Contextual Understanding of Academic-Related Responses Based on Enhanced Word Embeddings, Clustering, and Community Detection

Mary Joy P. Canon [1,*], Lany L. Maceda [1], and Nancy M. Flores [2]

[1] Computer Science and Information Technology Department, Bicol University, Legazpi City, Philippines
[2] College of Information Technology and Computer Science, University of the Cordilleras, Baguio City, Philippines
Email: mjpcanon@bicol-u.edu.ph (M.J.P.C.); llmaceda@bicol-u.edu.ph (L.L.M.); nancy@uc-bcf.edu.ph (N.M.F.)
*Corresponding author

***Abstract*—In strengthening educational policy, it is crucial to understand recipients' feedback and experiences. This paper presents an innovative method for generating contextual analysis of qualitative responses from the beneficiaries of a free education program. We employed a two-tier clustering approach using the K-means algorithm and Louvain community detection, based on enhanced word embeddings and a bi-gram network. Our methodology extends beyond traditional text analysis by combining Word2vec and Glove embeddings, which captured the semantic meaning of words within the dataset. Through the application of the K-means algorithm, we identified five distinct clusters with a notable silhouette score of 0.3477, corresponding to themes of "Support and Educational Opportunity", "Accessibility and Financial Relief", "Gratitude and Satisfaction", "Positive Evaluation with Suggestions for Improvement", and "Program Effectiveness". Further refinement of the clusters with the Louvain method, coupled with the use of a bi-gram text network instead of a uni-gram, achieved a higher modularity score of 0.637. While the transition from K-means clusters to Louvain communities resulted in slight thematic changes, it provided a more comprehensive view of the relationships among different response aspects. The two-tier clustering method highlights the methodology's strengths and effectiveness in revealing hidden patterns and themes in the text responses. The findings not only highlight the strengths of the free education program in providing support to the beneficiaries but also reveal certain areas needing attention and improvement, which are crucial in policy development and enhancement.**

***Keywords*—enhanced word embedding, clustering, community detection, text analysis, quality tertiary education, program evaluation**

I. Introduction

Higher education holds great importance and has a significant impact on individuals and society in general. Its role is crucial in alleviating youth poverty as well as uplifting the economic progress of a country [1]. Embedded as one of the global objectives established by the United Nations (UN) under the 17 Sustainable Development Goals (SDG) is SDG 4 for quality education aiming to guarantee inclusive and equitable quality education and promote opportunities for lifelong learning for all [2, 3]. Target 4.3 is focused on the development of higher education, which aims to guarantee that all individuals, regardless of gender, have equitable access to affordable and high-quality education at technical, vocational, and tertiary levels, including university studies [4].

In Philippines, access to quality education is an inalienable right of all Filipinos. As stated in the 1987 Philippine Constitution, the State shall protect and promote the right of all citizens to quality education at all levels [5]. In recognition of this, the Philippine government has been implementing programs to continuously improve the education system in the country. In 2017, the country enacted into law the Republic Act No. 10931 also known as the Universal Access to Quality Tertiary Education Act (UAQTE) which Mandates State Universities and Colleges (SUCs), Local Universities and Colleges (LUCs) and state-run Technical-Vocational Institutions (TVIs) to provide quality tertiary education to eligible Filipino students. It has four main components: (a) Free Higher Education (FHE) program which provides free tuition and other school fees in public Higher Education Institutions (HEIs); (b) free tuition in Technical Education and Skills Development Authority (TESDA) technical-vocational training institutes; (c) Tertiary Education Subsidy (TES); and (d) student loan programs (R.A. 10931). Commission on Higher Education (CHED) and Unified Student Financial Assistance System for Tertiary Education (UniFAST) are the primary implementing agencies and units of the project [6].

While the UAQTE has been acclaimed for its commitment to the delivery of the policy benefits to the intended recipients, it also faced challenges and criticisms since its introduction. Some argue that the program may

Manuscript received January 11, 2024; revised February 14, 2024; accepted February 26, 2024; published June 5, 2024.

have fallen short in effectively identifying those who truly need assistance. Disadvantages experienced by students from socioeconomic classes in primary and secondary education tend to remain barriers to their entrance to tertiary education [7], notwithstanding the provision of tuition subsidies.

Despite years of implementation of the UAQTE program, there remains a lack of comprehensive understanding regarding the recipients' perceived impact and overall feedback on the program. Given the complexity and scope of the program, it is necessary to conduct a comprehensive evaluation to improve the chances of achieving its objectives. The discussion paper [7] stressed the need to evaluate the implementation of R.A. 10932 even in its early stage of enactment. By conducting thorough assessments, the program can be strengthened to ensure its long-term sustainability and maximize its positive outcomes on the Philippine tertiary education system.

Public participation postulates an open, democratic form of planning and policy-making. It does not only engage the public in decision-making which results in better governance, but also one contributing factor to sustainable development [8]. In evaluating the UAQTE program, stakeholders' participation offers crucial feedback on the effectiveness and relevance of the program initiatives. By capturing the feedback and responses of the program beneficiaries, implementers can identify strengths, weaknesses, and areas for improvement to better serve the target population.

The challenge in analyzing and interpreting text responses, such as feedback from the UAQTE beneficiaries, lies in distilling meaningful patterns and themes present in the corpus. This feedback encompasses a wide range of experiences and perceptions that are critical in evaluating the program's effectiveness and impact. From the perspective of qualitative data analysis, natural language processing offers powerful approaches to analyzing, modeling, and processing text data. Incorporating this technology to analyze the recipients' responses related to the implementation of the UAQTE, offers a significant contribution to drawing various insights for understanding program outcomes and potential policy enhancement.

Seeing both challenges and opportunities related to the implementation of the UAQTE program, this paper intends to employ advanced text analysis techniques to automatically discover themes from the text responses, encompassing narratives on experiences, perceived impact, and overall feedback of UAQTE beneficiaries. The primary goal is to generate a contextual understanding of academic-related responses, by employing enhanced word embeddings, clustering techniques, word networks, and community detection.

II. Literature Review

A. Enhanced Text Embeddings

Word embeddings are the numerical representation of texts suitable as input features for natural language processing tasks. The introduction of word embeddings in a neural probabilistic language model by Bengio *et al.* [9] laid the groundwork for representing words in continuous vector spaces and capturing semantic relationships. Its significant advancement was marked by the evolution of word embeddings, particularly through Word2Vec [10] and Glove [11]. These models efficiently capture word associations and context, moving beyond mere word frequencies to understanding the subtleties of language.

Beyond the traditional application of these embeddings, several efforts and methodologies have been introduced to improve word vectors. For instance, Bojanowski *et al.* [12] further enhanced word embeddings by introducing subword information. This approach allows for a deeper understanding of word morphology and improves the handling of out-of-vocabulary words, which is particularly crucial in analyzing diverse academic texts where technical or specialized terms are common. In another work [13], a method that combines domain-based ontologies with word embeddings which enhances key phrase extraction from geological documents was proposed. This approach suggests that embedding models can be tailored to specific domains for more accurate results. Qi *et al.* [14] and Biswas and De [15] explored the effectiveness of pre-trained word embeddings in neural machine translation and the improvement of embedding models, respectively. These studies show that continuous enhancement and adaptation of embedding models are crucial for various Natural Language Processing (NLP) tasks.

Focusing on sentiment analysis tasks, Yu *et al.* [16] developed a method to refine pre-trained word embeddings using sentiment intensity scores, which aligns word vectors with semantically and sentimentally similar words. This model offers a nuanced understanding of sentiment compared to binary labels and improves sentiment classification performance. Complementing this, a refined model [17] enhances pre-trained vectors such as Word2Vec and GloVe for more accurate sentiment analysis in both binary and fine-grained classifications. Additionally, Rezaeinia *et al.* [18, 19] introduced Improved Word Vectors (IWV), a novel method that showcases the effectiveness of task-specific fine-tuning in sentiment analysis.

B. Clustering Technique in Analyzing Academic Data

Clustering analysis is used to find the useful and unidentified classes of patterns in a dataset. It involves organizing and partitioning a collection of data in such a manner that items within the same cluster exhibit greater similarity to each other than those in other clusters.

The diverse applications of clustering techniques in academic data analysis are well-documented in recent research. Zhang *et al.* [20] innovatively employed K-means clustering to dissect international education trends, with a particular focus on the ramifications of the COVID-19 pandemic. Zhao and Wang [21] applied clustering to categorize news texts related to international Chinese education, achieving a significant accuracy rate. In a different context, a model [22] that employs text clustering to effectively track and analyze public opinion trends in university networks was crafted, providing

insights into the dynamic nature of online educational discourse. Tao *et al*. [23] utilized cluster analysis to gain a deeper understanding of English language learning conceptions among Chinese university students. Meanwhile, the clustering technique was applied in literary analysis, specifically in the context of Shakespeare's stories, showcasing the method's versatility [24]. Additionally, prevailing themes and trends were identified on the application of gamification in education through bibliometric and text mining analysis [25]. Finally, Han and Lee [26] demonstrated the use of hierarchical cluster analysis to categorize and understand online learning types in software education, categorizing learners based on self-regulated learning characteristics. Collectively, these studies underscore the adaptability and effectiveness of clustering techniques in extracting meaningful patterns and insights across a variety of educational contexts and data types.

C. Community Detection Technique in Educational Research

The idea of community detection has gained prominence in network science as a technique for uncovering groups within complex systems through a graph representation [27]. In machine learning, community detection can be used to find groups with similar attributes and extract groups for a variety of reasons [28].

The application of community detection in educational research received significant attention in recent years. Kadry *et al*. [29] applied this technique to analyze user engagement in online learning networks, specifically within Khan Academy's extensive repository. An enhanced collaborative problem-based learning was proposed by Chen and You [30], combining community detection with opinion leader identification. Accordingly, insights into student preferences and patterns were discovered by applying community detection in student course selection in University admission [31]. In the context of distance education, Yassine *et al.* [32] provided a comprehensive review of community detection applications in online learning environments. Complementing this, Evgenia *et al*. [33] employed Social Network Analysis for community detection in Distance Education forums, highlighting the dynamics of online educational interactions. In understanding digital engagement, Saputri *et al.* [34] analyzed browsing behavior in educational institutions using Wi-Fi log data, demonstrating the application of community detection. Lastly, Shao [35] developed a novel approach for detecting student communities and recommending personalized learning paths. These studies collectively showcase the versatility and impact of community detection techniques in various educational contexts.

III. Data and Methods

This section describes the dataset used in the study, as well as the two-tier approach to discovering themes present in the text corpus. First, we describe the context of the dataset and its sources. Then, we discuss the cleaning and pre-processing steps undertaken on the collected responses. Subsequently, we describe the generation of enhanced word embeddings. Lastly, we explain how the clusters were generated through the K-means algorithm and refined using word network graphs and community detection.

A. Academic-Related Text Responses

To properly assess the implementation of the UAQTE program through an unsupervised approach, qualitative data that captures the feedback, experiences, and felt impact of the beneficiaries is necessary. In this paper, we made use of the dataset collected using the BosesKo application, a citizen's participation toolkit. Through this platform, the beneficiaries of the free tuition and education subsidy participated in a survey designed to capture the necessary information for assessing and evaluating the UAQTE program's impact. Specifically, responses to the following key questions were utilized in the experiments:

1) In 3–4 sentences, write your experience as one of the beneficiaries of the UAQTE.
2) Write a short description of the impact of the UAQTE on you and your family.
3) What is your overall feedback on the implementation of the UAQTE?

Participants of the survey are recent graduates and college students from public and private higher education institutions across the Philippines who availed of or are currently availing of the Free Higher Education (FHE) and Tertiary Education Subsidy (TES) components of the UAQTE program. The survey was completed by 3,150 beneficiaries between December 15, 2022 and December 8, 2023. A total of 8,536 responses to the three questions were used for text processing.

B. Text Pre-processing

An essential step to clean and transform unstructured text data in preparation for data analysis is the combination of various text preprocessing techniques. To enhance the data quality and relevance of our dataset, we employed different libraries of the Natural Language Toolkit (NLTK) in transforming the corpus. Most of the samples we collected are composed of single sentences. To standardize the sample length and to augment the sample size, responses underwent sentence tokenization, which resulted in 17,516 sentences. Initially, rows containing non-informative markers such as "N/A", "none", or words shorter than three letters were discarded. The texts were then standardized for uniformity and clarity by performing lemmatization, expanding contractions, and lowercasing. Regular expressions made it easier to remove digits, short words, and special characters. Furthermore, the process included filtering out stop words and non-English terms. The exclusion of non-English terms was crucial, as they often cluster together during modeling, potentially skewing the dataset's semantic meaning.

C. Enhanced Word Embeddings

To enrich our feature set, we utilized the combination of Word2vec and Glove methods and then reduced the dimensionality of the generated embeddings. Word2Vec [10] is a predictive embedding model trained to either predict a word given its context (Skip-gram

model) or to predict the context given a word (Continuous Bag of Words, CBOW model). This model generates embeddings such that words that occur in similar contexts are close to each other in the embedding space. Mathematically, the Skip-gram model aims to maximize the objective function in Eq. (1).

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c \le j \le c, j \ne 0} log\, P(w_{t+j}|w_t;\theta) \quad (1)$$

where *wt* is the input word, *wt+j* are the context words within a window of size *c*, and θ are the parameters of the model. The probability *p*(*wt+j*|*wt*) is defined using the softmax function in Eq. (2).

$$p(w_o|w_I) = \frac{\exp(v'_{w0}{}^T v_{w_I})}{\sum_{w=1}^{W}\exp(v'_{w0}{}^T v_{w_I})} \quad (2)$$

where v_w and v'_w are the input and output vector representations of word *w*, and *W* is the number of words in the vocabulary.

On the other hand, GloVe or Global Vectors for Word Representation is a count-based model, developed by Stanford researchers, that utilizes a global factorization method, a word-word co-occurrence matrix from a corpus [11]. The model effectively captures both global statistics and local context. Its objective is to reduce the difference between the logarithm of the chance of two words occurring together and the dot product of their embeddings. The GloVe objective function is presented in Eq. (3).

$$J = \sum_{i,j=1}^{V} f(X_{ij}) x (w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log X_{ij})^2 \quad (3)$$

where *Xij* is the number of times word *i* occurs in the context of word *j*, *f* is a weighting function that prevents learning only from extremely common word pairs, *w* and $\tilde{w}$ are the word vectors, and *b*, $\tilde{b}$ are scalar biases for each word.

For a compatible and consolidated semantic space, both embedding models generated using word2Vec and GloVe employed the same 300-dimensional space. The Word2Vec embeddings are obtained from a model pre-trained on the Google News dataset, which contains roughly 100 billion words, whereas the Glove embeddings are trained on a corpus aggregated from Wikipedia and Gigaword5 dataset, a collection of newswire text data.

After generating embedding models using Word2Vec and Glove methods, we merged the features by calculating the average of the vector representations for corresponding words. Refer to the formula in Eq. (4).

$$V_{combined}(w) = \frac{V_{w2v}(w) + V_{glove}(w)}{2} \quad (4)$$

where $V_{w2v}(w)$is the Word2Vec embedding for word *w*, $V_{glove}(w)$ is the GloVe embedding for word *w*, and $V_{combined}(w)$ is the resulting combined embedding vector.

This averaging process is expected to yield a more standardized representation by taking advantage of the strengths of both models. If a word is not found in either of the models, a zero vector of the specified size. However, combining embeddings using different models can increase the dimensionality of the vectors. This situation affects computational efficiency and increases data sparsity and noise within the vector space. To address this, we reduced the dimensionality of the embeddings using Principal Component Analysis (PCA) [36]. This is a statistical method that uses eigenvectors to capture the directions of maximum variance in the data, and eigenvalues to determine their magnitude. In other words, applying PCA to reduce dimensions can help distill the most relevant linguistic or semantic features from these combined embeddings, which is crucial for downstream tasks like clustering.

D. K-means Clustering and Theme Identification

Enhanced feature sets were fed to the K-means clustering algorithm. K-means is a centroid-based clustering algorithm that divides data into K distinct groups, each represented by the mean of its points [37]. Given its historical performance and ability to group several data sets in a fast and efficient computing period, K-means remains the best grouping algorithm available [38]. In the present study, the algorithm was used to group similar embeddings into the same cluster. K-means objective function in Eq. (5) quantifies the goal of the clustering process.

$$J = \sum_{j=1}^{K} \sum_{x_i \in S_j} ||x_i - \mu_j||^2 \quad (5)$$

K represents the total number of clusters. $\sum_{j=1}^{K}$ is the summation over all data points. S_j represents the set of all data points that are assigned to cluster *j*. Lastly, $||x_i - \mu_j||^2$ is the squared Euclidean distance between a data point x_i and the centroid μ_j of its cluster.

In exploring the performance of this algorithm, we used different k values, ranging from 2 to 9. We computed the silhouette score for each cluster configuration to determine an object's similarity to its cluster, or its cohesion versus its difference from other clusters, or its separation. This method is a way of tuning and validating the clustering models to find the best fit for the data.

The final generated clusters were labeled by four domain experts. Two of them are technical specialists from higher education and two are social scientists. They examined the instances for each cluster and identified the corresponding themes.

E. Word Network Graphs

Python's NetworkX and Matplotlib libraries, along with Scikit-learn for Term Frequency-Inverse Document Frequency (TF-IDF) vectorization were implemented to produce text network graphs. These graphs visually represent the relationship between words or bi-grams present in the text clusters. Constructing these networks involves the calculation of Term Frequency-Inverse Document Frequency (TF-IDF) scores for words or bi-grams. The TF-IDF score for a term is calculated using the formula in Eq. (6).

$$\text{TF} - \text{IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t) \quad (6)$$

Term Frequency TF operates on the principle that the frequent occurrence of a term *t* in a document *d* is indicative of *its* significance for *d*. Accordingly, Inverse Document Frequency (IDF) gauges the rarity of term *t* across the entire corpus [39], assigning higher values to less common terms. Put simply, it evaluates the significance of a term in a document compared to its relevance across the entire collection of documents.

Based on the TF-IDF scores, we identified the top words and top bi-grams. For each cluster, nodes representing the top terms are added to the graph. Each node's size is determined by the corresponding term's TF-IDF score, making more significant terms visually prominent. Edges are added between nodes within the same cluster to represent the association between terms. The edge creation effectively forms a subgraph for each cluster, where the nodes are the terms and the edges signify their co-occurrence or semantic closeness.

F. Community Detection

For further refinement, we applied the Louvain algorithm [40] for community detection on the word and bi-gram network graphs. This is a popular method in network science for extracting non-overlapping communities within a large network. It iteratively aggregates nodes into communities, trying to maximize the modularity score of the network. Each node is assigned to a community or cluster, to produce densely connected subgraphs within the larger network.

IV. Results and Discussion

This section presents the results of the conducted experiments and discusses some implications derived from these results.

A. Enhanced Word Embeddings

Based on the Word2Vec and Glove vocabularies, we identified 3,460 unique text embeddings present in our corpus. This number indicates a substantial lexical variety of the terms within the dataset. Table I presents the magnitude of the combined embeddings from Word2vec and Glove. The scores from each model reflect the magnitude of the respective vectors in the embedding space. The quantified measure of the embeddings from the two models is represented by the final scores. These numbers reflect the semantic richness and contextual relevance of words, appropriate as feature sets in the clustering process.

TABLE I. Sample Tokens and Their Corresponding Scores

Token	Word2, Vec score	Glove score	Final Score
assistance	2.8296824	6.2453046	3.544196
acknowledge	2.5976148	5.195564	2.9032726
motivate	2.8047934	5.780639	3.2699735
opportunity	2.695808	5.660927	3.13566
education	2.5700479	6.6734014	3.6078951
stipend	3.7024503	6.570337	3.747017
inequality	3.6119199	6.97027	4.0364656
grateful	3.6119199	6.97027	4.0364656
burden	3.486293	6.3713894	3.6992564
free	2.4175155	6.4717374	3.5070798

B. Generated Clusters Using K-means Algorithm

After testing various cluster configurations, we selected five as the *k* value to define the initial groupings of the text responses. This number of clusters obtained a silhouette score of 0.3477, slightly lower than $k = 3$ with 0.3512. This score indicates that partitioning with five clusters still maintains a relatively good level of separation and definition among clusters. On average, this number suggests that the clusters are reasonably well-defined. Fig. 1 illustrates the clustering of text responses into five groups using a t-distributed Stochastic Neighbor Embedding (t-SNE) plot. There is still some degree of overlap, but the data points in each cluster are closer to each other than they are to the data points in other clusters. Other considerations in choosing 5 as the *k* value, are the domain knowledge, and variety of responses representing aspects of feedback or experiences. These reasons warrant a finer division of clusters to capture the nuances of data, generating groups that are indeed meaningful to be analyzed separately.

Fig. 2 is the word clouds for each cluster. They display the key terms that are most representative of the sentiments and themes within each group. The interrelation of the themes is evident in the word clouds which means that some themes are not mutually exclusive. This can be attributed to the instances in the corpus sharing common features. For example, words like "helpful", "education", and "grateful" appear across multiple clusters, implying a shared aspect of the UAQTE program being discussed, and that the program generally fosters a sense of gratitude and satisfaction among beneficiaries. This type of overlap is generally acceptable in the clustering method [41] because real-world datasets have inherently overlapping clusters [42].

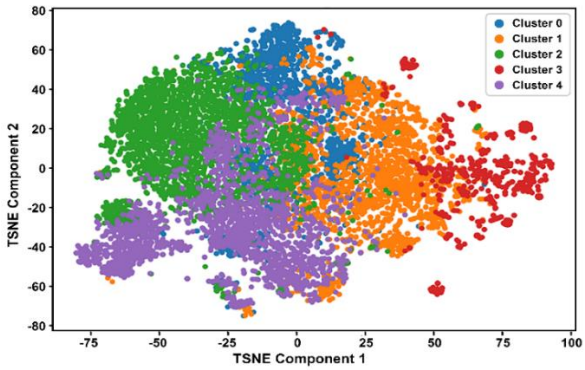

Fig. 1. t-SNE plot of the generated clusters using the K-means algorithm.

C. Identified Themes by Domain Experts

Table II lists the identified themes and sample responses related to the UAQTE program for each cluster. The domain experts labeled Cluster 0 with "Support and Educational Opportunity", which responses refer to financial assistance provided to scholars, expressions of being able to pursue higher education, and descriptions of enhanced academic focus. Cluster 1 is identified as "Accessibility and Financial Relief" which responses emphasize the impact of financial support on educational

aspirations, easing the financial burden of the family, and promoting equal opportunities for all. Both Cluster 1 and Cluster 2 convey the role of financial aid not just in supporting students through their education, but also in contributing to overall improvement of well-being. This implies a transformative effect of the UAQTE program on the beneficiaries and their families.

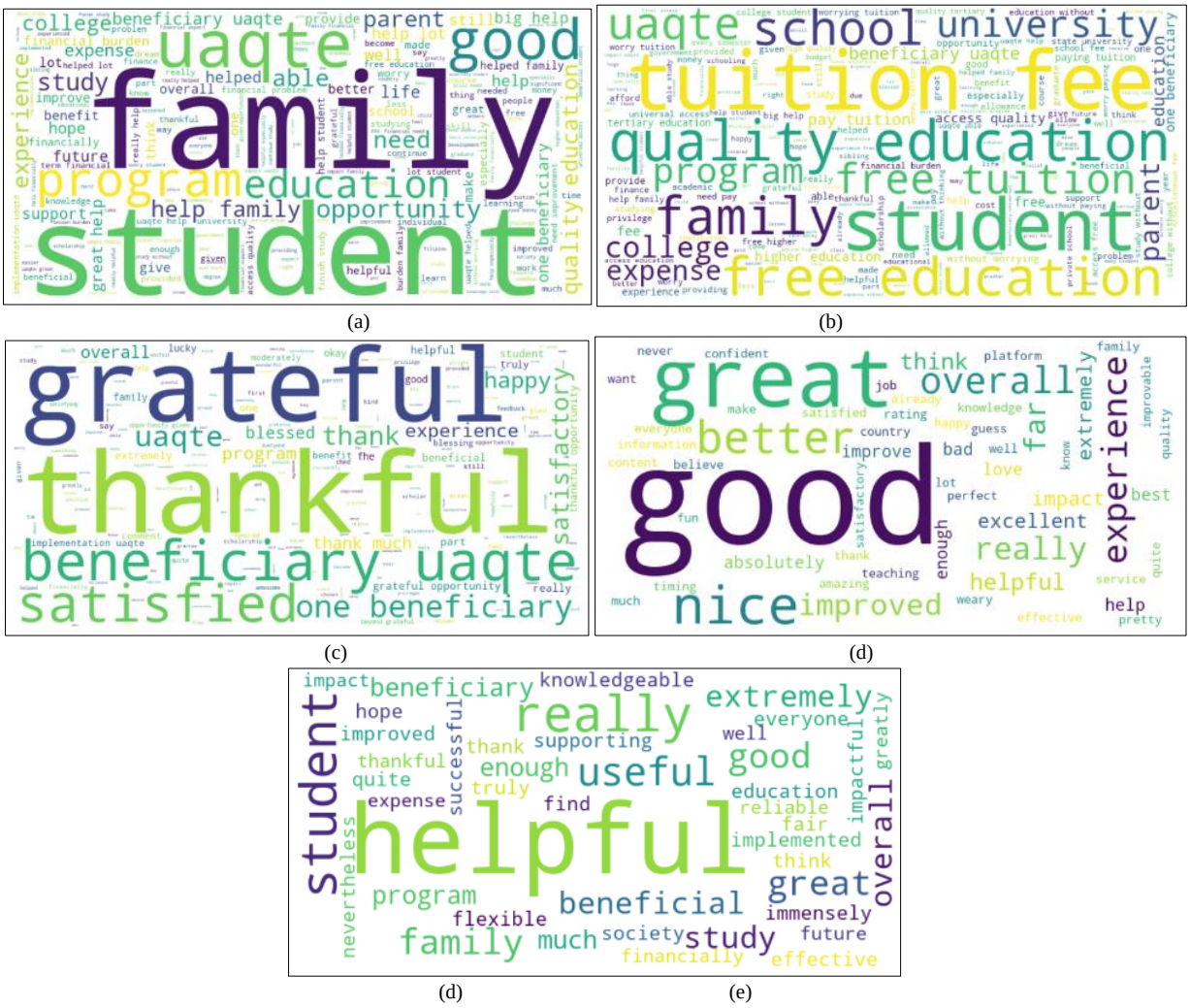

Fig. 2. Generated word clouds for each cluster, (a). Most important words in Cluster 0, (b) Most important words in Cluster 1, (c) Most important words in Cluster 2, (d) Most important words in Cluster 3, (e) Most important words in Cluster.

TABLE II. IDENTIFIED THEMES AND SAMPLE INSTANCES PER CLUSTER

Cluster	Identified Theme	Sample Instances
0	Support and Educational Opportunity	(1) helped lessen costs parents would shoulder (2) lifted huge burden shoulder allowing focus on study without worrying much finances (3) beneficiaries able to access high-quality education uaqte provided knowledge skills needed to pursue a dream (4) truly give hope young individual pursue their desire course despite coming poor family (6) moreover allowed focus academic help family financially (7) program helped overall productivity learning process lesser financial concern (8) inclusive educational opportunity acquired knowledge skill needed a secure fulfilling career
1	Accessibility and Financial Relief	(1) parents not worry much allowance going to school, (2) removing financial obstacles ensures fellow students regardless of socioeconomic status access to quality education, (3) one member low-income middle-class family free education beneficial wanted study graduate time to support the family financially, (4) lower burden family free education, (5) thankful help lighten family expense attending state college availing free tuition, (8) parent focus time providing money living academic related expense
2	Gratitude and Satisfaction	(1) thankful blessed one beneficiary unique, (2) unique truly beneficial, (3) glad one beneficiary unique, (4) beneficiary unique big blessing, (5) always thankful opportunity given ched, (6) thankful experience kind field (7) satisfied happy, (8) lucky part beneficiary rate (9) lucky part beneficiary rate
3	Positive Evaluation with suggestions for improvement	1)overall good, (2) good experience, (3) good, (4) quality teaching good, (5) good but improved, (6) good helps a lot, (7) excellent rating, (8) good but could better
4	Program Effectiveness	(1) helpful study, (2) really effective helpful, (3) helpful reliable, (4) helpful future, (5) immensely helpful, (6) good helpful, (7) helpful but not enough (8) helpful but hope flexible

Moreover, "Gratitude and satisfaction" is the assigned theme for Cluster 2. Responses in this cluster reflect a clear sense of gratitude and satisfaction, suggesting a positive impact on their lives. Cluster 3 is tagged as "Positive evaluation with suggestions for improvement", which instances seem to be associated with recipients' general positive feedback on their experiences as scholars and the services they have received with a desire for program improvement. Lastly, Cluster 4 is labeled as "Program Effectiveness", which instances focus on the implementation aspect. In this cluster, the program is described as helpful and effective, but with a slight implication of some needs not being fully met, which suggests a need for assessing the program's overall effectiveness.

The implication of the results offers valuable insights into UAQTE's impact and areas for potential policy improvement. The program delivered varied impacts, providing not just financial support, but also educational opportunities, both contributing to the overall well-being of the recipients. Despite the positive outcomes, analysis reveals areas for program enhancement, targeted interventions, or additional services.

D. Communities Detected from Clusters using Top Bi-Grams Text Network

Text network graphs for top words and top bi-grams were generated. Alongside network generation, we computed the modularity score of each graph to determine the strength of the network's division. The bi-gram text network achieved a higher modularity score of 0.637, indicating stronger community structures, compared to the text network using top words, which only scored 0.298. In network science, modularity scores closer to 1 imply a strong community structure [43]. In our experiments, the significantly higher modularity score of the bi-gram network signifies a well-structured network with distinct communities. Conversely, the lower score for the uni-gram network means that single words are more broadly used across different contexts and are less indicative of distinct communities on their own. This finding suggests that bi-grams, due to their ability to capture more contextual information, are more effective in revealing clear and meaningful patterns than single words.

Based on the computed TF-IDF scores for the bi-grams, we determined the top sequences for each cluster as shown in Table III. These top sequences were used for visualization of the text network graph depicted in Fig. 3. In this graph, the nodes correspond to the bi-grams, while the connecting edges map and indicate the relationships between the text pairs. There exists a commonality of bi-grams across clusters, revealing a thematic consistency within the data. Some of the identified common bi-grams are: "beneficiary unique", "implementation date", "help the family", "quality education" and "good helpful". For instance, the first common bi-gram implies that the beneficiaries of the UAQTE program are the central point of discussion among clusters. "Help family" bi-gram points to the program's social impact beyond individual recipients.

TABLE III. TOP BI-GRAMS FOR THE K-MEANS CLUSTERS AND LOUVAIN COMMUNITIES

Group	Top Bi-grams (K-Means)	Assigned Theme	Top Bi-grams (Community Detection)	Assigned Theme
0	access quality, beneficiary unique, burden family, family financial, financial burden, finish study, focus study, free education, help family, impact family, impact rate, implementation date, lessen financial, quality education	Support and Educational Opportunity	help a lot, great help, help student, big help, financial burden, helped family, helped a lot, financial problem, unique help, finish study	Support and Program Effectiveness
1	access education, access free, access quality, beneficiary unique, big help, college student, financial burden, free education, free higher, free tuition, help family, help student, pay tuition, quality education, school fee, state university, tertiary education, tuition fee, universal access, worry tuition	Accessibility and Financial Relief	beneficiary unique, implementation date, grateful beneficiary, thankful beneficiary, grateful opportunity, thankful opportunity, unique grateful, unique thankful, happy beneficiary, an opportunity given, overall satisfied, extremely satisfactory, thankful program	Gratitude and Satisfaction
2	beneficiary program, beneficiary unique, blessed beneficiary, extremely satisfactory, extremely satisfied, glad beneficiary, grateful opportunity, happy beneficiary, implementation date, overall satisfied, overall thankful, really thankful, satisfied beneficiary, satisfied implementation, unique blessing,	Gratitude and Satisfaction	overall good, good experience, good impact, experience good, well improved, good helpful, far good, really good, well better, extremely good, good far, really nice, well good, nicely improved	Positive Evaluation
3	good, experience good, extremely good, good experience, good helpful, good impact, good improved, good really, great weary, help lot, helpful better, helpful improved, know improve, make better, nice improved, overall good, quality teaching, really good, really great, really service good	Positive Evaluation with suggestions for improvement	quality education, tuition fee, help family, free education, free tuition, pay tuition, access quality, higher education, paying tuition, tertiary education, worry tuition, universal access, worrying tuition, free higher	Accessibility and Financial Relief
4	beneficial helpful, beneficiary helpful, extremely helpful, good helpful, great helpful, helpful expense, helpful family, helpful future, helpful impactful, helpful improved, helpful knowledgeable, helpful program, helpful society, helpful student, helpful study, helpful supporting, helpful thank, impact helpful, overall helpful, really effective, really helpful, successful helpful	Program Effectiveness	helpful student, really helpful, helpful family, helpful program, extremely helpful, great helpful, helpful study, overall helpful, beneficiary helpful, beneficial helpful, helpful thankful, helpful thank, helpful supporting, helpful society	Program Effectiveness

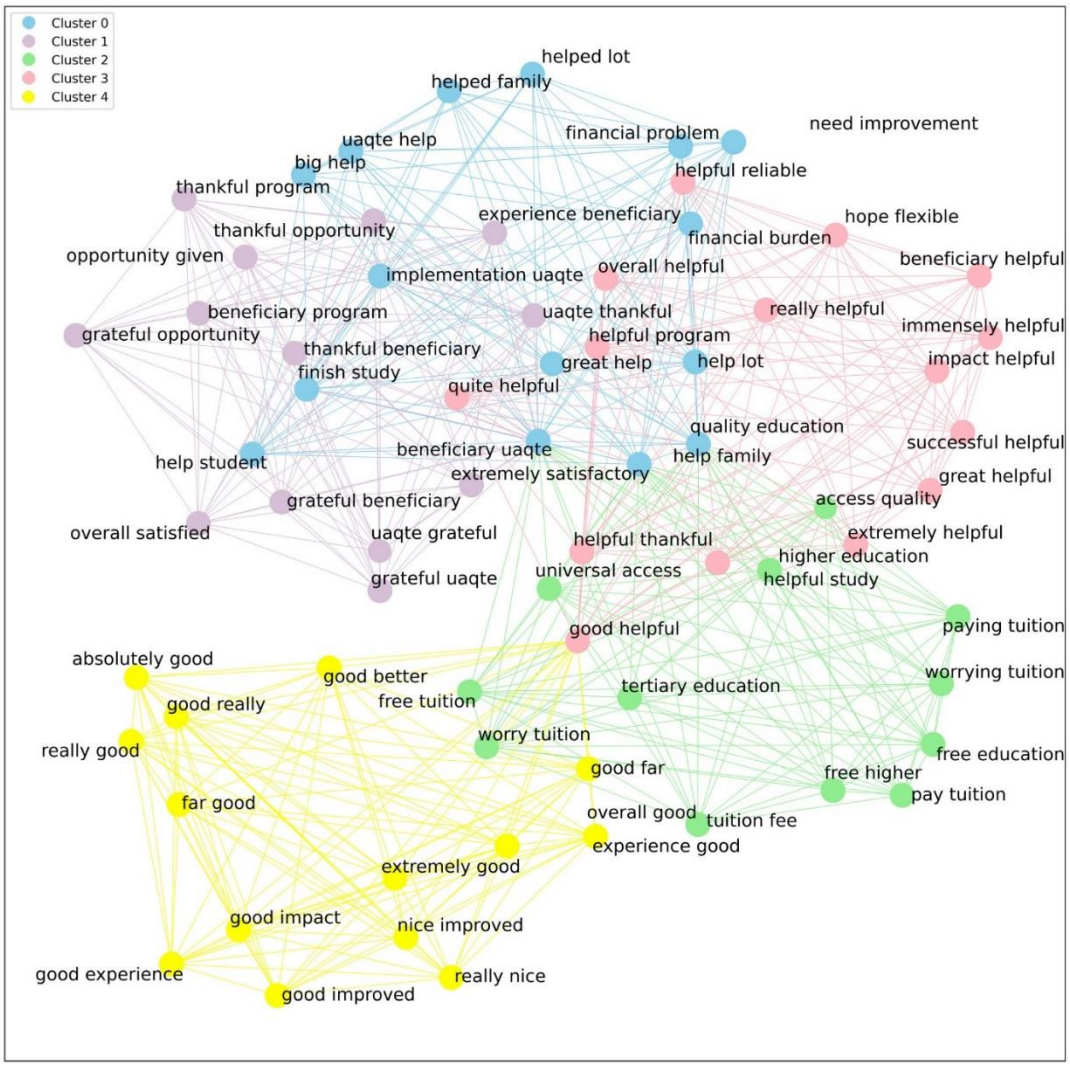

Fig. 3. Text network of top Bi-grams per cluster.

Furthermore, these combinations of words semantically strengthen the identified themes in the previous section. They highlight the aspects of positive feedback, educational opportunity, financial support, relief of burden, and, areas for enhancement in the implementation of the UAQTE program. Positive bi-grams suggest that the program is well-implemented overall, whereas bi-grams expressing challenges and shortcomings offer an opportunity for policy improvement and better service.

Through text network graphs, the generated clusters are further refined using the community detection method, particularly using the Louvain algorithm. The resulting groups are presented in Fig. 4. This graph revealed communities within the groups of bi-grams that are more frequently discussed together. The size of the nodes correlates with the importance of the bi-gram in the dataset. For instance, "free education" appears as a large node, an implication that it is a significant adjacent pair within the responses. The edges represent the connections between bi-grams. If two bi-grams share a common word or are frequently found together, they are connected by an edge. Another notable observation in the graph is the presence of bi-gram nodes that appear to bridge clusters or are centrally located between clusters. Examples of these nodes are "good helpful", "implementation date" and "beneficiary unique". The bi-gram "good helpful" is very apparent between Cluster 1 and Cluster 3. It is a common thread across responses expressing general satisfaction with the program implementation and positive sentiments toward the financial support received. The bi-gram "beneficiary unique" is not just one of the most important bi-grams, but also serves as a point of connection between different themes, especially Clusters 2, 4, and 5. This relationship shows that responses about program outcomes and experiences of the beneficiaries act as a bridge between policy-oriented feedback and personal experiences.

The transformation from K-means clusters to detected communities through the Louvain method is a shift from a heuristic grouping based on similarity to a more natural grouping based on actual relationships within the data. Compared to the K-means clusters, the communities produced through the Louvain method refined the relationships by identifying more meaningful connections and a clearer separation between different groups. The graph also highlights central bi-grams or those with many connections within the community, like "quality education", and peripheral bi-grams, or those with fewer connections, such as "helpful supporting". In terms of

themes, the Louvain method retained the main topics from the K-means clusters but has provided a more detailed and clearer view of the selection of bi-grams. The same domain experts examined the bi-gram communities and aligned their labeling with topics identified in the K-means clusters, with a slight modification. Cluster 0 originally labeled as "Support and Educational Opportunity" was tagged as "Support and Program Effectiveness". This adjustment was prompted by the prevalence of text pairs that specifically express assistance provided by the program and its effectiveness in achieving its goals of helping students.

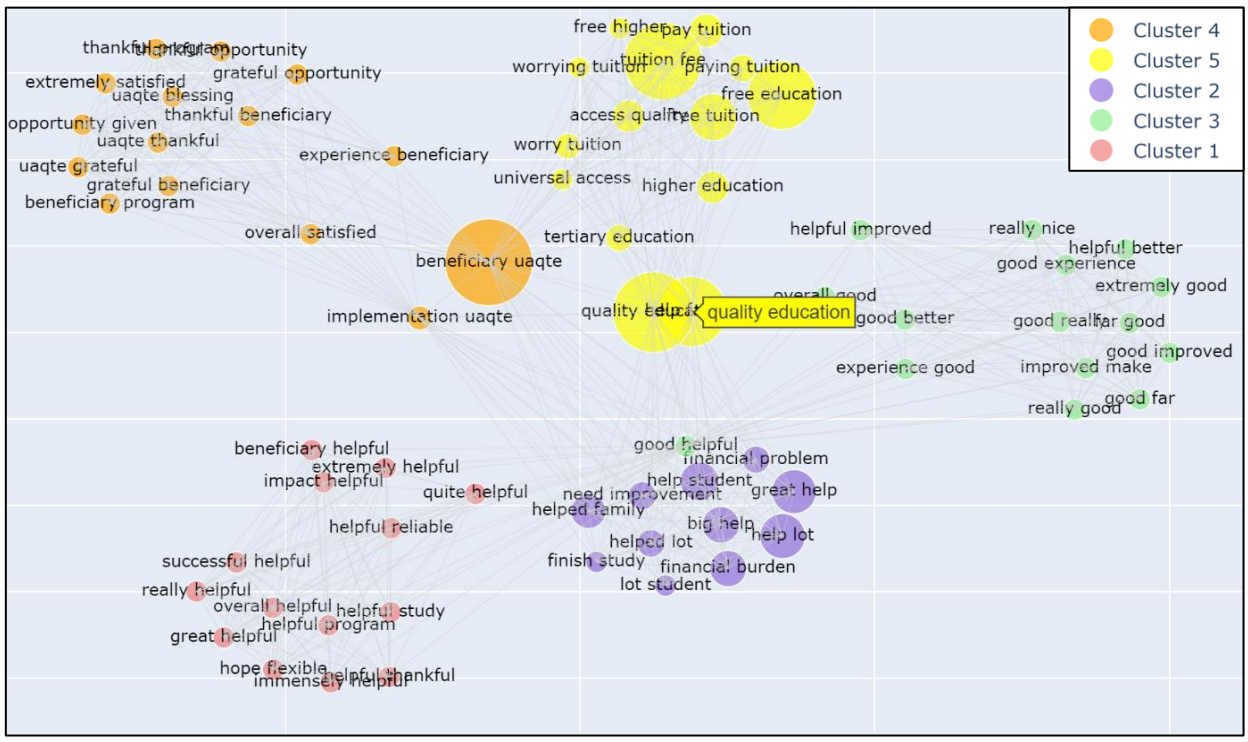

Fig. 4. Communities generated using the Louvain algorithm.

In summary, the methods employed uncover various strengths of the UAQTE program that significantly contribute to a positive impact on its beneficiaries. Among these is the program's provision of support and educational opportunities which alleviate financial burdens on students and their families. Recipients were able to pursue higher education and careers of their choice.

This financial support extends beyond accessibility to higher education by showing transformative effect on beneficiaries' lives. It is reflected through the sense of gratitude and satisfaction among the respondents.

While recognized for its strengths, several areas require attention to improve the program's overall effectiveness. One key area is the need to provide other services beyond financial aid, such as mentoring, counseling, career guidance, and academic assistance. It was highlighted that more efforts are essential to reach potential beneficiaries who may be overlooked due to geographical and socioeconomic barriers. Furthermore, continuous monitoring and evaluation are crucial for identifying areas for improvement and ensuring the effectiveness and relevance of the free higher education initiative. This kind of evaluation ensures that the program remains aligned with its goals and is responsive to the evolving needs of the beneficiaries. Addressing these areas through policy enhancement and development can significantly improve the effectiveness of educational support programs like UAQTE, making education more accessible for all students. In a wider scope, these improvements can strengthen the educational landscape in the Philippines, broadening the free education program's social and economic impact.

V. Conclusion

In this paper, we explored an innovative method for generating contextual analysis from qualitative responses related to the experiences, perceived impact, and overall feedback of the beneficiaries of the Universal Access to Quality Education (UAQTE) program. Our approach employed a two-tier clustering method, using the K-means algorithm and Louvain community detection, based on enhanced word embeddings and a bi-gram network. Our methodology extends beyond traditional text analysis by combining Word2vec and Glove embeddings. Through the application of the K-means algorithm and Louvain method in a bi-gram network, we identified five distinct clusters, corresponding to themes of "Support and Educational Opportunity", "Accessibility and Financial Relief", "Gratitude and Satisfaction", "Positive Evaluation with Suggestions for Improvement", and "Program Effectiveness". The results of the analysis highlight the strengths of the UAQTE program in providing support to the beneficiaries, achieving its goal of providing quality tertiary education. However, it also reveals certain areas

needing attention and improvement. These insights are crucial in policy development and enhancement.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization, Methodology, Data Curation and Experiments: Mary Joy Canon; Validation and Formal Analysis: Mary Joy Canon, Lany Maceda; Writing of original draft: Mary Joy Canon; Review and Editing: Lany Maceda and Nancy Flores. All authors approved the final version.

FUNDING

This paper is funded by the Philippine Commission on Higher Education (CHED)—Leading the Advancement of Knowledge in Agriculture and Science (LAKAS) Project. Particularly, supported by eParticipation 2.1, which focuses on harnessing Natural Language Processing (NLP) for community participation.

ACKNOWLEDGMENT

The authors wish to thank the CHED-LAKAS.

REFERENCES

[1] T. Kromydas, "Rethinking higher education and its relationship with social inequalities: Past knowledge, present state and future potential," *Palgrave Communications*, vol. 3, no. 1, pp. 1–12, 2017.

[2] I. Demirbağ and S. Sezgin, "Book review: Guidelines on the development of open educational resources policies," *The International Review of Research in Open and Distributed Learning*, vol. 22, no. 2, pp. 261–263, 2021.

[3] K. Shiohira. (2021). Understanding the impact of artificial intelligence on skills development. UNESCO-UNEVOC International Centre for Technical and Vocational Education and Training. [Online]. Available: https://unevoc.unesco.org/pub/understanding_the_impact_of_ai_on_skills_development.pdf

[4] A. Ashida, "The role of higher education in achieving the sustainable development goals," *Sustainable Development Disciplines for Humanity*, pp. 71–84, 2003. https://doi.org/10.1007/978-981-19-4859-6_5

[5] The 1987 constitution of the Republic of the Philippines. Article II: Declaration of Principles and State Policies. [Online]. Available: https://www.officialgazette.gov.ph/constitutions/the-1987-constitution-of-the-republic-of-the-philippines/the-1987-constitution-of-the-republic-of-the-philippines-article-ii/

[6] Republic of the Philippines. Seventeenth Congress of the Philippines. Act No. 10931. [Online]. Available: https://www.officialgazette.gov.ph/2017/08/03/republic-act-no-10931/

[7] M. K. P. Ortiz, K. A. M. Melad, N. V. V. Araos, A. C. Orbeta Jr., and C. M. Reyes, "Process evaluation of the universal access to quality tertiary education act (RA 10931): Status and prospects for improved implementation," *PIDS Discussion Paper Series*, Philippine Institute for Development Studies, 2019.

[8] C. Hao, M. S. Nyaranga, and D. O. Hongo, "Enhancing public participation in governance for sustainable development: Evidence from Bungoma county," *Kenya. SAGE Open*, vol. 12, no. 1, 2020.

[9] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in Neural Information Processing Systems*, vol. 13, 2000.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[11] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532−1543.

[12] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135−146, 2017.

[13] Q. J. Qiu, Z. Xie, L. Wu, and W. J. Li, "Geoscience keyphrase extraction algorithm using enhanced word embedding," *Expert Systems with Applications*, vol. 125, 2019.

[14] Y. Qi, D. S. Sachan, M. Felix, S. J. Padmanabhan, and G. Neubig, "When and why are pre-trained word embeddings useful for neural machine translation?" arXiv preprint, arXiv:1804.06323, 2018.

[15] R. Biswas and S. De, "A comparative study on improving word embeddings beyond Word2Vec and GloVe," in *Proc. 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, 2022, pp. 113−118.

[16] L. C. Yu, J. Wang, K. R. Lai, and X. J. Zhang, "Refining word embeddings for sentiment analysis," in *Proc. 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 534–539.

[17] L. C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings using intensity scores for sentiment analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 671–681, March 2018.

[18] S. M. Rezaeinia, R. Rahmani, A. Ghodsi, and H. Veisi, "Sentiment analysis based on improved pre-trained word embeddings," *Expert Systems with Applications*, vol. 117, 2019.

[19] S. M. Rezaeinia, A. Ghodsi, and R. Rahmani, "Improving the accuracy of pre-trained word embeddings for sentiment analysis," arXiv preprint, arXiv:1711.08609, 2017.

[20] W. Zhang, P. C. Wang, M. N. Wang, and L. Y. Chen, "Big data mining and analysis of hot issues in international education—Based on K-means algorithm of cluster analysis," in *Proc. 2020 International Conference on Information Science and Education (ICISE-IE)*, Sanya, China, 2020, pp. 1–4. doi: 10.1109/ICISE51755.2020.00008

[21] L. Yuan, H. Zhao, and Z. Wang, "Research on news text clustering for international chinese education," in *Proc. 2023 International Conference on Asian Language Processing (IALP)*, Singapore, Singapore, 2023, pp. 377−382. doi: 10.1109/IALP61005.2023.10337054

[22] Y. Zou, "Construction of hot spot tracking model of university network public opinion based on text clustering," in *Proc. 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Xi'an, China, 2021, pp. 76–80. doi: 10.1109/ITNEC52019.2021.9586841

[23] J. Tao *et al*., "Cluster analysis on Chinese university students' conceptions of English language learning and their online self-regulation," *Australasian Journal of Educational Technology*, vol. 36, no. 2, pp. 105–119, 2020.

[24] D. Agnihotri, K. Verma, and P. Tripathi, "Pattern and cluster mining on text data," in *Proc. 2014 Fourth International Conference on Communication Systems and Network Technologies*, Bhopal, India, 2014, pp. 428−432. doi: 10.1109/CSNT.2014.92

[25] M. J. Parreño *et al.*, "The use of gamification in education: a bibliometric and text mining analysis," *Journal of Computer Assisted Learning*, vol. 32, 2016.

[26] J. Y. Han and S. H. Lee, "Investigating online learning types based on self-regulated learning in online software education: Applying hierarchical cluster analysis," *The Journal of Korean Association of Computer Education*, vol. 22, no. 5, pp. 51–65, 2019.

[27] M. A. Javed, M. S. Younis, S. Latif, J. Qadir, and A. Baig, "Community detection in networks: A multidisciplinary review," *Journal of Network and Computer Applications*, vol. 108, pp. 87–111, 2018.

[28] A. Allman, W. Tang, and P. Daoutidis, "Towards a generic algorithm for identifying high-quality decompositions of optimization problems," *Computer Aided Chemical Engineering*, vol. 44, pp. 943–948, 2018.

[29] S. Yassine, S. Kadry, and M. A. Sicilia, "Application of community detection algorithms on learning networks. The case of Khan

academy repository," *Computer Applications in Engineering Education*, vol. 29, no. 2, pp. 411–424, 2021.
[30] C. M. Chen and Z. L. You, "Community detection with opinion leaders' identification for promoting collaborative problem-based learning performance," *British Journal of Educational Technology*, vol. 50, 2018.
[31] E. G. Sturludóttir *et al*., "Gaining insights on student course selection in higher education with community detection," in *Proc. 14th International Conference on Educational Data Mining*, 2021.
[32] S. Yassine, S. Kadry, and M. A. Sicilia, "Detecting communities using social network analysis in online learning environments: Systematic literature review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 1, 2022.
[33] P. Evgenia *et al*., "Community detection and social presence in students," *Discussion Fora*, pp. 879–891, 2023.
[34] M. S. Saputri *et al*., "Browsing behavior analysis from Wi-Fi logs based on community detection: Case study on educational institution," in *Proc. 2018 International Workshop on Big Data and Information Security (IWBIS)*, 2018, pp. 87–92.
[35] Y. Shao, "Student community detection and recommendation of customized paths to reinforce academic success," Master's thesis, University of Central Florida, USA, 2019.
[36] I. T. Jolliffe, "Principal component analysis," *Springer Series in Statistics*, vol. 101, 2002.
[37] D. A. Simovici and C. Djeraba, *Mathematical Tools for Data Mining*, Springer London, 2014.
[38] E. Umargono, J. E. Suseno, and S. K. V. Gunawan, "K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula," in *Proc. the 2nd International Seminar on Science and Technology*, 2019.
[39] R. K. Roul, J. K. Sahoo, and K. Arora, "Modified TF-IDF term weighting strategies for text categorization," in *Proc. 2017 14th IEEE India Council International Conference*, 2017, pp. 15–17.
[40] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical Mechanics: Theory and Experiment*, vol. 10, 2008.
[41] F. Bonchi, A. Gionis, and A. Ukkonen, "Overlapping correlation clustering," *Knowledge and Information Systems*, vol. 35, 2013.
[42] S. Khanmohammadi, N. Adibeig, and S. Shanehbandy, "An improved overlapping k-means clustering method for medical applications," *Expert Systems with Applications*, vol. 67, 2017.
[43] J. D. Holster. Introduction to R for Data Science: A LISA 2020 Guidebook. [Online]. Available: https://bookdown.org/jdholster1/idsr/