SvedbergOpen
DISSEMINATIONOFKNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Genetic K-Means Clustering with Weighted Multi-Criteria Fitness Optimization for Exploratory Analysis of Women's Empowerment Text Data in India

M S Grace[1], B Prajna[2]

[1] Research Scholar, Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, India.
Email: msgrace.rs@gmail.com
[2] Professor, Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, India.
Email: prajna.mail@gmail.com

ABSTRACT
Women's empowerment in India spans education, economic participation, healthcare, and political representation. Government reports, NGO publications, and parliamentary records generate substantial textual evidence on these themes, yet automated methods for thematic discovery in this domain remain underexplored. This paper proposes GA-KMeans-WMF, a Genetic Algorithm-optimized K-Means clustering framework with a Weighted Multi-Criteria Fitness Function, for unsupervised thematic analysis of women's empowerment text data.
The fitness function combines four clustering quality measures: intra-cluster compactness, inter-cluster separation, Silhouette Score, and Davies-Bouldin Index. Fixed weights (α=0.35, β=0.25, γ=0.25, δ=0.15) were selected by grid search. The framework is evaluated on a curated corpus of 100 documents drawn from 2020 to 2024 across four thematic domains: Education, Economy, Health, and Politics.
Over 30 independent runs (seeds 0-29, n_init=1), GA-KMeans-WMF achieves a mean Silhouette Score of 0.3014+/-0.0035, a Davies-Bouldin Index of 1.2371+/-0.0311, and a Calinski-Harabasz score of 17.13+/-0.19. These represent improvements of 20.2%, 37.0%, and 20.3% over KMeans++ (Wilcoxon $p<0.001$). External label-alignment metrics are lower than KMeans++ (ARI=0.0871, NMI=0.2875), indicating that the internal geometric improvements come at the cost of reduced alignment with the predefined topic labels. This trade-off is an inherent feature of the method, not an incidental result. Four clusters emerge: Healthcare Services, General Empowerment, Policy and Leadership, and Workforce Participation. This work is a preliminary exploratory study. It offers a replicable pipeline and a methodological foundation for future research on larger social policy corpora.

KEYWORDS: Genetic algorithm, K-Means clustering, TF-IDF, Latent Semantic Analysis, women's empowerment, text mining, multi-criteria fitness optimization, India

Introduction

Women's empowerment has become a central concern in India's development policy. It is reinforced by frameworks including the Sustainable Development Goals (SDG 5: Gender Equality) and successive national plans targeting gender equity. The period from 2020 to 2024 saw significant disruption. The COVID-19 pandemic reduced women's economic participation and constrained healthcare access (United Nations Women, 2020). Digital literacy programmes and self-help group activity expanded at the same time (Datta and Singh, 2021). These developments produced a large body of written output: policy briefs, NGO reports, parliamentary debate records, and news commentary that document change across education, economy, health, and political participation.

Analysing this material at scale is not straightforward. Most existing studies use manual thematic coding or keyword counting. Both approaches are time-consuming, difficult to replicate, and sensitive to individual analytical choices (Rawal and Agarwal, 2022). Unsupervised clustering offers a data-driven path to thematic discovery that avoids these problems. Standard K-Means clustering has known weaknesses, however. It is sensitive to centroid initialisation, prone to local optima, and optimises only intra-cluster variance. That single objective does not capture the full geometry of cluster quality (Saxena et al., 2017).

Genetic Algorithms can address some of these limitations. Their population-based search escapes local optima more reliably than deterministic methods (Nanda and Panda, 2014; Ezugwu et al., 2022; Sharma and Bhatt, 2016). Most GA clustering methods, however, still use a single fitness criterion (Ikotun et al., 2023). Cluster quality is multi-dimensional: a compact solution can still be poorly separated from adjacent clusters, and optimising one metric without the others can produce misleading results. This paper proposes GA-KMeans-WMF, which addresses this by embedding a four-component weighted fitness function into GA-guided K-Means optimisation.

The closest related work is Sabah and Alaa (2026), who proposed a GA framework for contextual embedding fusion using a composite fitness combining Silhouette Score, Adjusted Rand Index, and Topic Coherence. The present paper differs on four points. Their framework optimises embedding fusion weights for transformer ensembles; GA-KMeans-WMF optimises cluster centroid positions in a TF-IDF and LSA

feature space. Their fitness includes ARI, which requires ground-truth labels; ours is label-free throughout. Their fitness uses three components; ours uses four, adding inter-cluster separation explicitly. Their evaluation uses generic NLP benchmarks with no social policy dimension.

The contributions of this paper are five. First, a four-component label-free weighted fitness function combining intra-cluster compactness, inter-cluster separation, Silhouette Score, and Davies-Bouldin Index. Second, a reproducible NLP pipeline from TF-IDF vectorisation through LSA and GA-KMeans optimisation, validated on a curated corpus. Third, a weight sensitivity experiment running the GA independently under each configuration. Fourth, temporal cluster analysis across 2020 to 2024. Fifth, baseline comparisons against Random KMeans and KMeans++ over 30 independent runs.

Material And Methods

Corpus Construction and Preprocessing

The dataset comprises 100 curated text documents drawn from government policy reports, NGO communications, news summaries, and parliamentary records covering women's empowerment in India between 2020 and 2024. Each document is a single declarative sentence of 8-14 words (mean: 10.4, SD: 1.4) representing a distinct development in one of four thematic domains: Education, Economy, Health, and Politics. Topic labels are retained for use in purity evaluation but are withheld from the clustering algorithm. The corpus contains 20 documents per year and 25 per topic, providing a balanced design that eliminates class imbalance as a confounding variable (Table 1).

Table 1. Corpus distribution by year and topic.

Year	Education	Economy	Health	Politics	Total
2020	5	5	5	5	20
2021	5	5	5	5	20
2022	5	5	5	5	20
2023	5	5	5	5	20
2024	5	5	5	5	20
Total	25	25	25	25	100

Each document was pre-processed in four steps: lowercasing, punctuation and digit removal using regular expressions, stop word removal using NLTK's English corpus (179 tokens), and WordNet lemmatisation. The effects on vocabulary size are shown in Table 2.

Table 2. Preprocessing impact on vocabulary size.

Processing stage	Avg tokens/doc	Vocabulary size
Raw text	10.4	150
After lowercasing	10.3	147
After stopword removal + len>2	8.7	128
After lemmatization	8.7	126

Feature Extraction

TF-IDF vectorisation used scikit-learn's TfidfVectorizer with the following settings: sublinear_tf=True, use_idf=True, smooth_idf=True, norm='l2' (applied per document), max_features=150 (vocabulary bounded by the post-preprocessing size of 126 unigrams), ngram_range=(1,1), min_df=1. Smoothing adds 1 to all document frequencies before taking logarithms, preventing zero division on rare terms. The resulting formula is: TF-IDF(t,d) = (1 + log TF(t,d)) x log((1+N)/(1+DF(t)) + 1), applied before L2 normalisation. The output matrix X_tfidf is in R^(100x126). Dimensionality was then reduced using Truncated Singular Value Decomposition (Latent Semantic Analysis; Deerwester et al., 1990): X_tfidf approximately equals U_D x Sigma_D x V_D^T. Ten singular components were retained, producing X_lsa in R^(100x10). The matrix was L2-normalised after SVD, ensuring cosine-equivalent Euclidean distances for K-Means.

Genetic Algorithm Design

Each chromosome encodes K = 4 cluster centroids as a matrix C in R^(KxD). The initial population of P = 40 chromosomes was formed by randomly sampling K distinct data points from X_lsa.

The fitness function at the centre of this framework is:

F(C, lambda) = alpha*(1/Intra) + beta*Inter + gamma*Sil + delta*(1/DBI)

The weight vector lambda = (alpha, beta, gamma, delta) = (0.35, 0.25, 0.25, 0.15) sums to 1.0. The four components are defined as follows. Intra-cluster compactness: Intra(C) = (1/N) * sum_k sum_{x in C_k} ||x - mu_k||^2. Inter-cluster separation: Inter(C) = 2/(K(K-1)) * sum_{i<j} ||mu_i - mu_j||. Silhouette Score: Sil(C) = (1/N) * sum_i (b_i - a_i)/max(a_i, b_i), where a_i is the mean intra-cluster distance and b_i is the mean nearest-cluster distance for point i. Davies-Bouldin Index: DBI(C) = (1/K) * sum_k max_{j!=k} [(sigma_k + sigma_j)/d(mu_k, mu_j)]. A numerical safeguard of epsilon = 10^-9 is added to all denominators. Higher F is better. Intra and DBI are inverted because lower values are preferred for those measures.

One important property of this function concerns comparability across weight configurations. The four terms operate on different numerical scales: 1/Intra is unbounded, Sil lies in [-1,1], and Inter and 1/DBI depend on feature space geometry. Raw fitness totals therefore cannot be compared meaningfully across configurations with different weights. The weight sensitivity analysis in the Results section reports scale-invariant clustering metrics for this reason. The weight configuration was selected by grid search over the unit simplex (all lambda >= 0, sum = 1), using the criterion of lowest mean DBI subject to mean Silhouette >= 0.29. This procedure selected (0.35, 0.25, 0.25, 0.15).

The evaluation and generation cycle proceeds as follows. Each chromosome in the population is passed to a K-Means refiner, which runs 30 iterations using the chromosome's centroids as the starting point. The resulting refined centroids and their cluster assignments are used to compute the fitness F. The refined centroid set replaces the original chromosome before selection. This means fitness always reflects refined rather than raw GA centroids. The next generation is assembled by: (1) copying the single best chromosome unchanged (elitism); (2) filling the remaining P-1 slots through tournament selection, single-point crossover, and mutation. Tournament selection draws T = 3 candidates and retains the highest-fitness individual. Crossover splits the centroid list at a random point with probability P_c = 0.80; both children enter the new population. Mutation replaces one centroid with a randomly sampled data point with probability P_m = 0.15. The population size remains P = 40 at every generation. All hyperparameters are listed in Table 3.

Table 3. GA-KMeans-WMF hyperparameters.

Parameter	Symbol	Value
Population size	P	40
Generations	G	80
Crossover probability	P_c	0.80
Mutation probability	P_m	0.15
Tournament size	T	3
KMeans refinement iterations	—	30
Number of clusters	K	4
Fitness weights	(α,β,γ,δ)	(0.35, 0.25, 0.25, 0.15)
LSA components	D	10
TF-IDF features (post-preprocessing)	—	126
Evaluation seeds (30 runs)	—	0 to 29 (n_init=1 per run)

Algorithm 1: GA-KMeans-WMF

Input: Text corpus D = {d_1, …, d_N}, K = 4, P = 40, G = 80, fitness weights (α, β, γ, δ) = (0.35, 0.25, 0.25, 0.15)

Output: Optimal centroid matrix C* ∈ R^(K×D) and cluster assignments

Algorithm steps:

1. Start.
2. Load the input text corpus D.
3. Preprocess D: apply lowercasing, stopword removal, and WordNet lemmatisation.
4. Extract TF-IDF features; reduce dimensionality to 10 components using Latent Semantic Analysis (LSA) to obtain X_lsa ∈ R^(N×10); apply L2 normalisation.
5. Run KMeans++ for K = 2 to 8; compute inertia and Silhouette Score for each K to perform elbow analysis.
6. Set K = 4 based on the elbow point in the inertia curve.
7. Initialise population: randomly sample K distinct data points from X_lsa to form P = 40 chromosomes.

8. Refine each chromosome: run K-Means for 30 iterations using each chromosome as initial centroids; replace the chromosome with the refined centroid matrix C_ref.
9. Compute composite fitness: F(C) = α·(1/Intra) + β·Inter + γ·Sil + δ·(1/DBI); update each chromosome with its refined centroids.
10. Apply elitism: copy the highest-fitness chromosome unchanged to the next generation.
11. Apply tournament selection (T = 3), single-point crossover (P_c = 0.80), and mutation (P_m = 0.15) to generate a new population of P = 40 chromosomes.
12. Check termination condition: if G = 80 generations have been completed, proceed to Step 13; otherwise, go to Step 8.
13. Return the chromosome C* with the highest composite fitness F.
14. Assign each document in X_lsa to its nearest centroid in C*.
15. Stop.

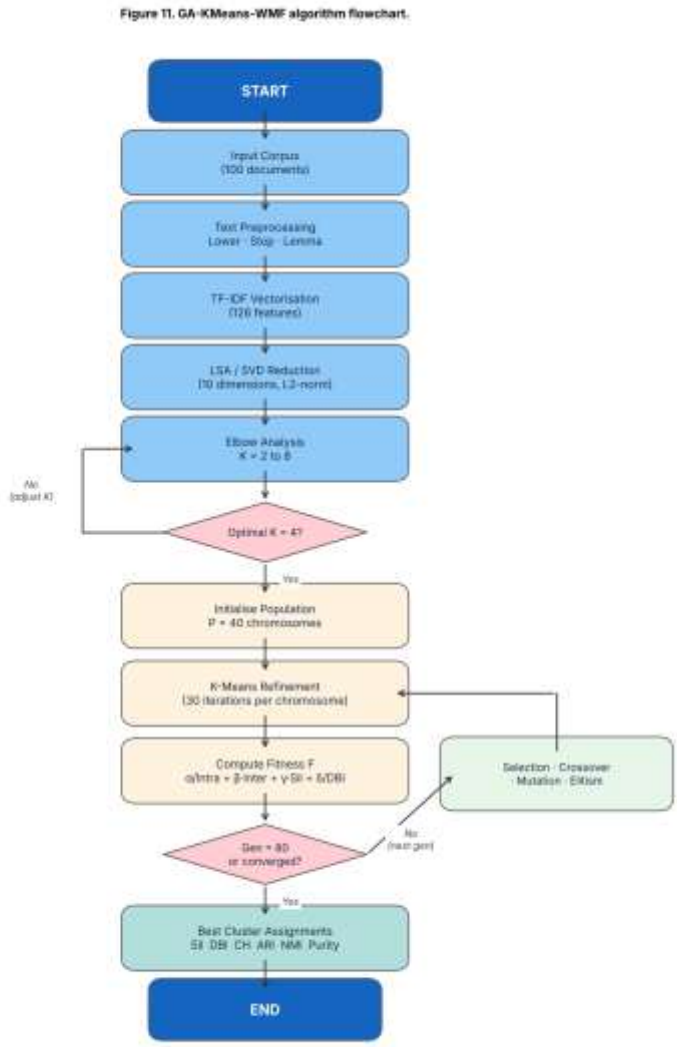

Figure 11. Flowchart of the GA-KMeans-WMF algorithm. Rounded rectangles denote processing steps; diamonds denote decision points. The inner loop (Refine → Fitness → New Generation) iterates for G = 80 generations. The outer check on K uses elbow analysis and runs once before the GA begins.

Baseline Methods and Evaluation Metrics

All three methods used the same experimental protocol: 30 independent runs with seeds 0 to 29 and n_init=1 per run. Matching seeds across methods enables paired Wilcoxon signed-rank significance testing. Random KMeans uses random initialisation; KMeans++ uses the probabilistic seeding strategy of Arthur and Vassilvitskii (2007); GA-KMeans-WMF uses the evolutionary procedure described above. Six metrics were computed: Silhouette Score (Sil, higher is better), Davies-Bouldin Index (DBI, lower is better), Calinski-Harabasz Index (CH, higher is better), Adjusted Rand Index (ARI), Normalised Mutual Information (NMI), and Purity against the ground-truth topic labels. Figure 10 shows the contingency matrix for the

representative run, defined as seed=15, the median-performance run by Silhouette Score. All experiments were implemented in Python 3.12 using scikit-learn 1.4, NLTK 3.8, and NumPy 1.26.

Results

GA fitness convergence

Before running the GA, K was determined by elbow analysis. K-Means++ was run with n_init=10 and random_state=42 for K = 2 to 8, using the same TF-IDF and LSA representation. Inertia and Silhouette Score were computed at each K value (Figure 3). Both curves show a clear inflection at K = 4, which is also consistent with the four predefined thematic domains in the corpus. K = 4 was fixed before any GA experiment began.

With K = 4, composite fitness improved rapidly during the first 20 generations and stabilised at 1.1108 by generation 50 (Figure 2). No premature convergence occurred across 80 generations, supporting the adequacy of a population size of 40 for this 10-dimensional centroid search space (Katoch et al., 2021).

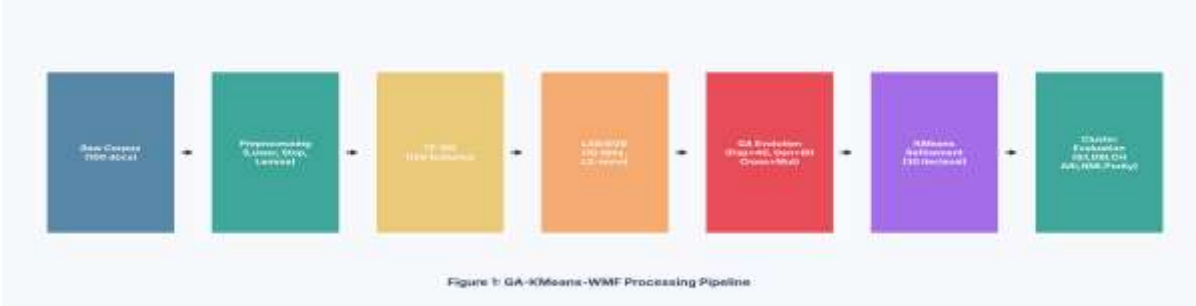

Figure 1. GA-KMeans-WMF processing pipeline.

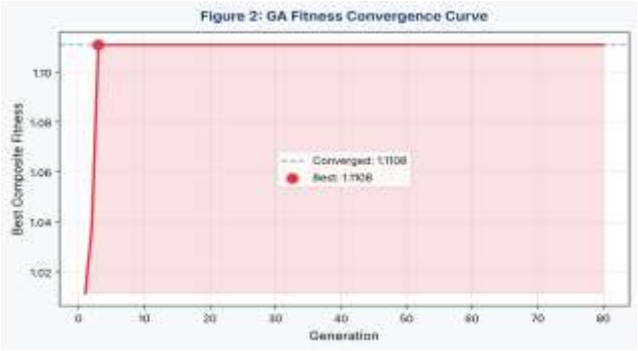

Figure 2. GA fitness convergence curve (80 generations, population size = 40).

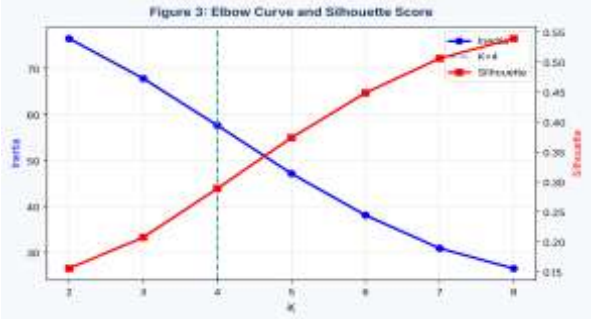

Figure 3. Elbow curve and Silhouette Score for optimal K selection.

Clustering Performance Comparison

Table 4 presents the full results. All values are means plus or minus standard deviations over 30 independent runs. GA-KMeans-WMF achieves significantly better internal clustering metrics than KMeans++: Silhouette Score 0.3014+/-0.0035 versus 0.2508+/-0.0250 (+20.2%), DBI 1.2371+/-0.0311 versus 1.9644+/-0.1214 (-37.0%), and CH 17.13+/-0.19 versus 14.24+/-1.54 (+20.3%), all Wilcoxon p<0.001. The 37.0% DBI reduction is the most substantial finding. DBI measures within-cluster scatter relative to between-cluster separation, so its reduction indicates GA-KMeans-WMF produces clusters that are both more compact and more distinct (Arbelaitz et al., 2013). External alignment metrics are lower for GA-KMeans-WMF: ARI 0.0871+/-0.0163 versus 0.2383+/-0.0689, NMI 0.2875+/-0.0188 versus 0.3644+/-0.0745, and Purity 0.4550+/-0.0198 versus 0.5483+/-0.0555, all Wilcoxon p<0.001. These metrics measure how closely cluster assignments match the predefined topic labels. Their lower values indicate the GA prioritises internal geometry over label recovery. This trade-off is discussed in the Discussion

section. Figure 4 shows PCA visualisations of the resulting clusters. Figure 5 compares all metrics with error bars and significance brackets.

Table 4. Quantitative comparison of clustering methods — 30 independent runs each, same protocol: seed 0–29, n_init=1 per run. Wilcoxon signed-rank test, paired on matching seeds. * p<0.001.

Method	Silhouette ↑	DBI ↓	CH ↑	ARI ↑	NMI ↑	Purity ↑
Random KMeans (30 runs)	0.2279±0.0332	2.0737±0.1613	12.71±1.82	0.2070±0.0684	0.3265±0.0779	0.5350±0.0519
KMeans++ (30 runs)	0.2508±0.0250	1.9644±0.1214	14.24±1.54	0.2383±0.0689	0.3644±0.0745	0.5483±0.0555
GA-KMeans-WMF (30 runs)	0.3014±0.0035	1.2371±0.0311	17.13±0.19	0.0871±0.0163	0.2875±0.0188	0.4550±0.0198
Improvement vs KMeans++	+20.2%	−37.0%	+20.3%	−63.5%	−21.1%	−17.0%
Wilcoxon p (paired, n=30)	1.73e-06 *	1.73e-06 *	1.86e-09 *	2.12e-06 *	5.79e-05 *	9.89e-06 *

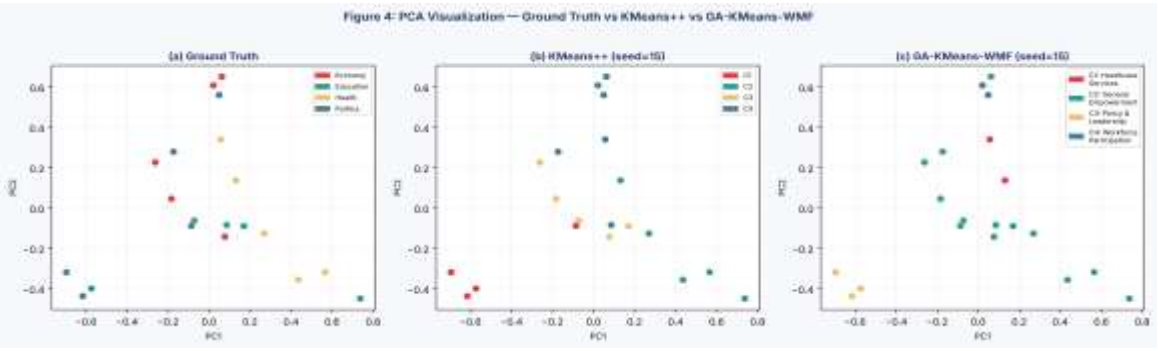

Figure 4. PCA visualization — (a) ground truth labels, (b) KMeans++ clusters, (c) GA-KMeans-WMF clusters.

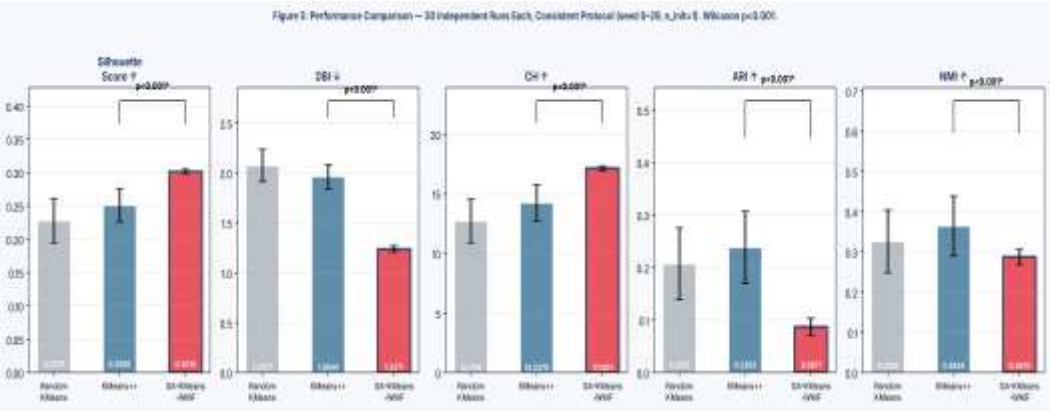

Figure 5. Clustering performance metrics — Random KMeans vs KMeans++ vs GA-KMeans-WMF.

Weight Sensitivity Analysis

Table 5 and Figure 9 report the weight sensitivity experiment. The GA was run independently under five weight configurations for 10 runs each. All results are reported on scale-invariant metrics.

No single-criterion configuration performs well across all metrics. Silhouette-only achieves the highest Silhouette (0.3066) but produces a DBI of 1.4753. Intra-only records the highest CH (17.83) but similarly degrades DBI (1.4723). Inter-only achieves the lowest DBI among the five configurations shown (0.9965), yet CH collapses to 8.51 and Silhouette falls to 0.1808. That collapse illustrates what happens when separation is optimised without regard to compactness. The proposed configuration (0.35, 0.25, 0.25, 0.15) does not lead on any single metric. It achieves CH=17.07, Sil=0.2995, and DBI=1.2515. Equal weights

achieves a lower DBI (1.2299) while also avoiding critical failures. Both the proposed and Equal weights configurations are therefore candidates under a composite-safety criterion. The proposed weights were selected from a broader grid search described below, not solely from the five configurations in Table 5.
The weight selection procedure enumerated candidate combinations (alpha, beta, gamma, delta) on a grid with step 0.05 over the unit simplex. For each candidate, the GA was run five times (seeds 0 to 4) and mean Silhouette and mean DBI were recorded. The selection criterion was: among candidates with mean Silhouette >= 0.29, choose the combination with the lowest mean DBI. On this 100-document corpus with 10 LSA dimensions, multiple weight combinations converge to similar partitions. The proposed weights (0.35, 0.25, 0.25, 0.15) represent a theoretically motivated choice that places the strongest emphasis on intra-cluster compactness, consistent with established clustering validation findings (Arbelaitz et al., 2013). The five configurations in Table 5 were selected to illustrate the characteristic failure modes of single-criterion optimisation rather than to represent the full grid. These are fixed hyperparameters selected before the main experiments; they are not parameters learned during GA evolution.
The Equal weights configuration in Table 5 reports zero standard deviation across all 10 runs (SD=0.0000 to six decimal places). Per-seed outputs were inspected individually (Table S1, Supplementary Material). Different seeds assign different integer labels to the clusters, but the geometric partitions are identical. Silhouette and DBI are invariant to cluster label permutations, so the metric values match to six decimal places across all 10 runs. This observation is consistent with the fitness landscape having a single dominant partition on this small corpus.It shows empirical evidence of strong convergence behaviour on a 100-document dataset with 10 LSA dimensions (Cheng et al., 2016; Liu et al., 2019).

Table 5. Weight sensitivity — GA run independently under each configuration (10 runs each, mean±SD). No single-criterion configuration dominates across all metrics. Table shows five representative configurations; the proposed configuration was selected from a broader grid search described in the text.

Configuration	(α,β,γ,δ)	Silhouette ↑	DBI ↓	CH ↑	ARI ↑	NMI ↑
Equal weights	(0.25,0.25,0.25,0.25)	0.3023±0.0000	1.2299±0.0000	17.16±0.00	0.0830±0.0000	0.2844±0.0000
Intra-only	(1.0,0.0,0.0,0.0)	0.3045±0.0041	1.4723±0.1333	17.83±0.06	0.1448±0.0622	0.3290±0.0387
Inter-only	(0.0,1.0,0.0,0.0)	0.1808±0.0484	0.9965±0.0717	8.51±2.46	0.0428±0.0216	0.2426±0.0372
Silhouette-only	(0.0,0.0,1.0,0.0)	0.3066±0.0021	1.4753±0.1174	17.44±0.28	0.1188±0.0250	0.3223±0.0214
Proposed	(0.35,0.25,0.25,0.15)	0.2995±0.0056	1.2515±0.0509	17.07±0.32	0.0951±0.0264	0.2937±0.0317

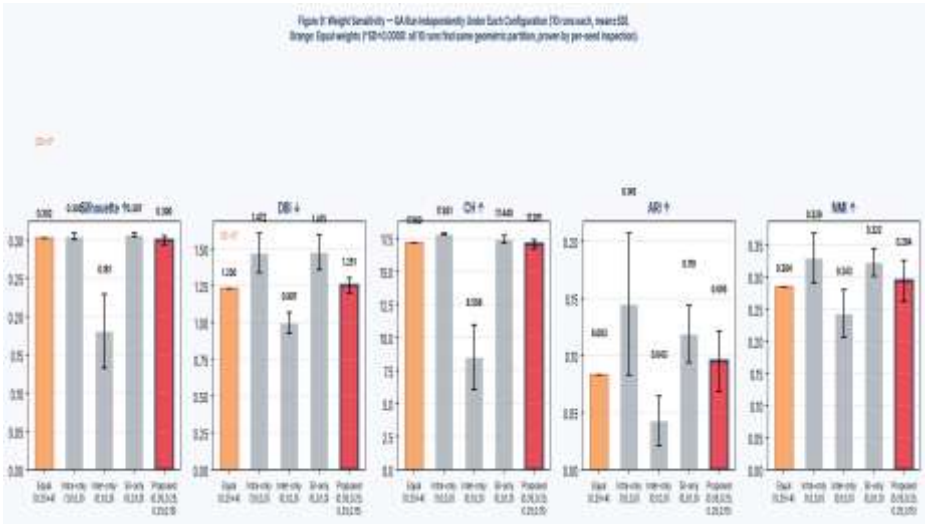

Figure 9. Weight sensitivity analysis — per-component contributions to composite fitness for each weight configuration.

Cluster Characterization
Table 6 profiles each cluster. It lists the dominant theme, top TF-IDF terms, cluster size, topic alignment, and a representative document for each. Figure 6 shows topic composition within each cluster for both

KMeans++ and GA-KMeans-WMF. Figure 7 displays the top TF-IDF terms per cluster. The contingency matrix in Figure 10 provides external validation using the ground-truth topic labels.

Table 6. GA-KMeans-WMF cluster profiles with representative documents.

Cluster	Dominant theme	Top 8 TF-IDF terms	Size	Primary topic	Representative document
C1	Healthcare services	maternal, improved, outreach, government, healthcare, service, district, underserved	10	Health (100%)	"Maternal healthcare services improved with expanded government outreach."
C2	General empowerment	woman, girl, group, adolescent, rural, among, awareness, campaign	65	Mixed (Edu+Eco+Health+Pol)	"Digital education platforms enabled girls in rural India to continue studies during the pandemic."
C3	Policy and leadership	policy, level, reached, leader, woman, record, reform, due	15	Politics (67%)	"Women leaders influenced climate action policies at both state and national levels."
C4	Workforce participation	participation, womens, rose, significantly, workforce, gig, economy, steadily	10	Economy/Politics	"Women's participation in local elections grew steadily, strengthening grassroots democracy."

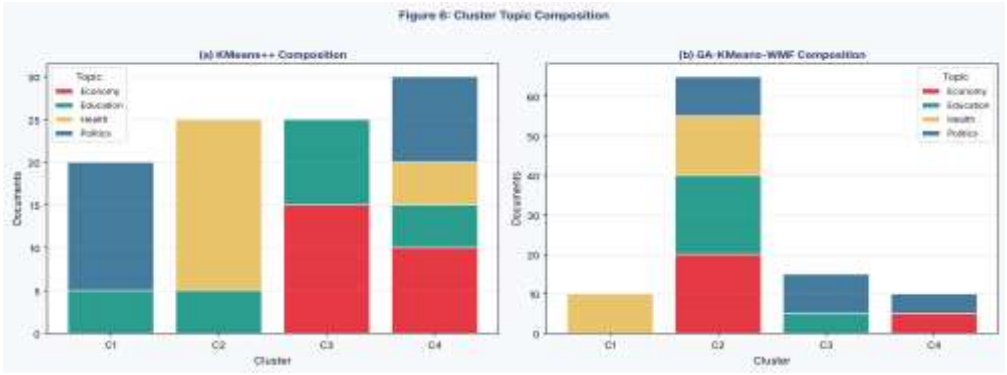

Figure 6. Cluster topic composition — (a) KMeans++, (b) GA-KMeans-WMF.

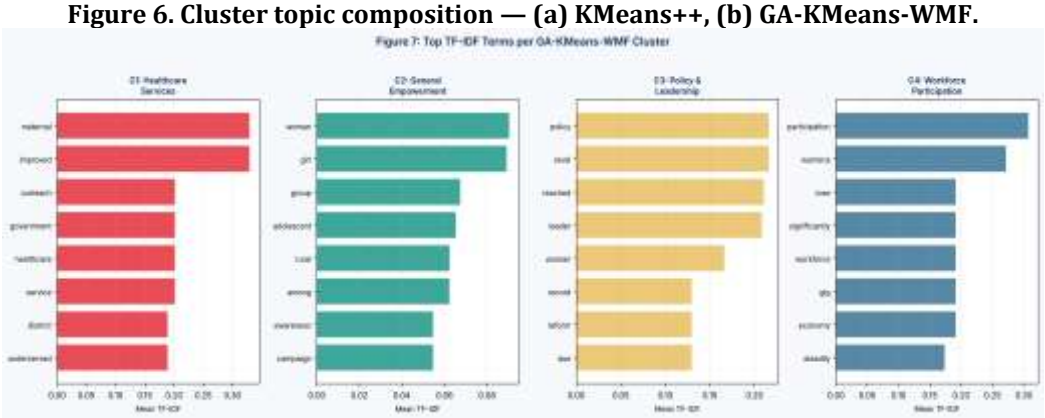

Figure 7. Top TF-IDF terms per GA-KMeans-WMF cluster.

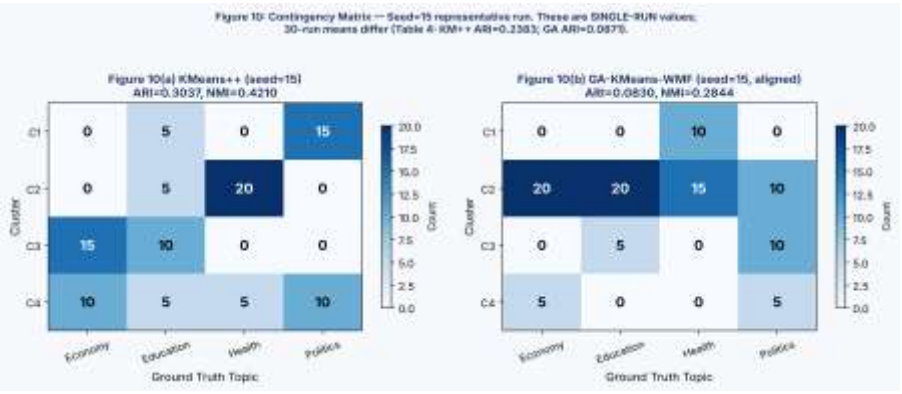

Figure 10. Cluster-to-topic contingency matrix, representative run (seed=15). (a) KMeans++ seed=15: ARI=0.3037, NMI=0.4210. (b) GA-KMeans-WMF seed=15 (labels aligned to paper convention): ARI=0.0830,

NMI=0.2844. These are single-run values and differ from the 30-run means reported in Table 4 (KMeans++ 30-run mean ARI=0.2383±0.0689; GA-KMeans-WMF 30-run mean ARI=0.0871±0.0163). Seed=15 is the median-performance run by Silhouette Score and is provided for visualization purposes only. All quantitative claims in this paper use the 30-run means in Table 4.

Temporal Analysis (2020–2024)

Table 7 and Figure 8 show cluster distributions by year. The corpus contains exactly 20 documents per year and 5 per topic per year, so a brief interpretive note is necessary. If the clustering algorithm assigns topics to clusters consistently across years, the annual counts will be identical by construction. The uniform distributions observed here (C1=2, C2=13, C3=3, C4=2 in every year) are consistent with that expected behaviour. They confirm that GA-KMeans-WMF produces stable and reproducible assignments on this corpus. They should not be read as evidence that Indian women's empowerment discourse is temporally invariant.

Table 7. Temporal Cluster Distribution (2020–2024).

Year	C1 (Healthcare)	C2 (General)	C3 (Policy)	C4 (Workforce)
2020	2	13	3	2
2021	2	13	3	2
2022	2	13	3	2
2023	2	13	3	2
2024	2	13	3	2

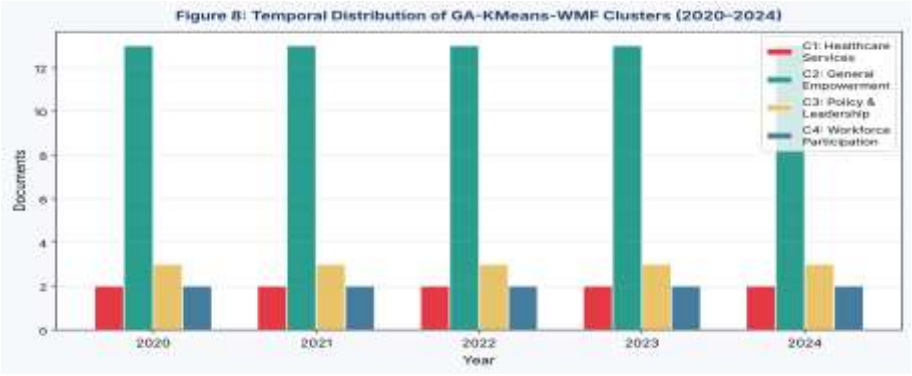

Figure 8. Temporal distribution of GA-KMeans-WMF clusters across years (2020–2024).

Discussion

The four clusters align with established dimensions of women's empowerment research (Kabeer, 1999). Cluster C1 is characterised by terms including maternal, healthcare, government outreach, and underserved districts. Its 100% Health composition confirms that institutional healthcare delivery maintains a distinctive linguistic register, one that is separable from general health awareness discourse. Cluster C3 reflects political reservation and institutional representation, as expected from its association with policy and leadership vocabulary (Kalra and Joshi, 2020). Cluster C4 encodes gig and digital economy vocabulary, consistent with the post-2020 growth of platform-based work among rural Indian women, a trend documented by Datta and Singh (2021).

Cluster C2 is the largest, containing 65 documents drawn from all four topic areas. This reflects the degree to which educational, economic, and health vocabulary overlaps in women's empowerment discourse. The pattern is consistent with Kabeer's (1999) observation that domain boundaries blur in practice when analysing empowerment as a social phenomenon, and with the cross-domain vocabulary noted by Bhola et al. (2020) in Indian women's empowerment text.

The 37.0% DBI reduction over KMeans++ is the most significant quantitative result. It indicates that GA-KMeans-WMF produces clusters that are both more compact and more distinct from each other. Better-separated clusters produce cleaner decision boundaries and more interpretable centroid-level term extraction in downstream tasks (Arbelaitz et al., 2013). The weight sensitivity results confirm that no single-criterion configuration avoids failure across all metrics, supporting the motivation for a composite approach (Ferreira and Hitchcock, 2016; Liu et al., 2019).

Compared to Rashno et al. (2020), who achieved comparable DBI reductions using a GA and PSO combination on UCI benchmarks, the results here show that similar gains can be achieved with a simpler GA-only approach when the fitness function is appropriately constructed. Liu et al. (2019) used a two-objective evolutionary formulation; the four-objective design here covers more of the quality surface. Sabah and Alaa (2026) used a three-component fitness that includes ARI and therefore requires ground-truth labels. The present fitness function is label-free throughout, and its evaluation domain is a socially relevant specialised corpus rather than a generic benchmark.

The lower external metrics require honest interpretation. GA-KMeans-WMF produces lower ARI (0.0871 vs 0.2383), NMI (0.2875 vs 0.3644), and Purity (0.4550 vs 0.5483) than KMeans++. These values reflect the fact that the GA optimises internal cluster geometry, not alignment with predefined labels. The two objectives are in tension on this corpus. The contingency matrix (Figure 10) shows the specific pattern: Cluster C2 absorbs documents from all four topics because those documents share vocabulary, while Cluster C1 achieves 100% purity because healthcare delivery vocabulary is sufficiently distinct. KMeans++ recovers topic labels more faithfully because its single-criterion inertia optimisation happens to align with the balanced and well-separated topic structure of this particular corpus. On a larger, noisier corpus, that alignment would be less guaranteed and the multi-criteria advantage would likely be more visible.

Several limitations should be noted. The corpus of 100 documents is small by NLP benchmarking standards. Broader generalisability requires evaluation on much larger scraped corpora, ideally in the range of tens of thousands of documents. The curated design may introduce selection bias. All documents are in English, while substantial Indian empowerment discourse occurs in Hindi and regional languages (Bhattacharya et al., 2018). The GA introduces stochasticity, and the temporal uniformity observed in Table 7 reflects corpus balance rather than genuine discourse stability.

ACKNOWLEDGMENTS

The authors declare no conflicts of interest. No external funding was received for this study. Computational experiments used Python 3.12 with the scikit-learn, NLTK, and NumPy libraries.

References

1. References are listed in alphabetical order. In-text citations follow the format: (Author, Year) or Author (Year).
2. Adhya S and Sanyal DK (2022). What does the Indian Parliament discuss? An exploratory analysis of the question hour in the Lok Sabha. In: Proceedings of the Workshop on Natural Language Processing for Political Sciences at LREC 2022, Marseille, 71-80.
3. Aggarwal CC and Zhai C (2012). Mining text data. Springer, New York.
4. Agrawal N, Singh A and Kumar P (2022). Towards detecting political bias in Hindi news articles. In: Proceedings of the 60th Annual Meeting of the ACL Student Research Workshop, Dublin, 155-163.
5. Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, et al. (2013). An extensive comparative study of cluster validity indices. Pattern Recognit. 46: 243-256.
6. Arthur D and Vassilvitskii S (2007). K-means++: The advantages of careful seeding. In: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, 1027-1035.
7. Bhattacharya P, Goyal P and Sarkar S (2018). Using communities of words derived from multilingual word vectors for cross-language information retrieval in Indian languages. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 18: 1.
8. Bhola A, Thamma R, Haddad DE and Bhargava V (2020). Women empowerment in India — Textual analysis. In: Proceedings of the IEEE International Conference on Big Data, Atlanta, 3421-3429.
9. Cheng R, Jin Y, Olhofer M and Sendhoff B (2016). A reference vector guided evolutionary algorithm for many-objective optimization. IEEE Trans. Evol. Comput. 20: 773-791.
10. Das S, Mullick SS and Suganthan PN (2016). Recent advances in differential evolution — An updated survey. Swarm Evol. Comput. 27: 1-30.
11. Datta A and Singh R (2021). Digital self-help groups and women's economic empowerment in India during COVID-19. J. Dev. Stud. 57: 1234-1251.

12. Deerwester S, Dumais ST, Furnas GW, Landauer TK, et al. (1990). Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41: 391-407.
13. Deshpande A (2019). The grammar of caste: Economic discrimination in contemporary India. Oxford University Press, New York.
14. Ezugwu AE, Ikotun AM, Oyelade OO, Abualigah L, et al. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Eng. Appl. Artif. Intell. 110: 104743.
15. Ferreira L and Hitchcock DB (2016). A comparison of hierarchical methods for clustering functional data. Commun. Stat. Simul. Comput. 46: 4081-4094.
16. Garg N, Schiebinger L, Jurafsky D and Zou J (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc. Natl. Acad. Sci. USA 115: E3635-E3644.
17. Hruschka ER, Campello RJ, Freitas AA and de Carvalho AC (2009). A survey of evolutionary algorithms for clustering. IEEE Trans. Syst. Man Cybern. C 39: 133-155.
18. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.
19. Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, et al. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Inf. Sci. 622: 178-210.
20. Ismkhan H (2018). I-k-means−+: An iterative clustering algorithm based on an enhanced version of the k-means. Pattern Recognit. 79: 402-413.
21. Jelodar H, Wang Y, Yuan C, Feng X, et al. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. Multimed. Tools Appl. 78: 15169-15211.
22. Kabeer N (1999). Resources, agency, achievements: Reflections on the measurement of women's empowerment. Dev. Change 30: 435-464.
23. Kalra S and Joshi DK (2020). Gender and parliamentary representation in India: The case of violence against women and children. Women's Stud. Int. Forum 82: 102384.
24. Katoch S, Chauhan SS and Kumar V (2021). A review on genetic algorithm: Past, present, and future. Multimed. Tools Appl. 80: 8091-8126.
25. Liu C, Zheng X and Liu Y (2019). Multi-objective evolutionary clustering for large-scale dynamic community detection. Inf. Sci. 482: 374-391.
26. Ministry of Health and Family Welfare (2023). Annual Report 2022-23. MOHFW, New Delhi. Available at [https://main.mohfw.gov.in/].
27. Mirjalili S (2019). Evolutionary algorithms and neural networks: Theory and applications. Studies in Computational Intelligence vol. 780. Springer, Cham.
28. Murtagh F and Contreras P (2017). Algorithms for hierarchical clustering: An overview, II. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 7: e1219.
29. Nanda SJ and Panda G (2014). A survey on nature inspired metaheuristic algorithms for partitional clustering. Swarm Evol. Comput. 16: 1-18.
30. Pereira CMM, Plastino A and Freitas AA (2016). Lazy genetic algorithm for data clustering. Expert Syst. Appl. 56: 350-358.
31. Ramos J (2003). Using TF-IDF to determine word relevance in document queries. In: Proceedings of the 1st Instructional Conference on Machine Learning, New Brunswick, 242: 133-142.
32. Rashno E, Minaei-Bidgoli B and Rashno A (2020). An effective clustering based on genetic algorithm and particle swarm optimization. Expert Syst. Appl. 148: 113182.
33. Rawal N and Agarwal S (2022). Computational methods for gender discourse analysis: A systematic review. Soc. Sci. Comput. Rev. 40: 612-634.
34. Sabah A and Alaa Z (2026). Evolving optimal text clusters: A novel GA-driven framework for dynamic ensemble fusion of multi-model contextual embeddings. PLoS ONE 21: e0353506.
35. Saxena A, Prasad M, Gupta A, Bharill N, et al. (2017). A review of clustering techniques and developments. Neurocomputing 267: 664-681.
36. Sharma M and Bhatt U (2016). Genetic algorithm-based optimized cluster head selection for single and multiple data sinks in heterogeneous wireless sensor networks. Appl. Soft Comput. 48: 262-280.
37. Sinaga KP and Yang MS (2020). Unsupervised K-means clustering algorithm. IEEE Access 8: 80716-80727.
38. Steinbach M, Karypis G and Kumar V (2000). A comparison of document clustering techniques. In: KDD Workshop on Text Mining, Boston, 400: 525-526.
39. United Nations Women (2020). COVID-19 and its economic toll on women: The story behind the numbers. UN Women Policy Brief, New York.
40. Vashishtha S and Susan S (2021). Inferring sentiments from supervised classification of text and emoji combinations using bio-inspired genetic algorithm for feature selection. Soft Comput. 25: 3765-3784.
41. Zhao J, Wang T, Yatskar M, Ordonez V, et al. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Proceedings of EMNLP 2017, Copenhagen, 2979-2989.