Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

ADVANCED INTERNATIONAL JOURNAL OF BUSINESS, ENTREPRENEURSHIP AND SMES (AIJBES)
www.gaexcellence.com/aijbes

INTEGRATING BIG DATA AND CRM FOR SMES: A HYBRID RFM AND LEXICON-BASED CLUSTERING APPROACH FOR CUSTOMER SEGMENTATION

Syaimak Abdul Shukor[1]*, Siti Hajar Zulkefly[2]

[1] Center for Artificial Intelligence Technology, Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.
syaimak@ukm.edu.my https://orcid.org/0000-0002-7694-154X
[2]Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.
hajarzulkefly@gmail.com https://orcid.org/0009-0007-4474-1539
*Corresponding Author

Article Info:

Article history:
Received date: 01.04.2026
Revised date: 22.04.2026
Accepted date: 26.05.2026
Published date: 15.06.2026

To cite this document:

Shukor, S. A., & Zulkefly, S. H. (2026). Integrating Big Data and CRM for SMEs: A Hybrid RFM and Lexicon-Based Clustering Approach for Customer Segmentation. *Advanced International Journal of Business Entrepreneurship and SMEs, 8* (28), 327-341.

Abstract:

In the era of Big Data, integrating Customer Relationship Management (CRM) with advanced analytics is essential for businesses to maintain a competitive edge. This study focuses on customer segmentation, a core CRM strategy, using a public dataset from a giftware retailer. Although previous CRM segmentation studies have extensively applied the RFM model and clustering algorithms, most focus mainly on transactional metrics without examining product-category purchasing behaviours embedded within unstructured product descriptions. Current studies rely heavily on raw product identifiers, which produce high-dimensional sparse data and limit the interpretability of customer preferences. Accordingly, there remains limited research integrating semantic product categorisation with RFM-based clustering to uncover cluster-specific purchasing behaviours within SME retail environments. To bridge this gap, this study proposes a methodology that combines the RFM (Recency, Frequency, Monetary) model with a Lexicon-based approach for product categorisation during data preprocessing. This approach effectively reduces the dimensionality of product varieties, allowing for a more meaningful analysis of buying behaviours. The study employed an unsupervised K-means clustering algorithm using engineered RFM features derived from transactional records. The optimal number of clusters was determined using the Elbow Method, while the model's validity was confirmed using the Silhouette Index and business logic. The results identified four distinct customer segments: Platinum, Gold, Silver, and Bronze, ranked by their monetary value. Findings specify that the Lexicon-based categorisation significantly enhances the interpretability of purchasing patterns within each cluster.

Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

This research proposes SMEs a scalable framework for customer profiling, targeted marketing, inventory optimisation, and strategic CRM decision-making through the integration of transactional analytics and semantic product categorisation.

DOI: 10.35631/AIJBES.828022 **Keyword:**

Big Data, CRM, Customer Segmentation. K-Means Clustering, Lexicon Approach, RFM Model

Introduction

The rapid evolution of Big Data technologies has empowered organisations to transition from traditional "knowing" entities to "learning" organisations. In the retail sector, Big Data emerges primarily from the automation of business activities, generating massive volumes of transactional, analytical, and customer-centric data. This data environment is characterised by high volume and variety, encompassing both structured data and quantifiable information stored in relational databases, as well as unstructured data such as text, audio, and web imagery. Unlike structured data, unstructured data lacks a fixed schema and requires sophisticated preprocessing before conventional machine learning models can use it.

In this data-saturated era, integrating Big Data with Customer Relationship Management (CRM) is no longer optional but essential. CRM strategies leverage this information to bolster sales, marketing, and customer service efforts. By synthesising customer interactions and market trends, organisations can deliver personalised experiences and tailored services (Anshari et al., 2019). While CRM success is enabled by technology, its core lies in its capacity to learn and adapt to customer needs, thereby fostering long-term loyalty and retention (Kebede & Tegegne, 2018). Ultimately, effective CRM serves as a profit-maximisation strategy by centring the business model on the quality of customer relationships.

The current literature categorises customer segmentation via data mining into methodology-oriented and application-oriented approaches (Namvar, Khakabimamaghani & Gholamian, 2011). Methodology-oriented studies focus on optimising specific algorithms, such as Naïve Bayes, Decision Trees, and Neural Networks, for classification and prediction. Conversely, application-oriented studies seek to define new variables or input sets to solve specific industry problems. Within this context, the Recency, Frequency, and Monetary (RFM) model remains a cornerstone of application-oriented segmentation, valued for its long-standing effectiveness in decoding buyer habits and behavioural patterns (Guney, Peker & Turhan, 2020).

Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

Most existing customer segmentation studies within CRM analytics prioritise transactional attributes such as spending frequency and purchase recency while overlooking semantic information contained in product descriptions. In retail environments with highly diverse product catalogues, particularly the giftware industry, raw product identifiers create sparse and fragmented datasets that obscure meaningful behavioural patterns. Consequently, prior segmentation models often fail to explain *what types of products* customers prefer across different loyalty segments, limiting the strategic usefulness of clustering outcomes for targeted marketing and inventory planning. This study addresses this limitation by pursuing two primary objectives: first, to generate customer segments using K-means clustering based on the RFM model; and second, to identify distinct purchasing patterns within these clusters through a novel Lexicon-based approach.

By applying Natural Language Processing (NLP) to unstructured text attributes (product descriptions), this research extracts latent information to categorise product varieties. The primary contribution of this study is the development of a specialised dictionary for the giftware industry, which reduces dimensionality and provides granular insights into cluster-specific purchasing behaviours. This hybrid approach demonstrates how SMEs can enhance standard RFM models with NLP to gain a more sophisticated understanding of their customer base (Wang, Tsai & Ciou, 2020).

In conclusion, this research provides a significant contribution to the field of data-driven CRM by demonstrating how SMEs can transcend the limitations of small or constrained datasets. By integrating the traditional RFM model with a novel Lexicon-based NLP approach, this study moves beyond simple classification to offer a nuanced understanding of customer purchasing behaviours in the giftware industry. The identification of four distinct segments, Platinum, Gold, Silver, and Bronze, proves that even with limited attributes, the application of unsupervised machine learning can yield actionable strategic intelligence. For practitioners and SME owners, this research offers a scalable, cost-effective framework to enhance customer retention and personalise marketing efforts, ultimately transforming raw transactional data into a sustainable competitive advantage in an increasingly digitised marketplace.

The novelty of this study lies in the integration of traditional RFM-based customer segmentation with a Lexicon-driven semantic categorisation framework for product descriptions. Unlike previous studies that depend solely on raw transactional attributes, this research introduces a domain-specific lexicon to reduce product complexity and sparsity through semantic grouping. This integration enhances cluster interpretability, enables cluster-specific purchasing pattern analysis, and improves the strategic usability of segmentation results for SME-oriented CRM analytics.

Literature Review

The literature review establishes the theoretical and empirical foundations of this study by synthesising current advancements in customer segmentation, machine learning, and data analytics within the retail landscape. It examines the structural evolution of clustering algorithms and the diverse data taxonomies ranging from transactional records to behavioural sequences that inform modern Customer Relationship Management (CRM) strategies. Furthermore, by evaluating the intersection of traditional frameworks such as the RFM model with emerging Natural Language Processing (NLP) techniques, this section identifies critical research gaps in product-specific behavioural analysis. Ultimately, this review provides the

Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

necessary context for the proposed hybrid lexicon-based approach, positioning it as a strategic solution for SMEs operating within the complexities of the Big Data era.

Foundations of Clustering Analysis

Clustering analysis is a fundamental unsupervised machine learning technique that partitions data into groups characterised by high intra-cluster similarity and high inter-cluster dissimilarity. This methodology is widely adopted across various sectors to identify valuable customer segments. Beyond the retail industry (Dogan, Hiziroglu & Seymen, 2020; Rahadian & Syairudin, 2020), clustering is instrumental in optimising credit card limits in banking (Zorina, 2019), identifying enrollment factors in informal education (Rahadian & Syairudin, 2020), analysing patient visitation patterns in healthcare (Arunachalam & Kumar, 2018), and predicting genre preferences in video-on-demand services (Guney, Peker & Turhan, 2020).

The efficacy of a clustering model depends on rigorous pre- and post-modelling considerations. Central to the pre-modelling phase is determining the optimal number of clusters (k). For algorithms such as K-means, the Elbow Method is frequently used to initialise the model by identifying the point at which the rate of decrease in the within-cluster sum of squares (WCSS) levels off (Kabasakal, 2020; Savitri, Bachtiar & Setiawan, 2018). Post-modelling, clusters must undergo validation using metrics such as the Silhouette Index and business logic to ensure practical relevance.

Taxonomy of Clustering Algorithms

Clustering approaches are generally divided into hierarchical and partitioning methods. Hierarchical clustering iteratively compares data points to build a nested structure, which is further categorised into agglomerative (bottom-up) and divisive (top-down) algorithms (Hamdan, Abu Bakar & Ahmad Nazri, 2018). Conversely, partitioning algorithms, such as K-means and K-medoids, require the prespecification of the number of clusters and group data based on proximity to the nearest centroid. The choice between these methods depends largely on the study's objectives and the dataset's dimensionality.

Categorisation of Cluster Input Data

Literature identifies four primary types of input data for customer segmentation: transaction data, value data, profile data, and sequence data (Liu & Chen, 2017).

- **Transaction Data:** Comprising numerical and temporal values (e.g., invoices, quantity, shipping), this is the most common input.
- **Customer Value Data:** Focuses on the customer's economic contribution to the firm. The **RFM (Recency, Frequency, Monetary) model** is the preeminent framework in this category, used to distinguish high-value loyalists from at-risk customers (Kabasakal, 2020; Wang et al., 2020).
- **Customer Profile Data:** Utilises demographic variables such as age, gender, and income to understand the "who" behind the purchase (Dogan et al., 2020).
- **Customer Purchase Sequence Data:** Analyses the chronological order of purchases to capture dynamic behavioural shifts (Liu & Chen, 2017).

Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

Research Gaps and the Lexicon Approach

A comparative analysis of previous studies as shown in Table 1 utilising the same giftware dataset reveals a significant research gap. While the existing literature has explored general patterns, there is a lack of granular analysis of product purchase behaviours within specific clusters. Griva et al. (2021) noted that product variation and taxonomy are critical factors influencing segmentation effectiveness, yet they are often overlooked due to their complexity. In retail analytics, raw product identifiers frequently generate sparse and fragmented feature spaces due to excessive product variation. This negatively affects cluster interpretability and limits the identification of meaningful purchasing patterns. Lexicon-based categorisation addresses this limitation by semantically grouping related products into interpretable categories, thereby reducing dimensional complexity while preserving behavioural relevance. Consequently, semantic categorisation improves segmentation quality, facilitates cluster interpretation, and supports more actionable CRM decision-making.

In the giftware industry, developing automated product categorisation models is hampered by a lack of specialised training datasets. To address this, this study proposes a manual Lexicon-based approach integrated with Natural Language Processing (NLP). By defining twelve distinct categories including Accessories, Home, Kitchen, and Stationery; the dimensionality of unstructured text attributes can be reduced. This hybrid methodology extends prior research by applying a robust classification framework to examine cluster-specific purchasing behaviours.

Table 1 Comparison previous research paper with same dataset

No	Research Paper, Year	Objective	Technique/ Algorithm	Cluster Validation
1	(Chen, Sain & Guo 2012)	To perform customer segmentation based on RFM Model to find meaningful customer segment.	K-means clustering, decision tree	RFM statistics (ie min, med, max)
2	(Christy et al. 2018)	To perform customer segmentation based on RFM Model by comparing three different algorithm..	K-means clustering, Fuzzy c-min clustering, Repetitive median k-min (RM K-Min)	iteration execution taken Silhouette Index
3	(Otiko, Odey & Inyang 2019)	To discover set of items frequently bought together and segmenting the product to find	Association Rules mining, Hierarchical clustering	Confidence support lift

Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

		meaningful product cluster.		
4	(Anitha & Patil 2020)	Perform customer segmentation using k-min 5 clustering and execute the m6odel on real time data.	K-means clustering	Silhouette Index
5	(Vohra et al. 2020)	To perform customer segmentation based on RFM Model. Then, develop two different models as performance benchmarks for future research.	K-means clustering, self-organizing maps (SOM)	Execution Time No of Iterations Space Complexity Time Complexity
6	(Piskunova & Klochko 2020)		Linear discriminant analysis (LDA), Support vector machine (SVM), Classification and regression trees (CART), k-nearest neighbors (KNN), Random forests (RF)	Elbow Method, Silhouette Index, Hubert, Dunn Index and 20 other criteria Accuracy statistics (minimum, maximum, average, quartile 1, quartile 3

In summary, while clustering and the RFM model are established pillars of customer segmentation, their application in the giftware sector remains limited by the unstructured nature of product descriptions. The synthesis of the existing literature underscores the need to move beyond basic transaction metrics and include product taxonomy. By bridging the gap between unsupervised learning and NLP-driven lexicon categorisation, this research provides a comprehensive framework that allows SMEs to derive deeper, more actionable insights from limited datasets, ultimately enhancing both theoretical understanding and practical CRM implementation.

Research Methodology

This section outlines the systematic research framework employed to segment the customer base of a UK-based online retailer. The study adopts a three-phased methodology: Data Preparation, Clustering Modelling, and Comparative Analysis. By integrating traditional RFM metrics with advanced Natural Language Processing (NLP), this approach moves beyond transactional analysis to include product-specific behavioural insights.

The study utilises a comprehensive customer transaction dataset from a United Kingdom-based online retail company, sourced from the UCI Machine Learning Repository (Dua & Graff, 2017). The retailer, established in 1981, specialises in a diverse range of giftware, including jewellery, kitchenware, and seasonal products. The dataset captures the company's pivotal transition period into e-commerce (2009–2011), providing a robust foundation for analysing early-stage digital customer behaviour (Chen, Sain & Guo, 2012).

Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

The raw data comprises 1,067,371 transaction records spanning from December 1, 2009, to December 9, 2011. Each record includes eight fundamental attributes: Invoice Number, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, and Country.

Phase 1: Data Preparation and Feature Engineering

The preparation phase involves two critical workflows: the creation of a Baseline Dataset (original transactional data) and a Lexicon Dataset (enriched with product categories).

Data Cleaning

To ensure high-quality inputs, the dataset underwent several preprocessing steps:

- Target Selection: Focusing on relevant transactional records while excluding non-customer data.
- Missing Value Handling: Identifying and removing records with missing Customer IDs to ensure segmentation accuracy.
- Noise and Outlier Removal: Managing negative values (returns/cancellations) and extreme statistical outliers that could skew the clustering results.

NLP and Lexicon Categorisation

The lexicon-based categorisation also functions as a semantic dimensionality reduction mechanism by consolidating thousands of heterogeneous product descriptions into twelve interpretable product categories. This process mitigates the curse of dimensionality commonly associated with retail product data and improves cluster explainability.

To handle the unstructured "Description", NLP techniques were applied to develop the Lexicon Dataset. The text was standardised by lowercase, punctuation removal, tokenisation, stopword removal, and lemmatisation. Using a manual Lexicon approach, products were mapped into twelve domain-specific categories (e.g., Accessories, Home, Kitchen, Stationery). This categorisation facilitates a more granular analysis of purchasing patterns within the resulting clusters.

RFM Extraction and Scaling

The core features for the clustering model were extracted based on the RFM Model. The attributes are calculated as follows:

- **Recency (*R*):** The number of days between the current date and the customer's last purchase date.
  Recency=Current Date – Late Purchased Data
- **Frequency (*F*):** The total number of unique transactions per customer.
- **Monetary (*M*):** The total revenue generated by the customer, calculated as:
  *Monetary=Quantity * UnitPrice*

Given that the K-means algorithm is sensitive to data scale and distribution, the RFM values were log-transformed to reduce skewness and standardised (Z-score scaling) to ensure all features contributed equally to the distance calculations.

Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

Phase 2: Modelling and Cluster Determination

The modelling phase utilises the K-means clustering algorithm, a popular unsupervised learning technique for large-scale segmentation. A critical step in this phase is determining the optimal number of clusters (k).

The Elbow Method was utilised to evaluate the Within-Cluster Sum of Squares (WCSS). Based on the resulting scree plot, a benchmark of $\boldsymbol{k = 4}$ was selected, corresponding to the point at which additional clusters provide diminishing returns in explaining data variance. The scaled RFM data served as the primary input for the final modelling process.

Phase 3: Cluster Analysis and Comparison

The final phase focuses on interpreting the generated segments to ensure they are both distinct and actionable. Each cluster is profiled based on its unique RFM characteristics and labelled (Platinum, Gold, Silver, and Bronze) according to its average monetary contribution.
A key highlight of this phase is the comparative analysis between the Baseline Data and the Lexicon Data. This comparison assesses whether including product categories (via the Lexicon approach) provides deeper insights into cluster-specific purchasing behaviours than transactional data alone.

In summary, the methodology transitions from raw transactional records to refined strategic segments through a rigorous pipeline of NLP-driven preprocessing and K-means clustering. By harmonising the simplicity of the RFM model with the descriptive power of Lexicon-based product categorisation, this research framework provides a holistic view of customer loyalty and buying habits. This dual-layered analysis ensures that the resulting segments are not only statistically valid but also offer practical marketing intelligence for SMEs in the competitive giftware sector.

Result and Discussion

This section presents the empirical results of the clustering process and evaluates the strategic insights derived from the RFM-K-means framework. The discussion is structured first to validate the model's statistical integrity and then to analyse the behavioural profiles of the identified segments. A critical component of this analysis is the comparison between raw transactional data (Baseline) and the NLP-enhanced (Lexicon) data, highlighting how product categorisation uncovers hidden consumer trends that are vital for SME decision-making.

Model Validation and Optimal Cluster Selection

Selecting the optimal number of clusters (k) is a pivotal step in unsupervised learning, requiring a balance between mathematical precision and business utility. Following the Elbow Method (Figure 1), a benchmark of $k=4$ was initially identified. To ensure robustness, the clustering process was iterated over multiple k values and evaluated using the Silhouette Index, with higher scores indicating better cluster separation.

Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

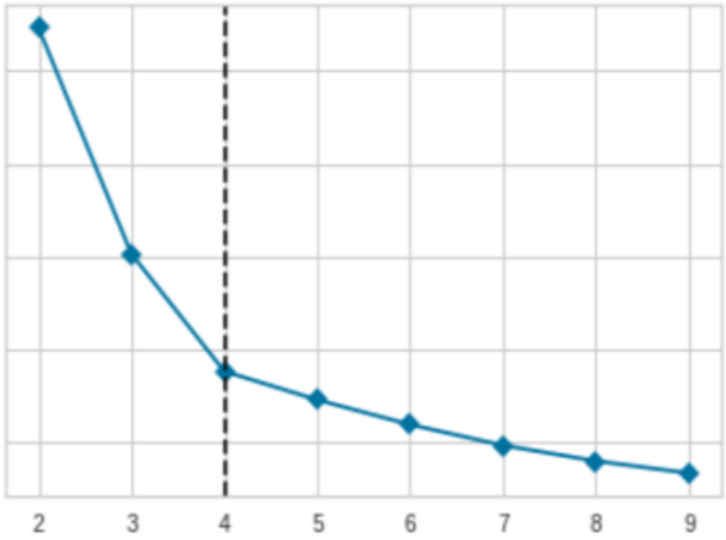

Figure 1 Elbow Method for Determining *K*

As illustrated in Table 2, while *k=2* yielded the highest Silhouette Index (0.41), a binary division of the customer base lacks the granularity required for sophisticated CRM strategies. In contrast, *k=4* emerged as the optimal "sweet spot," providing the second-highest Silhouette score while aligning with the Elbow Method and business logic (Piskunova & Klochko, 2020). This configuration allows for a nuanced four-tier segmentation that is statistically defensible and strategically actionable.

Table 2 Silhouette Index Values

	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9
Silhouette Index	0.41	0.31	0.33	0.30	0.28	0.29	0.29	0.28

Cluster Profiling and Behavioural Analysis

The model successfully partitioned the customer base into four distinct segments, categorised by their RFM characteristics. Figure 2 shows the average RFM values for each cluster:

- Platinum Cluster (Loyal Champions): This segment exhibits the highest Monetary (£10,801) and Frequency (19 visits) averages, with a low Recency of 26 days. These are the company's most valuable and loyal assets, contributing significantly to steady revenue.
- Gold Cluster (At-Risk High-Spenders): Characterised by high total spend (£1,922) and moderate frequency (5 visits), this group has a concerning average Recency of 218 days. They represent "potentially loyal" customers who are currently drifting toward churn, requiring urgent re-engagement campaigns.
- Silver Cluster (Emerging/New Customers): With the lowest Recency (21 days) and moderate spending (£849), this group represents the newest cohort. Their high activity levels suggest strong potential to convert to the Platinum tier through targeted loyalty programs.

Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

- Bronze Cluster (Lapsed/Inactive): This group shows the lowest engagement across all metrics (£329 spend, 1 visit, 371 days since last purchase). These are "lost" customers whose re-acquisition costs may exceed their projected lifetime value.

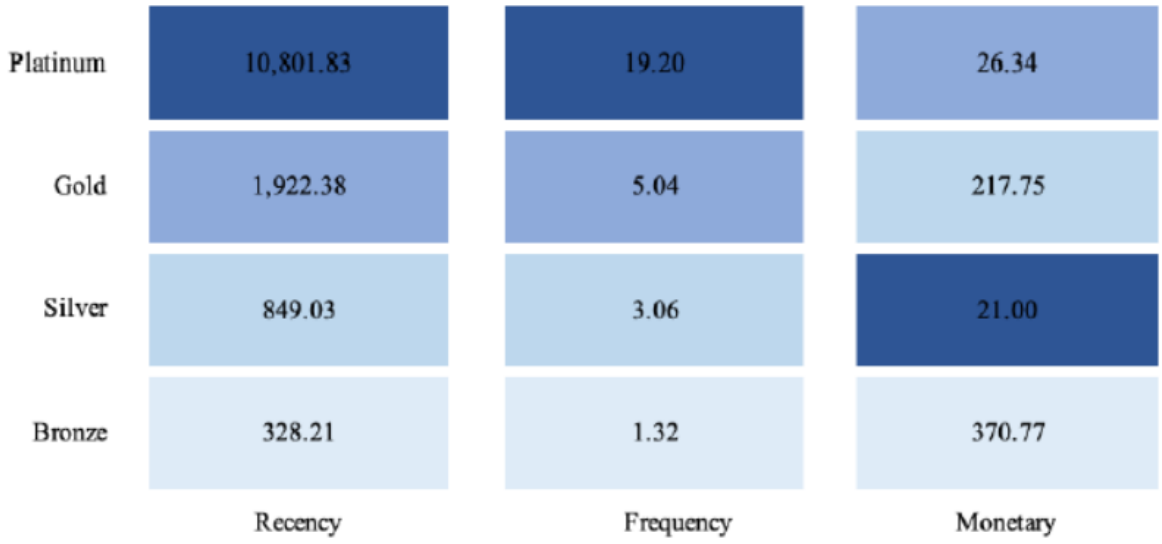

Figure 2 Average RFM values each cluster

Comparative Exploration: Baseline vs Lexicon Data

Baseline Data: The Limitations of Product ID Exploration

Initial analysis using raw Product IDs failed to yield significant strategic intelligence. When examining the top 12 products per cluster, the results remained largely consistent over the two years, suggesting stable but non-descriptive purchasing habits. High overlap—where 50% of the top items were identical across different clusters—rendered it impossible to distinguish the specific "tastes" or preferences of one segment over another. This highlights the "dimensionality curse" often found in retail Big Data, where raw IDs obscure rather than clarify consumer behaviour.

Lexicon Data: Category-Based Strategic Insight

In contrast, the Lexicon-based categorisation provided a transparent view of consumer preferences. Regardless of the year, a consistent hierarchy emerged across all clusters, with Kitchen, Home, and Bags dominating the giftware interest. While the types of items purchased are similar across tiers, the volume and value of these purchases serve as the primary differentiators. The semantic grouping mechanism substantially improved analytical interpretability compared to raw product-level analysis. Instead of analysing isolated stock identifiers with limited business meaning, the lexicon framework enabled the identification of broader consumer preference structures across functional product categories. This demonstrates the practical value of semantic reduction in SME-oriented Big Data analytics.

Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

Furthermore, the Lexicon approach revealed a critical temporal trend. A significant surge in transactions begins in August, driven by the Kitchen and Home categories. Notably, the Festive category peaks between August and October, followed by a sharp decline in December. This suggests that the giftware market operates on an "anticipatory buying" cycle, where customers complete holiday preparations months in advance. For SMEs, this insight is crucial: marketing spend for holiday stock should be concentrated in the late summer and early autumn rather than the traditional December window.

Summary of Findings

The results of this study underscore the superiority of hybrid analytical models over traditional transactional sorting. While standard K-means identifies which customers are valuable, integrating a Lexicon-based NLP approach explains what they buy and when. By reducing the noise of raw product IDs into twelve meaningful categories, this study demonstrates that SMEs can uncover actionable seasonal trends and segment-specific preferences. This methodological refinement allows businesses to move beyond broad-stroke marketing and toward high-precision, data-driven CRM strategies that optimise inventory and promotional timing.

Conclusion

This final section synthesises the empirical findings and articulates the strategic value of the research for the SME sector, providing a roadmap for managers to transition from descriptive analytics to prescriptive marketing actions while identifying critical avenues for subsequent scholarly inquiry. This study demonstrates the efficacy of a hybrid analytical framework for decoding complex consumer behaviour in the giftware industry. By triangulating the Elbow Method, the Silhouette Index, and business logic, the research identified that a four-cluster solution (*k4*) provides the optimal balance between statistical granularity and managerial interpretability. The resulting segments, Platinum, Gold, Silver, and Bronze, offer a clear hierarchy of customer value, primarily differentiated by their monetary contribution and engagement frequency. Furthermore, the integration of a Lexicon-based Natural Language Processing (NLP) approach proved superior to raw transactional analysis, successfully reducing the high dimensionality of product descriptions into twelve actionable categories. This methodological refinement uncovered significant seasonal purchasing cycles, particularly the "anticipatory" buying patterns for festive goods, which would have been obscured by traditional RFM modelling.

For SMEs, these findings provide a data-driven foundation for optimising Customer Relationship Management (CRM) strategies. Rather than using "one-size-fits-all" marketing, businesses can now deploy targeted interventions tailored to specific cluster needs. For instance, Platinum and Silver clusters can be prioritised for retention strategies, such as loyalty programs and early-access previews, to maintain high engagement levels. Conversely, the Gold cluster, characterised by high historical value but dwindling recent activity, should be the primary target for re-engagement tactics, including personalised "win-back" campaigns and incentivised offers. Finally, by identifying the bronze cluster as a low-value, high-churn group, managers can optimise resources by reallocating marketing budgets away from low-ROI segments and toward high-potential new customers. Ultimately, leveraging these insights allows SMEs to enhance customer lifetime value (CLV) and proactively mitigate churn in an increasingly competitive digital marketplace.

Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

While this study provides a robust framework for segmentation, several avenues for improvement remain. This research utilised a "hard" clustering approach, where each customer is assigned to a single, mutually exclusive group; however, future studies could explore Fuzzy C-Means or Soft Clustering algorithms. These methods allow customers to hold membership in multiple segments simultaneously, more accurately reflecting the fluid nature of consumer behaviour, where a single individual may exhibit multiple purchasing patterns. Additionally, while the manual Lexicon approach effectively handled the giftware domain's product taxonomy, it remains labour-intensive. Future research should investigate the use of Supervised Machine Learning or Deep Learning models, such as BERT or GPT-based embeddings, to automate product categorisation at scale. Expanding the dataset to include multi-channel touchpoints such as social media interactions and customer support logs would further enrich the model's predictive power, moving the field toward a more holistic, 360-degree view of the customer.

From a managerial perspective, the proposed framework enables SMEs to transform fragmented transactional data into strategically actionable customer intelligence. By combining behavioural scoring with semantic product analysis, organisations can improve customer retention strategies, optimise promotional timing, personalise marketing campaigns, and enhance inventory planning with greater analytical precision. The framework is particularly valuable for SMEs with limited analytical resources, as it provides an interpretable and scalable approach to Big Data-driven CRM implementation.

Acknowledgements: The authors would like to express their sincere gratitude to Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia for providing the necessary resources and support throughout the course of this research. Special appreciation is extended to colleagues and peers who contributed valuable insights and constructive feedback, which greatly enhanced the quality of this paper.

Funding Statement: This research received financial support from Universiti Kebangsaan Malaysia under Grant Number [TAP-K010342].

Conflict of Interest Statement: The authors declare that there is no conflict of interest regarding the publication of this paper. All authors have contributed to this work and approved the final version of the manuscript for submission to Advanced International Journal of Business, Entrepreneurship and SMEs (AIJBES).

Ethics Statement: This study did not involve any human participants, animals, or sensitive data requiring ethical approval. The authors confirm that the research was conducted in accordance with accepted academic integrity and ethical publishing standards.

Author Contribution Statement: All authors contributed significantly to the development of this manuscript. Syaimak Abdul Shukor was responsible for the conceptualization, methodology, critical revision of the manuscript and overall supervision of the study. Siti Hajar handled literature review, drafting, data collection, analysis, and interpretation of results. All authors read and approved the final version of the manuscript prior to submission.

Advanced International Journal of Business, Entrepreneurship and SME's
AIJBES
EISSN: 2682-8545
Volume: 8 Issue: 28 (June 2026) PP. 327-341

References

Anitha, P., & Patil, M. M. (2020). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2018.12.011

Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. (2019). Customer relationship management and big data enabled: Personalization & customization of services. *Applied Computing and Informatics*, *15*(2), 94–101. https://doi.org/10.1016/j.aci.2018.05.003

Arunachalam, D., & Kumar, N. (2018). Benefit-based consumer segmentation and performance evaluation of clustering approaches: Evidence of data-driven decision-making. *Expert Systems with Applications*, *111*, 11–34. https://doi.org/10.1016/j.eswa.2018.01.035

Bergström, S. (2019). *Customer segmentation of retail chain customers using cluster analysis* [Master's thesis, KTH Royal Institute of Technology]. DiVA Portal.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*(3), 197–208. https://doi.org/10.1057/dbm.2012.17

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2018). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2018.09.004

Dogan, O., Hiziroglu, A., & Seymen, O. F. (2020). Segmentation of retail consumers with soft clustering approach. In *International Conference on Intelligent and Fuzzy Systems* (pp. 39–46). Springer.

Dua, D., & Graff, C. (2017). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml

Griva, A., Bardaki, C., Pramatari, K., & Doukidis, G. (2021). Factors affecting customer analytics: Evidence from three retail cases. *Information Systems Frontiers*, 1–24. https://doi.org/10.1007/s10796-020-10040-2

Guney, S., Peker, S., & Turhan, C. (2020). A combined approach for customer profiling in video on demand services using clustering and association rule mining. *IEEE Access*, *8*, 185107–185120.

Hamdan, A. R., Abu Bakar, A., & Ahamd Nazri, M. Z. (2018). *Sains data penerokaan pengetahuan dari data raya*. Penerbit Universiti Kebangsaan Malaysia.

Kabasakal, I. (2020). Customer segmentation based on recency frequency monetary model: A case study in e-retailing. *Journal of Business and Economic Studies*, *13*, 47–56.

Kebede, A. M., & Tegegne, Z. L. (2018). The effect of customer relationship management on bank performance: In context of commercial banks in Amhara Region, Ethiopia. *Cogent Business & Management*, *5*(1). https://doi.org/10.1080/23311975.2018.1493915

Liu, Y. C., & Chen, Y. L. (2017). Customer clustering based on customer purchasing sequence data. *International Journal of Engineering Research and Application*, *7*(1), 49–58.

Namvar, M., Khakabimamaghani, S., & Gholamian, M. (2011). An approach to optimize customer segmentation and profiling using RFM, demographic features, and LTV. *International Journal of Electronic Customer Relationship Management*, *5*, 220–235.

Otiko, A. O., Odey, J. A., & Inyang, G. A. (2019). Conceptualisation of market segmentation and patterns for pre-Christmas sales in an online retail store. *International Journal of Research in Business and Social Science*.

Piskunova, O., & Klochko, R. (2020). Classification of e-commerce customers based on data science techniques. *Central European Management Journal*.

Rahadian, Y. R., & Syairudin, B. (2020). Segmentation analysis of students in X course with RFM model and clustering. *Jurnal Sosial Humaniora*, *13*(1), 1–12.

Savitri, A. D., Bachtiar, F. A., & Setiawan, N. Y. (2018). Segmentasi pelanggan menggunakan metode K-Means clustering berdasarkan model RFM pada klinik kecantikan (Studi kasus: Belle Crown Malang). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *2*(9), 2957–2966.

Vohra, R., Pahareeya, J., Hussain, A., Ghali, F., & Lui, A. (2020). Using self organizing maps and K-Means clustering based on RFM model for customer segmentation in the online retail business. *International Journal of Advanced Science and Technology*, *29*, 1641–1655.

Wang, S. C., Tsai, Y. T., & Ciou, Y. S. (2020). A hybrid big data analytical approach for analyzing customer patterns through an integrated supply chain network. *Journal of Industrial Information Integration*, *20*, 100177.

Widyadhan, D., Hastuti, R. B., Kharisudin, I., & Fauzi, F. (2021). Perbandingan analisis klaster K-Means dan average linkage untuk pengklasteran kemiskinan di Provinsi Jawa Tengah. *PRISMA, Prosiding Seminar Nasional Matematika*, 584–594.

Zorina, K. (2019). *Building segment based revenue prediction for CLV model*. [Thesis].