Book of Papers presented at the IFKAD 2026
Intelligent Knowledge for Sustainable Organizations
Budapest, Hungary 1-3 July 2026
ISBN: 978-88-96687-20-8

Contextual–RFM (C-RFM): a Multimodal Framework for the Fusion of Transactional and Semantic Data Through LLMs for Interactive Customer Segmentation

Davide Madonna, Simone Malacaria, Andrea De Mauro
Luiss University, Rome, Italy

Abstract

This study examines the evolution of customer segmentation from rule-based systems to multimodal AI architectures, emphasizing the necessity for transparent and context-aware CRM strategies. Despite the widespread adoption of RFM models, organizations often struggle to synthesize transactional metrics with multifaceted qualitative feedback, creating a numerical paradox that this research addresses by capturing the authentic voice of the customer within high-dimensional analytical spaces. This paper presents the Contextual-RFM (C-RFM) framework, an innovative multimodal architecture aimed at bridging the representational divergence between quantitative metrics and qualitative feedback.

The research operationalizes a comprehensive five-phase pipeline to ensure methodological transparency and replicability: (1) establishment of a statistical baseline, (2) narrative synthesis for tabular serialization, (3) multimodal vectorization utilizing advanced embeddings, (4) semantic analysis through Latent Dirichlet Allocation (LDA) and (5) implementation of the C-RFM Expert chatbot. This entire workflow is implemented using KNIME Analytics and Langflow, providing a robust and replicable process for both scholars and practitioners. A key distinction of the framework is its early-fusion approach. Unlike traditional late-fusion models that integrate data streams only at the final analysis stage, early-fusion retains critical cross-modal correlations during the initial processing phase, effectively addressing the semantic opacity common in traditional numerical-only datasets.

Empirical findings indicate that the C-RFM framework adeptly organizes linguistic complexity, evidenced by a significant improvement in structural validity when transitioning from purely transactional environments to topic-validated contexts. Moreover, this study introduces the Discrepancy Detection Rate (DDR) as a novel performance metric to quantify the misalignment between transactional behaviour and qualitative feedback, thereby identifying critical churn risks that conventional quantitative analyses often overlook.

The primary contribution of this research lies in the synergy achieved between human expertise and Artificial Intelligence. By deploying dynamic persona chatbots grounded in Retrieval-Augmented Generation (RAG), the framework facilitates a shift from opaque analytical outputs to actionable, customized recovery strategies. This glass-box methodology transforms intricate numerical complexities into auditable business narratives, empowering decision-makers to transcend statistical constructs and foster a deeper, dialogue-based understanding of consumers. The C-RFM framework represents a significant advancement in human-centred AI, promoting strategic marketing intelligence that is not only transparent but also finely tuned to the authentic voices of real-world consumer archetypes, thereby paving the way for the next generation of customer-centric strategies.

Keywords: C-RFM Framework, Large Language Models, Explainable AI, Multimodal Data Fusion, Human-Centred CRM

Paper type: Academic Research Paper

1 Introduction

Artificial Intelligence is fundamentally transforming the Customer Relationship Management (CRM) landscape by enabling a significant shift from standardized marketing approaches to highly personalized interactions (Naderipour and Hosseini Nasabian, 2025; Bukhari et al., 2022; Kandi and Basani, 2024). In this evolving framework, segmentation continues to play a crucial role, allowing organizations to effectively manage market density by identifying and isolating specific behavioural groups (Syahra, Fadlil and Yuliansyah, 2025; Akande, Asani and Dautare, 2024). Historically, the theoretical foundation of this field has been the Recency, Frequency, and Monetary (RFM) model. However, its evolution from rule-based systems to unsupervised machine learning, notably K-means clustering, has exposed a persistent numerical paradox: a reliance on purely transactional data that lacks the contextual depth necessary to capture the multifaceted nuances of the modern consumer journey (Lewaa, 2023; Gopakumar et al., 2021).

The technical roots of this limitation lie in traditional One-Hot Encoding (OHE), which often suffers from the dimensionality curse and statistical instability (Chung, 2024). More critically, OHE induces a condition of semantic opacity within datasets. As Bayrak (2022) and Liang (2025) highlight, by treating all categories as mathematically independent, this method fails to recognize latent behavioural similarities between consumer

ISBN: 978-88-96687-20-8

actions. Contemporary analytical frameworks are transitioning toward deep semantic embeddings to resolve these structural flaws. Smutek et al. (2024) utilize encoder-decoder structures to compress high-dimensional inputs into dense vectors where numerical proximity reflects semantic relationships. Validating these learned representations presents a significant challenge: while the Silhouette score is traditionally utilized as the primary metric for cluster validation, assessing structural quality by measuring both intra-cluster cohesion and inter-cluster separation (Gagolewski, Bartoszuk and Cena, 2021; Shutaywi and Kachouie, 2021), its effectiveness is often diminished in embedding spaces. The dimensionality curse and latent space sparsity make data points appear nearly equidistant, rendering traditional distance-based validation unreliable (Simpson et al., 2025). Researchers have increasingly adopted Latent Dirichlet Allocation (LDA) as a robust alternative to overcome these geometric limitations, thereby overcoming geometric limitations while demonstrating remarkable stability even when semantic relationships are intricate or data is sparse (Sidek and Ahmad, 2026).

Treating transactional data as a linguistic structure (Woltmann, Thiele and Lehner, 2018) positions Large Language Models (LLMs) as the next transformative frontier, leveraging their emergent abilities for data synthesis and optimization-based feature engineering (Nam et al., 2024). Despite this potential, LLM integration into analytical workflows faces a significant numerical paradox, where standard Byte-Pair Encoding (BPE) tokenizers fragment digits into subunits without inherent meaning, which, as Penttala (2025) states, strips numerical values of their quantitative significance. Hegselmann et al. (2023) demonstrate that serializing tabular data into natural language strings allows models to leverage their extensive prior knowledge to mitigate these errors, effectively interpreting transactional records as coherent behavioural narratives. This trajectory culminates in multimodal fusion strategies, which merge structured metrics with unstructured insights through sophisticated transformer-based architectures to capture cross-modal interactions and improve predictive transparency (Hangloo and Arora, 2025). Yet, even as these models enhance analytical depth, the transition from opaque outputs to actionable managerial trust remains a critical hurdle. This difficulty is primarily driven by the inherent lack of transparency in high-dimensional clustering, which frequently generates skepticism among strategic decision-makers (Wang and Ding, 2024). In critical CRM contexts, Explainable AI (XAI) offers a vital methodological framework for converting opaque computations into narratives that are comprehensible to humans (Patil, 2024). Beyond merely achieving predictive accuracy, XAI employs post-hoc tools such as SHAP and counterfactual analysis to uncover behavioural drivers, thereby bridging the gap between statistical constructs and strategic marketing intelligence while ensuring ethical and accountable insights (Almtrf, 2025). This integration is exemplified in dynamic persona chatbots, which transform the interaction between decision-makers and consumer archetypes from static descriptions to engaging, real-time dialogues (Salminen et al., 2026). By leveraging text embeddings models and direct LLM prompting to create persona chatbots, these agents accurately simulate consumer preferences, fostering a trust-based synergy between humans and AI to address complex marketing challenges (Li, Liu and Yu, 2025; Huang, Markovitch and Stough, 2024). This paper introduces the Contextual-RFM (C-RFM) framework to reconcile quantitative rigor with semantic interpretation. In addressing the heterogeneity gap that arises when comparing distinct representation spaces (Hangloo and Arora, 2025), C-RFM employs an early-fusion approach. By integrating transactional and semantic data during the initial preprocessing phase, the framework produces robust embeddings that overcome both the semantic limitations of traditional models and the numerical paradox associated with tokenization (Avnat et al., 2024). This methodology equips LLMs with a deeper contextual understanding while alleviating the information loss and managerial distrust often linked to high-dimensional clustering. By combining XAI with interactive chatbots, C-RFM enables an interactive data interpretation, addressing the following research questions:

- *RQ1: Which methodological approaches prove most effective for validating the multimodal segmentation of integrated transactional and semantic data?*
- *RQ2: How can an LLM-embedding framework, implemented through Explainable AI and interactive persona chatbots, effectively bridge the divide between opaque analytical outputs and actionable, transparent strategies?*

2 Methodology

The C-RFM framework is operationalized through a comprehensive five-phase pipeline, which serves as a procedural structure for converting the theoretical early-fusion strategy into a replicable analytical process.

2.1 Phase 1: Statistical baseline

This phase establishes a quantitative benchmark utilizing the Olist Brazilian E-Commerce dataset. A refined sample of 4,783 profiles was processed in KNIME Analytics, incorporating relational data from customers, orders, payments, and reviews. Procedural filters were applied to remove low-density reviews, while stratified sampling

ISBN: 978-88-96687-20-8

was employed to maintain variance within high-value segments. A baseline RFM analysis was performed using K-means clustering, optimized through a two-stage methodology, which identified three segments as the optimal configuration. Cluster validation was carried out using the Silhouette score, resulting in a mean coefficient of 0.474. Although this outcome is considered acceptable, it highlights the structural ambiguities and boundary overlaps present in raw transactional data.

2.2 Phase 2: Narrative synthesis

The limitations of traditional tokenization are mitigated by replacing statistical grouping with row-level textual representations. The dimensions of RFM were redefined through quantile binning into semantic categories, transforming numerical variance into contextually significant data for LLMs. As suggested by Hegselmann et al. (2023), this serialization enables models to utilize extensive prior knowledge to interpret transactional records as coherent behavioural narratives. By integrating these categories with the verified customer perspective, the framework preserves cross-modal correlations prior to high-dimensional processing.

2.3 Phase 3: Multimodal vectorization

This phase implements the early-fusion strategy by embedding synthesized narratives into a high-dimensional space using the OpenAI *text-embedding-3-small* model. To avoid data leakage and facilitate the identification of new latent relationships, narratives were cleansed of Phase 1 RFM labels. Dimensionality was managed through Principal Component Analysis (PCA), which retained 100 components, accounting for 76% of the variance, followed by K-means clustering with three clusters to ensure stable segmentation. The analysis indicated a notable transition between transactional and semantic profiles. As illustrated in the migration matrix (Figure 1), RFM clusters are predominantly aligned with embedding cluster 0, an occurrence attributed to the specific feedback distributions found within the Olist dataset. Despite this skewness, cross-tabulation analysis revealed a p-value of 0.001, indicating that high-value transactional segments are distributed across diverse semantic contexts. This finding highlights the limitations of relying solely on raw numerical data and emphasizes the need for enhanced semantic depth (Li, Liu and Yu, 2025).

Frequency Row Percent	0	1	2	Total
At-Risk Customers	998	314	533	1.845
	54,0921%	17,019%	28,8889%	
Lost Customers	1.096	386	531	2.013
	54,4461%	19,1754%	26,3785%	
Top Customers	449	166	310	925
	48,5405%	17,9459%	33,5135%	
Total	2.543	866	1.374	4.783

Figure 1 – Migration matrix of RFM and embeddings clusters.

2.4 Phase 4: Semantic analysis

Topic modelling using LDA revealed behavioural communities within the high-dimensional vector space. Following a text mining workflow designed to minimize linguistic noise and an evaluation of the perplexity index, three distinct themes emerged. Labels for these topics were assigned based on the semantic relevance of the most representative terms identified during the modelling process (Figure 2). Topic 0 was labelled as *delivery excellence and reliability*, indicating positive logistical experiences; topic 1 was characterized as *product discrepancy and material quality*, which highlighted concerns regarding the physical attributes of products; and topic 2 was designated as *logistic and administrative friction*, focusing on post-sale and bureaucratic challenges.

ISBN: 978-88-96687-20-8

Topic id (String)	Concatenate(Term) (String)
topic_0	recommend, quality, ahead, packaged, price, earlier, expectations, quickly, advertised, correct
topic_1	quality, missing, box, photo, material, original, watch, bad, little, fit
topic_2	delay, invoice, wrong, response, email, requested, day, refund, address, exchange

Figure 2 – Representative terms extracted for each topic

The assignment of narrative identities to the embedding clusters was informed by their distribution across these themes (Figure 3). As a result, cluster 0 was classified as *effortless brand ambassadors*, cluster 1 as *quality-focused skeptics,* and cluster 2 as *administrative hostages*. The structural integrity of these segments was further substantiated through unigram and bigram validation, culminating in the development of a high-fidelity customer narrative.

Semantic Cluster (String)	Delivery Excellence & Reliability... (Number (Integer))	Logistic & Administrative Friction... (Number (Integer))	Product Discrepancy & Material Quality (Number (Integer))
Administrative Hostages	411	805	158
Effortless Brand Ambassadors	2096	234	213
Quality-Focused Skeptics	165	130	571

Figure 3 – Distribution of embeddings clusters across semantic topics

This format serves as the foundational contextual input for the persona chatbot, integrating transactional data, experience scores, and semantic analysis into a cohesive behavioural profile (Figure 4).

```
"CUSTOMER PROFILE: This user is a '", $RFM Cluster$, "' (categorized as: ", $Recency Label$, ", ", $Frequency Label$, ", and ", $Monetary Value Label$, "). ",
"TRANSACTIONAL DATA: Last purchase: ", string($Recency$), " days ago. Total activity: ", string($Frequency$),
" orders with a total lifetime value of ", string($Monetary Value$), " Reais. ",
"EXPERIENCE DATA: Review Score: ", string($Review Score$), "/5 stars. Feedback: '", $Review Text$, "'. ",
"SEMANTIC ANALYSIS: Cluster: '", $Semantic Cluster$, "'. Topic: '", $Topic Name$, "'. ",
"CONTEXTUAL KEYWORDS: ", $Topic Keywords$, ". "
```

Figure 4 – Prompt engineering template for customer stories

2.5 Phase 5: Implementation of the C-RFM Expert

The final phase focuses upon the deployment of the C-RFM Expert through a RAG architecture, which serves as the primary interface for narrative explainable AI. The implementation is structured around three core processes: knowledge base construction, orchestration via Langflow, and agentic deployment.

The construction of the knowledge base began with 1,000 representative vectorized narratives. A distributed processing approach utilizing a LLMs was employed to conduct a discrepancy analysis, aligning transactional metrics with sentiments derived from qualitative feedback. This methodology facilitates the categorization of customers into semantic clusters and the development of recovery plans that are specifically tailored to the unique behavioural nuances identified within each profile narrative.

System orchestration was executed in Langflow, utilizing Chroma DB for long-term multidimensional storage and a *gpt-4o-mini* agent to independently search the vector store. The resulting C-RFM Expert persona operates under a system prompt designed to identify friction points and rationalize strategic interventions. This approach guarantees that strategic recommendations are both auditable and grounded in empirical customer narratives, effectively converting numerical complexity into actionable business intelligence.

3 Results and Discussion

The following sections outline the empirical findings of the C-RFM framework, emphasizing the validation of the multimodal approach and its strategic advantages compared to conventional transactional models. By integrating quantitative metrics with generative insights, the results illustrate how the numerical paradox can be effectively addressed through both structural and semantic alignment.

3.1 Quantitative performance and validation

The validation of the C-RFM framework employed a multi-faceted statistical approach, incorporating the Silhouette score to assess segment cohesion, the Adjusted Rand Index (ARI) to measure model novelty, and the Discrepancy Detection Rate (DDR), a novel metric specifically developed and implemented through this study, to quantify the business value of uncovering latent attrition risks.

ISBN: 978-88-96687-20-8

3.1.1 Multimodal structural validation

A significant finding emerged when traditional RFM clusters were projected onto the PCA-derived semantic space, resulting in a Silhouette score of 0.000. This outcome suggests that transactional segments are statistically indistinguishable from random noise when assessed through customer feedback, thereby providing empirical evidence that purely quantitative models lack semantic richness. Within this conventional framework, RFM categories do not share a unified linguistic or experiential narrative, highlighting that high spending does not necessarily equate to customer satisfaction and that transactional loyalty can often obscure underlying administrative challenges. In contrast, the C-RFM framework effectively organized this linguistic complexity. As illustrated Table 1, the transition from raw PCA space to probabilistic topic distributions significantly improved structural validity.

Table 1 – Comparative analysis of Silhouette scores across multimodal analytical spaces

Model/Space	**Silhouette score**	**Interpretation**
RFM clusters on PCA space	0.000	Total semantic opacity
C-RFM clusters on PCA space	0.161	Emergence of latent structure
C-RFM on topic matrix (LDA)	0.249	Probabilistic consolidation

3.1.2 Innovation and risk: ARI and DDR

The scientific novelty of the model was evaluated by calculating the Adjusted Rand Index (ARI) between the RFM clusters from Phase 1 and the semantic clusters from Phase 4. The primary aim of the ARI analysis was to assess the degree of similarity, or lack thereof, and potential redundancy between the established transactional model and the proposed multimodal framework (Sundqvist, Chiquet and Rigaill, 2023).

The computed ARI of 0.005 represents a significant finding, indicating that the two partitioning methods are stochastically independent. This near-zero value underscores a phenomenon of semantic opacity, suggesting an almost non-existent correlation between a customer's spending behaviour and their qualitative experiences. While an ARI close to 1 would indicate redundancy, this outcome confirms that the C-RFM framework reveals distinct behavioural dimensions, demonstrating its capacity to provide a non-redundant and unique strategic perspective.

Additionally, the framework presents the DDR as an innovative performance metric, which is established through the statistical cross-validation of transactional RFM segments and semantic topic clusters. This metric is defined by the following ratio:

$$DDR = \frac{|\{\mathrm{c} \in Starget : \omega(\mathrm{c}) \in Crisk\}|}{|Starget|}$$

- Starget: this refers to the group of customers identified within the highest transactional value segment.
- Crisk: this denotes the subset of semantic clusters that highlight critical friction points.
- ω(c): this is the semantic mapping function that assigns a customer, denoted as c, to a specific cluster based on their textual feedback.
- |·|: this operator indicates cardinality, representing the total number of individuals within the specified set.

Derived from the empirical values within the dataset, the computed DDR of 0.515 exposes a significant latent churn crisis, indicating that 51.5% of the company's most valuable customers are classified within high-risk clusters. This underlying threat underscores the fundamental limitations of conventional RFM analysis, which often fails to identify these critical churn risks (Lewaa, 2023). LDA topic modelling validates the internal differentiation within traditional RFM clusters, proving that customers are not a homogeneous entity but are instead fragmented across diverse experiential themes.

3.2 Managerial interpretability

The evolution from abstract clusters to persona chatbots signifies the peak of Human-Computer Interaction (HCI) (Salminen et al., 2026). By transforming statistical groupings into strategic archetypes, the framework effectively connects data science with actionable marketing strategies. The deployment of the C-RFM Expert agent illustrates that LLMs, when supported by a RAG vectorized knowledge base, can conduct discrepancy analysis and generate prescriptive strategies with high accuracy.

ISBN: 978-88-96687-20-8

This interpretability is implemented within the Langflow playground, where the agent enables interactive segmentation. Unlike conventional dashboards that present static data lists, the agent empowers managers to query the database using context-sensitive criteria. A pertinent example of this functionality is illustrated by the following strategic inquiry:

“*Identify a high-value customer in the database who had a negative experience with the last purchase and provide a brief report specifying their exact numerical Recency, Frequency, and Monetary values, their Semantic Cluster name, a quote of their actual Review Text, and a concise 3-step recovery plan to resolve the discrepancy between their high transactional value and negative feedback”.*

As demonstrated in Figure 5, the agent successfully retrieved a high-value profile, accompanied by feedback expressing significant frustration with delivery failures.

High-Value Customer Report

Customer Profile:

- **Recency:** 192 days since last purchase
- **Frequency:** 18 orders
- **Monetary Value:** 181.86 Reais
- **Semantic Cluster Name:** Administrative Hostages

Review Feedback:

- **Rating:** 1.0/5 stars
- **Quote:** "The delivery was severely lacking, leading to frustration and dissatisfaction."

Recovery Plan

1. **Immediate Resolution:**
   - Contact the customer to provide an update on their delivery status. If the product is confirmed lost or significantly delayed, offer a prompt refund or replacement. Ensure clear communication about the steps being taken to resolve the issue.
2. **Compensation Offer:**
   - As a gesture of goodwill, offer free expedited shipping on their next order and a 20% discount on their next purchase. This acknowledges their loyalty and addresses the logistical issues they faced.
3. **Follow-Up Communication:**
   - After resolving the issue, follow up with the customer to ensure their satisfaction with the resolution and to reinforce the commitment to improving their experience. This can help rebuild trust and encourage future purchases.

Figure 5 – C-RFM Expert prescriptive response

The output generated by the C-RFM agent represents a significant shift in analytical detail and segmentation effectiveness. While traditional RFM clustering provides a purely quantitative perspective on consumer value, remaining semantically opaque regarding the underlying causes of attrition, the C-RFM framework leverages the agent's ability to integrate transactional significance with detailed linguistic feedback.

By following the agent's reasoning, managers could obtain an auditable and evidence-based understanding of the proposed strategy. This glass-box methodology ensures that each recommendation is underpinned by a clear rationale, enabling decision-makers to transition from descriptive insights to prescriptive interventions in real-time (Patil, 2024). Such transparency effectively alleviates the managerial skepticism commonly associated with black-box automated systems (Wang & Ding, 2024).

3.3 Academic and practical implications

The results of this study provide a dual-layered contribution, connecting advanced data science with strategic marketing management.

3.3.1 Academic implications

This research resolves the numerical paradox by demonstrating that early-fusion textualization enables LLMs to preserve and interpret cross-modal correlations. The shift from a Silhouette score of 0.000 to 0.249 illustrates that semantic signals can be mathematically structured to enhance transactional data.

However, it is important to acknowledge the absolute value of these structural metrics. Although this transition signifies a measurable improvement, scores remain lower than traditional RFM baselines, aligning with literature on the dimensionality curse in high-dimensional embedding spaces. This structural compression arises because the Silhouette score on Euclidean distance, which often fails to capture true proximity within high-dimensional contexts where latent space sparsity complicates rigid partitioning (Simpson et al., 2025). The framework addresses this challenge by shifting from geometric rigidity to the probabilistic logic of LDA, supplemented by the ARI and DDR metrics. While ARI provides empirical proof of the framework's novelty by confirming the stochastic independence of transactional and semantic spaces, the DDR offers a non-geometric validation of the model's strategic utility. By modelling the relationship between documents and topics as a distribution, the framework navigates the inherent ambiguity and polysemy of natural language more effectively than traditional distance metrics. Thus, the observed spatial overlap should not be interpreted as a failure of the model, but rather as a reflection of the fuzzy and continuous nature of semantic meanings (Petukhova et al., 2024).

3.3.2 Practical implications

The C-RFM framework effectively addresses hidden customer attrition by shifting from aggregate, retrospective segmentation to proactive, hyper-personalized recovery strategies. While traditional RFM models often lack insight into the emotional and logistical challenges that accompany high-value transactions, this framework functions as a scalable tool that converts passive data points into dynamic, conversational intelligence.

Organizations can implement interventions that are both semantically and economically tailored by automating the identification of strategic archetypes. Such a transformation ensures that recovery efforts directly address the root causes of dissatisfaction, tackling specific friction points that numerical metrics alone may overlook. Additionally, the introduction of an agentic interface effectively mitigates the managerial skepticism. The framework establishes AI as an advanced strategic partner by integrating complex multidimensional insights into actionable and verifiable business narratives, thereby ensuring a human-centred and evidence-based response.

4 Conclusions and Future Research

The C-RFM framework operationalized in this study effectively reconciles the quantitative inconsistencies inherent in traditional segmentation methods with a transparent model of XAI through the implementation of an early-fusion architecture. Its primary contribution lies in the practical validation of this early-fusion approach, which minimizes information loss and generates integrated semantic embeddings, thereby preserving semantic richness.

In addressing RQ1, the study illustrates that a comprehensive validation strategy is crucial for effective integrated multimodal segmentation. By merging structured narrative synthesis with LDA topic modelling and structural metrics, the framework identifies behavioural segments that are both statistically robust and semantically distinct. This methodology ensures that the resulting clusters transcend mere mathematical constructs, representing meaningful and interpretable consumer archetypes.

Concerning RQ2, the framework bridges the divide between opaque analytical outputs and actionable strategies through an agentic RAG architecture. The C-RFM Expert persona exemplifies how interactive chatbots, rooted in validated narratives, can automate discrepancy assessments and generate customized recovery plans. This shift from static categorizations to dynamic dialogue guarantees that marketing strategies are contextually transparent, data-driven, and auditable by human stakeholders.

Additionally, the introduction of the DDR signifies a foundational innovation for identifying latent customer attrition. As a novel metric, it serves as a crucial starting point for iterative development, facilitating further refinement as it is applied across various organizational contexts to quantify the misalignment between transactional value and consumer sentiment.

Future research could enhance the C-RFM framework by tackling current challenges related to data acquisition and the computational demands associated with commercial LLM interface. By incorporating socio-demographic variables, subsequent studies can further enrich the depth and detail of customer narratives, yielding more comprehensive contextual insights. This framework can also serve as a significant advancement in human-centred CRM, fostering a more ethical and accountable approach that aligns data-driven accuracy with the complex realities of human behaviour.

5 Ethics Declaration

No ethical approval was required for this study.

6 AI Declaration

LLMs including Notebook LM were employed for content extraction, while Gemini was employed for linguistic refinement. Gemini specifically assisted in proofreading, structural clarification, and the formalization of the DDR formula, which is firmly grounded in the author's conceptual framework and empirical data. All substantive content, interpretations, and analytical contributions remain the author's original work. These tools provided purely editorial and formalization support, without influencing the intellectual or methodological foundations of the research.

References

Akande, O. N., Asani, E. O. and Dautare, B. T. (2024) "Customer segmentation through RFM analysis and K-means clustering: Leveraging data-driven insights for effective marketing strategy", *Ceddi Journal of Information System and Technology (JST)*, Vol. 3, No. 1, pp 14-25. https://doi.org/10.56134/jst.v3i1.81

Almtrf, A. (2025) "Integrating Explainable Ai (Xai) Into Decision Support Systems: A Framework for Enhancing Transparency and Trust in Managerial Decision-Making", *International Journal of Managerial Studies and Research (IJMSR)*, Vol. 13, No. 9, pp 9-22. https://doi.org/10.20431/2349-0349.1309002

Avnat, E., Levy, M., Herstain, D., Yanko, E., Ben Joya, D., Tzuchman Katz, M., Eshel, D., Laros, S., Dagan, Y., Barami, S., Mermelstein, J., Ovadia, S., Shomron, N., Shalev, V. and Abdulnour, R.-E.E. (2025). "Performance of large language models in numerical vs. semantic medical knowledge: Benchmarking on evidence-based Q&As". *Artificial Intelligence in Medicine*, 151, 102943. https://doi.org/10.2196/64452

Bayrak, A. T. (2022) "An application of Customer Embedding for Clustering", *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp 914–920. https://doi.org/10.1109/ICDMW58026.2022.00019

Bukhari, T. T., Oladimeji, O., Etim, E. D. and Ajayi, J. O. (2022) "Customer Lifetime Value Prediction Using Gradient Boosting Machines", *Gyanshauryam, International Scientific Refereed Research Journal*, Vol. 5, No. 4, pp 488–506. https://gisrrj.com/paper/GISRRJ225427.pdf

Chung, S. (2024) "Limitations and Problems of Binarization with One Hot Encoding", *CIS 660 Data Mining with Advanced Machine Learning*. Cleveland State University.

Gagolewski, M., Bartoszuk, M. and Cena, A. (2021). "Are cluster validity measures in fact valid?" *Information Sciences*, 581, 620–636. https://doi.org/10.1016/j.ins.2021.10.004

Gopakumar, A. B., Krishnan, G., Rozario, N. L., Nazreth, R. B. and Anilkumar, P. (2021) "Customer segmentation using K - Means Algorithm", *International Journal of Innovative Research in Technology (IJIRT)*, Vol. 8, No. 7, pp 127–132. https://ijirt.org/publishedpaper/IJIRT154245_PAPER.pdf

Hangloo, S. and Arora, B. (2025) "Multimodal fusion techniques: Review, data representation, information fusion, and application areas", *Neurocomputing*, Vol. 649, Art. 130827. https://doi.org/10.1016/j.neucom.2025.130827

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X. and Sontag, D. (2023) "TabLLM: Few-shot classification of tabular data with large language models", in Ruiz, F., Dy, J. and van de Meent, J. W. (eds.) *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*, Proceedings of Machine Learning Research, Vol. 206, pp. 5549–5581. https://proceedings.mlr.press/v206/hegselmann23a.html

Huang, D., Markovitch, D. G. and Stough, R. A. (2024) "Can chatbot customer service match human service agents on customer satisfaction? An investigation in the role of trust", *Journal of Retailing and Consumer Services*, Vol. 76, Art. 103600. https://doi.org/10.1016/j.jretconser.2023.103600

Kandi, A. and Basani, M. A. R. (2024) "Personalization and Customer Relationship Management in AI-Powered Business Intelligence", *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, Vol. 12, No. 11, pp. 714-715. https://doi.org/10.22214/ijraset.2024.65159

Lewaa, I. (2023) "Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis", *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS32021293

Li, Y., Liu, Y. and Yu, M. (2025) "Consumer segmentation with large language models", *Journal of Retailing and Consumer Services*, Vol. 82, Art. 104078. https://doi.org/10.1016/j.jretconser.2024.104078

Liang, Z. (2025). "Efficient representations for high-cardinality categorical variables in machine learning". In *Proceedings of the 2025 IEEE International Conference on Advanced Machine Learning and Data Science (AMLDS 2025).* IEEE. https://doi.org/10.1109/AMLDS63918.2025.11159409

Naderipour, M. and Hosseini Nasabian, H. (2025). "A comprehensive RFM-A framework: Integrating age for enhanced customer segmentation and marketing strategies". *Transactions on Soft Computing*, 1(1), 14–26. https://doi.org/10.48314/tsc.v1i1.34

Nam, J., Kim, K., Oh, S., Tack, J., Kim, J. and Shin, J. (2024) "Optimized Feature Generation for Tabular Data via LLMs with Decision Tree Reasoning", *In Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. https://doi.org/10.52202/079017-2932

Patil, D. (2024) "Explainable Artificial Intelligence (XAI) for industry applications: Enhancing transparency, trust, and informed decision-making in business operation", *Hurix Digital*. https://www.researchgate.net/publication/385881486

Penttala, J. (2025) "Performance of LLM models with numerical data", Master's Thesis, *University of Helsinki*. https://helda.helsinki.fi/items/aaa2fe24-fbcd-4cc4-a65d-912746098b92

Petukhova, A., Matos-Carvalho, J. P. and Fachada, N. (2024) "Text Clustering with Large Language Model Embeddings", *International Journal of Cognitive Computing in Engineering*. https://doi.org/10.1016/j.ijcce.2024.11.004

Salminen, J., Jung, S.-G., Chowdury, S. and Jansen, B. J. (2026) "AI-generated personas: Representing user needs with generative AI models", *International Journal of Human-Computer Studies*. https://doi.org/10.1016/j.ijhcs.2026.103745

Shutaywi, M. and Kachouie, N.N. (2021). "Silhouette analysis for performance evaluation in machine learning with applications to clustering." *Entropy*, 23(6), 759. https://doi.org/10.3390/e23060759

Sidek, Z. and Ahmad, S. S. S. (2026) "Exploring word embeddings and clustering algorithms for user reviews", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 41, No. 3, pp. 1017-1024. https://doi.org/10.11591/ijeecs.v41.i3.pp1017-1024

Simpson, C., Tabatsky, E., Rahil, Z., Eddins, D. J., Tkachev, S., Georgescauld, F., Papalegis, D., Culka, M., Levy, T., Gregoretti, I., Meehan, C., Schiller, C., Bestak, K., Schapiro, D., Chernyshev, A., Walther, G., Ghosn, E. E. B. and Orlova, D. (2025) "Lifting the curse from high-dimensional data: Automated projection pursuit clustering for a variety of biological data modalities", *GigaScience*, Vol. 14, pp. 1-20. https://doi.org/10.1093/gigascience/giaf052

Smutek, T., Sikora, J., Bogacki, S., Rutkowski, M. and Woźniak, D. (2024) "Use of Autoencoder and One-Hot Encoding for Customer Segmentation", *European Research Studies Journal*, Vol. XXVII, Special Issue 2, pp. 72-82. https://doi.org/10.35808/ersj/3388

Sundqvist, M., Chiquet, J. and Rigaill, G. (2023). "Adjusting the adjusted Rand Index: A multinomial story". *Computational Statistics*, 38(1), 327–347. https://doi.org/10.1007/s00180-022-01230-7

Syahra, Y., Fadlil, A. and Yuliansyah, H. (2025). "Customer segmentation using RFM and K-Means clustering to support CRM in retail industry". *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 9(3), 1120–1131. https://doi.org/10.33395/sinkron.v9i3.14907

Wang, P. and Ding, H. (2024) "The rationality of explanation or human capacity? Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance", *Information Processing and Management*, Vol. 61, No. 4, Art. 103732. https://doi.org/10.1016/j.ipm.2024.103732

Woltmann, L., Thiele, M. and Lehner, W. (2018) "Modeling Customers and Products with Word Embeddings from Receipt Data", *IDEAS 2018: 22nd International Database Engineering & Applications Symposium*, pp. 246-252. https://doi.org/10.1145/3216122.3229860