Triyan Agung Laksono[1]
Master in Information Technology
Faculty of Information Technology
Universitas Teknologi Digital Indonesia,
Yogyakarta, Indonesia
email: triyan31@stiesbi.ac.id

Widyastuti Andriyani
Master in Information Technology
Faculty of Information Technology
Universitas Teknologi Digital Indonesia,
Yogyakarta, Indonesia
email: widya@utdi.ac.id

Fadhlih Girindra Putra
Master in Information Technology
Faculty of Information Technology
Universitas Teknologi Digital Indonesia,
Yogyakarta, Indonesia
email: students.fadhlihgiri24@mti.utdi.ac.id

Ivonia Fatima Ruas da silva
Master in Information Technology
Faculty of Information Technology
Universitas Teknologi Digital Indonesia,
Yogyakarta, Indonesia
email: ruasdasilvaivoniafatima@gmail.com

Wiwi widayani
Faculty of Computer Science
Information Systems
email: wiwi.w@amikom.ac.id

Digital Activity Location Clustering Based on Twitter Geospatial Data for Spatiotemporal Business Intelligence

This research develops an approach for clustering digital activity locations based on Twitter geospatial data with the aim of supporting business intelligence spatiotemporal . By utilizing the Twitter Geospatial Data dataset containing more than 14 million tweets geo-tagged from the United States, this study implements and compares the DBSCAN and K- Means algorithms to identify spatial and temporal patterns of Twitter user activity. The research process begins with the data pre -processing stage using the Knowledge Discovery Database (KDD), followed by the implementation of the clustering algorithm , and ending with the integration of the results into the dashboard. business intelligence using Power BI . The results show that DBSCAN is able to detect irregular clusters that follow geographic patterns and population density, while K- Means produces a division of the region into three main clusters (West Coast, Central Region, and East Coast) with different temporal activity patterns. Integration of clustering results into a BI dashboard produces actionable business insights , such as identification of digital activity hotspots , optimal time for content delivery, geographic segmentation for marketing strategies, and temporal activity patterns for campaign scheduling. This research contributes to the development of an integrated spatiotemporal analysis pipeline to support data-driven decision making.

***KeyWords**: Clustering ; Geospatial Data ; Twitter; Business intelligence ; Spatiotemporal Analysis*

This Article was:
submited: 25-06-25
accepted: 09-07-25
publish on: 20-07-25

How to Cite:
T. A. Laksono, et al, "Digital Activity Location Clustering Based on Twitter Geospatial Data for Spatiotemporal Business Intelligence", Journal of Intelligent Software Systems, Vol.4, No.1, 2025, pp.22–27, 10.26798/jiss.v4i1.2005

1 Introduction

The development of information and communication technology has driven the rapid growth of digital data, especially from social media such as Twitter . The data generated by social media users is not only textual, but also contains very rich spatial and temporal information, such as location (longitude and latitude) and activity time (timestamp) [1]. This spatial-temporal data can be used for various analytical purposes, ranging from event detection, monitoring public opinion, to data-based business decision making (business intelligence) [2], [3].

With the increasing need for fast and accurate decision making, business intelligence (BI) is one of the fastest growing fields. BI leverages big data, including spatial-temporal data from social media, to generate insights that support an organization's business and operational strategies [4], [5]. However, the use of spatial-temporal data from social media for BI still faces major challenges, such as the very large volume of data, the diversity of spatial and temporal patterns, and the presence of noise and outliers in the data [6].

Based on this, one of the methods that is widely used to group spatial-temporal data is clustering . The K- Means Algorithm and DBSCAN are two algorithms clustering is popular and widely applied in various studies related to spatial and temporal data [7], [8]. The K- Means algorithm known for its simplicity and speed of processing, but still has shortcomings, namely it is less effective for data with irregular cluster shapes and is sensitive to outliers . On the other hand, DBSCAN is able to detect clusters with changing shapes and identify noise , although choosing fixed parameters is a challenge in itself [9], [10].

Various studies have developed methods or combinations of algorithms to improve the accuracy and effectiveness of spatial-temporal data clustering, for example by combining K- Means for initialization or determination of DBSCAN parameters [11], [12]. In addition, the integration of clustering results into the dashboard business intelligence using tools such as Power BI or Google Earth Engines are also starting to be widely applied to facilitate visualization and interpretation of analysis results [4], [5]. However, to date, there is still very little research that develops an integrated social media spatial-temporal analysis pipeline from the clustering process to BI visualization, especially those that are ready to be adapted to the Indonesian context [13].

Based on this background, this study aims to develop a spatial-temporal analysis pipeline of digital activity based on Twitter data using a combination of the DBSCAN and K- Means algorithms. , and integration of the results into the dashboard business intelligence . Although the dataset used is from the United States [14],

[1]Corresponding Author.

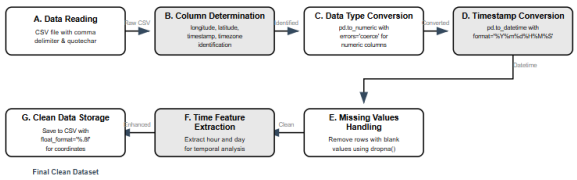

Fig. 1 Research procedure

the pipeline and methods developed have the potential to be applied to cases in Indonesia if local data is available. The novelty of this study lies in the comprehensive integration of spatial-temporal clustering processes into BI visualizations, as well as testing a combination of algorithms on social media spatial-temporal data to support data-driven decision making[15].

2 Method

This section explains the research stages systematically starting from data collection, data cleaning, clustering analysis process , to integrating the results into the dashboard. business intelligence . Each step is designed to ensure valid, accurate, and well-interpretable analysis results according to research objectives.

2.1 Data Sources and Pre-Processing. The dataset used in this study is Twitter Geospatial Data taken from UCI Machine Learning Repository [14]. This dataset contains more than 14 million tweets . geo-tagged data from the United States, including longitude , latitude , timestamp , and timezone attributes . Data pre-processing was done using Python with the library pandas , following the Knowledge approach Discovery Database (KDD) which includes data selection, data cleaning, data transformation, and normalization. This KDD approach was also applied by previous studies in their research on clustering book reading habits in students in Kendari City, where this process proved effective in producing meaningful [15] insights from user behavior data. Pre-processing is performed to ensure data quality, including:

a. Data Reading: Data is read from a CSV file with comma delimiter and quotechar .
b. Column Determination: Columns are identified as ' longitude ', ' latitude ', ' timestamp ', and ' timezone '.
c. Data Type Conversion: Numeric columns (longitude , latitude , timezone) are converted to numeric types using pd.to_numeric with the errors = ' coerce ' parameter to handle invalid values.
d. Timestamp Conversion : Timestamp column is converted to datetime format using pd.to_datetime with format '% Y%m%d%H%M%S '.
e. Missing Handling Values : Rows with blank values in important columns are deleted using dropna .
f. Time Feature Extraction: Hour and day features are extracted from timestamps for temporal analysis.
g. Clean Data Storage: Pre -processing results are stored in CSV format with proper decimal format for longitude and latitude using the float_format ='%.8f' parameter.

Due to the very large size of the dataset (14 million records), 100,000 data were randomly sampled (with random_state = 42) for computational efficiency at the clustering stage , because otherwise the computation would require resources. which is very big.

2.2 Clustering Process Spatiotemporal.

2.2.1 DBSCAN Algorithm and Parameter Determination. clustering process uses the DBSCAN (Density-Based Clustering) algorithm. Spatial Clustering of Applications with Noise) which is effective for detecting clusters with arbitrary shapes and identifying noise in spatial-temporal data [7].

DBSCAN Implementation is carried out in the following stages:

a. Sampling result data (100,000 rows) is loaded from the pre -processed CSV file .
b. The features used for clustering are taken, including longitude , latitude , hour , and day .
c. Feature standardization is done using StandardScaler from sklearn.preprocessing to standardize the feature scale.
d. DBSCAN algorithm was applied with parameters eps =0.5 and min_samples =2, which were determined based on data exploration and related literature [9]. esp and min_samples parameter values greatly affects the results of the DBSCAN algorithm , especially in detecting spatial density patterns in standardized data . The value of esp = 0.5 reflects the neighborhood radius in the standardized feature space, while min_samples = 2 allows the identification of small clustering but is also more susceptible to noise . This approach is in accordance with recommendations from previous studies that suggest the use of [9],k- distance analysis plot for estimating eps values and paying attention to the characteristics of the data distribution. Although these parameters successfully produced meaningful cluster patterns in early experiments, adjusting the min_samples value to the range of 5-10 is also highly recommended on large-scale spatial datasets to improve stability and reduce the number of outliers[10].
e. clustering results (cluster labels) are added to the dataframe as the ' DBSCAN _label ' column.
f. The final results are saved in CSV format with proper decimal format for longitude and latitude.

2.2.2 K- Means Algorithm and Combination of Methods. As a comparison, the K- Means algorithm is also applied to the same data with systematic implementation stages. The selection of K-Means as a clustering algorithm is based on its ability to segment data efficiently and produce clusters that can be interpreted clearly. This approach is in line with the implementation of K- Means carried out by previous research for campus promotion strategies, where the K-Means algorithm successfully identified user segments with similar characteristics and provided actionable insights for business decision making [16].

The implementation of K- Means is carried out with the following stages:

a. sampling result data (100,000 rows) is loaded from the pre -processed CSV file .
b. The same features (longitude , latitude , hour , day) are used for clustering .
c. Feature standardization is done using StandardScaler . d. K-Means algorithm was applied with parameters n_clusters = 3, random_state = 42, and n_init = 'auto'. In the K- Means algorithm , the determination of the number cluster (k) is a key element in the effectiveness of data segmentation. For Twitter spatial-temporal data, an exploratory visual approach and spatial domain considerations were taken, namely dividing the United States into three main geographic zones: the West Coast, the Central Region, and the East Coast. The value of k = 3 was used in the implementation by considering the efficiency and interpretability of the results, and in line with the practices adopted by previous studies in geographic area segmentation studies. The [8].clustering process was carried out on data that had been standardized using StandardScaler to equalize the scale between longitude , latitude , hour , and day features .
d. clustering results (cluster labels) are added to the dataframe as the ' kmeans_label ' column.
e. The final results are saved in CSV format with proper decimal format.business interpretation .

Clustering results from K- Means are then compared with the DBSCAN results to evaluate the advantages and disadvantages of each algorithm in the context of Twitter geospatial data , as well as to identify digital activity patterns that can be used for business decision making.

2.3 Clustering Results into Dashboard Business intelligence . Clustering results are visualized and further analyzed using a BI dashboard (Power BI), to support spatial-temporal interpretation and business decision making. The integration process includes:

a. Export clustering results (DBSCAN_results_sample.csv and kmeans_results_sample.csv) to Power BI.
b. Creation of spatial visualization (map) that displays cluster distribution based on longitude and latitude .
c. Creating a temporal visualization in the form of a line chart (distribution of tweets per day) and bar chart (distribution of tweets per hour) for each cluster .
d. Making pies chart to show the proportion of tweets per cluster .
e. Added interactive filters (slicers) based on cluster , day, and hour for more dynamic data exploration.
f. Analysis of insights generated from visualizations, for example identification of digital activity hotspots and temporal patterns per cluster. Integration with Power BI enables interactive data exploration and transformation of clustering results into actionable business insights .

3 Results and Discussion

This section presents the results of the clustering process that has been carried out using the DBSCAN and K- Means algorithms on the Twitter Geospatial Data dataset . The results are presented in several sections that include analysis of spatial distribution, temporal patterns, and integration of the results into the dashboard. business intelligence.

3.1 Spatial Clustering Results.

3.1.1 Cluster Distribution DBSCAN. DBSCAN algorithm with the parameters eps = 0.5 and min_samples = 2 on 100,000 sample data produces an interesting clustering pattern. Spatial visualization of clustering results DBSCAN is shown in Figures 2 and 3, which show the distribution of clusters based on longitude and latitude coordinates.

From the visualization, it can be seen that DBSCAN is able to identify tweet distribution patterns that follow the geographic shape of the United States. Some of the main characteristics of the clustering results DBSCAN includes:

(1) Clusters are formed following population density patterns, with high concentrations on the east coast, west coast, and several large cities in the center of the country.
(2) The algorithm successfully identified various clusters with irregular shapes that follow natural geographic patterns.
(3) Different colors in the visualization indicate different clusters , with the color gradient from blue to green indicating variations in tweet density.

3.1.2 Cluster Distribution K- Means. K-Means algorithm with the parameter n_clusters = 3 results in the division of data into 3 main clusters . Spatial visualization of clustering results K- Means is shown in Figures 4 and 5.

From the visualization, it can be seen that K- Means divides the data into 3 main clusters with the following characteristics:

(1) Cluster 1 (green): Covers most of the west coast of the United States, concentrated around California, Oregon, and Washington.

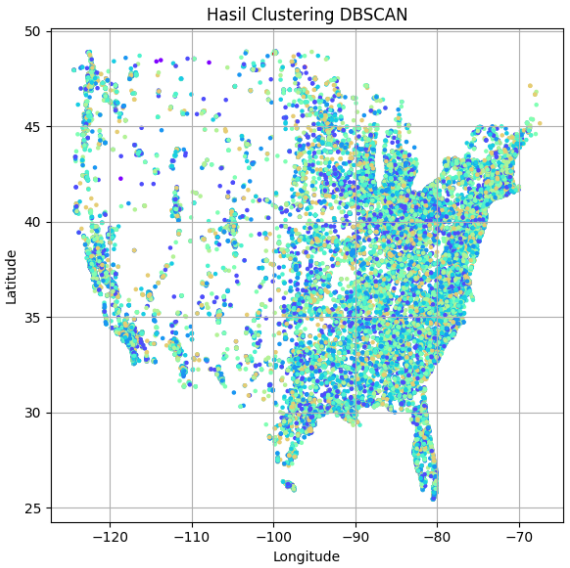

Fig. 2 Clustering Results DBSCAN showing the spatial distribution of tweets based on longitude and latitude coordinates (Python visualization)

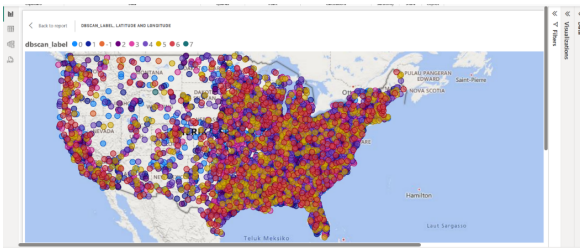

Fig. 3 Clustering Results DBSCAN showing the spatial distribution of tweets on an interactive map (Power BI visualization)

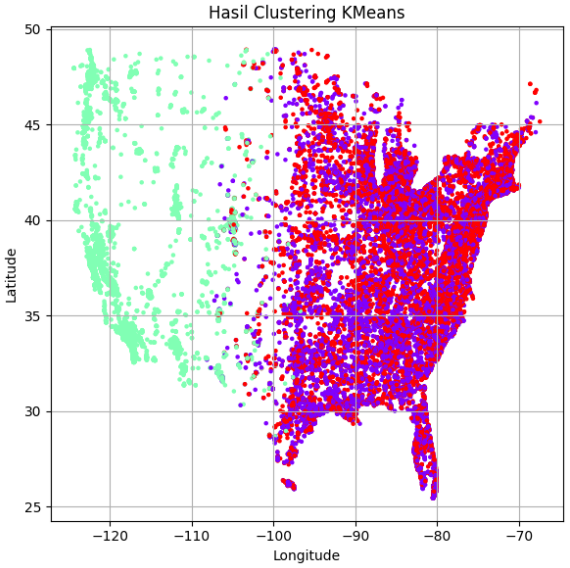

Fig. 4 Clustering Results K-Means showing the spatial distribution of tweets based on longitude and latitude coordinates (Python visualization)

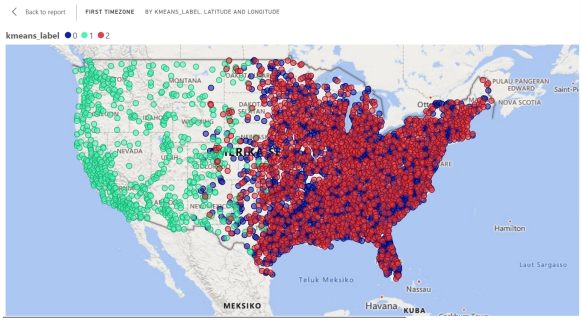

Fig. 5 Clustering Results K-Means showing the spatial distribution of tweets on an interactive map (Power BI visualization)

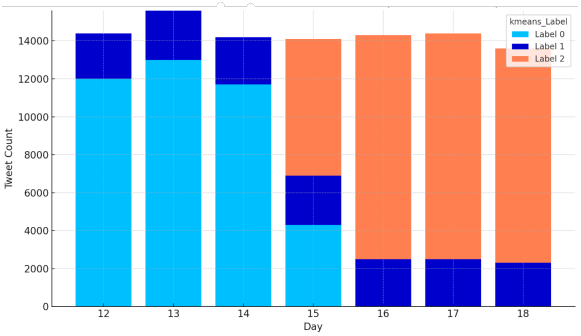

Fig. 6 Distribution of tweets by hour of the day

(2) Cluster 2 (blue): Covers the central region of the United States, including most of the states in the midwest and south .

(3) Cluster 3 (red): Covers the east coast of the United States, concentrated around New York , Boston, Washington DC, and Florida.

3.1.3 Spatial Clustering Results. spatial clustering results between DBSCAN and K- Means shows significant differences in terms of:

a. Cluster shape : DBSCAN is able to detect clusters with irregular shapes that follow geographic patterns and population density, while K- Means produces clusters with more regular shapes and is divided based on distance.
b. Noise handling : DBSCAN explicitly identifies noisy data , while K- Means allocates all data to one of three clusters .
c. Cluster size distribution : DBSCAN produces clusters of varying sizes based on data density, while K- Means tends to produce clusters of relatively balanced sizes.

3.2 Temporal Analysis. In addition to spatial analysis, this study also analyzes the temporal patterns of Twitter activity based on hours and days. Temporal features (hour and day) extracted from timestamps are used as part of the clustering process.

3.2.1 Tweet Distribution By Time. tweet distribution by time (hour) shows the pattern of Twitter user activity that varies throughout the day. Figure 6 shows the distribution of tweets by hour of the day for the entire dataset. From this visualization, it can be seen that :

a. Twitter activity peaks at 19:00-22:00 (local time), indicating high social media usage at night.
b. The lowest activity occurs at 03:00-05:00 (local time), which is in line with most people's sleep patterns.

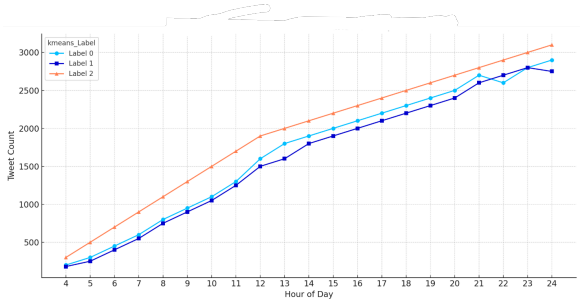

Fig. 7 Distribution of tweets by hour for each cluster . K-Means

c. There is a significant increase in activity between 07:00-09:00, indicating user activity in the morning before or during the commute to work.

3.2.2 Temporal Variation Between Clusters. Temporal analysis for each cluster generated by K- Means shows interesting differences in activity patterns. Figure 7 shows the distribution of tweets by hour for each cluster. K-Means.

From the visualization, it can be seen that: a. Cluster 1 (dark blue): Has a relatively low but consistent tweet volume throughout the day, with peak activity occurring at 20:00-23:00 local time. This cluster shows a more even activity pattern compared to the other clusters . b. Cluster 0 (light blue): Shows the highest tweet volume with peak activity at 19:00-22:00 local time. This cluster has a significant decrease in activity in the early morning hours (03:00-05:00). c. Cluster 2 (orange): Has a similar activity pattern to Cluster 0 but with a lower volume, with peak activity at 18:00-21:00 local time. The differences in activity patterns reflect the variations in time zones and lifestyles in different regions of the United States. Based on the geographic distribution seen in Figure 5, these differences correspond to the spatial location of the clusters , where Cluster 1 is dominant on the west coast, Cluster 0 in the mid-west, and Cluster 2 on the east coast of the United States.

3.3 Integration of Results into Dashboard Business intelligence.

3.3.1 Interactive Visualization with Power BI. Clustering results have been integrated into the dashboard Power BI to facilitate interactive exploration and analysis. Figure 8 shows a dashboard view that includes spatial and temporal visualizations of the clustering results.

Dashboard provides several interactive visualizations that are integrated with each other, including:

a. Interactive map showing the spatial distribution of clusters K- Means , with different colors for each cluster (blue for cluster 0, red for cluster 1, and orange for cluster 2). This map visualizes the geographic distribution of tweets across the United States.
b. Time series graph " Tweet Count by day and kmeans_label " which displays the temporal distribution of tweets per day for each cluster K- Means , allows analysis of weekly activity patterns.
c. Tweet "Bar chart Count by hour and kmeans_label " which displays the hourly distribution of tweets for each cluster , allowing identification of daily activity patterns.
d. Pie chart " Distribution of Tweets by Cluster " which displays the proportion of tweets per cluster K- Means , shows that Cluster 0 has 40.51% of the data, Cluster 2 has 42.14% of the data, and Cluster 1 has 17.36% of the data.

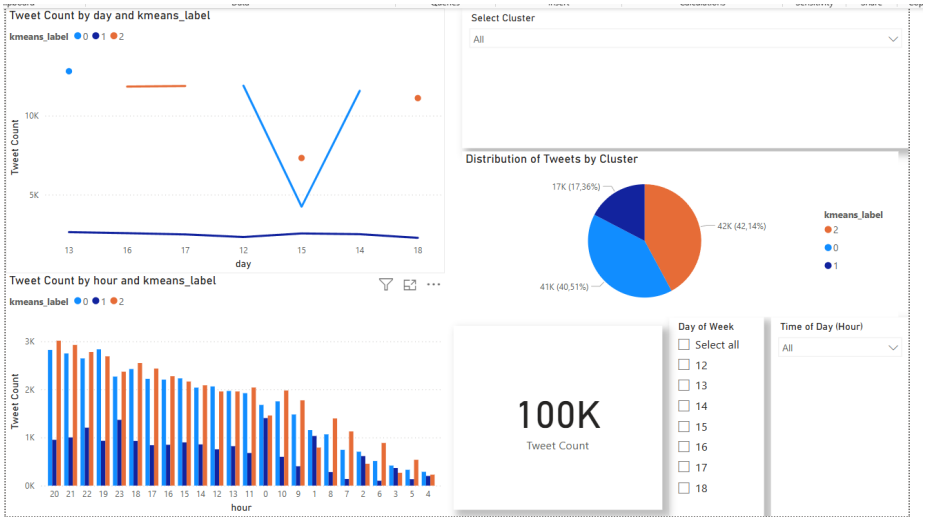

Fig. 8 Dashboard Business intelligence that displays clustering results K-Means

e. Interactive filter by cluster , day (Day of Week), and hours (Time of Day) for more dynamic and in-depth data exploration.

Although this visualization only shows the clustering results, K-Means , the same approach can be applied to DBSCAN results . Integration of spatial and temporal visualizations in one dashboard allows for comprehensive analysis of digital activity patterns based on location and time simultaneously, resulting in richer and more contextual business insights.

3.3.2 Insight from Clustering Results. BI dashboard exploration, several business insights that can be identified include:

a. Hotspots : Areas with the highest concentration of Twitter activity were identified in major cities such as New York , Los Angeles, Chicago, and San Francisco. These insights can be leveraged for digital marketing campaigns targeting areas with high engagement .
b. Optimal Times: Time periods with the highest user activity (19:00-22:00 local time) can be utilized for serving ads or promotional content to maximize reach and engagement .
c. Geographic Segmentation: Dividing the United States into 3 main clusters (West Coast, Central Region, and East Coast) can be used for marketing strategies tailored to the characteristics of each region.
d. Temporal Activity Patterns: Variations in activity patterns based on time and region can be leveraged for optimal content and campaign scheduling across time zones.

clustering results into BI dashboards enables decision makers to interactively explore data and identify actionable business insights without the need for technical expertise in data mining or machine learning. learning .

4 Conclusion

This research has successfully developed a spatiotemporal analysis pipeline for clustering digital activity locations based on Twitter geospatial data and integrated it into a dashboard. business intelligence . Through the implementation and comparison of DBSCAN and K- Means algorithms , several important conclusions were obtained. First, DBSCAN proved effective in detecting irregularly shaped clusters that follow natural geographic patterns and population density, while K-Means successfully divided the United States into three main clusters that reflect geographic differences (West Coast, Central Region, and East Coast). Second, temporal analysis showed that Twitter user activity patterns varied by time, with peak activity at 19:00-22:00 and lowest activity at 03:00-05:00, as well as variations in activity patterns between clusters that reflect differences in time zones and lifestyles.

clustering results into the dashboard business intelligence using Power BI enables interactive data exploration and generates actionable business insights , such as identifying digital activity hotspots for targeted marketing campaigns, determining optimal content delivery times, geographic segmentation for tailored marketing strategies, and temporal activity patterns for optimal campaign scheduling across time zones. This approach demonstrates the potential of spatiotemporal analytics in supporting data-driven business decision making.

clustering results of Twitter digital activity have important implications in the business context. intelligence , especially for institutions such as Bank Indonesia. By identifying locations with high digital activity density (hotspots), peak times of social interaction, and patterns of differences between regions, financial institutions can develop spatial monitoring systems for informal economic dynamics and population mobility. This supports the idea that spatial data-based BI can be a tool in formulating policies based on local and real- time data.For example, the concentration of digital activity in large cities at night can be interpreted as an indicator of informal sectoral economic activity that can be used in setting monetary communication strategies or targeting area-based incentive policies.

Although this study used a dataset from the United States, the methodology developed has the potential to be applied to the Indonesian context if local data is available. For further research, it is recommended to explore other combinations of clustering algorithms, integrate sentiment analysis and tweet content , and develop predictive models based on identified spatiotemporal patterns. In addition, the development of a more comprehensive BI dashboard with predictive analytics capabilities can also be a promising research direction.

This study has several limitations that need to be noted. First, the data used only comes from Twitter users who have activated geo-tagging , so it does not reflect the entire digital population. Second, the features used are limited to location and time, without considering the content or sentiment of the tweet . When the clustering evaluation method is not yet equipped with quantitative metrics such as Silhouette Score or Davies-Bouldin Index. In the future, research can be developed by combining text-based features using the Natural Language approach. Processing (NLP) , as well as applying deep learning techniques Autoencoder -based

clustering to handle large and complex spatial-temporal data.

References

[1] Karami, A., Lundy, M., Webb, F., and Dwivedi, Y. K., 2020, "Twitter and Research: A Systematic Literature Review through Text Mining," IEEE Access, **8**, pp. 67698–67717.

[2] Hu, T. et al., 2021, "Revealing public opinion towards covid-19 vaccines with twitter data in the united states: Spatiotemporal perspective," JMIR Publications Inc.

[3] Silva, R. A., Pires, J. M., Datia, N., Santos, M. Y., Martins, B., and Birra, F., 2019, "Visual analytics for spatiotemporal events," Multimedia Tools and Applications, **78**(23), pp. 32805–32847.

[4] Castro, G. M. M., Aitken, H. G. W., and Calvanapon, A. A., 2023, "Business intelligence Tools for a Digital Services Company in Peru, 2022," International Journal of Business Intelligence Research, **14**(1).

[5] Wark, J. D., 2022, "Power Up: Combining Behavior Monitoring Software with Business intelligence Tools to Enhance Proactive Animal Welfare Reporting," Animals, **12**(13).

[6] Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S., and Brisco, B., 2020, "Google Earth Engine for geo-big data applications: A meta-analysis and systematic review," Elsevier BV.

[7] Bushra, A. A. and Yi, G., 2021, "Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms," IEEE Access, **9**, pp. 87918–87935.

[8] Li, J., Zheng, A., Guo, W., Bandyopadhyay, N., Zhang, Y., and Wang, Q., 2023, "Urban flood risk assessment based on DBSCAN and K-Means clustering algorithm," Geomatics, Natural Hazards and Risk, **14**(1).

[9] Chen, S., Liu, X., Ma, J., Zhao, S., and Hou, X., 2019, "Parameter selection algorithm of DBSCAN based on K-means two classification algorithm," The Journal of Engineering, **2019**(23), pp. 8676–8679.

[10] An, X. et al., 2023, "STRP- DBSCAN: A Parallel DBSCAN Algorithm Based on Spatial-Temporal Random Partitioning for Clustering Trajectory Data," Applied Sciences (Switzerland), **13**(20).

[11] Li, P., Jiang, L., Zhang, S., and Jiang, X., 2022, "Demand Response Transit Scheduling Research Based on Urban and Rural Transportation Station Optimization," Sustainability (Switzerland), **14**(20).

[12] Madbouly, M. M., Darwish, S. M., Bagi, N. A., and Osman, M. A., 2022, "Clustering Big Data Based on Distributed Fuzzy K-Medoids: An Application to Geospatial Informatics," IEEE Access, **10**, pp. 20926–20936.

[13] Necochea-Chamorro, J. I. and Larrea-Goycochea, L., 2023, "Business intelligence Applied in the Corporate Sector: A Systematic Review," TEM Journal, **12**(4), pp. 2225–2234.

[14] E., G. Y. W. S. H. N. and Ma, P., 2015, "Twitter Geospatial Data," .

[15] Wahab, A. W., 2025, "Clustering of Book Reading Habits Among Students in Kendari City," JIKO (Journal of Informatics and Computers), **9**(1), p. 1.

[16] Latifah, U. W., Bahri, S., and Satriandhini, M., 2024, "Implementation Algorithm K-Means Clustering for IBISA Campus Promotion Strategy," JIKO (Journal Informatics and Computers), **8**(2), p. 292.