

# Participant-Driven Design of Community Program Topics for Aging Communities: A Text Clustering Approach to Activity Preferences

Indri Hapsari<sup>1\*</sup>, Devi Rachmasari<sup>1</sup>

<sup>1</sup>Universitas Surabaya, Surabaya

<sup>1</sup>[indri@staff.ubaya.ac.id](mailto:indri@staff.ubaya.ac.id)

\*Corresponding Author

Received 27 April 2026; accepted 08 July 2026

**Abstract.** The rapid growth of Indonesia's elderly population highlights the need for community programmes that not only extend lifespan but also improve healthspan. This study aims to design ageing activity topics based on participants' suggestions using Natural Language Processing (NLP) and unsupervised clustering. Data were collected from 213 open-ended responses from participants of the Center for Aging Wellness (CAW), University of Surabaya, covering liked aspects, improvement areas, and future topic suggestions. Text data were pre-processed (lowercasing, removal of punctuation, digits, and stopwords, and spelling normalization), then transformed using TF-IDF and clustered with K-Means. Five values of  $k$  (2–6) were tested, with  $k = 6$  selected based on the highest silhouette score and interpretability. Six thematic clusters emerged: knowledge literacy (12.2%), appreciation and learning (35.7%), leisure and social activities (8.9%), brain health and dementia (18.3%), physical and mental health (15.5%), and financial independence (9.4%). These clusters were mapped into broader need categories and aligned with the Indonesia Longitudinal Aging Survey (ILAS) 2023 domains. As an exploratory study, the results support thematic mapping and programme planning rather than prediction. The findings demonstrate that participant feedback can be transformed into an evidence-based roadmap for more targeted and context-sensitive ageing community programmes.

**Keywords:** Aging; Natural Language Processing; Text Clustering; K-Means; Community-based programmes

## 1 Introduction

The proportion of elderly in Indonesia continues to increase. According to Statistics Indonesia (BPS), the share of the elderly ( $\geq 60$  years) rose from 7.56%, or approximately 18 million people, in 2012 [1], to 11.75%, or around 32 million people, in 2023 [2]. This demographic shift has important implications for health, social, and economic systems and requires serious attention from both government and society. The present study is situated within the UN Decade of Healthy Aging (2021–2030), which calls for coordinated action to improve later-life health, participation, and dignity through age-friendly communities and person-centred services [3]. WHO also highlights that rapid population aging is accompanied by a growing gap between lifespan and healthy life expectancy, with many elderly people living longer but with multiple chronic conditions and functional limitations [4]. As age advances, the elderly become more vulnerable to declines in fitness, physical weakness, and various health

problems, which in turn increase their dependence on assistive devices and support from people around them. The decline in sensory functions, coupled with accumulating life stressors, can also disrupt communication and demand greater discipline in self-regulation. These health problems mean that some elderly individuals do not reach the national life expectancy, which was around 73.93 years in 2023. As a consequence, their long-standing friendship networks and community ties shrink, feelings of loneliness increase, and families are not always able to provide optimal support due to competing demands and busy schedules. Declining health and social conditions then spill over into economic aspects, such as more limited employment opportunities, reduced productivity, and shrinking social networks, ultimately affecting the mental health of the elderly. Classic equity frameworks emphasize that health in later life is shaped not only by medical care but also by social, economic, and environmental determinants, underscoring the importance of community-based interventions that improve access to support and participation for everyone [5].

The 2023 Indonesia Longitudinal Aging Survey [6] provides a detailed portrait of aging in Indonesia based on interviews with more than 4,000 middle-aged and participant responses across various regions, designed to be nationally representative. ILAS serves as a macro-level information source on the situation of the elderly in Indonesia. The survey shows that rising life expectancy does not fully translate into healthy life expectancy because around 70% of participant responses have at least one chronic condition, such as hypertension, metabolic disorders, or digestive problems, and many conditions remain undiagnosed. In the policy context, ILAS functions as a monitoring tool for the National Strategy for Aging [7] and as an instrument for evaluating micro-level indicators from the user perspective. ILAS conceptualises aging wellness as a multidimensional condition that includes physical health, cognitive function, psychological well-being, social relationships, economic security, access to services, and technology literacy. At present, aging wellness in Indonesia remains highly dependent on families and communities, with key challenges including undetected chronic diseases, limited long-term care, digital divides, and low formal economic protection. This situation calls for cross-sectoral approaches grounded in longitudinal data. Longitudinal data refer to data collected repeatedly from the same subjects at multiple time points, allowing researchers to examine patterns over time, such as how health, cognitive function, economic status, or social participation among the elderly progress over several years.

The University of Surabaya recognises both the challenges and opportunities posed by the aging population. On 20 March 2025, it established the Life Science Integrated Facilities and the Center for Aging Wellness (CAW), which are connected to various elderly communities and regularly organise in-person activities such as digital academies, dementia prevention sessions, and fashion styling for the elderly, as well as podcast programmes accessible to the public as an educational medium on aging-related issues. In each activity, CAW distributes evaluation questionnaires in paper and online formats, with items that specifically encourage participants in aging community to provide open feedback. Prior work on technology design stresses that elderly users should be actively involved throughout the research and design process, rather than treated as passive recipients of pre-defined solutions [8]. This study also adopts a similar participant-driven orientation for community programme planning. This study focuses on three key aspects: what participants liked, what they felt needed improvement, and suggested topics for future activities. After 213 questionnaires were collected, open-ended responses were processed systematically and used as an evidence-based foundation for designing more targeted programmes for the aging

communities.

Based on [9] and [10], Natural Language Processing (NLP) is used as a framework to transform the participant's open-ended responses, which are free-form, unstructured, and rich in casual language, into data that can be analysed systematically. The process includes several steps, starting with text cleaning (converting to lowercase, removing punctuation, digits, and special characters, and eliminating non-informative stopwords), tokenisation, and then numerical representation using techniques such as Term Frequency–Inverse Document Frequency (TF-IDF), so that each response can be mapped into a vector space reflecting its most important terms. TF-IDF measures how important a word is within a specific document relative to a collection of documents [11] [12]. Term Frequency (TF) captures how often a word appears in a response, assigning higher weights to more frequent words. Inverse Document Frequency (IDF) down-weights very common words that appear in many documents, such as “and” or “which”. The TF-IDF score is high for words that appear often in one response but rarely in others, helping identify important terms that differentiate themes across responses. In this way, feedback from participants about aspects they liked, areas needing improvement, and suggested topics can be analysed as quantitative data. Recurring thematic patterns can then be identified. These patterns are further grouped using unsupervised clustering to construct a thematic map of needs and activity preferences.

Unsupervised clustering, such as the K-Means algorithm, is then applied to discover groups or clusters in the data without requiring predefined labels or categories [13]. In the context of open-ended responses, each answer is represented as a TF-IDF vector and positioned as a point in a high-dimensional space. K-Means groups nearby points into the same cluster by minimising within-cluster distances and maximising separation between clusters. K-Means is used because it is a simple and efficient unsupervised method for grouping TF-IDF text representations into thematic clusters [14]. It also allows testing different values of  $k$  and evaluating them using the silhouette score.

The result is a set of response clusters, each reflecting a distinct theme, without the need to manually define categories in advance. The combined use of NLP and unsupervised clustering is particularly relevant for summarising patterns of needs and preferences among the participants, as it can uncover structure in evaluation data that initially appears random and highly heterogeneous. Open responses containing impressions, critiques, and topic suggestions are transformed into rational thematic groups, for example, needs related to digital literacy, desires for expanded social interaction, or interests in physical and cognitive health. Programme managers can therefore go beyond simple numerical ratings and obtain a richer thematic picture of what truly matters to the participants, enabling subsequent activities to be more focused, evidence-based, and aligned with the participants' own priorities.

Previous studies have shown that open-ended responses can be analysed using text mining and unsupervised modelling to identify recurring themes. However, limited research has used participant feedback from elderly community programmes as the basis for designing a thematic programme roadmap. In addition, studies linking micro-level community feedback to broader national ageing domains remain scarce. This study addresses that gap by combining NLP-based text analysis with unsupervised clustering to identify thematic patterns in participant feedback and then mapping these patterns to ILAS 2023 domains.

The novelty of this study lies in integrating macro-level evidence from ILAS with micro-level participant feedback from CAW within a single analytical framework. More specifically, the study employs NLP and unsupervised clustering to extract a

thematic map of aging activity preferences and to translate that map into a programme agenda grounded in participant voices rather than organiser assumptions alone. To the best of our knowledge, no prior study has combined open-ended feedback from aging community activities, NLP-based clustering, and ILAS national aging domains to develop a participant-driven program roadmap. This paper contributes both methodologically, by demonstrating a practical text-analytic approach for aging community programmes and by showing how community-based feedback can enrich national healthy aging priorities with concrete and actionable programme themes.

## 2 Method

This study was designed as an exploratory NLP-based thematic analysis, combining quantitative and qualitative text analysis to identify patterns of needs, preferences, and programme expectations emerging from participants in aging community activities based on open-ended responses in activity evaluation forms. It adopts a macro–micro framework. Findings and issue domains from ILAS are used as a macro-level reference for key dimensions of healthy aging. Evaluation data from CAW serve as the micro-level focus to examine how these issues appear in concrete activities. The activities were elderly-oriented, but the evaluation forms were completed by all participants who were willing to respond. Therefore the dataset represents the broader response ecosystem surrounding aging community programmes, not exclusively elderly individuals. The co-creation literature argues that designing services with users, not just for them, helps surface latent needs and supports shared ownership of resulting interventions [15]. This study operationalises that principle by using participants’ own textual feedback as the starting point for programme topic design.

Evaluation forms were distributed in three CAW events under the Aging Wellness Talk (AWT) series, with themes co-developed by organisers and partner institutions. The first AWT, “Akademi Digital Lansia”, was organised in partnership with Masyarakat Anti Fitnah Indonesia (MAFINDO) on 16 June 2025. The second event on 20 September 2025, “Alzheimer Awareness Month: Peduli, Pahami, dan Dampingi”, was held in collaboration with Rumah Sakit Ubaya and Alzheimer’s Indonesia. The third AWT, on 30 October 2025, focused on “Enhancing Medication Safety: The Importance of Pharmacovigilance” with the Ministry of Health as partner. The fourth AWT, held on 20 December 2025, involved Alzheimer’s Indonesia and the Muslimat Nahdlatul Ulama as community partners, with the theme Workshop Personal Branding, Make up and Fashion Styling. Except for the third AWT, which targeted health professionals, the other three events primarily invited participants from various communities, including University of Surabaya (Ubaya) retirees, the Urban Community Empowerment Centre (PUSDAKOTA), non-Ubaya retirees, and the general public, with attendance ranging from approximately 30 to 80 participants per event. Evaluation forms were distributed in two formats: paper forms and a QR code linking to a Google Form, to accommodate participants with different levels of comfort with digital tools. Forms were handed out at the beginning of the event and collected at the end. Paper-based responses were manually entered into an Excel file and merged with online responses. The resulting dataset included identification variables (name, phone number, age, gender, education, and community of origin). Micro-level data consisted of open-ended items on aspects liked, aspects needing improvement, and suggested activity topics, as well as closed-ended items for satisfaction ratings (1–5) and willingness to be contacted for future AWT events. Macro-level data were represented by ILAS domains, used as a conceptual reference without individual-level ILAS analysis. Figure 1 shows the flow of research methodology for this research.

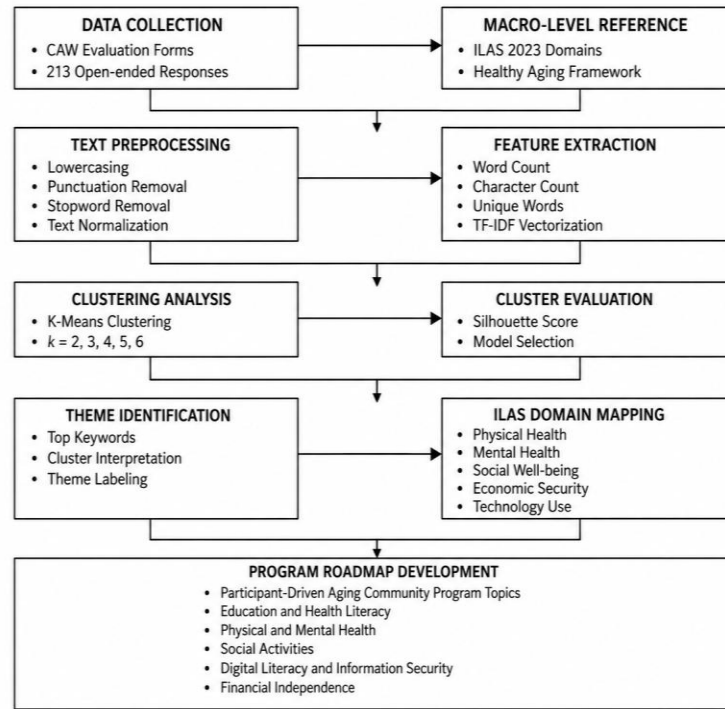


Figure 1. Research Methodology

Data processing began with NLP-based text pre-processing to prepare open-ended responses for further analysis. Standard NLP textbooks describe how tokenisation, normalisation, and vector-space models underpin downstream tasks such as document clustering and topic discovery [9], and survey work on text mining summarises standard pre-processing, TF-IDF representations, and clustering as core building blocks for transforming unstructured text into analysable features [16] [17]. The steps in this study included converting all text to lowercase, removing punctuation, digits, and special characters, eliminating non-informative stopwords, and performing spelling normalization so that different variants of the same word were treated as a single term. The cleaned text was retained at the level of each original column (liked aspects, aspects to be improved, suggested topics), and additionally linked together into a single combined text field for global thematic analysis. The cleaned outputs were stored in structured files to ensure consistent use in subsequent feature extraction.

In the feature extraction stage, open-ended responses were transformed into quantitative indicators. First, simple numerical features were computed for each response column, including word count, character count, and the number of unique words, to capture length and lexical diversity. Second, to represent semantic content, TF-IDF (Term Frequency-Inverse Document Frequency) features were generated from the combined text of the three open-ended columns, using parameters such as maximum number of features, minimum document frequency, and a predefined stopwords list. TF-IDF vector-space representations are a standard way to encode documents by weighting terms according to their frequency and discriminatory power across a corpus [13]. Text mining pipelines that combine pre-processing, vector representations, and unsupervised modelling are widely used to identify themes in large qualitative datasets in organisational research [18] [19]. TF-IDF and similar representations are also

commonly used in modern text mining workflows. Large-scale epidemiologic work has also shown that open-ended survey responses can be transformed into semantic representations, such as latent semantic space, to support automated clustering of health-related concerns [18]. The combination of numerical features and TF-IDF representations in this study provided the basis for clustering analysis.

Clustering was conducted using the K-Means algorithm as the primary unsupervised method to group participant responses without prior labels [21] [22] [23] [24]. Comprehensive reviews of text clustering algorithms position K-means with TF-IDF vectors as a widely used and computationally efficient baseline for document partitioning in high-dimensional spaces [25]. The K-means procedure used here follows the classical partitioning approach that iteratively assigns observations to clusters so as to minimise within-cluster variance [26]. Several values of  $k$  (e.g., 2 to 6 clusters) were tested, and the optimal configuration was selected based on a combination of silhouette scores and the substantive interpretability of the resulting themes in the context of aging activities. The silhouette score was used to assess clustering quality, with values ranging from -1 to 1. For each data point, the score compares its cohesion with its own cluster to its separation from the nearest alternative cluster; scores close to 1 indicate well-separated, coherent clusters, scores around 0 suggest borderline or overlapping clusters, and negative scores indicate potential misassignment. Clustering was primarily based on TF-IDF features so that clusters reflected semantic proximity in the textual content, with optional use of numerical features as a robustness check.

Following clustering, each cluster was interpreted and mapped onto meaningful themes. For each cluster, the most dominant keywords (top- $n$  terms) were extracted to identify recurring lexical patterns within that group of responses. Studies applying topic modelling to open-ended survey questions demonstrate that unsupervised text mining can compress rich qualitative feedback into coherent, interpretable themes to inform programme improvement [11] [27] and this study follows a similar logic using clustering rather than topic models. Based on the keyword sets, thematic labels were assigned to each cluster such as digital literacy, social interaction, physical and cognitive health, or service and activity design. Thematic or cluster labels were assigned based on dominant keywords and independently reviewed by two researchers familiar with aging-community programs. Discrepancies were discussed until consensus was reached.

These themes were then organised into a higher-level thematic map that grouped clusters into broader categories of elderly needs, facilitating programme design and priority setting. Integration with ILAS domains was achieved by mapping each CAW-derived thematic cluster onto issue domains identified in ILAS, such as physical health, mental health, social dimensions, economic security, and technology use. For instance, clusters related to digital literacy and information security were linked to technology use, whereas clusters focused on physical and cognitive health were aligned with health domains. This mapping allowed micro-level findings from CAW to be discussed in relation to national-level patterns observed in ILAS, either corroborating broader trends or enriching them with concrete examples of activity formats, preferences, and specific needs of elderly communities at the local level.

### **3 Results and Discussion**

The respondent demography can be viewed from two perspectives: all form entries (213 records) and unique respondents based on name (168 individuals). At the level of all entries, most forms were completed by women ( $\approx 78.9\%$ ; 168 entries), with men accounting for about 16.9% (36 entries) and a small number of entries without reported gender. Among those who reported age (138 entries), the mean age was

approximately 61.7 years, with a range of 21–86 years; respondents aged <60 years contributed about 38.4% of entries, those aged 60–69 years 34.1%, 70–79 years 26.1%, and ≥80 years around 1.4%. Educational backgrounds were diverse, with a substantial proportion missing (≈37.6%), followed by bachelor’s degree (S1, 22.5%), senior high school (SMA, 21.6%), master’s degree (S2, 9.9%), diploma (D3, 6.6%), and doctoral degree (S3, 1.9%). In terms of community origin, the largest contribution came from Muslimat NU (81 entries), followed by general participants (40 entries) and Ubaya alumni/retirees (38 entries), with other categories appearing in smaller numbers. When restricted to unique respondents (168 names), the demographic pattern is broadly similar. Women still dominate (≈81.0%; 136 individuals), with men representing about 19.0% (32 individuals). Among the 105 unique respondents who reported age, the mean age was approximately 62.1 years, with a relatively balanced distribution: <60 years (38.1%), 60–69 years (32.4%), 70–79 years (27.6%), and ≈1.9% aged ≥80 years. For higher education, about 39.9% remained missing, while those reporting senior high school (22.0%), bachelor’s degree (20.2%), master’s degree (8.3%), diploma (7.1%), and doctoral degree (2.4%) formed the remainder. Community origins among unique respondents were again dominated by Muslimat NU (51 individuals), general participants (42 individuals), and Ubaya alumni/retirees (34 individuals), with smaller groups such as “PUSDAKOTA community” and “Ubaya staff”.

Data processing began with a text pre-processing stage to prepare open-ended responses for analysis. The pipeline included converting all text to lowercase, removing punctuation, digits, and special characters, eliminating non-informative stopwords, and performing basic spelling normalisation so that different variants of the same word were treated as a single token. The text was retained at the level of each original column and also linked together into a single combined column for global thematic analysis. The cleaned text was then stored in structured files for consistent use in feature extraction. In terms of structure and completeness, there were 120 entries in the column “Things you liked about the activity”, 194 responses in “Things that could be improved”, and 63 entries in “Suggested activities/topics”. In the *Raw Data* sheet, these three columns were first combined into a single field named *Combinedtext*. In the *Lowercase* step, this combined text was converted to lowercase. The *NoPunctNum* step removed punctuation, digits, and special characters, leaving only letters and spaces, followed by whitespace normalisation. The *NoStopwords* step then removed common Indonesian stopwords. Finally, in the *Normalized* step, simple spelling normalisation was applied (e.g. *bnyk* → *banyak*, *utk* → *untuk*, *gak/nggak* → *tidak*, *sdh/udah* → *sudah*, *dr* → *dari*). The resulting dataset is stable and consistent for feature extraction and clustering, with minimal noise from case variation, punctuation, or highly variable spellings.

Open-ended responses were then transformed into quantitative indicators. First, simple numerical features were calculated for each response column: *word\_count* (number of words), *char\_count* (number of characters), and *unique\_words* (number of unique words), to capture length and lexical diversity. Second, to represent semantic content, TF–IDF features were generated from the combined text of the three columns, using parameters such as maximum number of features, minimum document frequency, and a predefined stopwords list. The combination of numerical features and TF–IDF formed the basis for the clustering analysis. For the simple *Fitur\_Numerik* stage, each entry received indicators for *word\_count*, *char\_count*, and *unique\_words*. In the “Things you liked” column, the average response length was 2.3 words (median 2, maximum 16), reflecting that many answers were very short phrases (e.g. “nambah ilmu”, “diskusi”, “narasumber bagus”). In the “Things that could be improved” column, the average length was 4.9 words (median 4, maximum 18), making it the longest column, as participants tended to write more when asked what should be improved. In

the “Suggested activities/topics” column, the average length was 1.2 words (median 0, maximum 10), indicating that only a minority of respondents proposed new topics, often in the form of a single keyword (e.g., “kesehatan”, “senam”, “digital”). For the *Combinedtext* column, the average length was 9.6 words (median 8, maximum 28). The mean number of unique words per response was 8.9, suggesting that despite their brevity, responses tended not to repeat the same word excessively and were more related to short lists of ideas. These numerical features imply that the clustering model should avoid overly high-dimensional representations to prevent overfitting, and that the “Things that could be improved” column is particularly informative for identifying “pain points” most strongly perceived by the participants.

Using the normalised *Combinedtext* column, TF-IDF features were then constructed in the *Fitur\_TFIDF* stage. The model used `max_features = 300`, yielding an effective set of about 200 salient terms, and `min_df = 2`, i.e., a word had to appear in at least two responses to be retained. These semantic features, together with the numerical indicators, were prepared for subsequent analysis. Examining average TF-IDF scores (excluding “nan”) showed that high-weight terms included *pengetahuan, kesehatan, mental, fisik, menjaga; teman, bersosialisasi, luang, waktu, ketrampilan, tambahan, memperoleh; finansial, bekerja, wiraswasta; and lansia, alzheimer*. This pattern suggests several global themes: enhancement of knowledge and health (activities perceived as a means to gain knowledge and maintain physical mental health); social interaction and meaningful use of leisure time (opportunities to socialise and engage in purposeful activities); skill and capacity building (interest in training that provides new skills, such as digital literacy or practical hobbies); financial and productive life in later years (concerns about economic independence and small scale entrepreneurship); cognitive issues and degenerative disease (visibility of Alzheimer’s/dementia prevention and education); and digital literacy and information security (concerns about online fraud and the need for digital competence).

Clustering was performed using the K-Means algorithm as the main unsupervised method to group responses without prior labels. Several values of *k* (2 to 6) were tested, and the best configuration was selected based on silhouette scores and the substantive interpretability of themes in the context of elderly activities. The silhouette score ranges from -1 to 1 and evaluates how well each data point fits within its assigned cluster relative to the nearest alternative cluster: values close to 1 indicate well separated, coherent clusters; values around 0 suggest fuzzy or overlapping cluster boundaries; and negative values indicate potential misassignments. Clustering was primarily based on TF-IDF features so that clusters reflected semantic proximity in the text, with numerical features used as an optional comparison. The clustering input was the *Data\_NonEmpty* subset, which contained only entries with non-empty text (`word_count > 0`), including respondent identifiers (No, Name) and *Gabungan\_Teks*, together with *Fitur\_Numerik\_NonEmpty* and *Fitur\_TFIDF\_NonEmpty* as the numerical and semantic feature sets. Table 1 presents the TF-IDF-based silhouette scores obtained from five K-Means clustering models with different numbers of clusters. The silhouette score was used to evaluate cluster quality and identify the most appropriate clustering solution for psychographic segmentation of tourist reviews.

The best-performing configuration was *k* = 6, which was selected as the main clustering solution. Although the corresponding silhouette score (0.157) is relatively low, such values are not uncommon for short and heterogeneous open-ended responses, where thematic boundaries tend to overlap and some responses may plausibly relate to more than one cluster. For this reason, silhouette score was used as one criterion for model selection, but not as the sole basis for choosing the final solution. The six-cluster

configuration was retained because it not only produced the highest silhouette score among the tested values of  $k$ , but also generated clusters that were more substantively interpretable and more useful for thematic mapping and programme planning. The TF-IDF-based clustering outputs were generated for all tested configurations (*TFIDF\_k2* to *TFIDF\_k6*), while *TFIDF\_Best\_k* summarised the selected solution ( $k = 6$ ) with respondent identifiers, combined text, and assigned cluster labels. A comparative analysis was also conducted using numerical-feature-based clustering with the same number of clusters.

Table 1. TF-IDF-Based Silhouette Score Evaluation

| Model   | Number of Clusters (k) | Cluster Labels     | Silhouette Score |
|---------|------------------------|--------------------|------------------|
| Model 1 | 2                      | {0, 1}             | 0.089            |
| Model 2 | 3                      | {0, 1, 2}          | 0.119            |
| Model 3 | 4                      | {0, 1, 2, 3}       | 0.141            |
| Model 4 | 5                      | {0, 1, 2, 3, 4}    | 0.155            |
| Model 5 | 6                      | {0, 1, 2, 3, 4, 5} | <b>0.157</b>     |

The thematic map derived from the clustering process shows that the needs and preferences of the elderly regarding CAW community activities are organised into several interrelated themes. The six thematic clusters have different sizes: Cluster 0 contains 26 responses (12.2%), Cluster 1 contains 76 responses (35.7%), Cluster 2 contains 19 responses (8.9%), Cluster 3 contains 39 responses (18.3%), Cluster 4 contains 33 responses (15.5%), and Cluster 5 contains 20 responses (9.4%). Figure 2 presents the distribution of respondents across these thematic clusters and helps identify which cluster represents the “critical mass” of elderly needs. Cluster 1, labelled “knowledge enhancement and appreciation of content”, is the largest cluster, indicating that the educational function of CAW (health literacy, new knowledge, perceived quality of content and speakers) is the primary reason participants in elderly-oriented communities attend its programmes.

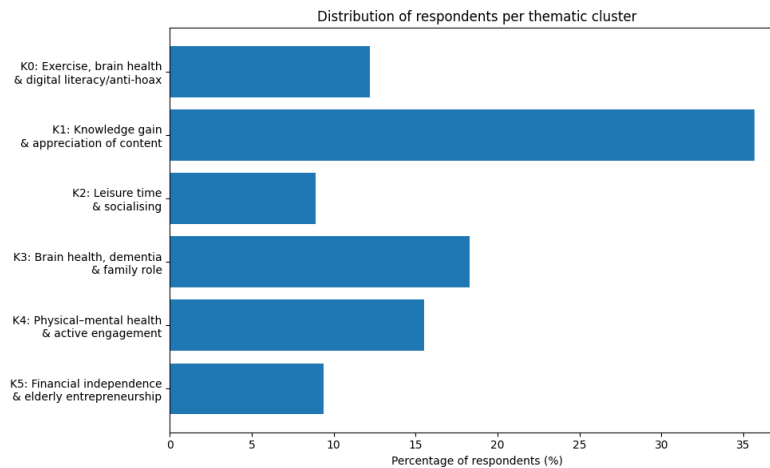


Figure 2. Distribution of respondents per thematic cluster

The other clusters complement this picture: Cluster 2 focuses on “filling leisure time and socialising”, Cluster 4 on “physical–mental health and active engagement”, Cluster 3 on “brain health, dementia and the role of family”, Cluster 0 combines “exercise, brain health and digital literacy/anti-hoax”, and Cluster 5 captures “financial independence and elderly entrepreneurship”. Taken together, each cluster reflects a concrete dimension of elderly needs such as knowledge, physical and cognitive health, social relations, digital safety, and economic aspects, that previously appeared scattered across individual open-ended responses. The heterogeneity of these clusters reinforces prior arguments that elderly users should not be treated as a homogeneous group, but as diverse contributors whose preferences ought to be explicitly mapped rather than assumed [8].

Based on top terms and thematic labels, these six clusters were further grouped into broader categories of needs: (1) Education and Health Literacy (including general knowledge, physical health and brain health/dementia), (2) Social Activities and Meaningful Use of Leisure Time, (3) Economic Independence and Productive Activities, and (4) Digital Literacy and Information Security.

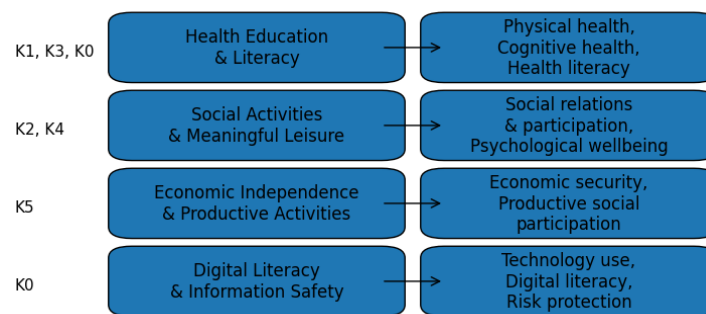


Figure 3. Thematic map of clusters linked to ILAS domains

Figure 3 shows the number of respondents falling into each of these broader categories, revealing that most participant’s responses concentrate on Education and Health Literacy, followed by social activities and physical–mental–cognitive health, while needs related to economic independence and digital literacy emerge as smaller but strategically important segments. This thematic map can be directly translated into a programme roadmap for CAW, for example: a health education and literacy strand (seminars, classes on nutrition, chronic disease, and dementia), a physical and mental health strand (regular exercise, brain gym, cognitively stimulating hobbies), a social and community strand (interest clubs, discussion groups, community gatherings), a digital literacy and safety strand (Digital Academy for the Elderly, anti-hoax classes, training on safe use of smartphones and digital services), and a financial independence and micro-entrepreneurship strand (home-based business training, simple financial management, elderly business groups). Community-based, technology-supported health services show how digital tools can support healthy aging in real-world settings [28]. The participatory approach used in these services also aligns with this study’s aim to design programme topics based on the preferences expressed by aging community members. Usability frameworks for mHealth tools further demonstrate that age-related sensory, physical, and cognitive changes introduce specific barriers that must be addressed when designing digital health solutions for the elderly [29], suggesting that

the digital literacy and safety strand in the CAW roadmap should be implemented with careful attention to these usability constraints. Linking usability barriers to age-related changes also implies that ongoing feedback from elderly users, such as the textual responses analysed here, is indispensable when designing digital extensions of community programmes. The roadmap is conceptual but practical. It can guide the annual CAW curriculum to balance education, physical and mental activities, social interaction, and economic support.

At the same time, this thematic map and roadmap are linked to the domains used in ILAS, thereby bridging the macro level of a national survey with the micro level of community voices. The category of Education and Health Literacy relates to physical health, cognitive health, and health literacy; Social Activities and Meaningful Leisure correspond to social relations, participation, and psychological well-being; Economic Independence and Productive Activities connect to economic security and productive social participation; while Digital Literacy and Information Security align with technology use, digital literacy, and protection from risk. This mapping shows that the need patterns emerging from CAW text responses are broadly consistent with major issues identified in ILAS, such as the burden of chronic disease, the importance of physical and cognitive activity, economic vulnerability, and the digital divide. At the same time, compared with the aggregated macro picture provided by ILAS, CAW data offer more concrete micro-level detail in the form of proposed activities such as brain gym sessions, anti-hoax classes, and home-based business training, thereby enriching understanding of how healthy ageing strategies can be operationalised at the community level. The participatory and data-driven approach adopted in this study is therefore valuable in connecting national policy priorities with grassroots programme design. However, several limitations should also be acknowledged. First, the dataset was based on a mixed sample, because the evaluation forms were collected from participants attending elderly-oriented community activities but were not limited exclusively to elderly respondents; thus, the findings may reflect the broader programme ecosystem rather than the preferences of elderly participants alone. Second, many open-ended responses were relatively short, limiting semantic richness and reducing the contextual information available for clustering [30] [31]. Third, the selected clustering solution produced a relatively low silhouette score, indicating that cluster separation was modest, which is consistent with the presence of overlapping themes and the interconnected nature of health, social, digital, and economic concerns in later life. Fourth, the study relies on K-Means clustering as a baseline method. Future studies may compare K-Means with alternative approaches such as Hierarchical Clustering, LDA Topic Modeling, BERTopic, or transformer-based embeddings to improve thematic separation and cluster quality. Finally, one of the analysed events primarily targeted health professionals rather than aging community members, which may have introduced additional variation into the dataset. Accordingly, the findings should be interpreted as exploratory and primarily useful for thematic mapping and programme design rather than as a definitive typology of elderly needs.

#### 4 Conclusion

This study shows that the needs and preferences of participants in aging community programmes are not homogeneous but cluster into several major themes: health education and literacy (physical, cognitive, dementia related), social activities and meaningful use of leisure time, economic independence and productive engagement, and digital literacy and information security. By combining NLP techniques with unsupervised text clustering, open-ended responses from evaluation forms were transformed into a structured thematic map and subsequently developed

into an operational roadmap for elderly programmes. When these thematic clusters are mapped onto the domains used in ILAS, the voices of participants in elderly-oriented community programmes at CAW both confirm and enrich national-level findings like the burden of chronic disease, the digital divide, and economic vulnerability. By translating them into highly concrete activity proposals at the community level, the study underscores the importance of data-driven social engineering grounded in the voices of the participants in aging community programmes textual feedback becomes both an administrative feedback mechanism and a basis for designing and adjusting programmes that are more targeted, sustainable, and aligned with the broader healthy aging agenda in Indonesia. Future studies may expand the dataset by including elderly communities from multiple regions and employ more advanced text-mining approaches such as topic modeling, BERTopic, or large language model-based semantic clustering. Comparative evaluation across different algorithms may provide deeper insights into thematic structures and improve clustering quality.

## References

1. Badan Pusat Statistik, *Population of Indonesia: Results of the 2010 Population Census*. Jakarta, Indonesia: Badan Pusat Statistik, 2012.
2. Badan Pusat Statistik, *Statistics of Elderly Population 2023*. Jakarta, Indonesia: Badan Pusat Statistik, 2024.
3. World Health Organization, *Ageing and Health*. Geneva, Switzerland: World Health Organization, 2022. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health>
4. World Health Organization, *WHO's Work on the UN Decade of Healthy Ageing (2021–2030)*. Geneva, Switzerland: World Health Organization, 2020. [Online]. Available: <https://www.who.int/initiatives/decade-of-healthy-ageing>
5. G. Dahlgren and M. Whitehead, *Policies and Strategies to Promote Social Equity in Health*. Stockholm, Sweden: Institute for Futures Studies, 1991.
6. Indonesia Longitudinal Aging Survey (ILAS) Team, *Indonesia Longitudinal Aging Survey (ILAS) 2023: Main Report*. Jakarta, Indonesia: Centre for Aging Studies, 2023.
7. Pemerintah Republik Indonesia, *Peraturan Presiden Republik Indonesia Nomor 88 Tahun 2021 tentang Strategi Nasional Kelanjutan Usia*, 2021. [Online]. Available: <https://peraturan.bpk.go.id/Details/178090/perpres-no-88-tahun-2021>
8. J. Vines, G. Pritchard, P. Wright, P. Olivier, and K. Brittain, "Inclusion of older adults in the research and design of digital technology," *Int. J. Environ. Res. Public Health*, vol. 16, no. 19, Art. no. 3718, 2019.
9. D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed., draft version, 2023. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/>
10. Indriati, M. A. Rahman, I. Arwani, and L. Muflikhah, "Classifying application user comments using the improved K-nearest neighbor and BM25F weighting methods," *J. Inf. Technol. Comput. Sci.*, vol. 9, no. 2, pp. 174–182, 2024, doi: 10.25126/jitecs.92622.
11. D. Buenaño-Fernández, S. Luján-Mora, and M. Divitini, "Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach," *IEEE Access*, vol. 7, pp. 168804–168818, 2019.
12. A. Nabilah, N. I. Sari, M. Afrina, and A. Ibrahim, "The sentiment analysis of Indonesian startup application reviews using TF-IDF + SVM and FastText: A comparative study," *J. Inf. Technol. Comput. Sci.*, vol. 10, no. 3, 2025, doi: 10.25126/jitecs.2025103807.
13. C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
14. La Gubu, D. Rosadi, and Abdurakhman, "A new approach for robust mean-variance portfolio selection using trimmed k-means clustering," *Ind. Eng. Manag. Syst.*, vol. 20, no. 4, pp. 782–794, 2021.
15. E. B.-N. Sanders and P. J. Stappers, "Co-creation and the new landscapes of design,"

- CoDesign*, vol. 4, no. 1, pp. 5–18, 2008.
16. S. Vijayarani and R. Janani, “Text mining: Open source tokenization tools, an analysis,” *Adv. Comput. Intell. Int. J.*, vol. 3, pp. 37–47, 2016.
17. Z. Dai, K. Li, H. Li, and X. Li, “An unsupervised learning short text clustering method,” *J. Phys.: Conf. Ser.*, vol. 1650, Art. no. 032090, 2020, doi: 10.1088/1742-6596/1650/3/032090.
18. V. B. Kobayashi, S. T. Mol, H. A. Berkens, G. Kismihók, and D. N. Den Hartog, “Text mining in organizational research,” *Organ. Res. Methods*, vol. 21, no. 3, pp. 733–765, 2017.
19. R. J. Wadnare, S. S. Sherekar, and V. M. Thakare, “Development of text clustering method with K-means for analysis of text data,” *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 7, no. 2, pp. 143–151, 2021, doi: 10.32628/CSEIT217237.
20. M. F. Ismail, T. A. Kurniawan, and S. R. Akbar, “Transforming requirement sentences in the form of requirement boilerplates to ontology using natural language processing,” *J. Inf. Technol. Comput. Sci.*, vol. 10, no. 2, 2025, doi: 10.25126/jitecs.2025102535.
21. La Gubu, E. Cahyono, Arman, H. Budiman, and M. K. Djafar, “Family of K-means clustering for robust mean-variance portfolio selection: A comparison of K-medoids, K-means, and fuzzy C-means,” *Ind. Eng. Manag. Syst.*, vol. 23, no. 3, pp. 342–356, 2024.
22. C. Lal, A. Ahmed, R. Siyal, S. K. Beejal, S. Aftab, and A. Hussain, “Text clustering using K-MEAN,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 4, pp. 2892–2897, 2021, doi: 10.30534/ijatcse/2021/371042021.
23. R. J. Wadnare, S. S. Sherekar, and V. M. Thakare, “Development of text clustering method with K-means for analysis of text data,” *Int. J. Res. Publ. Rev.*, vol. 7, no. 2, pp. 143–151, 2021.
24. D. Debaio, M. Yinxia, and Z. Min, “Analysis of big data job requirements based on K-means text clustering in China,” *PLoS ONE*, vol. 16, no. 8, Art. no. e0255419, 2021, doi: 10.1371/journal.pone.0255419.
25. C. C. Aggarwal and C. X. Zhai, “A survey of text clustering algorithms,” in *Mining Text Data*, C. C. Aggarwal and C. X. Zhai, Eds. New York, NY, USA: Springer, 2012, pp. 77–128.
26. J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, Berkeley, CA, USA, 1967, pp. 281–297.
27. J. P. Matruty, A. M. Adrian, and A. Angdresey, “Sentiment analysis of visitor reviews on star hotels in Manado City,” *J. Inf. Technol. Comput. Sci.*, vol. 8, no. 1, pp. 21–32, 2023, doi: 10.25126/jitecs.202381403.
28. B. Reeder et al., “A community-based, technology-supported health service for detecting and preventing frailty among older adults: A participatory design development process,” *J. Aging Soc. Policy*, vol. 25, no. 2, pp. 124–147, 2013.
29. G. A. Wildenbos, L. Peute, and M. Jaspers, “A framework for evaluating mHealth tools for older patients on usability,” *Telemed. e-Health*, vol. 21, no. 10, pp. 1–7, 2015.
30. S. Fatimi, C. El Saili, and L. Alaoui, “A framework for semantic text clustering,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, pp. 451–459, 2020, doi: 10.14569/IJACSA.2020.0110657.
31. K. Abdalgader, A. A. Matroud, and K. Hossin, “Experimental study on short-text clustering using transformer-based semantic similarity measure,” *PeerJ Comput. Sci.*, vol. 10, Art. no. e2078, 2024, doi: 10.7717/peerj-cs.2078.