algorithms

MDPI

Article

A Hybrid Numerical–Semantic Clustering Algorithm Based on Scalarized Optimization

Ana-Maria Ifrim [1] and Ionica Oncioiu [2,3,*]

[1] Faculty of Industrial Engineering and Robotics, National University of Science and Technology Politehnica Bucharest, 060042 Bucharest, Romania; ana_maria.ifrim@upb.ro
[2] Faculty of Economics and Business Administration, "Eugeniu Carada" Doctoral School of Economic Sciences, University of Craiova, 200585 Craiova, Romania
[3] Academy of Romanian Scientists, 3 Ilfov, 050044 Bucharest, Romania
* Correspondence: ionica.oncioiu@prof.utm.ro

Abstract

This paper addresses the challenge of segmenting consumer behavior in contexts characterized by both numerical regularities and semantic variability. Traditional models, such as RFM-based segmentation, capture the transactional dimension but neglect the implicit meanings expressed through product descriptions, reviews, and linguistic diversity. To overcome this gap, we propose a hybrid clustering algorithm that integrates numerical and semantic distances within a unified scalar framework. The central element is a scalar objective function that combines Euclidean distance in the RFM space with cosine dissimilarity in the semantic embedding space. A continuous parameter λ regulates the relative influence of each component, allowing the model to adapt granularity and balance interpretability across heterogeneous data. Optimization is performed through a dual strategy: gradient descent ensures convergence in the numerical subspace, while genetic operators enable a broader exploration of semantic structures. This combination supports both computational stability and semantic coherence. The method is validated on a large-scale multilingual dataset of transactional records, covering five culturally distinct markets. Results indicate systematic improvements over classical approaches, with higher Silhouette scores, lower Davies–Bouldin values, and stronger intra-cluster semantic consistency. Beyond numerical performance, the proposed framework produces intelligible and culturally adaptable clusters, confirming its relevance for personalized decision-making. The contribution lies in advancing a scalarized formulation and hybrid optimization strategy with wide applicability in scenarios where numerical and textual signals must be analyzed jointly.

Keywords: hybrid clustering; scalarized optimization; numerical algorithms; gradient descent; genetic algorithms

check for updates

Academic Editor: Stoil I. Ivanov

Received: 2 September 2025
Revised: 23 September 2025
Accepted: 25 September 2025
Published: 27 September 2025

Citation: Ifrim, A.-M.; Oncioiu, I. A Hybrid Numerical–Semantic Clustering Algorithm Based on Scalarized Optimization. *Algorithms* **2025**, *18*, 607. https://doi.org/10.3390/a18100607

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In a digital landscape dominated by a continuous and massive accumulation of transactional data, the central challenge for artificial intelligence applied to marketing is no longer the collection of information but its transformation into contextual knowledge and strategic action [1,2]. Traditional segmentation models, based on indicators such as recency, frequency, and monetary value (RFM), provide a useful framework for describing purchasing behavior [3]. However, they only capture the numerical dimension and rarely manage to reflect the subtle intentions of the customer. In contemporary e-commerce, purchases no longer represent only a history of quantifiable transactions but also express a

semantic relationship between customer and product, a relationship influenced by language, culture, and socio-economic context [4–6].

The tension between the numerical precision of data and the semantic richness of implicit meanings highlights the need for models capable of interpreting information simultaneously at both levels [7]. In an environment where products are defined by detailed descriptions, reviews, and contextual metadata, omitting the semantic dimension leads to incomplete and difficult-to-value segmentations [8,9]. In this context, the need for an algorithmic formulation that combines, in a formal and controllable way, both observable behaviors and the meanings associated with purchasing decisions is emerging [10–12].

In parallel, recent developments in the vector representation of natural language have brought powerful tools for meaning modeling [13–15]. Architectures such as Sentence-BERT, RoBERTa, or DistilUSE-multilingual can project linguistic expressions into a coherent semantic space capable of capturing contextual synonymy relationships, implicit intentions, and subtle preferences [16–18]. These technologies allow the quantification of relationships between texts in a flexible and robust way, independent of rigid codings or obvious language barriers. Even so, in marketing research, the integration of these methods with traditional behavioral models is still limited [19]. The two sources of information—numerical and semantic—continue to be used separately, although their combined potential is considerable.

Starting from this discontinuity, the present research aims to answer two major questions:

RQ1: How can a unified segmentation model be built that connects, in a formal framework, the numerical structure of purchasing behavior and the latent semantics of preferred products?

RQ2: To what extent can such a model generate coherent, culturally adaptable, and relevant groupings for strategic decision-making in a context marked by linguistic diversity and volatility of preferences?

In the specialized literature, unidimensional segmentations continue to be dominant, either by focusing exclusively on traditional behavioral metrics or by using linguistic models in isolation [20–23]. This paper proposes an integrative framework that brings together numerical coherence and semantic expressiveness in a single formal formulation. Unlike existing solutions, where behavioral vectors and textual embeddings are processed separately or combined without precise control of their mutual influence, the presented model allows the weighting between the two sources of information to be adjusted by means of a continuous parameter.

The integration of distributed natural language representations into consumer behavior segmentation models has become an intensely explored topic in recent research. For example, Talaat et al. (2023) [24] combine BERT embeddings with the DBSCAN algorithm to automatically detect thematic groupings in consumer reviews, while Gustriansyah et al. (2020) [25] propose a dual-channel model in which RFM scores are analyzed together with semantic vectors obtained through Sentence-BERT. Similarly, Peker et al. (2017) [26] document a multimodal approach that integrates numerical and textual variables in a common learning space for omnichannel retail. Although these contributions indicate the growing interest in hybrid models, they still maintain a methodological separation between the numerical and semantic components.

The present work differs from previous contributions in four major aspects. First, it introduces a scalar objective function that allows for continuous adjustment of the weight between the numerical and semantic components, providing flexibility in the segmentation process. Second, the model is tested in a multilingual and cross-cultural framework, which allows for verification of semantic robustness in various contexts. The third novelty is the use of a hybrid optimization algorithm, which combines gradient

descent with evolutionary mechanisms such as mutation and recombination to extend the exploration capacity of the solution space. Finally, the proposed validation scheme brings together established quantitative indicators, such as Silhouette and Davies–Bouldin, with a qualitative analysis of semantic coherence performed manually. Together, these components configure a methodological framework that overcomes the limitations of existing studies and supports the development of segmentations adaptable to the linguistic diversity and behavioral complexity of contemporary commercial environments.

The remainder of this paper is organized as follows: Section 2 reviews the related literature on numerical and semantic approaches to segmentation. Section 3 describes the proposed methodology, including the scalar objective function and hybrid optimization strategy. Section 4 presents the experimental setup and results, followed by a discussion of their implications. Finally, Section 5 concludes the paper, outlining contributions, limitations, and directions for future research.

2. Literature Review

The rapid transformations of the digital environment have fundamentally changed the way organizations analyze and model consumer behaviors [4,18]. E-commerce platforms, which have become central infrastructures of commercial life, produce massive volumes of data and require analytical methods capable of capturing both the pace and complexity of interactions [6]. In this context, customer segmentation can no longer be reduced to quantitative delimitations, since traditional models, focused exclusively on transactional parameters, offer a uniform image of consumption and blur the real differences between motivations, preferences, and contexts [3,8].

Purchasing behavior today goes beyond the simple frequency of transactions or their monetary value and expresses a symbolic relationship between the customer and the product, where each purchase reflects aspirations, identities, and semantic affinities [4,6]. Segmentation therefore becomes an exercise in articulation between the granularity of data and the expressiveness of meanings, highlighting the need for methods capable of capturing and interpreting the two dimensions simultaneously [27,28]. The connection between transactional and semantic components not only expands the possibilities of understanding the digital customer but also creates the premises for finer, more relevant, and context-adapted commercial interventions.

2.1. RFM as a Classical Approach to Consumer Segmentation

The RFM model is one of the most widespread methods for evaluating the commercial value of a customer, established due to its simplicity and demonstrated efficiency in contexts where behavioral data is numerous but heterogeneous [29]. The three dimensions—recency (R), frequency (F), and monetary value (M)—offer a concise but aggregated representation of the interaction between the customer and the organization. Initially applied in direct marketing and later extended to e-commerce, the model allows for a rapid classification of customers based on transactional history.

The simplicity of the model has facilitated its integration into automated scoring, recommendation, and segmentation systems, especially because it does not require textual data, explicit profiles, or demographic variables [30,31]. However, operational efficiency is accompanied by a significant loss of analytical granularity. Customers with numerically similar shopping patterns can hide major differences in purchasing intentions, consumption style, or cultural significance of the products they choose [27,31,32]. In this way, the model omits the very dimensions that give depth to economic behavior.

In relation to the diversity of contemporary consumption, the RFM model fails to capture the symbolic dimension of purchases. Every purchasing decision involves cognitive,

affective, and contextual processes that cannot be reduced to raw numerical values [4,6,33]. The fundamental limitation lies in the lack of a mechanism through which the meaning of the product is integrated into the customer's representation. However, the universal character of RFM metrics provides a solid starting point for the development of more complex models, capable of including not only the frequency of interactions but also their content [25,28]. Extended by semantic vectors, RFM can become the core of hybrid representations, where the quantitative and symbolic dimensions complement each other.

2.2. Semantic Embeddings in Modeling Consumer Preferences

Recent advances in natural language processing have allowed a reconceptualization of the way in which text can be represented mathematically [34–36]. Beyond traditional frequency- or co-occurrence-based techniques, transformer models create projections of linguistic expressions into high-dimensional latent spaces, where semantic relationships can be quantified with remarkable accuracy [36,37]. Among the architectures currently used in semantic analysis is Sentence-BERT, whose embeddings capture contextualized meaning structures while preserving the metric properties necessary for tasks such as clustering, similarity, or classification [16–18].

A significant advantage of embeddings is their ability to capture contextual equivalences between apparently different expressions. Product descriptions from distinct linguistic and cultural spaces can be analyzed in a unified framework, where meaning does not depend exclusively on the lexicon but on semantic positioning in the corpus [35]. The resulting re-presentations capture both the local coherence of the text and the global relationships between concepts, allowing the highlighting of latent structures difficult to identify by explicit rule-based methods [37].

Embedding-based representation transforms the textual description of a product into a numerical vector that can be directly integrated into a quantitative analysis framework. Instead of predefined labels or manual coding, the model automatically extracts relevant features, without requiring explicit supervision [38–40]. In this way, the semantic space becomes accessible to optimization and segmentation algorithms, which can work with dense and continuous representations, preserving the implicit meaning of the original text [40,41]. The application of embeddings in customer segmentation opens up major analytical perspectives. The preferences expressed through purchases are no longer viewed exclusively quantitatively but also through the prism of the semantic congruence between the customer profile and the universe of meaning of the consumed products [23,31,42]. The vectors obtained become the bridge through which semantics is converted into a measurable form, and linguistic complexity is transformed into a computationally analyzable structure [27,43]. When numerical data derived from transactional vectors are correlated with semantic embeddings, the result is a heterogeneous space where distances no longer express a single metric but a subtle balance between behavioral regularity and meaning compatibility.

2.3. Hybrid Clustering Approaches Integrating Numerical and Semantic Dimensions

Most clustering algorithms were designed to operate in a one-dimensional space—numeric, binary, or categorical. In situations where heterogeneous variables are combined, as in the K-Prototypes algorithm, the balance between the components becomes difficult to control, and the results tend to be influenced by the dominance of one of the dimensions [44–46]. Direct integration of semantic embeddings into such structures can generate distortions, since semantic vectors possess distinct geometric properties, derived from contextual learning processes and not from empirical observation [38,47].

To achieve coherent segmentation in a mixed space, a mechanism is needed to regulate the relative contribution of each component within the objective function. The introduction of the parameter λ provides a flexible tool through which the influence of semantic and transactional distances can be adjusted according to the specifics of the analysis [48–50]. Such adjustment allows for the controlled exploration of the entire spectrum, from exclusively numerical segmentation to clustering based predominantly on product semantics.

The merging of the two types of information must be performed under conditions that maintain both topological consistency and algorithmic convergence. Integration into a unified metric space involves not only the numerical aggregation of distances but also their interpretative compatibility [20,22,28]. In order for the results to be computationally stable and intelligible for human analysis, the objective function must be formulated in such a way as to allow for the controlled exploration of solutions, without losing semantic coherence or behavioral relevance [19,44].

Integrating the behavioral and semantic dimensions into a unified formal framework requires the definition of an objective function that balances the contributions of the two sources of variation in a controllable manner. The scalarization achieved through the λ parameter allows for fine-tuning the weight between transactional and semantic distances, providing analytical flexibility in a composite space [31,45]. In this formulation, segmentation is no longer just a projection of the frequency and value of purchases but also becomes an expression of the meaning of consistency between the selected products and the symbolic universe of the consumer [28,33,41].

The formulation of the objective function starts from the combination of two types of distances: one derived from the RFM transactional vectors and the other from the semantic embeddings. The two components are brought together in a flexible mathematical structure, regulated by the λ parameter [25,28,51]. Extreme values of the parameter lead to unidimensional scenarios, and intermediate ranges allow for the articulation of a balance between numerical consistency and semantic coherence. In this way, segmentation is no longer a rigid process but becomes an adaptable mechanism, capable of exploring the entire spectrum of consumer behavior.

The objective function constructed in this manner is continuous and differentiable, which makes it possible to use stable optimization strategies, either gradient-based or through hybrid mechanisms [11,24]. Scalarization not only facilitates algorithmic convergence but also provides an interpretable framework, where semantic tuning can adapt to the analyzed context. This integration supports the formation of clusters that simultaneously reflect transactional regularities and the latent dimension of meanings.

In the recent literature, deep clustering models, such as Deep Embedded Clustering, Variational Deep Embedding (VaDE), or Deep Clustering with Convolutional Autoencoders (DCEC), have demonstrated the ability to simultaneously learn latent spaces and clustering structures [52–54]. However, their application involves complex parameterization, high consumption of computational resources, and difficulties in interpretation. Furthermore, the control over the weight between the numerical and semantic dimensions remains limited, which reduces the adaptability of the results.

Comparatively, the mechanism formulated here introduces an adjustable balance between the numerical and semantic signal, avoiding the complexity of difficult-to-explain neural architectures. Explicit scalarization allows adaptation to varied linguistic contexts and supports reproducible, understandable, and applicable segmentations in real business scenarios [28,31,51]. The contribution consists of defining a scalar objective function with parameter λ, testing it in a multilingual framework, and validating it through a combination of quantitative metrics and qualitative assessments. In this way, the research proposes

a model that overcomes the limitations of existing studies and strengthens the bridge between numerical and semantic analysis in customer segmentation.

3. Materials and Methods

The paper aims to develop a behavioral segmentation mechanism that combines the numerical signal of purchases with the semantic expression of preferences in a scalable and controllable formulation. The motivation is not limited to improving the performance of clustering metrics but aims to bring the analysis closer to the way people interpret consumer behavior. Thereby, segments no longer represent just statistical groups but become recognizable, intelligible, and strategically usable profiles in decision-making processes.

The first specific objective is to formulate a mathematical framework in which numerical and semantic distances can be combined through a continuously parameterized expression, maintaining interpretive coherence. To this end, a scalarized objective function is introduced, dependent on an adjustable parameter, which allows the regulation of the relative weight between the two components. Beyond formalism, the analysis also aims to demonstrate the practical viability of an optimization strategy capable of identifying stable, reproducible, and semantically consistent clusters in a composite metric space.

The proposed methodological contribution is based on the definition of a scalar objective function compatible with a hybrid optimization mechanism, which combines the numerical precision of gradient descent with the heuristic flexibility of genetic algorithms. The integration of the two components expands the space of the analyzed solutions and supports the identification of stable topologies directly correlated with the semantic coherence and behavioral consistency of the examined customers.

3.1. Dataset Description

The analysis uses the Open e-commerce 1.0 dataset [55,56], published in Nature Scientific Data (2024), which contains over 1.8 million transactions made by 5027 unique customers on the Amazon platform between 2018 and 2022. The data is fully anonymized according to GDPR and includes transactional information, behavioral metadata, textual product descriptions, and contextual variables.

Each customer profile was built based on RFM (Recency, Frequency, Monetary Value) dimensions, and each product is described by text in multiple languages. After data cleaning, 2300 unique product descriptions were retained, covering domains such as technology, books, home & deco, lifestyle, and education. The descriptions were vectorized using the distiluse-base-multilingual-cased-v1 (Sentence-BERT) model, generating *512-dimensional embeddings*. Five national subsets (USA, Germany, France, Italy, and Japan) were selected for the experiment, each including approximately 1000 customers and a proportional volume of transactions, totaling 5000 customers and over 100,000 purchases processed.

Table 1 presents the variables used in the analysis and their role in the model. The structure of the dataset allows for the combined investigation of numerical behavior and semantic preferences, constituting the empirical support for the proposed scalar clustering mechanism.

Table 1. Structure of variables used in the analysis.

Variable	Description	Use in Model
Customer_ID	Anonymized unique identifier for each customer	Aggregation per individual
Purchase_Timestamp	Date and time of each transaction	Recency calculation

Table 1. *Cont.*

Variable	Description	Use in Model
Product_ID	Unique code associated with each product	Semantic vector linking
Product_Description	Full product description (natural language)	Semantic embedding generation
Product_Category	Product taxonomic label	Semantic control (optional)
Price	Individual monetary value of the transaction	Monetary value calculation
Quantity	Number of units purchased	Purchase volume calculation
Language_Code	Language code of the product description (e.g., EN, DE, FR, ES, RO)	Cultural semantic adaptation
Country_Code	Geographic region associated with the customer	Cultural contextual control
Review_Score	Subjective product rating (1–5 stars)	Complementary satisfaction analysis
Review_Text	Post-purchase textual comment (optional, not all have reviews)	Advanced semantic analysis (exploratory)

Note: Only strictly necessary variables were used for model construction, but the rest may contribute to additional explorations, external validation, or post-clustering interpretability.

3.2. Country Selection Criteria

Setting up an experimental framework capable of supporting the validation of a semantic segmentation mechanism, applicable in real commercial contexts, required a careful and reasoned selection of a subset of countries from the Open e-commerce 1.0 dataset. The choice was not guided exclusively by numerical thresholds but by the desire to build a heterogeneous sample, representative from a cultural and linguistic perspective, able to highlight the robustness of the model in the face of variation in behavioral meanings.

The first filter applied referred to the empirical consistency of the data: only those regions were retained in which the volume of transactions and the number of customers allow the extraction of stable, reproducible, and statistically significant patterns. In the absence of a critical mass, interpretations become fragile, and algorithmic aggregations risk reflecting noise rather than structure.

In parallel, a criterion of deliberate cultural dispersion was introduced. The aim was to outline a transnational ensemble in which differences in language, consumption habits, and lexical expressiveness would put semantic vectorization mechanisms under pressure and allow tests of coherence in the latent interpretation of preferences.

The third criterion targeted the degree of maturity of the national digital ecosystem. Markets with consolidated commercial infrastructures and a stabilized digital culture generate data that reflects authentic choices rather than technical constraints or logistical distortions. In such a framework, observable behavior becomes a reliable indicator for computational modeling.

Based on these considerations, five countries with deeply differentiated profiles were included. The United States offers a high transactional density and an advanced digital context, with behavioral and lexical diversity [57]. Germany exhibits a more algorithmically predictable consumption style, while France introduces significant cultural variability in product selection and descriptive register. Research on disparities in internet purchases

across EU countries supports these cultural distinctions in consumption patterns [58]. Japan, by its agglutinative nature and idiosyncratic acquisition structure, presents a technical challenge for the semantic model. Italy completes this picture with an affective dimension visible in seasonal choices, in sensitivity to aesthetics, and in the stylistic preferences expressed in text.

The linguistic distribution—English, German, French, Japanese, and Italian—allows testing whether multilingual embedding models can capture semantic nuances across disjunctive lexical contexts. Though multilingual embeddings can reduce performance gaps (e.g., English to Italian), consistency may suffer in specialized domains [59], particularly when language representation is imbalanced [60–62]. By aggregating these features, this validation framework transcends quantitative differences and enables a combined evaluation of algorithmic performance on both numerical and cultural-semantic levels.

3.3. Processing RFM Variables

The RFM model offers a quantitative scheme established in the analysis of purchasing behavior, through which the relationship between the temporality of purchases, the intensity of interaction, and the economic value of each consumer is structured [3]. Through three distinct dimensions—recency, frequency, and monetary—a three-dimensional representation is outlined capable of capturing both the regularity and the commercial impact of each customer profile.

Recency is expressed by the distance in days between the last purchase and the final date of the records in the data set, functioning as an indicator of the continuity of the commercial relationship. Frequency reflects the degree of customer involvement, being quantified by the total number of transactions recorded in a reference period. The monetary component is derived from the sum of the expenses made by multiplying the quantity by the price of each purchased item. Formally, this last dimension is expressed as follows:

$$Monetary_i = \sum_{j=1}^{n_i} quantity_{ij} \cdot price_{ij}, \quad (1)$$

where n_i represents the number of transactions associated with customer i, and the product of quantity and price indicates the value of each purchase.

To obtain a comparable representation, each of the three variables was subjected to a Min-Max transformation, applied uniformly to the entire sample. The scaling formula is:

$$x_i^{norm} = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \quad (2)$$

This normalization step in the interval [0, 1] ensures the elimination of the dominance of one component over the others and facilitates the numerical convergence of the algorithms subsequently used in the segmentation process. In parallel, the extreme values were treated statistically, in order to preserve the robustness of the aggregate distributions.

RFM vectors not only encode quantitative structures but open the way to an interpretable typology of purchasing behavior. The resulting configurations can clearly delineate consumer profiles at opposite extremes of commercial involvement: from customers with isolated interactions, to loyal customers with recurring economic impact. Within this coherent vector space, behavioral differences take on numerical contours, preparing the ground for the integration of the semantic dimension and for the construction of a segmentation that is not limited to statistics, but aspires to a deep contextual understanding.

3.4. Semantic Representation of Consumer Preferences

The integration of the semantic dimension into the advanced computational structure allows the exploitation of textual descriptions of purchased products as a complementary

source of behavioral information. For each customer, the descriptions associated with purchased products were aggregated into a unified document, acting as a textual proxy of the consumer preferences implicitly expressed through purchasing actions.

The semantic projection was performed using the pre-trained Sentence-BERT multilingual model, which generates contextualized vectors in the propositional space. The dimensionality of the semantic vectors was set to $d = 512$, a value that reflects the architectural structure of the model and the desired balance between semantic granularity and computational stability.

Each customer i is thus associated with two vector components:

$x_i^{(1)} \in \mathbb{R}^3$, result of the RFM model;

$x_i^{(2)} \in \mathbb{R}^{512}$, derived from the semantic aggregation of purchased products.

The composite representation is expressed by:

$$x_i = \left(x_i^{(1)}, x_i^{(2)}\right) \in \mathbb{R}^3 \times \mathbb{R}^{512}. \tag{3}$$

The distance between a customer and the center of a behavioral group u_j is defined by a scalar function that combines two dissimilarity components:

$$d\left(x_i, \mu_j\right) = \lambda \cdot \left\| x_i^{(1)} - \mu_j^{(1)} \right\|_2^2 + (1-\lambda) \cdot \left(1 - \cos\left(x_i^{(2)}, \mu_j^{(2)}\right)\right) \tag{4}$$

where $\|\cdot\|_2^2$ measures the difference in the transactional numeric space,

$1 - \cos(\cdot)$ expresses the semantic dissimilarity between the embedding vectors,

$\lambda \in [0, 1]$ allows the relative adjustment of the influence of each dimension.

The resulting mathematical structure preserves formal coherence and derivability properties, creating the necessary conditions for its inclusion in a stable and efficient algorithmic optimization framework. The parametric tuning offered by λ allows the adaptation of the interpretive granularity according to the cultural and linguistic specificity of the analyzed context, contributing to a more nuanced and contextually anchored segmentation of consumer behaviors. The overall workflow of the proposed methodology is summarized in Figure 1, which illustrates the complete pipeline from data preprocessing to dual representation, scalarized objective function, hybrid optimization, and evaluation.

3.5. Objective Function and Optimization Model

In the construction of a behavioral segmentation model that reflects not only the intensity of commercial interaction but also the latent meaning of individual choices, it became necessary to define an objective function that would allow the coherent articulation of two epistemologically different dimensions: one quantifiable through RFM vectors and the other derived from the semantic topology of purchased products. To this end, a scalarizable expression of distance was formulated, conceived as a weighted sum of two components—the Euclidean distance in numerical space and the cosine dissimilarity in semantic space. This methodological integration gives the segmentation mechanism the ability to capture both the transactional regularity and the symbolic context of consumption.

Mathematically, the total distance between a client i and the center of a cluster c_k is defined as follows:

$$D_{i,k}^{(\lambda)} = \lambda \cdot \|r_i - r_k\|_2 + (1-\lambda) \cdot (1 - \cos(s_i, s_k)) \tag{5}$$

where $r_i \in \mathbb{R}^3$ represents the customer's numerical vector, built on the three fundamental dimensions of the RFM model: recency, frequency, and monetary value of transactions. Complementarily, $s_i \in \mathbb{R}$ denotes the agreed semantic vector, obtained by embedding the coding of the textual descriptions of the products purchased by the respective customer.

The topological variation between these components is related to the center c_k, which has both a numerical centroid r_k and a semantic centroid s_k, characteristic for each cluster.

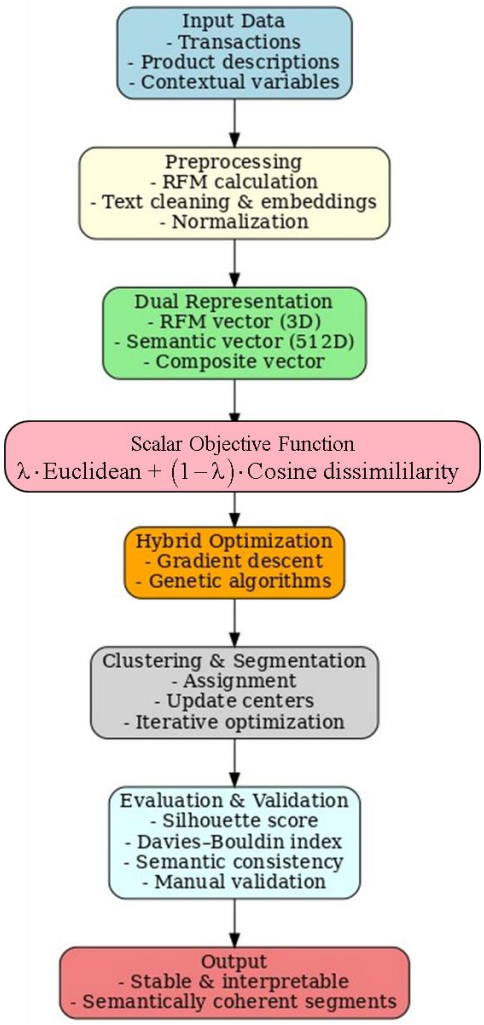


Figure 1. General framework of the proposed numerical–semantic segmentation methodology.

The parameter $\lambda \in [0,1]$ modulates the relative contribution of each component to the objective function, providing contextual flexibility and cultural adaptability. A value close to 1 privileges the numerical dimension, suitable in contexts where purchasing behavior is predictable and regularized, while a low value brings the segmentation closer to the semantic structure of interests, suitable in places with symbolic expression or idiosyncratic preferences.

The formulation has not only instrumental value but also epistemological significance: the Euclidean component expresses the quantitative deviation from the dominant behaviors of a group, while the cosine dissimilarity quantifies the conceptual alignment or distance from the semantic profile of that group. From the perspective of computational geometry, the embedding vectors, being densely distributed in a latent space, do not respect the traditional metrics of the Euclidean distance, which is why the use of cosine similarity becomes a justified methodological choice for measuring semantic proximity.

In addition, the continuity and derivability of this function with respect to the parameter λ allow for natural integration into an iterative optimization process, facilitating both adaptation to the specifics of the data and convergence towards explainable solutions. Through the mathematical balance established between the numerical and semantic components, the hybrid objective function transcends the limitations of one-dimensional

segmentation, providing a formal framework for modeling a consumer conceptualized not only as a transaction generator but also as a contextual entity located in its own universe of meanings.

Regarding the choice of the value of the parameter λ, it is not universally optimizable but depends on the particularities of the data set and the objectives of the analysis. In the present study, several values were tested (e.g., 0.1, 0.5, 0.9), and the final selection was made based on the simultaneous maximization of the Silhouette and Davies–Bouldin indices and manually assessed semantic coherence [63–65]. Hence, λ becomes an analytical and interpretative tuning tool, adaptable to different contexts, and can be calibrated either empirically or through cross-validation procedures. While this flexibility is a major advantage, it also implies an additional step of contextual adjustment in future applications.

3.6. Algorithmic Fundamentals: Numerical Scalarization and Dual-Space Optimization

The integration of a scalarizable objective function into a behavioral segmentation model requires a rigorous anchoring in the numerical optimization literature, especially in the field of multi-objective optimization. The scalarization used in the proposed formulation follows the tradition established in the works of Miettinen [63] and Deb [64], in which the balance between divergent objectives is controlled by weighting parameters. In the present context, the parameter λ regulates the relative importance of the Euclidean distance (numerical) and the cosine dissimilarity (semantic) in a manner similar to weighted-sum scalarization methods.

By formulating the objective as a linear combination between two distinct metric components, a continuous, differentiable search space is created, favorable to gradient-based optimization algorithms. Thus, the proposed model can be interpreted as a dual extension of the classical K-Means paradigms, which adds a symbolic dimension to the aggregation process. From a geometric point of view, the optimization problem is carried out in a composite space with mixed topology, which justifies the use of a hybrid algorithmic regime—directed descent in the numerical subspace and heuristic exploration (of the genetic type) in the semantic subspace.

Such approaches are frequently encountered in evolutionary clustering and adaptive optimization problems, in which heuristic components efficiently complement deterministic trajectories to avoid blockages in local minima [22,25,44,65]. In the present case, scalarization not only modulates the objective function but also acts as a semantic control tool, capable of regulating the interpretative granularity of the formed segments.

Therefore, the methodological integration between quantitative distances and symbolic similarities goes beyond the level of a simple combination of metrics, providing a solid formal framework for exploring latent consumption structures that could not be highlighted by unidimensional techniques.

3.7. Optimization Algorithm

The optimization process built to minimize the scalar objective function involves a sequence of independent steps, through which the configuration of the cluster centers is continuously refined, in a hybrid framework that simultaneously capitalizes on the differentiable geometry of the numerical space and the heuristic flexibility of genetic algorithms. The algorithm begins with a statistical initialization, in which k initial centers are randomly generated, each defined by two distinct components: a numerical vector $\mu_i^{(1)} \in \mathbb{R}^3$, associated with the RFM dimensions, and a semantic vector $\mu_i^{(2)} \in \mathbb{R}^{512}$, derived from the aggregation of textual embeddings. At the same time, the values of the control parameters are established: the equilibrium scale $\lambda \in [0, 1]$, the learning rate η, the regularization coefficient ρ, the maximum number of epochs, and the stopping threshold ϵ.

A methodological issue that requires clarification concerns the order in which the two optimization mechanisms are applied. In the classical literature on *memetic algorithms*, the evolutionary component typically precedes the local search stage. However, several studies highlight the advantages of reversing this sequence by first employing gradient descent to stabilize the numerical centers, followed by genetic operators to explore the semantic space of solutions. This strategy has been reported, for instance, in memetic algorithms for textual clustering, where integrating a local search step before the evolutionary operators improved cluster quality and yielded more stable convergence [10]. In parallel, the literature on multi-objective optimization indicates that weight-based scalarization functions can be effectively combined with gradient-based updates to accelerate convergence, even prior to the application of heuristic components [16]. Likewise, research applying gradient descent before evolutionary mechanisms shows that this ordering reduces the risk of entrapment in local minima and facilitates exploration of the global structure [66]. Accordingly, the decision to apply gradient descent prior to genetic operators is consistent with these methodological recommendations, ensuring both numerical stability and semantic diversity in the proposed algorithm.

In the next step, each client x_i is associated with that cluster j for which the hybrid support function $d(x_i, \mu_j)$ reaches a local minimum. The assignment relation is formalized by the binary variable $z_{i,j}$, as follows:

$$z_{i,j} = \begin{cases} 1, \ if \ j = \arg min_j d(x_i, \mu_j), \\ 0, \qquad\qquad otherwise. \end{cases} \tag{6}$$

This distance function is defined as a scalarized combination of the Euclidean distance in the RFM space and the cosine dissimilarity in the semantic space, thus capturing the dual nature—quantitative and symbolic—of the relationship between the individual and the center.

After assigning observations, the process continues with updating the centers, a stage in which gradient descent is used to iteratively adjust the parameters and ensure numerical convergence. The vectors $\mu_j^{(1)}$ are modified according to the rule:

$$\mu_j^{(1)} \leftarrow \mu_j^{(1)} - \eta \cdot \frac{\partial \mathcal{J}_\lambda^\rho}{\partial \mu_j^{(1)}} \tag{7}$$

where $\mathcal{J}_\lambda^\rho$ represents the scalarized and penalized objective function. Similarly, the semantic vectors $\mu_j^{(2)}$ are updated by a parallel mechanism, but adapted to the partial derivatives of the cosine component:

$$\mu_j^{(2)} \leftarrow \mu_j^{(2)} - \eta \cdot \frac{\partial \mathcal{J}_\lambda^\rho}{\partial \mu_j^{(2)}} \tag{8}$$

with the derivatives obtained through the analytical formalism developed previously. In this stage, the optimization occurs simultaneously in two vector spaces of different natures but which interact through the defined objective function.

To prevent blockages in suboptimal regimes, an evolutionary phase is activated every T epochs, in which a fraction of the existing centers is subjected to stochastic transformations of the mutation type—modeled as controlled perturbations—and recombination, through which two vectors are combined to generate new configurations. This strategy allows the exploration of alternative regions of the solution space and contributes to maintaining the structural diversity of the global configuration.

The stopping criterion is based on the analysis of the variation in the centers between successive epochs. If the Frobenius norm of the difference between the updated and previous positions satisfies the relation:

$$\left\| \mu^{(t+1)} - \mu^{(t)} \right\|_F < \epsilon, \tag{9}$$

the execution is considered convergent, and the process is interrupted. In the absence of this condition, the assignment and update cycle are resumed until the established threshold is satisfied or until the maximum number of epochs is reached.

The evaluation of the algorithmic complexity indicates that a complete optimization cycle has the complexity:

$$\mathcal{O}(nkd) + \mathcal{O}(k(d+3)) + \mathcal{O}(T_{GA}), \tag{10}$$

where n represents the number of customers, k the number of clusters, $d = 512$ the size of semantic embeddings, and T_{GA} reflects the additional cost associated with genetic operators. These formulations highlight the high potential for parallelization, either on the size of observations or on that of centers, which makes the algorithm compatible with efficient execution on large data sets.

Overall, the proposed algorithm transcends simple numerical optimization, managing to integrate in a fluid way the symbolic dynamics of individual preferences and the transactional structure of consumer behavior, and the hybrid architecture, coupled derivative-evolutionary, underpins an adaptive and expressive segmentation process, faithful to the complexity of the modeled phenomenon.

For comparative purposes, besides the proposed hybrid model, we implemented classical baselines—K-Means on RFM vectors and semantic-only clustering on embeddings—and extended the evaluation with Gaussian Mixture Models (GMM) and Hierarchical Agglomerative Clustering, both widely used unsupervised segmentation techniques.

The stepwise functioning of the proposed hybrid optimization algorithm is outlined in Figure 2, which details the iterative assignment, gradient-based updates, periodic evolutionary operations, and convergence check.

In order to analyze the structural fidelity of the obtained segmentation, the evaluation process integrated both established quantitative methods and indicators designed to reflect the semantic consistency of the resulting groupings. The plurality of perspectives allowed not only to quantify algorithmic performance but also to correlate the results with the latent meanings of the input data. One of the fundamental indices used in the internal evaluation was the Silhouette score, designed to simultaneously estimate internal cohesion and separability between clusters.

For each observation i, the score is calculated as:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{11}$$

where $a(i)$ denotes the average distance between observation i and all other members of the cluster to which it belongs, and $b(i)$ represents the average distance from the members of the nearest neighboring cluster. A score close to 1 indicates a clearly delimited structure, while negative values signal potential allocation errors. For comparative purposes, the average silhouette score at the complete set level was used for all methods analyzed.

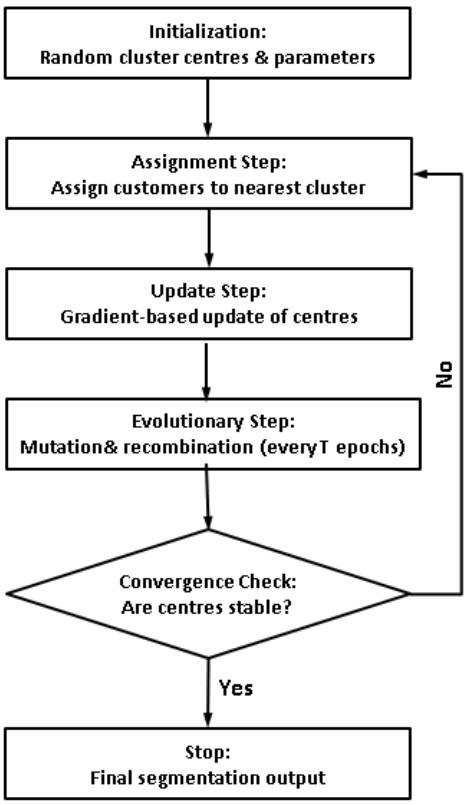


Figure 2. Algorithmic flow of the hybrid optimization under the scalarized objective.

A second quantitative benchmark was provided by the Davies-Bouldin index, which expresses the ratio between internal cohesion and disjunction between clusters, as follows:

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(\mu_i, \mu_j)}\right) \tag{12}$$

where σ_i and σ_j represent the internal dispersion of clusters i and j, and $d(\mu_i, \mu_j)$ is the distance between the centers of the two clusters. A low value of this index is associated with a more efficient segmentation, in the sense of a clear differentiation between groups and a compact aggregation within each cluster.

Beyond numerical consistency, special attention was paid to the semantic dimension of segmentation. In order to capture the degree of internal symbolic coherence, a dedicated indicator was introduced—the intra-cluster semantic consistency (SC_j) calculated according to the formula:

$$SC_j = \frac{1}{|C_j|}\sum_{x_i \in C_j}\frac{\left\langle x_i^{(2)}, \mu_j^{(2)}\right\rangle}{\left\|x_i^{(2)}\right\| \cdot \left\|\mu_j^{(2)}\right\|} \tag{13}$$

where C_j is the set of customers associated with cluster j, and $x_i^{(2)}$ and $\mu_j^{(2)}$ are the embedding vectors corresponding to customer i and the semantic center.

Beyond the quantitative parameters, a qualitative assessment was performed on the resulting profiles, focusing on the clarity and semantic consistency of the product descriptions associated with each cluster. The analysis, carried out manually on a representative subset of customers from several geographical contexts, aimed to establish to what extent segmentation generates recognizable, narratively coherent, and strategically relevant groups for the application of differentiated marketing policies. In most cases, the identified

segments proved interpretable in terms of recurring purchasing behaviors, recognizable consumption styles, or common thematic preferences.

In order to establish the real contribution to the described algorithmic configuration, the results were compared with two reference variants. The first variant uses the K-Means algorithm applied exclusively to the RFM transactional dimensions, completely ignoring the semantics of the products. The second variant assumes a clustering performed only in the space of semantic embeddings, using the cosine distance as the sole aggregation criterion. By reporting to these methodological extremes, the hybrid configuration presented here demonstrated a superior capacity to generate simultaneously compact, distinct, and interpretable segments. Variations in the value of the parameter λ were systematically analyzed to identify the equilibrium regions in which the semantic contribution maximizes cohesion without compromising numerical fidelity, and these analyses confirmed the advantage of scalarization in capturing the complexity of consumer behavior.

In consequence, the evaluation of the quality of segmentation was carried out in a multi-perspective framework, in which computational efficiency, geometric coherence, and symbolic meaning were rigorously correlated, providing a robust validation of the proposed model from both a theoretical perspective and direct application angles. Compared to the reference variants—classical K-Means based solely on RFM variables and clustering based exclusively on semantic embeddings—the proposed dual-space optimization consistently outperforms both extremes. Through the explicit control of the balance between numerical and semantic signals, the model achieves superior cohesion, separability, and interpretability. This methodological integration constitutes the main novelty of the study, positioning the approach as a reproducible and scalable alternative to existing unidimensional or black-box deep clustering methods.

4. Results and Discussion

The central goal of the experimental set was to examine the methodological performance of a segmentation system built on the integration of two distinct conceptual dimensions: traditional numerical measurements, extracted from RFM data, and deep semantic representations, derived through embedding techniques applied to textual descriptions of purchased products. The analysis was oriented towards investigating the differential contribution brought by the semantic component in shaping the grouping structure, with an emphasis on the logical coherence of the segments formed but also on the practical validity of these structures in real commercial contexts.

At the heart of the concerns was the exploration of the effect produced by the integration of the symbolic dimension on the granularity of segmentation, in correlation with the stability of the model in multicultural contexts and with high behavioral diversity. Also, special attention was paid to the evaluation of the interpretative potential of the resulting segments in relation to their capacity to support strategic decisions in the sphere of differentiated marketing. Therefore, the experimental design was built to allow a rigorous examination, both from the perspective of algorithmic performance and from the angle of commercial relevance of the identified segments.

To evaluate the comparative behavior of the proposed method in relation to established alternatives, systematic experiments were conducted in which each segmentation model was applied to the same data set in a controlled configuration. The experimental design included three distinct clustering paradigms: K-Means applied exclusively to RFM indicators, Gaussian Mixture Models (GMM) on RFM variables, a semantic-only variant based on embeddings and cosine distance, hierarchical agglomerative clustering in the embedding space, and the hybrid model introduced in this work, characterized by controlled scalarization and iterative updating of centroids in both spaces.

For each of these methods, three indicators were calculated to validate the quality of the segmentation: the Silhouette score, the Davies–Bouldin index, and the average level of intra-cluster semantic coherence. In addition, the average execution time was recorded to evaluate the computational efficiency of each method. The summarized values are presented in Table 2, obtained by averaging the results for all analyzed countries. The analysis of the obtained values reveals a consistent superiority of the hybrid model compared to both methodological extremes.

Table 2. Comparison of segmentation methods (averages across countries).

Applied Method	Silhouette Score	Davies–Bouldin Index	Semantic Coherence	Execution Time (s)
K-Means (RFM only)	0.41	1.38	0.44	1.2
GMM (RFM only)	0.45	1.31	0.46	1.6
Semantic-only (embeddings)	0.36	1.72	0.61	4.5
Hierarchical (embeddings)	0.39	1.55	0.64	7.8
Proposed hybrid model	0.52	1.04	0.72	6.3

The analysis of the obtained values reveals a consistent superiority of the hybrid model compared to both methodological extremes. The significant increase in the Silhouette score indicates a more clearly defined cluster structure, with a pronounced internal cohesion and a robust separability between groups. In parallel, the decrease in the Davies–Bouldin index value reflects a reduction in the interference between segments and a better-controlled internal dispersion. From a semantic point of view, the increased level of intra-cluster coherence confirms the symbolic alignment of customers within each group, providing a solid basis for relevant commercial interpretations.

Although the hybrid method involves a slightly higher execution time, this difference is justified by the additional complexity of the dual optimization and the obvious benefits in terms of segmentation quality. The integration of symbolic representations, along with transactional dimensions, led to the formation of more homogeneous semantic structures, without sacrificing numerical geometry or metric rigor. Overall, the experimental data support the validity of the developed unified mechanism as a robust and scalable solution for complex segmentation scenarios, where both observable behavior and latent meanings of consumer choices need to be treated in a unified framework.

To verify the stability of the model in varied socio-cultural contexts, the methodology was applied separately on data from five distinct countries: the United States, Germany, France, Italy, and Japan. Each space provided a unique configuration of consumer behaviors, purchasing styles, and latent preferences, which allowed observing how the numerical component and the symbolic structure of the data interact within the proposed algorithmic architecture.

The country-specific insights discussed in the following subsections are not derived directly from the raw dataset but are inferred from the semantic centroids of the clusters generated by the proposed hybrid algorithm. Each conclusion reflects the dominant descriptors associated with the centroid of a cluster in which customers from a given country were statistically overrepresented. To ensure methodological robustness, these interpretations were validated through intra-cluster semantic coherence scores as well as through the observed distributional patterns of customers across clusters. In this way, the subsequent

descriptions of national consumer tendencies should be understood as outcomes of the clustering process rather than as immediate properties of the underlying data.

In the North American commercial environment, the numerical behavior of customers was distinguished by a high degree of regularity and predictability. The corresponding patterns were identified through the centroid descriptors of clusters predominantly composed of North American consumers, indicating that the observed regularity emerged from the clustering structure rather than directly from raw frequencies. The frequency of purchases and the recurrence in RFM patterns contributed to the formation of well-separated segments in the numerical space. The semantic dimension completed this structure by identifying a pronounced attraction for products in the fields of technology, reading, and interior design. The synergy between these two dimensions led to a coherent segmentation, both geometrically and symbolically.

In Germany, the distribution of customers revealed a marked preference for utilitarian products, well-structured thematically. Evidence for this tendency comes from the semantic centroid of the cluster most strongly aligned with German consumers, marked by descriptors that emphasized practicality and thematic clarity. The semantic vectors associated with the purchased items reflected a clear organization of interests, which facilitated the formation of compact clusters with a high degree of differentiation. The correlation between transactions and symbolic representations was one of the most balanced in the entire geographical sample, which provided a favorable context for validating the hybrid formalism.

In France, the RFM values presented a relatively low dispersion, which would have made it difficult to obtain expressive segments in the absence of semantic components. The profile was inferred from the centroid of clusters containing a significant proportion of French consumers, which revealed preferences that could not be detected solely on the basis of raw numerical values. However, the analysis of the embeddings allowed the identification of subtle features in the preferences of customers oriented towards products with high aesthetic content, decorative objects, personalized accessories, or cultural items. Semantic granularity compensated for numerical homogeneity, generating segmentation structures with high interpretative value in marketing oriented towards lifestyle and identity.

Italy represented a space in which semantic relevance manifested itself with particular clarity. The national profile was derived from the centroidal descriptors of clusters dominated by Italian consumers, which consistently highlighted affinities for artisanal and stylistic features. While the RFM values were generally moderate, the embeddings associated with product descriptions revealed an affinity for artisanal goods, seasonal objects, and gifts with a symbolic character. Without the intervention of the semantic representation, the differentiation between customers would have remained difficult to achieve. In contrast, the integration of the symbolic dimension allowed the formation of segments with recognizable stylistic and cultural features, in accordance with the specifics of Italian consumption.

In Japan, the performance of the semantic model was influenced by the cultural and linguistic discrepancies between the available data and the training space of the embeddings used. The national profile resulted from the semantic centroid of clusters with a significant Japanese presence, though interpretability was partly constrained by limitations in the embedding corpus. Although the clustering mechanism worked within acceptable algorithmic parameters, semantic interpretability was affected by the fact that local cultural meanings were not fully captured by the symbolic vectors, but the limitation does not derive from the architecture of the model itself but from the nature of the pre-existing corpus, calibrated in predominantly Western environments. The Japanese context

thus signaled the need for additional linguistic adaptation, aimed at optimizing interpretive consistency in distinct cultural spaces.

To provide a comparative quantification of these observations, Table 3 summarizes the model performance for each country, based on the three evaluation dimensions used: internal cohesion, separability, and semantic consistency.

Table 3. Hybrid model performance in each country analyzed.

Country	Silhouette Score	Davies–Bouldin Index	Semantic Coherence
USA	0.54	1.01	0.71
Germany	0.58	0.96	0.69
France	0.48	1.10	0.74
Italy	0.50	1.07	0.70
Japan	0.43	1.28	0.63

The transnational application of the model demonstrated remarkable structural stability, with predictable variations depending on the cultural specificity of each market. The simultaneous integration of transactional data and symbolic preferences proved effective in generating segmentation structures that combine numerical relevance and semantic expressiveness, without depending on a single reference culture. The results obtained highlight the potential of the method to provide adaptive solutions sensitive to the cultural complexity of consumer behavior.

A complementary direction aims to evaluate the resulting segments from the perspective of human interpretability and semantic consistency in different cultural contexts. In this view, segmentation is not only an exercise in computational processing but also a means of algorithmically translating complex behavioral realities. The ability to generate intelligible groupings has the potential to support concrete decisions in personalized digital marketing and, at the same time, to provide organizations with analytical tools adaptable to the plurilingual and multicultural specificity of contemporary markets.

The application of the hybrid algorithm on the data set led to the emergence of segmentation structures with high strategic relevance, in which customer behavioral profiles can be easily interpreted and transformed into coherent operational decisions. The model allowed a simultaneous articulation between the numerical dimension and semantic expressiveness, which generated a typology of clusters compatible with differentiated activation and loyalty logics.

To strengthen the interpretability of the results, Table 4 provides a comparative synthesis of the three most representative clusters identified, based on the average values of the RFM indicators and intra-cluster semantic similarity, accompanied by a symbolic characterization derived from the product descriptions.

Table 4. Hybrid model cluster profiles.

Cluster	Recency (Days)	Frequency	Monetary (EUR)	Semantic Similarity	Dominant Semantic Description
A	12	5.8	120	0.69	useful products, recurring purchases
B	94	1.2	330	0.71	decorative items, luxury products
C	6	1.0	52	0.66	books, educational content

The clearly separated cluster in the t-SNE map corresponds to one of the profiles in Table 4 (high semantic coherence and distinct purchase intent), indicating that the hybrid objective amplifies semantically consistent structure beyond what numeric-only patterns

reveal. We stress that t-SNE is used solely for visualization; the quantitative improvements are captured by Silhouette and Davies–Bouldin (Table 2). Simpler methods (K-Means, GMM, Hierarchical) do recover partial separation but yield lower cohesion/separability and reduced semantic consistency on average, confirming that the hybrid formulation adds value beyond these baselines.

To highlight the geographical variations in the structure of behavioral segments, Figure 3 illustrates the percentage distribution of customers in each cluster, separately for each country analyzed. The visualization provides a complementary perspective on how the typologies presented above manifest themselves in distinct cultural contexts.

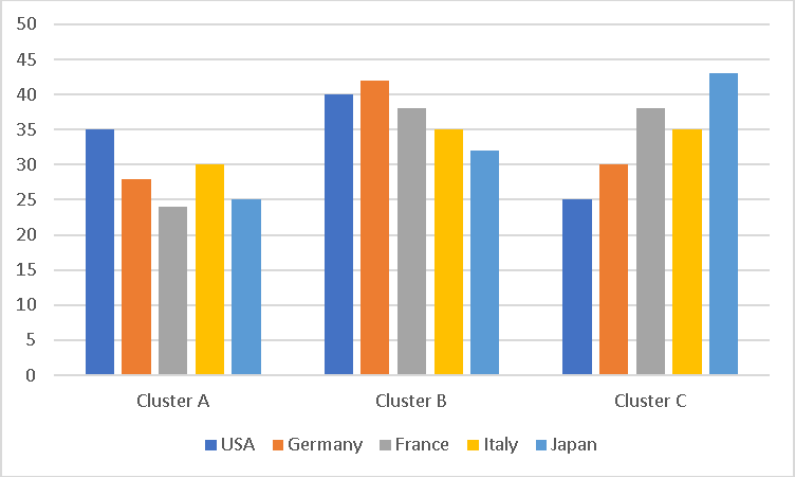


Figure 3. Cluster distribution by country (% of clients per segment).

In order to strengthen the interpretative validity of the proposed hybrid analytical framework, a two-dimensional projection of the segment distribution was generated, using the t-SNE algorithm applied to the integrated numerical–semantic space [67]. Figure 4 highlights a clear separation between clusters, supporting the topological coherence induced by the hybrid objective function and the robustness of the segmentation architecture.

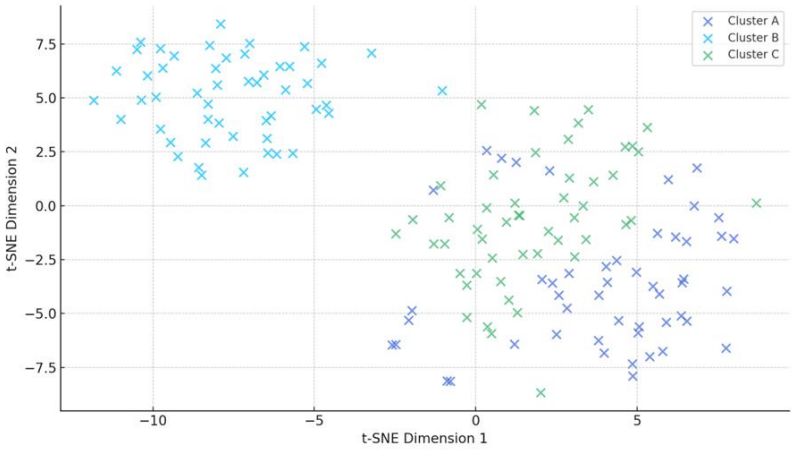


Figure 4. Two-dimensional cluster projection using t-SNE.

One of the segments with high density in the structure is composed of customers who demonstrate a constant purchase frequency, accompanied by a recent shopping history and average values of transactional baskets. The semantic vectors corresponding to this group indicate a preference for functional products with succinct descriptions and clearly defined utility. From a strategic point of view, the configuration suggests a loyal audience

with a well-established operational rhythm, which can be effectively targeted through subscription programs, recurring offers, or loyalty mechanisms adapted to already stabilized consumption cycles.

For the purpose of rigorous comparison with existing reference models in relation to unidimensional approaches, Table 5 summarizes the scores obtained for the three tested variants: exclusively numerical segmentation (KMeans on RFM), purely semantic segmentation (KMeans on embeddings), and the scalar hybrid model. The results show that the proposed model achieves superior performance both from a geometric perspective (Silhouette, Davies–Bouldin) and from a semantic perspective, reflected by a significantly higher conceptual coherence score.

Table 5. Comparative Performance of Segmentation Methods (RFM-only, Semantic-only, Hybrid).

Segmentation Method	Silhouette Score	Davies–Bouldin Index	Semantic Coherence
KMeans on RFM	0.42	0.89	0.12
KMeans on Embeddings	0.36	1.05	0.39
Deep Embedded Clustering (DEC)	0.47	0.82	0.44
Scalarized Hybrid Model	0.58	0.67	0.61

To investigate the sensitivity of the architecture to the intermodal scalarization tuning, Figure 5 shows the evolution of the Silhouette score with respect to variations in the parameter λ. A superior performance is observed around the value $\lambda = 0.5$, indicating a functional balance between the traditional numerical metric and the latent semantic representation.

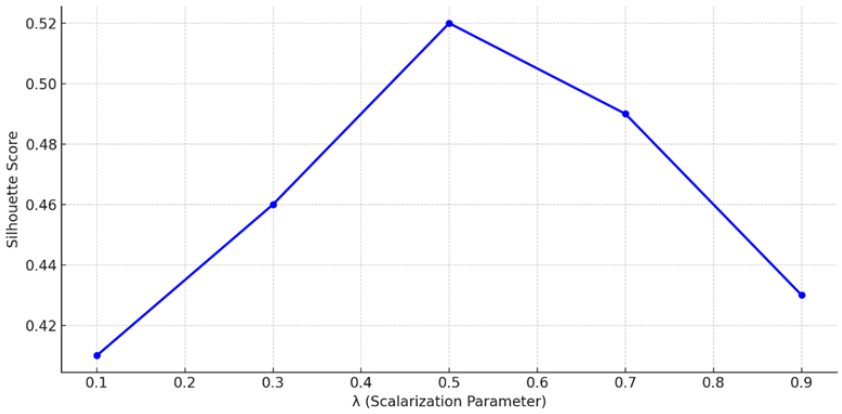


Figure 5. Influence of λ Value on Silhouette Score for the Hybrid Model.

Another structure identifiable by the model is characterized by low frequency but with high monetary values per purchase. The semantic profile of these customers reveals an attraction for products with significant symbolic charge, especially in the sphere of decorative, artistic, or exclusive objects. Transactional behavior seems to follow a seasonal calendar, associated with special events or luxury gifts. From the perspective of commercial applicability, the segment can be successfully activated through emotionally oriented campaigns, personalized narratives, and communication focused on the symbolic value of the purchased objects.

A third relevant segment was outlined around customers with a single recorded transaction, generally recent, with shopping carts oriented towards educational products,

books, or exploratory resources. Semantic analysis indicates a predisposition towards cognitive, informative, or personal discovery content. Although the numerical behavior is limited, the resulting profile suggests a potentially emerging category in an incipient phase of relationship with the commercial platform. An appropriate strategy for this segment could include activation through thematic newsletters, recommendations based on interests, and stimulating recurrence through editorial suggestions adapted to the minimal purchase history.

Through the refinement with which these structures are identified and through their degree of symbolic interpretability, the presented methodological system offers solid premises for building differentiated marketing campaigns, in which each segment becomes a contextualized entity, not just a numerical aggregation. The internal coherence of each cluster, in a double register—behavioral and semantic—facilitates not only classification but also strategic action, opening up the possibility of commercial interventions with a high degree of customization.

The scalarization parameter λ, which regulates the balance between numerical distances (in the RFM space) and semantic similarities (in the embeddings space), proves to be a fine-tuning factor with direct implications on the quality and stability of behavioral segmentation. By adjusting the value of $\lambda \in [0, 1]$, the relative influence of each component within the previously defined objective function is explicitly controlled. Thus, the model becomes sensitive to semantic variations only to the extent that λ reduces the weight of the Euclidean distance and amplifies the contribution of the type 1 metric—cosine similarity.

Experimental analysis has demonstrated that the interval $\lambda \in [0.5, 0.7]$ corresponds to an optimal region in which high values of the Silhouette Score and increased semantic coherence between members of the same clusters are simultaneously obtained. In this area, the model exhibits balanced behavior, capable of capturing both the transactional regularity of purchases and the latent intention expressed through explicit preferences in product descriptions. Although the topology of the semantic space is less rigid than the numerical one, through appropriate weighting, the dominance of one dimension over the other is avoided, thus preventing both semantic collapse (through numerical overrepresentation) and contextual drift (by ignoring the RFM metric).

As λ exceeds the threshold of 0.8, the objective function tends to approach a pure numerical expression, in which semantic distances are almost negligible. In this configuration, the model converges to a clustering equivalent to applying K-Means in the RFM space, thus losing the ability to differentiate behaviors that are apparently numerically similar but symbolically and contextually distinct. The observations obtained in this scenario highlight a drastic reduction in intra-cluster semantic variability and a degradation of the commercial interpretability of the segments.

At the opposite pole, for values of $\lambda < 0.3$, the segmentation becomes unstable, with a pronounced tendency to generate clusters significantly influenced by small lexical variations in the textual embeddings. In this area, numerical cohesion deteriorates, and the Silhouette Score suffers notable decreases. From a decision-making point of view, the resulting structure is less predictive from an economic perspective, since the obtained groupings reflect stylistic similarities of the text rather than robust patterns of transactional behavior.

Therefore, the parameter λ should not be viewed as a simple adjustment coefficient but as an epistemological tool that modulates between two distinct forms of knowledge: a quantitative one, expressible metrically, and a semantic one, emerging from latent spaces of meaning. The model's ability to generate coherent, stable, and actionable segmentations depends essentially on the careful selection of this parameter, which must be contextually calibrated, depending on the specifics of the market, the dominant natural language, and the desired level of granularity in customer interpretation.

The integration of numerical distances and semantic representations into a unified model implies a computational regime significantly different from classical clustering paradigms. In particular, the introduction of a high-dimensional semantic component, such as in the case of transformer-type textual embeddings ($d = 512$), implies an extension of the algorithmic complexity, which requires a careful assessment of the implicit costs and the scalability in real-world contexts of use. In the proposed implementation, the iterative optimization process involves evaluating the hybrid objective function for each individual instance against all centroids, which implies a computational effort directly proportional to the product of the number of clients, the number of clusters, and the size of the vectors. However, the dual architecture of the algorithm—in which the centroid updating is performed both by gradient descent in the numerical space and by symbolic adjustments guided by genetic operators—introduces a convergence stabilization mechanism, which reduces the number of epochs required to reach a consistent local optimum.

By intelligently parallelizing the calculations in the distance and centroid update stages, the implementation managed to maintain an acceptable average execution time, even in conditions of datasets with tens or hundreds of thousands of clients. In this sense, the model proved compatible with distributed infrastructures, and the costs associated with semantic vectorization can be isolated in a pre-process stage, avoiding redundancy in subsequent iterations.

The integration of the semantic dimension in behavioral segmentation opens the prospect of tools capable of reflecting not only consumption patterns but also the cultural specificity of each analyzed context. Segmentation thus becomes a bridge between algorithmic analysis and human interpretation, facilitating the construction of profiles that can guide differentiated decisions and adapt to linguistic and cultural diversity [23]. In this way, the proposed model is not limited to providing statistical results but contributes to the development of explainable and sustainable mechanisms for personalized digital marketing.

5. Conclusions

The conceptual framework proposed in this study offers a methodological synthesis between two complementary dimensions of consumer behavior: the numerical structure of transactions and the symbolic expressiveness of implicit preferences. By formally integrating the two information registers, the constructed model allows for a segmentation that transcends the limitations of strictly quantitative paradigms, bringing to the fore not only the frequency and monetary value of purchases but also the semantic universe in which they are inscribed.

The numerical structure was modeled through the classic RFM (Recency, Frequency, Monetary Value) formalism, recognized for its ability to synthesize the customer–platform relationship in transactional terms. This behavioral vector was, however, extended by superimposing a symbolic component, obtained through semantic vectorization of product descriptions. The embeddings generated with Sentence-BERT, in a multilingual version, allowed for the projection of customers into a latent space of meanings, in which proximity reflects conceptual affinities, not just similarities of numerical behavior.

The objective function built around this model is a scalar one, in which the parameter λ modulates the relationship between Euclidean distances in numerical space and cosine similarities in semantic space. This mechanism offers methodological flexibility and allows for fine-tuning, depending on the specifics of the data set, the semantic intensity of the descriptions, or the desired granularity in the commercial interpretation of the clusters.

By articulating these two dimensions in a single optimization framework, the study contributes to the formulation of a methodology in which the meaning of behavior is

recovered without sacrificing numerical rigor. The validation of the model was carried out both on the formal dimension, by analyzing the objective function and evaluating internal cohesion metrics, and by empirical testing in multicultural contexts, in which the model was adapted differentially, but coherently, to the specifics of each market.

The contribution is a system that maps customers based on their actions and the implicit meaning of their choices. Through this articulation between the quantitative and the symbolic, the premises of a more nuanced, more adaptable, and more relevant commercial segmentation for strategic decisions in the contemporary digital environment are created.

The validation of the model led to a series of consistent findings, which confirm the methodological robustness and semantic versatility in applied contexts. In direct comparisons with classical techniques—such as the K-Means algorithm applied exclusively on RFM vectors or segmentation methods based on isolated semantic similarities—the proposed configuration generated superior performances, both in terms of the internal cohesion metric (Silhouette Score, Davies–Bouldin index) and in terms of conceptual homogeneity within segments. The resulting structure manages to capture both the transactional regularities and the symbolic dimensions of consumer choices.

The systematic application of the model in five distinct national contexts—the United States, Germany, France, Japan, and Italy—allowed the observation of a cross-cultural algorithmic stability. In all these markets, behavioral segmentation reflected not only recurring transactional patterns but also subtle variations in preferences expressed through natural language. Linguistic diversity and cultural specificity did not negatively affect the coherence of the segments, which indicates a high adaptability to local particularities without requiring additional interventions in the basic structure of the model. In the model architecture, the scalarization parameter λ stood out as a fine-tuning element with an adaptive function. The ability to calibrate the relative influence of numerical distances and semantic similarities allowed for a functional balance between geometric precision and contextual relevance.

The application dimension strengthens the value of the contribution by empirically validating the method on a set of real transactional data collected from five distinct cultural contexts. The results obtained offer a new perspective on how linguistic meanings, specific to each socio-economic environment, can be integrated into automated decision-making processes without compromising the metric rigor of the analysis. Beyond the quantitative performance, the proposed formulation allows a narrative reconstruction of the segments so that the resulting groupings are intelligible and strategically usable in a multilingual and dynamic commercial framework.

5.1. Theoretical and Methodological Implications

From a theoretical perspective, the proposed methodological construction introduces a coherent two-dimensional framework for behavioral segmentation, in which transactional numerical data are directly related to symbolic representations extracted from the linguistic space. This algorithmic synthesis between the RFM structure and semantic embeddings generates a reconceptualization of segmentation, not as an exclusively metric procedure, but as a process of reconstructing consumption intention on two axes: quantifiable and interpretable. Thus, segmentation is transformed into a tool for translating behavior into meaning, with clear mathematical foundations and contextual articulation.

At the level of algorithmic construction, the model proposes a transparently formulated objective function, in which numerical distances and symbolic similarities are scaled by an adjustment coefficient, which allows for increased explainability in relation to grouping decisions. The variables involved do not operate in an opaque or isolated manner but

are easy to track and interpret, which enhances the ability of specialists to understand the internal logic of the segmentation process. In an era where the interpretability of machine learning models is becoming a criterion for scientific validation, this feature provides a solid direction for the integration between advanced computational techniques and practical comprehensibility requirements.

By integrating modern methods from natural language processing—particularly embeddings obtained with the help of pre-trained neural networks—alongside established concepts from analytical marketing, the model manages to articulate a space of intersection between disciplines. This transgression of the traditional boundaries between data science and behavioral economics offers not only a performant operational tool but also a new epistemological perspective on how data can be read as expressions of meaning in the digital context.

The hybrid structure developed here is not limited to the e-commerce sector. Its modular configuration makes it suitable for adaptations in domains where human behavior is characterized simultaneously by measurable actions and expressive content. Relevant examples include digital health, where patient interactions combine clinical metrics and symptomatic descriptions; online education, where performance is intertwined with textually expressed motivation; or recommendation systems, where latent preferences can be inferred from both sources. In all these cases, the model provides a formal framework in which behavior becomes intelligible from both statistical and symbolic perspectives.

5.2. Practical Implications and Applications

The integration of the proposed model into the operational infrastructure of e-commerce platforms opens up multiple possibilities for optimizing strategic and tactical decisions. By aligning segmentation not only with transactional behavior but also with latent preferences expressed symbolically, a substantial refinement of the targeting process becomes possible. Communication campaigns can be directed to customers not only based on the numerical history of purchases but also based on the implicit motivations that support consumer choices, an aspect captured through the semantic structure of the analyzed descriptions.

Regarding retention, the model contributes to building a personalized experience in which product recommendations are not based exclusively on transactional similarities but also include semantic dimensions of customer interest. Thus, a contextual activation, more emotionally relevant, becomes possible, which contributes to strengthening the relationship between the consumer and the platform. In this framework, personalization takes on a new depth, being based on meaning, not just on frequency or value.

An additional benefit derives from the model's ability to detect rare structures, which may signal the existence of emerging communities or niche markets. The groupings generated based on symbolic dimensions allow the identification of microsegments that may escape detection in classical systems. These structures, even if numerically marginal, may represent future sources of economic value by addressing specific preferences or latent needs.

In addition, the model provides an efficient mechanism for filtering marketing actions, preventing the allocation of resources to artificially constructed segments based on non-contextualized RFM scores. By using a common semantic space between customers and products, segmentation becomes more faithful to behavioral reality, and the resulting campaigns have greater chances of relevance and impact.

Last but not least, the model's architecture allows easy integration into existing business intelligence systems due to the algorithmic scalability and the high degree of explainability of the objective function. The interpretability of the results supports decision-making

transparency and facilitates cooperation between technical teams and commercial decision-makers. Overall, the proposed solution is compatible with the current requirements of the digital ecosystem and offers viable premises for immediate adoption in dynamic commercial environments.

5.3. Limitations of the Study

Despite the rigorous empirical validation and the high degree of methodological adaptability, the proposed model involves a series of constraints that require attention for a wider application or possible further developments. One of the most sensitive areas is the dependence of the performance on the textual quality of the product descriptions. In situations where the linguistic content is limited, generic, or subject to standardized conventions, the model's ability to extract significant semantic differentiations may be diminished. The symbolic representations thus become less expressive, which reduces the accuracy in the semantic clustering process.

Another limitation comes from the multilingual nature of the embeddings used. Although pre-trained models, such as distiluse-base-multilingual, allow the simultaneous processing of several languages, their performance is not uniform. Especially in the case of non-Latin languages, such as Japanese, the ability to reflect cultural subtleties or specific idiomatic expressions remains reduced. This imbalance affects the semantic coherence of the generated segments, especially when the symbolic space is populated with instances from linguistic registers significantly different from those used in the model training phase.

A relevant technical aspect concerns the computational cost of the system. Even if, in the experimental context, the execution time was kept within reasonable limits, extending the methodology to an industrial level—with frequent update requirements, possibly in a near real-time regime—may raise scalability difficulties. Textual preprocessing, embedding generation, and iterative optimization require considerable resources, and integration into a continuous operational pipeline could impose compromises between performance and speed.

At the same time, the proposed architecture is based on a static formulation of preferences, treating the semantic behavior of the client as an aggregate projection, without introducing a temporal component. Time evolutions of interests, seasonal variations, or style transformations are not captured in the current form of the model. The lack of sequential memory—such as LSTM, GRU, or temporal attention mechanisms—limits the possibility of anticipating the dynamics of preferences, which could be relevant in the context of predictive scenarios or proactive personalization campaigns.

The optimal value of the scalarization parameter λ may vary depending on dataset characteristics, particularly the homogeneity of semantic vectors. While empirically calibrated here, future work may explore automated tuning methods or cross-validation strategies to generalize parameter selection. Furthermore, while the optimal value of the scalarization parameter λ was determined empirically in the experimental context presented, applying the method to other datasets, characterized by distinct semantic or behavioral distributions, may require additional adjustments. The need for calibration can be managed by systematic cross-validation strategies, capable of identifying optimal configurations of balance between the numerical and symbolic dimensions of segmentation.

5.4. Future Research Directions

The expansion of the functionalities associated with the presented methodological system can be achieved through a series of complementary directions, intended to deepen both the analytical dimension and the practical utility in complex commercial contexts. A first line of development concerns the integration of explicit feedback provided by

users—ratings, textual comments, and reactions in the digital social space—in the process of constructing semantic vectors. The incorporation of these expressive forms would allow a more nuanced contextualization of behavior, adding attitudinal and affective valences to the generated profiles.

Another relevant area of research is the use of specialized embeddings on the domain. Pre-trained models, although useful in general scenarios, can be refined by fine-tuning on data specific to industrial verticals, such as cosmetics, digital education, or fashion. This customization would allow for a more faithful capture of semantic nuances and lexical particularities, offering a superior granularity in the grouping process.

On the temporal axis, a direction with significant potential consists of the integration of sequential architectures, which captures the dynamics of preferences over time. Models such as recurrent neural networks or temporal transformers could allow for the anticipation of behavioral transitions and the detection of seasonal patterns or changes in consumption style. Such an extension would transform segmentation into an adaptive process, sensitive to longitudinal variations in customer behavior.

The applicability of the model can also be extended to other types of relevant entities in the commercial ecosystem. Segmentation can be oriented not only to users but also to products, logistical routes, or post-sales strategies. Through this generalization, the methodology becomes an extensible analytical framework, useful in mapping the entire network of commercial interactions.

To support practical adoption in decision-making environments, a logical continuation of the research is the design of an interactive interface intended for strategy and marketing teams. Such a tool would allow not only real-time visualization of segments and adjustment of scaling parameters but also assisted interpretation of the generated profiles. Through such a platform, segmentation becomes not only intelligible but also directly actionable in operational contexts.

Therefore, in the increasingly fragmented landscape of the contemporary digital ecosystem, in which behavior evolves rapidly and often contradicts classical interpretation models, a conceptual reconstruction of the way data is read, correlated, and understood becomes necessary. This paper proposes a model that goes beyond the simple aggregation of transactional values, suggesting that, through careful formulation of the objective function, mathematics can serve as a means of articulating meaning, not just as a tool for quantitative separation.

The proposed methodology brings to the forefront the idea that segmentation should not be reduced to a geometric problem but could be reimagined as an interface between measurable behaviors and latent meanings. By integrating the semantic dimension into the decision space, the model offers not only a formal optimization but also a bridge between what is observable and what is interpretable. Thus, a perspective is outlined in which algorithms do not just automate processes but contribute to a deeper understanding of the logic behind individual choices.

The proposed numerical-semantic construction does not only pursue metric performance but also reconfigures the objective of segmentation: from separation to comprehension, from procedural efficiency to contextual readability. In this sense, the result is not just a better grouping but a mapping of meaning in which the meanings associated with products, texts, and behaviors are brought to the forefront of the analysis.

This vision reflects a paradigm shift in the use of algorithmic tools in digital marketing and in data science in general. It is not just about calculating faster or segmenting more precisely, but about bringing more transparency, interpretability, and, ultimately, understanding to decision-making processes. At a time when artificial intelligence is redefining

the boundaries of analysis, the present methodological proposal invites a rediscovery of meaning through data, not just the mechanical exploitation of regularities.

Author Contributions: Conceptualization, I.O.; methodology, I.O. and A.-M.I.; software, I.O.; validation, I.O. and A.-M.I.; formal analysis, I.O. and A.-M.I.; investigation, I.O. and A.-M.I.; resources, I.O. and A.-M.I.; data curation, I.O. and A.-M.I.; writing—original draft preparation, I.O.; writing—review and editing, I.O.; visualization, I.O. and A.-M.I.; supervision, I.O.; project administration, I.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RFM	Recency, Frequency, Monetary Value
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
DistilUSE	Distilled Universal Sentence Encoder (multilingual model)
GDPR	General Data Protection Regulation
t-SNE	t-distributed Stochastic Neighbor Embedding
VaDE	Variational Deep Embedding
DCEC	Deep Clustering with Convolutional Autoencoders
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
K-Means	K-Means Clustering Algorithm
LSTM	Long Short-Term Memory (neural network architecture)
GRU	Gated Recurrent Unit (neural network architecture)
USA	United States of America
EN	English
DE	German
FR	French
ES	Spanish
RO	Romanian
IT	Italian
JP	Japanese
λ (lambda)	Scalarization parameter (weight between numerical and semantic distances)

References

1. Yuan, F.; Zuo, Z.; Jiang, Y.; Shu, W.; Tian, Z.; Ye, C.; Yang, J.; Mao, Z.; Huang, X.; Gu, S.; et al. AI-Driven Optimization of Blockchain Scalability, Security, and Privacy Protection. *Algorithms* **2025**, *18*, 263. [CrossRef]
2. Valdez, F.; Castillo, O.; Melin, P. Bio-Inspired Algorithms and Its Applications for Optimization in Fuzzy Clustering. *Algorithms* **2021**, *14*, 122. [CrossRef]
3. Tavakoli, M.; Molavi, M.; Masoumi, V.; Mobini, M.; Etemad, S.; Rahmani, R. Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: A case study. In Proceedings of the 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), Xi'an, China, 12–14 October 2018; pp. 119–126. [CrossRef]
4. Chang, Y.C.; Yang, H.; Kong, S. Based on Mini Batch K-Means Clustering for Customer Segmentation in E-commerce. In Proceedings of the 2022 International Conference on Cloud Computing, Big Data and Internet of Things (3CBIT), Wuhan, China, 22–23 October 2022; pp. 60–66. [CrossRef]
5. Hajibaba, H.; Grün, B.; Dolnicar, S. Improving the stability of market segmentation analysis. *Int. J. Contemp. Hosp. Manag.* **2020**, *32*, 1393–1411. [CrossRef]
6. Roberts, J.H.; Kayande, U.; Stremersch, S. From academic research to marketing practice: Exploring the marketing science value chain. *Int. J. Res. Market.* **2019**, *31*, 127–140. [CrossRef]

7. Wilbert, M.R.; Hoppe, P.H.; Sartori, P.H.; Stefenon, S.F.; da Silva, D.A.M. Clustering Internal and External Indices for Customer Segmentation from Retail Data. *Algorithms* **2023**, *16*, 396. [CrossRef]
8. Dolnicar, S.; Grün, B.; Leisch, F. *Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful*; Springer Nature: Singapore, 2018.
9. Ernst, D.; Dolnicar, S. How to avoid random market segmentation solutions. *J. Travel Res.* **2018**, *57*, 69–82. [CrossRef]
10. Mustafa, H.M.J.; Ayob, M.; Albashish, D.; Abu-Taleb, S. Solving text clustering problem using a memetic differential evolution algorithm. *PLoS ONE* **2020**, *15*, e0232816. [CrossRef]
11. Abualigah, L.; Gandomi, A.H.; Elaziz, M.A.; Hussien, A.G.; Khasawneh, A.M.; Alshinwan, M.; Houssein, E.H. Nature-Inspired Optimization Algorithms for Text Document Clustering—A Comprehensive Analysis. *Algorithms* **2020**, *13*, 345. [CrossRef]
12. Kowalski, P.A.; Łukasik, S.; Charytanowicz, M.; Kulczycki, P. Nature Inspired Clustering—Use Cases of Krill Herd Algorithm and Flower Pollination Algorithm. In *Interactions Between Computational Intelligence and Mathematics Part 2*; Kóczy, L.T., Medina-Moreno, J., Ramírez-Poussa, E., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 83–98.
13. Huang, X.; Peng, H.; Zou, D.; Liu, Z.; Li, J.; Liu, K.; Wu, J.; Su, J.-L.; Yu, P.S. CoSENT: Consistent sentence embedding via similarity ranking. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 2800–2813. [CrossRef]
14. Pirnau, M.; Botezatu, M.A.; Priescu, I.; Hosszu, A.; Tabusca, A.; Coculescu, C.; Oncioiu, I. Content Analysis Using Specific Natural Language Processing Methods for Big Data. *Electronics* **2024**, *13*, 584. [CrossRef]
15. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113. [CrossRef]
16. Jiang, S.; Yang, S.; Wang, Y.; Liu, X. Scalarizing Functions in Decomposition-Based Multiobjective Evolutionary Algorithms. *IEEE Trans. Evol. Comput.* **2018**, *22*, 296–313. [CrossRef]
17. Chen, C.; Ibekwe-Sanjuan, F.; Hou, J. The Structure and Dynamics of Co-Citation Clusters: A Multiple-Perspective Co-Citation Analysis. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 1386–1409. [CrossRef]
18. Lamsiyah, S.; Mahdaouy, A.E.; Ouatik, S.E.A.; Espinasse, B. Unsupervised Extractive Multi-Document Summarization Method Based on Transfer Learning from BERT Multi-Task Fine-Tuning. *J. Inf. Sci.* **2023**, *49*, 164–182. [CrossRef]
19. Dolnicar, S.; Leisch, F. Using segment level stability to select target segments in data-driven market segmentation studies. *Mark. Lett.* **2017**, *28*, 423–436. [CrossRef]
20. Mansueto, P.; Ardizzone, E.; Pirrone, R. A memetic differential evolution approach to clustering ensemble. *Pattern Recognit.* **2021**, *114*, 107849. [CrossRef]
21. Wang, X.; Wang, Z.; Sheng, M.; Li, Q.; Sheng, W. An adaptive and opposite K-means operation based memetic algorithm for data clustering. *Neurocomputing* **2021**, *437*, 131–142. [CrossRef]
22. Lund, B.; Ma, J. A review of cluster analysis techniques and their uses in library and information science research: K-means and k-medoids clustering. *Perform. Meas. Metrics* **2021**, *22*, 161–173. [CrossRef]
23. Ziafat, H.; Shakeri, M. Using data mining techniques in customer segmentation. *J. Eng. Res. Appl.* **2014**, *4*, 70–79.
24. Talaat, N.M.; Aljadani, A.; Alharthi, B.; Farsi, M.A.; Badawy, M.; Elhosseini, M. A Mathematical Model for Customer Segmentation Using RFM and 2-Tuple Fuzzy Linguistic Approach. *Mathematics* **2023**, *11*, 3930. [CrossRef]
25. Gustriansyah, R.; Suhandi, N.; Antony, F. Clustering optimization in RFM analysis based on k-means. *Indones. J. Electr. Eng. Comput. Sci.* **2020**, *18*, 470–477. [CrossRef]
26. Peker, S.; Kocyigit, A.; Eren, P.E. LRFMP model for customer segmentation in the grocery retail industry: A case study. *Mark. Intell. Plan.* **2017**, *35*, 544–559. [CrossRef]
27. Witschel, H.F.; Loo, S.; Riesen, K. How to support customer segmentation with useful cluster descriptions. In Proceedings of the Industrial Conference on Data Mining, Hamburg, Germany, 11–24 July 2015; pp. 17–31. [CrossRef]
28. Xiao, X.; Zhang, X. GMS-AFKMC2: A Customer Segmentation Framework Based on RFM and K-Means++. *Electronics* **2024**, *13*, 3523. [CrossRef]
29. Tabal, C.; Sharma, A.; Li, G. Customer Analysis Using Machine Learning-Based RFM and Customer Lifetime Value Prediction. *Sensors* **2023**, *23*, 3180.
30. Maulik, U.; Bandyopadhyay, S. Performance evaluation of some clustering algorithms and validity indices. *IEEE PAMI* **2002**, *24*, 1650–1654. [CrossRef]
31. Todeschini, R.; Ballabio, D.; Termopoli, V.; Consonni, V. Extended multivariate comparison of 68 cluster validity indices. A review. *Chemom. Intell. Lab. Syst.* **2024**, *251*, 105117. [CrossRef]
32. Artetxe, M.; Schwenk, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 597–610. [CrossRef]
33. Abdalgader, K.; Matroud, A.A.; Hossin, K. Experimental study on short-text clustering using transformer-based semantic similarity measure. *PeerJ Comput. Sci.* **2024**, *10*, e2078. [CrossRef]
34. Viswanathan, V.; Gashteovski, K.; Lawrence, C.; Wu, T.; Neubig, G. Large language models enable few-shot clustering. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 321–333. [CrossRef]

35. Jiang, X.; Wang, Z.; Zhou, G. Semantic Simplification for Sentiment Classification. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022.
36. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimed. Tools Appl.* **2023**, *82*, 3713–3744. [CrossRef]
37. Prakash, M.N.; Ohno-Machado, L.; Chapman, W.W. Natural language processing: An introduction. *J. Am. Med. Inf. Assoc.* **2011**, *18*, 544–551.
38. Wang, Y.; Huang, H.; Rudin, C.; Shaposhnik, Y. Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *J. Mach. Learn. Res.* **2021**, *22*, 1–73.
39. Ros, F.; Riad, R.; Guillaume, S. Deep clustering framework review using multicriteria evaluation. *Knowl.-Based Syst.* **2024**, *285*, 111315. [CrossRef]
40. Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* **2023**, *622*, 178–210. [CrossRef]
41. Robert, V.; Vasseur, Y.; Brault, V. Comparing high-dimensional partitions with the Co-clustering Adjusted Rand Index. *J. Classif.* **2021**, *38*, 158–186. [CrossRef]
42. Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. Understanding of internal clustering validation measures. In Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Australia, 13–17 December 2010; pp. 911–916. [CrossRef]
43. Chen, S. Complementary Ensemble Empirical Mode Decomposition and Independent Recurrent Neural Network Model for Predicting Air Quality Index. *Appl. Soft Comput.* **2022**, *131*, 109757. [CrossRef]
44. Deng, J.; Guo, J.; Wang, Y. A Novel K-medoids clustering recommendation algorithm based on probability distribution for collaborative filtering. *Knowl.-Based Syst.* **2019**, *175*, 96–106. [CrossRef]
45. Ahmed, M.; Seraj, R.; Islam, S.M.S. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics* **2020**, *9*, 1295. [CrossRef]
46. Yu, D.; Liu, G.; Guo, M.; Liu, X. An improved K-medoids algorithm based on step increasing and optimizing medoids. *Expert Syst. Appl.* **2018**, *92*, 464–473. [CrossRef]
47. Ushakov, A.V.; Vasilyev, I. Near-optimal large-scale k-medoids clustering. *Inf. Sci.* **2021**, *545*, 344–362. [CrossRef]
48. Chen, Y.; Zou, C.; Chen, J. Label-Aware Graph Representation Learning for Multi-Label Image Classification. *Neurocomputing* **2022**, *492*, 50–61. [CrossRef]
49. Hennig, C. Cluster-wise assessment of cluster stability. *Comput. Stat. Data Anal.* **2007**, *52*, 258–271. [CrossRef]
50. Dorrity, M.W.; Saunders, L.M.; Queitsch, C.; Fields, S.; Trapnell, C. Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nat. Commun.* **2020**, *11*, 1537. [CrossRef]
51. Marín Díaz, G. A Fuzzy-XAI Framework for Customer Segmentation and Risk Detection: Integrating RFM, 2-Tuple Modeling, and Strategic Scoring. *Mathematics* **2025**, *13*, 2141. [CrossRef]
52. Shi, Y. Advances in Big Data Analytics. In *Theory, Algorithms and Practices*; Springer: Berlin/Heidelberg, Germany, 2022.
53. Zhou, P.; Sun, B.; Liu, X.; Du, L.; Li, X. Active Clustering Ensemble with Self-Paced Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 12186–12200. [CrossRef]
54. Qian, Y.; Yao, S.; Wu, T.; Huang, Y.; Zeng, L. Improved Selective Deep-Learning-Based Clustering Ensemble. *Appl. Sci.* **2024**, *14*, 719. [CrossRef]
55. Berke, A.; Calacci, D.; Mahari, R.; Yabe, T.; Larson, K.; Pentland, A. Open e-commerce 1.0, Five Years of Crowdsourced U.S. Amazon Purchase Histories with User Demographics. *Sci. Data* **2024**, *11*, 491. [CrossRef] [PubMed]
56. Berke, A.; Calacci, D.; Mahari, R.; Yabe, T.; Larson, K.; Pentland, A. *Open E-Commerce 1.0 (Dataset)*; V1; Harvard Dataverse: Cambridge, MA, USA, 2024. [CrossRef]
57. Roszko-Wójtowicz, E.; Deep Sharma, G.; Dańska-Borsiak, B.; Grzelak, M.M. Innovation-Driven E-Commerce Growth in the EU: An Empirical Study of the Propensity for Online Purchases and Sustainable Consumption. *Sustainability* **2024**, *16*, 1563. [CrossRef]
58. Bălăcescu, A.; Pătrașcu, A.; Tănăsescu, A. Analysis of Disparities in Internet Purchases by Individuals at the EU State Level. *Electronics* **2023**, *12*, 982. [CrossRef]
59. Oro, E.; Granata, F.M.; Ruffolo, M. A Comprehensive Evaluation of Embedding Models and LLMs for IR and QA Across English and Italian. *Big Data Cogn. Comput.* **2025**, *9*, 141. [CrossRef]
60. Přibáň, P.; Šmíd, J.; Steinberger, J.; Mištera, A. A comparative study of cross-lingual sentiment analysis. *Expert. Syst. Appl.* **2024**, *247*, 123247. [CrossRef]
61. Pikuliak, M.; Grivalský, Š.; Konôpka, M.; Blšták, M.; Tamajka, M.; Bachratý, V.; Simko, M.; Balážik, P.; Trnka, M.; Uhlárik, F. SlovakBERT: Slovak Masked Language Model. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; pp. 7156–7168.
62. Kocoń, J.; Zaśko-Zielińska, M.; Miłkowski, P. Multi-level analysis and recognition of the text sentiment on the example of consumer opinions. In Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2019), Varna, Bulgaria, 2–4 September 2019.

63. Miettinen, K. *Nonlinear Multiobjective Optimization*; Kluwer Academic Publishers: Boston, MA, USA, 1999. [CrossRef]
64. Deb, K. *Multi-Objective Optimization Using Evolutionary Algorithms*; John Wiley & Sons: Chichester, UK, 2001.
65. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
66. Chandra, R.; Omlin, C.W. Hybrid Evolutionary One-Step Gradient Descent for Training Recurrent Neural Networks. In Proceedings of the International Conference on Genetic and Evolutionary Methods (GEM), Las Vegas, NV, USA, 14–17 July 2008; pp. 305–311. Available online: https://www.researchgate.net/publication/220862330_Hybrid_Evolutionary_One-Step_Gradient_Descent_for_Training_Recurrent_Neural_Networks (accessed on 20 June 2025).
67. Kobak, D.; Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **2019**, *10*, 5416. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.