# SIRIEMA: A Framework to Enhance Clustering Stability by Fusing Multimodal Data

Anonymous ACL submission

## Abstract

In marketing, customer segmentation is critical for creating content tailored to specific consumer groups. The stability of these segments, hinging on an algorithm's ability to form similar groupings consistently, is essential for effective marketing strategies and higher conversion rates. Traditionally, segment stability can be improved by relying on structured data like age and purchase history and integrating this data with textual information, such as social mnedia posts and product reviews. This study presents SIRIEMA, a multimodal framework de**SI**gned to enhance cluste**RI**ng stability by fusing cat**E**gorical, nu**M**eric**A**l, and textual data. Our proposal utilizes a transformer-based model for text, data fusion techniques, and generative models like variational autoencoders and generative adversarial networks. Using real-world datasets, SIRIEMA showed enhanced clustering stability and quality compared to existing methods. This research represents a novel approach to customer segmentation and paves the way for future exploration of data fusion techniques in the context of marketing and other applications.

## 1 Introduction

Customer segmentation provides valuable insights into customer preferences and behaviors, allowing for a more refined understanding of distinct consumer groups (Varadarajan, 2020). By acquiring these insights, marketers can tailor content to address each segment's unique needs and challenges (Leung et al., 2022).

A fundamental aspect of effective customer segmentation is clustering stability. This term refers to an algorithm's ability to consistently generate similar customer segments across various runs or data subsets, a feature crucial for ensuring customer grouping based on enduring traits or behaviors (Von Luxburg et al., 2010). Stable clustering not only bolsters the effectiveness of marketing campaigns but also significantly elevates conversion rates (Cortez et al., 2021; Ray, 2019; Ko et al., 2022). Conversely, instability in clustering, even with careful data preparation, can result in misleading marketing strategies. This instability often leads to campaigns that fail to connect with the target audience, resulting in decreased revenue and diminished customer satisfaction (Xie et al., 2016; Akay and Yüksel, 2018; He and Yu, 2019).

The literature has proposed various methods to enhance clustering stability, explicitly focusing on structured data like categorical and numeric data (Hajibaba et al., 2020; He and Yu, 2019; Lee et al., 2022). A popular method is the Deep Embedding Clustering With Mixed Data Using Soft-Target Network (Mixed DEC + SU), an algorithm that leverages a deep learning framework for clustering (Lee et al., 2022). This method uses a stacked autoencoder to learn latent feature representations and perform a clustering task using a soft assignment procedure. Although the Mixed DEC + SU strategy is quite effective, it faces challenges when applied to multimodal data encompassing structured and textual forms.

Building on this, existing research posits that integrating textual data with structured data could further enhance clustering stability in customer segmentation (Balducci and Marinova, 2018; Fresneda et al., 2021; Vo et al., 2021). Such integration is supported by evidence showing that textual data offer rich, contextual insights beyond what structured data alone can provide (Tay et al., 2021; Vaswani et al., 2017). The fusion of textual and structured data holds promise for enhancing clustering stability and providing a deeper, more nuanced insight into customer segments (Balducci and Marinova, 2018).

Therefore, in this article, we introduce SIRIEMA, a novel multimodal framework designed to enhance clustering stability by fusing categorical, numerical, and textual data. Our solution consists of three principal components: a transformer-based embedding model, a data fusion component, and a generative-based model.

The transformer-based embedding model is essential for converting textual data into meaningful embeddings, capturing intricate patterns and relationships. The data fusion component fuses the derived embeddings with categorical and numerical data to form a comprehensive feature space. Taking its output, a generative-based model, such as Variational Autoencoder (VAE) or Generative Adversarial Network (GAN), is then employed to refine the clustering process further. By capturing the intricate relationships within the data, generative models ensure that clusters are cohesive and consistent, reducing variance and leading to more stable clustering outcomes (Yang et al., 2020; Harshvardhan et al., 2020).

We employed five established stability measures to evaluate its effectiveness: Adjusted Rand Index (ARI), Adjusted Mutual Information Score (AMIS), BagClust (BG), Hierarchical Agglomerative Nesting (HAN), and Optimal Transport Alignment (OTA) — each one renowned for assessing cluster stability across varied contexts (Liu et al., 2022; Peyvandipour et al., 2020; Lall et al., 2021). The Davies–Bouldin Score (DBS) metric also evaluates cluster quality and separation. We selected the K-means algorithm for our evaluations, due to its straightforward nature and acknowledged instability when juxtaposed with other methods, such as hierarchical techniques (Zhou et al., 2022).

In our evaluation, we used real-world datasets, namely: *Yelp Dataset* (Dataset, 2014), *Melbourne Airbnb dataset* (Xie, 2019), *PetFinder.my* (Kaggle and PetFinder.my, 2019), and *Women's clothing reviews* (Brooks, 2017). To assess the robustness of our model, we benchmarked it against four prevailing strategies. The first strategy, **Structured**, strictly employs numerical and categorical data. The second, **Textual**, focuses exclusively on text embeddings. The third approach, **Combined Dataset - Structure Textual (CD-ST)**, integrates both structured and textual datasets, while the fourth, **Mixed DEC + SU**, assimilates mixed data categories to enhance convergence stability (Lee et al., 2022).

Our main contributions are as follows:

- We introduce SIRIEMA, a novel framework that effectively integrates categorical, numerical, and textual data, significantly enhancing clustering stability in multimodal environments;
- We demonstrate that by integrating categorical and numerical data with textual data within our multimodal framework, we can significantly improve the stability of clustering algorithms;
- We achieve state-of-the-art clustering stability with our multimodal framework, advancing the field of multimodal learning through enhanced data integration techniques;
- To the best of our knowledge, we are the first to integrate categorical, numerical, and textual data in a multimodal framework, significantly enhancing clustering stability.

The remainder of this article is organized as follows. Section 2 presents a state-of-the-art synthesis and discussions. Section 3 presents SIRIEMA. Section 4 details the experimental evaluations conducted. Section 5 contains our discussion. Finally, Section 6 presents the conclusions and directions for future work.

## 2 Related Work

Methods have been proposed in the literature to improve clustering stability with emphasis on categorical, numeric, and text data (Hajibaba et al., 2020; He and Yu, 2019; Lee et al., 2022; Prasad et al., 2015).

A discussion on clustering mixed panel datasets using Gower's distance and k-prototypes algorithms is offered in Akay's study (Akay and Yüksel, 2018). Panel datasets are commonly used in economics to analyze complex economic phenomena. The panel data matrix is constructed by combining data from different periods and different individuals or entities. The clustering method is applied to panel data analysis to solve the heterogeneity question of the dependent variable, which belongs to panel data, before the analysis. However, they need to consider incorporating textual data into the clustering process, such as customer reviews, which can offer valuable insights and enhance the clustering results, particularly in domains where sentiment or opinion analysis is crucial (Balducci and Marinova, 2018).

An Evolutionary K-Means (EKM) algorithm that uses clustering stability to evaluate partitions, namely Clustering Stability-based Evolutionary

KMeans (CSEKM), was proposed by He and Yu (2019). It addresses the initiation problem of K-Means by suggesting using at least one initial center from each underlying cluster. It uses cluster stability to evaluate partitions, making it more robust to noise and challenging clusters. However, while CSEKM focuses on addressing the initiation problem and incorporating clustering stability, it does not explicitly consider integrating multimodal data. Multiple modalities may capture richer patterns and relationships, improving clustering stability and potentially more accurate and reliable clustering results (Balducci and Marinova, 2018).

The development of a strategy to increase the stability of market segmentation solutions derived from binary empirical consumer data was proposed by Hajibaba et al. (2020). Through the combination the variable selection method proposed by Brusco (2004) and the global stability analysis introduced by Dolnicar and Lazarevski (2009), the strategy simultaneously selects the segmentation variables and the number of segments leading to high global stability levels. Although binary data can provide simplicity and ease of analysis, it may not convey the complexity and subtleties of consumer behavior; by restricting the analysis to binary variables, it is possible to neglect valuable information or subtleties in consumer preferences or attitudes. This could lead to a less comprehensive understanding of market segments and potentially suboptimal marketing strategy decision-making.

A novel non-linear Deep Encoder-Decoder framework to capture the cross-domain information for mixed data types is proposed by Sahoo and Chakraborty (2020). The authors discuss the challenge of representing data that contain mixed variable types, such as numerical and categorical variables. The joint distribution of mixed variables lies in a complex non-linear product space, making it challenging to represent the data in a suitable feature space. The representation of the data points can be carried out in a supervised or unsupervised manner. However, the proposed model's non-linear space can introduce complexities when dealing with cross-domain information, particularly when incorporating unstructured text data. This complexity can hinder the overall performance of the model (Balducci and Marinova, 2018).

A method called Deep Embedded Clustering (DEC) that simultaneously learns feature representations and cluster assignments using deep neural networks was proposed by Xie et al. (2016). The method learns a mapping from the data space to a lower-dimensional feature space in which it iteratively optimizes a clustering objective. The experimental evaluations on image and text corpora significantly improve over state-of-the-art methods. However, if the difference between soft assignment and target values is significant, DEC applications may suffer from convergence problems. To overcome these limitations, it was proposed by Lee et al. (2022) a deep embedded clustering framework, called Mixed DEC + SU, that can utilize mixed data to increase the convergence stability using soft-target updates derived from an enhanced deep Q-learning algorithm utilized in reinforcement learning. Integrating diverse data modalities and enhanced representation learning capabilities can provide a more accurate and reliable foundation for clustering analysis, resulting in better cluster assignments and more insightful clustering results, which are not seen in these works.

A new algorithm called uCLUST, which identifies clusters in unstructured data by capturing pattern similarity among objects was proposed by Prasad et al. (2015). The results demonstrate that uCLUST effectively clusters unstructured data and can be used in various fields such as libraries, insurance, and the world wide web. However, the proposed work considers only the frequency of words to calculate the similarity measure; language semantics and context of terms are not considered for clustering the document.

The strengths and weaknesses of these studies defined our approach. In particular, SIRIEMA improves upon these efforts by incorporating categorical, numerical, and textual features, resulting in a more complete representation of the data and significantly enhanced clustering stability.

## 3 Enhancing Clustering Stability in Multimodal Data Environments

This section introduces SIRIEMA, a framework that integrates a transformer-based model, a data fusion component, and a generative-based model to optimize data clustering.

### 3.1 Multimodal Framework

SIRIEMA has three key components: a transformer-based model, a data fusion component, and a generative-based model. Figure 1 provides a visualization of our framework.

We describe in detail each component that fol-

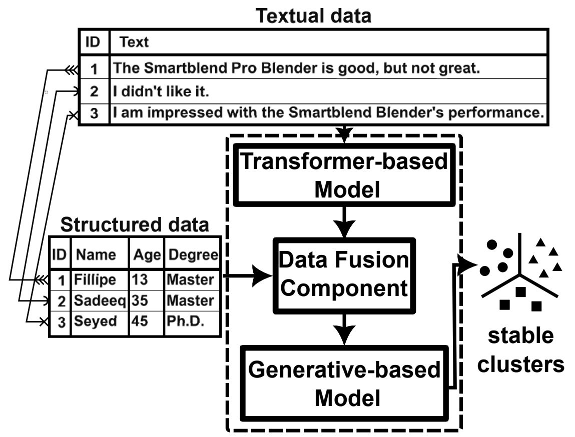


Figure 1: SIRIEMA combines text data with categorical and numerical features for enhanced clustering stability. It uses the Bidirectional Encoder Representations for Transformers (BERT) model for text features and adds a data fusion component to merge the BERT model's output with categorical and numerical features. These enriched features are then used within a VAE model.

lows:

1. **Transformer-based Model Component:** This component employs a pre-trained transformer-based model, including, but not limited to, BERT[1], GPT-3.5[2], Llama[3], and others[4]. Without specialized heads, these models are exclusively for embedding purposes, leveraging their extensive pre-existing knowledge. We denote the output of this process as $\mathbf{x}$, which provides our framework with robust encoding capabilities for textual information, thereby delivering significant advantages (Lin et al., 2022).

2. **Data Fusion Component:** This component receives the transformer-based model component's output ($\mathbf{x}$), along with categorical ($\mathbf{c}$) and numerical ($\mathbf{n}$) features as input, and produces an output denoted by $\mathbf{m}$, which any generative-based model then receives.

   We explored eight methods to integrate these features, each addressing the unique characteristics of their respective feature spaces. Drawing inspiration from the recent advancements in multimodal data fusion (Gao et al., 2020), these methods span from straightforward strategies such as simple concatenation to more intricate techniques leveraging Multilayer Perceptron (MLP). Table 1 presents all those methods.

3. **Generative-based Model Component:** This component processes the output $\mathbf{m}$ from the data fusion component. It employs well-established generative models, such as VAEs and GAN, to foster more cohesive and stable clustering solutions by deeply understanding the underlying data distributions. With this approach, we aim to establish robust clusters that capture subtle patterns and relationships within the data, ensuring consistent and reproducible outcomes across various data scenarios.

### 3.2 Datasets

We employed datasets encompassing various domains, including social media, tourism, pet services, and e-commerce.

The first dataset is from the Yelp public dataset challenge[5], which is a collection of user reviews and other related details from the Yelp platform. It involves structured features such as user-generated numerical details, including review counts and average ratings, and unstructured elements represented by the review texts. Next is the Melbourne Airbnb Open dataset[6], which gives a detailed insight into Airbnb listings in Melbourne, Australia. It encompasses structured details like price, number of reviews, review scores, and unstructured data in the listing descriptions and host information. Following this, the PetFinder.my Adoption Prediction dataset[7] offers structured information detailing the numerical and categorical characteristics of pet listings, in addition to unstructured data captured in the pet descriptions penned by the caretakers. Lastly, the Women's E-Commerce Clothing Reviews dataset[8] comprises customer reviews and ratings of women's clothes sold online, including structured data such as age, rating, and categorical details like department and class name. It also contains unstructured data, which comes as detailed review texts.

[1] huggingface.co/docs/transformers/model_doc/bert#transformers.BertModel.

[2] huggingface.co/spaces/yizhangliu/chatGPT.

[3] huggingface.co/meta-llama/Llama-2-7b.

[4] huggingface.co/docs/transformers/index.

[5] www.kaggle.com/datasets/yelp-dataset/yelp-dataset.

[6] www.kaggle.com/datasets/tylerx/melbourne-airbnb-open-data.

[7] www.kaggle.com/competitions/petfinder-adoption-prediction.

[8] www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews.

| # | Method | Equation |
|---|---|---|
| **1** | text only | $\mathbf{m} = \mathbf{x}$ |
| **2** | concatenation | $\mathbf{m} = (\mathbf{x}, \mathbf{c}, \mathbf{n})$ |
| **3** | MLP on categorical then concatenate | $\mathrm{m} = (\mathbf{x},\ MLP(\mathbf{c}),\ \mathbf{n})$ |
| **4** | individual MLP on categorical and numerical features then concatenate | $\mathbf{m} = (\mathbf{x}, MLP(\mathbf{c}), MLP(\mathbf{n}))$ |
| **5** | MLP on concatenated categorical and numerical features then concatenate | $\mathbf{m} = (\mathbf{x}, MLP(\mathbf{c}, \mathbf{n}))$ |
| **6** | attention on categorical and numerical features | $\mathbf{m} = \alpha_{x,x}\mathbf{W}_x\mathbf{x} + \alpha_{x,c}\mathbf{W}_c\mathbf{c} + \alpha_{x,n}\mathbf{W}_n\mathbf{n}$<br>$\alpha_{i,j} = \frac{exp(LeakyReLu(\mathbf{a}^T(\mathbf{W}_i\mathbf{x}_i, \mathbf{W}_j\mathbf{x}_j)))}{\sum_{k \in \{x,c,n\}} exp(LeakyReLu(\mathbf{a}^T(\mathbf{W}_i\mathbf{x}_i, \mathbf{W}_k\mathbf{x}_k)))}$ |
| **7** | gating on categorical and features and then sum (Rahman et al., 2020)(Gating) | $\mathbf{m} = \mathbf{x} + \alpha\mathbf{h}$<br>$\mathbf{h} = \mathbf{g}_c \odot (\mathbf{W}_c\mathbf{C}) + \mathbf{g}_n \odot (\mathbf{W}_n\mathbf{n}) + b_h$<br>$\alpha = min(\frac{\|\mathbf{x}\|_2}{\|\mathbf{h}\|_2}) * \beta, 1)$<br>$\mathbf{g}_i = R(\mathbf{W}_{gi}(\mathbf{i}, \mathbf{x}) + b_i)$<br>where $\beta$ is a hyperparameter and R is an activation function |
| **8** | weighted feature sum on text, categorical, and numerical features (Weighted Sum) | $\mathbf{m} = \mathbf{x} + \mathbf{w}_c \odot \mathbf{W}_c\mathbf{c} + \mathbf{w}_n \odot \mathbf{W}_n\mathbf{n}$ |

Table 1: Feature integration methods. Uppercase bold letters represent 2D matrices, lowercase bold letters represent 1D vectors, and non-bold, lowercase letters are scalar values.

The richness and diversity of these datasets provide a solid ground for performing a robust stability analysis.

### 3.3 Evaluation

To assess the model's effectiveness, we used five stability measures: ARI, AMIS, BG, HAN, and OTA. The ARI and AMIS measure clustering similarity, with ARI adjusting for chance in paired element clustering and AMIS based on mutual information. BG evaluates clustering consistency across data subsets, while HAN applies bootstrap techniques to estimate cluster stability. Lastly, OTA, the Optimal Transport Alignment algorithm, compares clusterings using the theory of optimal transport. These measures provide a multifaceted view of our model's performance, emphasizing clustering stability and effectiveness. Additionally, DBS was used to evaluate clustering compactness and separation.

This approach is further complemented by employing the K-means algorithm for cluster computation. We compared four distinct strategies across two sample sizes: 500 and the entire dataset. The strategies include **Structured**, which incorporates only numerical and categorical data, and **Textual**, which utilizes embeddings derived from text. For generating embeddings for all datasets, the bert-base-uncased model[9] is employed. Additionally, **CD-ST** employs a concatenation of structured and textual data; **Mixed DEC + SU** utilizes mixed data from categorical and numerical features to increase convergence stability using soft-target updates.

### 3.4 Experimental Settings

This section provides a detailed account of implementing our multimodal framework. Central to this framework is the BERT model, serving as the foundational model for text features and tokenization[10]. It incorporates a data fusion component that combines the BERT model output with categorical and numerical features, generating specific multimodal attributes. We used these enriched features as the final model within a VAE model.

The loss function merges reconstruction loss with Kullback-Leibler (KL) divergence, adding structure to the latent space for better generalization. The VAE was designed with layers of 768, 500, 300, and 200 units. The training was conducted for 15 epochs at a $3x10^{-3}$ learning rate using the AdamW optimizer. These adjustments enable the model to effectively manage complex categorical, numerical, and textual data clustering, ensuring stable and reliable performance (Lim et al., 2020). We divided the entire dataset into 80% for

[9] huggingface.co/bert-base-uncased.

[10] huggingface.co/docs/transformers/model_doc/bert#transformers.BertModel.

training, 10% for validation, and 10% for testing.

In our initial experiments, we assessed the performance of data fusion component methods using two sample sizes: 500 and the entire dataset. Table 2 shows each method's mean validation loss and 95% confidence interval, highlighting key results with underlining. This comprehensive testing is essential to determine the most effective method of integration suited to the diverse characteristics of the data.

We excluded methods **3** (*MLP on categorical then concatenate*) and **4** (*individual MLP on categorical and numerical features then concatenate*) for the Yelp dataset due to their lack of categorical features. During our evaluation, method **7** (*gating*) emerged as the optimal approach for both the Yelp and Airbnb datasets, whereas method **5** (*MLP on concatenated categorical and numerical features, then concatenate*) and method **6** (*attention on categorical and numerical features*) performed best for the PetFinder.my and Women's Clothing Reviews datasets, respectively.

In the second phase of our experiments, we selected the most effective method from the data fusion component for each dataset, taking into account various sample sizes. Following this selection, we performed a comprehensive evaluation of our multimodal model's results, which involved a comparative analysis of four distinct strategies by applying the five stability metrics we had established earlier.

The model was developed using PyTorch[11] and is made available at a GitHub repository[12]. It ran on a system equipped with two Titan X Graphics processing unit (GPU)s, each having 12 GiB of Random Access Memory (RAM). The architecture, including the methods in data fusion component, was inspired by Gu and Budhkar (2021).

## 4 Experimental Results

Table 3 presents the stability metric results for the test dataset, using both the 500 sample size and the entire dataset with strategies as follows: **A**: Structured; **B**: Textual; **C**: CD-ST; **D**: Mixed DEC + SU; and **E**: SIRIEMA. Underlined values highlight the best outcomes, whereas bold values denote results from SIRIEMA. We performed the experiment ten times for each sample and metric and reported the mean results with a 95% confidence interval. For the Yelp dataset, SIRIEMA excelled in ARI and AMIS metrics for 500 samples and the entire dataset, showing robust clustering of multimodal data. It also outperformed in BG, HAN, and OTA metrics, emphasizing its proficiency in larger datasets. In the Airbnb dataset, SIRIEMA demonstrated superior performance and consistency across ARI, AMIS, BG, HAN, and OTA metrics for 500 samples and the entire dataset, highlighting its precision and adaptability. For the PetFinder.my dataset, SIRIEMA outshone alternatives in ARI, AMIS, BG, and HAN for both 500 samples and the entire dataset while ranking second in the OTA metric for 500 samples. In the Women's Clothing reviews dataset, SIRIEMA showed robust scalability and robustness in ARI and AMIS for 500 samples and the entire dataset. It also maintained superiority in BG, HAN, and OTA metrics, confirming its effectiveness in handling complex datasets.

Table 4 presents the DBS for each strategy across different sample sizes; underlined scores are the best in each row, while bold ones highlight the results of SIRIEMA. In the Yelp dataset, the Structured strategy reached a DBS score of 0.66 ± 0.08 for the entire dataset, while SIRIEMA excelled with a score of 0.20 ± 0.01, indicating high efficacy. In the Airbnb dataset, the CD-ST method achieved its highest score of 2.79 ± 0.19 for 500 samples. SIRIEMA showcased notable performance with a mean score of 0.10 ± 0.01 over the entire dataset, demonstrating its effectiveness in grouping similarity. For the PetFinder.my dataset, SIRIEMA consistently decreased the DBS as the sample size grew, nearly reaching optimal clustering at the entire dataset level, signifying excellent adaptability and efficient cluster separation. SIRIEMA demonstrated superior clustering efficiency and reliability in the Clothing dataset, indicated by the lowest and most consistent DBS scores across both 500 samples and the entire dataset.

## 5 Discussion

SIRIEMA demonstrated superior effectiveness in the metrics of the ARI and AMIS, consistently outperforming alternatives for 500 samples and the entire dataset. These consistently high scores highlight the model's robustness and precision, especially in handling large, complex multimodal datasets, making it ideal for applications requiring

[11] pytorch.org.

[12] anonymous.4open.science/r/SIRIEMA-6AD3/README.md. (note: this link will be replaced by the GitHub one in the case of paper's acceptance.)

| | **Yelp** | | **Airbnb** | | **PetFinder.my** | | **Clothing** | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **500 s.** | **Entire dataset** | **500 s.** | **Entire dataset** | **500 s.** | **Entire dataset** | **500 s.** | **Entire dataset** |
| **1** | 246.73 ± 7.37 | 239.41 ± 5.35 | 158.05 ± 7.76 | 136.27 ± 4.51 | 70.29 ± 6.59 | 66.05 ± 7.84 | 35.97 ± 9.04 | 38.58 ± 4.17 |
| **2** | 241.86 ± 5.49 | 232.5 ± 3.71 | 149.78 ± 9.83 | 122.78 ± 7.45 | 82.82 ± 8.62 | 77.45 ± 5.88 | 48.27 ± 5.5 | 44.63 ± 9.23 |
| **3** | - | - | 144.79 ± 4.68 | 119.92 ± 5.87 | 94.66 ± 3.59 | 88.82 ± 9.31 | 20.73 ± 6.72 | 23.77 ± 6.4 |
| **4** | - | - | 146.71 ± 6.7 | 120.18 ± 3.93 | 80.22 ± 7.34 | 80.54 ± 6.72 | 19.28 ± 8.24 | 17.64 ± 3.97 |
| **5** | 243.54 ± 5.13 | 237.57 ± 7.94 | 151.6 ± 7.76 | 123.6 ± 3.58 | 45.28 ± 8.94 | 42.29 ± 8.28 | 48.6 ± 8.44 | 50.10 ± 6.78 |
| **6** | 241.55 ± 3.46 | 231.79 ± 3.88 | 155.11 ± 5.15 | 129.63 ± 6.01 | 60.99 ± 5.46 | 57.34 ± 4.91 | 14.88 ± 5.11 | 16.55 ± 6.79 |
| **7** | 189.09 ± 8.22 | 185.7 ± 5.25 | 107.1 ± 3.54 | 85.23 ± 3.84 | 70.05 ± 6.18 | 66.59 ± 6.28 | 56.06 ± 5.62 | 54.21 ± 8.58 |
| **8** | 245.43 ± 9.09 | 239.57 ± 9.62 | 125.17 ± 8.52 | 105.86 ± 7.59 | 74.4 ± 6.86 | 68.59 ± 8.35 | 63.04 ± 8.37 | 62.25 ± 9.7 |

Table 2: The mean validation loss, accompanied by a 95% confidence interval, is provided for all methods in the data fusion component across all sample sizes for all datasets, with the best results underlined.

| | | **Yelp** | | **Airbnb** | | **PetFinder.my** | | **Clothing** | |
|---|---|---|---|---|---|---|---|---|---|
| | | **500 s.** | **Entire Dataset** | **500 s.** | **Entire Dataset** | **500 s.** | **Entire Dataset** | **500 s.** | **Entire Dataset** |
| ARI | **A** | 0.56 ± 0.08 | 0.55 ± 0.10 | 0.77 ± 0.02 | 0.89 ± 0.02 | 0.50 ± 0.02 | 0.52 ± 0.05 | 0.54 ± 0.05 | 0.99 ± 0.00 |
| | **B** | 0.84 ± 0.02 | 0.96 ± 0.00 | 0.93 ± 0.00 | 0.95 ± 0.00 | 0.51 ± 0.02 | 0.67 ± 0.07 | 0.67 ± 0.06 | 0.89 ± 0.03 |
| | **C** | 0.51 ± 0.10 | 0.54 ± 0.10 | 0.80 ± 0.02 | 0.88 ± 0.02 | 0.49 ± 0.03 | 0.56 ± 0.03 | 0.60 ± 0.02 | 0.90 ± 0.03 |
| | **D** | 0.30 ± 0.02 | 0.93 ± 0.03 | 0.52 ± 0.03 | 0.79 ± 0.01 | 0.49 ± 0.02 | 0.52 ± 0.01 | 0.46 ± 0.07 | 0.70 ± 0.01 |
| | **E** | **0.96 ± 0.01** | **1.00 ± 0.00** | **0.87 ± 0.03** | **1.00 ± 0.00** | **1.00 ± 0.00** | **0.99 ± 0.00** | **1.00 ± 0.00** | **1.00 ± 0.00** |
| AMIS | **A** | 0.53 ± 0.07 | 0.52 ± 0.10 | 0.71 ± 0.02 | 0.83 ± 0.02 | 0.55 ± 0.01 | 0.58 ± 0.04 | 0.53 ± 0.06 | 0.98 ± 0.00 |
| | **B** | 0.78 ± 0.02 | 0.93 ± 0.01 | 0.89 ± 0.01 | 0.91 ± 0.00 | 0.60 ± 0.02 | 0.73 ± 0.05 | 0.63 ± 0.05 | 0.84 ± 0.03 |
| | **C** | 0.48 ± 0.10 | 0.51 ± 0.11 | 0.74 ± 0.01 | 0.83 ± 0.02 | 0.56 ± 0.02 | 0.63 ± 0.02 | 0.58 ± 0.02 | 0.85 ± 0.03 |
| | **D** | 0.26 ± 0.03 | 0.88 ± 0.03 | 0.59 ± 0.01 | 0.73 ± 0.01 | 0.57 ± 0.02 | 0.60 ± 0.02 | 0.40 ± 0.07 | 0.67 ± 0.01 |
| | **E** | **0.93 ± 0.02** | **0.99 ± 0.00** | **0.84 ± 0.03** | **1.00 ± 0.00** | **1.00 ± 0.00** | **0.99 ± 0.00** | **1.00 ± 0.00** | **1.00 ± 0.01** |
| BG | **A** | 0.92 ± 0.02 | 1.00 ± 0.0 | 0.93 ± 0.01 | 0.95 ± 0.05 | 0.72 ± 0.02 | 0.73 ± 0.13 | 0.79 ± 0.02 | 0.76 ± 0.06 |
| | **B** | 0.97 ± 0.01 | 0.99 ± 0.00 | 0.99 ± 0.00 | 0.99 ± 0.00 | 0.72 ± 0.02 | 0.83 ± 0.02 | 0.93 ± 0.01 | 0.97 ± 0.02 |
| | **C** | 0.89 ± 0.04 | 0.94 ± 0.04 | 0.95 ± 0.01 | 0.97 ± 0.01 | 0.71 ± 0.02 | 0.76 ± 0.09 | 0.90 ± 0.01 | 0.97 ± 0.02 |
| | **D** | 0.76 ± 0.02 | 0.99 ± 0.01 | 0.69 ± 0.02 | 0.94 ± 0.02 | 0.68 ± 0.02 | 0.78 ± 0.04 | 0.82 ± 0.02 | 0.83 ± 0.03 |
| | **E** | **0.99 ± 0.01** | **1.00 ± 0.00** | **0.97 ± 0.00** | **0.97 ± 0.01** | **0.79 ± 0.04** | **0.85 ± 0.02** | **0.90 ± 0.02** | **0.97 ± 0.05** |
| HAN | **A** | 0.85 ± 0.07 | 0.97 ± 0.08 | 0.87 ± 0.01 | 0.94 ± 0.00 | 0.61 ± 0.03 | 0.63 ± 0.03 | 0.84 ± 0.03 | 0.92 ± 0.05 |
| | **B** | 0.92 ± 0.01 | 0.98 ± 0.00 | 0.96 ± 0.00 | 0.97 ± 0.00 | 0.61 ± 0.03 | 0.77 ± 0.01 | 0.86 ± 0.01 | 0.95 ± 0.01 |
| | **C** | 0.79 ± 0.04 | 0.81 ± 0.05 | 0.90 ± 0.01 | 0.94 ± 0.01 | 0.55 ± 0.03 | 0.63 ± 0.03 | 0.81 ± 0.03 | 0.95 ± 0.02 |
| | **D** | 0.61 ± 0.01 | 0.96 ± 0.01 | 0.58 ± 0.02 | 0.90 ± 0.01 | 0.54 ± 0.03 | 0.62 ± 0.04 | 0.70 ± 0.04 | 0.85 ± 0.02 |
| | **E** | **0.99 ± 0.0** | **1.00 ± 0.00** | **0.94 ± 0.02** | **1.00 ± 0.00** | **0.85 ± 0.02** | **0.86 ± 0.02** | **1.00 ± 0.00** | **1.00 ± 0.00** |
| OTA | **A** | 0.49 ± 0.01 | 0.50 ± 0.00 | 0.70 ± 0.05 | 0.73 ± 0.00 | 0.17 ± 0.03 | 0.13 ± 0.00 | 0.40 ± 0.08 | 0.40 ± 0.01 |
| | **B** | 0.73 ± 0.00 | 0.73 ± 0.01 | 0.80 ± 0.01 | 0.80 ± 0.01 | 0.37 ± 0.04 | 0.48 ± 0.02 | 0.60 ± 0.04 | 0.67 ± 0.03 |
| | **C** | 0.65 ± 0.05 | 0.50 ± 0.00 | 0.74 ± 0.01 | 0.73 ± 0.00 | 0.16 ± 0.05 | 0.45 ± 0.01 | 0.59 ± 0.04 | 0.67 ± 0.02 |
| | **D** | 0.54 ± 0.04 | 0.67 ± 0.00 | 0.20 ± 0.01 | 0.09 ± 0.01 | 0.25 ± 0.03 | 0.15 ± 0.05 | 0.59 ± 0.03 | 0.57 ± 0.01 |
| | **E** | **0.88 ± 0.02** | **0.89 ± 0.00** | **0.77 ± 0.01** | **0.88 ± 0.01** | **0.60 ± 0.06** | **0.77 ± 0.23** | **0.64 ± 0.25** | **0.89 ± 0.01** |

Table 3: Comparing the stability metrics of various strategies across different sample sizes on four distinct datasets on the test dataset. Underlined scores are the best in each row, while bold ones highlight the results of SIRIEMA. The strategies are as follows: **A**: Structured; **B**: Textual; **C**: CD-ST; **D**: Mixed DEC + SU; and **E**: SIRIEMA.

| | **Yelp** | | **Airbnb** | | **PetFinder.my** | | **Clothing** | |
|---|---|---|---|---|---|---|---|---|
| **Strategy** | **500 s.** | **Entire dataset** | **500 s.** | **Entire dataset** | **500 s.** | **Entire dataset** | **500 s.** | **Entire dataset** |
| **Structured** | 2.51 ± 0.12 | 0.66 ± 0.08 | 3.38 ± 0.17 | 3.43 ± 0.17 | 1.88 ± 0.03 | 1.96 ± 0.09 | 1.88 ± 0.13 | 1.84 ± 0.11 |
| **Textual** | 3.10 ± 0.21 | 3.14 ± 0.09 | 2.59 ± 0.10 | 2.63 ± 0.12 | 2.89 ± 0.11 | 3.01 ± 0.17 | 2.85 ± 0.17 | 2.87 ± 0.17 |
| **CD-ST** | 0.64 ± 0.10 | 0.84 ± 0.14 | 2.79 ± 0.19 | 2.66 ± 0.11 | 3.46 ± 0.16 | 3.52 ± 0.11 | 3.1 ± 0.19 | 3.08 ± 0.22 |
| **Mixed DEC + SU** | 2.39 ± 0.28 | 0.71 ± 0.21 | 1.57 ± 0.06 | 0.61 ± 0.03 | 1.53 ± 0.04 | 0.82 ± 0.03 | 2.19 ± 0.17 | 1.28 ± 0.10 |
| **SIRIEMA** | **0.30 ± 0.03** | **0.20 ± 0.01** | **0.13 ± 0.01** | **0.10 ± 0.01** | **0.02 ± 0.01** | **0.01 ± 0.001** | **0.48 ± 0.07** | **0.32 ± 0.08** |

Table 4: The DBS metric for each strategy across different sample sizes. Underlined scores are the best in each row, while bold ones highlight the results of SIRIEMA.

stable clustering and accurate information retrieval. In the BG metric, SIRIEMA was proficient for the entire dataset, indicating its ability to provide reliable and accurate clustering for extensive datasets. This performance affirms its effectiveness in scenarios demanding effective cluster separation and robustness. Our model also showed consistent superiority in the HAN metric for both 500 samples and the entire dataset. This underscores its capacity to generate stable and reliable clusters, proving its robustness and scalability and making it well-suited for various clustering tasks. In the OTA metric assessment, SIRIEMA emerged superior for the entire dataset, reaffirming its reliability and adaptability across different data volumes. Its consistent high OTA scores emphasize its suitability for maintaining clustering stability and agreement. Our proposed multimodal model displayed the lowest and most consistent DBS scores for 500 samples and the entire dataset, indicating its strong and consistent clustering patterns and making it a preferred choice for achieving clustering consistency and efficacy.

## 6 Conclusion

This research presented SIRIEMA, an innovative approach to customer segmentation by integrating structured and textual data. Our findings enhanced clustering stability in heterogeneous data contexts by developing a novel multimodal model building on BERT and a unique data fusion component coupled with a VAE. This is a significant advancement in addressing the challenge of clustering instability, which has previously plagued traditional methods, even when preprocessing and normalizing the data.

Our experiments employed real-world datasets such as Yelp, Melbourne Airbnb, PetFinder, and Women's clothing reviews, demonstrating the flexibility and robustness of the proposed model across diverse contexts. We assessed the model's effectiveness against five distinct stability measures and the DBS, revealing its superiority over conventional strategies, including the state-of-the-art Mixed DEC + SU method. By comparing four strategies, our results provide compelling evidence for our proposed model's soundness in achieving enhanced clustering stability, quality, and separation.

Our proposed multimodal clustering model demonstrated superior effectiveness, showcasing its utility in various applications involving intricate and diverse datasets where reliable clustering is paramount. Its adaptability across different sample sizes makes it a versatile tool for scenarios with varying data volumes, such as decision support systems, recommendation engines, and data-driven insights. While our model outperforms alternatives, further research can explore its applicability to different multimodal datasets and assess its limitations in specific contexts. Overall, our findings emphasize the importance of multimodal clustering in effectively handling complex data and contribute to advancing data analytics and clustering techniques, opening new avenues for data-driven decision-making and knowledge discovery.

Future work will further explore the optimization of the model and its applicability across diverse industries and contexts, including the potential integration of other types of unstructured data, such as images and audio. This exploration will include utilizing more Large Language Models (LLMs), such as Generative Pre-trained Transformer (GPT) and BERT variants, to enhance text processing and semantic understanding. By continuing to refine and expand this model, we aim to provide a versatile and powerful tool that can adapt to the rapidly evolving landscape of customer segmentation and targeted marketing.

## 7 Limitations

The SIRIEMA framework, while providing significant advancements in clustering stability for multimodal data, has limitations that need to be acknowledged. Integrating transformer-based models, data fusion techniques, and generative models contributes to a complex architecture. This complexity may lead to increased computational requirements, including higher memory consumption and longer processing times, posing challenges for real-time applications or environments with constrained computational resources. Also, its effectiveness is limited by how well models like BERT and GPT-3.5 match the target data and domain.

Another aspect is the challenge of data fusion. Effectively combining different data types, such as textual, categorical, and numerical, remains complex and may impact the clustering effectiveness if not executed optimally. Furthermore, while SIRIEMA shows promise, its ability to generalize across diverse datasets and domains has yet to be thoroughly validated, potentially limiting its effectiveness with varying data characteristics.

Scalability is another concern, as the framework's performance with massive datasets, particularly those with high-dimensional multimodal inputs, has yet to be extensively explored. There is a risk of overfitting on specific datasets, potentially harming generalization. Also, its robustness to data quality issues like missing values or noise still needs to be tested, affecting real-world applicability. Finally, hyperparameter tuning in the generative model requires time-consuming experimentation with potentially inconsistent results. Moreover, current stability measures may only partially reflect the framework's effectiveness in complex multimodal situations.

## 8 Ethical Considerations

SIRIEMA's usage of diverse datasets requires stringent data privacy and confidentiality measures, especially for personal and sensitive information. Compliance with data protection laws through anonymization, de-identification, and necessary consent from data subjects is critical. Addressing bias in datasets and model outputs is also essential. Proactive steps should be taken to identify and mitigate biases, ensuring fairness, particularly in multimodal environments. Transparency and explainability of complex models like transformers and generative are crucial. This includes clear documentation and the development of interpretative methods for model outputs. Additionally, SIRIEMA's environmental impact, due to high energy consumption and carbon emissions, necessitates improved computational efficiency and the use of green computing solutions.

### Code Availability

The source code used to generate all the results presented in this paper is available at https://anonymous.4open.science/r/SIRIEMA-6AD3/README.md.

## Acknowledgements

Omitted due to double-blind review.

## References

Özlem Akay and Güzin Yüksel. 2018. Clustering the mixed panel dataset using gower's distance and k-prototypes algorithms. *Communications in Statistics-Simulation and Computation*, 47(10):3031–3041.

Bitty Balducci and Detelina Marinova. 2018. Unstructured data in marketing. *Journal of the Academy of Marketing Science*, 46:557–590.

Nick Brooks. 2017. Women's e-commerce clothing reviews. https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews.

Michael J Brusco. 2004. Clustering binary data in the presence of masking variables. *Psychological Methods*, 9(4):510.

Roberto Mora Cortez, Ann Højbjerg Clarke, and Per Vagn Freytag. 2021. B2b market segmentation: A systematic review and research agenda. *Journal of Business Research*, 126:415–428.

Yelp Dataset. 2014. Yelp dataset. http://www.yelp.com/dataset_challenge.

Sara Dolnicar and Katie Lazarevski. 2009. Methodological reasons for the theory/practice divide in market segmentation. *Journal of marketing management*, 25(3-4):357–373.

Jorge E Fresneda, Thomas A Burnham, and Chelsey H Hill. 2021. Structural topic modelling segmentation: a segmentation method combining latent content and customer context. *Journal of Marketing Management*, 37(7-8):792–812.

Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864.

Ken Gu and Akshay Budhkar. 2021. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, Mexico City, Mexico. Association for Computational Linguistics.

Homa Hajibaba, Bettina Grün, and Sara Dolnicar. 2020. Improving the stability of market segmentation analysis. *International Journal of Contemporary Hospitality Management*, 32(4):1393–1411.

GM Harshvardhan, Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray. 2020. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38:100285.

Zhenfeng He and Chunyan Yu. 2019. Clustering stability-based evolutionary k-means. *Soft Computing*, 23(1):305–321.

Kaggle and PetFinder.my. 2019. Petfinder.my adoption prediction. https://www.kaggle.com/c/petfinder-adoption-prediction/data/.

Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. 2022. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141.

Snehalika Lall, Debajyoti Sinha, Abhik Ghosh, Debarka Sengupta, and Sanghamitra Bandyopadhyay. 2021. Stable feature selection using copula based mutual information. *Pattern Recognition*, 112:107697.

Yonggu Lee, Chulwung Park, and Shinjin Kang. 2022. Deep embedded clustering framework for mixed data. *IEEE Access*, 11:33–40.

Fine F Leung, Flora F Gu, Yiwei Li, Jonathan Z Zhang, and Robert W Palmatier. 2022. Influencer marketing effectiveness. *Journal of Marketing*, 86(6):93–115.

Kart-Leong Lim, Xudong Jiang, and Chenyu Yi. 2020. Deep clustering with variational autoencoder. *IEEE Signal Processing Letters*, 27:231–235.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open*.

Tianmou Liu, Han Yu, and Rachael Hageman Blair. 2022. Stability estimation for unsupervised clustering: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(6):e1575.

Azam Peyvandipour, Adib Shafi, Nafiseh Saberian, and Sorin Draghici. 2020. Identification of cell types from single cell data using stable clustering. *Scientific reports*, 10(1):12349.

D Venkatavara Prasad, Sathya Madhusudanan, and Suresh Jaganathan. 2015. uclust-a new algorithm for clustering unstructured data. *ARPN Journal of Engineering and Applied Sciences*, 10(5):2108–2117.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.

Susmita Ray. 2019. A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 35–39. IEEE.

Saswata Sahoo and Souradip Chakraborty. 2020. Learning representation for mixed data types with a nonlinear deep encoder-decoder framework. *arXiv preprint arXiv:2009.09634*.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2021. Scale efficiently: Insights from pretraining and fine-tuning transformers. *arXiv preprint arXiv:2109.10686*.

Rajan Varadarajan. 2020. Customer information resources advantage, marketing strategy and business performance: A market resources based view. *Industrial Marketing Management*, 89:89–97.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Nhi NY Vo, Shaowu Liu, Xitong Li, and Guandong Xu. 2021. Leveraging unstructured call log data for customer churn prediction. *Knowledge-Based Systems*, 212:106586.

Ulrike Von Luxburg et al. 2010. Clustering stability: an overview. *Foundations and Trends® in Machine Learning*, 2(3):235–274.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR.

Tyler Xie. 2019. Melbourne airbnb open data. https://www.kaggle.com/tylerx/melbourne-airbnb-open-data/version/7.

Lin Yang, Wentao Fan, and Nizar Bouguila. 2020. Clustering analysis via deep generative models with mixture models. *IEEE Transactions on Neural Networks and Learning Systems*, 33(1):340–350.

Sheng Zhou, Hongjia Xu, Zhuonan Zheng, Jiawei Chen, Jiajun Bu, Jia Wu, Xin Wang, Wenwu Zhu, Martin Ester, et al. 2022. A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. *arXiv preprint arXiv:2206.07579*.

# A Appendix

In this Appendix Section, we present the hyperparameters used for the best model, validation results for all datasets and explain the approach we adopted to determine the optimal number of clusters for each dataset.

## A.1 Hyperparameters

For the purpose of enhancing the reproducibility of our research, we provide Table 5. This table details the hyperparameters employed in the top-performing model across various experiments. We employed Grid Search[13] to methodically assess a range of hyperparameters, ensuring the selection of the most effective combination for our model.

## A.2 Stability analysis across validation datasets

Table 6 showcases the performance results across all validation dataset obtained using ARI, AMIS,

[13] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

| Hyperparameters | Value |
|---|---|
| Batch size | 768 |
| Maximum token length | 768 |
| Optimizer | Adam |
| Weight decay | 0.01 |
| Adam $\epsilon$ | 1e-8 |
| Adam $\beta$ | [0.7, 0.9] |
| Learning rate schedule | 1e-8 |
| Maximum learning rate | 4e-5 |
| Minimum learning rate | 1e-5 |
| # Steps | 2000 |

Table 5: Hyperparameters used in the fine-tuning process.

BG, HAN, and OTA metrics. The evaluation strategies employed are delineated as follows: **A**: Structured Approach; **B**: Textual Approach; **C**: CD-ST; **D**: Mixed DEC + SU; and **E**: SIRIEMA. We present the results as the mean, accompanied by a 95% confidence interval; underlined values highlight the best outcomes, whereas bold values denote results from our proposed multimodal model.

The SIRIEMA model exhibited superior efficacy in multiple metrics in the Yelp dataset. In the ARI metric, it scored $0.94 \pm 0.01$ for 500 samples and $1.0 \pm 0.0$ for the entire dataset. In AMIS, it registered $0.9 \pm 0.02$ for 500 samples and $0.99 \pm 0.1$ for the entire dataset, indicating robust mutual information alignment. For BG, the scores were $0.99 \pm 0.0$ for 500 samples and $1.0 \pm 0.0$ for the entire dataset, showing stable clustering. In HAN, SIRIEMA scored $0.99 \pm 0.01$ for 500 samples and $1.0 \pm 0.0$ for the entire dataset, outperforming other models. Lastly, in OTA, it achieved $0.86 \pm 0.04$ and $0.89 \pm 0.01$, demonstrating effective handling of multimodal data. In the Airbnb validation dataset, our model achieved an ARI of $0.91 \pm 0.01$ for 500 samples and demonstrated perfect stability with $1.0 \pm 0.0$ for the entire dataset. It scored $0.88 \pm 0.01$ in AMIS for 500 samples, reaching $1.0 \pm 0.0$ for the full dataset. For BG, the model registered $0.98 \pm 0.0$ for 500 samples and $0.97 \pm 0.01$ overall. In HAN, it attained $0.96 \pm 0.01$ for 500 samples and $1.0 \pm 0.0$ for the entire dataset. Lastly, the model showed improvement in the OTA metric, scoring $0.78 \pm 0.01$ for 500 samples and $0.91 \pm 0.01$ for the full dataset. In the Pertinder.my dataset, our proposal achieved $1.0 \pm 0.0$ for 500 samples and $0.79 \pm 0.03$ for the entire dataset in the ARI metric. In AMIS, it scored $0.98 \pm 0.01$ for 500 samples and $0.90 \pm 0.01$ for the entire dataset. The model showed strong performance in the BG metric, reaching $0.98 \pm 0.01$, and in HAN, with $0.84 \pm 0.05$ for 500 samples, maintaining consistent effectiveness for the entire dataset. In the OTA evaluation, it scored $0.44 \pm 0.13$ for 500 samples and $0.5 \pm 0.13$ for the entire dataset, reflecting its adaptability and areas for improvement. In the Clothing dataset, our model achieved perfect scores in ARI and AMIS, with $1.0 \pm 0.0$ and $1.0 \pm 0.01$, respectively, for both 500 samples and the entire dataset. In the BG metric, it ranked third with $0.89 \pm 0.01$ for 500 samples but achieved the top score of $1.0 \pm 0.01$ for the entire dataset. The model demonstrated robust clustering in the HAN metric, achieving $1.0 \pm 0.0$ for both sample sizes. Finally, in OTA, it scored $0.68 \pm 0.23$ for 500 samples and $0.89 \pm 0.01$ for the entire dataset.

### A.3 Clustering analysis

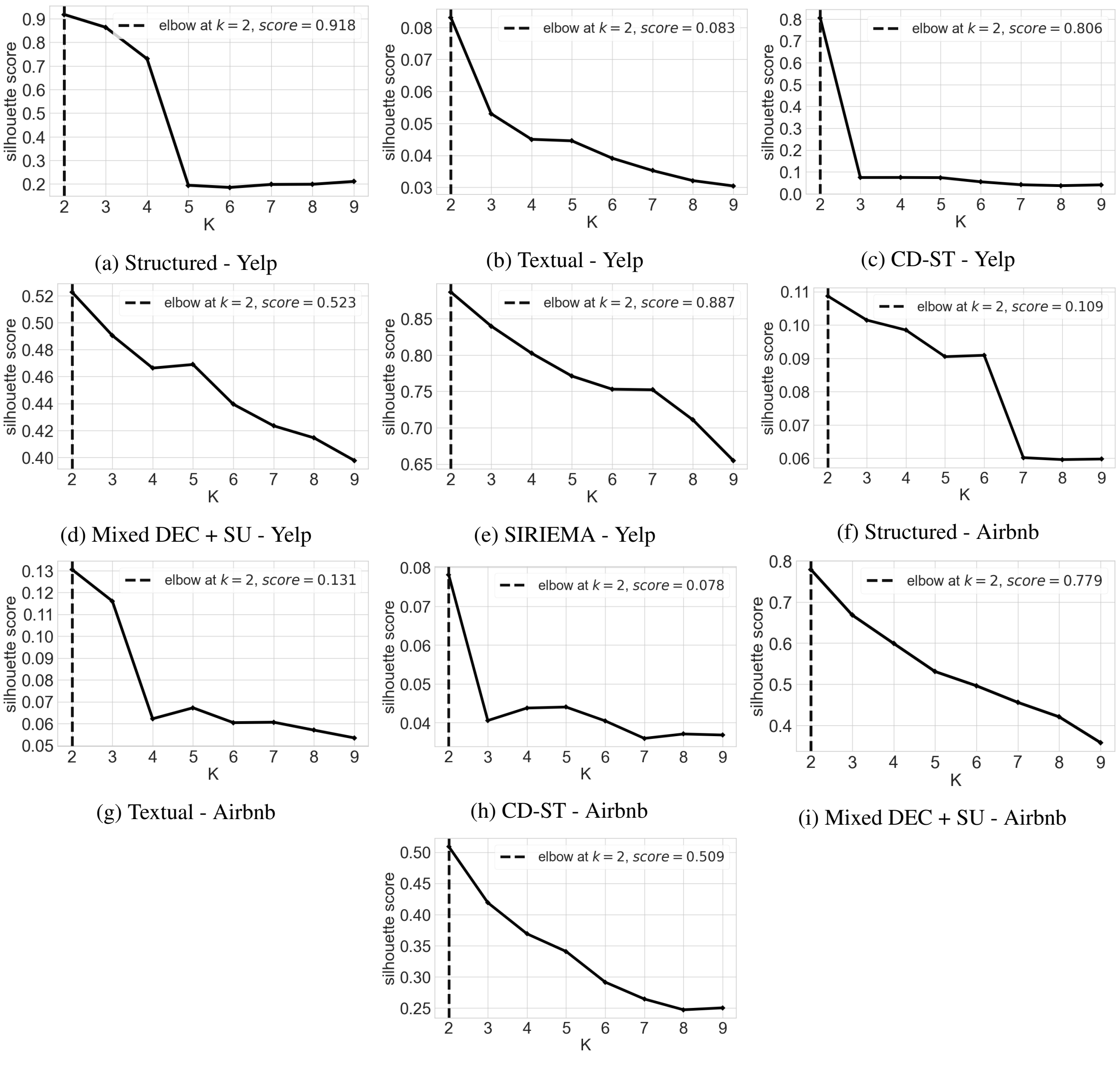
We determined the optimal cluster count for each strategy, using silhouette scores to analyze the clustering of Yelp and Airbnb datasets with $k$ ranging from 2 to 9. Silhouette score, ranging from -1 to 1, measures an object's fit to its cluster versus others, with higher values indicating better clustering.

Maximizing intra-cluster similarity and inter-cluster dissimilarity, the two-cluster configuration consistently showed the highest scores, indicating robust and distinct clustering. Figure 2 displays Silhouette scores for Yelp and Airbnb datasets, reflecting cluster quality and distinctiveness.

In contrast, we determined the optimal number of clusters for PetFinder.my and Clothings through their labels. The PetFinder.my dataset has *same day*, *1-7 days*, *8-30 days*, *31-90 days*, and *more than 100 days*. Women's E-Commerce clothing reviews dataset has *Not Recommended* and *Recommended labels*.

| | | **Yelp** | | **Airbnb** | | **PetFinder.my** | | **Clothing** | |
|---|---|---|---|---|---|---|---|---|---|
| | | **500 s.** | **Entire Dataset** | **500 s.** | **Entire Dataset** | **500 s.** | **Entire Dataset** | **500 s.** | **Entire Dataset** |
| ARI | **A** | 0.61 ± 0.07 | 0.53 ± 0.06 | 0.55 ± 0.05 | 0.69 ± 0.05 | 0.55 ± 0.02 | 0.68 ± 0.08 | 0.47 ± 0.04 | 0.92 ± 0.08 |
| | **B** | 0.84 ± 0.01 | 0.96 ± 0.00 | 0.92 ± 0.00 | 0.95 ± 0.01 | 0.47 ± 0.02 | 0.58 ± 0.02 | 0.58 ± 0.03 | 0.58 ± 0.14 |
| | **C** | 0.53 ± 0.05 | 0.40 ± 0.07 | 0.73 ± 0.08 | 0.89 ± 0.02 | 0.42 ± 0.01 | 0.52 ± 0.04 | 0.67 ± 0.08 | 0.56 ± 0.08 |
| | **D** | 0.25 ± 0.02 | 0.95 ± 0.01 | 0.51 ± 0.02 | 0.74 ± 0.02 | 0.53 ± 0.03 | 0.67 ± 0.04 | 0.52 ± 0.05 | 0.78 ± 0.02 |
| | **E** | **0.94 ± 0.01** | **1.00 ± 0.00** | **0.91 ± 0.01** | **1.00 ± 0.00** | **1.00 ± 0.0** | **0.79 ± 0.02** | **1.00 ± 0.00** | **1.00 ± 0.00** |
| AMIS | **A** | 0.59 ± 0.06 | 0.49 ± 0.06 | 0.51 ± 0.04 | 0.63 ± 0.04 | 0.57 ± 0.02 | 0.68 ± 0.06 | 0.45 ± 0.04 | 0.89 ± 0.07 |
| | **B** | 0.79 ± 0.01 | 0.93 ± 0.01 | 0.88 ± 0.01 | 0.92 ± 0.03 | 0.56 ± 0.02 | 0.65 ± 0.02 | 0.56 ± 0.04 | 0.55 ± 0.11 |
| | **C** | 0.50 ± 0.04 | 0.37 ± 0.06 | 0.68 ± 0.08 | 0.83 ± 0.02 | 0.50 ± 0.01 | 0.59 ± 0.03 | 0.63 ± 0.07 | 0.53 ± 0.07 |
| | **D** | 0.21 ± 0.02 | 0.91 ± 0.01 | 0.59 ± 0.02 | 0.69 ± 0.01 | 0.59 ± 0.02 | 0.69 ± 0.03 | 0.45 ± 0.05 | 0.74 ± 0.01 |
| | **E** | **0.90 ± 0.02** | **0.99 ± 0.00** | **0.88 ± 0.01** | **1.00 ± 0.00** | **0.98 ± 0.01** | **0.90 ± 0.01** | **1.00 ± 0.00** | **1.00 ± 0.00** |
| BG | **A** | 0.97 ± 0.01 | 0.99 ± 0.01 | 0.92 ± 0.01 | 0.94 ± 0.01 | 0.76 ± 0.01 | 0.81 ± 0.07 | 0.76 ± 0.02 | 0.77 ± 0.04 |
| | **B** | 0.97 ± 0.01 | 0.99 ± 0.00 | 0.98 ± 0.00 | 0.99 ± 0.00 | 0.69 ± 0.01 | 0.79 ± 0.05 | 0.92 ± 0.01 | 0.95 ± 0.03 |
| | **C** | 0.89 ± 0.02 | 0.95 ± 0.04 | 0.96 ± 0.01 | 0.97 ± 0.00 | 0.66 ± 0.01 | 0.73 ± 0.02 | 0.94 ± 0.01 | 0.95 ± 0.03 |
| | **D** | 0.72 ± 0.02 | 0.99 ± 0.00 | 0.73 ± 0.03 | 0.93 ± 0.02 | 0.70 ± 0.02 | 0.81 ± 0.10 | 0.83 ± 0.02 | 0.87 ± 0.02 |
| | **E** | **0.99 ± 0.00** | **1.00 ± 0.00** | **0.98 ± 0.00** | **0.97 ± 0.01** | **0.98 ± 0.01** | **0.79 ± 0.02** | **0.89 ± 0.01** | **1.00 ± 0.01** |
| HAN | **A** | 0.83 ± 0.07 | 0.99 ± 0.00 | 0.83 ± 0.03 | 0.84 ± 0.02 | 0.67 ± 0.01 | 0.77 ± 0.03 | 0.77 ± 0.03 | 0.91 ± 0.05 |
| | **B** | 0.92 ± 0.01 | 0.98 ± 0.00 | 0.96 ± 0.00 | 0.98 ± 0.01 | 0.56 ± 0.02 | 0.68 ± 0.07 | 0.82 ± 0.02 | 0.88 ± 0.01 |
| | **C** | 0.77 ± 0.06 | 0.78 ± 0.04 | 0.91 ± 0.02 | 0.95 ± 0.01 | 0.51 ± 0.02 | 0.59 ± 0.02 | 0.87 ± 0.02 | 0.86 ± 0.04 |
| | **D** | 0.55 ± 0.02 | 0.97 ± 0.00 | 0.59 ± 0.03 | 0.85 ± 0.01 | 0.56 ± 0.03 | 0.70 ± 0.04 | 0.75 ± 0.02 | 0.86 ± 0.03 |
| | **E** | **0.99 ± 0.01** | **1.00 ± 0.00** | **0.96 ± 0.01** | **1.00 ± 0.00** | **0.84 ± 0.05** | **0.78 ± 0.04** | **1.00 ± 0.00** | **1.00 ± 0.00** |
| OTA | **A** | 0.50 ± 0.01 | 0.50 ± 0.00 | 0.68 ± 0.03 | 0.70 ± 0.01 | 0.26 ± 0.05 | 0.10 ± 0.07 | 0.49 ± 0.07 | 0.34 ± 0.13 |
| | **B** | 0.73 ± 0.00 | 0.73 ± 0.00 | 0.78 ± 0.01 | 0.78 ± 0.01 | 0.43 ± 0.02 | 0.49 ± 0.00 | 0.53 ± 0.06 | 0.62 ± 0.21 |
| | **C** | 0.66 ± 0.04 | 0.63 ± 0.01 | 0.74 ± 0.01 | 0.73 ± 0.01 | 0.28 ± 0.06 | 0.42 ± 0.26 | 0.61 ± 0.06 | 0.68 ± 0.01 |
| | **D** | 0.52 ± 0.02 | 0.69 ± 0.01 | 0.26 ± 0.03 | 0.08 ± 0.03 | 0.28 ± 0.04 | 0.25 ± 0.03 | 0.61 ± 0.03 | 0.56 ± 0.12 |
| | **E** | **0.86 ± 0.04** | **0.89 ± 0.01** | **0.78 ± 0.01** | **0.91 ± 0.00** | **0.44 ± 0.12** | **0.50 ± 0.10** | **0.68 ± 0.23** | **0.89 ± 0.01** |

Table 6: Comparing the stability metrics of various strategies across different sample sizes on four distinct datasets on the validation dataset. Underlined scores are the best in each row, while bold ones highlight the results of SIRIEMA. The strategies are as follows: **A**: Structured; **B**: Textual; **C**: CD-ST; **D**: Mixed DEC + SU; and **E**: SIRIEMA.



(a) Structured - Yelp

(b) Textual - Yelp

(c) CD-ST - Yelp

(d) Mixed DEC + SU - Yelp

(e) SIRIEMA - Yelp

(f) Structured - Airbnb

(g) Textual - Airbnb

(h) CD-ST - Airbnb

(i) Mixed DEC + SU - Airbnb

(j) SIRIEMA - Airbnb

Figure 2: Silhouette score for Yelp and Airbnb datasets.