PROCEEDING OF RESEARCH AND CIVIL SOCIETY DESEMINATION

ISSN 3024-8426, Volume 3, No. 1, Pages 310-316

DOI: https://doi.org/10.37476/presed.v3i1.131.

Application of HDBSCAN Algorithm for WhatsApp Lead Segmentation in Sales Strategy Optimization in CV. Multi Engineering Partners

Wanda Cahyani[1*]; Raden Wirawan[2]; Nurkhalik Wahdanial Asbara[3]
[1,2,3]Institut Teknologi dan Bisnis Nobel Indonesia
*Correspondence: wandacahyanileo@gmail.com

Abstract: The development of the use of WhatsApp as the main communication channel in the pre-sales process for SMEs produces large, complex, and heterogeneous lead data, thus requiring an analytical approach to improve the effectiveness of lead management. This study applied the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm to segment WhatsApp leads on CVs. Multi Engineering Partners used observation-based simulated datasets for the January-June 2025 period. The research follows the CRISP-DM framework through the stages of business understanding, data preparation, modeling, evaluation, and deployment. The main variables analyzed included the number of chats, average response time, frequency of follow-ups, source of leads, and closing status. The results of the analysis showed that HDBSCAN was able to form multiple clusters automatically without determining the number of clusters at the beginning, while detecting noise that represented passive or non-potential leads. Clusters with high communication intensity show a greater closure rate so that they can be prioritized for follow-up strategies. These findings confirm that HDBSCAN is effective in handling data leads with varied characteristics and can be a solution for SMEs in optimizing data-driven sales strategies. This research contributes in the form of a Business Intelligence pipeline to support decision-making and recommendations to improve sales performance.

Keywords: HDBSCAN, Leads, WhatsApp, Clustering, CRM

A. Introduction

The rapid development of digital communication channels has allowed small and medium enterprises (SMEs) to acquire a large number of leads through instant media such as WhatsApp. In Indonesia, many SMEs including CV. Multi Teknik Partners who use WhatsApp as the main channel for pre-sales interactions due to their ease of access, low cost, and one-to-one nature with customers. However, large volumes of interactions result in highly heterogeneous lead data, differing in response speed, chat frequency, lead origin, and follow-up history, making it difficult for sales teams to manually manage and identify high-value leads efficiently. (Mobile phone et al., 2024)

In the context of modern business management, Business Intelligence (BI) and Customer Relationship Management (CRM) systems play a critical role in turning customer interaction logs into strategic decisions that support follow-up

prioritization, customer segmentation, and time-to-sales management. Research in Indonesia shows that the integration of data analytics capabilities in CRM systems can increase effectiveness and profitability through data-driven personalization and prioritization. Globally, the implementation of BI and predictive analytics in CRM has also been proven to provide a competitive advantage through lead ranking, auto-routing, and follow-up recommendation systems. Research, which says that entrepreneurial competence and the use of technology have a positive and significant effect on the entrepreneurial spirit. Therefore, the implementation of data-driven lead segmentation is needed so that SMEs can increase efficiency and effectiveness in the sales process. (Chitra & Heikal, 2024) (Blachowicz et al., 2025) (Hidayat et al., 2021)

In the customer segmentation literature, most studies still use centroid-based clustering techniques such as K-Means or the RFM (Recency Frequency Monetary) methodology to segment customers based on transactions or purchasing behaviors. While these methods have proven to be useful, criticism arises when the data has a very heterogeneous distribution, the shape of the cluster is not round, or there are many noise elements such as inactive leads or invalid numbers; in these conditions, K-Means or similar methods can provide segmentation results that are less than optimal or less interpretable. In response to these limitations, density-based and hierarchical-density-based methods such as HDBSCAN are gaining attention for their ability to detect clusters in arbitrary shapes, handle noise explicitly, and do not require a predetermined number of clusters. Although methodologically promising, the application of HDBSCAN for lead segmentation through instant messaging channels such as WhatsApp, especially in the context of SMEs in Indonesia, is still rarely done and has not been proven empirically. This study aims to fill this gap by applying the HDBSCAN (São Paulo & Sugiharti, 2024) (Nhat, 2024) (González-Alemán et al., 2022) *(Hierarchical Density-Based Spatial Clustering Of Applications with Noise) algorithm* on observation-based dummy data on WhatsApp leads in CV. Multi Engineering Partners for the January-June 2025 period. The dataset used was built simulatively based on the results of observations of the sales team's interaction patterns with potential customers, including characteristics such as the number of messages, average response time, frequency of follow-ups, sources of leads, and the status of closing transactions. Specifically, this study aims to design and extract behavioral features of customer interaction that are relevant to the sales process, then apply the HDBSCAN algorithm to automatically find lead segments and identify leads with low potential or classified noise. Furthermore, this study also evaluates the characteristics of each segment in relation to the conversion rate (closing rate) and prepares operational recommendations so that the sales team can prioritize follow-up based on the segmentation results obtained.

The main contribution of this study is to present empirical evidence that the HDBSCAN algorithm is able to produce interpretive and applicable lead segmentation in the context of WhatsApp-based communication. In addition, this research also builds a Business Intelligence (BI)-based analytics pipeline consisting of feature engineering, clustering implementation, and strategic insight extraction. This approach is expected to be adapted by Small and Medium Enterprises (SMEs) who face similar challenges in

managing lead data and optimizing sales strategies.

B. Materials and Methods

This study uses a descriptive quantitative approach with a data experiment method based on CRISP-DM (Cross Industry Standard Process for Data Mining). This approach was chosen because it provides a systematic framework ranging from business understanding to model application and evaluation, so that research results can be scientifically replicated and further developed. (Judiana et al., 2023)

In addition, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm is used in WhatsApp's lead segmentation process, as it is able to identify complex cluster structures and detect noise without requiring a predetermined number of clusters. This makes HDBSCAN suitable for analyzing heterogeneous lead data. (Stewart & Al-Khassaweneh , 2022) (Dwi Stuttgart & Rosyida , 2023)

1. Research Data and Sources

   The dataset used in this study is dummy (synthetic), but it is constructed based on the results of direct observation of digital marketing activities in CV. Multi-Teknik Partners during the January-June 2025 period. Observation is carried out by paying attention to the interaction patterns of the sales team through the WhatsApp Business channel, which functions as the main means of communication between the company and potential customers (leads).

   The resulting data consists of about 500 entries, with the proportion of lead sources resembling empirical conditions in the field as can be seen in the image below.

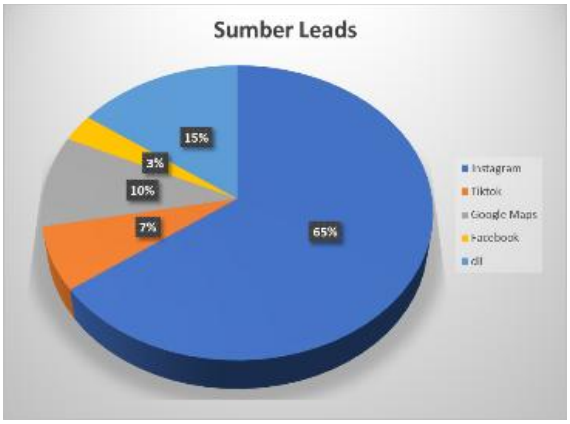

Image 1. Lead source

Each entry has the following attributes:

a) Lead_ID
b) Name
c) Nomor_WA (fictitious and disguised to maintain privacy)
d) Tanggal_Masuk
e) Jumlah_Chat
f) Rata_Waktu_Respon
g) Status_Pesan
h) Frekuensi_Followup
i) Sumber_Leads
j) Closing_Status

All data is built simulatively Using the python programming language in Google Colaboratory with the help of pandas, numpy and datetime libraries. This dataset will be stored and shared openly through the Kaggle platform, so that it can be accessed for further replication and development by other researchers.

2. Research Data and Sources

   The data analysis process is carried out through six main stages of CRISP-DM, namely:

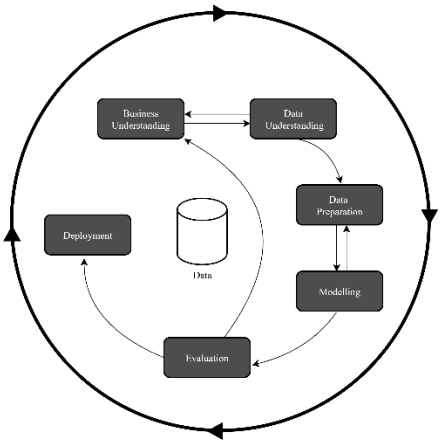

Image 2. CRISP-DM Analysis Process

a) Business Understanding
This stage aims to understand the context of the problem in CV. Multi Technique Partners, namely difficulties in managing and analyzing large and diverse WhatsApp lead data to determine potential customers.

b) Data Understanding
Dummy data was explored to understand the distribution of each variable (number of chats, response time, and lead source) and detect outliers and noise.

c) Data Preparation
It includes the process of data cleaning, elimination of blank values, coding of categorical variables (Sumber_Leads, Status_Pesan, Closing_Status), and normalization of numerical features so that they are ready to be used in clustering algorithms.

d) Modelling
At this stage, the HDBSCAN algorithm is applied to group leads based on similarity in communication and interaction behavior. The main parameters used include min_cluster_size and min_samples.

e) Evaluation
The results of clustering were evaluated by reviewing the composition of each cluster against Closing_Status variables to assess the effectiveness of the model in identifying high-potential leads.

f) Deployment
The segmentation results are visualized in the form of a Business Intelligence (BI)-based analytics dashboard to support marketing team decision-making. All modeling results and scripts will be published via Kaggle for transparency and replication.

3. Aspects of Research Ethics
This study did not involve Human or animal subjects directly. All data is anonymous and does not contain any Personal Information. Thus, no normal ethical permission is required, but the principles of integrity, transparency, and openness of scientific data are maintained throughout the research process.

C. Results and Discussion

The analysis process was carried out using the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm to group WhatsApp CV lead data. Multi-Engineering Partners for the period of January-June 2025. The dataset used consists of 500 entries with the main variables: Number of Chats, Average Response Time, Follow-up Frequency, Message Status, Lead Source, and Closing Status.

The clustering results showed that the HDBSCAN algorithm was able to form multiple groups of leads naturally without

first determining the number of clusters. This process provides an overview of lead distribution based on interaction behavior and potential conversion rate (closing rate). In general, the model produces 4 main clusters and 1 noise group, which indicates a significant variation in the behavioral characteristics of potential customers.

1. Cluster Distribution
   Two-dimensional visualization Using PCA shows a fairly clear separation between clusters. Leads with high engagement rates and *follow-up* frequencies more often tend to cluster in clusters with high *closing* rates, while leads with slow response or low interaction frequency tend to fall into the noise category or low-value clusters.

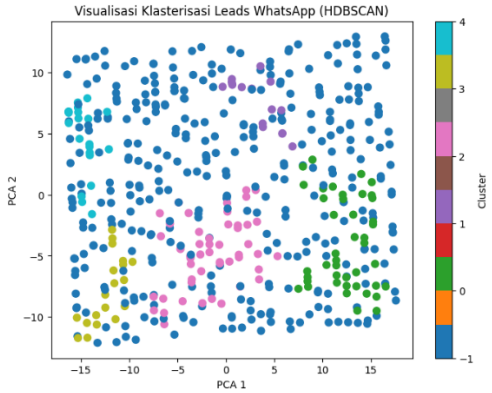

Image 3. Visualization of HDBSCAN clustering results Using two-dimensional PCA

2. Cluster Characteristics
   Descriptive analysis of each cluster shows patterns that are relevant to the sales strategy.

   a) Cluster 0 (High Potential Leads): Has the highest average number of chats (±28 times) and follow-up frequency >4 times, with a *closing* rate of 85%. Most of them come from Instagram.
   b) Cluster 1 (Moderate Leads): Tends to be responsive but with limited interaction (10-15 messages). Closing potential is moderate (~50%).
   c) Cluster 2 (Low Leads): Dominated by no reply or slow response, with an average high response time (≥20 minutes).
   d) Noise (-1): Represents inactive leads, invalid numbers, or data that doesn't show a clear pattern.

3. Business Interpretation and Implications
   These findings suggest that HDBSCAN is effective in separating leads based on interaction behavior and conversion potential, even without assuming cluster numbers. The results of this segmentation can be used to:

   1. Follow-Up Priorities
      Sales teams can focus on clusters with high interaction levels to increase closing chances.
   2. Time Allocation Optimization
      Leads with *noise characteristics Can be ignored or redirected to other marketing channels.*
   3. Personal Communication Strategy
      By understanding the character of each segment, the communication approach (frequency of messages, response time, etc.) can be adjusted to increase sales effectiveness.

4. Research novelty
   The results of this study are consistent with a study that states that the application of unsupervised clustering can improve the efficiency of lead management in the SME sector. However, the HDBSCAN approach used in this study offers advantages because it can detect noise and produce more flexible segmentation than traditional methods such as K-Means. In addition, these results support the finding (Rahmadani et al., 2024) that the use of

density-based algorithms provides higher accuracy in grouping customer data with heterogeneous interaction patterns.

5. Dataset and Notebook Access Links
   All code and experimental results can be accessed openly through the following Kaggle Notebook: https://www.kaggle.com/code/wandacahyani/segmentasi-leads-whatsapp-menggunakan-hdbscan. The dummy dataset used was built based on observations of CV's digital marketing activities. Multi Engineering Partners and is kept publicly for study replication purposes.

D. Conclusions and Suggestions

This study applies the HDBSCAN (Hierarchical Density Based Spatial Clustering of Applications With Noise) algorithm for WhatsApp lead segmentation on CVs. Multi Teknik Partners with observation-based dummy data on digital marketing activities for the January-June 2025 period. The results of the analysis showed that HDBSCAN was able to group *leads* without specifying the number of clusters at the beginning, as well as detecting *the noise* that presented *passive leads*.

This algorithm generates multiple clusters with different characteristics based on the *number of chats, the frequency of follow-ups,* and *the average response time.* Clusters with high interaction Clusters with high interaction and regular follow-up showed greater closure rates, so HDBSCAN proved effective in recognizing complex customer interaction patterns. These findings are in line with research that confirms the advantages of density-based methods over K-Means in handling heterogeneous data. (Auliani, 2024)

Hail research is expected to assist sales teams in determining follow-up priorities, optimizing time allocation, and developing data-driven communication strategies. For further research, it is recommended to use actual data from the CRM and add comparisons of other algorithms such as OPTICS or GMM. The development of *a Business Intelligence (BI) dashboard* is also recommended so that the results of the analysis can be directly used in real time by the marketing team.

Bibliography

Auliani, S. N. (2024). Implementation of Density-Based Spatial Clustering of Applications with Noise and Fuzzy C – Means for Clustering Car Sales. *The Indonesian Journal of Computer Science*, *13*(4). https://doi.org/10.33022/ijcs.v13i4.4135

Blachowicz, T., Wylezek, J., Sokol, Z., & Bondel, M. (2025). Real-Time Analysis of Industrial Data Using the Unsupervised Hierarchical Density-Based Spatial Clustering of Applications with Noise Method in Monitoring the Welding Process in a Robotic Cell. *Information (Switzerland)*, *16*(2). https://doi.org/10.3390/info16020079

Chitra, J., & Heikal, J. (2024). Customer segmentation using the K-Means Clustering algorithm in Foreign Banks in Indonesia. In *Indonesia Accounting Research Journal* (Vol. 11, Issue 4).

Dwi Handayani, F., & Rosyida, I. (2023). Clustering Review of Zenius Application Users on Google Play Store Services Using DBSCAN and HDBSCAN Methods. *Emerging Statistics and Data Science Journal*, *1*(2).

González-Alemán, R., Platero-Rochart, D., Rodríguez-Serradet, A., Hernández-Rodríguez, E. W., Caballero, J., Leclerc,

F., & Montero-Cabrera, L. (2022). MDSCAN: RMSD-based HDBSCAN clustering of long molecular dynamics. *Bioinformatics*, *38*(23), 5191–5198. https://doi.org/10.1093/bioinformatics/btac666

Handijono, A., Irawan Gunarto, R., & Sutrisna, E. (2024). *Utilizing Whatsapps Business for Promotion and Sales*. *4*(1). https://doi.org/10.37481

Nhat, N. M. (2024). Applied Density-Based Clustering Techniques for Classifying High-Risk Customers: A Case Study of Commercial Banks in Vietnam. *Journal of Applied Data Sciences*, *5*(4), 1639–1653. https://doi.org/10.47738/jads.v5i4.344

Nisak, C., & Sugiharti, E. (2024). Customer Lifetime Value Clustering Using K-Means Algorithm with Length Recency Frequency Monetary Model to Enhance Customer Relationship Management ARTICLE HISTORY. *Journal of Advances in Information Systems and Technology*, *6*(1).

Stewart, G., & Al-Khassaweneh, M. (2022). An Implementation of the HDBSCAN* Clustering Algorithm. *Applied Sciences (Switzerland)*, *12*(5). https://doi.org/10.3390/app12052405

Yudiana, Y., Yulia Agustina, A., & Nur Khofifah, and. (2023). *Customer Churn Prediction Using the CRISP-DM Method in the Telecommunications Industry as an Implementation of Customer Retention* (Vol. 8, Issue 1). https://doi.org/doi.org/10.30631/ijoieb.v8i1.1710