Identifying Improvement Strategic from User Application Reviews Group Using K-Means Clustering and TF-IDF Weighting

Khairunnisa Nurul Istiqomah [a,1], Imam Djati Widodo [b,2,*], Nisrina Faiza Mufid [b,3], Qurtubi [b,4]

[a] Master of Industrial Engineering, Faculty of Industrial Technology, Universitas Islam Indonesia, Yogyakarta 55584, Indonesia
[b] Department of Industrial Engineering, Faculty of Industrial Technology, Universitas Islam Indonesia, Yogyakarta 55584, Indonesia
[1] khairunnisa.istiqomah01@students.uii.ac.id; [2] imamdjati@uii.ac.id *; [3] nisrina.mufid@students.uii.ac.id@students.uii.ac.id; [4] qurtubi@uii.ac.id
*corresponding author

ARTICLE INFO

Article history
Received 29 Jul 2023
Revised 08 Sep 2023
Accepted 12 Des 2023

Keywords
Categorical Data
Elbow Method
Silhouette Score
TF-IDF

ABSTRACT

In the digital era, companies must pay attention to all their facilities, especially digital ones. One of the facilities that makes it easier for company customers to access information and needs for company services is through an application. Good application performance will benefit the company, such as assessing company performance. This needs attention because current assessments of services or products are based on ratings and reviews from previous user experiences. This research will group reviews based on the similarity of sentences and discussions. So that each cluster can be given the right strategy to improve ratings that were not previously expected, grouping using K-Means Clustering requires determining the number of clusters at the beginning first. Determining the optimal number of clusters in this research uses the elbow method and TF-IDF weighting because the data is in text form. The results of determining the number of clusters were strengthened using silhouette scores, and the optimal number of clusters in this study was 2 clusters. The improvement strategy for each cluster is adjusted from the analysis of cluster characteristics and possible causes for the appearance of the review.

1. Introduction

The company must develop following existing technological advances. One of them is that companies need to pay attention to and meet the needs of facilities that can make things easier for their users. In this online era, companies can provide a facility in the form of an application to make it easier for customers to order the services/products offered without having to come to an offline store [1]. There are many benefits when the company's customers use the application to make transactions. Applications can be a tool to store all information related to user activities while opening them. The application can also connect users with companies and serve as a tool for analyzing CRM (Customer Relationship Management). A good CRM will increase customer loyalty to the product/service or company, as will the applications developed by the company [2]. When the performance value of an application is good, then not only does the application have good value, but the company can also gain the trust of new users [3]. New users or customers will check the performance before using it. Customers can check performance by reviewing reviews from users who have used it. So, when a company has an unsatisfactory rating, it will reduce new users' interest in using its services/products. Companies need to maintain all the facilities they offer, such as applications. Especially if the application has a poor rating/review, this research will look at the

results of reviews of companies in the service sector that offer applications to make it easier for customers to order tickets. There is no specific discussion stating what a reasonable period is, so it is essential to review public opinion. This study used reviews only from the last year to find the most updated opinions. Especially in the past year, there have been several new things, such as moving online activities offline so that they require various forms of transportation.

Analysis of a review can use the application of text mining, also known as sentiment analysis. Sentiment analysis is one of the best methods to derive expressed emotions from unstructured texts by transforming the data into a structured format [4]. Several studies regarding application reviews have been conducted previously. The methods used in existing research are Support Vector Machine (SVM) and Naïve Bayes [5]. This research compared the advantages of the two algorithms: SVM and Naïve Bayes. The final result of the study found that the value of the SVM algorithm was higher than the Naïve Bayes algorithm. The same research has also been conducted to determine the sentiment results with the algorithm where it is known that Naïve Bayes shows better recall values and using SVM shows better precision values [6]. However, the two algorithms did not show significant differences in results. Research was also conducted using other methods to determine whether the difference in classification performance results was insignificant [7]. According to research conducted by Miles, this research is included in the practical knowledge gap [8] because it adds word weight analysis after clusters are formed and strategies for each cluster by the review group. Although research related to this method has been carried out, this research does not only stop at the grouping results but also provides strategies from the results of sentence analysis. This study did not classify reviews by labels first. However, reviews are grouped based on similarities in sentence form and similarity of words in them using the K-Means Clustering algorithm.

The Clustering approach itself is classified as an unsupervised learning approach, a category of machine learning that does not require labels in the dataset and does not require identification of classes. So, the group results will not always be divided into two or three groups as in similar research related to K-Means Clustering on text data [9]. Different optimal values for the number of clusters will be produced in the case of different data. This research will use K-Means Clustering to determine the number of review groups at the beginning and the Elbow method by giving weights to the reviews to determine the number of review groups or the optimal number of clusters [10]. The Elbow method is commonly used in numerical data types because it uses a percentage of data variance [11]. So, if the data to be analyzed is in the form of categories, it is necessary to do weighting first. The intended weighting depicts how vital the word in the sentence is. Using TF-IDF (Term Frequency-Inverse Document Frequency), the weighting is seen from the frequency of occurrence of sentences in the term. In addition, logarithmic calculations are carried out to the ratio of the total amount of data processed with data from a previous period [12]. It will produce a value in each sentence processed; in this case study, weighting is done by looking at the number of reviews. Reviews with similar weighting will later be identified into one group using the K-Means Clustering algorithm. The results of the number of clusters from the elbow method will be re-evaluated using a silhouette score so that the results are not objective. So, the evaluation model in this research uses the help of a silhouette score. After the clustering or grouping processes obtain the sentiment results, an analysis will be carried out on each review group. So, the study is carried out by looking at what factors cause the emergence of negative reviews. In previous research, review groups were only divided based on negative, neutral, and positive sentiment or only negative and positive sentiment without providing improvement strategies. If an improvement or strategy is to be carried out to increase and correct negative reviews, it is mainly based on the results of negative sentiment in that one cluster. This clustering aims to find an appropriate marketing strategy [13]. So that the sentiment in each review group can be known, as well as an analysis of the causal factors and strategies that can be provided. This is intended so that improvements can be carried out in stages according to the type of review that will be repaired first.

2. Method

2.1 Research Mapping

This kind of research has been carried out before but only stopped at the final result, sentiment or method accuracy. Research on application reviews has never identified sample sentences or reviews as improvements for the company. This research will show that the number of clusters in categorical

data grouping can be determined first according to the similarity of the reviews and provide more detailed identification regarding the characteristics of each cluster. So that companies can focus more on making decisions in determining improvements to review results, such as creating new strategies. Plans and strategies are essential for the sustainability of the company. The implementation of the strategy can be done by paying attention to efficiency and operations, alternative strategies, or the company's short- and long-term growth initiatives. Companies can also determine strategies through the results of answering fundamental questions [14].

2.2 Research Method

Before the data is processed using the main method, pre-processing must be done. Data processing was done to remove unnecessary information from primary data using Google Collaboratory. The main method for dividing data into groups is using K-Means Clustering. Clustering is a data exploration method to obtain hidden characters that aim to group data into several clusters by maximizing similarities between objects in one cluster and minimizing the similarity of metric distances between clusters determined in high-dimensional space [15]. The K-Means Clustering method is known for its efficiency in grouping large data. However, the average-based method on the distance between these data is constrained when dealing with data with categorical features [16]. Categorical data processing requires TF-IDF weighting before determining the optimal number of clusters using the elbow method. TF-IDF is obtained by multiplying term frequency (TF) with inverse document frequency (IDF) [17]. Inverse document frequency will identify that the occurrence of a term with a high frequency in data or documents is significant. Likewise, a document's low occurrence of words or terms is considered unimportant because the discussion needs to represent the document [12]. After carrying out these steps, the data can be processed using the elbow method. This method uses the percentage of variance, which functions as the number of clusters [11]. This method is used to group the number of clusters more effectively and efficiently and to get the best value. The best cluster value can later be determined from the angled shape of a graph resulting from analysis using the Elbow method [18]. However, when the elbow shape of the graphic results is too smooth or invisible, this method cannot be used alone or requires other supporting methods to determine the number of clusters or elbow points more accurately[19]. The model evaluation was carried out to review the accuracy of determining the number of clusters based on the silhouette score. The highest silhouette score indicates the optimal number of clusters formed [20].

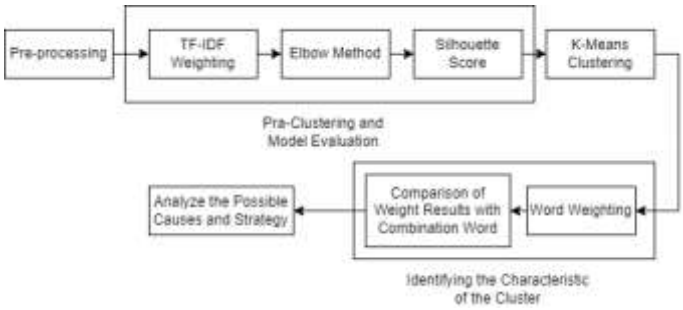

Fig. 1. Research Step

After obtaining the optimal number of clusters, the data grouping method is carried out. Data samples with the highest word frequency were taken from the grouping results for characteristic analysis. The analysis was carried out by giving word weights twice as proof that the first weight analysis was representative of the characteristics of a cluster.

3. Results and Discussion

3.1. Pre-processing Result

Several steps from the pre-processing stages are carried out according to the needs of the data. The proposed pre-processing methods are Cleaning Data, Tokenization, Stop Word Removal, Stemming, and Deleting Data Null.

1) Cleaning Data

Cleaning data referred to in this research is done by deleting data variables that are not needed. Hence, it only focuses on the data that is required, such as the content of the review and the time the review was created. After that, the data is duplicated, such as the results of reviews being sent twice or due to other errors being deleted, so the data used is only single reviews from users.

Table 1. Cleaning Data

Input	Output
"Please fix it again, the app is already difficult to use unlike it used to be"	"Please fix it again, the app is already difficult to use unlike it used to be"
"Please fix it again, the app is already difficult to use unlike it used to be"	
" It's often slow, and keeps failing., time is running out."	" It's often slow, and keeps failing., time is running out."
" It's often slow, and keeps failing., time is running out."	
" Often slow, please improve"	" Often slow, please improve"
" Often slow, please improve"	

2) Tokenization

Tokenization is used to investigate sentences and create a list of possible tokens used as input for the next algorithm. This process is also included with removal of elements such as brackets, hyphens and other punctuation, and also changes all letters to lowercase.

Table 2. Tokenization

Input	Output
"Please fix it again, the app is already difficult to use unlike it used to be"	"please; fix; it; again; the; app; is; already; difficult; to; use; unlike; it; used; to; be;"
"It's often slow, and keeps failing., time is running out."	"its; often; slow; and; keeps; failing; time; is; running; out;"
"Often slow, please improve"	"often; slow; please; improve;"

3) Stop Word Removal

A pre-processing stage that is carried out by removing words does not contain important and significantly different information or a word that appears frequently and has no influence on research on that sentence.

Table 3. Stop Word Removal

Input	Output
"please; fix; it; again; the; app; is; already; difficult; to; use; unlike; it; used; to; be;"	"please; fix; again; app; difficult; use; unlike;"
"it; is; often; slow; and; keeps; failing; time; is; running; out;"	"often; slow; keep; failing; time; running; out;"

4) Stemming

This pre-processing technique removes noise from text data, performs stemming, and trimms the size of text data. In the previous pre-processing, there were still several sentences that did not directly address the basic words, which were removed to focus more on the problem.

3.2. Formation of the Number of Clusters

Programming cannot read data if it is still in text form for analysis by the Elbow method. So with TF-IDF (Term Frequency & Inverse Document Frequency), which functions to sort important words with more weight with higher values. So, words that appear more frequently in the overall data will be more relevant to the data. The best determination of the number of clusters can be seen from the formation of elbows on the graph. From the graph below, it can be seen that the elbow shape is in cluster 2. Another determination can also be seen from the difference in values that have changed most significantly.

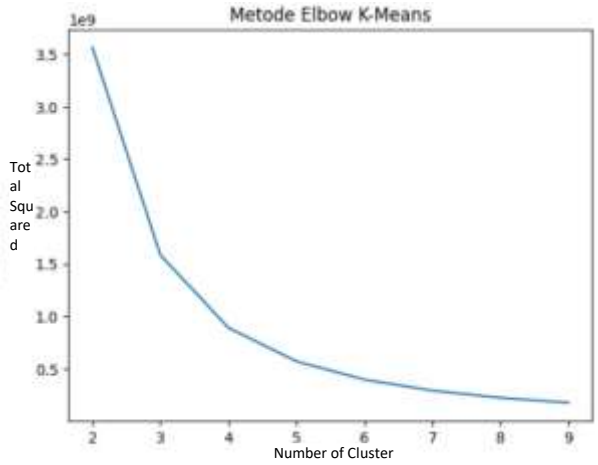

Fig. 2. Elbow K-Means Graph

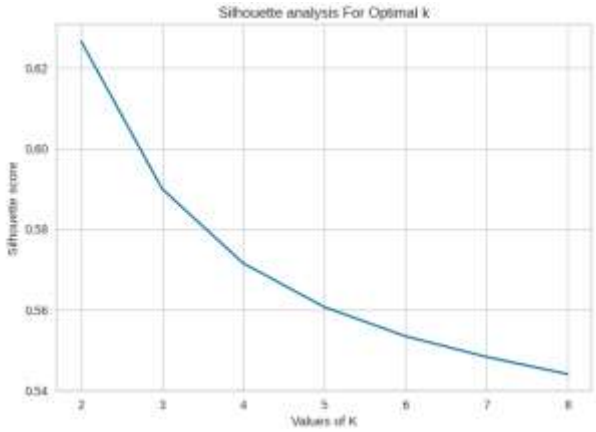

Fig. 3. Sillhoutte Score

Based on the results of the silhouette score evaluation, determining the number of clusters as 2 is the correct determination. This is because the silhouette score has the highest value compared to the number of other clusters. So the formation of these 2 clusters is the most optimal. The silhouette shows a score that gets smaller when the number of clusters increases. The number of clusters formed can change according to the size and type of data.

3.3. Clustering Results and Word Weights

After the optimal number of clusters is obtained, cluster formation is carried out using the k-means algorithm. After the data was divided into 2 clusters, the amount of data in cluster 1 was 3770 data, and in cluster 2, it was 1000 data. The next step is giving weight to words in clusters to know the characteristics of each cluster. This characteristic analysis took sample sentences and words with the highest frequency in the cluster. In each cluster group, the 20 words with the greatest frequency of occurrence were given weight to determine the sentiment formed. Words with positive meanings will be given a weight of 1, words with negative meanings will be given a weight of -1, and words that do not contain negative and positive meanings are neutral or equal to 0. The following are the sentiment analysis result on one of the clusters and the word samples.

Table 4. Greatest Frequency Word Cluster 2

Word	Frequency	Weight	Total Weight
Application	1359	0	0
Nice	324	1	324
Order	216	0	0
Error	158	-1	-158
Please	139	0	0
Pay	125	0	0
Open	100	0	0

Difficult	100	-1	-100
Choose	91	0	0
Good	92	1	92
Easy	77	1	77
Update	75	-1	-75
Fast	67	1	67
Bug	60	-1	-60
Thanks	23	1	23
Schedule	6	0	0
Practical	6	1	6
Use	5	0	0
Qris (a payment method)	4	0	0
Login	4	0	0
Total			196

To ensure the correctness of the results of the weighting analysis, the words mentioned above were re-weighted. The following results show that the previous sentiment analysis, which had a positive value still shows a consistent positive value even though the numbers are different. The analysis of another group (cluster 1) also showed consistent results in the first weighting, negative values and the second weighting.

Table 5. Word Combination Weighting

Word in a Sentence		Frequency	Weight	Total Weight
Application	Help	44	1	44
Application	Easy	6	1	6
Application	Good	7	1	7
Help	Thanks	16	1	16
Accept	Thanks	15	1	15
Help	Pay	10	1	10
Help	Street	9	1	9
Help	Order	10	1	10
Please	Help	8	-1	-8
Help	Thanks	8	1	8
Total				117

The total weight results show the tendency of the content of the review topics in the cluster to show what kind of sentiment. The two tables above are calculations for the second cluster. In the second cluster, the formed reviews tend to s Even though there were some negative responses, it didn't become the main topic and didn't dominate.

3.4. Strategy Analysis

Each word with the most significant frequency in each cluster was analyzed at its source or original review. This review determines possible causes for the emergence of positive or negative words. So that we can provide the best strategy to restore good reviews and improve the company's image. The strategy in this research results from previous studies similar to related causes.

Table 6. Strategy Analysis

Cluster	Possible Causes	Strategy
Cluster 1	Users find it difficult to log in, pay and register. Users feel that the application is not updated. Payment steps are impractical and need to be more complex. Payment methods are limited to only a few types.	Improve the payment process or registration user needs [21]. Provide maintenance to the application by paying attention to updating information on the application [22]. Create a payment process that is easy for users to understand [23] Provide payment methods according to user needs and those that most users commonly use [24].
Cluster 2	Users in cluster 3 do have the ability to understand application features quickly. Application users are loyal customers who already understand the difficult things for other users. The server network is too full or requires network development because too many people are accessing the application.	Perform maintenance to satisfy application users further [22]. Offer several promos so that application users don't worry too much about minor errors in application services [25]. Improve the server network and change it to better quality [26].

Several features still need to be more accessible for application users.	Analyze features that can be improved or developed, then improve these features [27].

4. Conclusion

Grouping reviews of an application can be done using K-Means Clustering with the help of weighting by TF-IDF (Term Frequency-Inverse Document Frequency) and determining the number of clusters using the elbow method. The results of grouping reviews in the text are formed into two groups. This research uses a silhouette score to ensure that the determination of the optimal number of clusters is correct. This research show that the analysis of the elbow and silhouette score methods shows the same number of clusters. Review the characteristics of each cluster by looking at the frequency of words that are the main discussion in that group. After that, refer to the existing combination of words and confirm the meaning of the original sentence. This determination is made because the data is in text form, so it is easier to review the original sentence of the words that appear.

This research has a weakness in the accuracy of determining characteristics because the review examines the words one by one in the sentence. So, the determining characteristics and strategies is based on reviewing the sentences in each cluster. Strategy determination is carried out by referring to previous research when dealing with similar problems. Deficiencies in reviews in each cluster can be given appropriate strategies to improve them by improving possible causes based on references to previous research.

References

[1] A. Sharma, S. Sharma, and M. Chaudhary, "Are small travel agencies ready for digital marketing? Views of travel agency managers," *Tour Manag*, vol. 79, Aug. 2020, doi: 10.1016/j.tourman.2020.104078.

[2] V. Rosalina, J. Raya, S.-C. Km, T. Kopassus, S. Banten Indonesia, and A. Triayudi, "Electronic Customer Relationship Management (E-CRM) Application as Efforts to Increase Customer Retention of Micro Small and Medium Enterprises (MSMEs) in Banten Indonesia," 2019.

[3] S. Sunder, K. H. Kim, and E. A. Yorkston, "What Drives Herding Behavior in Online Ratings? The Role of Rater Experience, Product Portfolio, and Diverging Opinions," *J Mark*, vol. 83, no. 6, pp. 93–112, Nov. 2019, doi: 10.1177/0022242919875688.

[4] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment Analysis of COVID-19 Tweets by Deep Learning Classifiers—A Study to Show How Popularity is Affecting Accuracy in Aocial Media," *Applied Soft Computing Journal*, vol. 97, Dec. 2020, doi: 10.1016/j.asoc.2020.106754.

[5] H. -, A. Y. Kuntoro, and T. Asra, "Klasifikasi Keluhan Pengguna KAI Access untuk Pemesanan Tiket dengan Algoritma SVM dan Naive Bayes," *JIKA (Jurnal Informatika)*, vol. 6, no. 2, pp. 161–169, Jun. 2022, Accessed: Dec. 13, 2023. [Online]. Available: https://jurnal.umt.ac.id/index.php/jika/article/view/6187

[6] H. Mustakim and S. Priyanta, "Aspect-Based Sentiment Analysis of KAI Access Reviews Using NBC and SVM," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 16, no. 2, p. 113, Apr. 2022, doi: 10.22146/ijccs.68903.

[7] H. Tan, "Machine Learning Algorithm for Classification," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Aug. 2021. doi: 10.1088/1742-6596/1994/1/012016.

[8] D. A. Miles, "A Taxonomy of Research Gaps: Identifying and Defining the Seven Research Gaps," 2017.

[9] O. Iparraguirre-Villanueva, V. Guevara-Ponce, F. Sierra-Liñan, S. Beltozar-Clemente, M. Cabanillas-Carbonell, and R. Palma, "Sentiment Analysis of Tweets using Unsupervised Learning Techniques and the K-Means Algorithm." [Online]. Available: www.ijacsa.thesai.org

[10] D. Marutho, S. Hendra Handaka, and E. Wijaya, "The Determination of Cluster Number at k-mean using Elbow Method and Purity Evaluation on Headline News," 2018.

[11] P. Bholowalia and A. Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN," *Int J Comput Appl*, vol. 105, no. 9, pp. 975–8887, 2014.

[12] C. T. Yu, K. Lam, and G. Salton, "Term Weighting in Information Retrieval Using the Term Precision Model," 1982.

[13] G. Punj and D. W. Stewart, "Cluster Analysis in Marketing Research: Review and Suggestions for Application."

[14] M. Del Pero, "The Importance of Strategic Planning," *Reinforced Plastics*, vol. 57, no. 2, pp. 16–18, 2013, doi: 10.1016/S0034-3617(13)70054-7.

[15] R. Dubes and A. K. Jain, "Validity Studies in Clustering Methodologies," *Pattern Recognit*, vol. 11, no. 4, pp. 235–254, 1979.

[16] Z. Huang, "Clustering Large Data Sets with Mixed Numeric and Categorical Values," in *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1997. [Online]. Available: https://api.semanticscholar.org/CorpusID:3007488

[17] N. Z. Dina, R. Triwastuti, and M. Silfiani, "TF-IDF Decision Matrix to Measure Customers' Satisfaction of Ride Hailing Mobile Application Services: Multi-Criteria Decision-Making Approach," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 17, pp. 104–118, 2021, doi: 10.3991/ijim.v15i17.22509.

[18] T. Soni Madhulatha, "An Overview on Clustering Methods," vol. 2, no. 4, pp. 719–725, 2012, Accessed: Dec. 13, 2023. [Online]. Available: www.iosrjen.org

[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, Dec. 2021, doi: 10.1186/s13638-021-01910-w.

[20] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, Jun. 2021, doi: 10.3390/e23060759.

[21] N. Singh and N. Sinha, "How Perceived Trust Mediates Merchant's Intention to use a Mobile Wallet Technology," *Journal of Retailing and Consumer Services*, vol. 52, Jan. 2020, doi: 10.1016/j.jretconser.2019.101894.

[22] W. T. Wang, W. M. Ou, and W. Y. Chen, "The Impact of Inertia and User Satisfaction on the Continuance Intentions to Use Mobile Communication Applications: A Mobile Service Quality Perspective," *Int J Inf Manage*, vol. 44, pp. 178–193, Feb. 2019, doi: 10.1016/j.ijinfomgt.2018.10.011.

[23] A. K. Kar, "What Affects Usage Satisfaction in Mobile Payments? Modelling User Generated Content to Develop the 'Digital Service Usage Satisfaction Model,'" *Information Systems Frontiers*, vol. 23, no. 5, pp. 1341–1361, Sep. 2021, doi: 10.1007/s10796-020-10045-0.

[24] O. Tounekti, A. Ruiz-Martinez, and A. F. Skarmeta Gomez, "Users Supporting Multiple (Mobile) Electronic Payment Systems in Online Purchases: An Empirical Study of Their Payment Transaction Preferences," *IEEE Access*, vol. 8, pp. 735–766, 2020, doi: 10.1109/ACCESS.2019.2961785.

[25] A. Subiyantoro, R. Dwi Astuti, H. Agung Nugroho, A. Manajemen Administrasi Yogyakarta, and U. Janabadra Yogyakarta, "Pengaruh Strategi Pemasaran, Pelayanan dan Promositerhadap Keputusan Pembelian Tiket Kereta Api," 2022. [Online]. Available: http://Jurnal.amayogyakarta.ac.id/index.php/albama

[26] T. F. Abdelzaher and N. Bhatti, "Web content adaptation to improve server overload behavior," 1999.

[27] F. Alqahtani and R. Orji, "Insights from user reviews to improve mental health apps," *Health Informatics J*, vol. 26, no. 3, pp. 2042–2066, Sep. 2020, doi: 10.1177/1460458219896492.