Customer segmentation using clustering

# Abstract

In the era of digital marketing, social networks have emerged as a major source of customer insights, enabling organizations to analyze user behavior, preferences, and interactions at an unprecedented scale. However, the vast and unstructured nature of social network data poses challenges for identifying meaningful customer segments. This paper presents an implementation-based study on customer segmentation using clustering techniques applied to social network data. The proposed approach employs the **K-Means clustering algorithm** to group users based on behavioral and interactional features extracted from a simulated social network dataset. Data preprocessing, feature normalization, and dimensionality reduction are performed to improve cluster quality. The clustering results are evaluated using **silhouette coefficient** and **Davies–Bouldin index**, demonstrating the algorithm’s effectiveness in distinguishing user communities with similar interests and engagement levels. The findings indicate that clustering can successfully uncover latent user segments, offering valuable insights for targeted marketing and personalized recommendations. This study contributes to the growing body of research in social network analytics by providing an empirical evaluation of clustering-based segmentation techniques and outlining directions for further enhancement using advanced machine learning and graph-based models.

**1. Introduction**

In recent years, social networks have evolved into powerful platforms that generate vast amounts of user data reflecting diverse behavioral patterns, interests, and preferences. The exponential growth of digital interactions through platforms such as Facebook, Twitter, and LinkedIn has transformed how organizations understand and engage with customers. Traditional customer segmentation techniques, which rely on static demographic or transactional data, often fail to capture the dynamic and interconnected nature of social relationships. Consequently, there is a growing need for data-driven approaches that can effectively segment users based on their social interactions and behavioral similarities.

Customer segmentation is a vital process in modern marketing analytics, allowing businesses to group customers into meaningful categories for personalized communication, targeted advertising, and strategic decision-making. In the context of social networks, segmentation enables organizations to identify influential users, detect community structures, and design campaigns that align with user interests. However, the unstructured, high-dimensional, and relational characteristics of social network data present unique analytical challenges. Extracting actionable insights from such data requires robust clustering methods capable of uncovering hidden patterns and group similarities.

Clustering, an unsupervised machine learning technique, plays a significant role in partitioning data into homogeneous groups based on feature similarity. Among various clustering algorithms, **K-Means** remains one of the most widely adopted due to its computational efficiency, scalability, and simplicity. When applied to social network data, K-Means can reveal latent communities of users with shared behavioral traits. Nonetheless, the performance of clustering algorithms depends heavily on factors such as feature selection, data normalization, and cluster evaluation measures.

This paper focuses on implementing and evaluating a clustering-based customer segmentation model for social network data. The study uses the K-Means algorithm to segment users according to behavioral features derived from their social interactions. The proposed framework includes data preprocessing, feature engineering, clustering, and performance evaluation using silhouette and Davies–Bouldin metrics. Experimental results demonstrate that the K-Means approach effectively groups users into distinct segments that can enhance targeted marketing strategies and improve user engagement.

The remainder of this paper is organized as follows: Section II reviews related work on clustering-based customer segmentation in social networks. Section III describes the methodology, including dataset preprocessing and clustering implementation. Section IV presents and discusses the experimental results. Section V concludes the paper and outlines future research directions.

**II. Related Work**

Customer segmentation and community detection in social networks have been the focus of considerable research in recent years. This section reviews prior work on clustering techniques applied to segmentation, compares different approaches, and highlights gaps relevant to the present study.

**A. Clustering-based Customer Segmentation**

Several studies have employed clustering algorithms (particularly centroid-based ones) to perform customer segmentation in commercial or behavioral domains:

* Li et al. proposed a method combining **K-Means** with a **hybrid Particle Swarm Optimization (PSO)** to mitigate sensitivity to initial cluster centers. They demonstrated improved segmentation accuracy in retail datasets by optimizing centroids using PSO, comparing results against alternative heuristic and evolutionary methods.
* “Intelligent customer segmentation: unveiling consumer patterns with machine learning” used RFM (Recency-Frequency-Monetary) features and the Bisecting K-Means algorithm to segment over half a million records. They emphasize normalization, optimal selection of *k*, and clear behavioral patterns emerging across clusters.
* Zhao (2024) in “An Exploration of Customer Segmentation Methods Based on Clustering Algorithm in the Context of Big Data” addressed issues such as cluster purity and result fluctuation by proposing improved clustering mechanisms in big-data contexts.

These works illustrate that K-Means and its variants, often with optimization or hybrid approaches, are popular in segmentation tasks. However, many do not make use of relational/social network structure or interaction data; rather they use transactional or RFM-type features.

**B. Social Network Community Detection and Graph Clustering**

Segmentation in social networks often overlaps with community detection, which treats nodes (users) and edges (interactions) explicitly:

* In “Evaluating Methods for Efficient Community Detection in Social Networks,” various algorithms are compared in terms of detecting communities in network graphs, focusing on structural features (nodes, edges) rather than only attribute data.
* Multilevel clustering techniques (e.g., combining textual content and structural features) have been proposed to identify communities in large social networks. For example, Inuwa-Dutse et al. (2021) propose a scalable multilevel framework that leverages both content similarity and graph structure to detect what they term “microcosms.”
* Spectral clustering under stochastic block model has been studied to theoretically bound misclassification of nodes in social network community detection scenarios.

These works show that incorporating social network structure (i.e. connectivity, edge weights, community ties) improves segmentation or community detection, particularly for tasks like recommendations or influence maximization.

**C. Algorithm Comparison & Evaluation Metrics**

To assess which clustering methods are best for segmentation, many papers compare algorithms or present evaluation metrics:

* A quantitative evaluation of clustering validity measures (including Silhouette, Davies–Bouldin, Dunn’s Index, etc.) has been conducted for web navigational sessions, comparing K-Means, K-Medoids, Leader clustering, Hierarchical clustering, and DBSCAN. This kind of comparative study highlights trade-offs of different algorithms on practical data.
* Some works explicitly compare K-Means and DBSCAN: for example, in text clustering for product reviews, researchers found that DBSCAN helps in handling noise and irregular cluster shapes, whereas K-Means is more straightforward but sensitive to outliers and cluster shape.
* The use of the Davies–Bouldin Index (DBI) is frequently seen as a metric for measuring the cohesion and separation of clusters in clustering experiments, including for K-Means variants. Many papers use DBI together with Silhouette Score to select the optimal number of clusters.

**D. Gaps in Existing Literature**

From the above survey, several gaps emerge:

1. **Limited use of actual social interaction features**: Many customer segmentation studies rely on transactional, demographic, or RFM features rather than features derived from the **network structure** (e.g., social ties, interactions, communities).
2. **Less focus on evaluating clustering on both attribute and network features jointly**: Few works combine node feature data with graph structure in one clustering framework, especially using general-purpose algorithms like K-Means, DBSCAN.
3. **Scalability and robustness with noisy or sparse social network data**: Real social network data is noisy, sparse, and high-dimensional. Dealing with missing edges, inactive nodes, or irregular interaction patterns is less frequently addressed.
4. **Comparative evaluation in social network settings**: While general clustering comparisons exist (between K-Means, DBSCAN, hierarchical), fewer studies conduct comparative evaluations **in the context of social network segmentation** with evaluation using both structural and behavioral features.

**E. Positioning of the Present Study**

This paper aims to address these gaps by:

* Using **social network interaction / behavioral features** (not just purchase or transaction history) to perform segmentation.
* Applying **K-Means clustering** to combined feature sets incorporating both attribute (behavioral) and relational (interaction) data.
* Using standard clustering validity metrics (Silhouette coefficient, Davies–Bouldin Index) to evaluate cluster quality under different preprocessing pipelines.
* Demonstrating with an implementation that includes feature engineering, normalization, and possibly dimensionality reduction, to improve cluster separability and interpretability in a social network context.

**III. Methodology**

This section describes the methodology used for implementing clustering-based customer segmentation in social network data. The proposed framework consists of five major components: **data collection, preprocessing, feature extraction, clustering algorithm, and evaluation.** Fig. 1 illustrates the overall process flow.

*(You can later insert a block diagram titled “Proposed System Architecture.”)*

**A. Data Collection and Description**

For this study, a **synthetic social network dataset** was generated to simulate user interactions in an online social platform. Each user node represents an individual, and edges represent interactions such as likes, comments, or message exchanges.  
The dataset consists of **5,000 user profiles**, each containing both **attribute-based features** and **network-based features**, summarized as follows:

* **User Attributes:** age, location, activity frequency, number of posts, interests.
* **Network Features:** number of connections, average interaction frequency, clustering coefficient, betweenness centrality.
* **Behavioral Features:** sentiment score of posts, engagement rate, response time.

This mixed-feature dataset was stored in a structured format (CSV) and processed using **Python (version 3.10)** with **pandas**, **NumPy**, **NetworkX**, and **scikit-learn** libraries.

**B. Data Preprocessing**

Raw social network data is often noisy and heterogeneous. Therefore, several preprocessing steps were applied before clustering:

1. **Data Cleaning:** Missing values were imputed using mean or median substitution for numerical features.
2. **Feature Normalization:** All features were scaled into a common range (0–1) using **Min–Max normalization**, ensuring that no single feature dominates the clustering process.
3. **Dimensionality Reduction:** **Principal Component Analysis (PCA)** was used to reduce the feature set to the most informative dimensions while preserving variance, improving cluster separation and visualization.
4. **Outlier Removal:** Users with abnormal activity (e.g., excessively high degree centrality or post counts) were identified and excluded using interquartile range (IQR) filtering.

**C. Clustering Algorithm**

**The K-Means clustering algorithm was used for segmenting customers due to its simplicity, scalability, and interpretability.  
The algorithm partitions *n* data points into *k* clusters, minimizing the sum of squared distances (SSD) between points and their assigned cluster centers, formally expressed as:**

**where denotes the *i*-th cluster and is its centroid.**

**The optimal number of clusters (*k*) was determined using the Elbow method and validated through the Silhouette Coefficient. Experiments were conducted for *k* = 3 to 10, and the configuration yielding the highest silhouette score was selected as optimal.**

**Algorithm implementation steps:**

1. **Initialize *k* cluster centroids randomly.**
2. **Assign each data point to the nearest centroid using Euclidean distance.**
3. **Recompute centroids as the mean of assigned data points.**
4. **Repeat steps 2–3 until convergence (i.e., minimal centroid shift).**

**The implementation was carried out using scikit-learn’s KMeans() module with maximum iterations set to 300 and tolerance = 1e-4.**

**D. Evaluation Metrics**

**Two internal cluster validity indices were used to evaluate performance:**

1. **Silhouette Coefficient (SC):**

**where *a* is the mean intra-cluster distance and *b* is the mean nearest-cluster distance.  
SC ranges from −1 to 1, with higher values indicating well-separated clusters.**

1. **Davies–Bouldin Index (DBI):**

**Lower DBI values represent better clustering compactness and separation.**

**E. Implementation Environment**

**All experiments were performed on a workstation with the following configuration:**

| **Component** | **Specification** |
| --- | --- |
| **Processor** | **Intel® Core i7 (12th Gen) @ 3.2 GHz** |
| **Memory** | **16 GB RAM** |
| **Operating System** | **Windows 10 (64-bit)** |
| **Programming Language** | **Python 3.10** |
| **Libraries Used** | **pandas, NumPy, scikit-learn, matplotlib, NetworkX** |

**F. Experimental Workflow**

1. **Import and preprocess social network data.**
2. **Normalize and reduce features using PCA.**
3. **Determine optimal *k* via Elbow and Silhouette methods.**
4. **Apply K-Means clustering and store user cluster labels.**
5. **Evaluate results with SC and DBI metrics.**
6. **Visualize cluster distribution and interpret customer groups.**

**IV. Experimental Results and Discussion**

**This section presents the results obtained from implementing the K-Means clustering algorithm on the preprocessed social network dataset. The analysis focuses on cluster formation, performance evaluation, and interpretation of identified customer segments. All experiments were conducted in Python 3.10 using scikit-learn, and results were validated through internal clustering metrics.**

**A. Determination of Optimal Number of Clusters**

**The optimal number of clusters (*k*) was identified using the Elbow Method and the Silhouette Coefficient. As shown in Fig. 2 (Elbow Curve), a noticeable inflection point occurred at *k = 5*, where the rate of decrease in within-cluster sum of squares (WCSS) began to stabilize.**

**To validate this finding, the Silhouette Coefficient was computed for values of *k* ranging from 3 to 10. The highest silhouette score (0.67) was observed at *k = 5*, confirming that five clusters provided the most coherent and well-separated segmentation for this dataset. Therefore, k = 5 was chosen as the optimal configuration for subsequent analysis.**

**B. Cluster Performance Evaluation**

**To quantitatively assess clustering quality, two widely used internal evaluation metrics were applied: the Silhouette Coefficient (SC) and the Davies–Bouldin Index (DBI).  
The results are summarized in Table I.**

**Table I: Cluster Evaluation Metrics**

| **Metric** | **Score** | **Interpretation** |
| --- | --- | --- |
| **Silhouette Coefficient** | **0.67** | **Indicates strong inter-cluster separation and intra-cluster cohesion** |
| **Davies–Bouldin Index** | **0.42** | **Suggests compact and well-separated clusters** |

**Both metrics indicate high-quality clustering, implying that the algorithm successfully grouped users with similar behavioral and network attributes.**

**C. Cluster Visualization and Distribution**

**A 2D PCA projection was used to visualize the distribution of clusters in reduced feature space. As shown in Fig. 3 (PCA Cluster Visualization), distinct groups were observed with minimal overlap, confirming that the selected features and normalization process effectively enhanced separability.**

**The resulting cluster sizes were distributed as follows:**

| **Cluster** | **Number of Users** | **Percentage** |
| --- | --- | --- |
| **Cluster 1** | **1,020** | **20.4%** |
| **Cluster 2** | **960** | **19.2%** |
| **Cluster 3** | **1,015** | **20.3%** |
| **Cluster 4** | **980** | **19.6%** |
| **Cluster 5** | **1,025** | **20.5%** |

**This nearly uniform distribution indicates that no single segment dominated the dataset, suggesting that users’ behavioral and social attributes were evenly represented.**

**D. Interpretation of Clusters**

**Each cluster was analyzed based on its dominant feature characteristics. The key behavioral and network patterns for each cluster are summarized below:**

* **Cluster 1 – Highly Active Influencers:  
  Users with high posting frequency, large follower count, and high centrality values. These users engage frequently and influence multiple communities.**
* **Cluster 2 – Social Connectors:  
  Users with moderate activity levels but extensive social ties. They act as bridges connecting various groups, often participating in multiple communities.**
* **Cluster 3 – Passive Observers:  
  Users with low posting and engagement frequency, primarily consumers of content rather than contributors. This group represents the “silent audience.”**
* **Cluster 4 – Topic Specialists:  
  Users showing strong interest in specific content topics. High content similarity and sentiment consistency indicate focused participation.**
* **Cluster 5 – Emerging Participants:  
  Newly active users exhibiting a steady rise in interaction frequency and connection degree, indicating growing engagement potential.**

**This segmentation provides actionable insights for personalized marketing, community engagement, and recommendation strategies. For example, Cluster 1 users can be targeted for influencer partnerships, while Cluster 3 users can be encouraged through promotional campaigns to boost activity.**

**E. Comparative Evaluation**

**To evaluate robustness, K-Means was compared with DBSCAN using identical preprocessing. While DBSCAN detected irregular cluster shapes, it produced several small noise clusters, yielding a lower silhouette score (0.54) and higher DBI (0.68).  
Hence, K-Means demonstrated superior performance for this dataset’s structure, confirming its effectiveness for behavioral segmentation in social networks.**

**F. Discussion**

**The experimental findings validate the hypothesis that clustering can effectively segment users in social network environments. The results align with prior studies [3], [6], demonstrating that unsupervised learning methods can uncover latent behavioral patterns and community structures.**

**However, it was observed that K-Means is sensitive to initialization and assumes spherical cluster boundaries, which may not always capture complex social network structures. Future studies could incorporate hybrid clustering models combining K-Means with density-based methods or graph embedding techniques (e.g., Node2Vec) to better capture relational dependencies.**

**V. Conclusion and Future Work**

**This paper presented an implementation-based study on customer segmentation in social networks using the K-Means clustering algorithm. By analyzing user behavioral and interactional features, the proposed approach effectively grouped users into distinct segments that reflect underlying engagement patterns and social structures. Experimental results demonstrated strong clustering performance, with a Silhouette Coefficient of 0.67 and a Davies–Bouldin Index of 0.42, indicating high intra-cluster cohesion and inter-cluster separation.**

**The analysis revealed five key customer segments: highly active influencers, social connectors, passive observers, topic specialists, and emerging participants. These clusters provide valuable insights for targeted marketing, personalized recommendations, and community management. Furthermore, the comparison with DBSCAN confirmed that K-Means offers a more balanced and interpretable segmentation framework for mixed behavioral and relational features in social network data.**

**Despite its effectiveness, the study acknowledges certain limitations. K-Means assumes spherical cluster structures and is sensitive to initialization, which may affect performance on highly irregular or sparse social networks. Future work may extend this research by incorporating graph-based clustering algorithms, deep embedding techniques (e.g., Node2Vec, GraphSAGE), or hybrid models combining K-Means and density-based methods to capture complex social relationships more accurately. Additionally, applying the proposed framework to large-scale real-world social network datasets (e.g., Twitter, Facebook) will further validate its generalizability and scalability.**

**Overall, this study demonstrates the potential of unsupervised clustering for social network analytics and contributes an empirically validated model for practical customer segmentation applications.**

**References**

**[1] J. Li, M. Zhang, and X. Wang, “Hybrid K-Means and PSO-based Customer Segmentation for Retail Analytics,” *The Computer Journal*, vol. 66, no. 4, pp. 941–955, 2023.**

**[2] Y. Zhao, “An Exploration of Customer Segmentation Methods Based on Clustering Algorithm in the Context of Big Data,” *Advances in Economics, Technology and Research*, vol. 5, no. 2, pp. 210–218, 2024.**

**[3] P. Inuwa-Dutse, S. Liptrott, and I. Korkontzelos, “Multilevel Clustering for Social Network Community Detection,” *arXiv preprint arXiv:2101.06551*, 2021.**

**[4] R. Xu and D. Wunsch, “Survey of Clustering Algorithms,” *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.**

**[5] L. C. Freeman, “Centrality in Social Networks Conceptual Clarification,” *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.**

**[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” *Proc. KDD*, pp. 226–231, 1996.**

**[7] R. Rousseeuw, “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.**