INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

How to Deeply Analyze the Content of Online Newspapers Using Clustering and Correlation

Yeni Rokhayati [a,*], Sartikha [b], Nur Zahrati Janah [b]

[a] *Multimedia and Network Engineering, Politeknik Negeri Batam, Jl. Ahmad Yani Batam Kota, Batam, 29641, Indonesia*
[b] *Informatics Engineering, Politeknik Negeri Batam, Jl. Ahmad Yani Batam Kota, Batam, 29641, Indonesia*
*Corresponding author: *yeni@polibatam.ac.id*

***Abstract*— The increase in the number of visitors is one of the keys to increasing income for online newspapers, whether to increase the number of ads, Google AdSense, or customer trust. Therefore, finding which news categories increase the number of visitors needs to be known and analyzed more deeply. Because it is very common to add content to online newspaper sites every day, even for hours, this pattern analysis is not the same as analyzing regular website content patterns. This study intends to add methods in the world of research on how to analyze website content, especially online news, by using the clustering method to classify what news categories bring high, medium, or a low number of visitors and then analyzing the correlation to explore the depth of the relationship between the variables, namely which parameters have a large or low effect on the increase in the number of visitors. A local Batam-based online newspaper company is used as a case study for this research. Data is collected, preprocessed first, and analyzed using the clustering and correlation method. This analysis of the news content readership suggests what news categories should be optimized because it provides an increase in the number of visitors. A summary of the analysis steps in this study is presented. We also provided some suggestions if other online newspaper owners or researchers are interested in a similar analysis of online news content.**

***Keywords*— Clustering; content analysis; correlation; data mining; news category; online newspapers.**

Manuscript received 15 Jan. 2022; revised 23 Mar. 2022; accepted 12 Apr. 2022. Date of publication 31 May 2022.

I. INTRODUCTION

Online newspapers can earn from Paywall [1], [2], Google AdSense [3], [4], sponsored contents [5], [6], and banner ads [7], [8]. The more visitors who read the news in the newspaper, the more income they earn. For online newspapers, increasing the number of visitors is one of the main goals. There are various ways to increase the number of web visitors, including Search Engine Marketing (SEM) and Optimization (SEO) [9]–[13], social media marketing [14]–[18], visitor behavior analysis [19]–[23], and content analysis [24]–[26].

In online newspapers, where the content is often produced hourly, it is very appropriate if the policy is taken based on an in-depth content analysis pattern. These include how the contents should be, what news categories should be optimized, etc. The interview with 9 out of 10 top Reuter's news companies [24] concluded that web analysis could guide editorial decisions and journalists to take risks and change what should be and what news serves the public interests. It also means that content written by journalists plays an important role in increasing the rating of their news sites. Akhirina et al. [25] predicted website content based on visitor data with a data mining approach for improving the quality of targeted content according to the interests of website visitors. A classification model was formed based on the attributes of the entrance, page view, unique page view, session duration, and page title. This model was able to predict which page titles were popular. However, this page title prediction model is more appropriate for websites with monotonous page titles. The page title that is predicted to be popular will continue to be raised/prioritized on the website. This approach is not suitable for online news websites where the news titles are always increasing and changing every day.

Meanwhile, Sjøvaag [26] designed a quantitative online news content analysis method by making a computer program, where the algorithm is adjusted to the needs of the desired analysis results. This approach requires strong coding and data mining skills, while not all companies have such human resources.

This paper proposes a method to analyze website contents, especially online news, by using the clustering method and then proceeds with analyzing the correlation by utilizing easy-

to-use software already available, i.e., RStudio for the clustering and MS Excel for the correlation analysis. The results of this study include how the stages in analyzing online news website content are using clustering and correlation. A case study is implemented as it is suitable for exploratory research [27] and appropriate to answer "how" or "why" research questions [28]. By employing case study research, researchers can explain or clarify a particular situation and get a clear understanding of a certain phenomenon [28], [29].

The case studies in implementing this method are also expected to inform what news categories can increase the number of visitors. These results can be used as considerations in optimizing the selected news categories. Additionally, it is also predicted that the popularity of the news category is influenced by the scope of the news site, whether at the local, national, or international level.

II. Materials and Method

This research begins with data collection activities that are used as case studies. The data are collected daily for approximately three months periods. Before clustering using the K-Means algorithm, the data were preprocessed by data generalization. Additionally, correlation analysis is directly carried out on the data to determine the value of the relationship between the variables. Correlation analysis is also carried out in each cluster, which continues the clustering stage. Figure 1 is the flow chart of this research.

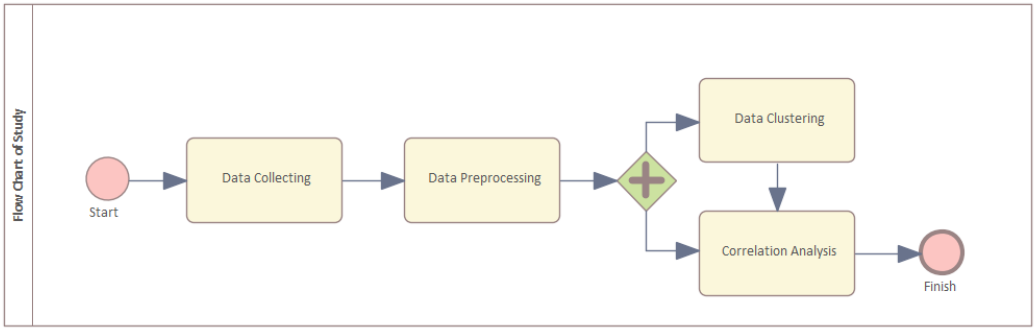

Fig. 1 The Flow Chart of The Study

A. Data Collecting

Our intern student collected the data from a local online newspaper company in Batam. The data collected every day for approximately three months using the Google Analytics tool [30]–[32] is the percentage of readers from each news headline published.

B. Data Preprocessing

Data preprocessing is one of the important stages in the data mining process [33], [34], which is related to data preparation that makes the knowledge mining process run smoothly and efficiently [35]–[37]. Data preprocessing has various ways, including data cleaning, integration, transformation, and reduction [38], [39]. One of the data reduction methods is data generalization [40], [41], which is used in this study. Data generalization creates a more broad categorization of data to obtain a more general picture of insights.

We obtained data on the percentage of readers from each news title during the data collection stage. Next, we preprocessed the data by generalizing each news title into 12 categories. These 12 categories are Health (Local; National), Economics (Local; National), Politics (Local; National), Law (Local; National), Criminal (Local; National), and Events (Local; National). Table 1 is an example of the generalization process of the data. Therefore, the raw daily readership data were compounded into categories instead of news article basis.

TABLE I
Data Generalization Process

News Title	Category											
	Local						National					
	Health	Economy	Politics	Law	Criminal	Event	Health	Economy	Politics	Law	Criminal	Event
Floods hit several areas of Singapore												√
Cipkon operation in Karimun targets teenagers hanging out					√							
Roro KMP Ark of the Archipelago Heritage of Kepri-Kalimantan		√										
Iskandar Syah insinuated to the Regent about the discourse of the Natuna Province			√									
The elderly are starting to threaten Singapore							√					
6 Important Things to Know About Herpes							√					
Suspicion of airport officials when these two women were intercepted					√							
Note! This is the Calendar of Linga Tourism Events Throughout 2020						√						
Roro KMP Bahtera Nusantara Connects Riau Islands-Kalimantan		√										

So Rajagukguk accuses Rudi of failing, Head of BP Batam: 100 days is not a benchmark		√			
China claims Natuna as..., Foreign Minister: Unfounded		√			
Buwas answered 5 questions, prayed for the President		√			
The uniqueness of the car in Batam plate X, Z, V, U and Y?			√		
Japanese Boss Attends The Launching of PT SIS Tomoe			√		
Does your name start with the letter S? This is the meaning			√		
Numerology: Numbers that reveal your love life	√				
Want to go home on a roro? This is the route and the boat fare	√				
Use agate in the cafe, eat as much free as you like	√				
Several areas of Singapore were hit by floods					√

In addition to 12 categories of news readership data, we add one more parameter, i.e., the number of readers. Therefore, the total number of parameters that are used is 13. They are:

- P1: Local Health
- P2: National Health
- P3: Local Economy
- P4: National Economy
- P5: Local Politics
- P6: National Politics
- P7: Local Law
- P8: National Law
- P9: Local Criminal
- P10: National Criminal
- P11: Local Event
- P12: National Event
- P13: The number of readers

Finally, we obtained 60 data ready for the clustering and the correlation analysis process. We presented a sample of 10 data in Table 2. Each row represents readership data in a day, and columns P1 to P12 represent the percentage of daily readership in those news categories. Meanwhile, P13 is the total number of readers on that day.

TABLE II
SAMPLE DATA

No	ID	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
1	J-07	0,00	2,11	1,89	0,00	2,11	0,00	0,00	0,00	4,22	0,00	1,32	6,84	95
2	J-09	0,00	1,24	0,00	0,00	0,00	5,79	0,83	0,00	0,83	3,72	0,83	1,82	121
3	J-10	2,50	0,83	1,94	0,00	1,67	0,00	0,00	0,83	4,72	3,33	0,83	2,50	120
4	J-13	0,00	0,00	0,00	0,00	1,43	0,00	0,48	0,00	5,92	0,00	1,73	0,95	210
5	J-14	3,45	0,00	1,38	0,00	1,73	0,00	0,00	0,00	1,61	0,00	4,05	1,38	145
6	J-16	0,00	0,00	3,91	0,00	4,69	0,00	1,56	0,00	0,78	0,00	2,21	1,96	128
7	J-17	0,00	1,32	2,05	0,00	0,00	0,00	0,00	0,00	0,88	0,00	2,44	1,32	114
8	J-18	0,00	0,00	2,10	0,00	5,95	0,00	0,70	0,70	2,10	0,00	3,08	1,75	143
9	J-20	0,00	0,00	4,84	0,00	1,61	0,00	0,92	0,00	2,92	0,00	4,51	1,07	217
10	J-21	0,00	0,00	3,33	0,00	2,31	0,00	3,08	0,77	2,31	0,00	2,69	1,03	130

C. Data Clustering

Data clustering is a method in data mining that clusters data into n groups [42]. In contrast to the classification that has a class label (supervised), clustering does not have a group label (unsupervised) [43], [44]. In this study, the clustering method with the K-Means algorithm was used. After being preprocessed, the data is clustered using the K-Means algorithm with the number of clusters = 3. The clustering process is carried out as follows [44]:

- Determine the number of clusters.
- Determine the center point of the cluster (use a random number for the initialization).
- Calculate the distance of each data point from the center of the cluster.
- Determine which data point will follow which cluster center by looking for the closest distance.
- Update the new cluster center by calculating the average of each parameter of all cluster members.
- Go back to step 3. The iteration stops if the cluster members are the same (unchanged) as the previous iteration.

The clustering process in this study utilizes the functions available in the RStudio library [45]. We can focus on the analysis process as the clustering process can be carried out by simply calling the needed functions.

D. Correlation Analysis

Correlation analysis is a term used to show associations or relationships or information about the proximity of two variables [46]. In the data preprocessing section, it is stated that there are 13 parameters used in this study. The correlation analysis carried out in this study determines the relationship between parameters P1-P12 (news category) and P13 (number of readers). This relationship can show which news categories may increase the readers' number. It reveals insights for owners or managers of online news sites in optimizing the contents based on news categories.

The correlation value is calculated using the correlation function in Excel [47] as in equation 1.

$$\text{Correlation Value}=\text{CORREL}(\text{array1};\text{array2}) \quad (1)$$

Array 1 in equation 1 is the independent variable, while array2 is the dependent variable. Several combinations of variables were correlated to find an interesting relationship among the variables. The correlation value ranges from -1 to 1. If the value is close to -1 or 1, the relationship is very strong.

Otherwise, the relationship is weak if close to 0 [48]. Figure 2 shows the strength and direction of the correlation values.

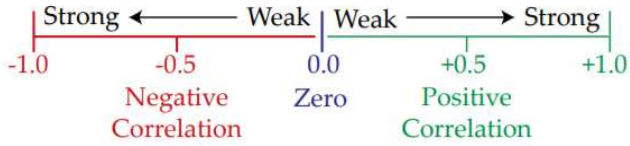

Fig. 2 Strength and Direction of Correlation

If the correlation value is close to -1, the relationship is inversely proportional. If the independent variable is higher, the value causes the dependent variable to be lower, and vice versa. On the other hand, if the correlation value is close to 1, the variable relationship is said to be directly proportional, which means the higher the value of the independent variable, the higher the value of the dependent variable, and vice versa.

III. Results and Discussion

The results and discussion in this section are divided into two sections, i.e., the cluster results section and the correlation analysis.

A. Clustering Results

After the data was clustered using the K-Means algorithm assisted by RStudio software, 3 clusters were obtained with 30 members of Cluster 1 (data number 4, 9, 12, 15, 19, 20, 22, 23, 24, 25, 26, 27, 29, 34, 35, 36, 37, 38, 39, 40, 43, 44, 45, 46, 47, 52, 53, 55, 56, 60), 24 members of Cluster 2 (data number 1, 2, 3, 5, 6, 7, 8, 10, 11, 13, 14, 16, 28, 30, 31, 41, 42, 48, 49, 50, 51, 54, 57, 58), and 6 members of Cluster (data number 17 , 18, 21, 32, 33, 59). Visualization of the results of this clustering can be seen in Figure 3. The calculation of the clusters average data is shown in Table 3.

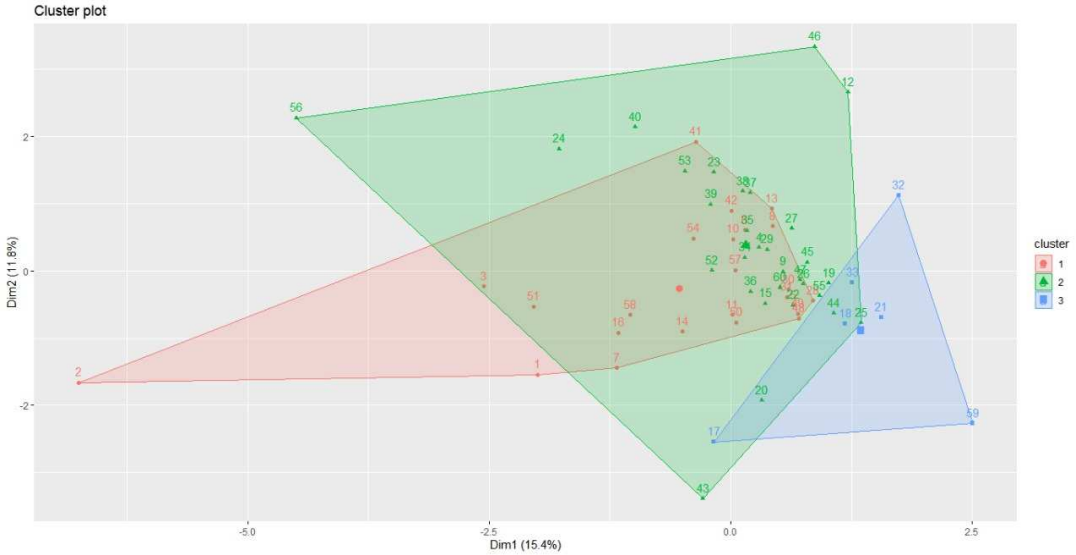

Fig. 3 Clustering Visualization

TABLE III
CLUSTERS AVERAGE

Cluster	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
1 (Middle)	2,25	0,11	2,73	0,68	2,40	0,06	3,39	0,03	3,17	0,28	2,74	0,56	266,67
2 (Low)	1,91	0,40	2,52	0,07	2,10	0,36	1,37	0,45	2,61	0,39	2,57	1,81	145,21
3 (High)	5,31	0,08	0,12	0,00	1,25	0,00	0,91	0,12	1,93	0,00	3,64	8,82	564,50

Table 3 shows that Cluster 1 produces a moderate number of readers, Cluster 2 produces a low number of readers, and Cluster 3 produces a high number of readers. During the 3-month duration of data collection, it turns out that only six days (number of Cluster 3 members) resulted in a high number of readers. To reveal more deeply the characteristics of each cluster, we conducted a correlation analysis.

B. Correlation Analysis

Calculating correlation values aims to determine the relationship between the dependent and the independent variables. In this study, the correlation value was calculated in several data groups: all data (Fig. 4), Cluster 1 data (Fig. 5), Cluster 2 data (Fig. 6), and Cluster 3 (Fig. 7).

1) Correlation of All Data: The correlation value obtained from all data is weak because it only has a value between -0.23 to 0.18. By carefully observing all the data again, this weak relationship occurs because there are many data with a value of 0, which means there is no news of that category published in the online newspaper. The highest inversely proportional correlation value is -0.23, owned by P2, the national health category. In comparison, the highest directly proportional correlation value is 0.18, owned by P1, which is the local health category. In this local-level online newspaper, readers tend to read local-level health instead of national-level health news.

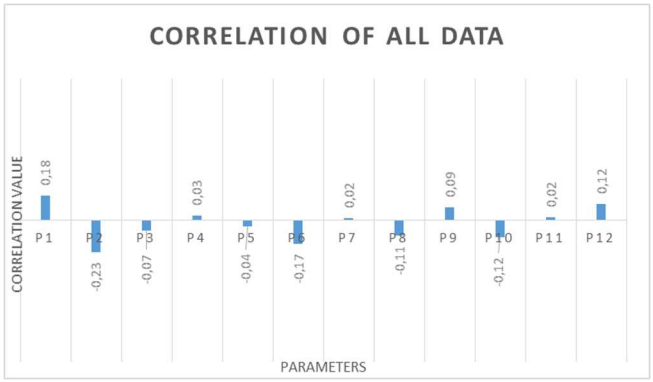

Fig. 4 Correlation of All Data

2) Correlation of Cluster 1: Looking at the Cluster 1 correlation value in Fig. 5, the relationship formed is not very strong, with the correlation value only ranging between -0.29 and 0.37. It is due to a large number of data having a value of 0, which means that news for that category was not published on that day. The highest inversely proportional correlation value is -0.29, owned by P12, the national event category, and it implies that news on national events does not increase the number of readers. Meanwhile, the highest correlation value is 0.37, owned by P3, the local economic news category. It means that local economic news can draw readers' interest.

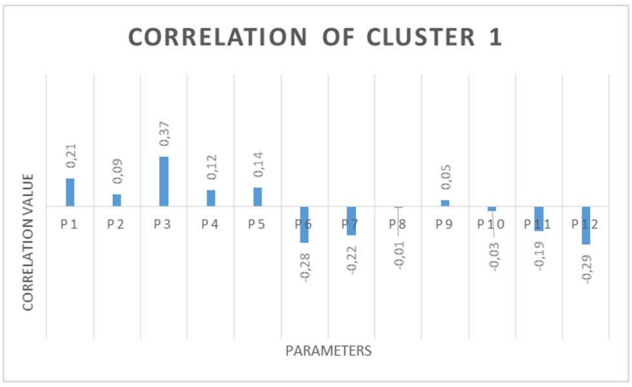

Fig. 5 Correlation of Cluster 1

These results also explain that reading national events news from local-level online newspaper sites is not preferable by the readers. On the other hand, the readers prefer to read local economic news from this local-level online newspaper.

3) Correlation of Cluster 2: The correlation value in Cluster 2 is not strong because the correlation value only ranges from -0.28 to 0.44. The highest inversely proportional correlation value is -0.28 for the national politics category, while the highest directly proportional correlation value is 0.44 for the local events category. It means that readers are more likely to read local events in this local-level online newspaper, while national politics news is less likely to be read in this local newspaper.

4) Correlation of Cluster 3: The correlation value in Cluster 3 ranges from -0.65 to 0.96. This relationship is quite strong because the correlation value is close to 1. The local event category (P11) has the highest inversely proportional correlation value of -0.65. The highest correlation value is 0.96 in the local crime news category. In addition, there is also another quite strong correlation value of 0.89, owned by P7, the local law news category. It can be concluded that the news categories that play a very important role in generating high readership for this local-level online newspaper are local criminal as well as local law categories. A careful look at the data in the Cluster 3 reveals that only six days out of 60-days data, the local online newspaper succeeded in gaining a high number of readers. Further data is needed to find patterns on when and how the high readership is made, whether it occurs periodically or happens only after major events that draw people's attention.

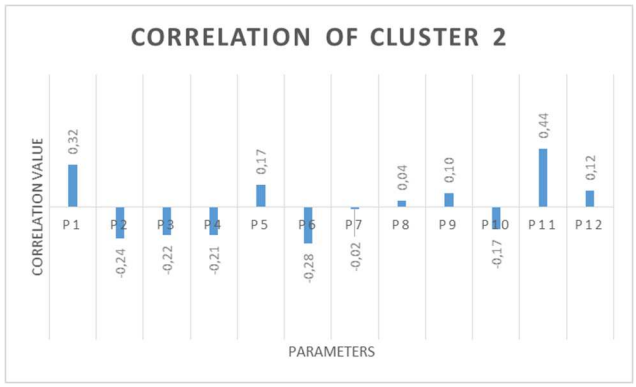

Fig. 6 Correlation of Cluster 2

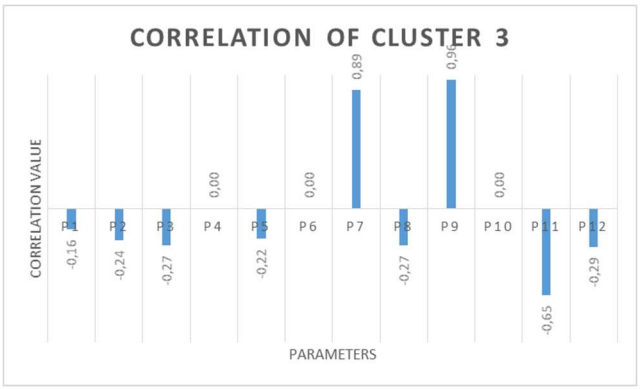

Fig. 7 Correlation of Cluster 3

IV. Conclusion

The clustering analysis formed 3 clusters, and each of these clusters provides information on their different characteristics. Cluster 1 is a cluster with a medium number of readers, cluster 2 is a group with a low number of readers, and cluster 3 has a high number of readers. Also, it is found that cluster 3 only consists of 6 data, which means that from 3 months, only six days have a high number of readers.

Correlation analysis was carried out from all data and each data cluster. The results infer readers' tendency to read news from a local-level online newspaper in Batam. The readers tend to look for local economic news, local events, local crime, local law, and local health. Meanwhile, the readers tend to choose other news sources for national-level news, such as national events, national politics, and national health.

Thus, the information gained from the results of this study can provide policy views to local online newspaper managers to further optimize local news, especially economics, events, crime, and law. It is to increase the number of readers, which can also increase the amount of companies' income.

Finally, some important points on how to better analyze online news content using clustering and correlation are as follows: Complete all data, which means that all news categories should be presented so that the data is complete and provides clearer analysis results; To focus on the results of the analysis of news contents, use case studies from various levels of online newspapers, be it local, national, or even international. More data inclusion enriches the dimension of comparison and analysis. The workflow of this study can be utilized in similar analyses in the future. The analysis process consists of collecting data, generalizing the data, part of the preprocessing, and clustering.

The clustering process is not limited to a certain algorithm, but researchers may use any clustering algorithm matched with the case study. Finally, the correlation analysis is done to deepen the analysis process. It can be carried out in the overall data and the clustered data.

ACKNOWLEDGMENT

The authors are grateful to the online newspaper company and Ester Afrida Dhuha for providing the data. We are also obliged to Politeknik Negeri Batam, our educational institution, to support facilities and funding for this research.

REFERENCES

[1] H. Sjøvaag, "Introducing the paywall: A case study of content changes in three online newspapers," *Journal. Pract.*, vol. 10, no. 3, pp. 304–322, 2016.

[2] R. K. Olsen and M. K. Solvoll, "Bouncing off the paywall--Understanding misalignments between local newspaper value propositions and audience responses," *Int. J. Media Manag.*, vol. 20, no. 3, pp. 174–192, 2018.

[3] J. A. Penny, "What Would Google Do?," *Pers. Psychol.*, vol. 63, no. 3, p. 809, 2010.

[4] R. Moro Visconti, "The Valuation of Newspaper Headings, Publishing Titles, and Copyright," in *The Valuation of Digital Intangibles*, Springer, 2020, pp. 267–291.

[5] D. Palau-Sampio, "Sponsored Content in Spanish Media: Strategies, Transparency, and Ethical Concerns," *Digit. Journal.*, vol. 9, no. 7, pp. 908–928, 2021.

[6] C. J. Vargo and M. A. Amazeen, "Agenda-Cutting Versus Agenda-Building: Does Sponsored Content Influence Corporate News Coverage in US Media?," *Int. J. Commun.*, vol. 15, p. 22, 2021.

[7] W.-Y. Lee, Y. Hur, D. Y. Kim, and C. Brigham, "The effect of endorsement and congruence on banner ads on sports websites," *Int. J. Sport. Mark. Spons.*, 2017.

[8] H. Kindermann, "A short-term twofold impact on banner ads," in *International Conference on HCI in Business, Government, and Organizations*, 2016, pp. 417–426.

[9] R. Sen, "Optimal search engine marketing strategy," *Int. J. Electron. Commer.*, vol. 10, no. 1, pp. 9–25, 2005.

[10] S. Das, *Search engine optimization and marketing: A recipe for success in digital marketing*. CRC Press, 2021.

[11] T. B. Clarke, J. Murphy, L. R. Wetsch, and H. Boeck, "Teaching search engine marketing through the google ad grants program," *Mark. Educ. Rev.*, vol. 28, no. 2, pp. 136–147, 2018.

[12] R. S. Bhandari and A. Bansal, "Impact of search engine optimization as a marketing tool," *Jindal J. Bus. Res.*, vol. 7, no. 1, pp. 23–36, 2018.

[13] C. Jie, Z. W. Da Xu, L. Wang, and W. Shen, "Bidding via clustering ads intentions: an efficient search engine marketing system for e-commerce," *arXiv Prepr. arXiv2106.12700*, 2021.

[14] T. L. Tuten and M. R. Solomon, *Social media marketing*. Sage, 2017.

[15] D. Evans, S. Bratton, and J. McKee, *Social media marketing*. AG Printing \& Publishing, 2021.

[16] R. Felix, P. A. Rauschnabel, and C. Hinsch, "Elements of strategic social media marketing: A holistic framework," *J. Bus. Res.*, vol. 70, pp. 118–126, 2017.

[17] Y. K. Dwivedi *et al.*, "Setting the future of digital and social media marketing research: Perspectives and research propositions," *Int. J. Inf. Manage.*, vol. 59, p. 102168, 2021.

[18] W. Tafesse and A. Wien, "Implementing social media marketing strategically: an empirical assessment," *J. Mark. Manag.*, vol. 34, no. 9–10, pp. 732–749, 2018.

[19] S. Sharma and M. Rai, "Customer Behaviour Analysis using Web Usage Mining," *Int J Sci Res Comput Sci Eng*, vol. 5, no. 6, pp. 47–50, 2017.

[20] M. Munk, A. Pilkova, L. Benko, P. Blazekova, and P. Svec, "Methodology of stakeholders' behaviour modelling based on time," *MethodsX*, vol. 8, p. 101570, 2021.

[21] S. Sharma, M. Rai, and others, "Comparative analysis of various tools to predict consumer behaviour," *J. Comput. Theor. Nanosci.*, vol. 16, no. 9, pp. 3860–3866, 2019.

[22] S. Mowla and N. P. Shetty, "Analysis of web server logs to understand internet user behaviour and develop digital marketing strategies," *Int. J. Eng. Technol.*, vol. 7, no. 4.41, pp. 15–21, 2018.

[23] S. J. Miah, H. Q. Vu, J. Gammack, and M. McGrath, "A big data analytics method for tourist behaviour analysis," *Inf. \& Manag.*, vol. 54, no. 6, pp. 771–785, 2017.

[24] V. Belair-Gagnon and A. E. Holton, "Boundary Work, Interloper Media, And Analytics In Newsrooms: An analysis of the roles of web analytics companies in news production," *Digit. Journal.*, vol. 6, no. 4, pp. 492–508, 2018, doi: 10.1080/21670811.2018.1445001.

[25] T. Y. Akhirina, A. Rusmardiana, D. Yulistyanti, F. G. Febrinanto, C. Dewi, and A. Triwiratno, "Popular Content Prediction Based on Web Visitor Data With Data Mining Approach Popular Content Prediction Based on Web Visitor Data With Data Mining Approach," pp. 0–7, doi: 10.1088/1742-6596/1641/1/012105.

[26] H. Sjøvaag and E. Stavelin, "Web media and the quantitative content analysis: Methodological challenges in measuring online news content," *Convergence*, vol. 18, no. 2, pp. 215–229, 2012, doi: 10.1177/1354856511429641.

[27] I. Benbasat, D. K. Goldstein, and M. Mead, "The Case Research Strategy in Studies of Information Systems," *MIS Q.*, vol. 11, no. 3, pp. 369–386, 1987, [Online]. Available: http://www.jstor.org/stable/248684.

[28] K. M. Eisenhardt, "Building theories from case study research," *Acad. Manag. Rev.*, vol. 14, no. 4, pp. 532–550, 1989.

[29] R. K. Yin, *Case study research: Design and methods*, vol. 5. sage, 2009.

[30] B. Plaza, "Google Analytics for measuring website performance," *Tour. Manag.*, vol. 32, no. 3, pp. 477–481, 2011.

[31] B. Clifton, *Advanced web metrics with Google Analytics*. John Wiley \& Sons, 2012.

[32] B. Mangold, *Learning Google AdWords and Google Analytics*. Loves Data, 2018.

[33] S. Garc\'\ia, J. Luengo, and F. Herrera, *Data preprocessing in data mining*, vol. 72. Springer, 2015.

[34] S.-A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Data preprocessing in predictive data mining," *Knowl. Eng. Rev.*, vol. 34, 2019.

[35] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017.

[36] S. Garc\'\ia, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Syst.*, vol. 98, pp. 1–29, 2016.

[37] Y. Rokhayati, U. H. B. Rusdi, D. E. Kurniawan, N. Z. Janah, and S. Irawan, "Analysis of SP students using AHP-Apriori combination," in *International Conference On Applied Science and Technology 2019-Social Sciences Track (iCASTSS 2019)*, 2019, pp. 214–219.

[38] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, "New data preprocessing trends based on ensemble of multiple preprocessing techniques," *TrAC - Trends Anal. Chem.*, vol. 132, p. 116045, 2020, doi: 10.1016/j.trac.2020.116045.

[39] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Anal.*, vol. 1, no. 1, p. 9, 2016, doi: 10.1186/s41044-016-0014-0.

[40] A. H. Bokhari, A. Y. Al-Dweik, F. D. Zaman, A. H. Kara, and F. M. Mahomed, "Generalization of the double reduction theory," *Nonlinear Anal. Real World Appl.*, vol. 11, no. 5, pp. 3763–3769, 2010, doi: 10.1016/j.nonrwa.2010.02.006.

[41] M. D. Sikirić, A. Schürmann, and F. Vallentin, "A Generalization of Voronoi's reduction theory and its application," *Duke Math. J.*, vol. 142, no. 1, pp. 127–164, 2008, doi: 10.1215/00127094-2008-003.

[42] P. Berkhin, "Survey of Clustering Data Mining Techniques," pp. 1–56.

[43] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," *33rd Int. Conf. Mach. Learn. ICML 2016*, vol. 1, pp. 740–749, 2016.
[44] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
[45] W. N. Arifin, "Introduction to R and RStudio IDE R and RStudio R packages Help," 2019.
[46] S. Senthilnathan, "Usefulness of Correlation Analysis," *SSRN Electron. J.*, no. July, 2019, doi: 10.2139/ssrn.3416918.
[47] D. Baus, "Correlation Analysis with Excel Handout," 2017.
[48] N. J. Gogtay and U. M. Thatte, "Principles of correlation analysis," *J. Assoc. Physicians India*, vol. 65, no. MARCH, pp. 78–81, 2017.