SAGA: Journal of Technology and Information Systems
Vol 4, Issue 2, May 2026, Page 599-606
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i2.686

# Implementation of K-Means Algorithm to Classify Instagram Reels and Carousel Content Performance Based

Shelomitha Trinitia Wowor[1], Christa Gabriella Putri Tumbol[2], Jimmy H. Moedjahedy[3*], Green Arther Sandag[4]

Faculty of Computer Science, Universitas Klabat, Minahasa Utara, Indonesia

Author Email: s22210267@student.unklab.ac.id[1], s22210400@student.unklab.ac.id[2], jimmy@unklab.ac.id[3], greensandag@unklab.ac.id[4]

**Abstract.** With the increasing popularity of digital marketing, Instagram became one of the top platforms where the audience can be reached. It is important to gain an insight into the performance of different types of contents to ensure that the marketing efforts bear fruit. This research will apply the K-Means algorithm to classify Instagram Reels and Carousel contents based on performance by taking into account such factors as likes, comments, shares, and saves. For the purposes of the study, the data were collected from a variety of accounts both personal and of a business nature. The number of clusters was defined by the Elbow Method, after which they were categorized depending on their performance such as high, medium, and low. The results indicate that the classification based on performance provided by the K-Means algorithm can provide insights into marketing practices on Instagram. Consequently, the present research will contribute to the development of digital marketing studies, particularly in the area of content analysis, within the field of data mining.

**Keywords:** Instagram Reels, Carousel, K-Means Clustering, Engagement Metrics, Performance Classification

## 1 Introduction

As a result of constant development of information and communication technology, the Internet has become the leading means of communication, having substituted the traditional systems for digitally-based ones [1]. It can be supplemented by smartphones that allow using various types of services (social networks and messengers), available regardless of time and place [2]. Thanks to the popularity of numerous social media websites since the beginning of the new millennium [3], information dissemination has become cheaper, faster, and more convenient. However, this process is also connected with significant changes to behavior patterns and culture in society [4].

In the contemporary era, social media plays an important role in people's lives as the source of relevant information and news [5]. Among popular platforms for information dissemination, Instagram stands out thanks to the ability of presenting information in the form of images or videos [6, 7]. The capabilities provided by such tools as Feeds, Stories, and Reels make Instagram an effective tool for information dissemination [8].

The same trends are observed in local media where the functioning of social media is illustrated by an ABC account. Observations made throughout the internship demonstrate that, despite regular publications of news, the level of user engagement remains variable. This phenomenon is typical for managing social media channels where determining the effectiveness of content creation is necessary to ensure active audience participation [9].

When it comes to evaluating the performance of content creators, one faces certain difficulties when it comes to measuring their effectiveness. Evaluating each publication individually might be considered easy; however, when analyzing all posts created by an individual in a given period, this activity requires more effort. Standard features offered by Instagram Insights provide a creator with only the statistics or basic graphs but not a grouping by common traits [10]. At the same time, no automatic topic-based classification exists.

At the early stage of data cleaning and description, Excel is used. However, Excel was not designed for automatic data grouping, and, hence, grouping is based on assumptions made manually [11]. In order to solve these problems, K-Means clustering algorithms have been suggested. K-Means clustering is an unsupervised machine learning algorithm that groups numerical data based on similarity characteristics, where prior labeling is not needed [12]. This way, an objective and automated evaluation can be performed [13]. Thus, the main purpose of this research is implementing K-Means clustering algorithms in order to objectively group Instagram content performance, specifically Reels and Carousels, by their engagement metrics. It is hoped that this research will allow conducting a more objective content analysis and will uncover characteristics of audience engagement. By

SAGA: Journal of Technology and Information Systems
Vol 4, Issue 2, May 2026, Page 599-606
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i2.686

using a more objective approach, one does not have to perform manual analysis of posts one by one. Thus, it is aimed at using K-Means clustering algorithm in order to group Instagram content performance (Reels and Carousels), taking into account like rate, comment rate, view rate, share rate, save rate, repost rate, and reach.

## 2 Methods

The research methodology follows the methodology of Knowledge Discovery in Databases (KDD), starting from the detection of the technical problems associated with content performance evaluation. In general, Knowledge Discovery in Databases (KDD) is an analytical technique used to extract valuable knowledge and patterns from raw data.

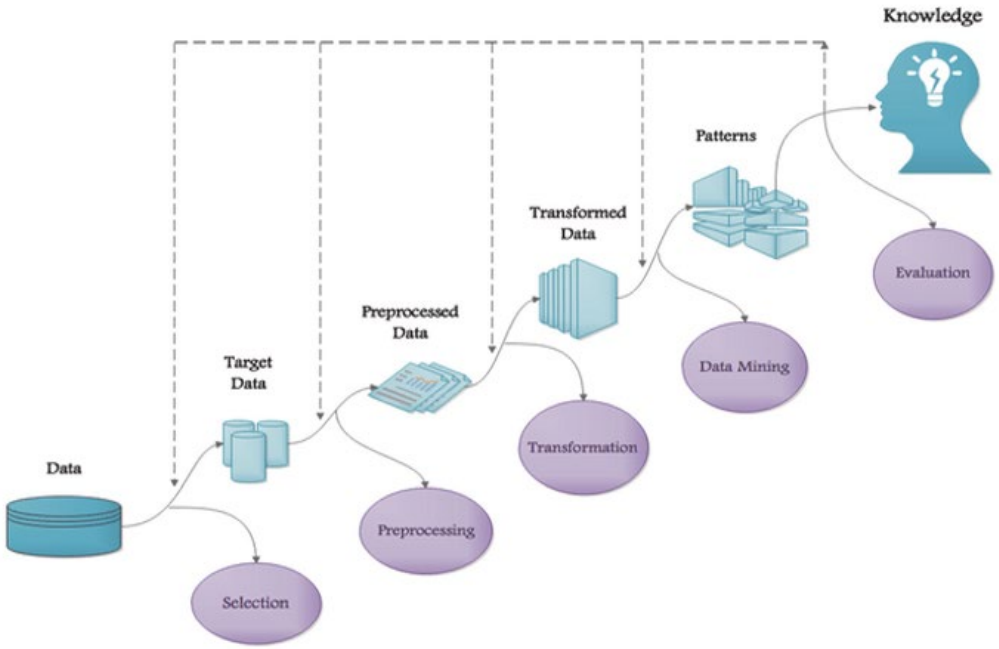


**Figure 1.** Knowledge Discovery in Databases (KDD) framework

The research methodology uses the KDD methodology, consisting of several phases: data selection, data pre-processing, data transformation, data mining, and interpretation/evaluation. Figure 1 shows the conceptual flow of the KDD process utilized in this research project.

### 2.1 Data Selection

The data selection process consists of selecting data relevant for research purposes from a large set of data. The selection criteria involve: selecting content uploaded from August 11, 2025 to December 21, 2025; filtering out Reels and Carousel format; excluding all non-original contents; and making sure that each piece of content selected has full interaction metrics.

### 2.2 Data Preprocessing

The data pre-processing stage includes various techniques applied to the dataset to ensure high-quality data for further analysis. This stage implies performing such actions as: exclusion of duplications, filling up missing data or gaps in datasets, data consistency checking, and data accuracy confirmation.

### 2.3 Data Transformation

At the data transformation stage, numerical features of interest undergo normalization using Min-Max scaling. It is performed to equalize ranges of different value variables' ranges and avoid any one variable dominating distance calculations in the K-means algorithm. The Min-Max scaling method allows normalizing variables in the range from 0 to 1 based on min-max values for each particular variable.

SAGA: Journal of Technology and Information Systems
Vol 4, Issue 2, May 2026, Page 599-606
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i2.686

### 2.4 Data Mining

During the data mining process, K-means algorithm is used for clustering the data based on similar characteristics by calculating data distances. An appropriate number of clusters is chosen following the Elbow Method that considers inertia values for a range of possible numbers of clusters. Then the algorithm identifies the elbow point, where a sharp decrease occurs. The steps of the clustering algorithm consist of: determining a possible number of clusters; using Elbow Method for obtaining the optimal number of clusters K; choosing initial centroids for clusters; calculating the Euclidean distance from each point to a center; assigning points to the closest centroid; calculating new centroids of the cluster as the mean of all points belonging to a cluster; and repeating the procedure for several iterations until centers' positions stop changing. The reason why K-Means was selected over other possible approaches includes the following. Hierarchical Clustering is too computationally costly on such large datasets and is not efficient enough to scale well. For DBSCAN one needs to specify its density parameters – epsilon and MinPts which are prone to the great variability observed in social media interactions. Gaussian Mixture Model would provide a better opportunity for differentiating clusters' shapes, but more rigorous assumptions about distributions and more computing power are needed; for this exploratory study the simplicity of K-Means has been preferred. The assumption made for K-Means about spherical shape of clusters was verified through PCA scatter plot and calculation of variance within clusters. The obtained clusters have shown a sufficient level of compactness and clear distinction, which validates the suitability of the K-Means approach for this particular dataset.

### 2.5 Interpretation / Evaluation

The last phase involves clustering assessment and pattern interpretation based on mean interaction values among the clusters' properties. Additionally, the distribution of content topics within each cluster will be analyzed to detect trends in the interactions' performance concerning different topics.

## 3 Result and Discussion

### 3.1 Data Selection and Preprocessing

The selected research dataset was specifically created based on Instagram performance metrics between August 11, 2025, and December 21, 2025. Specifically, Reels and Carousels were prioritized in the process, and such numerical interaction values were chosen: Likes, Comments, Shares, Saves, Reposts, Account Reach (AR), and Views. Within the process of preprocessing, data from several monthly CSV documents were unified and preprocessed. This stage implied the standardization of column names, the elimination of unnecessary summary rows, and topic filtering to only include such topics: Entertainment, National, Local News (Berita Daerah), and Economy. As for missing values in interaction features, these were replaced with zeros to avoid calculation mistakes. Thus, the preprocessed data set included 546 observations.

### 3.2 Data Transformation

The data transformation involved Min–Max scaling of interaction values to range from 0.0 to 1.0. It is important to mention this stage because K-Means algorithm is distance-based, and thus, without normalization, such metrics as Views, which have much larger values than, for example, Comments, would prevail in determining cluster centers. These seven measures have been chosen since they represent the entire suite of quantitative engagement metrics that can be obtained through Instagram Insights for the analyzed account, each measure representing a unique aspect of the user audience behavior (passive approval – Likes, participation in discussion – Comments, active sharing – Shares and Reposts, intentional bookmarking – Saves, and algorithmic amplification – Views and Account Reach). The Pearson correlation matrix was constructed using the raw data before proceeding with clustering. It was found that there are high correlations between Views and Account Reach ($r = 0.91$), as well as between Likes and Views ($r = 0.74$), meaning that there may be overlap. However, despite this, all variables will be used since normalization would ensure that they have equal influence in clustering regardless of their scales. It must be noted that this study, based on data from a single Instagram account over the course of only four months, is particular to the characteristics of the audience of that account and to the circumstances surrounding the creation of its Reels and Carousels.

SAGA: Journal of Technology and Information Systems
Vol 4, Issue 2, May 2026, Page 599-606
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i2.686

### 3.3 Data Mining and Optimal Cluster Selection

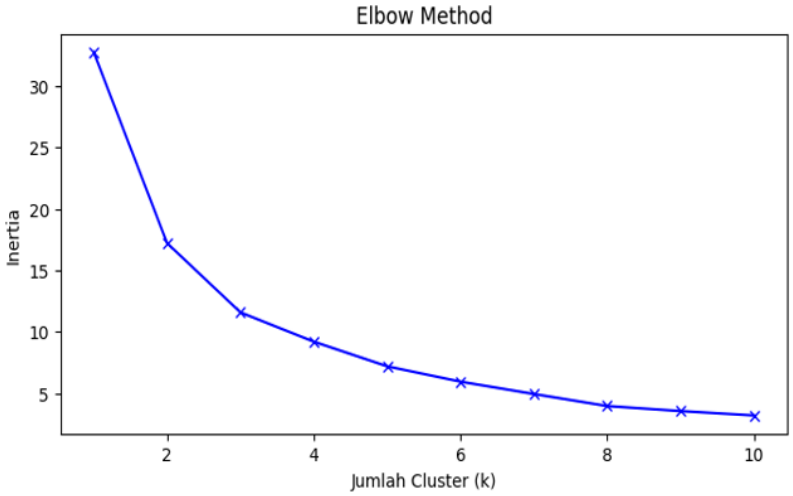


**Figure 2.** Elbow Method

The Elbow method was used to determine the best number of clusters (k). Through analyzing inertia for various k, a clear elbow can be seen at k=3, as seen on Figure 2. When applying the K-Means model with k=3, three clusters emerged based on the data analyzed. The clusters were labeled by their account reach (AR) in the form of clusters with the highest account reach, high account reach, and low account reach. In order to determine cluster quality in addition to the Elbow Method's inertia criteria, the Silhouette Score and Davies-Bouldin Index (DBI) were calculated for k = 3. The former measures how much every observation within the dataset is similar to its cluster compared to other clusters, with +1 values being a sign of good clustering, whereas the latter is a measure of the average similarity between clusters, with low scores denoting high separation. Overall, these metrics have helped confirm the appropriateness of k = 3 since they show that the obtained clusters are sufficiently separated from one another and consist of coherent observations, thus validating the Elbow Method result. Concerning the criteria for labeling the clusters, Account Reach (AR) was selected as the main sorting factor as this measure encompasses the total number of unique accounts that see the post—thus reflecting algorithm behavior related to the amplification of content, exactly as expected. In other words, Account Reach encapsulates the results of algorithm activity, namely likes, saves, shares, and comments, and is therefore used for labeling purposes.

### 3.4 Interpretation and Evaluation

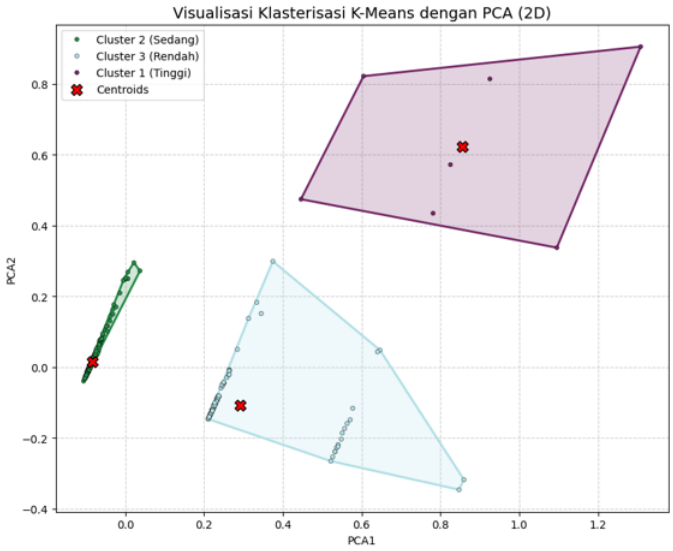


**Figure 3.** Clustering Visualization with PCA

SAGA: Journal of Technology and Information Systems
Vol 4, Issue 2, May 2026, Page 599-606
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i2.686

K-Means Clusters Visualized with PCA: The effectiveness of discrimination between clusters was evaluated through the use of Principal Component Analysis (PCA), whereby it reduced the data to two dimensions. Based on Figure 3, there is non-overlapping convex hulls indicating effective discrimination between clusters with High, Medium, and Low performance. Figure 3 shows the clusters mapping on a 2-dimensional basis with the use of PCA. Euclidean Distance Analysis: This metric assesses how much each data point conforms to its cluster through Euclidean distance. The PCA visualizing technique has been used in this case for exploration purposes to offer an easy visual representation of clusters in two dimensions, and not for validating the clusters. PCA maps high dimensional vectors into components which have maximum variability; therefore, the two-dimensional visualization will likely not maintain all pairwise distances from the initial seven dimensions. In view of the foregoing, it should be noted that the main criterion of cluster validity will be quantitative measures provided earlier (Silhouette Score and Davies-Bouldin Index) in addition to Euclidean Distance calculation with regard to each centroid. The PCA diagram has been added as supplementary visualization for understanding.

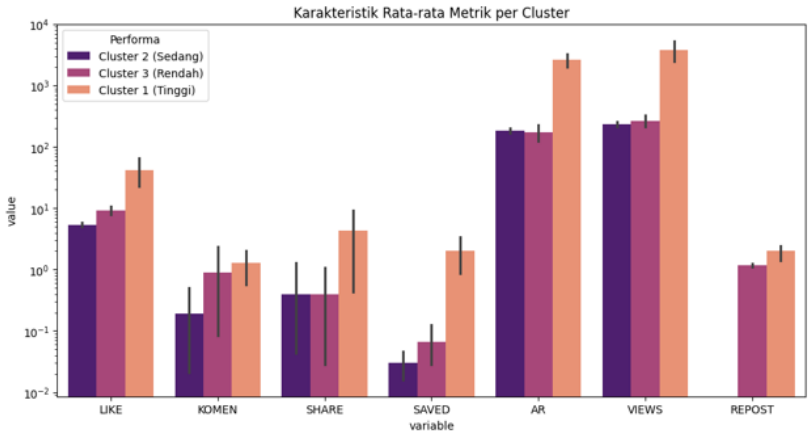


**Figure 4.** Metric Characteristics

A small distance to a cluster centroid denotes great compliance with that cluster. With the use of mathematical means, each Instagram post will be categorized into a suitable performance category, as seen on Figure 4. Metric Characteristics: Since there is a huge disparity in magnitude across the different metrics, a logarithm base 10 was adopted to display the average metrics in each cluster.

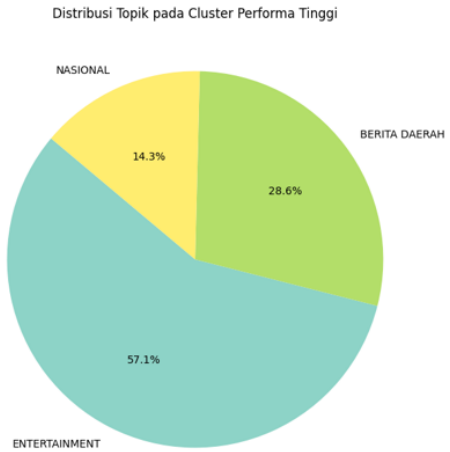


**Figure 5.** High Performance Cluster 1

Cluster 1 (High) performed better than any other clusters in the metrics with a notable highlight in Views and Shares. Figure 4 demonstrates the Interaction Metric Characteristics for Each Cluster. Topic Distribution Analysis: High Performance (Cluster 1) entails mainly Entertainment (57.1%) and Local News (28.6%). The posts are in the Reels format.

SAGA: Journal of Technology and Information Systems
Vol 4, Issue 2, May 2026, Page 599-606
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v4i2.686

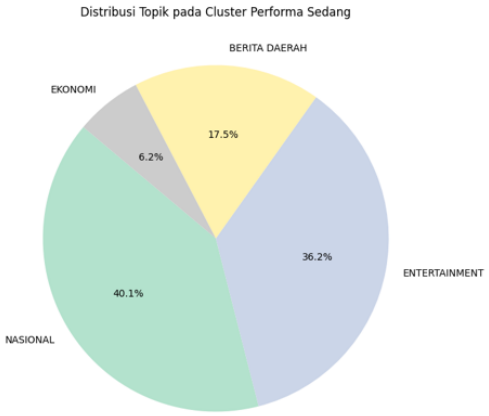


**Figure 6.** Medium Performance Cluster 2

Medium Performance (Cluster 2): Covers a broader range of topics with National (40.1%) and Entertainment (36.2%) being the most frequent.

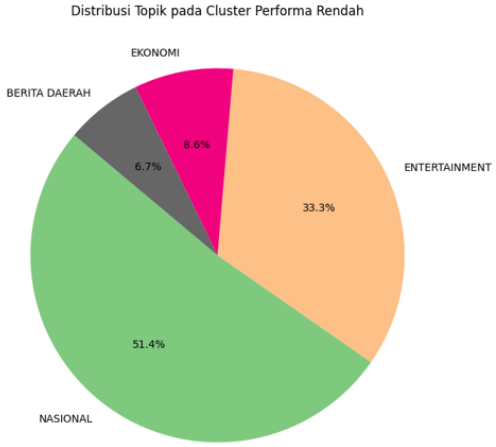


**Figure 7.** Low Performance Cluster 3

### 3.5 Extracted Knowledge

These cluster findings lead to tangible conclusions regarding the strategy that should be applied to the content produced. First, the High cluster reveals that the Reels related to the theme of Entertainment work best at generating engagement. On the other hand, the Low cluster signifies that specific themes and/or formats such as Carousels require reassessment to move to the next level. It must be noted that the differences in observed performances cannot be solely explained by the type of topic or type of content. The application of K-means clustering for this project means that posts are classified based on the collective performance pattern with regard to all seven metrics mentioned. Variables related to the publication date and audience behavioral tendencies were not considered when performing cluster analysis and were not isolated as factors affecting the performance. In light of this information, the high proportion of Reels associated with the category of Entertainment in Cluster 1 can be interpreted as a correlation between the two variables instead of causality. Further research should involve taking publication date into consideration as another variable and employing the method of comparisons with regard to topics posted as Reels or Carousels on the same date.

## 4 Conclusion

Moreover, the application of K-Means clustering algorithm within the KDD methodology helps overcome the problems connected with Instagram post performance assessment. As discussed above in the introduction, a purely manual assessment of dozens of posts proves to be an ineffective and biased process. This paper proves that through the use of the K-Means algorithm one can divide the content into three clusters that represent high, medium, and low performances. Thus, Cluster 1 (High) is characterized by dominance of the Reels video format

and entertainment topics while Cluster 2 (Medium) represents a high volume of moderately performing posts. Lastly, Cluster 3 (Low) indicates the necessity of strategic analysis of the carousel formats and some national topics which demonstrate a relatively low level of performance. Therefore, these findings prove to be useful for content creators as the basis for choosing the right topic to write about. As for future work, it may be suggested to expand the data sample period in order to take into account seasonality changes. Moreover, a sentiment analysis of user comments might provide additional information about the reasons behind high engagement of certain posts. First, in connection with the algorithms' assumptions, K-Means was chosen considering its efficiency, interpretability for practitioners, and applicability to numerical engagement metrics since the distance-based grouping is aligned with the goal of the paper, i.e., clustering of posts based on their performance. In addition, although the cluster assumption does not hold perfectly in the considered case, the preliminary PCA visualization together with analysis of the Euclidean distance between cluster centers demonstrates that all three clusters are quite compact and well-separated. Further studies might focus on the testing of alternative methodologies, such as HC, DBSCAN, and GMM, to reveal if non-spherical clusters or those based on density differ substantially in terms of clustering results. Second, concerning variable selection, the seven metrics (Likes, Comments, Shares, Saves, Reposts, Reach, and Views) were selected since they are all native signals provided by Instagram Insights. Moreover, they cover the spectrum of interaction signals, both active and passive ones. As shown by the preliminary Pearson correlation analysis, some metrics are highly correlated (Views and Reach). Nevertheless, all variables have been included since removal of correlated metrics leads to a decrease in the quality of content performance representation, while normalizing and using distances in K-Means reduce the risk of domination of any variable. Lastly, the scope of the analyzed data should be discussed. First of all, the dataset includes only 546 posts published by one regional media account (ABC) during approximately four months. Hence, the obtained results should be seen as specific to the characteristics of this particular account in terms of demographics of followers, content posted and frequency, and other aspects. In the given context, the aim of this work can be considered achieved due to the exploratory nature of the paper, i.e., providing the ABC team with a decision support tool.

## Acknowledgement

The authors would like to express their deepest gratitude to the Faculty of Computer Science, Universitas Klabat, for creating a proper atmosphere for conducting scientific research. We are very thankful to the management and social media department of the ABC account for providing us with data needed for our investigation. Additionally, we would like to thank our tutors and colleagues for the constructive comments and proofreading provided during the creation of this article.

## References

[1] R. Haholongan *dkk.*, “Pemanfaatan Internet Sebagai Media Pembelajaran: Transisi Dari Sistem Konvensional Ke Sistem Digital Dalam Kegiatan Belajar Pada Siswa SMP Pembangunan Jakarta Timur,” *Jurnal Pengabdian Masyarakat Ilmu Terapan*, vol. 6, no. 1, hlm. 7–12, Apr 2024. [Online]. Tersedia pada: https://jpmit.uho.ac.id

[2] L. Nitami, “Perkembangan Media Sosial Terhadap Perubahan Sosial Masyarakat Di Indonesia Tahun 2000-Sekarang”, *KAMACA*, vol. 11, no. 2, pp. 69-74, Dec. 2023.

[3] C. Noventa, I. Soraya, dan A. Muntazah, “Pemanfaatan Media Sosial Instagram BuddyKu Sebagai Sarana Informasi Terkini,” *JKOMDIS: Jurnal Ilmu Komunikasi Dan Media Sosial*, vol. 3, no. 3, hlm. 626–635, Sep 2023, doi: 10.47233/jkomdis.v3i3.1124.

[4] M. I. Mansiz dan Z. Fatah, “Pengelompokan Pengguna Media Sosial Berdasarkan Pola Interaksi Menggunakan K-Means,” hlm. 388–397, Nov 2024, doi: 10.59435/gjmi.v2i11.1100.

[5] A. Ahmed and A. S. Imran, "The role of large language models in UI/UX design: A systematic literature review," 2025, arXiv:2507.04469. [Online]. Available: https://arxiv.org/abs/2507.04469.

[6] W. P. Dananjaya, G. H. Prathama, dan K. Darmaastawan, “User-Centered Design Approach in Developing User Interface and User Experience of Sculptify Mobile Application,” *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 6, no. 3, hlm. 1089–1097, 2024.

[7] N. R. Wiwesa, “User Interface dan User Experience untuk Mengelola Kepuasan Pelanggan,” *J. Sos. Hum. Terap.*, vol. 3, no. 2, hlm. 18–19, 2021.

[8] D. Wojtal and P. Powroźnik, “Analysis of the impact of selected user interface elements on its usability”, J. Comput. Sci. Inst., vol. 35, pp. 113–120, Jun. 2025.

[9] R. H. S. Mezan El-Khaeri Kesuma, *Design Thinking UI/UX Teori dan Praktik*. Deli Serdang, Sumatera Utara: *PT. Mifandi Mandiri Digital,* 2024.

[10] B. A. . Pratama, U. . Proboyekti, and K. . Wijana, "Penerapan Metode User Centered Design (UCD) Dalam Pembangunan Layanan Online Jual Beli Barang Bekas", *JUTEI,* vol. 4, no. 1, pp. 33–43, Jul. 2021., doi: 10.21460/jutei.2020.41.192.

[11] Y. V. Akay, A. J. Santoso, and F. L. R. S. ahayu, "Metode User Centered Design (UCD) Dalam Perancangan Sistem Informasi Geografis Pemetaan Tindak Kriminalitas (Studi Kasus : Kota Manado)", *Prosiding Seminar Nasional ReTII*, Jan. 2017.

[12] H. Sulastri, R. N. Shofa, A. U. Rahayu, dan N. Hiron, Implementation of User Center Design (UCD) in Achieving Design by Focusing on End Users in the Caribi Mobile Application. *Atlantis Press International BV*, 2023. doi: 10.2991/978-94-6463-180-7.

[13] N. W. Beben Sutara, "Application Of User Centered Design (Ucd) Method For Developing User Interface And User Experience In The Kaia-Pay Application," vol. 1, no. 1, pp. 1–13, 2024. doi: 10.69933/jocsit.v1i1.2

[14] Ronni Sahat Hutabarat and Ketut Sudaryana, "User-Centered Design pada User Interface (UI) / User Experience (UX) Prototyping Aplikasi E-Commerce", *JPTIS*, vol. 2, no. 4, pp. 89–99, Dec. 2024.