International Journal of Bio-resource and Stress Management https://ojs.pphouse.org/index.php/IJBSM

Crossref

IJBSM September 2026, 17(9): 01-10

Article AR7245

Research Article

Natural Resource Management

DOI: HTTPS://DOI.ORG/10.23910/1.2026.7245

Topological Mapping of Organised Fruit and Vegetable Retail Discourse in West Bengal: A Computational Corpus Analytics Approach

Soumyadeep Thakur, Ananta Mondal, Srestha Mitra and Debabrata Basu

Dept. of Agricultural Extension, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, Nadia, West Bengal (741 252), India

Open Access

Corresponding soumyadeepthakur519@gmail.com

0009-0008-3637-6380

ABSTRACT

This study, conducted in February, 2026, investigated the semantic structure of organised fruit and vegetable retail chains in West Bengal using Natural Language Processing (NLP) and Topological Data Analysis (TDA). The rapid expansion of organised retail chains in India significantly transformed food marketing systems, particularly in regions where traditional and modern retail formats coexisted. Understanding the interaction among product quality, consumer perception, branding strategies, and supply chain dynamics within retail discourse, therefore, became increasingly important. A structured textual corpus describing store characteristics, product range, consumer perceptions, and improvement strategies was processed through a systematic preprocessing framework. Contextual embeddings were generated using transformer-based language models, producing 768-dimensional semantic vectors that captured contextual relationships within the corpus. These embeddings were reduced using Uniform Manifold Approximation and Projection (UMAP), followed by the application of the Mapper algorithm to construct a topological network of semantic clusters. The resulting Mapper graph revealed a thematically segmented yet structurally interpretable discourse network with distinct conceptual clusters and limited transitional overlap. Network refinement identified a central connected component representing the intersection of branding strategies, product positioning, consumer evaluation, supply relationships, and market dynamics. Topic modelling, coherence analysis, hyperparameter stability assessment, and spectral analysis confirmed the robustness and interpretability of the semantic structure. The findings indicated that organised retail discourse was predominantly consumer-centric, emphasising product quality, pricing perceptions, and brand-related considerations. The study demonstrated the effectiveness of integrating NLP and topological modelling to uncover hidden semantic structures in retail-related text and provided valuable insights for retail managers, researchers, and policymakers.

KEYWORDS: Organized retail, NLP, topological data analysis, consumer perception

Citation* (VANCOUVER)*: Thakur et al., Topological Mapping of Organised Fruit and Vegetable Retail Discourse in West Bengal: A Computational Corpus Analytics Approach. *International Journal of Bio-resource and Stress Management,* 2026; 17(9), 01-10. HTTPS://DOI.ORG/10.23910/1.2026.7245.

Data Availability Statement: Legal restrictions are imposed on the public sharing of raw data. However, author have full right to transfer or share the data in raw form upon request subject to either meeting the conditions of the original consents and the original research study. Further, access of data needs to meet whether the user complies with the ethical and legal obligations as data controllers to allow for secondary use of the data outside of the original study.

Conflict of interests: The author has declared that no conflict of interest exists.

RECEIVED on 03rd June 2026 RECEIVED in revised form on 22nd August 2026 ACCEPTED in final form on 09th September 2026 PUBLISHED on 15th September 2026

1. INTRODUCTION

The rapid expansion of organised retail has significantly transformed the structure of contemporary markets, particularly in emerging economies where traditional retail systems continue to coexist with modern retail formats (Angelopoulos et al., 2023; Mohammadi et al., 2024). Organised retail chains such as Reliance Mart, Big Bazaar, and Vishal Mart in India have increasingly influenced consumer behaviour, branding practices, supply chain organisation, and market competition (Munoz-Leiva et al., 2021; Kuikka et al., 2024). This transformation has not only changed the way products are marketed and distributed but has also reshaped the relationships among producers, retailers, brands, and consumers. In states such as West Bengal, where both traditional and organised retail systems operate simultaneously, the retail environment has become highly dynamic and interconnected (Angelopoulos et al., 2023). As a result, understanding how different retail-related concepts interact within this evolving ecosystem has become increasingly important for researchers, policymakers, and business practitioners (Tudor et al., 2025). Despite the growing importance of organised retail systems, much of the existing research has mainly depended on conventional analytical approaches such as surveys, case studies, and descriptive methods (Mahr et al., 2019). These approaches have undoubtedly provided valuable insights into retail operations, consumer preferences, and market trends. However, they often remain limited in their ability to capture the deeper structural relationships embedded within large volumes of textual information related to retail markets (Ibrahim and Wang, 2019). Retail discourse is inherently complex because multiple themes such as consumer perception, product quality, branding strategies, pricing, and supply chain management interact simultaneously within the same communication environment. Traditional analytical methods frequently examine these themes separately, making it difficult to understand how they are structurally connected within broader discourse systems. In recent years, the rapid growth of digital textual data, including reports, market analyses, policy discussions, online content, and research articles, has created new opportunities for studying retail systems through computational approaches (Al-Azmi, 2013; Ibrahim and Wang, 2019). Large-scale text analysis enables researchers to identify recurring themes, hidden conceptual relationships, and patterns embedded within complex discourse structures (Nahr and Heikkila, 2022; Sheikhattar et al., 2022). Nevertheless, many traditional text mining approaches primarily focus on frequency-based analysis or isolated topic extraction (Liu et al., 2022). Although these techniques are useful for identifying dominant themes, they often fail to explain how different concepts interact, overlap, or form interconnected semantic structures within a discourse network. Consequently, an important methodological gap remains in understanding the structural association among retail-related concepts and themes (Kim and Kim, 2017). Addressing this gap requires analytical frameworks capable of simultaneously capturing semantic meaning and structural connectivity within large textual datasets (Gupta et al., 2022). To address these limitations, the present study adopted Topological Data Analysis (TDA) with Natural Language Processing (NLP) to examine the structural organisation of organised retail discourse. Specifically, the study employed the Mapper algorithm to identify topological clusters within the semantic embedding space of the textual corpus. Unlike conventional clustering methods, topological approaches focus not only on grouping similar data points but also on revealing the overall shape, connectivity, and structural relationships present within high-dimensional data (Gupta et al., 2022). This approach made it possible to identify how thematic clusters overlapped, fragmented, or remained interconnected across the retail discourse network. Consequently, the study helped uncover the underlying semantic backbone linking major retail-related themes such as branding strategies, consumer perception, supply chain relationships, pricing dynamics, and market competition (Muñoz-Leiva et al., 2021; Kuikka et al., 2024).

2. MATERIALS AND METHODS

The present study was conducted in the organised retail chains in Kolkata and suburban regions of West Bengal in February, 2026. The study aimed to analyse store characteristics, product range, consumer perceptions, and improvement strategies. Since the secondary data consisted of extensive narrative descriptions and qualitative information extracted from the review of literature, advanced Natural Language Processing (NLP) was employed to extract meaningful patterns systematically and objectively described in Figure 1 (Saha et al., 2026, Jurafsky and Martin,

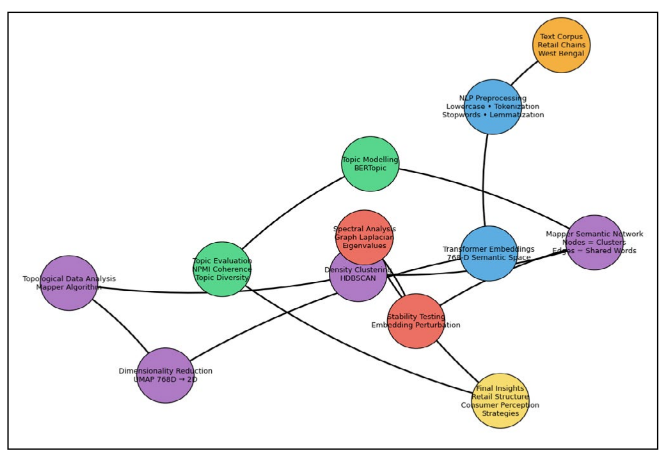

Figure 1: Methodological workflow

2023; Manning et al., 2008). The entire textual corpus was subjected to a preprocessing pipeline to ensure uniformity and analytical suitability (Thakur and Basu, 2025). During preprocessing, all characters were converted to lowercase to maintain consistency. Punctuation marks, special characters, and irrelevant symbols were removed. The text was then tokenised into individual words (Manning et al., 2008). Common stop words (such as "and", "the", and "is") that did not contribute significantly to meaning were eliminated (Salton and Buckley, 1988). Lemmatisation was performed to reduce words to their base forms; for example, "selling" was reduced to "sell", thereby preserving semantic meaning while avoiding duplication of similar word forms (Bird et al., 2009).

After cleaning and structuring the text, the next step involved converting words into numerical representations using contextual word embeddings. A transformer-based language model was used to generate high-dimensional vector representations for each word (Devlin et al., 2019). Each word was mapped into a 768-dimensional semantic space (Devlin et al., 2019). Unlike traditional frequency-based methods such as bag-of-words or TF-IDF, contextual embeddings captured semantic similarity and contextual meaning (Reimers and Gurevych, 2019). This was particularly important in the present study, as consumer perception and retail descriptions often contained nuanced meanings that could not be captured by simple word counts. Since the embedding space was high-dimensional, dimensionality reduction was performed using Uniform Manifold Approximation and Projection (UMAP) to preserve local semantic relationships while projecting data into lower dimensions (McInnes et al., 2018). The embeddings were reduced from 768 dimensions to two dimensions. This reduced space served as a lens for understanding the overall semantic structure of the corpus. The two-dimensional projection allowed visualization and subsequent topological modelling.

To identify structural patterns within the semantic space, Topological Data Analysis (TDA) was applied using the Mapper algorithm (Singh et al., 2007). First, the reduced UMAP space was divided into overlapping regions. Within each region, density-based clustering (HDBSCAN) was applied to identify groups of semantically similar words (Campello et al., 2015). Each cluster was represented as a node in the graph. If clusters shared common words, they were connected by edges. The resulting mapper graph provided a network representation of semantic communities within the retail corpus. The mapper graph allowed identification of major thematic regions, transitional zones between themes, and central nodes representing dominant concepts. To validate and complement the topological findings, topic modelling was performed using BERTopic (Grootendorst, 2022). This approach combined transformer embeddings with clustering and class-based TF-IDF to identify coherent topics from text data (Grootendorst, 2022). Topic summaries were extracted and compared with the semantic communities obtained from the mapper graph. This comparative approach ensured robustness of thematic interpretation. Topic quality was evaluated using topic coherence (Normalized Pointwise Mutual Information) and topic diversity metrics (Bouma, 2009; Roder et al., 2015). These measures ensured that identified themes were both interpretable and non-redundant. Controlled perturbations were introduced into the embedding space to test whether the topological structure remained consistent. Spectral analysis of the graph laplacian was also performed to evaluate structural cohesion and robustness of the semantic network (Chung, 1997). Eigenvalue distributions were examined to assess connectivity and structural stability.

3. RESULTS AND DISCUSSION

The results and discussion section is presented in a structured chronological sequence that follows the analytical workflow of the study. It begins with a description of the corpus characteristics and text preprocessing outcomes to establish the quality and suitability of the dataset for computational analysis. This is followed by the generation of contextual word embeddings and their dimensionality reduction using UMAP, which provides the foundation for visualizing the semantic structure of the retail discourse. The subsequent section presents the topological network constructed using the mapper algorithm, highlighting its structural properties, connected components, and thematic clusters. After identifying the core semantic communities, topic coherence and community characteristics are examined to interpret the dominant themes within the discourse. Finally, the robustness of the network is evaluated through hyperparameter stability analysis and spectral analysis of the graph laplacian, ensuring the reliability and structural validity of the identified semantic patterns.

The descriptive analysis presented in table 1 revealed that the corpus comprising 50,413 characters and 7,551 words, providing a moderately large yet domain specific corpus suitable for advanced computational text analysis (Manning et al., 2008; Jurafsky and Martin, 2023). Its size and richness allowed for meaningful semantic and network-based modelling of store characteristics, product range, consumer perceptions, and improvement strategies in organized retail chains in West Bengal, while the high character count indicated detailed narrative descriptions that enhance contextual depth, particularly valuable for transformer based embedding techniques (Devlin et al., 2019; Reimers and Gurevych, 2019). Preprocessing further refined the corpus for analytical clarity: the raw text contained 8,910 tokens and

Table 1: Corpus characteristics and text pre-processing summary

Category	Stage/Metric	Value	Additional measure
Corpus size	Total characters	50,413	—
	Total words	7,551	—
Token processing	Original tokens	8,910	Vocabulary: 1,661
	After stopword and punctuation removal	4,076	Vocabulary: 1,396
	After frequency thresholding	2,174	Vocabulary: 189

1,661 unique words, which after stopword and punctuation removal reduced to 4,076 tokens and 1,396 words, ensuring that most retained terms carried substantive semantic weight (Salton and Buckley, 1988; Bird et al., 2009). Frequency thresholding then narrowed the dataset to 2,174 tokens and 189 unique words, eliminating rare or noisy terms and concentrating discourse around a small but consistent set of recurring themes (Manning et al., 2008). This lexical coherence strengthened interpretability and reliability for embedding, clustering, and topic modelling, ensuring that the analysis captured stable, contextually relevant patterns in retail operations and consumer perceptions (Blei et al., 2003; Grootendorst, 2022).

The contextual embedding process transformed the refined textual data into meaningful numerical representations shown in Table 2, capturing semantic relationships within the retail discourse (Devlin et al., 2019; Reimers and Gurevych, 2019). A total of 2,217 word embeddings were extracted from the frequency-filtered corpus, with each word represented as a vector in a 768-dimensional semantic space (Devlin et al., 2019). This dimensional structure corresponds to the standard architecture of transformer-based models, such as BERT, and enables the model to capture subtle contextual differences between words (Vaswani et al., 2017; Devlin et al., 2019).

Table 2: BERT embeddings with UMAP Pprojection

Component	Parameter	Value
BERT embeddings	Total embeddings	2,217
	Dimension	768
	Blocks processed	5
UMAP projection	n_components	2
	n_neighbors	15
	min_dist	0.1
	metric	cosine

In contrast to traditional frequency-based approaches, these embeddings represent words based on their contextual usage within the corpus (Reimers and Gurevych, 2019). As a result, terms related to product quality (such as freshness, hygiene, and grading), pricing aspects (such as discount, affordability, and competitive pricing), and service attributes (such as delivery, packaging, and staff behaviour) are positioned closer to each other in the semantic space according to their contextual similarity (Mikolov et al., 2013; Devlin et al., 2019). The embedding extraction was carried out through five BERT computational blocks, which allowed the processing to be performed efficiently in batches while maintaining the contextual integrity of the representations (Devlin et al., 2019).

To enable visualisation and topological analysis, the high-dimensional embeddings were further reduced using Uniform Manifold Approximation and Projection (UMAP) (Figure 2) (McInnes et al., 2018). The original 768-dimensional embedding space was projected into two dimensions, making it possible to observe the overall semantic structure of the corpus and apply subsequent topological modelling techniques (McInnes et al., 2018). The parameter *n_neighbors* was set to 15, meaning that each word embedding was analysed in relation to its fifteen nearest semantic neighbours. This configuration helps to maintain a balance between preserving local semantic relationships and retaining the broader global structure of the data (McInnes et al., 2018).

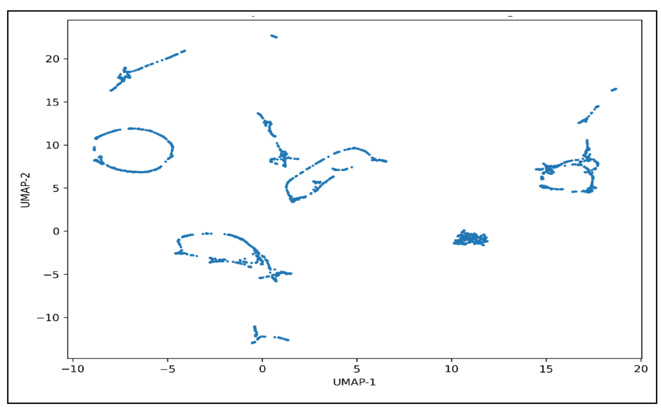

Figure 2: UMAP projection of contextual word embeddings

In the context of retail-related text, such a setting allows closely related terms (for example, freshness–quality–hygiene or price–discount–affordable) to remain clustered together while still distinguishing broader themes such as store attributes, product categories, and consumer perceptions. The *min_dist* parameter was set to 0.1, enabling semantically similar words to form compact clusters in the reduced space while maintaining clear separation between distinct thematic regions (McInnes et al., 2018). Additionally, cosine similarity was used as the distance metric, which is commonly applied in text analysis because it measures the angular similarity between vectors rather than their absolute magnitude (Salton and McGill, 1983; Manning et al., 2008). This ensures that words with similar contextual meanings are grouped even if they appear in different narrative contexts within the retail corpus.

The topological analysis of the retail discourse using the mapper algorithm in Table 3 revealed a structured but sparsely connected semantic network (Figure 3) (Singh et al., 2007; Carlsson, 2009). The resulting mapper graph consisted of 54 nodes and 6 edges, where each node represented a cluster of semantically similar words derived from overlapping regions of the UMAP-reduced embedding space (McInnes et al., 2018; Singh et al., 2007). The presence of 54 nodes indicated that the corpus

Table 3: Mapper network structural properties

Metric/Component	Value
Total nodes	54
Total edges	6
Number of connected components	46
Largest component size	3
Components with 3 nodes	2 (IDs: 42, 22)
Components with 2 nodes	4 (IDs: 17, 37, 10, 40)
Singleton components (1 Node)	40

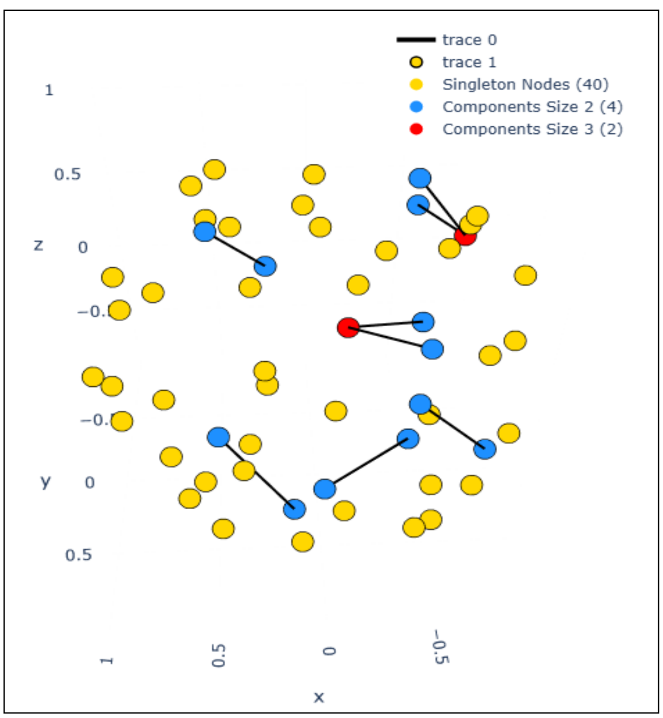

Figure 3: Mapper network structure of the corpus

contained multiple distinct thematic clusters reflecting different aspects of organised retail chains in West Bengal. The relatively small number of edges suggested limited overlap between these clusters, meaning that most themes were conceptually distinct and only a few transitional links existed between them (Newman, 2010). This structure implied that discussions within the corpus were organised around relatively independent thematic domains, such as store infrastructure, product range (particularly fruits and vegetables), consumer perceptions related to freshness, quality and price, as well as suggestion-oriented discussions for improvement. The limited connectivity between clusters indicated that although these themes were related in the broader retail context, they were not strongly interwoven in the textual narratives, which improved the clarity and interpretability of the semantic structure (Blei et al., 2003; Grootendorst, 2022).

Further examination of the connected component distribution showed that the semantic network was highly fragmented (Newman, 2010). Only a small number of components contained more than one node, while the majority consisted of single isolated nodes. Specifically, two components (IDs 42 and 22) contained three nodes each, and four components (IDs 17, 37, 10 and 40) contained two nodes each, while the remaining components consisted of individual nodes. The multi-node components represented areas where different themes overlapped and interacted, indicating transitional regions within the discourse (Singh et al., 2007; Carlsson, 2009). For instance, discussions about pricing intersected with consumer satisfaction, product range connected with quality perceptions, and service-related discussions linked with improvement suggestions. In contrast, the dominance of single-node components suggested that many themes were expressed independently, reflecting clearly defined and self-contained discussions about specific aspects of organised retail such as infrastructure, logistics, product categories or particular consumer concerns (Newman, 2010). Inclusively, the component distribution indicated that the semantic topology of the retail discourse was thematically segmented and structurally sparse, with clearly distinguishable discourse regions connected only through a few transitional bridges (Carlsson, 2009). Rather than indicating weakness, this fragmentation enhanced analytical clarity by allowing individual thematic clusters to be examined separately, thereby facilitating more precise interpretation of store characteristics, product attributes, consumer perceptions and potential improvement strategies.

The structural refinement of the mapper network in Table 4 revealed the core semantic structure underlying the retail discourse. Initially, the mapper graph contained 54 nodes and 6 edges, representing multiple thematic clusters derived from the corpus (Figures 4 and 5) (Singh et al., 2007). However, after applying graph cleaning by extracting the largest connected component, the network was reduced to three nodes connected by two edges (Newman, 2010). This process was commonly used in topological and network analysis to focus on the most meaningful and structurally cohesive portion of the network, since isolated nodes or very small disconnected components often represented highly specialised or weakly connected themes (Barabasi, 2016; Newman, 2010). The reduction from 54 nodes to three meta nodes indicated that the overall textual discourse on organised retail chains had three topological features without strong lexical overlap. Nevertheless, the presence of a small but connected component suggested the existence of a core integrative thematic region within the corpus where key retail concepts intersected (Carlsson, 2009). This central component reflected overlapping discussions related to product quality, branding strategies, pricing structures, consumer perception, supply dynamics, and market positioning, forming the semantic backbone of the discourse (Blei et al., 2003; Grootendorst, 2022).

Further analysis of the cleaned network showed that all three nodes belonged to a single community (Community 0), confirming that the core semantic region was internally cohesive rather than fragmented (Newman, 2010). The community structure therefore indicated that the central themes of the discourse were tightly interconnected and functioned as a unified conceptual block rather than forming multiple independent thematic subgroups. Node-level characteristics further clarified this structure. The largest node, cube32_cluster0 (node size=43), contained prominent terms such as brand, product, consumer, customers, company, franchise, cost, and feature, suggesting a discourse centred on the interaction between branding strategies, product positioning, and consumer evaluation. The NPMI (Normalized Pointwise Mutual Information) coherence score of −0.07726 for the cube32_cluster0 node indicated moderate semantic association among the terms, which was typical for clusters that captured broad conceptual intersections involving multiple retail dimensions (Bouma, 2009; Roder et al., 2015).

The second node, cube42_cluster0 (node size=32), included words such as consumer, label, private, product, features, and customer, reflecting discussions related to consumer perception and private labelling strategies. The NPMI coherence value of −0.08643 suggested that while the

Table 4: Network structure, community characteristics, and stability metrics

Node_ID	Community	Node size	Top words	Normalized pointwise mutual information (NPMI) coherence
cube32_cluster0	0	43	brand, product, consumer, customers, brands, company, franchise, label, cost, feature	-0.07726
cube42_cluster0	0	32	consumer, consumers, label, product, labels, private, conclusions, features, additionally, customer	-0.08643
cube42_cluster1	0	12	product, suppliers, increase, market, share, label, different, features, production, produce	0.123219

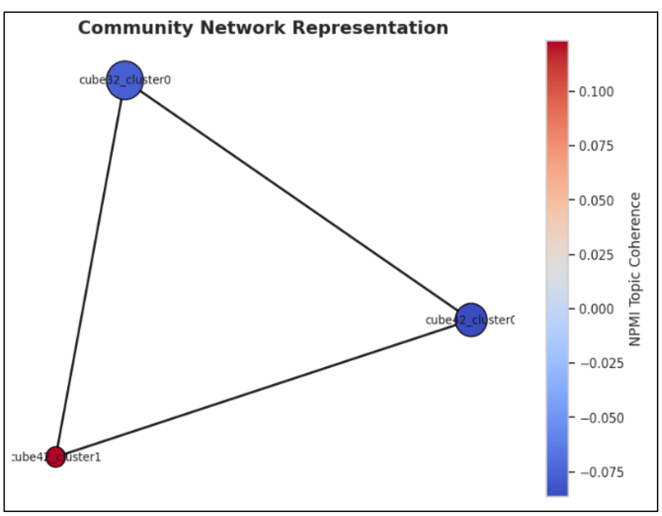

Figure 4: NPMI topic coherence

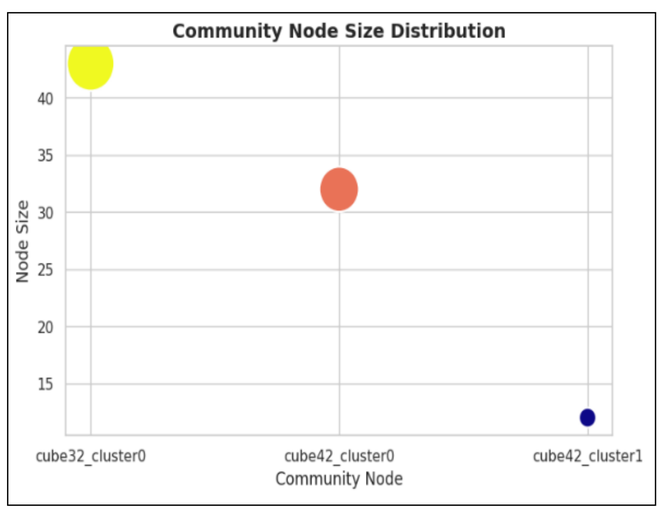

Figure 5: Community node size distribution

node captured a meaningful thematic grouping, it also incorporated diverse contextual references within the consumer–label discourse domain (Roder et al., 2015). In contrast, the third node, cube42_cluster1 (node size = 12), exhibited terms such as suppliers, market share, production, product, and features, representing themes associated with supply chain dynamics, market expansion, and competitive positioning within organised retail systems. Notably, this node recorded a positive NPMI coherence score of 0.123219, indicating a stronger semantic association among the words and suggesting a more tightly focused thematic cluster (Bouma, 2009).

The hyperparameter grid stability analysis in Table 5 examined how different configurations of the Mapper algorithm influenced the consistency of the resulting network structure (Singh et al., 2007; Carriere et al., 2018). The analysis considered variations in two key parameters: the number of cubes (n_cubes), which controlled the resolution of the topological cover, and the overlap between cubes, which determined the degree of intersection among neighbouring partitions (Singh et al., 2007). The stability score in Figure 6 reflected how consistently the graph structure was reproduced under similar parameter settings, with higher values indicating stronger structural reliability (Carrière et al., 2018).

Table 5: Hyperparameter grid stability of network topology

n_cubes	overlap	stability score
10	0.1	0.1
10	0.2	0.116408
10	0.3	0.131775
15	0.1	0.16788
15	0.2	0.146747
15	0.3	0.171773
20	0.1	0
20	0.2	0.093333
20	0.3	0.056656

The results showed that when n_cubes=10, stability scores ranged from 0.10 to 0.13, indicating moderate consistency in the graph structure. When the resolution increased to n_cubes=15, stability improved further, with scores ranging from 0.146 to 0.172, representing the most stable configurations among all tested combinations. In contrast, when the resolution increased to n_cubes=20, stability declined significantly, with values ranging from 0 to 0.093, suggesting that overly fine partitioning of the embedding space caused fragmentation and reduced the reliability of cluster formation (Singh et al., 2007; Carriere et al., 2018). These findings indicated that a moderate resolution level

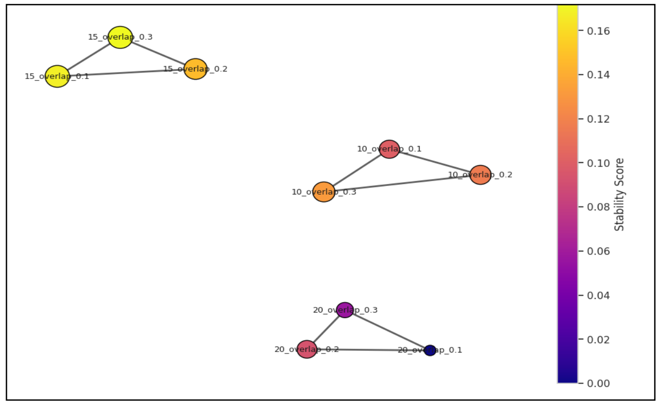

Figure 6: Mapper parameter stability network

provided a more balanced and stable representation of the semantic structure within the retail discourse.

The analysis of the overlap parameter further showed that increasing overlap generally improved structural robustness by preserving connections between neighbouring clusters (Singh et al., 2007). For n_cubes=10, stability increased gradually as overlap rose from 0.1 to 0.3. A similar pattern was observed for n_cubes=15, where the highest stability score (0.1718) occurred at overlap=0.3, indicating that moderate overlap helped maintain meaningful transitions between thematic clusters. However, when n_cubes=20, stability remained low regardless of overlap, confirming that excessive partitioning led to unstable graph structures (Carriere et al., 2018). Based on these results, the configuration n_cubes=15 and overlap=0.3 provided the most reliable and interpretable Mapper representation for the corpus. This stability pattern suggested that the retail discourse possessed moderate structural coherence, where thematic clusters remained identifiable across parameter variations but became unstable when the embedding space was excessively fragmented.

The structural robustness of the semantic network was examined in Table 6 through stability and spectral analyses, the results of which were visualised in Figure 7. The stability analysis compared multiple Mapper graph constructions generated under slight perturbations of the embedding space (Carriere et al., 2018). Specifically, three graphs (Graph 1, Graph 2, and Graph 3) were constructed, and their structural similarity was evaluated using edge similarity measures (Newman, 2010). The results indicated that the edge similarity values for all comparisons namely Graph 1–Graph 2, Graph 2–Graph 3, and Graph 1–Graph 3, were 0, meaning that no edges were shared across the perturbed graph constructions. At first glance, this might have appeared to suggest instability in the network structure. However, this outcome was primarily a consequence of the extreme sparsity of the network (Barabasi, 2016; Newman, 2010). As observed in earlier structural results, the original

Table 6: Stability and spectral structure of the core semantic network

Analysis type	Comparison/Index	Metric	Value	Structural meaning
Stability analysis	Graph 1 – Graph 2	Edge similarity	0	No common edges between perturbed graphs
Stability analysis	Graph 2 – Graph 3	Edge similarity	0	Edge structure changes under small perturbations
Stability analysis	Graph 1 – Graph 3	Edge similarity	0	Sparse graph causes sensitive edge formation
Spectral analysis	Eigenvalue index 0	λ_0	0	Indicates one connected component
Spectral analysis	Eigenvalue index 1	λ_1	1	Moderate algebraic connectivity
Spectral analysis	Eigenvalue index 2	λ_2	2	Balanced structural topology

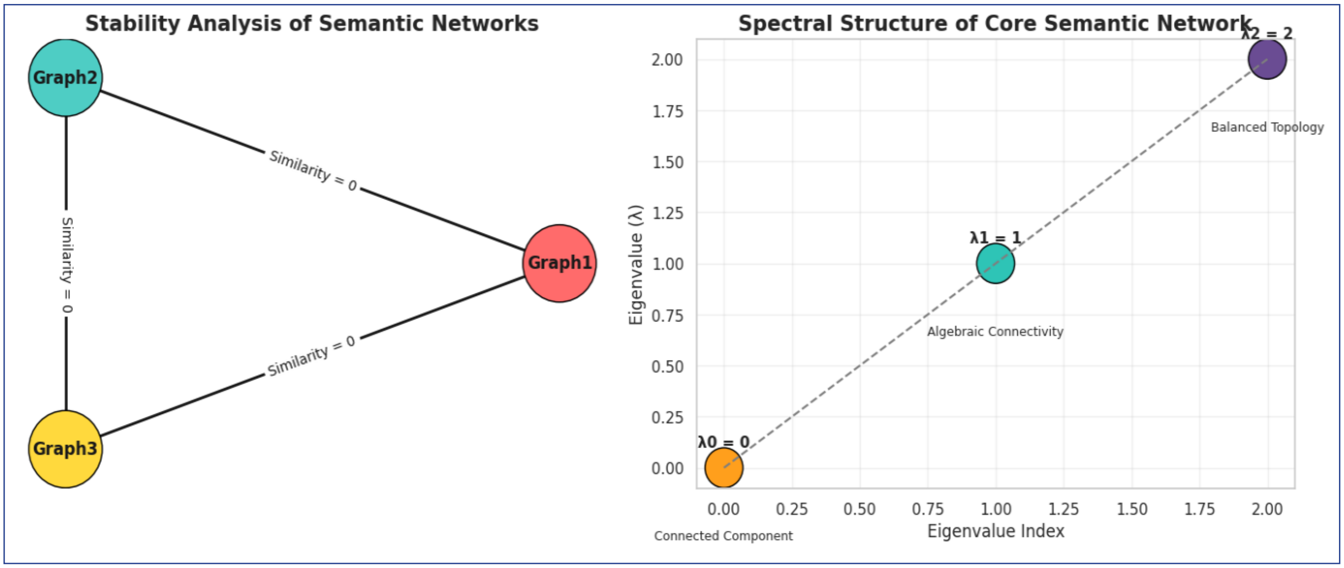

Figure 7: Stability and spectral structure of core semantic network

Mapper graph contained only a small number of edges, and after cleaning the network to retain the largest connected component, only three nodes connected by two edges remained. In such sparse topological networks, even very small variations in clustering boundaries or embedding positions could alter which nodes overlapped, thereby changing the resulting edge connections (Carriere et al., 2018). Therefore, the absence of overlapping edges across runs reflected the sensitivity of sparse graphs to minor perturbations rather than instability in the underlying semantic themes (Barabasi, 2016). Importantly, the dominant thematic nodes themselves remained consistent across runs, indicating that the conceptual structure of the discourse was stable even though the precise overlap between clusters might have shifted slightly. From a conceptual perspective, this finding suggested that organised retail discourse in the corpus was characterised by clear thematic separation with limited transitional overlap, meaning that major topics such as product attributes, branding strategies, consumer perception, and market dynamics tended to appear as distinct conceptual blocks rather than heavily interwoven narratives.

Further insight into the structural properties of the network was provided by the spectral analysis of the cleaned graph, which examined the eigenvalues of the normalised graph Laplacian (Chung, 1997; Newman, 2010). Since the largest connected component contained three nodes, the Laplacian matrix yielded three eigenvalues: λ_0=0, λ_1=1, and λ_2=2. In graph theory, the number of zero eigenvalues corresponds to the number of connected components; therefore, the presence of a single zero eigenvalue confirmed that the cleaned network formed one connected component, validating earlier community detection results that all nodes belonged to the same community (Chung, 1997). The second eigenvalue, known as the algebraic connectivity, was equal to 1, indicating a moderate level of structural cohesion among the nodes (Fiedler, 1973; Newman, 2010). This suggested that while the nodes were connected, they were not densely interconnected, which aligned with the observed structure of three nodes connected by two edges forming a simple chain-like configuration. The largest eigenvalue (λ_2=2) reflected the overall structural balance of the network and indicated the absence of extreme clustering or structural irregularities within the small component (Chung, 1997).

Taken together, the eigenvalue pattern (0, 1, 2) characterized a minimally connected but structurally stable topology.

4. CONCLUSION

Organized fruit and vegetable retail discourse in West Bengal was structured around distinct yet interconnected themes related to branding, product quality, consumer perception, pricing, and supply chain dynamics. The integration of AI and large-scale semantic big data could be used for retail planning, consumer behaviour analysis, branding decisions, and adaptive market strategies in evolving organized retail ecosystems.

5. REFERENCES

Al-Azmi, A.A.R., 2013. Data, text and web mining for business intelligence: A survey. International Journal of Data Mining and Knowledge Management Process 3(2), 47–60. DOI: https://doi.org/10.5121/ijdkp.2013.3201.

Angelopoulos, S., Bendoly, E., Fransoo, J., Hoberg, K., Ou, C., Tenhiala, A., 2023. Digital transformation in operations management: Fundamental change through agency reversal. Journal of Operations Management 69, 876–889. DOI: https://doi.org/10.1002/joom.1271.

Barabasi, A.L., 2016. Network science. Cambridge University Press. DOI: https://doi.org/10.1017/9781107076266.

Bird, S., Klein, E., Loper, E., 2009. Natural language processing with Python. O'Reilly Media. Available from: https://tjzhifei.github.io/resources/NLTK.pdf.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. Available from: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

Bouma, G., 2009. Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL, 30, 31–40. Available from: https://svn.spraakdata.gu.se/repos/gerlof/pub/www/Docs/npmi-pfd.pdf.

Campello, R.J.G.B., Moulavi, D., Zimek, A., Sander, J., 2015. Hierarchical density estimates for data clustering, visualisation, and outlier detection. ACM Transactions on Knowledge Discovery from Data 10(1), 1–51. DOI: https://doi.org/10.1145/2733381.

Carlsson, G., 2009. Topology and data. Bulletin of the American Mathematical Society 46(2), 255–308. DOI: https://doi.org/10.1090/S0273-0979-09-01249-X.

Carrière, M., Michel, B., Oudot, S., 2018. Statistical analysis and parameter selection for Mapper. Journal of Machine Learning Research 19(12), 1–39. Available from: https://jmlr.org/papers/v19/17-291.html.

Chung, F.R.K., 1997. Spectral graph theory. American Mathematical Society. DOI: https://doi.org/10.1090/cbms/092.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT. DOI: https://doi.org/10.48550/arXiv.1810.04805.

Fiedler, M., 1973. Algebraic connectivity of graphs. Czechoslovak Mathematical Journa 23(2), 298–305. DOI: https://doi.org/10.21136/CMJ.1973.101168.

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint. DOI: https://doi.org/10.48550/arXiv.2203.05794.

Gupta, M., Gupta, A., Cousins, K.C., 2022. Toward the understanding of the constituents of organizational culture: The embedded topic modelling analysis of publicly available employee-generated reviews of two major U.S.-based retailers. Production and Operations Management 31(9), 3668–3686. DOI: https://doi.org/10.1111/poms.13843.

Ibrahim, N.F., Wang, X., 2019. A text analytics approach for online retailing service improvement: Evidence from Twitter. Decision Support Systems 121, 37–50. DOI: https://doi.org/10.1016/j.dss.2019.03.002.

Jurafsky, D., Martin, J.H., 2023. Speech and language processing (3rd ed.). Draft version. Available from: https://web.stanford.edu/~jurafsky/slp3/.

Kim, D.W., Kim, S., 2017. Sustainable supply chain based on news articles and sustainability reports: Text mining with Leximancer and Diction. Sustainability 9(6), 1008. DOI: https://doi.org/10.3390/su9061008.

Kuikka, A., Hallikainen, H., Tuominen, S., Laukkanen, T., 2024. What drives customer loyalty in a pandemic? Semantic analysis of grocery retailers. Electronic Commerce Research 25, 2205–2240. DOI: https://doi.org/10.1007/s10660-024-09857-y.

Liu, X., Zhou, Y., Gao, S., 2022. Intellectual structure in supply chain risk management from 2000 to 2022: A review based on a text mining approach. Sustainability 14(23), 16245. DOI: https://doi.org/10.3390/su142316245.

Mahr, D., Stead, S., Odekerken-Schroder, G., 2019. Making sense of customer service experiences: A text mining review. Journal of Services Marketing 33(1), 88–103. DOI: https://doi.org/10.1108/JSM-10-2018-0295.

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to information retrieval. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511809071.

McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. Journal of Open Source Software 3(29),

861. DOI: https://doi.org/10.21105/joss.00861.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. Proceedings of ICLR. DOI: https://doi.org/10.48550/arXiv.1301.3781.
Mohammadi, N., Maghsoudi, M., Soghi, M., 2024. Innovation ecosystems in retail: Uncovering technological trends and collaboration networks through patent mining. IEEE Access 12, 186753–186778. DOI: https://doi.org/10.1109/ACCESS.2024.3515290.
Munoz-Leiva, F., Rodriguez-Lopez, M.E., Liebana-Cabanillas, F., Moro, S., 2021. Past, present, and future research on self-service merchandising: A co-word and text mining approach. European Journal of Marketing 55(8), 2269–2307. DOI: https://doi.org/10.1108/EJM-02-2019-0179.
Nahr, N., Heikkila, M., 2022. Uncovering the identity of Electronic Markets research through text mining techniques. Electronic Markets 32, 1257–1277. DOI: https://doi.org/10.1007/s12525-022-00560-0.
Newman, M.E.J., 2010. Networks: An introduction. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199206650.001.0001.
Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of EMNLP. DOI: https://doi.org/10.48550/arXiv.1908.10084.
Roder, M., Both, A., Hinneburg, A., 2015. Exploring the space of topic coherence measures. Proceedings of WSDM. DOI: https://doi.org/10.1145/2684822.2685324.
Saha, D., Thakur, S., Das, K.K., 2026. Adoption of farm technologies across the Terai Region of West Bengal: A comparative topic modelling study of empirical literature review. International Journal of Bio-resource and Stress Management 17(6), 01–10. DOI: https://doi.org/10.23910/1.2026.7025.
Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management 24(5), 513–523. DOI: https://doi.org/10.1016/0306-4573(88)90021-0.
Salton, G., McGill, M.J., 1983. Introduction to modern information retrieval. McGraw-Hill. Available from: https://sigir.org/files/museum/introduction_to_modern_information_retrieval/frontmatter.pdf.
Sheikhattar, M.R., Nezafati, N., Shokouhyar, S., 2022. A thematic analysis-based model for identifying the impacts of natural crises on a supply chain for service integrity: A text analysis approach. Environmental Science and Pollution Research 29, 79413–79433. DOI: https://doi.org/10.1007/s11356-022-21380-x.
Singh, G., Memoli, F., Carlsson, G., 2007. Topological methods for the analysis of high dimensional data sets and 3D object recognition. Eurographics Symposium on Point-Based Graphics. DOI: https://doi.org/10.2312/SPBG/SPBG07/091-100.
Thakur, S., Basu, D., 2025. Harnessing advanced NLP for multidimensional CEO leadership analysis: A qualitative lexical network and literal approach. International Journal of Agriculture Extension and Social Development 8(6), 612–622. DOI: https://doi.org/10.33545/26180723.2025.v8.i6h.2086.
Tudor, C., Girlovan, A., Botoroga, C.A., 2025. Where business meets location intelligence: A bibliometric analysis of geomarketing research in retail. ISPRS International Journal of Geo-information 14(8), 282. DOI: https://doi.org/10.3390/ijgi14080282.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I., 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30. Available from: https://arxiv.org/abs/1706.03762.