

Transforming Education Policy: Evaluating UAQTE Program Implementation through LDA, BoW and TF-IDF Techniques

Christian Y. Sy*, Lany L. Maceda*, Thelma D. Palaoag**, Mideth B. Abisado***

*Bicol University, Legazpi City, Philippines

** University of the Cordilleras, Baguio City, Philippines

*** National University, Manila, Philippines.

cysy@bicol-u.edu.ph, llmaceda@bicol-u.edu.ph, tdpalaoag@uc-bcf.edu.ph, mbabisado@national-u.edu.ph

Abstract— This research aimed to deepen our understanding of the issues and challenges faced by beneficiaries of the Philippines' Universal Access to Quality Tertiary Education (UAQTE) program by utilizing Latent Dirichlet Allocation (LDA) complemented by Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) techniques. The study used the "Boses Ko" participatory toolkit, a digital portal to collect student responses from the various Higher Education Institutions (HEIs). The analysis showcased the LDA model's proficiency in identifying distinct clusters, underscoring its ability to generate coherent and meaningful topics that provide interpretable and insightful dataset representations. Evaluation of the models included Silhouette and Coherence Scores, supplemented by manual assessments by domain experts. This comprehensive approach led to the identification of key themes, including "Academic Difficulties," "Financial Difficulties," "Pandemic-Related Challenges," "Grant Disbursement," and "Program Implementation." The identified themes present actionable policy recommendations geared towards strengthening academic support, financial assistance, flexible grant disbursement, addressing pandemic-related challenges, and establishing a structured feedback mechanism. These measures collectively aim to enhance the overall implementation of the UAQTE program.

Keywords— Topic Modeling, Latent Dirichlet Allocation (LDA), Bag-of-Words, TF-IDF, UAQTE program

I. INTRODUCTION

Philippines, where economic and social inequalities persist, tertiary education emerges as a transformative force. It becomes a powerful channel for social mobility, granting individuals from less privileged economic backgrounds equal access to higher education opportunities traditionally reserved for the more affluent. The emphasis on tertiary education aligns with the global agenda, resonating with the United Nations' fourth Sustainable Development Goal (SDG), which emphasizes "Quality Education" as a catalyst for sustainable development [1]–[3].

In the quest for equitable education, the Universal Access to Quality Tertiary Education (UAQTE) program, also known as Republic Act No. 10931 was passed on August 13, 2017, requiring all public higher education institutions (HEIs) and government-run technical-vocational institutions (TVIs) to

provide free quality tertiary education to eligible Filipino students. This initiative is instrumental in helping individuals break free from the cycle of poverty and catalyze national progress, thereby playing a vital role in the country's socioeconomic development [4], [5]. The program's commitment to providing free tertiary education is fundamental in equipping students with essential knowledge, skills, and critical thinking abilities that are indispensable tools for overcoming the intricate challenges of today's world [6], [7]. This educational empowerment broadens students' perspectives, unveiling various career possibilities and empowering graduates to contribute substantially to their communities and the country [8], [9].

As the UAQTE program expands, it becomes increasingly important to assess its implementation holistically by considering the perspectives of student beneficiaries. This approach allows for a more thorough evaluation of its effectiveness, identifies potential areas for improvement, and gathers valuable insights. To achieve this, our research aims to leverage topic modeling techniques implementing Latent Dirichlet Allocation (LDA), Bag of Words (BoW), and Term Frequency-Inverse Document Frequency (TF-IDF) to uncover patterns, and correlations within the dataset, providing a comprehensive understanding of the UAQTE program's implementation and its impact on stakeholders.

Topic modeling is a statistical technique within natural language processing domain that is used to uncover underlying themes or topics in a textual data collection. It facilitates the organization and categorization of extensive unstructured text, enabling the extraction of meaningful insights [10], [11]. Through topic modeling, we can effectively reveal the underlying structure in the textual data by grouping words that tend to co-occur frequently within the same context [12]–[15].

Central to this study is the "Boses Ko" participatory toolkit, a digital platform developed through a collaborative effort between Bicol University and National University, funded by the Commission on Higher Education - Leading the Advancement of Knowledge in Agriculture and Sciences (CHED-LAKAS). This tool is crucial in collecting the viewpoints of student beneficiaries of the UAQTE program, serving as a pivotal avenue for students to articulate the encountered issues and challenges.

Building upon the pre-processed dataset, this study integrates a multifaceted qualitative modeling approach, employing Latent Dirichlet Allocation (LDA), a well-established methodology, complemented by the use of Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) techniques, both of which play a crucial role in the topic modeling process. The Bag of Words (BoW) model lays the groundwork for representing textual data in this study, analyzing documents by the frequency of words while setting aside their order or contextual significance [16]–[18]. This simplification of the dataset aids Latent Dirichlet Allocation (LDA) in efficiently identifying and classifying topics based on the occurrence of words.

Consecutively, the study employs the Term Frequency-Inverse Document Frequency (TF-IDF) technique to elevate the analysis. TF-IDF measures the importance of each word within a document relative to a collection of documents, assigning a weight to the frequency of each word. This method effectively diminishes the impact of commonly occurring words across the dataset, instead highlighting words that are more distinctive to individual documents [19]–[21]. Such weighted differentiation enhances LDA's ability to uncover and distinguish more relevant and unique topics in the dataset.

The generated models are evaluated combining automated metrics using the Silhouette and Coherence Score with manual topic assessment guided by domain experts' insights. The Silhouette Score is instrumental in determining the clustering quality of the model by measuring how similar each data point is to its own cluster compared to neighboring clusters [22], [23]. This metric is crucial for assessing the effectiveness of the clustering approach taken by the LDA model. The Coherence Score, on the other hand, is used to evaluate the interpretability and quality of the topics identified by the model. It quantitatively measures the semantic relationships among the most significant words within each topic, ensuring they form a coherent and meaningful theme [24], [25]. These automated metrics offer objectivity and efficiency in the initial stages of model assessment, allowing for rapid refinement and baseline evaluations of the model's performance.

The domain experts' insights capture the depth and nuances of the model's output. They bring a vital perspective to the evaluation process, considering aspects such as the relevance, interpretability, and real-world applicability of the topics generated by the model [26]–[28]. Their specialized knowledge and experience in the domain allow them to provide a more distinct analysis, aligning the model's findings with the intricacies and specific requirements of the research objectives.

Ultimately, this research aims to provide evidence-based insights to enhance and optimize the UAQTE program implementation, benefiting students and the wider educational sector. It highlights the importance of collaboration and stakeholder involvement, enabling education sectors to actively shape the UAQTE program and influence future tertiary education policies. This strategy demonstrates a commitment to ongoing improvement and adaptive policymaking, ensuring the program stays dynamic, aligns with the evolving educational landscape, and meets the diverse needs of its beneficiaries.

II. METHODOLOGY

This section details the methodology used. Figure 1 illustrates the information processing phases, which include data collection and dataset pre-processing, feature extraction, topic modeling, topic interpretation and labeling, and model evaluation.



Figure 1. Information Processing Phases

A. Data Collection

The "Boses Ko" toolkit has played a pivotal role in implementing a bottom-up approach to data collection from student beneficiaries, emphasizing the importance of insights from those directly involved in the UAQTE program. Tailored to capture the unique perspectives of these students, the central qualitative question of this study is: "What are the specific issues and challenges faced by beneficiaries of the UAQTE program?"

The study included a sample of 2,800 student beneficiaries from various State Universities and Colleges (SUCs) across the Bicol region. This broad selection ensured a thorough evaluation of the UAQTE program's effectiveness and impact in different institutional settings.

B. Data Pre-processing

The preparation of the collected responses for feature extraction began with critical data pre-processing steps. This included removing non-English, duplicate, non-grantee, and empty responses, followed by text standardization. Standardization involves converting text to lowercase and removing special characters, punctuation, and digits, aiming to refine the textual representation for clearer analysis and reduced noise. Further processing utilized the Natural Language Toolkit (NLTK) for tokenization and the removal of common stopwords (e.g., "the," "is," "and"). This step segmented the text into individual words or tokens, and removed frequently occurring but semantically limited words, enhancing the clarity and interpretability of the topics. Additionally, domain-specific terms like "UAQTE," "issues," "challenges," and "beneficiaries" were treated as additional stopwords to focus the text analysis and produce more meaningful results by highlighting key terms.

The study intentionally avoided using stemming and lemmatization techniques to retain the nuanced meanings of words. Applying these methods could risk oversimplifying the language and result in a loss of textual clarity, as they reduce words to their base forms or substitute them with similar-meaning terms. For instance, transforming "synchronous" into "synchron" or "better" into "good" might change the original intent or context, potentially leading to misinterpretation or loss of critical information.

C. Feature Extraction

In the feature extraction of Latent Dirichlet Allocation (LDA), two predominant techniques are utilized: the Bag of Words (BoW) model and the Term Frequency-Inverse Document Frequency (TF-IDF) transformation. These methods are instrumental in uncovering underlying structures and enhancing insights from text data [29], [30]. BoW models each document as a vector, quantifying the frequency of words within it [31], [32]. This enables LDA to translate textual data into numerical features, facilitating the identification of underlying topics and their distribution across documents.

TF-IDF augments feature representation quality, building upon the BoW model [33]. It assigns weights to words based on their relevance in individual documents and the entire corpus, balancing word significance. This step, following BoW, produces a more informative and distinct feature matrix, aiding LDA in assigning importance to words unique to specific documents or topics. By doing so, it diminishes the weight of common words.

Integrating TF-IDF after the BoW method is crucial for refining feature representation, reducing noise, and improving topic distinction within textual data [34]. This methodology ensures that LDA more accurately captures the distinctive elements of the dataset, resulting in more precise and meaningful categorization of topics.

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

Equation 1 assesses the importance of the term 't' within a specific document 'd' throughout a collection of documents. This evaluation incorporates two primary factors: the frequency of the term within the document ('tf') and its scarcity or distinctiveness across the entire collection of documents ('idf').

D. Topic Modeling/Extraction

Building on the initial feature extraction phase using the Bag of Words (BoW), where documents are represented as frequency vectors of words, LDA fundamentally uncovers hidden topics within the text corpus. Operating on the document-term matrix derived from BoW, LDA interprets each document as a composition of various topics. Each word occurrence within a document is associated with one of these topics. The algorithm dynamically refines its understanding, adjusting topic associations for each word by analyzing word co-occurrence patterns throughout the corpus.

The integration of the TF-IDF methodology refines the BoW representation. This approach assigns specific weights to terms, considering their frequency within individual documents and their overall relevance in the entire corpus. This weighted TF-IDF matrix then serves as the input for LDA, enabling it to delve into the significance and distribution of terms to pinpoint underlying themes. By evaluating the prominence of particular terms across different documents, the algorithm seeks to grasp the semantic interconnections among words. This approach enriches the topic modeling process, ensuring a more sophisticated and accurate interpretation of the textual content.

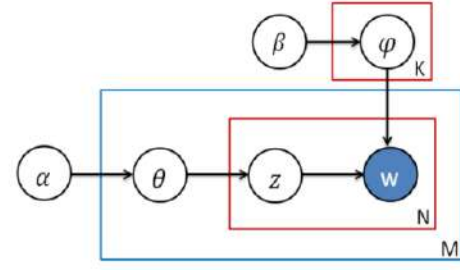


Figure 2. Topic Extraction using LDA

Figure 3 presents the Latent Dirichlet Allocation (LDA) model's generative mechanism, offering a systematic document formation perspective. This includes the process of topic selection, word distribution within these topics, and the role of hyperparameters α and β in shaping the distributions of topics and words.

- α (alpha): This parameter influences the Dirichlet prior for topic distributions within each document (θ). It determines how topics are varied across individual documents.
- β (beta): This parameter governs the Dirichlet prior for word distributions within each topic (ϕ). It specifies how words are distributed across topics.
- $\theta(\theta)M$: This represents the distribution of topics in document M , depicting the range of topics that make up the document.
- $\phi(\phi)K$: This denotes the distribution of words within topic K , showing the probability of each word appearing in topic K .
- Z_{mn} : Indicates the topic allocation for specific words, particularly the topic assigned to the n th word in document m .
- W_{mn} : Represents the actual words in the corpus. Each W_{mn} is a unique word in the document, and combined with Z_{mn} , it shows the particular word associated with a specific topic in that document.

E. Hyper-Parameters

The following are the hyper-parameters used:

- Number of Topics (K). determines the level of detail and variety of themes extracted from the text corpus.
- Number of Words. Indicates the total number of distinct words accounted for in the analysis, defining the breadth of the vocabulary considered.
- Passes. Dictates how many times the algorithm reviews the entire corpus during its training phase. Each pass involves an update in the distributions of topics for documents and the distributions of words for topics.
- Iterations. Denotes the frequency of iterations within each training pass. This controls how often the model reassesses the entire dataset to refine its topic assignments.
- Alpha (α). Affects how topics are distributed across documents. A higher alpha value results in documents featuring a wider range of topics, whereas a lower alpha

leads to more concentrated topic distributions in each document.

- Eta (η). Impacts the density of word distributions within each topic. Higher eta values foster a more expansive distribution of words across topics, while lower values encourage a tighter, more focused word distribution for each topic.

TABLE 1. HYPER-PARAMETERS USED

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

Table 1 illustrates the range of hyperparameters examined to ascertain the optimal configurations for the LDA model experiments. These configurations are pivotal in producing coherent, interpretable, and contextually relevant topics derived from the dataset. Fine-tuning these hyperparameters is critical to align the topic modeling process with the dataset's unique attributes, significantly impacting the clarity and relevance of the extracted topics.

F. Identification of Appropriate Themes

The main objective in labeling the topic model was assigning pertinent labels to each topic, guided by the observed word similarities. This approach aimed to amplify topic clarity and strengthen effective communication, substantially enhancing their applicability within the UAQTE framework. Working in close partnership with domain experts from various disciplines, including CHED administrators, social scientists, data scientists, and recipients of the UAQTE program, greatly aided in choosing topic labels that were clearly understandable and closely aligned with the research objectives. Table 2 displays the labels established through this collaborative process, reflecting a well-rounded and research-focused understanding of the topics.

TABLE 2. DOMAIN EXPERTS' IDENTIFIED LABELS

Categories	Description
Financial Difficulties	Refers to the challenges faced due to limited financial resources, including difficulties in budgeting, paying for miscellaneous fees, boarding, daily expenses, and additional educational costs not covered by the program or scholarship.
Grant Disbursement	Refers to issues of scholarships or financial aid being delayed in disbursement, causing financial strain and difficulties meeting educational expenses.
Academic Difficulties	Refers to the challenges in fulfilling various academic obligations, such as preparing and submitting requirements, managing coursework, and striving to achieve satisfactory academic performance. Other factors like lack or poor facilities and infrastructures are likewise part of this category.

Pandemic-Related Challenges	Refers to the obstacles and difficulties arising from the COVID-19 pandemic, including the transition to online learning, lack of access to reliable internet connectivity and necessary devices, disruptions in academic schedules, and the impact on mental health, stress, anxiety, and feelings of isolation.
Program Implementation	Refers to the diverse viewpoints regarding the program's implementation, incorporating a range of positive and negative perspectives.

G. Model Evaluation

The Silhouette score is calculated as follows:

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

The silhouette score (s) of a data point is calculated by comparing its average distance (a) to points in the same cluster and (b) to points in the nearest neighboring cluster. A score close to +1 indicates strong cluster alignment and distinctiveness from neighboring clusters, suggesting clear and well-separated groupings. Scores near zero suggest the data point is on the boundary between clusters, indicating ambiguity in its assignment. Scores approaching -1 imply possible misplacement of the object in the wrong cluster, hinting at issues like cluster overlap or indistinct boundaries. Overall, silhouette scores provide insights into cluster separation and clarity, helping evaluate cluster quality and potential overlaps in the dataset.

The Coherence score is calculated as follows:

$$C_V(T) = 2 / (|T|(|T| - 1)) * \sum_{i=1}^{|T|} \sum_{j \neq i} \{sim(Word_i, Word_j)\} \quad (3)$$

The coherence score ($CV(T)$) for topic T is determined by considering the number of words in the topic, the representation of word pairs ($Word_i$ and $Word_j$) in the topic, and their similarity measure. This score assesses the connectedness and meaningfulness of words within a topic. A score near 0 indicates poor word connections, leading to difficult interpretation. Scores around 0.2 to 0.4 suggest limited coherence and interpretability. Scores between 0.4 to 0.6 denote reasonably coherent topics, improving interpretability. Scores from 0.6 to 0.8 reflect well-defined, highly coherent topics that are easy to interpret. Scores approaching 0.8 to 1 indicate exceptionally coherent topics with closely related words, offering high interpretability.

Domain experts are crucial in the evaluation phase, bringing essential contextual insights and specialized knowledge. Their input is vital given the multifaceted requirements of topic models, which include statistical integrity, semantic coherence, and relevance to the specific field. As key validators, they ensure that the generated models align with the research objectives.

Combining expert insights with silhouette and coherence scores establishes a robust evaluation method. Good silhouette and coherence scores, alongside expert validation of topic

interpretability, collectively indicate the high quality of the generated topics. Integrating quantitative metrics with the judgment of domain experts creates a well-rounded assessment framework. It guarantees a nuanced and dependable evaluation of the topic model's overall quality and effectiveness, ensuring that the models are not just statistically sound but rich in context and relevance to the UAQTE domain.

III. RESULTS AND DISCUSSION

This section presents key results and findings obtained through topic modeling of the UAQTE dataset implementing LDA. Table 3 highlights the configurations that achieved acceptable silhouette and coherence scores from the numerous experiments. Both scores aim to provide quantitative measures of the quality or usefulness of the output produced by unsupervised learning algorithms. Higher scores in both cases generally indicate better results, with higher silhouette scores implying well-defined clusters and higher coherence scores indicating more interpretable topics.

LDA model analysis investigated various configurations, including topics, words, passes, iterations, alpha, and eta, leading to variations in silhouette and coherence scores under different parameter settings, as indicated in Table 3. In particular, configurations with 8 topics, 50 passes, and 3000 iterations showcased higher silhouette scores, implying the model's effectiveness in creating more distinct data clusters. These settings reflect the LDA's strength in delineating clear groupings within the dataset.

TABLE 3. LDA HYPERPARAMETERS AND EVALUATION SCORES

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

Although the coherence scores ranged between 0.591 and 0.645, indicating some challenges in capturing complex semantic relationships due to the probabilistic nature of LDA, the model still showed a considerable degree of effectiveness in associating topics and words. This was instrumental in generating relevant labels. Despite the relatively modest coherence scores, LDA demonstrated its capability to formulate coherent and meaningful topics. This underscores its value in delivering interpretable and insightful dataset representations, highlighting its practical importance in data analysis.

Table 4 features the LDA model labeled by domain experts, with the highest Silhouette and Coherence Scores among all the LDA experiments. This model is characterized by its

hyperparameters: it includes 8 topics, the top 10 words, 50 passes, and 3000 iterations, along with an alpha value set at 0.01 and an eta value of 0.01. The model attained a Silhouette Score of 0.789 and a Coherence Score of 0.623 in this configuration.

TABLE 4. LDA LABELED MODEL

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

The extensive LDA experiments highlighted key themes such as "Academic Difficulties," "Financial Difficulties," "Pandemic-Related Challenges," "Grant Disbursement," and "Program Implementation." The analysis underscores the significance of these core topics, illustrating the model's ability to identify and prioritize crucial themes relevant to the dataset.

Integrating Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) methods with LDA remarkably enhances the model's interpretability. BoW contributed to improved topic identification and model stability, ensuring clear and consistent topic representation. On the other hand, TF-IDF refined topic descriptors and balanced word representation, offering nuanced and contextually relevant topics by focusing on unique words.

IV. CONCLUSIONS

Investigating various parameters and evaluation metrics in LDA topic modeling yielded valuable insights. Key performance metrics such as silhouette and coherence scores highlighted the importance of parameter optimization in LDA, particularly its ability to define distinct topic clusters. This underlines the necessity of tailoring modeling approaches to align with specific research objectives and the unique characteristics of datasets.

Critically, the involvement of domain experts in assigning labels to the generated LDA models has been instrumental in uncovering key insights for enhancing the UAQTE program. Their domain-specific knowledge guided the interpretation of themes such as "Academic Difficulties," "Financial Difficulties," "Grant Disbursement," "Pandemic-Related Challenges," and "Program Implementation." This has led to robust recommendations, including strengthening academic

support, enhancing financial assistance, flexible grant disbursement policies, addressing pandemic-related challenges, and establishing a structured feedback mechanism for program improvement.

These recommendations provide actionable directions for policy reforms within the UAQTE program, aiming to improve student support and the overall efficiency of the educational sector. It is crucial to note that these policy recommendations are part of an iterative process, requiring ongoing evaluation and flexibility to adapt to evolving circumstances, ensuring the long-term impact and relevance of the UAQTE program.

ACKNOWLEDGMENT

The researchers extend their 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 for the eParticipation 2.1 initiative: "Harnessing Natural Language Processing (NLP) for Community Participation." We deeply appreciate the financial support from this project, which was crucial in facilitating our research.

REFERENCES

- [1] E. G. García, E. C. Magaña, and A. C. Ariza, "Quality education as a sustainable development goal in the context of 2030 agenda: Bibliometric approach," *Sustainability* (Switzerland), vol. 12, no. 15, Aug. 2020, doi: 10.3390/SU12155884.
- [2] E. Boeren, "Understanding Sustainable Development Goal (SDG) 4 on 'quality education' from micro, meso and macro perspectives," *International Review of Education*, vol. 65, no. 2, pp. 277–294, Apr. 2019, doi: 10.1007/s11159-019-09772-7.
- [3] R. Nazar, I. S. Chaudhry, S. Ali, and M. Faheem, "Role of Quality Education for Sustainable Development Goals (SDGS)," *PEOPLE: International Journal of Social Sciences*, vol. 4, no. 2, pp. 486–501, 2018, doi: 10.20319/pijss.2018.42.486501.
- [4] O. Legusov, R. L. Raby, L. Mou, F. Gómez-Gajardo, and Y. Zhou, "How community colleges and other TVET institutions contribute to the united nations sustainable development goals," *J Furth High Educ*, vol. 46, no. 1, pp. 89–106, 2022, doi: 10.1080/0309877X.2021.1887463.
- [5] E. Lalith, A. P. S. K., S. Sampath, and R. Lakshmi, "Creating a psychological paradigm shift in students choice for tertiary education in Sri Lanka: The influence of socioeconomic factors," *International Journal of Educational Administration and Policy Studies*, vol. 14, no. 1, pp. 1–16, Jan. 2022, doi: 10.5897/ijeaps2021.0722.
- [6] C. Hursen, "The Effect of Problem-Based Learning Method Supported by Web 2.0 Tools on Academic Achievement and Critical Thinking Skills in Teacher Education," *Technology, Knowledge and Learning*, vol. 26, no. 3, pp. 515–533, Sep. 2021, doi: 10.1007/s10758-020-09458-2.
- [7] T. Kromydas, "Rethinking higher education and its relationship with social inequalities: Past knowledge, present state and future potential," *Palgrave Commun*, vol. 3, no. 1, Dec. 2017, doi: 10.1057/s41599-017-0001-8.
- [8] M. P. S. Mousavi et al., "Stress and Mental Health in Graduate School: How Student Empowerment Creates Lasting Change," *J Chem Educ*, vol. 95, no. 11, pp. 1939–1946, Nov. 2018, doi: 10.1021/acs.jchemed.8b00188.
- [9] B. A. Burt, B. D. Stone, R. Motshubi, and L. D. Baber, "STEM validation among underrepresented students: Leveraging insights from a STEM diversity program to broaden participation," *J Divers High Educ*, 2020, doi: 10.1037/dhe0000300.
- [10] Abdou Khadre Diop, Serban Meza, Mihaela Gordan, and Aurel Vlaicu, "LDA based classification of video surveillance sequences using motion information," 2018 20th International Conference on Advanced Communication Technology (ICACT), 2018.
- [11] M. Hasan, A. Rahman, M. Razaul, M. Saikat Islam Khan, M. Razaul Karim, and M. Jahidul Islam, "Normalized Approach to Find Optimal Number of Topics in Latent Dirichlet Allocation (LDA)," 2021. [Online]. Available: <https://www.researchgate.net/publication/344803887>
- [12] J. A. Lossio-Ventura, S. Gonzales, J. Morzan, H. Alatrística-Salas, T. Hernandez-Boussard, and J. Bian, "Evaluation of clustering and topic modeling methods over health-related tweets and emails," *Artif Intell Med*, vol. 117, Jul. 2021, doi: 10.1016/j.artmed.2021.102096.
- [13] A. Masoud, W. Cheruiyot, K. Ogada, C. Author, and A. Masoud, "Topic of Interest Discovery on Social Media Using Knowledge Base and Term Frequency-Inverse Document Frequency Techniques," 2018, 2018, doi: 10.9790/1813-0710030120.
- [14] A. Meddeb and L. Ben Romdhane, "Using Topic Modeling and Word Embedding for Topic Extraction in Twitter," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 790–799. doi: 10.1016/j.procs.2022.09.134.
- [15] S. M. Ozdemirci and M. Turan, "Case Study on well-known Topic Modeling Methods for Document Classification," in *Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021*, Institute of Electrical and Electronics Engineers Inc., Jan. 2021, pp. 1304–1309. doi: 10.1109/ICICT50816.2021.9358473.
- [16] M. Yousef and D. Voskergian, "TextNetTopics: Text Classification Based Word Grouping as Topics and Topics' Scoring," *Front Genet*, vol. 13, Jun. 2022, doi: 10.3389/fgene.2022.893378.
- [17] J. Wang and Y. Dong, "Measurement of text similarity: A survey," *Information (Switzerland)*, vol. 11, no. 9. MDPI AG, pp. 1–17, Sep. 01, 2020. doi: 10.3390/info11090421.
- [18] W. H. Park, I. F. Siddiqui, C. Chakraborty, N. M. F. Qureshi, and D. R. Shin, "Scarcity-aware spam detection technique for big data ecosystem," 2023.
- [19] X. Li, S. Wang, R. Malekian, S. Hao, and Z. Li, "Numerical Simulation of Rock Breakage Modes under Confining Pressures in Deep Mining: An Experimental Investigation," *IEEE Access*, vol. 4, pp. 5710–5720, 2016, doi: 10.1109/ACCESS.2016.2608384.
- [20] Institute of Electrical and Electronics Engineers, Morocco Section, and Jāmi'at al-Hasan al-Thānī. Ecole Normale Supérieure de l'Enseignement Technique de Mohammedia, ICOA 2018 Optimization: proceedings of the 2018 International Conference on Optimization and Applications (ICOA): April 26-27, 2018, Mohammedia, Morocco. 2018.
- [21] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int J Comput Appl*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.
- [22] R. Vangara et al., "Finding the Number of Latent Topics with Semantic Non-negative Matrix Factorization," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3106879.
- [23] M. V. Mantyla, M. Claes, and U. Farooq, "Measuring LDA topic stability from clusters of replicated runs," in *International Symposium on Empirical Software Engineering and Measurement*, IEEE Computer Society, Oct. 2018. doi: 10.1145/3239235.3267435.
- [24] H. M. Alash and G. A. Al-Sultany, "Improve topic modeling algorithms based on Twitter hashtags," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Nov. 2020. doi: 10.1088/1742-6596/1660/1/012100.
- [25] P. Tijare and P. J. Rani, "Exploring popular topic models," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Dec. 2020. doi: 10.1088/1742-6596/1706/1/012171.
- [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, Feb. 2023, doi: 10.1016/j.compedu.2022.104682.
- [27] E. M. Rinke, T. Dobbrick, C. Löb, C. Zirn, and H. Wessler, "Expert-Informed Topic Models for Document Set Discovery," *Commun Methods Meas*, vol. 16, no. 1, pp. 39–58, 2022, doi: 10.1080/19312458.2021.1920008.
- [28] Q. Fu, Y. Zhuang, J. Gu, Y. Zhu, and X. Guo, "Agreeing to Disagree: Choosing Among Eight Topic-Modeling Methods," *Big Data Research*, vol. 23, Feb. 2021, doi: 10.1016/j.bdr.2020.100173.
- [29] A. P. Pimpalkar and R. J. Retna Raj, "Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features," *ADCAIJ: Advances in Distributed Computing and*

- Artificial Intelligence Journal, vol. 9, no. 2, pp. 49–68, Jun. 2020, doi: 10.14201/adcaij2020924968.
- [30] P. Celard, A. S. Vieira, E. L. Iglesias, and L. Borrajo, “LDA filter: A Latent Dirichlet Allocation preprocess method for Weka,” PLoS One, vol. 15, no. 11, November, Nov. 2020, doi: 10.1371/journal.pone.0241701.
- [31] B. Ozyurt and M. A. Akcayol, “A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA,” Expert Syst Appl, vol. 168, Apr. 2021, doi: 10.1016/j.eswa.2020.114231.
- [32] C. Yang, J. Barth, D. Katumullage, and J. Cao, “Wine Review Descriptors as Quality Predictors: Evidence from Language Processing Techniques,” Journal of Wine Economics, vol. 17, no. 1, pp. 64–80, Feb. 2022, doi: 10.1017/jwe.2022.3.
- [33] G. Kontonatsios, S. Spencer, P. Matthew, and I. Korkontzelos, “Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews,” 2020, doi: 10.1016/j.eswa.2020.10.
- [34] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, S. M. Al-Ghuribi, and F. A. Ghanem, “Enhancing Big Social Media Data Quality for Use in Short-Text Topic Modeling,” IEEE Access, vol. 10, pp. 105328–105351, 2022, doi: 10.1109/ACCESS.2022.3211396.



Christian Y. Sy, is working on his DIT degree at the University of the Cordilleras, received his MIT degree in the same institution. An associate professor IV of Bicol University, his interests lie in the dynamic blend of machine learning, data analytics, and natural language processing (NLP). Published several machine learning and natural language processing papers in various scientific journals.



Dr. Lany L. Maceda, received her DIT and MIT degrees at the University of the Cordilleras. Her fields of research include Machine Learning, Data Analytics, and Soft Computing. Her professional affiliations include Computing Society of the Philippines-Special Interest Group on Natural Language Processing (NLP), Special Interest Group in Women in Computing (WIC), and Commission on Higher Education (CHED) Regional Quality Assessment Team (RQAT).



Dr. Thelma D. Palaoag, Director of the Innovation and Technology Transfer Office of the University of the Cordilleras, she is a visionary leader, accomplished researcher, and trailblazer in the field of Information Technology (IT). With an unwavering commitment to advancing technological frontiers, she has dedicated her career to pushing the boundaries of IT research and fostering innovation within academic institutions.



Dr. Mideth B. Abisado, received her DIT degree at the Technological Institute of the Philippines, MCS at the Mapua Institute of the Philippines. Her fields of research include Artificial Intelligence, Adaptive Learning, E-Learning, Emphatic Computing, and Machine Learning. Her professional affiliations include Computing Society of the Philippines, National Research Council of the Philippines, and Association for Computing Machinery.