Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 3, No 2, March 2024
Hal 208−214
DOI 10.47065/comforch.v3i2.1207

Utilizing K-Means Clustering to Understanding Audience Interest in SEO-Optimized Media Content

Erlin Windia Ambarsari[1,*], Dedin Fathudin[2], Gravita Alfiani[3]

[1,*]Faculty of Engineering and Computer Science, Informatics Engineering, Universitas Indraprasta PGRI, Jakarta, Indonesia
[2]Faculty of Computer Science, Informatics Engineering, Universitas Pamulang, Tangerang Selatan, Indonesia
[3]Director, Media Have Fun, Tangerang Selatan, Indonesia
Email: [1,*]erlinunindra@gmail.com, [2]dosen00398@unpam.ac.id, [3]gravita@mediahavefun.com
Email Penulis Korespondensi: erlinunindra@gmail.com

Abstract−This study observes k-means clustering for segmenting SEO data to understand audience interests, identifying the elbow method as crucial for determining the optimal number of clusters. It highlights notable differences in content engagement across clusters, emphasizing the need for refined SEO strategies and a deeper understanding of audience segmentation. Based on the purification algorithm, we obtained a random sample of 327 entries for traffic queries and a random selection of 100 entries for the landing pages. Then, we got four maximum clusters from the elbow method, which is divided by K-mean clustering into four characteristics. Cluster 1 (red) shows low engagement, suggesting an interest in niche content. Cluster 2 (Green) demonstrates a discrepancy between clicks and impressions, indicating highly relevant content to its audience. Cluster 3 (Blue) is marked by significant interest, as evidenced by solid impressions, but lacks corresponding engagement through clicks, suggesting a passive interest. Finally, Cluster 4 (Yellow) exhibits high clicks and impressions, signifying content with broad appeal and popularity among the audience. We concluded from the results of K-Mean clustering that this methodology provides a strong foundation for enhancing content strategies despite challenges like SEO's dynamic nature and data reliance. Future research suggestions include cross-platform data integration, longitudinal studies, sentiment analysis, content experimentation, user experience (UX) focus, and monitoring algorithm updates to develop more adaptive content and SEO strategies aligned with changing audience behaviors.

Keywords: K-Means; Elbow; SEO; Audience

1. INTRODUCTION

Tourism significantly contributes to a country's foreign exchange earnings. However, the global spread of COVID-19 severely restricted international and domestic travel, impacting these revenues. Consequently, this led to a significant stagnation in monetary circulation within affected countries—a notable example of this phenomenon was Sri Lanka in early 2022 [1].

In 2023, countries such as Indonesia began to emerge from the economic downturn wrought by the global pandemic. The Indonesian government's decision to repeal the PPKM (Implementation of Community Activity Restrictions) was pivotal in this recovery, significantly stimulating the tourism industry. This policy change allowed people to travel freely [2]. The resurgence was evident in the revival of public events, such as music concerts and cosplay gatherings. These events marked a return to pre-pandemic normalcy and played a vital role in revitalizing the nation's cultural landscape and economic health.

The revitalization of public gatherings has also profoundly affected the media, with the MICE (Meeting, Incentive, Conference, and Exhibition) sector. This impact extends beyond mere coverage, fostering a symbiotic relationship that promotes the events industry and media engagement. Therefore, for media online such as Media Have Fun (MHF), the resurgence of public events presents both a challenge and an opportunity to mend and enhance relationships with the events industry, which had languished during the protracted quietude of the COVID-19 era. This period necessitates a strategic realignment towards disseminating information about cultural events and the contributions of small and medium enterprises (SMEs), recapturing its audience's interest by publishing compelling articles.

To enhance viewership metrics, MHF strategically promotes tourist promote tourist destinations, highlights local culture, and spotlights SMEs renowned for their unique souvenirs and innovative products. Consequently, a thorough analysis of historical article trends on the mediahavefun.com website is imperative to ascertain the types of content that resonate most with their audience. This approach enables MHF to tailor its future publications to align with reader interests, thereby supporting the broader objective of cultural promotion. MHF needs to readjust its publications due to the vacuum in publishing articles during the pandemic, which may result in changes in reader interest patterns that make it necessary to carry out strategic planning. Neglecting to adapt content strategies in the post-pandemic era could result in significant declines in audience engagement and advertising revenue, slowing local economic recovery supported by tourism and cultural events [3]. To resolve the issue, we categorized data to streamline the analysis, encompassing article content data and audience interaction. These data sets are derived from SEO [4]-[7] analytics within the website, facilitating a comprehensive understanding of user engagement and content effectiveness.

To construct the categorization of data, we employ the k-means clustering algorithm. This method systematically groups data based on similarities in article content and audience interactions, enabling a data-driven approach to content strategy optimization. Through this analytical process, we discern the specific types of articles that captivate readers' attention. Applying k-means clustering provides a nuanced understanding of reader preferences and identification of content themes and styles that significantly influence engagement levels. Analyzing these patterns can tailor MHF's content to align with audience interests, optimizing reader engagement and satisfaction.

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 3, No 2, March 2024
Hal 208−214
DOI 10.47065/comforch.v3i2.1207

Our decision to utilize the k-means clustering algorithm was firmly grounded in the previous literature that underscored the algorithm's broad applicability and effectiveness in diverse research domains. [8] demonstrated its utility in customer segmentation within SMEs, enabling more targeted marketing strategies. Studies [9] in the educational field have also shown k-means' capability to segment students based on academic performance and study habits, providing valuable insights for educational strategy optimization. Furthermore, the algorithm has been employed in credit portfolio management for SMEs, illustrating its effectiveness in risk mitigation and business expansion efforts [10]. The classification of universities based on various attributes further highlights k-means' adaptability to academic evaluation. Additional studies, such as the analysis of COVID-19 spread within districts and the global ranking of universities, emphasize the algorithm's versatility in handling public health data and international educational standards [11]. These varied implementations across sectors reinforce the suitability of k-means for our research, which aims to categorize content and audience interaction data efficiently and uncover patterns that elucidate reader preferences and content engagement trends. Additionally, because previous research has focused on implementing k-means in business and education, it rarely has explored its application in post-pandemic online media content strategies, especially publishing articles about tourism and events.

Consequently, by employing k-means clustering to observe MHF effectivity, the analytical approach is anticipated to increase understanding of post-pandemic audience content preferences and provide strategic recommendations for online media as a tool that promotes tourism and SMEs.

2. RESEARCH METHODOLOGY

2.1 Research Stages

In the study, we retrieve "data-export Search Console Trafik and Queri.csv" and "data-export halaman landing.csv" from MHF's SEO toolkit to explore audience interests. The features collected from both datasets provide a comprehensive overview of user interaction and content performance on the platform. These metrics, including organic clicks and impressions from Google Search, click-through rates, average positions in search results, user engagement sessions, conversion rates, and ad revenue, offer detailed insights into how users find and interact with content. Additionally, landing page-specific data such as session counts, user engagement, new users, average interaction time per session, conversions, and total revenue further enhance our understanding of content effectiveness and audience behavior. Analyzing these datasets allowed us to track how visitors found MHF's website and what captured their attention. Furthermore, we carry out the following analysis steps for K-means Clustering in Figure 1:

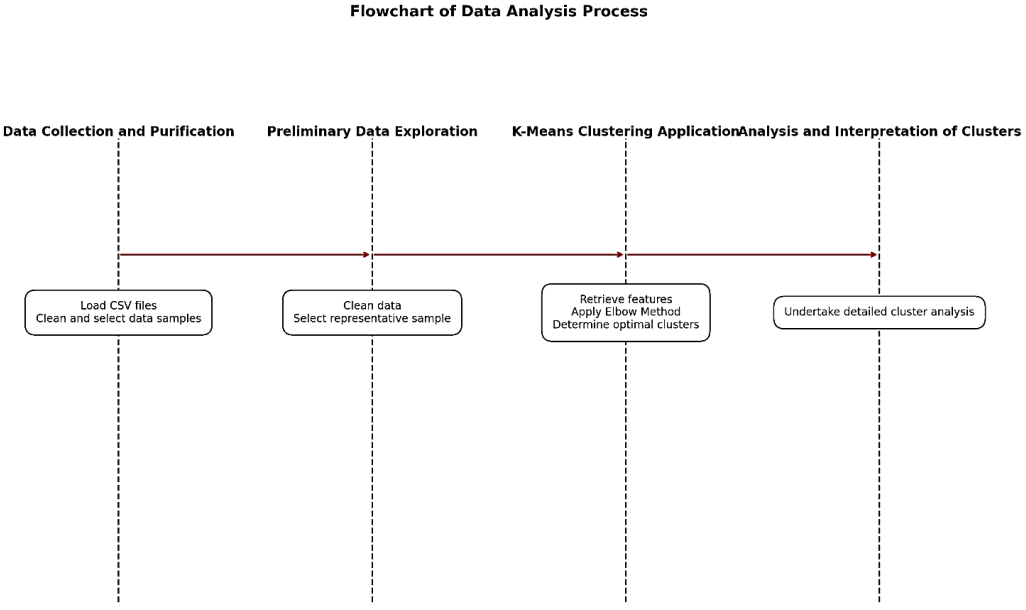

Figure 1. Research Stages using the K-Means Clustering Method

2.1.1 Data Collection and Purification

Gathering data from "data-export Search Console Trafik and Queri.csv" alongside "data-export halaman landing.csv" involved cleansing the dataset to eliminate any anomalies or irrelevant entries, ensuring a clean dataset for analysis. The following algorithm cleans the data and selects representative data samples for further analysis:

```
import pandas as pd

# Load the datasets
```

ISSN 2808-375X (Media Online)
Vol 3, No 2, March 2024
Hal 208−214
DOI 10.47065/comforch.v3i2.1207

```
df_traffic_queries = pd.read_csv('/mnt/data/data-export Search Console Trafik and Queri.csv')
df_landing_pages = pd.read_csv('/mnt/data/data-export halaman landing.csv')

# Display the first few rows of the datasets to understand their structure
df_traffic_queries.head(), df_landing_pages.head()
```

Based on the defined algorithm, through the "data-export Search Console Trafik and Queri.csv," a random sample of 327 entries was obtained. Similarly, for the "data-export halaman landing.csv" dataset, a random selection of 100 entries was made.

2.1.2 Preliminary Data Exploration

Conduct an exploratory data analysis to decipher patterns and distributions within the data, such as the frequency of queries and the popularity of specific landing pages:

```
# Data cleaning: remove inconsistent instances or missing values
df_traffic_queries_clean = df_traffic_queries.dropna().reset_index(drop=True)
df_landing_pages_clean = df_landing_pages.dropna().reset_index(drop=True)

# Determine a representative sample size if the data is too large
# For demonstration purposes, we will use a random sample of 10% of the data if the number of rows is > 1000
sample_frac = 0.1 if len(df_traffic_queries_clean) > 1000 else 1
df_traffic_queries_sample = df_traffic_queries_clean.sample(frac=sample_frac, random_state=42)

sample_frac_lp = 0.1 if len(df_landing_pages_clean) > 1000 else 1
df_landing_pages_sample = df_landing_pages_clean.sample(frac=sample_frac_lp, random_state=42)

# Displays sample information to be used for analysis
(df_traffic_queries_sample.info(), df_landing_pages_sample.info())
```

2.1.3 K-Means Clustering Application

Employed the K-Means clustering algorithm [13], [14] to categorize the data into groups based on search query similarities and landing page interactions. We retrieved the Elbow Method to ascertain the most appropriate number of clusters, which algorithms as follows:

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import numpy as np

# Retrive the 'Google Search Organic Clicks' and 'Google Search Organic Impressions' features for clustering
#  This approach sheds light on audience engagement patterns, revealing how effectively different content attracts and retains viewers based solely on organic search performance.
features = df_traffic_queries_sample[['Google Search Organic Clicks', 'Google Search Organic Impressions ']].values

# The Elbow Method is utilized to determine the optimal number of clusters
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0)
    kmeans.fit(features)
    wcss.append(kmeans.inertia_)

# Elbow Method Visualization
plt.figure(figsize=(10, 5))
plt.plot(range(1, 11), wcss)
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS') # within cluster sum of squares
plt.show()
```

The Elbow Method determines the optimal number of clusters by plotting the within-cluster sum of squares (WCSS) against the number of clusters. By identifying where the WCSS decreases more slowly, resembling an "elbow," we can select the most appropriate number of clusters for the k-means algorithm [14]-[19]. This method is crucial for balancing detail with generalization in clustering, ensuring neither overfitting nor oversimplification. Subsequency, the

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 3, No 2, March 2024
Hal 208−214
DOI 10.47065/comforch.v3i2.1207

elbow values were used to determine the number of K-Mean clusters, which we explain in the Result and Discussion section.

2.1.4 Analysis and Interpretation of Clusters

We undertake a detailed cluster analysis to reveal distinct characteristics and preferences of each audience segment, inferred from their search behaviors and landing page visits.

3. RESULT AND DISCUSSION

After completing the Data Collection and Purification and Preliminary Data Exploration, we determined the number of clusters using the elbow method. This method effectively identifies the optimal number of clusters by analyzing the rate of decrease in the within-cluster sum of squares (WCSS) as the number of clusters increases. The "elbow" point—where further increases in clusters result in only minimal reductions in WCSS—signals the most suitable number of clusters. The Elbow Visualization is as follows in Figure 2.

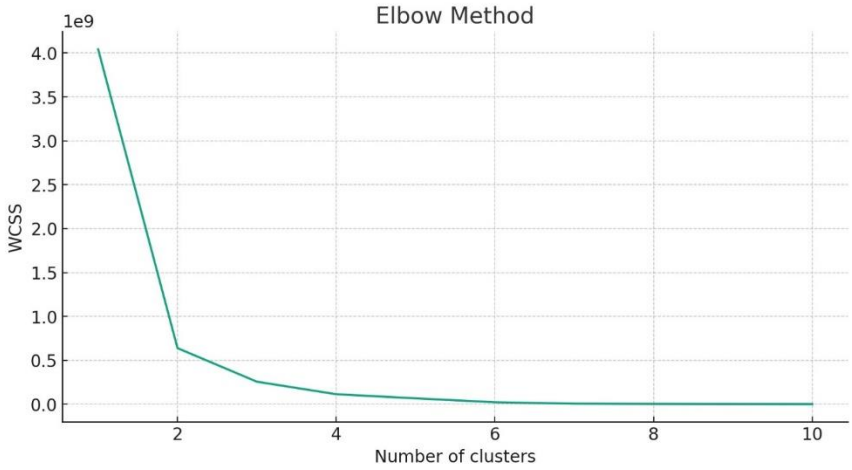

Figure 2. Elbow Method Visualization

In Figure 2, the Elbow Method visualization indicates the change in the Total Within-Cluster Sum of Squares (WCSS) with varying numbers of clusters. The plot presents an "elbow" formation around 3 or 4 clusters, suggesting this possibility is the optimal number of clusters. This finding implies that additional clusters do not significantly decrease the variance but construct 3 or 4 clusters sufficient to represent the data adequately.

Furthermore, we applied the K-Means Clustering with the number of clusters chosen based on observations from the Elbow Method. Following the specified K-Means algorithm, we selected four as the optimal number of clusters for this observation.

```
# K-Means Clustering with 4 clusters
kmeans_optimal = KMeans(n_clusters=4, init='k-means++', max_iter=300, n_init=10, random_state=0)
y_kmeans = kmeans_optimal.fit_predict(features)

# The Result K-Means Clustering Visualization
plt.figure(figsize=(10, 5))
colors = ['red', 'green', 'blue', 'yellow']
for i in range(4):
 plt.scatter(features[y_kmeans == i, 0], features[y_kmeans == i, 1], s = 100, c = colors[i], label = f'Cluster {i+1}')

plt.scatter(kmeans_optimal.cluster_centers_[:, 0], kmeans_optimal.cluster_centers_[:, 1], s = 300, c = 'purple', label = 'Centroids')
plt.title('Clusters of Audience Interest')
plt.xlabel('Organic Clicks from Google Search')
plt.ylabel('Organic Impressions from Google Search')
plt.legend()
plt.show()
```

Afterward, we visualized the clustering results and analyzed the content that attracts readers based on these clusters presented in Figure 3.

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 3, No 2, March 2024
Hal 208−214
DOI 10.47065/comforch.v3i2.1207

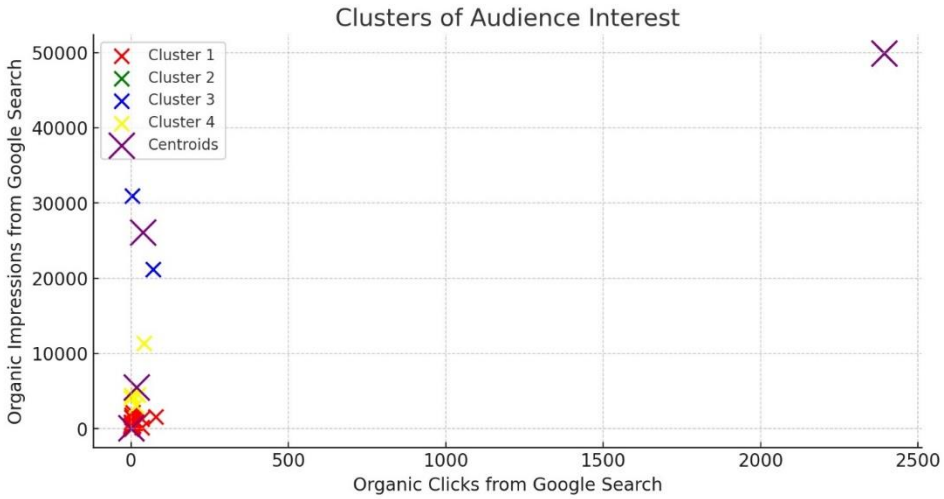

Figure 3. K-Means Clustering for Audience Interest

Based on Figure 3, the audience is divided into four clusters based on organic clicks and impressions from Google Search, each with distinct characteristics:

a. Cluster 1 (Red): The cluster presented low engagement in clicks and impressions, indicating a selective interest in less exposed content.
b. Cluster 2 (Green): Audiences have a higher click rate than impressions, suggesting they found the content highly relevant.
c. Cluster 3 (Blue): Characterized by deep impressions but relatively low clicks, indicating interest that does not fully motivate engagement.
d. Cluster 4 (Yellow): This represents the segment with very high clicks and impressions, indicating popular and widely appealing content.

3.1 Discussion

Based on the clustering analysis in Figure 3, it is evident that content genres captivating readers vary significantly across different interaction levels. Specifically, Cluster 2 content, such as feature keywords or topics highly relevant to a niche audience, leads to significant conversion rates per impression. Conversely, Cluster 3 content may benefit from further SEO or optimization to boost click-through rates. Several points that underline:

a. Well-Optimized Content: Cluster 4's high clicks and impressions indicate topics or keywords with broad appeal. Maintaining and applying effective SEO strategies for such content is crucial.
b. High-Potential Content: Cluster 2's content, yielding high clicks from fewer impressions, suggests specific keywords or topics hold significant relevance for the target audience.
c. Content Optimization: Content in Cluster 3 necessitates SEO evaluation and optimization to enhance engagement, possibly through keyword research, content quality improvement, or engaging formats.
d. Audience Understanding: This clustering aids in comprehending various audience segments and their differing preferences, which is valuable for crafting more targeted content to boost overall engagement.

This analysis underscores the importance of leveraging data and clustering analysis in refining content and SEO strategies. Online media can develop more effective strategies to enhance engagement and achieve business objectives by understanding content that resonates with audiences. Interestingly, Media Have Fun's attention is on Cluster 1, where audiences engage with non-tourism-related content, such as history or myths, which SEO often overlooks. Cluster 1's correlation is depicted in Figure 4.

Journal of Computing and Informatics Research
ISSN 2808-375X (Media Online)
Vol 3, No 2, March 2024
Hal 208−214
DOI 10.47065/comforch.v3i2.1207

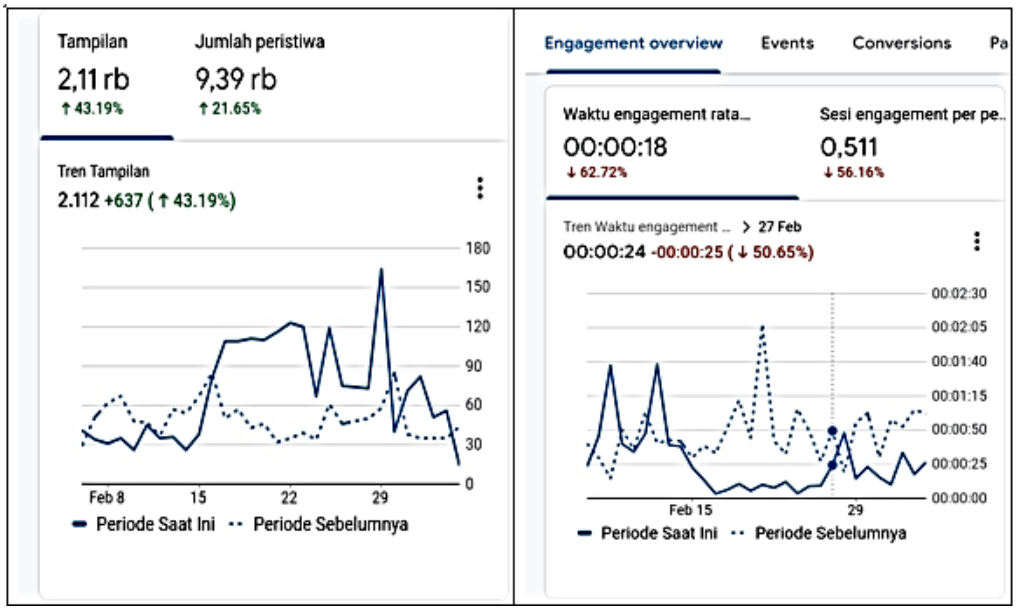

Figure 4. MHF's Engagement Overview

Based on Figure 4, While impressions and total events have increased, indicating greater reach, the average engagement time and the number of sessions have decreased sharply. We conclude that although more people see the content, they spend less time with it and engage less frequently. It could be due to many factors, including the audience not finding the article they are looking for or the article the reader is looking for not being recommended by SEO. This insight suggests a regular posting and SEO optimization strategy to enhance visibility and engagement.

Applying k-means clustering to discern audience interests in SEO-optimized content reveals limitations, especially within the SEO context, such as data dependency, SEO variability, interpretation challenges, external factor impacts, and multi-platform SEO considerations. K-Means Clustering can still identify distinct audience interest groups, laying a foundation for informed content and SEO strategies. However, recognizing its limitations and considering additional analytical techniques or complementary data is crucial for a holistic understanding of reader interests. Adapting to interest dynamics over time and adopting a flexible, adaptive approach to content analysis and strategy may further succeed in analyzing and meeting reader interests.

4. CONCLUSION

This study validates the utility of k-means clustering for SEO data segmentation to understand audience interests, pinpointing the elbow method as key for identifying the optimal cluster number. It uncovers notable differences in content engagement across clusters, underscoring the importance of refining SEO strategies and deepening audience segmentation insights, such as MHF's case that the audience could not find the article they were looking for or the article the reader was looking for, not being recommended by SEO. Despite challenges like SEO's evolving nature and data reliance, this approach establishes a robust basis for improving content strategies. Recommendations for future research include embracing cross-platform data, conducting longitudinal studies, incorporating sentiment analysis, experimenting with content, focusing on UX (User Experience), and tracking algorithm updates, all aimed at crafting more responsive content and SEO strategies to align with changing audience preferences.

REFERENCES

[1] R. Andrianto, “Ini yang Bikin Sri Lanka Chaos, Semoga Tak Terjadi di RI,” *CNBC Indonesia*, Jakarta, Jul. 11, 2022. Accessed: Mar. 04, 2024. [Online]. Available: https://www.cnbcindonesia.com/news/20220711141658-4-354659/ini-yang-bikin-sri-lanka-chaos-semoga-tak-terjadi-di-ri

[2] OECD, "Mitigating the impact of COVID-19 on tourism and supporting recovery," 2020.

[3] A. A. Aburumman, "COVID-19 impact and survival strategy in business tourism market: the example of the UAE MICE industry," *Humanit Soc Sci Commun*, vol. 7, no. 1, p. 141, 2020, doi: 10.1057/s41599-020-00630-8.

[4] C. Lopezosa, L. Codina, J. Díaz-Noci, and J.-A. Ontalba, "SEO and the digital news media: From the workplace to the classroom," *Comunicar*, vol. 28, no. 63, pp. 65–75, Apr. 2020, doi: 10.3916/C63-2020-06.

[5] D. Giomelakis, C. Karypidou, and A. Veglis, "SEO inside Newsrooms: Reports from the Field," *Future Internet*, vol. 11, no. 12, p. 261, Dec. 2019, doi: 10.3390/fi11120261.

[6] R. S. Bhandari and S. Bansal, "An Analysis Between Search Engine Optimization Versus Social Media Marketing Affecting Individual Marketer's Decision--Making Behavior," *Jindal Journal of Business Research*, vol. 8, no. 1, pp. 78–91, Jun. 2019, doi: 10.1177/2278682119829607.

ISSN 2808-375X (Media Online)
Vol 3, No 2, March 2024
Hal 208−214
DOI 10.47065/comforch.v3i2.1207

[7] M. Poturak, D. Keco, and E. Tutnic, "Influence of search engine optimization (SEO) on business performance," *International Journal of Research in Business and Social Science (2147- 4478)*, vol. 11, no. 4, pp. 59–68, Jun. 2022, doi: 10.20525/ijrbs.v11i4.1865.

[8] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster," *IOP Conf Ser Mater Sci Eng*, vol. 336, p. 012017, Apr. 2018, doi: 10.1088/1757-899X/336/1/012017.

[9] A. Zaki, I. Irwan, and I. A. Sembe, "Penerapan K-Means Clustering dalam Pengelompokan Data (Studi Kasus Profil Mahasiswa Matematika FMIPA UNM)," *Journal of Mathematics Computations and Statistics*, vol. 5, no. 2, p. 163, Oct. 2022, doi: 10.35580/jmathcos.v5i2.38820.

[10] I. A. Yorinda and A. B. Raharjo, "Analysis of Customer Profile Charasteristic with Credit Quality Using the Clustering Method for Risk Mitigation and Small Medium Enterprise Credit Portofolio Expansion Planning," *Business and Finance Journal*, vol. 8, no. 2, pp. 181–191, Nov. 2023, doi: 10.33086/bfj.v8i2.5225.

[11] F. Dikarya and S. Muharni, "Penerapan Algoritma K-Means Clustering untuk Pengelompokan Universitas Terbaik di Dunia," *Jurnal Informatika*, vol. 22, no. 2, pp. 124–131, Dec. 2022, doi: 10.30873/ji.v22i2.3324.

[12] D. Triyansyah and D. Fitrianah, "Analisis Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing," *Jurnal Telekomunikasi dan Komputer*, vol. 8, no. 3, p. 163, Oct. 2018, doi: 10.22441/incomtech.v8i3.4174.

[13] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf* Sci *(N Y)*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[14] N. Jayanti, C. G. Selan, and M. W. Prihatmono, "Analisis Data Pengunjung Mall Nipah Mendukung Strategi Digital Marketing Menggunakan K-Means Clustering," *Jurnal Penelitian Inovatif*, vol. 2, no. 1, pp. 209–220, May 2022, doi: 10.54082/jupin.65.

[15] I. Ramadhaniati, "Product Clustering using K-MEANS Method in CV. JAYA ABADI," *Jurnal TAM (Technology Acceptance Model)* , vol. 14, no. 1, pp. 91–97, 2023.

[16] A. Ali and S. A. M. Uktutias, "Optimasi Hasil Clustering Data Rekam Medis Balita di Desa Jumput Rejo dengan Metode Elbow dalam Menunjang Program Pemerintah Mengatasi Stunting," *Joutica : Journal of Informatic Unisla*, vol. 8, no. 1, 2023.

[17] K. C. Gunawan, "Analisis Peserta Online Test dengan Pembentukan Cluster dengan Algoritma K-Means dan Analisis Tabulasi," *Jurnal Titra*, vol. 8, no. 2, pp. 353–360, 2020.

[18] H. Humaira and R. Rasyidah, "Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm," in *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia*, EAI, 2020. doi: 10.4108/eai.24-1-2018.2292388.

[19] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP J Wirel Commun Netw*, vol. 2021, no. 1, p. 31, 2021, doi: 10.1186/s13638-021-01910-w.

[20] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," *Matrix : Jurnal Manajemen Teknologi dan Informatika*, vol. 9, no. 3, pp. 102–109, Nov. 2019, doi: 10.31940/matrix.v9i3.1662.