ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Optimizing the Design of a Multilingual Customer Service System for Business English Using Clustering Algorithms

Yang Wang[1,*]

[1] College of Humanities and Foreign Languages, Zhejiang Shuren University, Hangzhou 310015, Zhejiang, China

SUMMARY: *Addressing the challenge of precise customer value segmentation in multilingual customer service systems for business English, this paper primarily analyzes the application of an enhanced RFM model and optimized clustering algorithms within such systems. The study first constructs an RSA customer value assessment model, using the Recency (R), Stability (S), and Average Spending (A) as core indicators. The CRITIC weighting method is employed to scientifically determine the weight of each dimension, overcoming the issues of indicator collinearity and uniform weighting in traditional RFM models. Building upon this foundation, an improved K-means algorithm based on K-nearest neighbors and density peaks was designed. By optimizing the selection of initial cluster centers, the algorithm's convergence speed and classification accuracy were enhanced. Empirical validation using 366 active customers from a company identified four distinct customer segments: 43 high-value loyal customers, 127 high-potential new customers, 142 latent-value customers, and 54 low-value general customers. Analysis of variance revealed highly significant differences across all clusters in R, S, and A metrics, with F-values of 80.014, 92.816, and 117.607 respectively ($p<0.001$), confirming the statistical validity of the segmentation results. System performance testing demonstrated excellent efficiency and stability, with a response time of only 34.27 seconds even when processing 10,000 records.*

KEYWORDS: *RFM model; K-means algorithm; customer value assessment; merchant English customer service*

1 Introduction

With the acceleration of globalization and the expansion of cross-border e-commerce, the number of multinational projects has increased, driving corporate internationalization. Enterprises now engage with client groups from diverse countries and regions who speak different languages [1, 2]. In multinational projects, language often poses a critical communication barrier. Relevant language services aim to overcome this obstacle, ensuring accurate and seamless communication among all project stakeholders [3]. Whether in business negotiations, technical exchanges, or daily collaboration, these services deliver high-quality linguistic support, facilitating smooth project progression and achieving objectives such as timely delivery of products or services, successful partnership agreements, and enhanced overall project efficiency and effectiveness [4-6]. Consequently, multilingual customer service systems have emerged as vital tools for enterprises to elevate customer service efficiency and quality [7]. Through iterative advancements in machine translation, traditional human-based

*fortune868686@163.com
https://doi.org/10.65102/is20261037 Published: 30 April, 2026

customer service challenges—such as response delays and high operational costs—have been addressed. Multilingual customer service systems can automatically recognize and interpret inquiries in various languages, delivering accurate and timely responses in the corresponding language [8-10]. Market research indicates the global smart multilingual customer service market reached $20 billion in 2024 and is projected to grow to $40 billion by 2029, with a compound annual growth rate exceeding 15%. Current mainstream multilingual customer service systems can meet the international business needs of most enterprises.

Reference [11] integrated machine translation to design a 24/7 chatbot for e-commerce stores, ensuring translation quality while reducing costs, though it only covered four languages including English. Reference [12] developed a customizable and deployable multilingual chatbot service system for SMEs, supporting real-time translation, comprehension, and responses in English, Malay, and Chinese. Reference [13] addresses politeness issues in multilingual customer service by building an empathetic conversational bot using deep learning frameworks. This bot generates responses in different languages based on customer preferences. However, the method only analyzes interactions between two languages, lacking broad adaptability. Reference [14] developed a dialogue system integrating an English-taught Mike system with a Google Translate chatbot, covering 103 languages. However, testing revealed disparities in response fluency and comprehension between European and Asian languages. Reference [15] introduced a Transformer-based multilingual customer service system that enhances customer comprehension and responsiveness. This neural network-based system enables real-time translation alongside sentiment analysis, achieving personalized, efficient, and highly inclusive interactions. Reference [16] employs neural machine translation, multimodal translation, and domain-adaptive translation within a deep learning architecture to optimize the accuracy, response speed, adaptability, and coverage of language translation systems on cross-border e-commerce platforms, enabling real-time multilingual translation and multimodal conversion. Reference [17] reports a multilingual chatbot for customer service, international trade communication, and online services. By identifying user languages and translating for interaction, it displays both the user's original language and the translated response, enabling multilingual, multi-scenario, and multifunctional communication. Reference [18] shares a multilingual intelligent service system enhanced with a dynamic domain adaptation mechanism. By visualizing cross-language logistics through knowledge graphs, it enables real-time translation, risk alerts, and route optimization for cross-border logistics services across 12 languages.

Reference [19] leverages the natural language understanding capabilities of large language models to enhance multilingual dialogue processing for intelligent agents in business activities. By integrating named entity recognition to capture transaction details, it optimizes multilingual dialogue tasks in commerce. Reference [20] employs large language models to extract intents and keywords from multilingual environments in international business, introducing a reasoning chain-based approach to generate interpretable clustering results, thereby improving intent recognition performance in multilingual dialogue systems. Reference [21] embeds a natural language processing (NLP)-enabled multilingual chatbot within intelligent enterprise management software, allowing customers to communicate in multiple languages. With the assistance of a chatbot agent, customers can interact in their native language. Meanwhile, Reference [22] demonstrates how NLP technology can effectively simulate human language processing capabilities, enabling service systems to understand and generate natural language. This advances the interactive skills and service efficiency of intelligent customer service systems. However, NLP-based machine translation still struggles to support mixed communication scenarios involving three or more languages, often leading to service path disruptions and semantic fragmentation, which negatively impacts the overall performance and

user experience of customer service systems [23, 24].

Clustering algorithms group data such that instances within the same cluster exhibit high similarity, while instances across different clusters show significant divergence [25]. In natural language processing, clustering algorithms are widely applied to tasks such as text classification, text clustering, and text summarization [26, 27]. Fundamental clustering algorithms including K-means, DBSCAN, and hierarchical clustering can effectively address multilingual clustering and semantic clustering within business processes, classifying language-specific intents and sentiments. Reference [28] applied unsupervised clustering algorithms to analyze semantic similarities among business process activities like bicycle rentals and cargo tracking, generating specific microservices to optimize online services. Reference [29] employed a clustering ensemble algorithm combining different clustering methods to efficiently identify emoji and keyword semantics across diverse linguistic contexts in social media, thereby extracting sentiment. Reference [30] developed a deep learning-based clustering algorithm to analyze customer reviews on e-commerce platforms, capturing and classifying customer sentiment to uncover implicit emotional experiences. Reference [31] combined K-Means with heuristic-based capsule networks to establish a user intent detection model, addressing shortcomings in natural language processing technologies for robots and intelligent agents.

This paper aims to optimize the design of multilingual customer service systems by leveraging advanced clustering algorithms. Its core approach involves using data mining techniques to automatically identify customer groups with similar characteristics and needs from multidimensional data—including interaction histories, service requests, and consumption behaviors. This enables the formulation of differentiated service strategies for distinct groups, thereby enhancing overall service efficiency and customer satisfaction. The paper first outlines the data mining workflow within Customer Relationship Management (CRM), covering goal definition, data preprocessing, and data transformation. This prepares high-quality, model-ready datasets—the foundation for subsequent analysis. It then delves into the fundamentals of cluster analysis, exploring its applicability and limitations in customer segmentation, and demonstrates its unique value as an unsupervised learning algorithm for automatically identifying customer cohorts. To further enhance segmentation accuracy, an improved R (Recency), S (Spacing), A (Amount) model is proposed to address limitations of traditional RFM models. The CRITIC weighting method is introduced to assign weights to each metric, enabling more scientific valuation of multilingual customers. Finally, an enhanced K-means algorithm based on K-nearest neighbors and density peaks is designed. By optimizing the selection of initial cluster centers, it overcomes the traditional algorithm's sensitivity to initial values and susceptibility to local optima, thereby improving the precision and stability of customer value segmentation results.

2 Theoretical Foundations of Data Mining and Cluster Analysis in Business English CRM

2.1 Application Process of Data Mining in Customer Relationship Management

In the design and optimization of multilingual customer service systems, data mining technology plays a crucial role. By analyzing and processing vast amounts of customer interaction data, enterprises can more accurately identify customer needs, optimize service processes, and enhance the efficiency and quality of cross-language services. This section systematically introduces the application process of data mining in Customer Relationship

Management (CRM), laying the theoretical foundation for the subsequent specific application of clustering algorithms in multilingual customer service systems for business English.

2.1.1 Define Analysis and Forecasting Objectives

Applying data mining techniques in customer relationship management requires first defining the business objectives the enterprise aims to achieve. Data collection and preprocessing should then be conducted based on these objectives. This process forms an initial understanding of the data and reveals its distribution patterns, enabling the subsequent development of predictive models. Only then can data mining technology effectively serve within the customer relationship management system, helping businesses address real-world challenges.

Organizations must establish structured data collection processes and coordinate relevant departments to select data based on actual requirements. This ensures extraction of valuable insights from big data, enabling the selection of project-specific data to build tailored datasets. Within vast datasets, we must analyze and identify specific feature information suitable for data analysis. This data should then be stored in appropriate databases.

2.1.2 Data Preprocessing

After establishing the dataset, the next step is data preprocessing to make this data usable. Data preprocessing enhances accuracy, completeness, and consistency. Its primary purpose is to prepare for further data mining by filtering useful data from existing sources, ensuring its authenticity and validity.

Data mining requires extensive preliminary preparation, with over half the time and effort dedicated to data preprocessing. This lays the groundwork for subsequent data mining tasks. Data preprocessing primarily involves the following four aspects:

(1) Data Cleaning

Incomplete or erroneous information within the database must be removed. Only accurate and consistent data is retained and loaded into the data warehouse.

(2) Data Integration

Centralize data from diverse sources, formats, and characteristics, organizing it systematically to facilitate enterprise-wide sharing.

(3) Data Reduction

Given the often massive scale of data samples, mining the entire dataset would be prohibitively labor-intensive. To reduce workload, we can analyze smaller sample sets. Data reduction techniques enable this by:

Select samples that closely approximate the fundamental characteristics of the original data, then perform data mining on the reduced dataset. This approach minimally impacts mining outcomes. Three primary methods exist: dimensionality reduction, quantitative agreement, and data compression. Quantitative agreement replaces original data with smaller data substitutes. Dimensionality reduction employs wavelet transforms and principal component analysis. Data compression, used for multimedia like images and videos, is categorized as lossy or lossless.

2.1.3 Data Conversion

This process primarily involves transforming data formats to make them suitable for data mining. Data is converted into features that can accurately describe the data. To achieve optimal learning algorithms, real-valued data can be transformed through smoothing, clustering, normalization, and data generalization. Discretization and conceptual hierarchy transformation are also crucial for data conversion.

For information within databases, selecting appropriate analytical tools and applying diverse methods yields valuable insights. Classic data mining algorithms are employed to analyze and process the data.

2.2 Application of Cluster Analysis in Multilingual Customer Segmentation

After outlining the application process of data mining in customer relationship management, this article further focuses on one of its core technologies—clustering analysis. This section will elaborate on the fundamental principles of clustering analysis and its specific applications in customer classification, providing methodological support for subsequent integration with multilingual customer service scenarios.

2.2.1 Principles of Cluster Analysis

Cluster analysis is an unsupervised learning method that measures similarities among data points. It groups highly similar data into the same cluster, ensuring that data within a cluster exhibit strong correlations while data across clusters show weaker correlations. This enables a deeper understanding of the characteristics within each data category. First, we will provide a detailed introduction to cluster analysis.

Cluster analysis refers to the classification of unlabeled data objects using a similarity measurement method. Data points with high similarity are grouped into the same cluster, while those with low similarity are assigned to different clusters. This process uncovers latent features hidden within the data center, yielding more valuable insights. Definition of cluster analysis: Given a dataset X containing m data objects, each data object $x_i(i=1,2,3\cdots m)$ possesses n feature attributes, i.e., $x=\{x_{i1},x_{i2},x_{i3}\cdots\cdots x_{in}\}$. Cluster analysis involves partitioning the entire dataset X into K clusters $(C_1,C_2,C_3,\cdots\cdots C_k)$ based on some similarity measure among data objects x_i such that the following conditions are satisfied:

$$\begin{cases} X=C_1\cup C_2\cdots C_\kappa \\ C_i\cap C_j=\varnothing(i\neq j) \end{cases} \tag{1}$$

The above definition of cluster analysis is merely one widely accepted approach. For fuzzy clustering methods, data objects are classified based on their membership degrees, and a single data object may be assigned to different clusters, thus not satisfying the aforementioned conditions.

Cluster analysis primarily involves the following processes: (1) Data preprocessing: Selecting representative feature values to reduce data redundancy and representing data objects using appropriate data types. (2) Measuring similarity between data: Calculating similarity between data objects using a specific similarity metric. This measurement can be completed entirely before clustering or computed during clustering as needed. (3) Clustering or grouping data objects: Using a partitioning method and similarity measures to assign highly similar data to the same cluster and less similar data to different clusters. (4) Presentation of clustering results: Visualizing the clustering outcomes graphically or directly outputting the feature information of each cluster.

2.2.2 Advantages and Disadvantages of Cluster Analysis in Customer Classification

The principles and practical applications of cluster analysis reveal its distinct advantages, primarily manifested in the following three aspects.

(1) Since its inception, cluster analysis has matured significantly through extensive scholarly research. Widely adopted across various commercial sectors, its effectiveness as a customer segmentation method is substantiated by abundant practical outcomes.

(2) At its core, cluster analysis is a data modeling technique. It not only effectively categorizes customers but also processes and mines data. The diversity of cluster analysis techniques enables a broad range of applications.

(3) If the clustering results generated by this technology in business contexts can be readily interpreted through a business lens, it becomes particularly crucial for data-driven operations, thereby fostering further business development.

Despite these advantages, cluster analysis also exhibits shortcomings in customer classification, specifically in the following aspects.

(1) When performing cluster analysis on customer data, the number of clusters must be determined based on experience. If relevant practical experience is lacking in the business context, repeated testing of the data is required to accurately determine the number of clusters, which consumes significantly more time.

(2) Rapid clustering algorithms all utilize the mean to classify data, making customer data sensitive to outliers and noise points. This results in clusters that are difficult to converge and exhibit low accuracy.

2.3 Customer Segmentation Algorithm Based on an Enhanced RFM Model

Although cluster analysis excels in customer segmentation, it still has certain limitations, especially when dealing with complex multilingual customer behaviors. To address this, this section introduces an enhanced RFM model. By adjusting the metric structure and weight settings, it better adapts to customer value assessment needs in multilingual environments.

The RFM model serves as a crucial theoretical framework for corporate customer relationship management. The traditional RFM model evaluates customers primarily based on three metrics: the interval since their last purchase, purchase frequency over a specific period, and total purchase amount. By analyzing these metrics, customer groups are classified to facilitate targeted marketing strategies. The simplicity of the RFM model's principles and the ease of obtaining required data have led to its widespread adoption across industries for customer classification. However, the traditional RFM model also exhibits significant limitations.

First, the two indicators of purchase frequency and purchase amount within the traditional RFM model exhibit multicollinearity. The RFM model posits that higher purchase frequency signifies greater customer loyalty, while higher purchase amount reflects greater customer value. In practice, however, purchase frequency and purchase amount are often correlated—customers with higher purchase frequency typically spend more. This leads to redundancy in customer value assessments.

Second, traditional RFM models typically analyze customer value using data from only a specific time period. This results in a biased analysis that fails to accurately assess the value of customers with consistent purchasing patterns but longer purchase cycles. These customers often exhibit high loyalty and are precisely the ones businesses should invest effort in retaining. Therefore, adjustments to the traditional RFM model's metrics are necessary.

Furthermore, the traditional RFM model assumes equal importance for the three dimensions: time since last purchase, purchase frequency, and purchase amount. However, in real-world applications across different industries, the significance of these three indicators for customer value segmentation varies. For example, in industries like hotels and airlines—where customer purchase frequency is inherently low—the single transaction amount is more critical for customer value classification. Therefore, assigning distinct weighting coefficients based on the relative importance of each RFM metric is essential to achieve precise customer value segmentation. Furthermore, the traditional RFM model produces overly complex segmentation results, making it difficult for companies to formulate strategies based on such intricate outcomes. These limitations necessitate improvements.

Addressing the constraints of the traditional RFM model, this paper refines the original three metrics to construct an enhanced RFM model termed the RSA model. This model describes customer behavior characteristics through three metrics: time since last purchase (R), purchase cycle (S), and average transaction value (A). The improved model's metrics are represented by variables R, S, and A: R denotes the time since last purchase, indicating the interval from the most recent transaction to the present. A lower R value signifies a higher likelihood of repeat purchases and greater customer value. S denotes the purchase frequency, representing the average number of purchases per transaction from the customer's first purchase to the present. A smaller S value indicates higher customer loyalty. A represents the average transaction amount, calculated as the mean value of each transaction since the customer's first purchase. A larger A value signifies stronger purchasing power. The calculation formulas for the three indicators of the improved RFM model are shown in (2).

$$\begin{aligned} R &= t_{now} - t_{last} \\ S &= \frac{t_{now} - t_{first}}{total} \\ A &= \frac{price}{total} \end{aligned} \tag{2}$$

Here, t_{now} represents the current time, t_{last} denotes the most recent transaction time, t_{first} indicates the first transaction time, total signifies the number of transactions within the specified timeframe, and price reflects the total transaction amount within the specified timeframe.

To accurately reflect the differentiated characteristics among customers and fully account for the varying degrees of importance of different indicators in customer value assessment, this paper employs the CRITIC weighting analysis method to assign weights to each indicator. Finally, a clustering algorithm is utilized to segment customer value, enabling more precise customer classification results.

2.4 Design Approach for Improving the K-means Clustering Algorithm

To further enhance clustering performance, this section proposes an improved K-means algorithm based on K-nearest neighbors and density peaks. This approach optimizes the selection of initial cluster centers, thereby accelerating convergence speed and improving classification accuracy in multilingual customer data. This provides technical assurance for the system's practical deployment.

This algorithm integrates density peak detection while incorporating K-nearest neighbor information when calculating local density. It also applies weighted Euclidean distance calculations to optimize the selection of initial cluster centers. The core approach involves calculating the local density within the K-nearest neighbor range of sample i. Samples with

higher local density are selected for inclusion in the candidate set for initial cluster centers. The final initial cluster centers are then chosen from this candidate set. To ensure sufficient spatial separation between each cluster center, the selection is completed using a maximum-minimum distance algorithm.

2.4.1 Determining the Initial Cluster Center Candidate Set

Determining the initial cluster centers first requires calculating the distances between sample points. The improved K-means algorithm incorporates attribute weight factors into Euclidean distance calculations, employing weighted Euclidean distances to compute sample distances. This enhances the degree of differentiation between attributes, as shown in formula (3).

$$d_w(x_i, x_j) = \sqrt{\sum_{h=1}^{m} w_{ih}(x_{ih} - w_{jh})^2} \tag{3}$$

Here, w_{ih} denotes the weight of sample i in dimension h, while x_{ih} and x_{jh} represent the values of sample points x_i and x_j in dimension h.

The density of samples is measured by the distance between them to calculate the local density of samples. To eliminate the influence of the cutoff distance setting, the local density of samples is computed using a density peak algorithm based on K-nearest neighbors. The formula for calculating the local density of samples in this algorithm is shown in (4).

$$\rho_i = \sum_{j \in KNN(i)} \exp(-d_{ij}) \tag{4}$$

Here, $KNN(i)$ denotes the set of k nearest neighbor samples for sample i, and d_{ij} represents the distance between sample i and sample j.

As shown in Equation (4), the local density ρ_i increases as the Euclidean distance between sample point i and its K nearest neighbors decreases. Compared to traditional density peak algorithms, the calculation scope for sample local density is reduced from the entire dataset to the K nearest neighbors, yielding more precise local density results.

The local density obtained via the K-nearest neighbor-based density peak algorithm does not guarantee that the sample point with the highest local density is necessarily a cluster center. Therefore, the average local density of a sample point must be calculated. Sample points with local densities exceeding this average are marked as high-density points and added to the initial cluster center candidate set. The average local density refers to the mean of the local densities of all points within the K-nearest neighbor range of any given sample point, as expressed by Equation (5).

$$avg\,\rho_i = \frac{\sum_{j \in KNN(i)} \rho_i}{k} \tag{5}$$

2.4.2 Determining Initial Cluster Centers

Since high-density points may belong to the same cluster, this paper employs the maximum-minimum distance algorithm to select distant sample points from the initial cluster center candidate set as initial cluster centers. The Density Bee-Hive algorithm constructs a decision graph for the sample set based on the local density ρ_i and relative distance δ_i of each sample

point i. The relative distance δ_i is calculated using Equations (6) and (7).

$$\delta_i = \min_{j:p_j>p_i}(d_{ij}) \tag{6}$$

If sample i has the maximum local density, then

$$\delta_i = \max_{i\neq j}(\delta_j) \tag{7}$$

Points in the decision graph where both ρ_i and δ_i are relatively high are closest to the true cluster centers. Therefore, we compute $\rho_i \times \delta_i$ as shown in formula (8).

$$\beta_i = \rho_i \times \delta_i \tag{8}$$

From the candidate set, select the sample point with the largest β_i as the first cluster center. Then, using the maximum-minimum distance algorithm, select the remaining $k-1$ initial cluster centers. This approach prevents initial cluster centers from being assigned to nearby points, ensuring the selection results closely approximate the true values.

3 Empirical Study on Customer Value Based on the RSA Model and an Improved K-Means Algorithm

After completing the theoretical framework for improving the RSA model and optimizing the clustering algorithm, this chapter will focus on the empirical analysis section. It aims to validate the practical effectiveness of the proposed methods in multilingual customer service systems using real enterprise data. Based on a company's customer data, this study employs the RSA model and the improved K-means algorithm introduced earlier to achieve a scientific segmentation of customer value. This provides data support and decision-making basis for formulating multilingual customer service strategies.

3.1 Enterprise Customer Value Evaluation Based on the RSA Model

3.1.1 Data Sources

This section focuses on sample data from representative clients of a certain enterprise. The company's customer database contained 534 clients. Applying predefined screening criteria, 168 inactive clients were excluded, ultimately retaining 366 relatively active clients as the analytical sample. This provides a robust data foundation for computing the enhanced RFM model—the RSA model proposed herein. These sample data originate from extensive customer information and detailed transaction records accumulated during the company's long-term operations, ensuring high accuracy and comprehensiveness.

Regarding data source determination, this study meticulously selected three quantitative indicators: recent purchase interval, purchase cycle, and average transaction amount. These metrics were precisely extracted through direct access to the company's internal databases and financial systems. The selection of quantitative indicators was based on their significance and reliability in reflecting customer characteristics, transaction behavior, and market performance, ensuring the accuracy and validity of the data analysis.

3.1.2 Calculation of RSA Indicator Values

Building upon the RSA model established in the preceding section, this chapter will perform weighted calculations to generate a composite score across three dimensions, thereby supporting subsequent cluster analysis and customer management strategies. After filtering and normalizing the customer sample data, 366 relatively active customers from Company A were selected as the analysis sample.

Using the CRITIC weighting method, the weights for the three dimensions of the RSA model are calculated as [WR, WS, WA] = [0.273, 0.328, 0.399]. Subsequently, based on Equation (2), the data for the three dimensions—the interval since the last purchase (R), purchase cycle (S), and average transaction amount (A) using the RSA model. Weighted calculations yielded each customer's comprehensive evaluation score under the RSA model. Table 1 displays the normalized RSA scores for selected customers.

Table 1: The specific RSA scores of some customers after normalization

	R	S	A
C1	0.401	0.365	0.288
C2	0.218	0.592	0.428
C3	0.978	0.395	0.684
C4	0.135	0.768	0.215
C5	0.735	0.074	0.486
C6	0.825	0.035	0.900
C7	0.224	0.735	0.044
C8	0.011	0.512	0.247
C9	0.439	0.071	0.186
C10	0.851	0.432	0.126
……	……	……	……
C366	0.119	0.157	0.966

3.2 Analysis of Corporate Customer Value Based on K-Means Clustering

Building upon the comprehensive customer value scoring under the RSA model, this section introduces an enhanced K-means clustering algorithm to segment customers and further identify groups with similar behavioral characteristics. By using weighted scores from the R, S, and A dimensions as clustering inputs, combined with the initial center optimization strategy described earlier, we achieve automatic and precise segmentation of customer groups. This provides a classification basis for differentiated service strategies within multilingual customer service systems.

3.2.1 Determining the K Value

This paper employs the elbow method to determine the K value, calculating the sum of squared errors (SSE) corresponding to all possible K values. This metric measures the error of all samples after clustering. Specifically, SSE represents the sum of the squares of distances from all samples to their respective cluster centers (centroids). A graph illustrating the relationship between SSE and K values is plotted. By observing the line chart of SSE versus K, an “elbow point” is identified—the point where the rate of decrease in SSE as K increases undergoes a significant change. This point is typically regarded as the optimal K value, as it signifies that the marginal benefit of adding more cluster centers (i.e., larger K values) toward reducing clustering error begins to diminish significantly. The SSE values corresponding to different K

values in this experiment are shown in Figure 1.

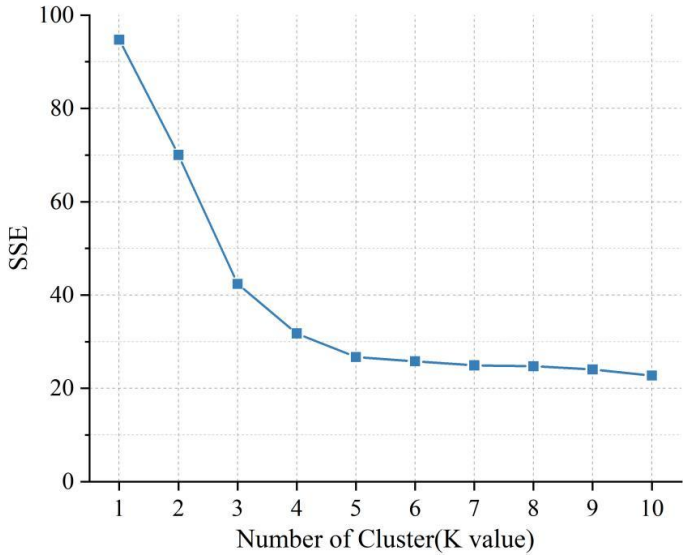

Figure 1: SSE values corresponding to different K values

It can be observed that when K=4, the curve exhibits an inflection point, and the decline in SSE begins to flatten. This indicates that increasing the number of clusters beyond this point yields diminishing marginal benefits in reducing SSE. Therefore, this experiment selects 4 clusters for the cluster analysis. Choosing a smaller K value while achieving a low SSE helps maintain the model's simplicity and interpretability. Simultaneously, it further reduces the average distance between data points and cluster centers, thereby enhancing the clustering effectiveness. This enables enterprises to better understand the needs and characteristics of different customer groups, facilitating the development of more personalized marketing strategies and service plans.

3.2.2 Initial Cluster Center Selection

When selecting initial cluster centers, care should be taken to avoid choosing outliers or abnormal values as cluster centers, as these points may adversely affect the clustering results. This study utilizes SPSS software to automatically select initial cluster centers. After specifying the number of clusters K=4, SPSS automatically selects four representative samples as initial cluster centers based on the actual characteristics of the sample data. This method is simple and efficient, and for the dataset in this study, it yields reasonably appropriate initial cluster centers. The initial cluster centers are shown in Table 2.

Table 2: Initial clustering center

	R	S	A
Cluster1	0.190	0.134	0.773
Cluster2	0.076	0.809	0.696
Cluster3	0.869	0.123	0.799
Cluster4	0.962	0.575	0.187

3.2.3 Selection of Final Cluster Centers

After confirming the K value, the K-means clustering algorithm can be performed using SPSS software. Iterations are conducted to adjust the cluster centers until the algorithm reaches a

stable state. The final cluster centers and clustering results are shown in Table 3 and Figure 2, respectively.

Table 3: Final clustering center

	R	S	A	Number of case
Cluster1	0.112	0.183	0.874	43
Cluster2	0.089	0.735	0.642	127
Cluster3	0.856	0.224	0.758	142
Cluster4	0.943	0.681	0.205	54

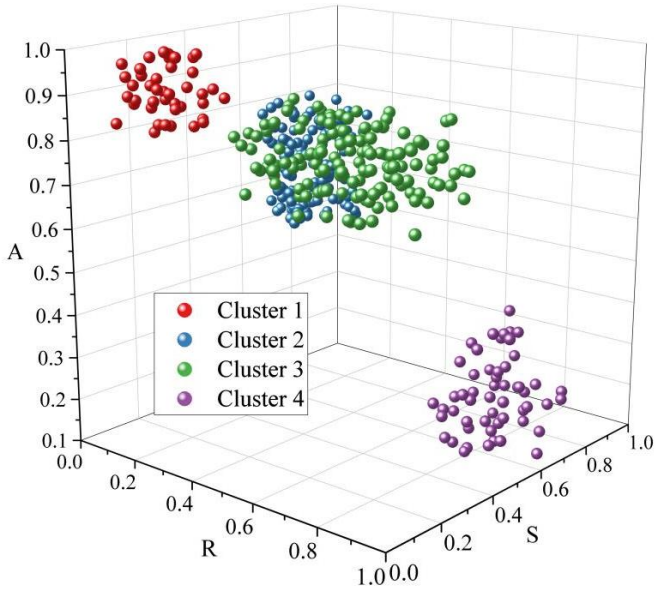

Figure 2: The final clustering results

The final cluster center values identified four distinct customer segments. Cluster 1 comprises 43 customers with an exceptionally low R value of 0.112, indicating recent active spending. Their S value of 0.183 reflects high purchase frequency and strong loyalty, while their highest A value of 0.874 signifies large average transaction amounts and significant overall contribution. These customers represent the company's most valuable assets, classified as high-value loyal clients.

Cluster 2 comprises 127 customers. The R-value center of 0.089 indicates recent consumption with extremely high activity; the S-value of 0.735 suggests unstable consumption frequency; and the A-value of 0.642 indicates moderate spending power. These customers may be new or intermittent spenders with high conversion potential. They are characterized as recently active but with unstable spending frequency, representing high-potential new customers.

Cluster 3 comprises 142 customers. The high R value of 0.856 indicates long intervals between recent purchases and low activity. The low S value of 0.224 suggests historically high loyalty. An A value of 0.758 indicates strong purchasing power. This suggests that although these customers are currently inactive, they possess high historical loyalty and spending capacity, categorizing them as dormant value customers.

Cluster 4 comprises 54 customers. Both R (0.943) and S (0.681) values are high, indicating low consumption activity and frequency; An A value of 0.205 indicates limited purchasing power. These customers have overall low value, likely representing occasional buyers or users of low-priced products. They are defined as low-value general customers.

To further validate whether differences across clusters are significant across various dimensions, this study conducted one-way ANOVA. The results for R (interval since last purchase), S (purchase cycle), and A (average transaction amount) are shown in Table 4.

Table 4: Analysis of variance for dimensions R, S and A

		R	S	A
Cluster	Mean square	2.891	2.516	1.393
	df	3	3	3
Error	Mean square	0.034	0.033	0.337
	df	276	276	276
F		80.014	92.816	117.607
Sig.		0.000	0.000	0.000

The F-values for the R, S, and A dimensions are 80.014, 92.816, and 117.607 respectively, all with significance levels of 0.000 and p-values < 0.001. This indicates extremely significant intergroup differences across all clusters for the R, S, and A metrics. The between-group mean square for the A dimension far exceeded the error mean square, indicating that average single-transaction amount exhibited the highest discriminatory power across customer segments. For dimension S, the between-group mean square was 2.516 and the error mean square was 0.033, indicating strong discriminative power for consumption cycle. For dimension R, the between-group mean square was 2.891 and the error mean square was 0.034. Although the error was slightly larger, significant between-group differences were still evident. Overall, the clustering results are highly statistically significant, with differences between cluster centers not attributable to random factors. This further validates the effectiveness and scientific rigor of the proposed RSA model and improved clustering algorithm in multilingual customer segmentation.

3.3 Cluster Analysis Based on Customers' Current and Potential Value

Through the improved K-means clustering analysis in Section 3.2, we successfully segmented customers into four distinct value groups with significant characteristics, achieving precise identification of their static behavioral traits. To deepen our dynamic understanding of customer value and provide more forward-looking insights for multilingual customer service strategies, this section introduces a two-dimensional mapping model of current value and potential value based on the aforementioned clustering results. We conduct an in-depth analysis of customer value distribution under different weighting assignments to guide the scientific optimization of resource allocation and the implementation of personalized services.

Based on the RSA model and K-means clustering analysis, Company A's 366 customers were grouped into four clusters: High-Value Loyal Customers, High-Potential New Customers, Latent-Value Customers, and Low-Value General Customers. Based on current and potential value, these can be described as:

Cluster 1: High current value + high potential value;

Cluster 2: High current value + low potential value;

Cluster 3: Low current value + high potential value;

Cluster 4: Low current value + low potential value.

Cluster analysis is now conducted using different weighting factors for mapping each customer's current and potential value.

3.3.1 Cluster Analysis with Weight Mapping [2/3:1/3]

When the weight ratio between the current value and potential value of a customer is [2/3:1/3],

the clustering analysis results are shown in Figure 3.

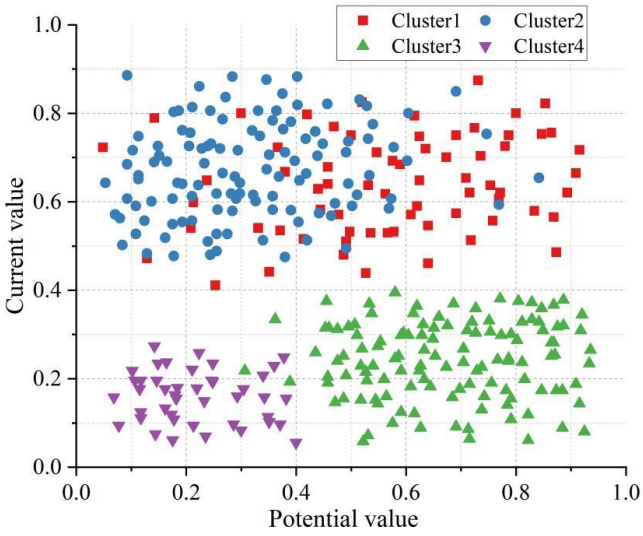

Figure 3: The clustering analysis results when the weight ratio is [2/3: 1/3]

At this point, the model identifies 63 customers in Cluster 1 (“High Current Value + High Potential Value”), 107 in Cluster 2 (“High Current Value + Low Potential Value”), 145 in Cluster 3 (“Low Current Value + High Potential Value”), and 51 in Cluster 4 (“Low Current Value + Low Potential Value”). These figures differ from the actual customer characteristics. The mapping weights indicate the company's prioritization of each value category. When prioritizing current customer value, the mapping range for current value is larger than that for potential value. Both Cluster 1 and Cluster 2 customers exhibit high current value, with the primary difference lying in their potential value. When the mapping range for current value is narrow, the distinction becomes less pronounced, making it difficult for the clustering model to effectively differentiate between Cluster 1 and Cluster 2 customers.

3.3.2 Cluster Analysis with a Weight Mapping of [1:1]

When the weight ratio between the customer's current value and potential value is [1:1], the clustering analysis results are shown in Figure 4.

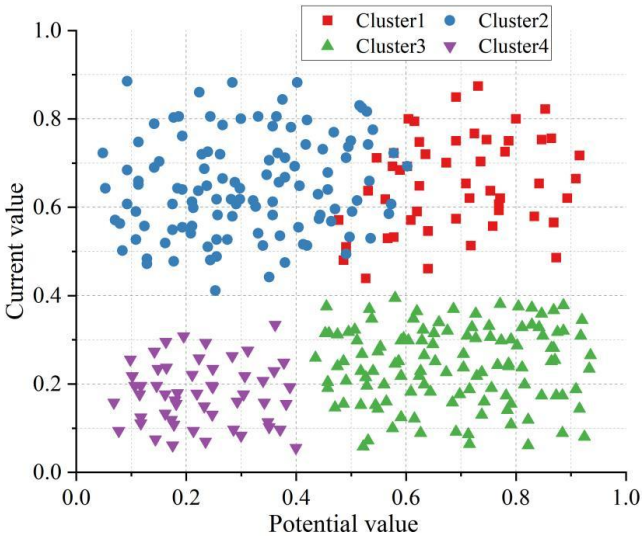

Figure 4: The clustering analysis results when the weight ratio is [1:1]

The four clusters identified by the current value-potential value clustering under a 1:1 weighting ratio are 45, 125, 142, and 54, respectively. These results closely align with the RSA model clustering outcomes, achieving an identification accuracy of 96.99%. When a customer's current value and potential value fall within the same mapping range, their relative importance in terms of current and potential value is comparable. In such cases, the clustering results more closely correspond to those derived from the RSA model-based classification.

4 Application of a Clustering Algorithm-Based Multilingual Customer Service System for Business English

Based on Company A's data, this study empirically validates the effectiveness and scientific rigor of the RSA model and the improved K-means algorithm in customer value segmentation. To further demonstrate the practical value of these methods in real business systems—particularly in multilingual customer service scenarios for business English—this chapter integrates the clustering analysis model into a customer service system using a real-world case study from a bank. It details the implementation and performance of this application.

4.1 Implementation of System Customer Management

Extract relevant data from the bank's database tables, establish associations using customer ID as the primary key, and construct a star schema data warehouse. The selected research data spans January 2023 to January 2024, comprising 28,342 total records. After preprocessing—including data cleansing, transformation, and attribute extraction—five metrics were applied: P (total customer assets), WP (total wealth management product value), R (interval since last transaction), S (transaction frequency), and A (single wealth management product transaction amount). These were analyzed using an improved clustering data mining algorithm. Customer statistics based on the improved K-means algorithm are presented in Table 5.

Table 5: Customer statistics based on the improved K-means algorithm

	P/RMB	WP/RMB	P/days	S	A/RMB
Cluster1	223,668	79,943	14	24	13,945
Cluster2	142,363	12,577	43	8	11,201
Cluster3	153,607	30,810	164	14	8,956
Cluster4	2,645	2,007	384	2	859

Cluster 1: High-Value Loyal Customers These customers hold the highest total assets and wealth management product holdings, averaging ¥220,000 in total assets and nearly ¥80,000 in wealth management products. They exhibit short intervals between recent transactions, high consumption frequency (24 transactions annually), and substantial single-transaction amounts. This indicates strong financial capacity, active trading, and high loyalty, making them the bank's most core premium customers. They warrant priority maintenance and in-depth service.

Cluster 2: High-Potential New Customers: Though their total assets and wealth management holdings are lower than high-value customers, their average transaction amount reaches ¥11,201, indicating strong spending power and investment intent. While their transaction intervals are short, their transaction frequency is low, suggesting they are new customers or have yet to establish stable transaction habits. They possess high conversion and growth potential, requiring the bank to strengthen guidance and nurturing.

Cluster 3: Latent Value Customers: With mid-range total assets and wealth management

product value, their most recent transaction occurred 164 days ago, indicating low recent activity. Historically high transaction frequency (14 times) suggests past loyalty. These customers are in a “dormant” or “semi-churn” state but retain substantial asset bases and spending power. Banks should employ reactivation strategies to regain their attention.

Cluster 4: Low-value general customers: These customers exhibit the lowest levels of total assets, wealth management product value, transaction frequency, and average transaction amount. They also have the longest consumption intervals, exceeding one year, resulting in overall low value. These customers may be occasional users or low-frequency consumers. Banks can provide low-cost standardized services or guide them toward gradually increasing value through basic products.

In summary, by integrating the enhanced K-means algorithm with the RSA model, banks can effectively identify customer segments across different value tiers and develop differentiated service strategies for each category. This approach enables optimized resource allocation and maximizes customer value. The methodology demonstrates high practicality and scalability within multilingual customer service systems for business English.

4.2 System Performance Testing

After implementing the customer management functionality, this section will conduct comprehensive performance testing on the multilingual customer service system integrated with the clustering algorithm to evaluate its processing efficiency and stability.

System performance testing primarily focuses on response times to user operations. A key feature of this system is its ability to classify bank customers using the K-means method. As the number of processed entries increases, processing time correspondingly rises. Test results are shown in Figure 5.

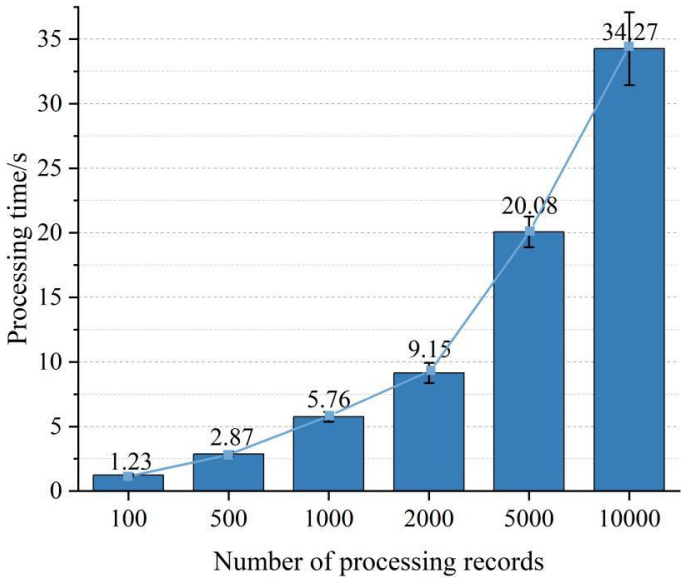

Figure 5: System performance test results

As the number of processed records increases, the system processing time shows an upward trend. However, in absolute terms, it remains within an acceptable range. For instance, processing 100 records takes only 1.23 seconds, while processing 10,000 records requires just 34.27 seconds. These results demonstrate the system's excellent scalability and stability, enabling it to efficiently handle bank-level customer data volumes and meet the real-time and response speed requirements of multilingual customer service systems.

5 Conclusion

This paper systematically proposes and validates a customer value segmentation method based on an improved RSA model and an optimized clustering algorithm, addressing the optimization requirements of multilingual customer service systems in business English.

Through empirical analysis of 366 active customer records from a company, the algorithm clearly segments customers into four distinct groups: 43 high-value loyal customers (cluster centers R=0.112, S=0.183, A=0.874), 127 high-potential new customers, 142 dormant value customers, and 54 low-value general customers. Single-factor ANOVA further confirmed highly significant differences across all clusters in the three dimensions of R, S, and A, indicating the segmentation results possess high statistical significance.

Finally, testing this model within a real-world banking system demonstrated that the customer service system integrating this algorithm exhibits excellent performance and practicality. When processing 10,000 records, the system achieved a response time of only 34.27 seconds, meeting real-time and stability requirements in a multilingual environment. This research provides effective theoretical methods and data support for gaining precise insights into customer value and formulating differentiated service strategies in multilingual contexts.

Funding

This work was supported by Industry-University Cooperation and Collaborative Education Project entitled "Research on the Teaching Reform of Integrating Moral Elements into the Comprehensive Business English Course". (Project No.: 220602054283203)

About the Author

Yang Wang was born in Jiayu, Hubei Province, China, in 1979.He is a lecturer in Zhejiang Shuren University. He received his bachelor's degree from Zhongnan University of Economics and Law and his master's degree from Shanghai International Studies University. His research interests include translation, ESL teaching methodology and business English. E-mail: fortune868686@163.com

References

[1] Yan, Z., Lu, X., Chen, Y., & Wang, K. (2023). Institutional distance, internationalization speed and cross-border e-commerce platform utilization. Management decision, 61(1), 176-200.

[2] Ma, S., Chai, Y., & Zhang, H. (2018). Rise of Cross-border E-commerce Exports in China. China & World Economy, 26(3), 63-87.

[3] Holmqvist, J., Van Vaerenbergh, Y., & Grönroos, C. (2017). Language use in services: Recent advances and directions for future research. Journal of Business Research, 72, 114-118.

[4] Xiao, P., Luo, X., & Daly, S. P. (2020). Language skills in business negotiation from the perspective of adaptation. International Journal of multidisciplinary and current educational research, 2(4), 181-187.

[5] Topal, I. H. (2024). Tandem language exchange application: A telecollaborative experience of linguistic and cultural exchange. Journal of Digital Educational Technology, 4(1), ep2408.

[6] Tseng, M. L., Wu, K. J., Chiu, A. S., Lim, M. K., & Tan, K. (2018). Service innovation in sustainable product service systems: Improving performance under linguistic preferences. International Journal of Production Economics, 203, 414-425.

[7] Javed, R. (2023). MULTILINGUAL CHATBOTS IN CUSTOMER SERVICE: A COMPUTATIONAL LINGUISTICS AND INFORMATION SYSTEMS STUDY. Multidisciplinary Research in Computing Information Systems, 3(4), 229-239.

[8] Risku, H., Pichler, T., & Wieser, V. (2017). Transcreation as a translation service: Process requirements and client expectations. Across Languages and Cultures, 18(1), 53-77.

[9] Pombal, J., Agrawal, S., & Martins, A. F. (2024, November). Improving context usage for translating bilingual customer support chat with large language models. In Proceedings of the Ninth Conference on Machine Translation (pp. 993-1003).

[10] Woydack, J. (2019). Language management and language work in a multilingual call center: An ethnographic case study. Revista Internacional de Organizaciones, (23), 79-105.

[11] Wołk, A., Skowrońska, H., & Skubis, I. (2021). Multilingual Chatbot for E-Commerce: Data Generation and Machine Translation. PACIS 2021 Proceedings, 232, 1-14.

[12] Kasinathan, V., Mustapha, A., & Bin, C. K. (2021). A customizable multilingual chatbot system for customer support. Annals of Emerging Technologies in Computing (AETiC), 5(5), 51-59.

[13] Firdaus, M., Ekbal, A., & Bhattacharyya, P. (2020, May). Incorporating politeness across languages in customer care responses: Towards building a multi-lingual empathetic dialogue agent. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 4172-4182).

[14] Vanjani, M., Aiken, M., & Park, M. (2019). Chatbots for multilingual conversations. Journal of Management Science and Business Intelligence, 4(1), 19-24.

[15] Shaik, M. (2024). Advanced Neural Networks for Multilingual Customer Service. IJLRP-International Journal of Leading Research Publication, 5(10).

[16] Yu, H., Zhang, Y., & Liu, Y. (2024). Application and optimization analysis of language translation function in cross-border E-commerce platforms. Journal of Computational Methods in Sciences and Engineering, 14727978251371937.

[17] Benita, J., Kumar, R. M. R., & Reddy, G. M. (2024, December). Multi-Language Conversational Agent for Tech Support: Design and Implementation. In 2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA) (pp. 1-7). IEEE.

[18] Lin, J., & Lei, F. (2025, March). Design and Simulation of Multilingual Intelligent

Language Service System for Cross-border Logistics Based on Multilingual Pre-training Modeling. In Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence (pp. 1482-1488).

[19] Ilagan, J. B., Ilagan, J. R., Zulueta, P. Y., & Rodrigo, M. M. (2024, May). Optimizing Conversational Commerce Involving Multilingual Consumers Through Large Language Models' Natural Language Understanding Abilities. In International Conference on Human-Computer Interaction (pp. 47-59). Cham: Springer Nature Switzerland.

[20] Chow, R., Suen, K. Y., & Lam, A. Y. (2025). On Leveraging Large Language Models for Multilingual Intent Discovery. ACM Transactions on Management Information Systems, 16(1), 1-17.

[21] Adeniyi, A. E., Olagunju, M., Awotunde, J. B., Abiodun, M. K., Awokola, J., & Lawrence, M. O. (2022, July). Augmented intelligence multilingual conversational service for smart enterprise management software. In International Conference on Computational Science and Its Applications (pp. 476-488). Cham: Springer International Publishing.

[22] Qi, S. (2025). Application and optimization of natural language processing technology in intelligent customer service system. Journal of Theory and Practice of Management Science, 5(4), 5-8.

[23] Delbaz, A. (2018). Natural Language Processing for Multilingual Chatbots: Techniques, Challenges, and Applications. International Journal of Artificial Intelligence and Machine Learning, 1(2).

[24] Amiri, S. M. H. (2025). Beyond language barriers: Multilingual NLP and voice recognition for global connectivity. International Journal of Science and Research Archive, 15, 406-419.

[25] Nigro, L., Cicirelli, F., & Fränti, P. (2023). Parallel random swap: an efficient and reliable clustering algorithm in Java. Simulation Modelling Practice and Theory, 124, 102712.

[26] Chen, Z. L. (2022). Research and application of clustering algorithm for text big data. Computational Intelligence and Neuroscience, 2022(1), 7042778.

[27] Nagaraj, P., Birunda, S. S., Venkatesh, R., Muneeswaran, V., Narayanan, S. K., Shree, U. D., & Sunethra, B. (2022, January). Automatic and Adaptive Segmentation of Customer in R framework using K-means Clustering Technique. In 2022 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-5). IEEE.

[28] Oumoussa, I., & Saidi, R. (2025). Automated Microservices Identification through Business Process Analysis: A Semantic-driven Clustering Approach. IEEE Access.

[29] Saadi, W., Laallam, F. Z., Mezati, M., Youmbai, D. L., & Messaoudi, N. E. (2024). Enhancing emotion detection on Twitter: an ensemble clustering approach utilizing emojis and keywords across multilingual datasets. Studies in Engineering and Exact Sciences, 5(2), e10548-e10548.

[30] Kim, W., Nam, K., & Son, Y. (2023). Categorizing affective response of customer with novel explainable clustering algorithm: The case study of Amazon reviews. Electronic

Commerce Research and Applications, 58, 101250.

[31] Magoo, C., & Singh, M. (2023). A novel hybrid approach for intent creation and detection using K-Means-Based topic clustering and heuristic-based capsule network. International Journal of Information Technology & Decision Making, 22(06), 1923-1960.