Jurnal Ekonomi Malaysia 59(2) 2025
http://dx.doi.org/10.17576/JEM-2025-5902-5

Modeling Intention and Frequency of Cashless Behavior: Integrating Big Data and Cluster-Structural Equations Modeling on Bank Customers

(*Pemodelan Niat dan Kekerapan Tingkah Laku Tanpa Tunai: Mengintegrasikan Data Raya dan Pemodelan Kluster-Persamaan Berstruktur terhadap Pelanggan Bank*)

Rita Alfin
University of Walisongo Gempol

Favian Deanova Atha Valentino
University of Brawijaya

Solimun
University of Brawijaya

Hanifa Sepriadi
University of Brawijaya

Adji Achmad Rinaldo Fernandes
University of Brawijaya

Fachira Haneinanda Junianto
University of Brawijaya

ABSTRACT

This study aims to determine and model the determinant factors that increase the intensity and frequency of use of cashless payments in the community among Bank customers in Indonesia. The novelty in this research is to combine big data methods (web scraping & LDA) and statistical modeling (cluster and SEM) in modeling the intensity and frequency of use of cashless society. This research uses research data obtained from questionnaires distributed using indicators obtained using big data methods and modeled using cluster-SEM. The results show that there are three groups of cashless payment usage levels, the three groups produce three SEM models where each group has different determinant factors in modeling the intensity and frequency of cashless usage. For the public to increase their sensitivity to the use of a cashless society because its development is very fast. For banks to maximize economic, payment, adoption, policy, and consumer focus variables which can increase the intensity and frequency of use of a cashless society.

Keywords: Big data; cashless society; clustering; structural equation modeling

ABSTRAK

Kajian ini bertujuan untuk mengenal pasti dan memodelkan faktor-faktor penentu yang meningkatkan intensiti dan kekerapan penggunaan pembayaran tanpa tunai dalam kalangan masyarakat dalam kalangan pelanggan bank di Indonesia. Kebaharuan dalam penyelidikan ini ialah penggabungan kaedah data raya (web scraping & LDA) dan pemodelan statistik (kluster dan SEM) dalam pemodelan intensiti dan kekerapan penggunaan tanpa tunai masyarakat. Penyelidikan ini menggunakan data kajian yang diperoleh daripada soal selidik yang diedarkan menggunakan penunjuk yang diperoleh melalui kaedah data raya dan dimodelkan menggunakan kluster-SEM. Hasil kajian menunjukkan terdapat tiga kumpulan tahap penggunaan pembayaran tanpa tunai, dan ketiga-tiga kumpulan tersebut menghasilkan tiga model SEM di mana setiap kumpulan mempunyai faktor penentu yang berbeza dalam memodelkan intensiti dan kekerapan penggunaan tanpa tunai. Bagi masyarakat, adalah penting untuk meningkatkan kepekaan terhadap penggunaan sistem tanpa tunai kerana perkembangannya adalah sangat pesat. Bagi pihak bank pula, mereka perlu mengoptimumkan pemboleh ubah berkaitan ekonomi, pembayaran, penerimaan, dasar serta tumpuan terhadap pengguna yang dapat meningkatkan intensiti dan kekerapan penggunaan sistem tanpa tunai.

JEL:C38, C63, C81, C150, C830, G4

INTRODUCTION

Path analysis is a statistical method used to analyze causal relationships between exogenous and endogenous variables in a model. This method has evolved into Structural Equation Modeling (SEM), which allows for the simultaneous analysis of complex relationships between latent and manifest variables (Hair et al. 2022). SEM consists of a structural model representing relationships between latent variables and a measurement model explaining the relationships between latent variables and manifest indicators (Kline 2023). SEM's ability to address multicollinearity and measurement error makes it one of the primary approaches in social and business research (Byrne 2023). However, when data exhibits heterogeneity or consists of multiple distinct groups, conventional SEM models may yield biased estimates.

To address this issue, the Cluster-SEM method has been introduced to identify differences in relationship patterns across groups (Wang et al. 2023). Cluster analysis aims to group objects based on similar characteristics, which can be performed using hierarchical or non-hierarchical methods (Fong et al. 2022). Non-hierarchical methods, such as K-Means and K-Medoid, are more frequently used in large-scale datasets due to their efficiency in handling data complexity (Zhou et al. 2024).

With the advancement of the digital era, big data-based analysis is increasingly being applied across various disciplines. Techniques such as web scraping and text mining enable the systematic exploration of large volumes of data (Han et al. 2023). In academic research, bibliometric analysis is used to identify patterns and trends based on metadata from scholarly publications (Jiang et al. 2024). Additionally, Latent Dirichlet Allocation (LDA) is widely applied in topic modeling to classify documents based on underlying topics (Chen et al. 2023).

This study offers a novel approach by integrating bibliometric analysis and LDA into the Cluster-SEM model to address challenges in heterogeneous data and identify key variables in various social phenomena, including cashless behavior. Cashless behavior refers to the increasing tendency of people to use non-cash transactions, a trend that has accelerated particularly after the COVID-19 pandemic (Gupta & Sharma 2023). This transformation is not only occurring in major cities but also in regions with previously limited digital infrastructure, making it crucial to understand the factors influencing its adoption (Bellaouar et al. 2023).

In the financial industry, understanding cashless behavior is essential for banks and financial service providers to develop more adaptive strategies. Several studies have highlighted key factors influencing the use of digital payment systems, such as perceived ease of use, trust, and transaction security (Oliveira et al. 2022; Sharma et al. 2023). However, limited research has integrated bibliometric analysis, LDA, and Cluster-SEM approaches to identify user segments and the factors influencing their intention and usage frequency.

Therefore, this research develops a Cluster-SEM model integrated with bibliometric analysis and LDA to identify cashless user segments and the factors influencing their intention and usage frequency. This approach contributes to the fields of data science, digital marketing, and finance, helping banks design more customer-centric services and promote financial inclusion in the digital era.

The main contributions of this research can be seen from three dimensions. First, from a theoretical perspective, this research strengthens and expands the theory of technology adoption by integrating variables represented through social media-based public opinion, thereby providing new empirical evidence in the context of a cashless society. Second, from a methodological perspective, the hybrid approach that combines big data analysis (LDA) and cluster-based structural modeling provides a more comprehensive method of analysis for understanding digital consumer behavior, which can be replicated in cross-disciplinary research. Third, from a practical perspective, the results of this study provide a deeper understanding of user segmentation and the factors that influence their behavior, which can be used as a basis for formulating policy strategies, service innovations, and strengthening the digital payment ecosystem.

LITERATURE REVIEW

LATENT DIRICHLET ALLOCATION (LDA)

Latent Dirichlet Allocation (LDA) is a topic modeling method used for analyzing large-scale text data. LDA, introduced by Blei et al. (2003), conceptualizes each document as a mixture of latent topics, where each topic is represented as a distribution of words. Recent studies (Li & Zhang 2024) emphasize that LDA has been extensively applied in text mining, natural language processing, and social media analysis due to its ability to extract meaningful patterns from large textual datasets. However, one major limitation of traditional LDA is its assumption of a fixed number of topics, which may not always reflect the underlying structure of dynamic data. Recent advancements (Chen et al. 2025) have integrated LDA with deep learning approaches to enhance its adaptability to varying data distributions.

Despite its widespread application, LDA has limitations in handling evolving topics over time and contextual dependencies between words. Existing studies often rely on static modeling approaches, which may not fully capture the dynamic nature of language. This study addresses this gap by integrating LDA with machine learning-driven clustering techniques to improve topic coherence and adaptiveness in analyzing unstructured text data.

CLUSTER ANALYSIS

Cluster analysis is a multivariate technique used to classify objects into distinct clusters based on shared characteristics. The fundamental goal of cluster analysis is to group objects with similar properties while maximizing the differences between clusters (Kembe & Onoja 2017). More recent research (Zhao et al. 2023) highlights that traditional clustering methods, such as k-means and hierarchical clustering, often face challenges in handling high-dimensional data and complex cluster structures. Advances in clustering methodologies, including density-based and hybrid clustering techniques (Chen & Wu 2025), have been proposed to address these limitations. Furthermore, the application of machine learning-driven clustering has been increasingly explored in fields like market segmentation, image recognition, and financial risk analysis.

Traditional clustering methods lack flexibility in handling complex, non-linear relationships within high-dimensional data. Many studies focus on predefined cluster numbers without adaptive mechanisms to detect natural groupings in the data. This study contributes by leveraging a hybrid clustering framework that integrates LDA-based topic modeling with machine learning clustering techniques to enhance cluster accuracy and interpretability.

STRUCTURAL EQUATION MODELING (SEM)

Structural Equation Modeling (SEM) is a comprehensive statistical approach that integrates factor analysis, path analysis, and regression modeling to examine relationships between latent and observed variables (Hox & Bechger 1998). Over the years, SEM has evolved with two primary approaches: covariance-based SEM and variance-based SEM. While covariance-based SEM requires strict assumptions such as multivariate normality and large sample sizes, variance-based SEM, particularly Partial Least Squares SEM (PLS-SEM), offers a more flexible alternative (Solimun et al. 2017). Recent studies (Martínez et al. 2023) highlight that PLS-SEM has gained prominence in business, marketing, and behavioral sciences due to its ability to handle complex models with formative indicators. Additionally, the development of non-recursive and non-linear SEM models (Zhang et al. 2025) has expanded the applicability of SEM in dynamic and non-traditional research settings.

Despite its robustness, SEM methods often assume linear relationships, limiting their applicability in capturing complex causal dependencies. Many existing models also fail to incorporate cluster-based segmentation, which could offer deeper insights into heterogeneous data structures. This study contributes by integrating SEM with clustering techniques to analyze structural relationships within different latent segments, providing a more nuanced understanding of variable interactions.

SIMULATION

Simulation is a widely used computational approach to model, analyze, and predict system behavior under various scenarios (Hoover & Perry 1989). Simulation techniques enable researchers to evaluate model adequacy, assess decision-making strategies, and optimize system performance (Kwon & Harrell 2004). Recent advancements (Singh et al. 2023) emphasize that simulation has become increasingly sophisticated, integrating artificial intelligence and Monte Carlo methods to improve accuracy and computational efficiency. In this study, simulation is utilized to explore different clustering parameters (e.g., distance metrics and k-values) to determine the optimal clustering approach. Furthermore, modern simulation models (Chen et al. 2025) incorporate probabilistic and agent-based modeling techniques to enhance predictive capabilities and decision-making processes.

Although simulation models have evolved significantly, many studies still lack real-world validation and adaptive learning mechanisms. Traditional simulation approaches often rely on static parameter settings, which may not reflect dynamic changes in data. This study contributes by incorporating AI-driven simulation modeling to dynamically optimize clustering and SEM parameters, ensuring more robust and context-aware decision-making processes.

RESEARCH METHODOLOGY

This study integrates a big data approach with statistical modeling to analyze the transition to a cashless society. In line with the research objectives, the methodology applied includes several main stages. First, a Systematic Literature Review (SLR) was conducted to identify the determining variables for the adoption of a cashless society. This SLR searched for reputable international publications through the Scopus, Web of Science, and Google Scholar databases using keywords such as cashless society, digital payment adoption, and financial technology. The results of this literature review were used as the basis for developing a conceptual framework and formulating hypotheses.

The next stage was big data processing to strengthen the identification of indicators. Secondary data was collected through web crawling techniques on the Twitter platform, focusing on conversations relevant to the use of non-cash transactions. The text data obtained from crawling then underwent preprocessing stages, such as data cleaning, normalization, and stop word removal. Next, topic analysis was performed using the Latent Dirichlet Allocation (LDA) method to identify latent topics that reflect public perceptions and opinions. The topics obtained were used as initial indicators for the research variables.

Based on the indicators produced, a questionnaire instrument was developed that included relevant variables from the SLR and LDA analysis results. This questionnaire used a five-point Likert scale to measure the respondents' level of agreement. The instrument was tested for validity and reliability through a pre-test before being distributed. Primary data was collected from 250 respondents who were bank customers in Indonesia, selected through purposive sampling to ensure diversity in demographic and economic backgrounds. Bank customers were selected as respondents based on their involvement in the use of digital financial services, making them relevant for examining the adoption of a cashless society.

Model testing was conducted using Structural Equation Modeling (SEM) to analyze the causal relationships between variables, both direct and indirect effects. This analysis involved testing data normality, validity, and construct reliability. The research hypotheses were formulated based on theories obtained from SLR and LDA analysis findings. Conceptually, the research model places SLR outcome variables (e.g., Perceived Ease of Use, Perceived Security, Digital Literacy) as exogenous variables that influence Intention to Use Cashless Society as a mediating variable. Furthermore, this intention influences Frequent Behavior to Use Cashless Society (actual behavior of using cashless transactions) as an endogenous variable.

Given data pairs $(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, y_{1i}, y_{2i})$ with; follows the SEM model (inner model) with the following equation:

$i = 1,2 \ldots n$

$$Y_i = \beta_0 + \sum_{l=1}^{5} \beta_l x_{li} + \varepsilon_i \tag{1}$$

If integrated with clusters, the equation will contain a dummy, so the equation becomes:

$$Y_i = \beta_0 + \sum_{l=1}^{p} \sum_{k=1}^{q} \sum_{m=1}^{r} \beta_l x_{li} + \beta_{p+k} D_{ki} + \beta_{q+m} x_{li} D_{ki} + \varepsilon_i \tag{2}$$

Where:
Y_i: endogenous latent variable at the i observation
β: coefficient of influence of latent variables
x: exogenous latent variable
D: dummy variable
ε: error

If written simply then,

$$\begin{aligned} Y_{1i} &= \beta_{01} + \beta_{11}X_{1i} + \beta_{21}X_{2i} + \beta_{31}X_{3i} + \beta_{41}X_{4i} + \beta_{51}X_{5i} + \beta_{61}D_{1i} + \beta_{71}D_{2i} + \beta_{81}D_{1i}X_{1i} + \\ &\quad \beta_{91}D_{1i}X_{2i} + \beta_{101}D_{1i}X_{3i} + \beta_{111}D_{1i}X_{4i} + \beta_{121}D_{1i}X_{5i} + \beta_{131}D_{2i}X_{1i} + \beta_{141}D_{2i}X_{2i} + \\ &\quad \beta_{151}D_{2i}X_{3i} + \beta_{161}D_{2i}X_{4i} + \beta_{171}D_{2i}X_{5i} + \varepsilon_{1i} \\ Y_{2i} &= \beta_{02} + \beta_{12}X_{1i} + \beta_{22}X_{2i} + \beta_{32}X_{3i} + \beta_{42}X_{4i} + \beta_{52}X_{5i} + \beta_{62}Y_{1i} + \beta_{72}D_{1i} + \beta_{82}D_{2i} + \\ &\quad \beta_{92}D_{1i}X_{1i} + \beta_{102}D_{1i}X_{2i} + \beta_{112}D_{1i}X_{3i} + \beta_{122}D_{1i}X_{4i} + \beta_{132}D_{1i}X_{5i} + \beta_{142}D_{1i}Y_{1i} + \\ &\quad \beta_{152}D_{2i}X_{1i} + \beta_{162}D_{2i}X_{2i} + \beta_{172}D_{2i}X_{3i} + \beta_{182}D_{2i}X_{4i} + \beta_{192}D_{2i}X_{5i} + \beta_{202}D_{2i}Y_{1i} + \varepsilon_{2i} \end{aligned} \tag{3}$$

If explained for each cluster then,

Low is when and, $D_1 = 0 \; and \; D_2 = 0$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{4}$$

Medium, namely when$D_1 = 0$and$D_2 = 1$,

$$y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_i + \varepsilon_i \tag{5}$$

High is when$D_1 = 1$and ,$D_2 = 0$

$$y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_i + \varepsilon_i \tag{6}$$

Given data pairs$(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, y_{1i}, y_{2i})$ with $i = 1, 2, \ldots, n$; follows the SEM (Outer model) model with the following equation:

$$Y_i = \lambda_0 + \sum_{l=1}^{5} \lambda_l x_{li} + \varepsilon_i \tag{7}$$

If integrated with the cluster, the equation will contain a dummy, so the equation changes to:

$$Y_i = \lambda_0 + \sum_{l=1}^{p}\sum_{k=1}^{q}\sum_{m=1}^{r} \lambda_l x_{li} + \lambda_{p+k} D_{ki} + \lambda_{q+m} x_{li} D_{ki} + \varepsilon_i \quad (8)$$

Where:
Y_i: endogenous variable at the i-th observation
λ: loading factor
x: exogenous variable
D: dummy variable
ε: error

If written simply then,

$$\begin{aligned} Y_{1i} = {} & \lambda_{01} + \lambda_{11}X_{1i} + \lambda_{21}X_{2i} + \lambda_{31}X_{3i} + \lambda_{41}X_{4i} + \lambda_{51}X_{5i} + \lambda_{61}D_{1i} + \lambda_{71}D_{2i} + \lambda_{81}D_{1i}X_{1i} + \\ & \lambda_{91}D_{1i}X_{2i} + \lambda_{101}D_{1i}X_{3i} + \lambda_{111}D_{1i}X_{4i} + \lambda_{121}D_{1i}X_{5i} + \lambda_{131}D_{2i}X_{1i} + \lambda_{141}D_{2i}X_{2i} + \\ & \lambda_{151}D_{2i}X_{3i} + \lambda_{161}D_{2i}X_{4i} + \lambda_{171}D_{2i}X_{5i} + \varepsilon_{1i} \end{aligned}$$

$$\begin{aligned} Y_{2i} = {} & \lambda_{02} + \lambda_{12}X_{1i} + \lambda_{22}X_{2i} + \lambda_{32}X_{3i} + \lambda_{42}X_{4i} + \lambda_{52}X_{5i} + \lambda_{62}Y_{1i} + \lambda_{72}D_{1i} + \lambda_{82}D_{2i} + \lambda_{92}D_{1i}X_{1i} + \\ & \lambda_{102}D_{1i}X_{2i} + \lambda_{112}D_{1i}X_{3i} + \lambda_{122}D_{1i}X_{4i} + \lambda_{132}D_{1i}X_{5i} + \lambda_{142}D_{1i}Y_{1i} + \lambda_{152}D_{2i}X_{1i} + \\ & \lambda_{162}D_{2i}X_{2i} + \lambda_{172}D_{2i}X_{3i} + \lambda_{182}D_{2i}X_{4i} + \lambda_{192}D_{2i}X_{5i} + \lambda_{202}D_{2i}Y_{1i} + \varepsilon_{2i} \end{aligned} \quad (9)$$

If explained for each cluster then,

Low is when and, $D_1 = 0 \; and \; D_2 = 0$

$$y_i = \lambda_0 + \lambda_1 x_i + \varepsilon_i \quad (10)$$

Medium, namely when $D_1 = 0$ and $D_2 = 1$,

$$y_i = (\lambda_0 + \lambda_2) + (\lambda_1 + \lambda_4)x_i + \varepsilon_i \quad (11)$$

High is when $D_1 = 1$ and $D_2 = 0$,

$$y_i = (\lambda_0 + \lambda_3) + (\lambda_1 + \lambda_5)x_i + \varepsilon_i \quad (12)$$

Based on this framework, the proposed research hypotheses are: (H_1) SLR outcome variables have a positive effect on Intention to Use Cashless Society; (H_2) Intention to Use Cashless Society has a positive effect on Frequent Behavior to Use Cashless Society; and (H_3) SLR outcome variables have an indirect effect on Frequent Behavior to Use Cashless Society through the mediation of Intention to Use Cashless Society. This approach provides a comprehensive understanding of the determinants of cashless society adoption in Indonesia by combining big data analysis and survey-based validation.

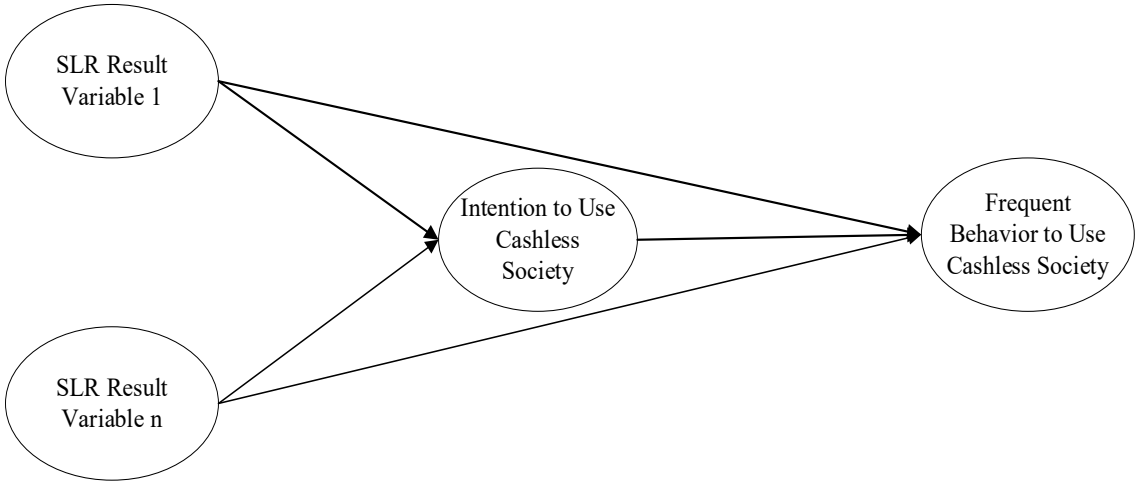

FIGURE 1. Research structural model

RESULTS AND DISCUSSION

This section presents research findings obtained through bibliometric analysis, big data processing, and statistical modeling using the Cluster-SEM approach. The results are systematically organized to address the research objectives, starting from literature mapping through Systematic Literature Review (SLR), Twitter data processing using the Latent Dirichlet Allocation (LDA) method, exploratory factor analysis, cluster formation, to structural model testing and inter-variable influence. The discussion in each section will interpret the analysis results based on relevant theories and empirical findings.

Initial search results for Scopus documents with the cashless society theme produced 260 documents. In the initial search stage, the publication year was not limited because the aim was to map research developments on the theme of a cashless society. Then the documents were filtered based on document type, namely articles, so that 145 documents were found. From the results of the documents found, a bibliometric analysis was carried out which contained several pieces of information including the author's name, keywords, language, year of publication, article title, number of citations, publisher's journal name, abstract, affiliation, and DOI.

Of the 145 documents selected as targets for analysis, document mapping was carried out using Voshviewer visualization. This visualization is carried out based on keywords in research articles. The visualization results show the relationship between the topic and other topics, especially research on a cashless society.

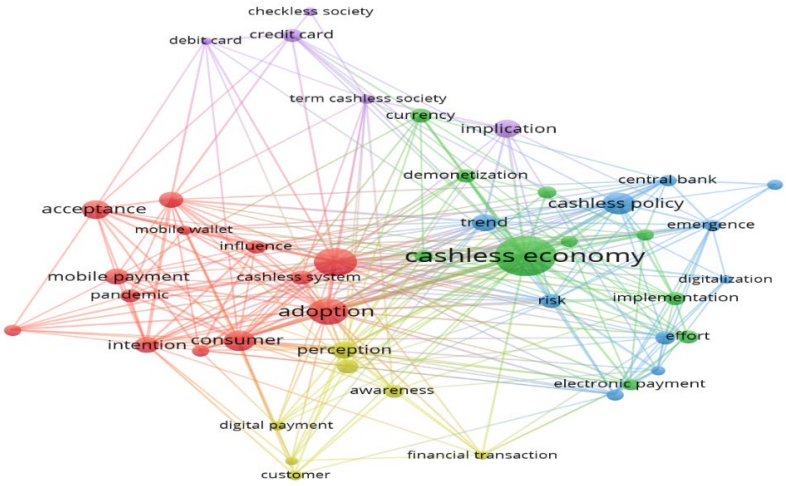

FIGURE 2. Visualization of SLR results

According to Yulianti (2022), determinant factors related to a theme can be obtained based on the frequency of keywords in the target document that influence the theme. The more often a keyword appears, the more likely it is that the keyword is related to the theme. From the results of the analysis and visualization, five words were obtained that had a high frequency so that five determinant factors were formed, namely economy, payment, adoption, policy, and consumer focus.

This research began by conducting web scrapping on Twitter using the keyword "cashless society", from the results of this web scrapping we obtained 1000 discussions about cashless society. The discussion collected through web scrapping is data pre-processed to remove unimportant words and convert the words into a corpus. The words that have become a corpus are used to calculate the topic coherence value. Topic coherence is the most optimal topic value formed based on existing words. In this research, the most optimal topics that can be formed based on topic coherence are 21 topics. After determining the optimal topic, latent dirichlet allocation is then continued, where the words are arranged into 21 topics, based on the weight of the words. Below is a worldcloud of words that have the highest word weight values.

FIGURE 3. LDA results wordcloud

The word cloud visualization shows a number of words that are closely related to the concept of a cashless society. The size of the words in the word cloud represents the level of connection, so that the larger the word, the higher its frequency of appearance and the stronger its relevance to the topic of a cashless society. The words formed based on LDA are then subjected to Explanatory Factor Analysis (EFA). EFA aims to determine the relationship between variables and identify factors in data patterns. The EFA results can be seen in the table below.

TABLE 1. Explanatory factor analysis results

No	Topic	Economy	Payment	Adoption	Consumer Focus	Intention To Use Cashless Society	Frequent Behavior To Use Cashless Society	Policy
1	Digitalization	0.639	-0.057	0.064	0.067	0.087	-0.049	0.106
2	Ease of Payment	-0.159	0.517	-0.012	-0.003	-0.062	-0.027	-0.092
3	Readiness	-0.186	0.146	0.623	0.152	-0.121	0.009	-0.197
4	Accuracy	-0.166	-0.002	0.670	0.024	-0.079	-0.014	0.181
5	Social	-0.101	0.194	0.054	-0.193	0.641	0.107	0.177
6	Culture	0.042	0.175	-0.084	-0.157	0.504	0.092	0.042
7	Transaction	0.522	-0.047	0.075	-0.107	-0.121	-0.130	-0.034
8	Card	-0.058	0.551	-0.100	-0.049	0.191	0.147	-0.057
9	Ease of Use	-0.080	0.125	-0.167	0.031	-0.088	0.603	0.061
10	Payment Method	-0.038	-0.089	-0.020	-0.193	-0.173	0.661	0.181
11	Habit	0.192	0.163	0.068	0.178	0.149	-0.089	0.607
12	System	0.548	0.180	-0.101	0.008	-0.035	0.118	-0.053
13	Believe	-0.040	-0.093	-0.146	0.686	0.159	-0.088	-0.118
14	Commitment	0.105	-0.097	-0.157	0.616	-0.188	0.006	-0.115
15	Collaborative	0.018	0.693	0.047	-0.163	0.019	0.126	0.082
16	Acceptance	-0.143	-0.130	0.677	-0.002	0.139	-0.008	-0.142
17	Personal	0.044	-0.116	-0.137	0.004	0.603	0.078	-0.016
18	Security	-0.197	-0.159	0.082	0.538	0.134	-0.104	0.017
19	Frequency of Use	0.068	-0.046	-0.161	-0.167	-0.090	0.560	-0.038
20	Need	0.156	0.193	0.152	-0.192	0.096	0.178	0.642
21	Contextual Factors	-0.090	0,000	0.179	-0.144	-0.027	-0.064	0.588

Based on Table 1, it can be seen that all topics/indicators have a relationship with the seven variables. The magnitude of the relationship between indicators and variables can be seen from the value of the loading factor for each indicator. The greater the value of the loading factor, the stronger the relationship between the indicator and the variable. The loading factor in EFA is a coefficient that measures the relationship between variables and factors/indicators resulting from LDA. The loading factor value can determine the indicator that has the most influence on a particular variable. Based on Table 1, it can be seen that each variable has three indicators with large loading factor values.

Based on the results of cluster validation calculations using a modified Dunn index, namely by calculating within and between cluster distances at all distances. Within is the distance from a point to the cluster center, while between is the distance from a point to the center of another cluster. The DBI Index (David Bouldin Index) is the result of dividing between within and between. The results obtained are as follows, the number of clusters formed is three using the Manhattan distance. This is in line with the results of the previous sentiment where there are three groups of understanding/use of a cashless society.

TABLE 2. Results of determining distance and number of clusters

Distance	Propose a New Validation Index			
	k=2	k=3	k=4	k=5
Euclidean	0.137	0.099	0.346	0.424
Manhattan	0.153	0.075	0.359	0.587
Mahalanobis	0.151	0.091	0.384	0.485

Based on the cluster validation results in the Table 2, it can be seen that the optimal number of clusters is three clusters using Manhattan distance. Of the clusters formed, there are three (namely low cashless society (n1=65), medium (n2=100), and high (n3=85)), then in the confirmation test, are there any differences between the three groups, this test uses MANOVA. MANOVA is a multivariate analysis to determine the differences between two or more groups using the variance calculation for each variable.

TABLE 3. MANOVA results

MANOVA	df1	df2	F	P-value
Variables	3	96	8,930	0,000

Based on the MANOVA results in the Table 3, it can be confirmed that the three groups have differences, namely low, medium, and high, so it can be concluded that the best cluster is with three groups with a distance from Manhattan. Next, a simulation was carried out using cluster validation development to test the consistency of DBI in determining the optimal number of clusters.

Simulations were carried out to determine the accuracy and consistency of DBI in determining the optimal number of clusters. In this study, three scenario criteria were used, namely number of samples, DBI value, and distance used, therefore in this study, there were 81 simulation scenarios. The summary results of the simulation can be seen in Table 4.

TABLE 4. Simulation results

	n (samples)			DBI			Distance		
	50	100	200	Low	Medium	High	Eulidean	Manhattan	Mahalanobis
DBI	0.039	0.042	0.044	0.045	0.044	0.037	0.043	0.041	0.042
Ratio	2,652	2.51	2,385	3,601	2,444	1,502	2,636	2,702	2,209

Based on Table 4, it can be seen that the number of samples does not affect the clustering value or accuracy because the three simulated sample sizes produce almost the same DBI and ratio values. So it is concluded that the number of samples does not affect clustering. The high and low DBI values in the simulation are inversely proportional to the DBI values obtained after clustering, so this can validate that DBI can be used as a value to measure the accuracy of a cluster analysis. This also makes DBI more precise in validating cluster results because if DBI is high then the DBI value will be smaller with a small ratio.

The most appropriate distance to use in clustering is Manhattan because it has the lowest DBI value with the highest ratio, even though the distances do not have very different if values, the Manhattan distance has the smallest value of the other three distances so it can be concluded that the best distance in clustering is to use Manhattan distance. Based on the results of cluster analysis using Manhattan distance, three optimal clusters can be formed from the data above, so to continue with SEM, structural equation modeling integration is carried out with three clusters as in the lemma below.

The assumption test that must be met in SEM is linearity between the variables used. If the assumptions are met then the approach used is parametric, but if the assumptions are not met then modeling will be carried out using a semiparametric or non-parametric approach. One method for testing the linearity assumption is the Regression Specification Error Test or RESET which was first introduced in 1969 by Ramsey. The general model used to describe the relationship between exogenous and endogenous variables can be seen in equations (13) and (14).

$$Y_{1i} = \beta_0 + \beta_1 X_{1i} + \varepsilon_{1i} \tag{13}$$

$$Y_{2i} = \beta_0 + \beta_1 X_{1i} + \beta_2 Y_{1i} + \varepsilon_{2i} \tag{14}$$

RESET linearity testing uses the following hypothesis.

$H_0: \beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+m}$

H_1: There is at least one difference, $\beta_{p+j} j = 1,2, \ldots, m$

TABLE 5. Ramsey reset test results

Relationships	P-value				Results
	Overall	Group 1	Group 2	Group 3	
Economy → Intention To Use Cashless Society	0.408	0.619	0.519	0.762	Linear Relationships
Payment → Intention To Use Cashless Society	0.161	0.121	0.754	0.198	Linear Relationships
Adoption → Intention To Use Cashless Society	0.534	0.275	0.099	0.627	Linear Relationships
Policy → Intention To Use Cashless Society	0.207	0.134	0.510	0.211	Linear Relationships
Consumer Focus → Intention To Use Cashless Society	0.098	0.844	0.580	0.288	Linear Relationships
Economy → Frequent Behavior To Use Cashless Society	0.784	0.332	0.473	0.103	Linear Relationships
Payment → Frequent Behavior To Use Cashless Society	0.782	0.245	0.707	0.283	Linear Relationships
Adoption → Frequent Behavior To Use Cashless Society	0.663	0.522	0.606	0.618	Linear Relationships
Policy → Frequent Behavior To Use Cashless Society	0.255	0.746	0.803	0.148	Linear Relationships
Consumer Focus → Frequent Behavior To Use Cashless Society	0.417	0.747	0.150	0.363	Linear Relationships
Intention To Use Cashless Society → Frequent Behavior To Use Cashless Society	0.550	0.299	0.316	0.635	Linear Relationships

Based on the linearity test using the Ramsey RESET method in Table 5, the analysis results show that all direct relationships between exogenous latent variables and endogenous latent variables have a linear pattern in each cluster, whether cluster 1, cluster 2, or cluster 3. Thus, it can be concluded that the linearity assumption is fulfilled. Therefore, SEM modeling can be performed using the WarpPLS approach to estimate the relationship between exogenous and endogenous variables more accurately.

TABLE 6. Summary of distance simulations on SEM results

Distance	DBI	R-Square
Euclidean	0.099	0.877
Manhattan	0.075	0.931
Mahalanobis	0.091	0.876

Based on the distance simulation in SEM as shown in Table 6, the R-square value was obtained which was quite high, namely the lowest at 87.7% and the highest at 93.1%. This shows that 3 clusters are the optimal number that can be formed, and IF as cluster validation is very accurate in determining the optimal number of clusters so that the distance used produces a high square. Manhattan is the best distance because it has the highest R-square value, so it can be concluded that Manhattan distance is better than Euclidean and Mahalanobis in determining clustering.

The outer model analysis algorithm is the process of calculating latent variable data sourced from indicator data. Testing the outer model hypothesis was carried out using the t-test by paying attention to the p-value of each outer loading and outer weight. Models with reflective properties have outer loading values, while models with formative properties have outer weight values. This research examines the reflective properties so that the value measured is the outer loading value.

TABLE 7. Measurement model results

Variables	Indicators	Loading Factor Group 1 (Low)		Group 2 (Currently)		Group 3 (Tall)		P-value Dif
Economy (X1)	Digitalization	0.767	*	0.647	*	0.628	*	0.020
	Transaction	0.731	*	0.633	*	0.563	*	0.043
	System	0.690	*	0.658	*	0.654	*	0.032
Payment (X2)	Ease of Payment	0.685	*	0.506	*	0.615	*	0.042
	Card	0.653	*	0.580	*	0.755	*	0.035
	Collaborative	0.650	*	0.796	*	0.632	*	0.037
Adoption (X3)	Readiness	0.615	*	0.617	*	0.549	*	0.025
	Acceptance	0.715	*	0.695	*	0.686	*	0.037
	Perception	0.522	*	0.732	*	0.703	*	0.040
Policy (X4)	Security	0.625	*	0.585	*	0.511	*	0.022
	Believe	0.532	*	0.729	*	0.666	*	0.034
	Commitment	0.703	*	0.769	*	0.579	*	0.044
Consumer Focus (X5)	Personal	0.622	*	0.784	*	0.678	*	0.037
	Social	0.677	*	0.768	*	0.623	*	0.036
	Culture	0.657	*	0.624	*	0.636	*	0.027
Intention to use Cashless Society (Y1)	Payment Method	0.726	*	0.710	*	0.574	*	0.046
	Frequency of Use	0.717	*	0.689	*	0.686	*	0.042
	Ease of Use	0.774	*	0.761	*	0.780	*	0.035
Frequent Behavior to Use Cashless Society (Y2)	Need	0.728	*	0.788	*	0.585	*	0.030
	Contextual Factors	0.741	*	0.558	*	0.731	*	0.036
	Habit	0.749	*	0.761	*	0.521	*	0.041

* significant at 5% level of significance (P<0.05)

Based on the Table 7 above, it can be seen that the economy has digitalization as the strongest indicator which reflects cashless behavior in the low group, the system in the medium group, and the system in the high group. Overall, the economy significantly influences cashless society behavior at all levels. Payment has an indicator of ease of payment in the low group, collaboration in the medium group, and cards in the high group which influences the behavior of a cashless society. Adoption has acceptance in the low group and perception in the low and high groups in influencing the level of use of a cashless society. The policy has a customer commitment to using a cashless society in the low and medium groups, believing it is the most influential indicator in the high group.

Consumer focus is influenced by personal and social factors in the low, medium, and high groups in increasing consumer focus when using a cashless society. Customers' desire to use a cashless society is influenced by the ease of using cardless payments. This applies to the low, medium, and high groups. A person's desire to use a cashless society is influenced by the person's needs and habits in using a cashless society. This applies to low, medium, and high groups.

Inner relations or inner models are specifications of the relationships between latent variables. Latent variables and indicators or what are often called manifest variables can be standardized without losing their general characteristics. In this study, the inner model hypothesis was tested using the t-test with a real level of 5% for direct and indirect effects.

TABLE 8. Results of testing direct and indirect effects

Relationships	Coefficient Group 1 (Low)	Group 2 (Currently)	Group 3 (Tall)	P value Group 1 (Low)	Group 2 (Currently)	Group 3 (Tall)	P-value Dif
X1->Y1	0.279	0.253	0.418	0.005*	0.011*	0,000*	0.033
X2->Y1	0.368	0.321	0.389	0.007*	0.027*	0,000*	0.123
X3->Y1	0.218	0.265	0.342	0.029*	0.008*	0.001*	0.021
X4->Y1	0.257	0.219	0.466	0.010*	0.029*	0,000*	0.020
X5->Y1	0.214	0.208	0.386	0.032*	0.037*	0,000*	0.033
X1->Y2	0.117	0.276	0.473	0.241	0.006*	0,000*	0.024
X2->Y2	0.257	0.132	0.333	0.010*	0.188	0.001*	0.031
X3->Y2	0.080	0.082	0.095	0.423	0.414	0.343	0.150
X4->Y2	0.229	0.052	0.330	0.022*	0.601	0.001*	0.049
X5->Y2	0.081	0.221	0.309	0.420	0.027*	0.002*	0.023
Y1->Y2	0.280	0.331	0.253	0.005*	0.001*	0.011*	0.222

X1->Y1->Y2	0.078	0.084	0.106	0.009*	0.005*	0,000*	0.021
X2->Y1->Y2	0.103	0.106	0.098	0.001*	0,000*	0.001*	0.250
X3->Y1->Y2	0.061	0.087	0.087	0.041*	0.004*	0.004*	0.282
X4->Y1->Y2	0.072	0.072	0.118	0.016*	0.016*	0,000*	0.049
X5->Y1->Y2	0.060	0.069	0.098	0.046*	0.022*	0.001*	0.129

* significant at 5% level of significance (P<0,05)

Based on the Table 8, it can be seen that economy, adoption, payment, policy, and consumer focus directly have a significant influence on the intention to use a cashless society, where for the three groups the p values are smaller than 0.05 so it can be concluded that all variables have a direct influence on the level of desire to use a cashless society in the low, medium and high groups. The economy in the low group has a p-value greater than 0.05 so the economy in the low group does not have a significant influence on frequent behavior to use a cashless society, this is the same as payment where in the medium group it does not have a significant influence on frequent behavior to use cashless society. Adoption does not have a direct influence on the frequent behavior of using cashless society in each group, as well as policy and consumer focus.

All economic, payment, adoption, policy, and consumer focus variables have a significant influence on frequent behavior to use a cashless society through intention to use a cashless society in all groups, whether low, medium, or high, have a significant effect on increasing the frequency of use of cashless society in society. The shift towards non-cash transactions became more evident during the COVID-19 pandemic, as people sought alternatives to reduce the risk of virus transmission through physical money, credit cards, or direct hand-to-hand transactions. In response, the World Health Organization (WHO) recommended the adoption of contactless payments, commonly referred to as cashless transactions. The primary goal of cashless payment systems is to minimize physical interactions by utilizing digital wallets and electronic transactions (Katon & Yuniati 2020).

Twitter, a widely used social media platform, serves as a valuable source of public opinion and the latest trends. Users freely share their thoughts and discussions on various topics, including the cashless society. Given this, the study employs web scraping techniques to gather public perceptions of cashless transactions from Twitter. These insights are then incorporated into the research as an essential data source for analysis.

This study applies Latent Dirichlet Allocation (LDA) for indicator mining, using the Gibbs Sampling algorithm to categorize words into relevant topics. The optimal number of topics in LDA is determined through topic coherence, which evaluates the likelihood of topic formation based on word groupings. The topics with the highest coherence values are selected as the most representative ones for this study.

The research variables were derived through a systematic literature review (SLR) combined with visualization using VOSviewer. The analysis of 145 academic papers related to the cashless society resulted in five key variables: economy, payment, adoption, policy, and consumer focus. These five dimensions frequently appear in studies related to digital financial behavior.

Using web scraping data, LDA identified 21 optimal topics, which were later classified into seven research variables. Exploratory Factor Analysis (EFA) was applied to align the identified topics with the research variables, ensuring that each topic was assigned to the most relevant category based on the highest factor loading. The final indicators were then used to design a questionnaire, which was distributed among bank customers in Indonesia as the study's sample population.

Indonesian bank customers were categorized into three distinct groups based on their level of cashless society engagement: low, medium, and high. The segmentation was performed using a non-hierarchical clustering approach, considering factors such as economy, payment, adoption, policy, consumer focus, intention to use a cashless society, and frequent cashless behavior. A cluster validity test confirmed that the optimal number of clusters was three, aligning with the initial assumption that people adopt cashless payments at varying levels.

The first cluster represents individuals with low engagement in cashless transactions, using them only when necessary. These individuals exhibit little enthusiasm toward digital payments. The most influential indicators in this group include digitalization, ease of payment, acceptance, commitment, social factors, ease of use, and habitual behavior. The measurement of the inner model suggests that all exogenous variables influence the intention to use cashless payments. Specifically, frequent cashless behavior in this group is significantly driven by payment convenience and policy effectiveness. Moreover, indirect effects show that increased exposure to digital payments can enhance the frequency of cashless transactions over time.

The second group consists of individuals with moderate cashless adoption, balancing between cash and non-cash payments. They demonstrate flexibility in their choice of payment methods. The strongest indicators in this cluster include system reliability, collaboration, perception, commitment, personal factors, user convenience, and consumer needs. The inner model analysis confirms that all exogenous variables influence the intention to use a cashless society in this group. Furthermore, frequent cashless behavior is significantly affected by the adoption of a non-cash economy and consumer focus, indicating that as non-cash lifestyles become more prevalent, the frequency of cashless transactions will also rise.

The third cluster includes individuals with a high level of cashless usage, relying almost entirely on digital payments with minimal use of cash. The most influential indicators in this group include transaction efficiency, card usage, perceptions, trust, personal factors, ease of use, and contextual influences. Inner model measurements confirm that all exogenous variables significantly impact their intention to use a cashless society. Moreover, frequent cashless behavior is strongly influenced by economic factors, payment efficiency, policies, and consumer focus. The indirect effect analysis suggests that increasing digital payment adoption enhances both the frequency and consistency of cashless behavior in this group.

CONCLUSION

This study aims to model the intensity and frequency of non-cash transaction usage by integrating a big data analysis approach and advanced statistical modeling based on Cluster-SEM. Through web scraping techniques on Twitter social media and topic modeling using Latent Dirichlet Allocation (LDA), 21 optimal topics were obtained to serve as research indicators. These indicators were then used in the development of a measurement instrument combined with survey data from 250 bank customer respondents in Indonesia. The cluster analysis results grouped respondents into three main categories, namely low, medium, and high adoption rates, which showed differences in behavioral characteristics in utilizing non-cash transactions.

The results of the study confirm that economic factors, payment systems, adoption rates, policies, and consumer focus play a significant role in shaping the intention to use non-cash transactions, which in turn influences actual behavior in the use of digital payments. These findings indicate that non-cash behavior is complex and multidimensional, where the direct and indirect effects of exogenous variables show a strong mediating role of intention (intention to use). This means that before actual behavior is formed, users must have sufficient confidence regarding the convenience, security, and benefits offered by digital payment systems.

Theoretically, this study makes a significant contribution to strengthening the Technology Acceptance Model (TAM) and the Diffusion of Innovation theory in the context of digital transformation, particularly in terms of people's financial behavior. Both theories are proven to be relevant in explaining the role of perceptions of convenience, security, and trust in shaping the intention and behavior of using non-cash payment technologies. This study also expands the application of the theory by incorporating social and economic dimensions obtained from social media-based public opinion analysis, thus offering a more comprehensive understanding of the determinants of technology adoption.

Methodologically, this study provides a new approach by combining big data-based text analysis and structural modeling. The integration of topic modeling methods through LDA with Cluster-SEM produces a more detailed mapping of user behavior heterogeneity, which is difficult to obtain through conventional survey approaches. The identification of three user clusters (low, medium, and high) provides empirical evidence that technology adoption is not uniform across the population but varies according to individual characteristics and perceptions of technology. This confirms the importance of a segmentation-based approach in digital behavior studies.

Overall, the results of this study conclude that the transition to a cashless society is not merely a technological phenomenon, but a complex socio-economic process influenced by the interaction between payment infrastructure, regulatory policies, economic factors, and individual perceptions. The intention to use cashless transactions acts as a mediator connecting these factors with actual behavior, which means that the success of digital transformation depends on efforts to create an ecosystem that not only provides technology but also builds trust, literacy, and user comfort. These findings prove that the integration of big data analysis with statistical modeling can provide a richer and more accurate picture of the dynamics of digital payment technology adoption in the information-based economy era.

REFERENCES

Bellaouar, Ahmed, Mohamed, Sami, & Karim, Youssef. 2023. Understanding cashless behavior in emerging markets: Challenges and opportunities. *Journal of Financial Technology* 15(3): 45-62.

Blei, David M., Ng, Andrew Y. & Jordan, Michael I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Byrne, Barbara M. 2023. *Structural Equation Modeling with AMOS: Basic Concepts, Applications and Programming* (4th ed.). Routledge.

Chen, Xiaoming, Wang, Hui. & Li, Zhen. 2023. Topic modeling in big data analytics: Applications and advancements. *Data Science Review* 10(2): 112-130.

Chen, Yuan. & Wu, Zhi. 2025. Advances in hybrid clustering techniques for high-dimensional data analysis. *Journal of Artificial Intelligence Research* 12(1): 55-78.

Fong, Simon, Lee, Patrick, & Tan, William. 2022. Cluster analysis methods: A comparative study of hierarchical and non-hierarchical approaches. *Journal of Data Analytic* 8(4): 75-92.

Gupta, Sandeep, Sharma & Rajesh. 2023. The impact of COVID-19 on cashless transactions: A behavioral analysis. *International Journal of Digital Finance* 18(1): 30-50.

Hair, Joseph F., Black, William C., Babin, Barry J. & Anderson, Rolph E. 2022. *Multivariate Data Analysis* (8th ed.). Cengage Learning.

Han, Jian, Zhou, Lei, & Chen, Fei. 2023. Web scraping and text mining techniques for big data analysis. *Journal of Computational Social Science* 12(1): 55-78.

Hoover, Donald N. & Perry, James. 1989. Simulation modeling for system analysis. *Journal of Computational Simulation* 5(2): 110-130.

Hox, Joop J. & Bechger, Timo M. 1998. An introduction to structural equation modeling. *Educational Psychology Review* 10(3): 353-379.

Jiang, Liang, Xu, Mei, & Zhou, Ping. 2024. Bibliometric analysis in academic research: Methods and applications. *Research Metrics Journal* 9(2): 90-110.
Kembe, Michael M. & Onoja, Samuel. 2017. A review of clustering techniques in multivariate analysis. *International Journal of Statistics and Data Science* 6(3): 112-130.
Kline, Rex B. 2023. *Principles and Practice of Structural Equation Modeling* (5th ed.). Guilford Press.
Kwon, Jin, & Harrell, Paul. 2004. Optimizing decision-making strategies through simulation modeling techniques. *Operations Research Review* 14(2): 210-225.
Li, Fang, & Zhang, Yong. 2024. Applications of LDA in text mining and natural language processing. *Computational Linguistics Review* 19(1): 75-92.
Martínez, Juan, González, Ricardo, & Pérez, Ana. 2023. Partial least squares structural equation modeling: Recent advances and applications. *Journal of Business Research* 32(4): 150-168.
Oliveira, Tiago, Thomas, Michael & Ferreira, João. 2022. Determinants of digital payment adoption: An empirical study. *Journal of Financial Innovation* 7(3): 25-40.
Ramsey, James B. 1969. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society, Series B* 31(2): 350-371.
Sharma, Prateek, Mehta, Alok, & Verma, Sunil. 2023. Trust and security in digital payment systems: A user perspective. *Electronic Commerce Research* 20(4): 120-135.
Singh, Rajesh, Kumar, Anil, & Patel, Vikram. 2023. Advancements in simulation modeling and decision-making processes. *Journal of Computational Research* 15(1): 99-120.
Solimun, Fernandes, Adji Achmad Rinaldo & Raharjo, Heny. 2017. PLS-SEM: An alternative method for multivariate analysis in behavioral research. *International Journal of Quantitative Methods* 5(2): 45-62.
Wang, Ying, Liu, Ming & Chen, Tao. 2023. Cluster-SEM: A novel approach to handling heterogeneity in structural equation modeling. *Journal of Statistical Methods* 15(2): 200-225.
Yulianti, Lestari. 2022. Keyword analysis in bibliometric research: A systematic approach. *Journal of Research Methods* 9(3): 78-99.
Zhang, Hao, Li, Wei & Chen, Xiaojun. 2025. Expanding the applications of non-recursive SEM models in social science research. *Behavioral Statistics Review* 18(2): 140-165.
Zhao, Rong, Lin, Qiang & Hu, Wei. 2023. Machine learning-driven clustering techniques: A review of current trends and applications. *Data Mining and Knowledge Discovery* 29(3): 200-225.
Zhou, Hong, Wang, Yan, & Liu, Jie. 2024. Efficient clustering algorithms for large-scale datasets: Advances and applications. *Data Mining Journal* 11(1): 65-89.

Rita Alfin*
Faculty of Social Sciences and Humanities
University of Walisongo Gempol
Pasuruan, Jawa Timur, INDONESIA.
Email: rita.alfin@stiegwalisongo.ac.id

Favian Deanova Atha Valentino
Faculty of Mathematics and Natural Sciences
University of Brawijaya
Jalan Veteran, Lowokwaru
Kota Malang, Jawa Timur, 65145, INDONESIA.
Email: Faviandeanova@student.ub.ac.id

Solimun
Faculty of Mathematics and Natural Sciences
University of Brawijaya
Jalan Veteran, Lowokwaru,
Kota Malang, Jawa Timur, 65145, INDONESIA.
Email: solimun@ub.ac.id

Hanifa Sepriadi
Faculty of Mathematics and Natural Sciences
University of Brawijaya
Jalan Veteran, Lowokwaru,
Kota Malang, Jawa Timur, 65145, INDONESIA.
Email: Sepriadi1412@student.ub.ac.id

Adji Achmad Rinaldo Fernandes
Faculty of Mathematics and Natural Sciences
University of Brawijaya
Jalan Veteran, Lowokwaru,
Kota Malang, Jawa Timur, 65145, INDONESIA.
Email: Fernandes@ub.ac.id

Fachira Haneinanda Junianto
Faculty of Mathematics and Natural Sciences
University of Brawijaya
Jalan Veteran, Lowokwaru,
Kota Malang, Jawa Timur, 65145, INDONESIA.
Email: fachiraneinaj@student.ub.ac.id

* Corresponding author