Comparison of the Performance of the DBSCAN and ST-DBSCAN Text Mining Algorithms for the Distribution of Ornamental Fish Sales

Atika Putri[1*], Putri Yuli Utami[2], Rizki Surtiyan Surya[3]

[1,2,3] *Department of Information System, Universitas Muhammadiyah Pontianak, Indonesia*

*corr-author: 211230032@unmuhpnk.ac.id

Abstract - The ornamental fish trade in Indonesia is growing rapidly through e-commerce and social media. The large, diverse, and unstructured sales data complicates accurate market mapping. This study analyzes the distribution of ornamental fish sales using density-based clustering algorithms, namely DBSCAN (Density-Based Spatial Clustering of Applications with Noise), which identifies clusters based on data density, enabling the formation of global distribution patterns while detecting outliers, and ST-DBSCAN (Spatial-Temporal DBSCAN), an extension of DBSCAN, which incorporates spatial dimensions to provide more detailed mapping based on geographic proximity. Data were collected through e-commerce platforms and social media such as Twitter. The results revealed that ornamental fish distribution is primarily concentrated in Java and also distributed in Sumatra, Kalimantan, Bali, and Sulawesi. DBSCAN produced five clusters with two noise points and achieved a Silhouette Coefficient of 0.791. Meanwhile, ST-DBSCAN produced six clusters with three noise points, a Silhouette Coefficient of 0.656, and a Davies-Bouldin Index of 0.985. Overall, DBSCAN effectively represents global distribution patterns and detects anomalies, while ST-DBSCAN enhances the analysis with spatial insights that highlight geographic variations. Together, these two algorithms provide a more comprehensive understanding of the distribution of ornamental fish sales by species and location.

Keywords: Clustering, DBSCAN, ornamental fish, ST-DBSCAN, Word2Vec.

I. INTRODUCTION

The ornamental fish trade is an economic subsector that has experienced significant growth in recent years, both globally and nationally. In Indonesia, public interest in ornamental fish continues to rise, not only as a hobby but also as a promising investment. This growth is reflected in the increasing number of online stores selling various types of ornamental fish through e-commerce platforms such as Shopee, Lazada, Blibli, and Tokopedia, as well as the increasing number of ornamental fish-themed discussions shared by users on social media platforms like Twitter [1]. However, this growth has also generated large, diverse, and unstructured sales data, posing challenges in mapping market distribution patterns [2]. Conventional methods tend to be limited in capturing complex variations in digital data, such as the presence of outliers, uneven distribution, and non-uniform text formats [3]. Therefore, a more adaptive analytical approach is needed that utilizes alternative data from e-commerce and social is required [4].

The fish distribution domain was selected for this study due to its strong economic relevance and prominent role in Indonesia's digital marketplace [5]. Indonesia's position as a major global exporter of ornamental fish underscores the strategic importance of this sector [6]. In addition, the digital trade of ornamental fish generates substantial unstructured text data (fish names, product descriptions) and clear spatial distribution patterns across regions such as Java, Sumatra, and Kalimantan [7]. These characteristics make this domain particularly suitable for analytical methods that combine semantic and spatial attributes. The mixed nature of the data aligns well with the capabilities of DBSCAN and ST-DBSCAN: DBSCAN identifies global density-based patterns [8], while ST-DBSCAN incorporates spatial radius parameters to capture localized distribution structures [9]. Thus, this domain provides an ideal case for evaluating the performance of density-based clustering on text spatial data.

Several previous studies have demonstrated the effectiveness of text mining and clustering approaches in extracting information from unstructured text data [10]. Research exploring public sentiment toward the film Dirty Vote using the Naïve Bayes algorithm and Logistic Regression showed that Logistic Regression had a higher accuracy (95.5%) than Naïve Bayes (91.1%), while emphasizing the importance of selecting the right analysis method to produce accurate text classification

[11]. Another study comparing the K-Means and K-Medoids algorithms in clustering student thesis titles found that K-Medoids produced better cluster quality (DBI 1.56), while K-Means excelled in computational speed. These findings suggest that each algorithm has specific advantages, depending on the context of its application [12]. Furthermore, research conducting sentiment analysis to predict presidential election results via Twitter using Lexicon-Based and Random Forest approaches showed that Random Forest was superior in detecting negative sentiment (94%), while the Lexicon-Based method was better at predicting positive and neutral sentiment [13]. This domain is well suited for comparing DBSCAN and ST-DBSCAN because ornamental fish sales exhibit both semantic variation (fish names, product descriptions) and spatial variation (location-based distribution), making it an ideal case for evaluating density-based clustering on mixed text-spatial data.

Thus, selecting the fish distribution domain is both economically and methodologically relevant for applying text mining and clustering techniques to unstructured data. Based on this foundation, this study compares DBSCAN and ST-DBSCAN to evaluate their performance in identifying semantic and spatial distribution patterns. DBSCAN effectively captures global density-based clusters, whereas ST-DBSCAN introduces spatial attributes that enable more detailed identification of local distribution patterns. This study aims to assess the impact of integrating spatial components on clustering outcomes and to evaluate the theoretical and practical effectiveness of spatial-semantic clustering approaches [14].

II. METHOD

This research was conducted through several main stages arranged systematically. A flowchart of the research stages can be seen in Fig. 1 [15].

A. Data Collection

The dataset used in this study comes from secondary data obtained through e-commerce platforms such as Shopee, Lazada, Blibli, and Tokopedia, in real-time through a scraping process, as well as public conversations on Twitter about ornamental fish using the official Twitter API. The Twitter data collection process was carried out without a time limit so that the data obtained represents public conversations based on the results available from the API. The collected data consists of 1,056 entries, with two main attributes: fish name and seller location. Both data sources were collected through keyword searches for "Kalimantan ornamental fish", "freshwater ornamental fish", and "Pontianak ornamental fish" [10].

B. Pre-Processing Data

In this study, text data processing is used to organize fish name information, while location data processing is focused on processing geographic information about sellers [16].

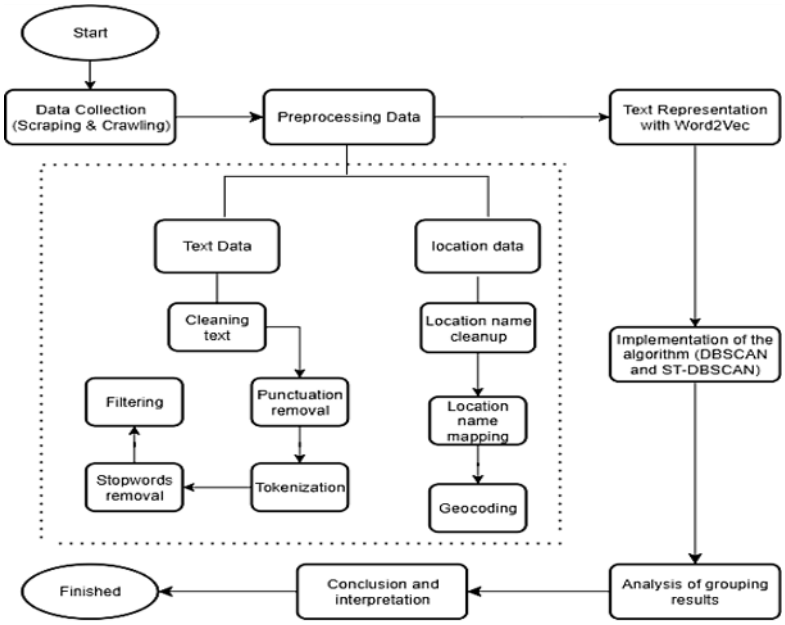

Fig. 1 Research stages

1) Text Data: Text processing was performed to organize the fish name information. The pre-processing stages included text cleaning from irrelevant characters, lowercase normalization, tokenization, stopword removal, and token filtering using a list of valid ornamental fish names. This stage ensures that only words relevant to the research object are retained.

2) Location Data: Location data is processed in two steps: extracting city/region names from the location text, then converting them to latitude and longitude coordinates through geocoding using the Geopy library. This coordinate representation is necessary to support spatial analysis in the clustering process.

C. Representation Word2Vec

The pre-processed text data is then used in the Word2Vec process with the Skip-Gram approach to generate vector representations of ornamental fish names. Each text entry is converted into a vector by calculating the average of the word vectors contained within it. The method for obtaining document vector values is shown in (1) and (2).

$$\frac{1}{T} \quad \sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log p\left(w_{t+j} \mid w_t\right) \qquad (1)$$

where w_t represents the center word being analyzed, $w_{(t+j)}$ refers to the context words that appear around the center word, c indicates the context window size, which determines how far the words to the left and right of the center word are considered; and t denotes the total number of words in the corpus used during the model training process.

$$vec_{c,j} = \quad \frac{1}{n_j} \sum_{i=1}^{m} w_i v \quad (w_i) \qquad (2)$$

where vec(c,j) represents the j-th document vector in class c; n_j is the number of features contained in document j; $v(w_i)$ describes the feature word vector w, which represents specific characteristics within the document; and w_i indicates the feature vector value of the word w.

D. Clustering with DBSCAN

The DBSCAN algorithm in this study was applied to data including text representations of ornamental fish names using Word2Vec and sales location coordinates [17]. The DBSCAN implementation workflow in this study is shown in Fig.2. The process begins by determining the epsilon (eps) value as the radius for searching the nearest points and MinPts as the minimum number of points within that radius that are considered part of a cluster. Once these parameters are defined, a starting point is randomly selected from the dataset. The distance from this starting point to all other points in the dataset is then calculated using the Euclidean Distance function. The basic formula for the Euclidean Distance function is given in (3).

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (pi - qi)^2} \qquad (3)$$

where d(p, q) represents the distance between points p and q; $(p_i - q_i)$ is the 1st coordinate of the two points used in the distance calculation; and n represents the number of data dimensions involved in the calculation process.

If the number of points within the eps radius meets or exceeds the MinPts value, the point is identified as a core point, and the cluster expansion process is performed by including other points that share the same density. Conversely, if a point does not have enough neighboring points according to the MinPts requirement, it is classified as noise. The algorithm then continues by processing the remaining unvisited points until all points in the dataset have been examined.

E. Clustering with ST-DBSCAN

The ST-DBSCAN workflow in this study essentially follows the same principles as DBSCAN. This approach allows the formation of clusters that represent the spatial distribution patterns of ornamental fish, with density calculations that consider the differences between spatial dimensions and non-spatial attributes, unlike DBSCAN which treats all dimensions equally [18].

III. RESULT AND DISCUSSION

This section presents a comparative analysis of the parameter testing results of the DBSCAN and ST-DBSCAN algorithms. Prior to the comparison, each algorithm was tested to obtain the best parameter configuration. Cluster quality evaluation was performed using the Silhouette Coefficient (SC) value, which ranges from –1 to 1, with values closer to 1 indicating good cluster quality. In addition to the SC, this study also considered the Davies–Bouldin index (DBI), the number of clusters formed, the amount of noise, and the consistency of the results with respect to parameter variations. PCA visualization and distribution maps were also used to practically assess the clarity of the cluster patterns. The parameter testing results for both algorithms are shown in Table I and Table II [19].

A. Testing the Parameters of the DBSCAN and ST-DBSCAN Algorithms

Parameter testing was conducted to determine the optimal configuration for each algorithm. In DBSCAN,

the parameters tested included the epsilon value (eps) as the search radius and MinPts as the minimum number of points to form a core point. In ST-DBSCAN, the parameters tested included the spatial radius (Eps1), the non-spatial radius (Eps2), and MinPts. The workflow of both algorithms was used as the basis for parameter testing in this study [20].

1) DBSCAN Test Results: Based on Table I, a small eps value (0.2) produces many clusters (13) with high noise (21) and poor separation (SC = 0.692641). As eps increases, the number of clusters and noise decreases, while the cluster quality improves. An eps value of 0.4 gives the highest SC (0.827320) but the distribution is unstable. The eps configuration of 1.0 and MinPts 3 was chosen as the best result because it produces 5 clusters, low noise (2), and good separation (SC = 0.791307). If eps exceeds 1.0, clusters tend to merge, thus reducing the clustering quality.

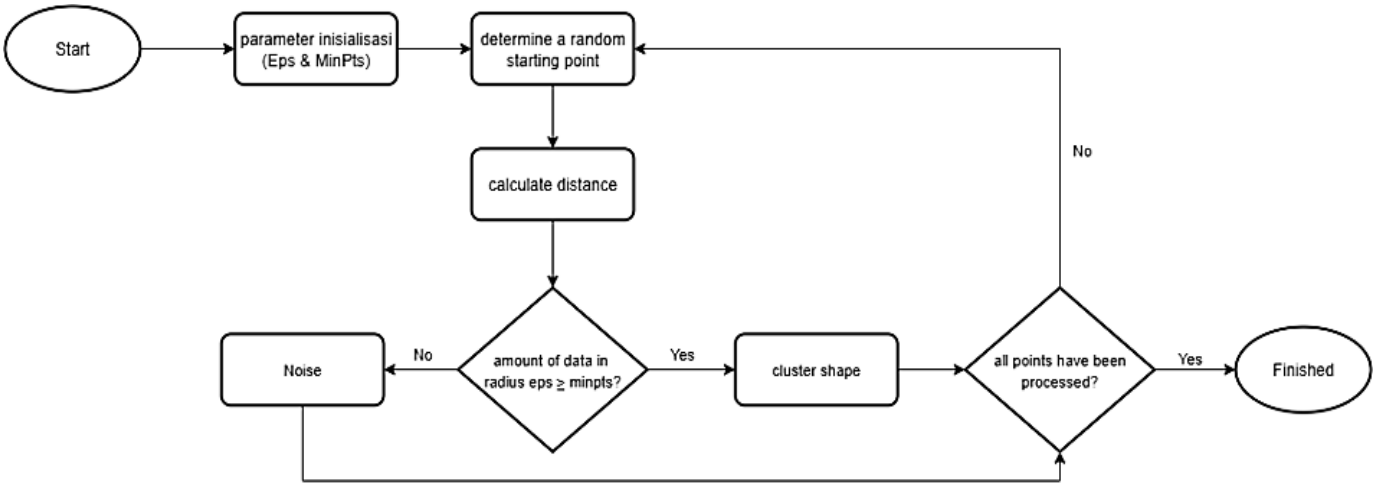

Fig. 2 DBSCAN workflow

TABLE I
DBSCAN PARAMETER TESTING

Eps	MinPts	Number of Clusters	Noise Amount	Sillhouette Index
0.2	3	13	21	0.692641
0.3	3	9	15	0.812501
0.4	3	8	13	0.827320
0.5	3	6	10	0.722933
0.6	3	5	8	0.778416
0.7	3	5	8	0.778416
0.8	3	5	7	0.774742
0.9	3	5	4	0.789946
1.0	3	5	2	0.791307

TABLE II
ST-DBSCAN PARAMETER TESTING

Eps1	Eps2	MinPts	Number of Clusters	Noise Amount	*Silhouette coefficient*	*Davies-Bouldin Index (DBI)*
6.0	2.0	5	5	10	0.654857	1.500872
6.0	3.0	5	5	7	0.656181	1.050187
7.0	2.0	5	5	5	0.656993	2.267939
7.0	3.0	2	6	2	0.656698	1.449330
7.0	2.0	2	6	3	0.656484	0.985725

2) ST-DBSCAN Test Results: The results in Table II show that the combination of Eps1 = 7.0, Eps2 = 2.0, and MinPts = 2 is the optimal parameter, resulting in 6 clusters with 3 noise levels. A low DBI value (0.985725) indicates clear cluster boundaries, while SC = 0.656484 indicates fairly good quality. Other alternative configurations such as Eps1 = 6.0 and Eps2 = 3.0 are also stable, but have higher DBI, resulting in less clear cluster boundaries. Therefore, the first combination was chosen because it provides a realistic number of clusters, minimal noise, and the best spatial-semantic visibility.

B. *Comparison of Parameter Test Results*

The results of testing the parameters of the two algorithms are presented in Table III, which presents a comparison of the number of clusters, the proportion of data containing noise, and the Silhouette Coefficient value for each parameter configuration.

Parameter comparison (Table III) shows that DBSCAN (Eps = 1.0, MinPts = 3) produces 5 clusters with 2 noises and Silhouette = 0.791, while ST-DBSCAN (Eps1 = 7.0, Eps2 = 2.0, MinPts = 2) forms 6 clusters with 3 noises and Silhouette = 0.656. DBSCAN groups data based on global density, while ST-DBSCAN adds a spatial dimension so that it is able to capture more specific distribution patterns at certain locations.

C. *Visualization of Clustering Results*

The clustering results visualization includes 2D PCA, word clouds, and location distribution maps. Fig. 3 and Fig. 4 show the PCA visualizations of DBSCAN and ST-DBSCAN, while Fig. 5 and 6 display the word clouds of each cluster [21].

TABLE III
BEST ALGORITHM PARAMETERS

	Eps	Eps1	Eps2	MinPts	Cluster	Noise	Silhoutte Coefficient
DBSCAN	1.0	-	-	3	5	2	0.791307
ST-DBSCAN	-	7.0	2.0	2	6	3	0.656484

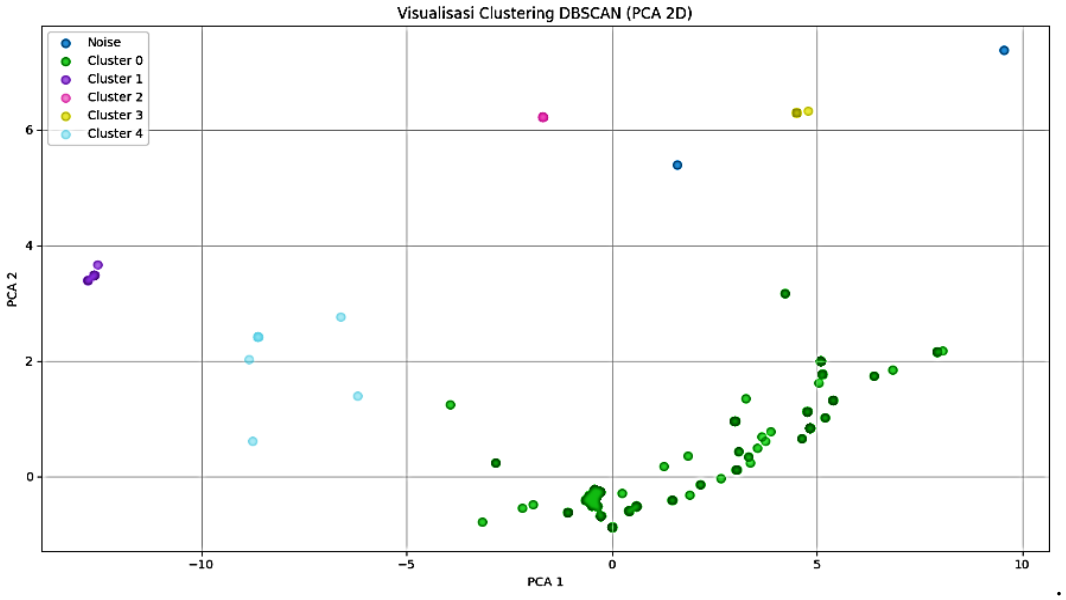

Fig. 3 2D PCA DBSCAN results

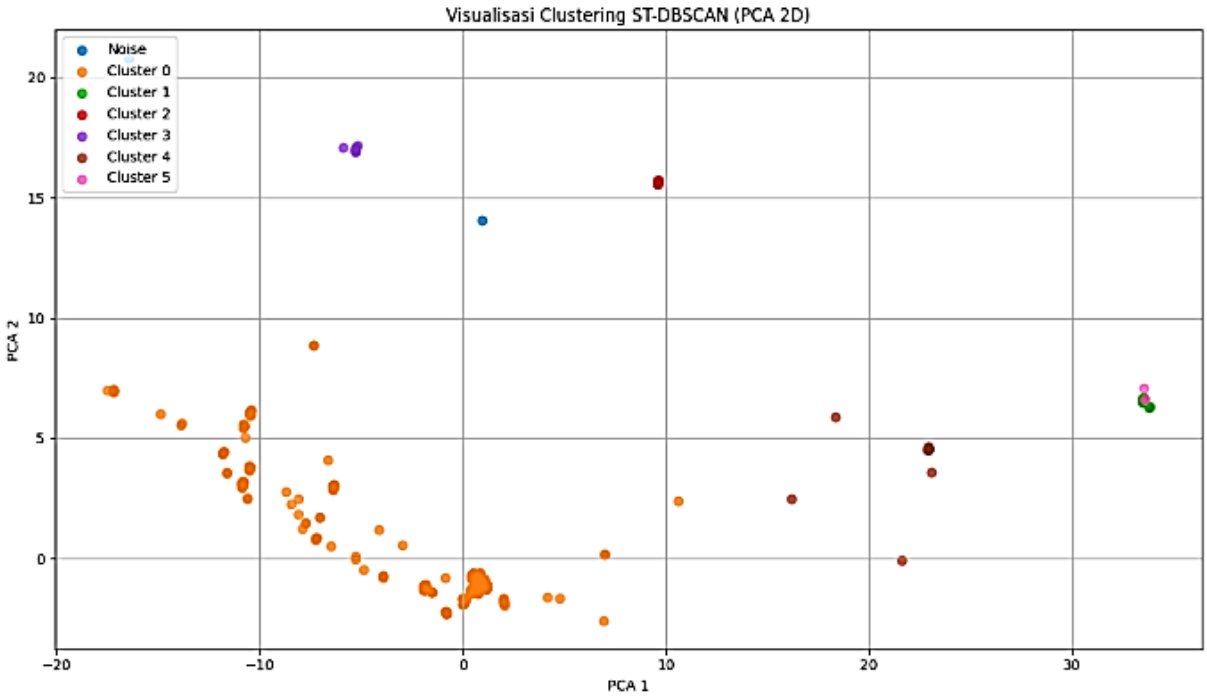

Fig. 4 2D PCA ST-DBSCAN results

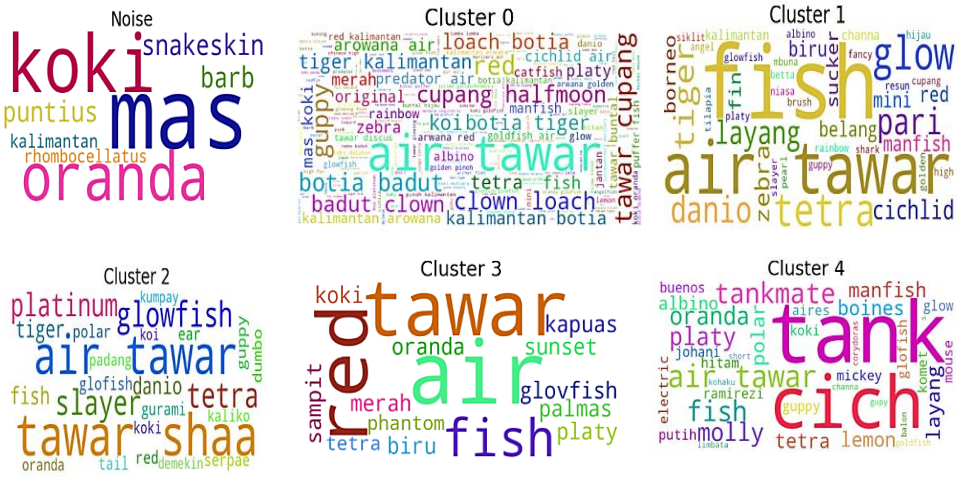

Fig. 5 DBSCAN word cloud results

1) DBSCAN Visualization Results: The PCA visualization in Fig. 3 shows five clusters, with cluster 0 being the largest and densest. Points that do not follow the density pattern are detected as noise. The word cloud in Fig. 5 shows the dominant words in each cluster. Cluster 0 is dominated by freshwater fish terms such as tetra, betta, platy, and botia. Cluster 1 contains brightly colored fish such as glows and danios, while cluster 2 displays unique variations such as platinum and calico. Cluster 3 refers to regional freshwater fish, and cluster 4 describes aquarium companion fish such as guppies, mollies, and cichlids [22].

2) ST-DBSCAN Visualization Results: The PCA in Fig. 4 shows six clusters with a more specific distribution in certain regions. The word cloud in Fig. 6 illustrates the grouping based on semantic similarity and location. Clusters 0 and 3 are dominated by freshwater fish such as bettas, cluster 1 contains luminous fish such as tetras and danios, cluster 2 contains unique variations such as glowfish and platinumfish, cluster 4 contains aquarium companion fish, and cluster 5 contains fish with special morphological characteristics such as albinos and sharks. Some points that do not meet the spatial-semantic proximity are detected as noise [18].

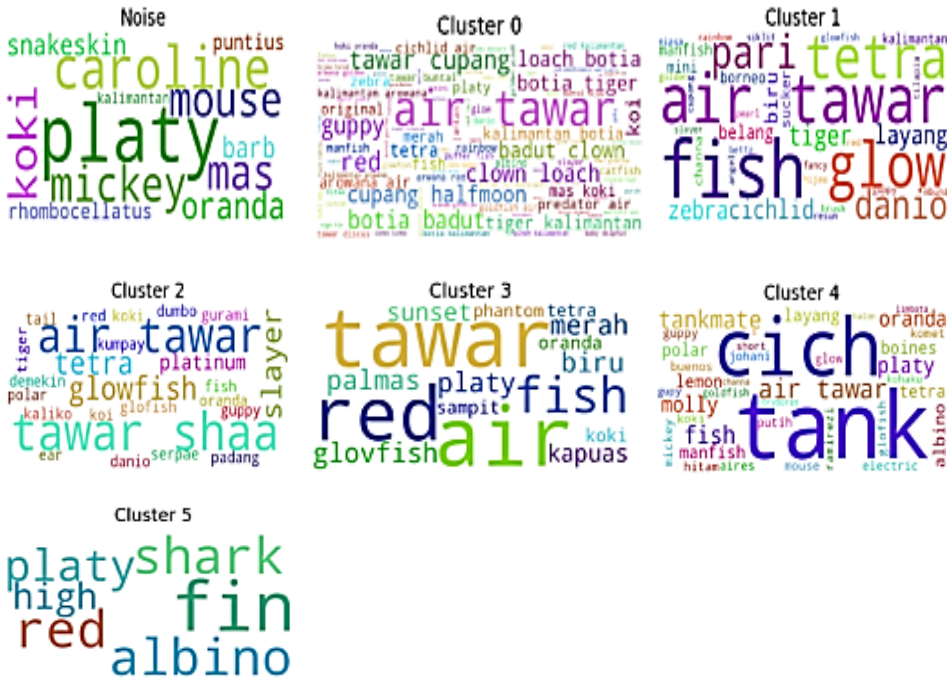

Fig. 6 ST-DBSCAN word cloud results

3) Visualization of DBSCAN and ST-DBSCAN Location Distribution Map: The distribution map visualizations shown in Fig.7 and Fig. 8 demonstrate the clustering results of both algorithms, including points identified as disturbances and cluster concentration areas. This allows for direct observation of differences in ornamental fish distribution patterns across regions and also serves as a basis for comparing the effectiveness of each algorithm's parameters in clustering sales data [23].

The DBSCAN distribution map in Fig. 7 shows that cluster 0 represents the largest activity centers, mostly spread across the islands of Java and Bali, while other clusters are spread across Sumatra, Kalimantan, and Sulawesi. This pattern suggests that DBSCAN tends to cluster data based on global density [17]. In contrast, the ST-DBSCAN map in Fig. 8 depicts a more geographically specific distribution pattern. Large clusters emerge in East Java, Central Java, and Jakarta, with variations in other clusters across Kalimantan and Sumatra [11]. Comparing the two algorithms in this study confirms that their use provides different perspectives on the data. DBSCAN places greater emphasis on global point density, while ST-DBSCAN adds spatial information, allowing for more detailed analysis of distribution patterns in areas with limited data [24,25,26].

Fig. 7 DBSCAN distribution map

Fig. 8 ST-DBSCAN distribution map

IV. CONCLUSION

This study successfully implemented and analyzed the DBSCAN and ST-DBSCAN algorithms for clustering ornamental fish sales data, utilizing Word2Vec text representation and spatial data in the form of location coordinates. DBSCAN, with parameters Eps = 1.0 and MinPts = 3, generated five clusters with two noise points and a Silhouette Coefficient of 0.791. These results demonstrate DBSCAN's ability to form global distribution patterns and effectively detect outliers. Meanwhile, ST-DBSCAN, with parameters Eps1 = 7.0, Eps2 = 2.0, and MinPts = 2, generated six clusters with three noise points, a Silhouette Coefficient of 0.656, and a Davies-Bouldin Index of 0.985. By considering the spatial dimension, ST-DBSCAN is able to describe the data distribution in greater detail, including the formation of specific clusters in Sumatra and Kalimantan that were not visible through text-based analysis alone. Overall, DBSCAN was effective in forming global clusters and detecting anomalies, while ST-DBSCAN extended the analysis with spatial information, resulting in more geographically specific clusters. The research findings also showed that ornamental fish sales activity still concentrated on Java, while Sumatra and Kalimantan experienced lower intensity. These results provide insights for industry players and policymakers regarding marketing strategies and the potential for regional equity, while also confirming that unstructured data analysis can be utilized to understand the dynamics of the digital economy in the ornamental fish trade. Future research can be developed by adding a temporal dimension to examine annual distribution changes and integrating environmental, aquatic, or socioeconomic data. The application of deep learning algorithms also has the potential to improve accuracy in detecting more complex spatial and semantic clusters.

REFERENCES

[1] I. B. Mahendra, I. M. G. Sunarya, and I. M. A. Wirawan, "Comparison of Multinomial, Bernoulli, and Gaussian Naïve Bayes for Complaint Classification in Pro Denpasar Application," *JUITA J. Inform.*, vol. 13, no. 1, pp. 77–86, 2025, doi: 10.30595/juita.v13i1.24828.

[2] M. F. Akbar and L. Zahrotun, "K-Means Centroid Optimization with Genetic Algorithm for Clustering Micro , Small , Medium Enterprises in Yogyakarta," *JUITA J. Inform.*, vol. 13, no. 2, pp. 87–97, 2025, doi: 10.30595/juita.v13i2.25480.

[3] A. R. Samsudin, D. H. Fudholi, and L. Iswari, "Temporal Spatial Property Profiling and Identification of Earthquake Prone Areas Using St-Dbscan and K-Means Clustering," *J. Tek. Inform.*, vol. 5, no. 3, pp. 917–929, 2024, doi: 10.52436/1.jutif.2024.5.3.1293.

[4] N. A. Azizah and T. Ernawati, "Analisis Sentimen Ulasan Pelanggan Terhadap Produk Roti Macan Artisan Di E-Commerce Tokopedia Menggunakan Metode Clustering," *J. Inf. Syst. Applied, Manag. Account. Res.*, vol. 8, no. 3, p. 580, 2024, doi: 10.52362/jisamar.v8i3.1576.

[5] A. D. B. T. A, M. H. B, S. J. C, and N. Zulbainarni, "Technology , Market , and Complexity Market-based dynamic capabilities for MSMEs : Evidence from Indonesia ' s ornamental fish industry," *J. Open Innov. Technol. Mark. Complex.*, vol. 9, no. 3, p. 100123, 2023, doi: 10.1016/j.joitmc.2023.100123.

[6] A. Dompak, B. Tarihoran, M. Hubeis, and S. Jahroh, "Competitiveness of and Barriers to Indonesia ' s Exports of Ornamental Fish," vol. 15, no. 11, p. 8711, 2023, doi:

10.3390/su15118711.

[7] S. G. Akmal, B. P. D. Záme, R. E. Prabowo, and A. M. Khatami, "Living Resources Marine ornamental trade in Indonesia," *Aquat. Living Resour.*, vol. 33, no. 25, p. 8, 2020, doi: 10.1051/alr/2020026.

[8] G. Earthquakes, D. Using, and D. Algorithm, "PENGELOMPOKAN DATA GEMPA BUMI MENGGUNAKAN GROUPING EARTHQUAKES DATA USING," pp. 31–38, 2021, doi: 10.31172/jmg.v22i1.738.

[9] D. K. Widyawati, A. Fauzan, and U. I. Indonesia, "An ST-DBSCAN Approach to Spatio-Temporal Clustering of Earthquake Events in West Java , Indonesia," *J. Teor. dan Apl. Mat.*, vol. 9, no. 4, pp. 1292–1308, 2025, doi: 10.31764/jtam.v9i4.33079.

[10] T. Taka and K. Tsuda, "Trends in consumer evaluations of tuna: Text mining of online reviews," *Procedia Comput. Sci.*, vol. 246, no. C, pp. 706–713, 2024, doi: 10.1016/j.procs.2024.09.489.

[11] M. M. Wikantari, Y. Sibaroni, and A. F. Ihsan, "Clustering Content Types and User Motivation Using DBSCAN on Twitter," *J. Comput. Syst. Informatics*, vol. 4, no. 4, pp. 741–748, 2023, doi: 10.47065/josyc.v4i4.3750.

[12] H. Xu and C. Tian, "Attitudes and preferences of the Chinese public towards products made from recycled materials: A text mining approach," *Resour. Conserv. Recycl. Adv.*, vol. 24, no. October, p. 200234, 2024, doi: 10.1016/j.rcradv.2024.200234.

[13] R. D. Himawan and E. Eliyani, "Perbandingan Akurasi Analisis Sentimen Tweet terhadap Pemerintah Provinsi DKI Jakarta di Masa Pandemi," *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 1, p. 58, 2021, doi: 10.26418/jp.v7i1.41728.

[14] W. Arisandi and S. Anggai, "Analisis Sentimen Ulasan Pengguna Aplikasi Media Sosial X Di Play Store Menggunakan Algoritma Long Short- Term Memory (LSTM) Dan Gated Recurrent Unit (GRU)," vol. IX, no. September, pp. 63–72, 2025, doi: 10.47970/siskom-kb.v9i1.875.

[15] N. Nasution and F. Rakhmawati, "Segmentasi Pengguna E-Wallet Dengan Menggunakan Metode Dbscan (Density Based Spatial Clustering Application With Noise) Di Kota Medan," *J. Lebesgue J. Ilm. Pendidik. Mat. Mat. dan Stat.*, vol. 4, no. 2, pp. 1386–1392, 2023, doi: 10.46306/lb.v4i2.415.

[16] J. Muliawan and E. Dazki, "SENTIMENT ANALYSIS OF INDONESIA'S CAPITAL CITY RELOCATION USING THREE ALGORITHMS: NAÏVE BAYES, KNN, AND RANDOM FOREST," *J. Tek. Inform.*, vol. 4, no. 5, pp. 1227–1236, 2023, doi: 10.52436/1.jutif.2023.4.5.347.

[17] A. S. Ritonga and I. Muhandhis, "Application of Text Mining and DBSCAN Method for Clustering E-Commerce Tweet Data," *Pros. Semin. Nas. UNIMUS*, vol. 4, pp. 113–123, 2021.

[18] S. H. Hastuti, A. Septiani, H. Hendrayani, and W. P. Nurmayanti, "Penerapan Metode OPTICS dan ST-DBSCAN untuk Klasterisasi Data Kesehatan," *Edumatic J. Pendidik. Inform.*, vol. 8, no. 1, pp. 252–261, 2024, doi: 10.29408/edumatic.v8i1.25765.

[19] D. J. Manalu, R. Rahmawati, and T. Widiharih, "PENGELOMPOKAN TITIK GEMPA DI PULAU SULAWESI MENGGUNAKAN ALGORITMA ST-DBSCAN (Spatio Temporal-Density Based Spatial Clustering Application with Noise)," *J. Gaussian*, vol. 10, no. 4, pp. 554–561, 2021, doi: 10.14710/j.gauss.v10i4.29499.

[20] A. S. Valente, T. Trunfio, M. Aiello, D. Baldi, M. Baldi, S. Imbò, M. A. Russo, C. Cavaliere, M. Franzese, Monica., "Text mining approach for feature extraction and cartilage disease grade classification using knee MRI radiology reports," *Comput. Struct. Biotechnol. J.*, vol. 24, no. October, pp. 622–629, 2024, doi: 10.1016/j.csbj.2024.10.003.

[21] A. A. Rio Wirawan a, Erly Krisnanik a, "Penambangan Teks untuk Peramalan Berita di Situs Web Turnback Hoax," *J. Int. PADA Vis. Inform.*, vol. 8, no. 1, 2024, doi: 10.62527/joiv.8.1.1939.

[22] J. Krishnan, B. Bhattacharjee, M. Pratap, J. K. Yadav, and M. Maiti, "Survival strategies for family-run homestays: Analyzing user reviews through text mining," *Data Sci. Manag.*, vol. 7, no. 3, pp. 228–237, 2024, doi: 10.1016/j.dsm.2024.03.003.

[23] K. A. Wijaya and R. S. Oetama, "Visualization And Forecast For Pollution Death Threats In ASEAN," vol. 4, no. 6, pp. 1622–1633, 2023, doi:10.46729/ijstm.v4i6.970.

[24] A. L. Ylber Januzaj, Edmond Beqiri, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *Encycl. Optim.*, vol. 1, pp. 687–694, 2023, doi: 10.1007/978-0-387-74759-0_123.

[25] T. Cahya Herdiyani and A. U. Zailani, "Sentiment Analysis Terkait Pemindahan Ibu Kota Indonesia Menggunakan Metode Random Forest Berdasarkan Tweet Warga Negara Indonesia Sentiment Analysis Related to Transportation of Indonesian Capital City Using Random Forest Method Based On Tweet Of Indonesian Citizens," vol. 3, no. 2, pp. 154-165, 2022, doi: 10.35957/jtsi.v3i2.2920..

[26] R. Nadlifatin, S. F. Persada, A. Beatrice, N. P. Putri, M. N. Young, Y. T. Prasetyo, A. A. N. P. Redi., "Exploring the Effectiveness of Work from Home: A Text Mining Analysis of Employee Perceptions and Experiences," *Procedia Comput. Sci.*, vol. 234, pp. 448–454, 2024, doi: 10.1016/j.procs.2024.03.026.