International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I6P118

Volume 72 Issue 6, 170-177, June 2024

Original Article

Clustering with Enhanced Word Embeddings for Contextual Analysis in Academic Texts

Mary Joy P. Canon[1*], Lany L. Maceda[1], Christian Y. Sy[1]

[1]*Computer Science and Information Technology Department, Bicol University, Legazpi City, Philippines.*

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

Received: 25 February 2024 Revised: 15 May 2024 Accepted: 19 May 2024 Published: 29 June 2024

Abstract *- To provide deserving Filipino students access to higher education, the Universal Access to Quality Education (UAQTE) program was enacted into law. However, despite its years of implementation, there remains a lack of comprehensive understanding of its perceived impact and feedback among its recipients. This paper explored an advanced text analysis approach in contextual understanding of text responses related to the implementation of the UAQTE by employing enhanced word embeddings from Word2Vec and Glove vectors, K-Means clustering algorithm and bi-gram word network. The combination of Word2vec and Glove embeddings captured the semantic meaning of words within the dataset. Five distinct groups were identified using the K-means algorithm which gained a decent silhouette score of 0.3477. Based on the computed TF-IDF scores for the bi-grams, top sequences for each cluster were used for the visualization of a text network graph. Accordingly, domain experts labeled the clusters of responses as "Support and Educational Opportunity", "Accessibility and Financial Relief", "Gratitude and Satisfaction", "Positive Evaluation with Suggestions for Improvement" and "Program Effectiveness". This approach not only highlights the strengths of the UAQTE program in providing support to the beneficiaries but also reveals certain areas needing attention and improvement, which are crucial in policy development and enhancement. Future work may focus on diversified data by incorporating feedback from other stakeholders, such as program implementers and educators.*

Keywords *- Clustering, Enhanced word embedding, Program Evaluation, Quality tertiary education, Text analysis.*

1. 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 [1] [2] as well as uplifting the economic progress of a country. 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 [3] [4]. Directly aimed at the development of higher education is target 4.3 of the 17 Sustainable Development Goals (SDG), which intends to ensure equal access for all women and men to affordable and quality technical, vocational, and tertiary education, including university [5]. In the Philippines, access to quality education is an inalienable right of all Filipinos [6]. Recognizing this right, in 2017, the country enacted into law the Republic Act No. 10931 also known as the Universal Access to Quality Tertiary Education Act (UAQTE). This law 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 TESDA technical-vocational training institutes; c. Tertiary Education Subsidy (TES); and d. student loan programs (RA 10931). 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. Notwithstanding the provision of tuition subsidies, some entities questioned the program's effectiveness in reaching those in need and overcoming pre-existing educational barriers [7]. This highlights a significant research gap: there is a lack of comprehensive understanding regarding the recipients' perceived impact and overall feedback on the program.

Despite years of implementation, the UAQTE program in the Philippines has not been thoroughly evaluated from the perspective of the beneficiaries. Given the complexity and scope of the UAQTE program, a comprehensive evaluation is necessary to improve the chances of achieving its objectives. Previous research, such as the discussion paper by [7], stressed the need to evaluate the implementation of the program even in its early stages. 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. 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 efficacy and impact. In the perspective of qualitative data analysis, Natural Language Processing(NLP) offers powerful approaches to analyze, model, and process text data. Incorporating this technology in modeling and analyzing 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 and a bi-gram word network. This approach offers novelty by leveraging advanced NLP techniques to provide deeper insights compared to traditional qualitative analysis methods used in previous studies.

2. Literature Review

2.1. Enhanced Word 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 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. In another work [13], an OEWE method that combines domain-based ontologies with word embeddings, enhancing keyphrase extraction from geological documents, was proposed. This approach suggests that embedding models can be tailored to specific domains for more accurate results. Two studies [14] [15] have 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 NLP tasks.

2.2. 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. [16] innovatively employed K-means clustering to dissect international education trends, with a particular focus on the ramifications of the COVID-19 pandemic. Yuan, Zhao, and Wang [17] adeptly applied clustering to categorize news texts related to international Chinese education, achieving a significant accuracy rate. In a different context, a model [18] that employs text clustering to track and analyze public opinion trends in university networks effectively was crafted, which provides insights into the dynamic nature of online educational discourse. This is complemented by Tao et al. [19] by utilizing 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 [20]. Additionally, prevailing themes and trends were identified on the application of gamification in education through bibliometric and text mining analysis [21].

2.3. Word Network Graphs

Word network graphs are a powerful tool for visualizing and analyzing the relationships between words in a text corpus. These graphs represent words as nodes and their co-occurrences as edges, creating a network that captures the semantic and syntactic structure of the text [22]. By analyzing the patterns and structures within these networks, researchers can gain insights into the thematic and conceptual organization of the text.

These graphs have been utilized in a variety of applications, including the analysis of literary texts, identification of key themes in academic papers, and exploration of social media discourse. For example, Wang et al. [23] proposed an unsupervised keyword extraction method that enhances traditional word graph networks by

incorporating word embeddings to assess semantic relevance between words, resulting in improved keyword extraction for both Chinese and English texts. In another study, Chinotaikul and Vinayavekhin [24] applied bibliometric and co-word network analysis to objectively identify influential articles and provide the analysis of intellectual structure in exploring digital transformation as a field in business and management research. Recent advancements in the field have focused on enhancing the interpretability and scalability of word network graphs. Techniques such as community detection and centrality measures have been employed to identify clusters of related words and key nodes within the graph, respectively [25]. Additionally, the integration of word embeddings with network analysis has been explored to improve the semantic representation of words and the accuracy of thematic extraction [26].

3. Data and Methods

This section describes the dataset used in the study, as well as the methodology for discovering themes present in the text corpus using enhanced word embeddings and clustering.

3.1. 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, the researchers made use of the dataset collected using the BosesKo application, a citizen's participation toolkit, through which beneficiaries of the free tuition and education subsidy participated in a survey. 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. 3,150 beneficiaries completed the survey between December 15, 2022, and December 8, 2023. A total of 8,536 responses were used for text processing.

3.2. Text Pre-processing

To enhance the data quality and relevance of our dataset, different libraries of the Natural Language Toolkit (NLTK) were employed in transforming the corpus. Most of the collected samples 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 in analysis, potentially skewing the dataset's semantic meaning.

3.3. Enhanced Word Embeddings

To create a richer set of features, the researchers leveraged the combination of Word2vec and Glove methods and then reduced the dimensionality of the generated embeddings. Word2Vec [10], [27]-[28] is a predictive embedding model, which is trained to either predict a word given its context or to predict the context given a. The model generates embeddings such that words that occur in similar contexts are close to each other in the embedding space. 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.

To have 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 were 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, using a function, the researchers combined the features by averaging the vectors of the corresponding words. Refer to the formula (1) below:

$$V_{combined}(w) = \frac{V_{w2v}(w)+V_{glove}(w)}{2} \quad (1)$$

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 is used. The researchers then reduced the dimensionality of the embeddings using Principal Component Analysis (PCA) [29]. Applying PCA to reduce dimensions can help distil the most relevant linguistic or semantic features from these combined embeddings, which is crucial for downstream tasks like clustering.

3.4. 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 [30]. 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 [31]. In the context of the present study, the algorithm tries to group similar embeddings into the same

cluster. K-means objective function in formula (2) quantifies the goal of the clustering process.

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

In exploring the performance of this algorithm, we used different k values, ranging from 2 to 9. The researchers computed the silhouette score for each cluster configuration to determine an object's similarity to its own 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 social scientists who examined the instances for each cluster and identified the corresponding themes.

3.5. Generated Word Network Graph

Python's NetworkX and Matplotlib libraries, along with Scikit-learn for Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, were implemented in producing text network graphs. These tools are necessary to visually represent the relationship between words or bi-grams present in the text clusters. The key part of the network construction 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 (3).

$$\text{TF} - \text{IDF}(t, d) = \text{TF}(t, d) \text{ x IDF}(t) \quad (3)$$

Term Frequency TF operates on the principle that the frequent occurrence of a term *t* in a document *d* is indicative of *t*'s significance for *d*. Accordingly, Inverse Document Frequency (IDF) gauges the rarity of term *t* across the entire corpus [32], assigning higher values to terms that are less common. 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, the researchers identified the 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.

4. Results and Discussion

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

4.1. Enhanced Word Embeddings

In reference to the Word2Vec and Glove vocabularies, the researchers identified 3,460 unique text embeddings present in our corpus. This number indicates a substantial lexical variety of the terms within the dataset. Table 1 presents the magnitude of the combined embeddings generated using Word2vec and Glove methods. Word2Vec score and Glove score 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. The use of enhanced word embeddings in the study, specifically through the combination of Word2Vec and GloVe vectors, contributed to a more nuanced understanding of the text data compared to traditional word embeddings. These embeddings capture both local context (via Word2Vec) and global co-occurrence statistics (via GloVe), allowing for a deeper analysis of semantic relationships and leading to more accurate thematic extraction and keyword identification.

Table 1. Sample tokens and their corresponding scores

Token	Word2Vec score	Glove score	Final score
assistance	2.8296824	6.2453046	3.544196
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
free	2.4175155	6.4717374	3.5070798

4.2. Clustering Experimental Results

The experimental results of clustering, as presented in Table 2, compare the performance of K-means and Agglomerative Clustering algorithms using the silhouette score as a measure of cluster quality. The silhouette score ranges from -1 to 1, with higher values indicating better-defined clusters. For K-means clustering, the highest silhouette score is 0.3512 for 3 clusters, suggesting that this configuration achieves the best balance of cohesion within clusters and separation between clusters.

However, a slightly lower silhouette score of 0.3477 was observed for 5 clusters, which was chosen due to the additional context and detail it provided, capturing the nuances in the data more effectively despite the slight drop in score. On the other hand, Agglomerative Clustering shows generally lower silhouette scores across all configurations compared to K-means, with the highest score being 0.3396 for 3 clusters. This indicates that K-means performs better in defining distinct clusters in this particular dataset. The results highlight the importance of selecting an appropriate number of clusters and clustering methods based on the specific characteristics and requirements of the dataset. While K-means provided better overall results, Agglomerative Clustering could be useful in scenarios where hierarchical relationships among data points are of interest. The choice of 5 clusters for detailed analysis was driven by the need to capture more granular themes and insights relevant to the UAQTE program.

Table 2. Experimental results of clustering

Algorithm	No. of clusters	Silhouette Score
K-Means Clustering	2	0.3409
	3	0.3512
	4	0.3482
	5	0.3477
	6	0.3278
Agglomerative Clustering	2	0.3238
	3	0.3396
	4	0.3387
	5	0.3290
	6	0.2836

4.3. Generated Clusters using K-means Algorithm

After testing various cluster configurations, the researchers selected the K-Means algorithm with five as the *k* value to define the groupings of the text responses. This number of clusters obtained a silhouette score of 0.3477, slightly lower than *k=3* with 0.3512. The 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.

Visualized in Figure 1 is the clustering of text responses into five distinct 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.

The researchers identified that five clusters provided the most meaningful insights. This configuration balanced the need for detailed analysis with statistical rigor, ensuring that the clusters were both interpretable and statistically significant. Unlike some state-of-the-art techniques that might rely on predetermined cluster numbers or less flexible algorithms, our approach allowed for a more tailored and insightful segmentation of the data, directly reflecting the diverse themes within the UAQTE program.

The word clouds for each cluster (Figures 2-6) 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 presented 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. Overlaps are generally acceptable in the clustering method [33] because real-world datasets have inherently overlapping clusters [34].

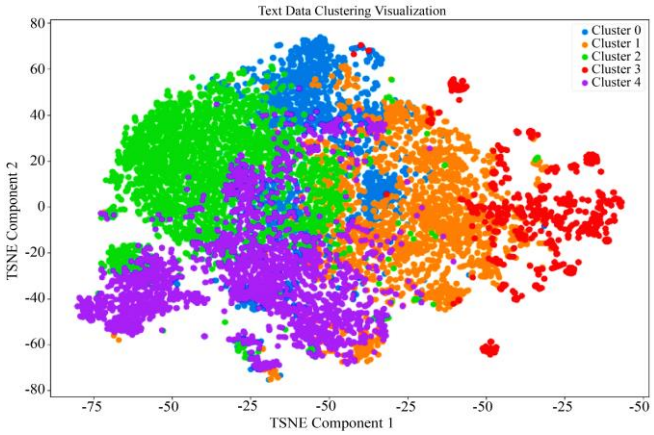

Fig. 1 t-SNE plot of the generated clusters using the K-means algorithm

4.4. Identified Themes by Domain Experts

Table 3 shows 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 give emphasis on 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 the overall improvement of well-being. This implies a transformative effect of the UAQTE program on the beneficiaries and their families.

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 being 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.

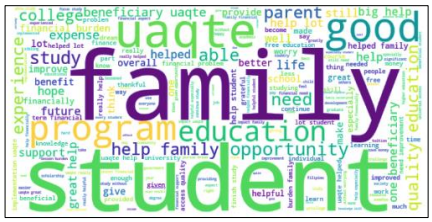

Fig. 2 Most important words in Cluster 0

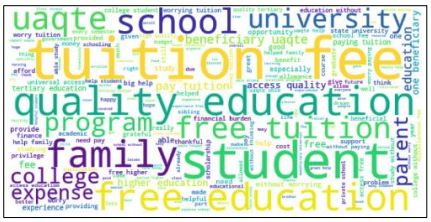

Fig. 3 Most important words in Cluster 1

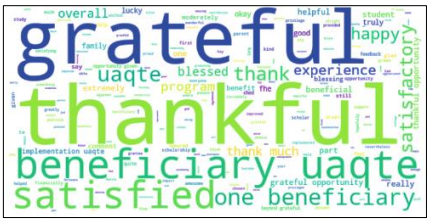

Fig. 4 Most important words in Cluster 2

Fig. 5 Most important words in Cluster 3

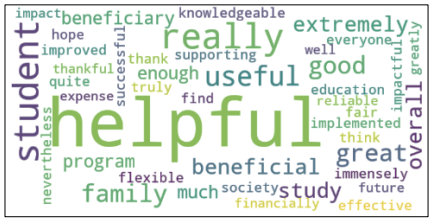

Fig. 6 Most important words in Cluster 4

Table 3. Identified themes and sample instances per cluster

Cluster	Identified Theme	Sample Instances
0	Support and Educational Opportunity	(1) helped lessen costs parents would shouldered (2) lifted huge burden shoulder allowing focus study without worrying much finance (3) beneficiaries able to access high-quality education UAQTE provided knowledge skills needed to pursue their dream
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 support family financially
2	Gratitude and Satisfaction	(1) thankful blessed one beneficiary UAQTE, (2) UAQTE truly beneficial (3) satisfied happy
3	Positive Evaluation with suggestions for improvement	1)overall good, (2) good experience, (3) good but improved, (4) good really helps a lot, (5) good but could be better
4	Program Effectiveness	(1) really effective helpful, (2)helpful reliable, (3)helpful future, (4) helpful but hope flexible

4.5. Bi-gram Text Network

A text network graph for top bi-grams from the clusters was created. Based on the computed TF-IDF scores for the bi-grams, the researchers determined the top sequences for each cluster and used these terms for visualization of the text network graph depicted in Figure 7. In this graph, the nodes correspond to the top 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 uaqte”, “implementation uaqte”, “help 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. Furthermore, these combinations of words semantically strengthen the identified themes in the previous section. They highlight the aspects of positive feedback, educational opportunities, 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.

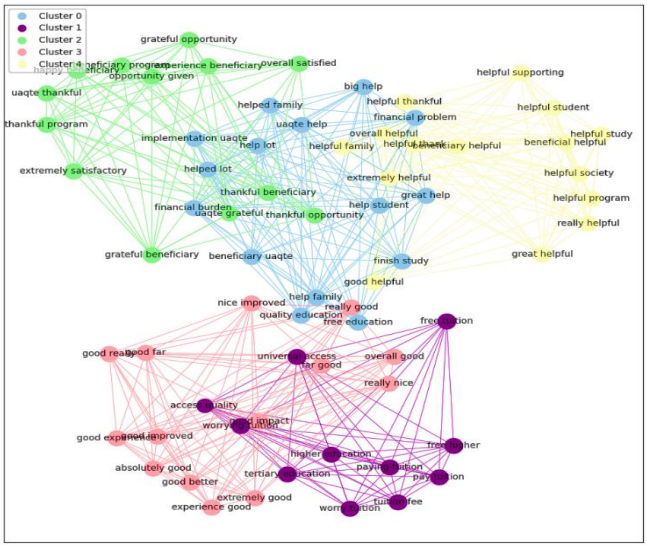

Fig. 7 Text network of Top Bi-grams per Cluster

5. Conclusion

In this paper, the researchers explored advanced text analysis approaches in the contextual understanding of texts by employing enhanced word embeddings using Word2Vec and Glove models, K-Means clustering algorithm, and bi-gram word network. Mainly, using the mentioned techniques, the researchers uncovered themes and contextual understanding from text responses encompassing narratives on the experiences, perceived impact, and overall feedback of the UAQTE beneficiaries.

The combination of Word2vec and Glove embeddings captured the semantic meaning of words within the dataset, generating rich input features for clustering. The first level of text partitioning using 5 clusters gained a relatively good silhouette score of 0.3477. Accordingly, domain experts labeled these clusters as "Support and Educational Opportunity", "Accessibility and Financial Relief", "Gratitude and Satisfaction", "Positive evaluation with Suggestions for Improvement" and "Program Effectiveness". Moreover, based on the computed TF-IDF scores for the bi-grams, the researchers determined the top sequences for each cluster and used these for visualization of the text network graph.

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. It also emphasizes the need for continuous evaluation of the education programs to meet the changing needs of the recipients. Future work may focus on the diversity of data by incorporating responses and feedback from other stakeholders, such as program implementers and educators.

Funding Statement

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

Acknowledgment

The authors wish to thank the Philippine Commission on Higher Education (CHED) Leading the Advancement of Knowledge in Agriculture and Science (LAKAS) for funding.

References

[1] Fengqin Liu et al., "Retracted Article: Role of Education in Poverty Reduction: Macroeconomic and Social Determinants form Developing Economies," *Environmental Science and Pollution Research*, vol. 28, pp. 63163-63177, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] R. Ukwueze Ezebuilo, and O. Nwosu Emmanuel, "Does Higher Education Reduce Poverty among Youths in Nigeria?," *Asian Economic Financial Review*, vol. 4, no. 1, pp. 1-19, 2014. [Google Scholar] [Publisher Link]

[3] İrem Demirbağ, and Sedef 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. [CrossRef] [Google Scholar] [Publisher Link]

[4] Shiohira Kelly, "*Understanding the Impact of Artificial Intelligence on Skills Development. Education 2030*," 2021. [Google Scholar] [Publisher Link]

[5] Akemi Ashida, *The Role of Higher Education in Achieving the Sustainable Development Goals*, Sustainable Development Disciplines for Humanity, Springer, Singapore, pp. 71-84, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Official Gazette, The 1987 Constitution of the Republic of the Philippines – Article II. [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/

[7] P. Ortiz Ma. Kristina et al., "*Process Evaluation of the Universal Access to Quality Tertiary Education Act (RA 10931): Status and Prospects for Improved Implementation*," Philippine Institute for Development Studies, Quezon City, 2019. [Google Scholar] [Publisher Link]

[8] Chen Hao, Maurice Simiyu Nyaranga, and Duncan O. Hongo, "Enhancing Public Participation in Governance for Sustainable Development: Evidence From Bungoma County, Kenya," *Sage Open*, vol. 12, no. 1, pp. 1-15, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent, "A Neural Probabilistic Language Model," *Advances in Neural Information Processing Systems*, vol. 13, 2000. [Google Scholar] [Publisher Link]

[10] Tomas Mikolov et al., "Distributed Representations of Words and Phrases and their Compositionality," *Advances in Neural Information Processing Systems*, vol. 26, 2013. [Google Scholar] [Publisher Link]

[11] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, "Glove: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, pp. 1532-1543, 2014. [Google Scholar] [Publisher Link]

[12] Piotr Bojanowski et al., "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135-146, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[13] Qinjun Qiu et al., "Geoscience Keyphrase Extraction Algorithm Using Enhanced Word Embedding," *Expert Systems with Applications*, vol. 125, pp. 157-169, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Ye Qi et al., "When and Why are Pre-trained Word Embeddings Useful for Neural Machine Translation?," *arXiv*, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[15] Rajdeep Biswas, and Suman De, "A Comparative Study on Improving Word Embeddings Beyond Word2Vec and GloVe," *2022 Seventh International Conference on Parallel, Distributed and Grid Computing*, Solan, Himachal Pradesh, India, pp. 113-118, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[16] Weili Zhang et al., "Big Data Mining and Analysis of Hot Issues in International Education—Based on K-Means Algorithm of Cluster Analysis," *2020 International Conference on Information Science and Education*, Sanya, China, pp. 1-4, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Liangjie Yuan, Huizhou Zhao, and Zhimin Wang, "Research on News Text Clustering for International Chinese Education," *2023 International Conference on Asian Language Processing*, Singapore, Singapore, pp. 377-382, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Yuxiang Zou, "Construction of Hot Spot Tracking Model of University Network Public Opinion Based on Text Clustering," *2021 IEEE 5th Information Technology,Networking,Electronic and Automation Control Conference*, Xi'an, China, pp. 76-80, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jing 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. [CrossRef] [Google Scholar] [Publisher Link]

[20] Deepak Agnihotri, Kesari Verma, and Priyanka Tripathi, "Pattern and Cluster Mining on Text Data," *2014 Fourth International Conference on Communication Systems and Network Technologies*, Bhopal, India, pp. 428-432, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[21] J. Martí-Parreño, E. Méndez-Ibáñez, and A. Alonso-Arroyo, "The Use of Gamification in Education: A Bibliometric and Text Mining Analysis," *Journal of Computer Assisted Learning*, vol. 32, no. 6, pp. 663-676, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[22] Wei Jin, and Rohini Kesavan Srihari, "Graph-Based Text Representation and Knowledge Discovery," *Proceedings of the 2007 ACM Symposium on Applied Computing*, Seoul Korea, pp. 807-811, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[23] Hongbin Wang et al., "Unsupervised Keyword Extraction Methods Based on a Word Graph Network," *International Journal of Ambient Computing and Intelligence*, vol. 11, no. 2, pp. 68-79, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Punthira Chinotaikul, and Sukrit Vinayavekhin, "Digital Transformation in Business and Management Research: Bibliometric and Co-word Network Analysis," *2020 1st International Conference on Big Data Analytics and Practices*, Bangkok, Thailand, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[25] Didier A. Vega-Oliveros et al., "A Multi-Centrality Index for Graph-Based Keyword Extraction," *Information Processing & Management*, vol. 56, no. 6, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[26] Simon Briscoe, Rebecca Abbott, and G.J. Melendez-Torres, "Expert Searchers Identified Time, Team, Technology and Tension as Challenges when Carrying Out Supplementary Searches for Systematic Reviews: A Thematic Network Analysis," *Health Information & Libraries Journal*, vol. 41, no. 2, pp. 182-194, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] Tomas Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," *International Conference on Learning Representations*, 2013. [Google Scholar] [Publisher Link]

[28] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever, "Exploiting Similarities among Languages for Machine Translation," *arXiv*, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[29] I.T. Jolliffe, *Principal Component Analysis*, Springer, pp. 1-487, 2002. [Google Scholar] [Publisher Link]

[30] Dan A. Simovici, and Chabane Djeraba, *Clustering*, *Mathematical Tools for Data Mining*, Advanced Information and Knowledge Processing, pp. 767-817, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[31] Edy Umargono, Jatmiko Endro Suseno, and S.K Vincensius Gunawan, "K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula," *Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[32] Rajendra Kumar Roul, Jajati Keshari Sahoo, and Kushagr Arora, "Modified TF-IDF Term Weighting Strategies for Text Categorization," *2017 14th IEEE India Council International Conference*, Roorkee, India, pp. 1-6, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[33] Francesco Bonchi, Aristides Gionis, and Antti Ukkonen, "Overlapping Correlation Clustering," *Knowledge and Information Systems*, vol. 35, pp. 1-32, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[34] Arindam Banerjee et al., "Model-Based Overlapping Clustering," *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago Illinois USA, pp. 532-537, 2005. [CrossRef] [Google Scholar] [Publisher Link]