al., "Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence," *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, pp. 1-16, 2021. [Google Scholar] [Publisher Link]

[45] Eickhoff Matthias, and Wieneke Runhild, "Understanding Topic Models in Context: A Mixed-Methods Approach to the Meaningful Analysis of Large Document Collections," *Proceedings of the 51st Hawaii International Conference on System Sciences*, pp. 903-912, 2018. [Google Scholar] [Publisher Link]

[46] Mekhail Mustak et al., "Artificial Intelligence in Marketing: Topic Modeling, Scientometric Analysis, and Research Agenda," *Journal of Business Research*, vol. 124, pp. 389-404, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[47] Bhart Gupta, P. Prakasam, and T. Velmurugan, "Integrated BERT Embeddings, BiLSTM-BiGRU and 1-D CNN Model for Binary Sentiment Classification Analysis of Movie Reviews," *Multimedia Tools and Applications*, vol. 81, no. 23, pp. 33067-33086, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[48] Akshay Khatri, P. Pranav, M. Anand Kumar, "Sarcasm Detection in Tweets with BERT and GloVe Embeddings," *arXiv*, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[49] Eirik Duesund Helland, "*Tackling Lower-Resource Language Challenges: A Comparative Study of Norwegian Pre-Trained BERT Models and Traditional Approaches for Football Article Paragraph Classification*," Master Theses, Norwegian University of Life Sciences, pp. 1-97, 2023. [Google Scholar] [Publisher Link]

[50] Simon Roth, "*Biased Machines in the Realm of Politics*," Doctoral Theses, Konstanz University, pp. 1-125, 2022. [Google Scholar] [Publisher Link]