Implementation of the K-Nearest Neighbor (KNN) Algorithm in Making a Web-Based Article Topic System

Fanny Ramadhani[1], Oris Krianto Sulaiman[2]
[1]Department of Computer Science, Universitas Negeri Medan, Indonesia
[2] Department of Informatics Engineering, Universitas Islam Sumatera Utara, Indonesia

ABSTRACT

This study aims to apply the K-Nearest Neighbor algorithm in making an article topic determination system. The K-Nearest Neighbor algorithm is one of the oldest and most popular NN-based methods for categorizing text. In determining class label predictions on test data it is determined by the value of k which represents the number of nearest neighbors. Of the k closest neighbors selected, voting is carried out by selecting the class with the most number as the predicted class label on the test data. Classification is considered the best method in processing when the training data is closest to the object. The way the KNN works requires determining the input in the form of training data, test data and k values.

Keyword : Article, K-Nearest Neighbor, Topic.

Corresponding Author:
Fanny Ramadhani,
Department of Computer Science,
Universitas Negeri Medan
Jalan William Iskandar Ps. V, Kenangan Baru, Sumatera Utara 20221 , Indonesia.
Email: fannyramadhani@unimed.ac.id

1. INTRODUCTION

Technology is very useful to support daily activities. The development of technology makes it easier for humans to communicate and get information. One of the information that is easy to obtain is information contained in the form of articles.

An article is a fact that is analyzed so as to give rise to the author's opinion or view of the fact being analyzed or it can also be called an opinion conveyed by a writer about an actual problem that has caught the public's attention. Articles contain ideas that aim to inform, influence, convince, and entertain [1].

Previous research has discussed the application of determining article topics using the K-Nearest Neighbor (KNN) algorithm. One of them is the journal made by Yoseph Samuel, et al (2015) in his journal it is explained that news documents will be made topical so that searching for news according to the topic desired by readers is easier. The K-Nearest Neighbor (KNN) algorithm usually uses a majority vote as a basis for determining where a document is classified. But here, the researcher replaces the use of majority vote with a decision rule with the hope that the use of the K-Nearest Neighor algorithm can be maximized. The news that will be used as training data is taken from three websites, namely: bbc.com, cnn.com, and foxnews.com. The incoming news will be categorized based on sports topics which are divided into 7 subtopics namely: Soccer, Formula 1, Basketball, Motorsport, Baseball, Tennis, and NFL. Meanwhile, the news used as training data is 280 news and 95 new news will be used for testing [2].

Based on some of the previous studies above, the author will conduct research by applying the K-Nearest Neighbor algorithm to determine the topic of an article based on knowledge clusters.

2. RESEARCH METHOD/MATERIAL AND METHOD/LETERATURE REVIEW (10 PT)

The research methodology in this final report is as follows:

a. Method of collecting data

The data is needed to support solving problems that arise based on research focus. These data were obtained using the method of literature study. Literature study is a method used to collect data or

sources related to the topic raised in this research. Literature studies can be obtained from various sources, such as journals, documentation books, the internet and libraries.

b. System Development Process Model

The process in developing this system uses the modified waterfall model. The following is a picture of the modified waterfall model [3].

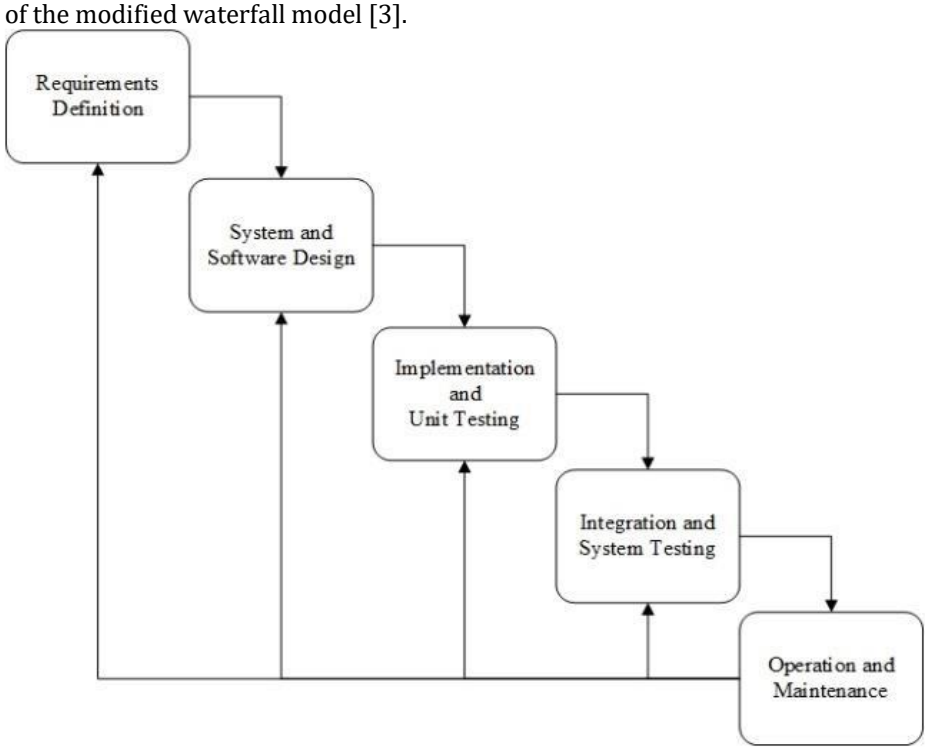

Figure 1. System Development Process Model

Based on the waterfall model above, the stages of the research procedure that will be carried out are Requirement Definition, System and Software Design, Implementation and Unit Testing, Integration and System Testing, and Operation and Maintenance stages.

The system to be built is an article topic determination system that uses the K-Nearest Neighbor algorithm. As for the general system flowchart, it can be seen in Figure 2.

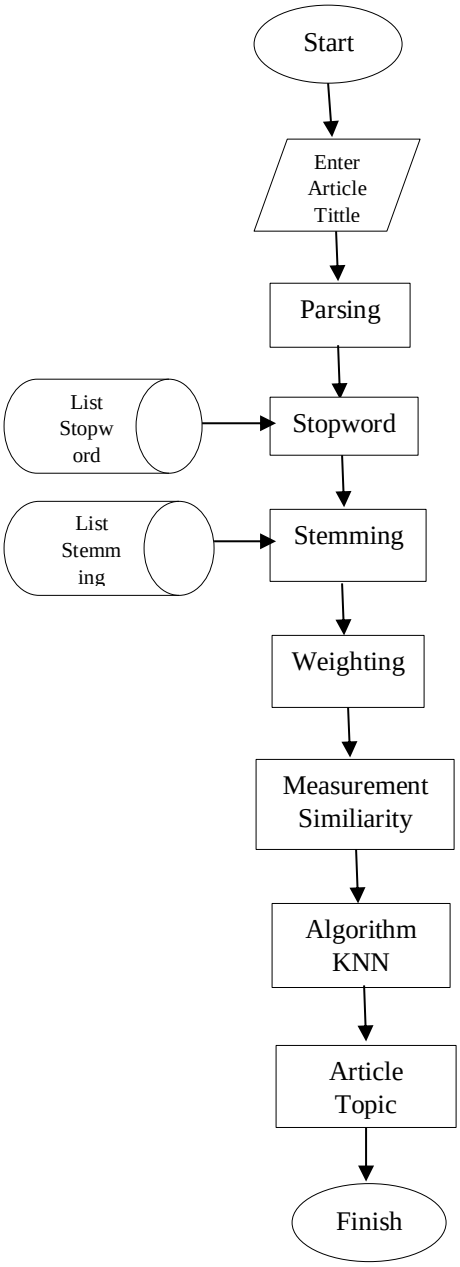

Figure 2. Flowchart for determining the topic of the article

Figure 2 shows the entire process of the system starting from inputting the title of the article, parsing, removing stopwords, stemming, weighting, measuring similarity using dicesimilarity, the KNN algorithm and the output being the topic of the article.

The system process stages described in Figure 2 consist of pre-processing processes, namely parsing, stopwords, and stemming.

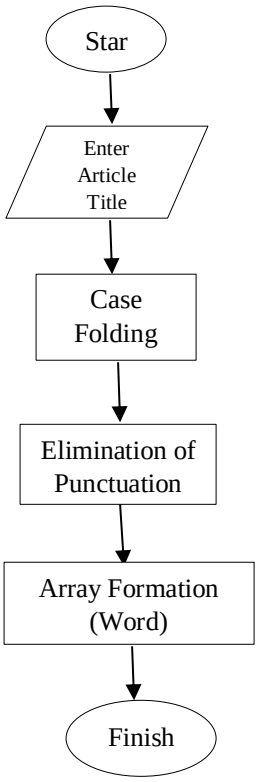

Figure 3. Flowchart parsing process

Figure 3 is a parsing process flowchart. In the early stages of the parsing process, namely case folding, which is to make all the letters in the sentence or text inputted lowercase, this is done to reduce the size of the database on the index. Then the next process is removing punctuation marks such as commas (,), periods (.), and so on. After that, the input text in the form of sentences is split into words or arrays. The output generated in this process is an array that will be used for the next process.

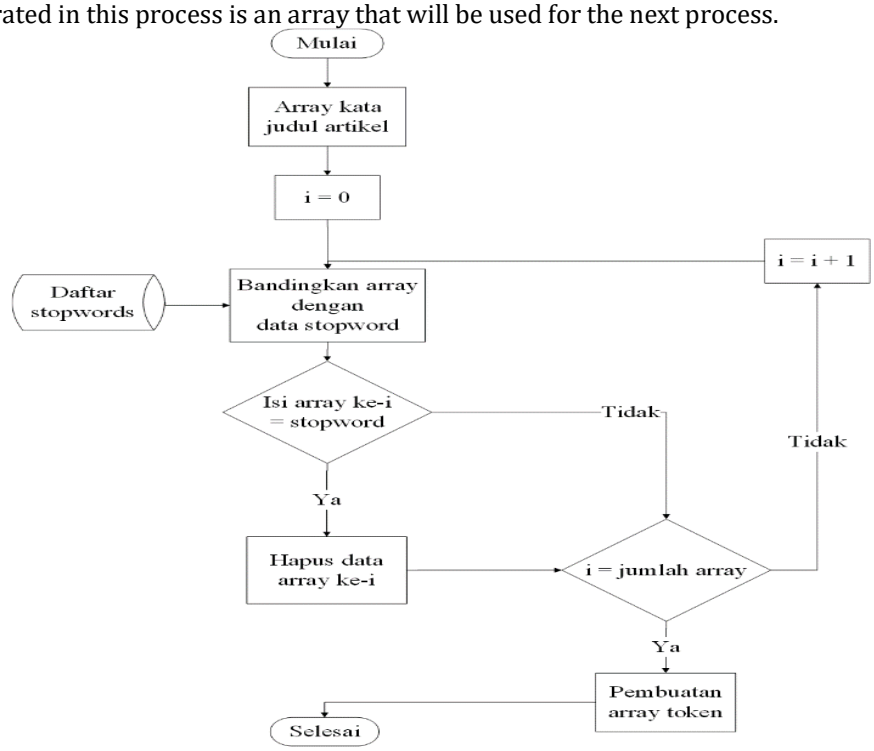

Figure 4. Flowchartstopword

Figure 4 is the process of removing stopwords which is done to remove words that are considered not important in the input sentence. Read the array of words obtained from the parsing process.

1. Retrieve a list of stopwords in the database.
2. Compare the array with the list of stopwords. If the word is included in the stopwords list, then remove the word from the array.
3. Repeat the third step until the last array.

The next process is a stemming process that is carried out to find the base word of a word. The stemming process is carried out by removing all affixes consisting of prefixes and suffixes. Stemming is obtained from the array of words resulting from the process of removing stopwords. The algorithm is as follows:

1. Read the array of words obtained from the process of removing stopwords.
2. Remove particles and remove pronouns
3. If the array has prefix 1 then delete prefix 1, if not there, then delete prefix 2 then delete suffix.
4. If the array has a suffix, then delete the suffix and then remove the prefix 2
5. Then match with the basic word dictionary in the database.
6. Repeat step 2 until the last word array.

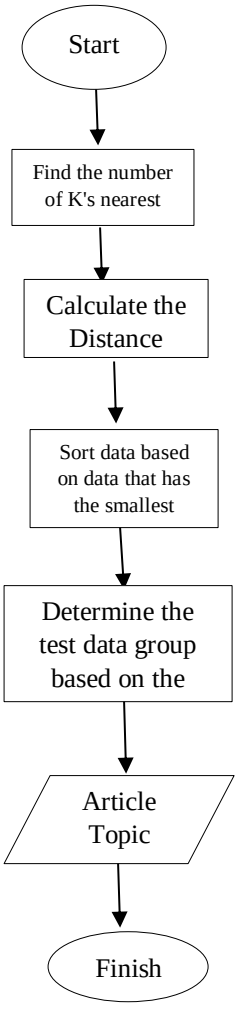

Figure 5. Flowchart KNN

3. RESULTS AND DISCUSSION

System Implementation

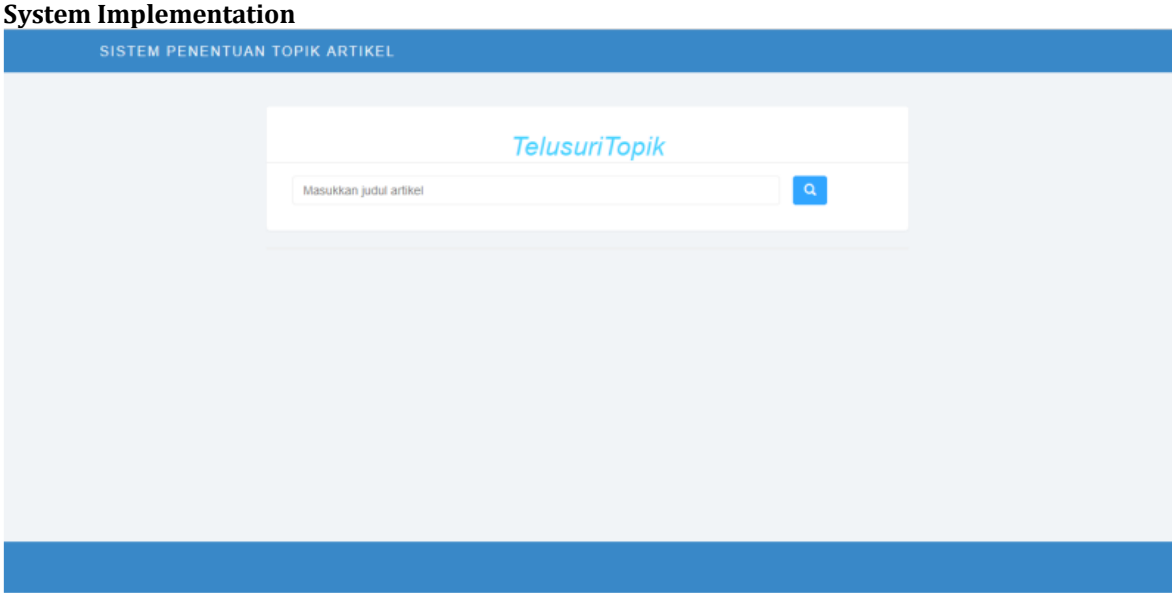

Figure 6. Usser View

Figure 6 displays the user display page where there is a column for entering the title of the article that will display the topic and pressing the search button to display the topic search results page.

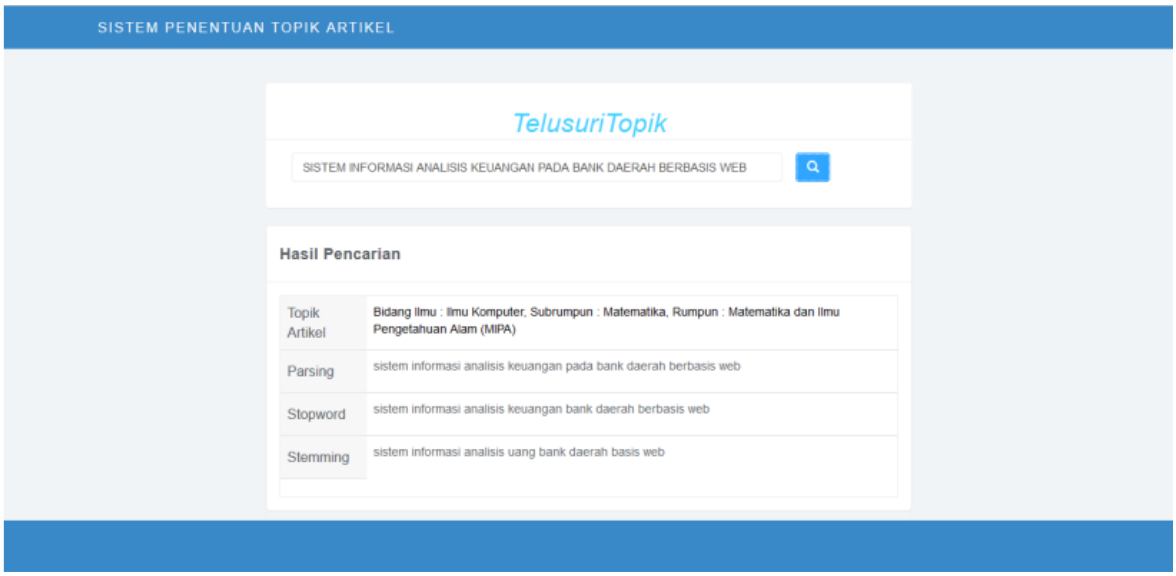

Figure 7 Display of Search Results

Furthermore, in Figure 7 it will display a search page from the input query that was previously entered. In the search results, the topic of the article will appear, namely the science field, subgroup of knowledge, and the family of the article title entered. Then there are several processes that will also be shown to the user, namely the parsing process, the stopword removal process, and the stemming process.

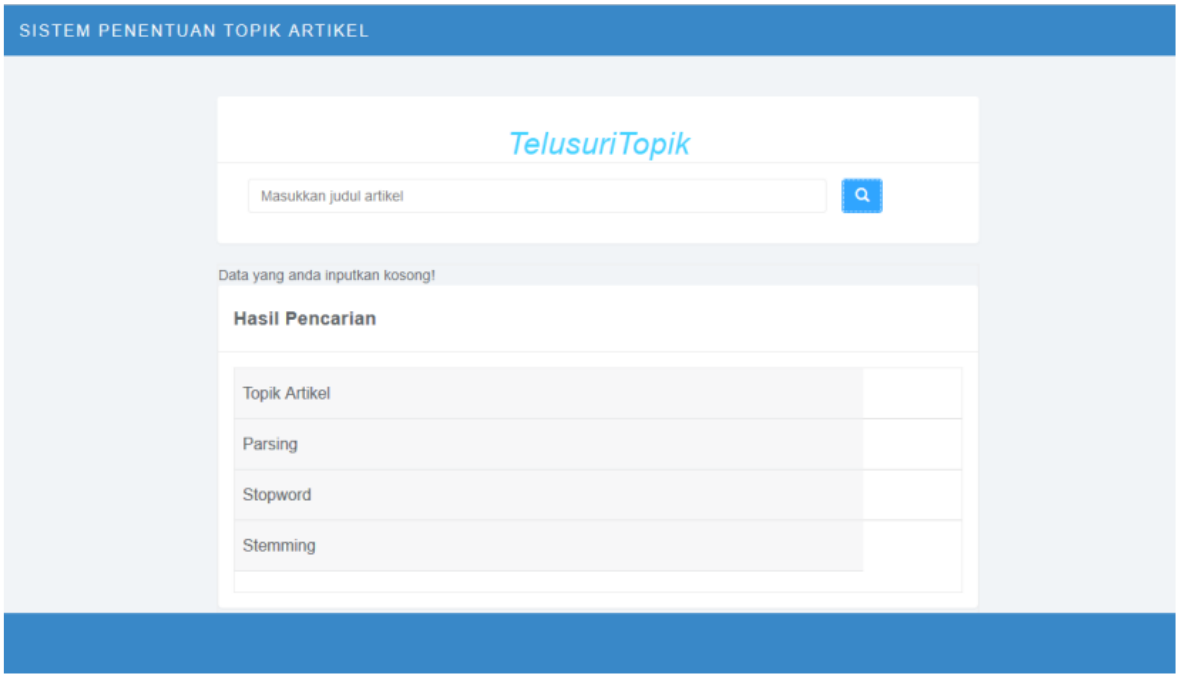

Figure 8 The Input Display is Empty

Figure 8 shows the page that appears if the user searches for the topic of the article but has not previously entered an input query in the column provided.

4. CONCLUSION

Based on the results of the research and implementation of the article topic determination system using the K-NearestNeighbor algorithm, it can be concluded that:

1. The results showed that the algorithm applied, namely the K-Nearest Neighbor algorithm, was successfully implemented in the article topic determination system.
2. The system for determining the topic of the article is used to find out the topic of an article based on the field of science, sub-group of science, and science family.

REFERENCES

[1] Sari, I.P., Al-Khowarizmi, A., & Batubara, I.H (2021). Cluster Analysis Using K-Means Algorithm and Fuzzy C-Means Clustering For Grouping Students' Abilities In Online Learning Process. *Journal of Computer Science, Information Technology and Telecommunication Engineering*, 2(1), 139-144.

[2] Lubis, A. R., Lubis, M. & Al-Khowarizmi (2020). Optimization of distance formula in K-Nearest Neighbor method. *Bulletin of Electrical Engineering and Informatics*, *9*(1), 326-338.

[3] Sari, I.P., Batubara, I.H., & Al-Khowarizmi, A (2021). Sensitivity Of Obtaining Errors In The Combination Of Fuzzy And Neural Networks For Conducting Student Assessment On E-Learning. *International Journal of Economic, Technology and Social Sciences (Injects)*, 2(1), 331-338.

[4] Y. Samuel, R. Delima, and A. Rachmat, "Implementasi Metode K-Nearest Neighbor dengan Decision Rule untuk Klasifikasi Subtopik Berita," J. Inform., vol. 10, 2015.

[5] Sari, I.P., Fahroza, M.F., Mufit, M.I., & Qathrunad, I.F (2021). Implementation of Dijkstra's Algorithm to Determine the Shortest Route in a City. *Journal of Computer Science, Information Technology and Telecommunication Engineering*, 2(1), 134-138.

[6] Batubara, I.H., Saragih, S., Syahputra, E., Armanto, D., Sari, I.P., Lubis,B.S., & Siregar, E.F.S (2022). Mapping Research Developments on Mathematics Communication: Bibliometric Study by VosViewer. *AL-ISHLAH: Jurnal Pendidikan* 14(3), 2637-2648.

[7] Sari, I.P., Al-Khowarizmi, A.K., & Batubara, I.H. (2021). Analisa Sistem Kendali Pemanfaatan Raspberry Pi sebagai Server Web untuk Pengontrol Arus Listrik Jarak Jauh. *InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan,* 6 (1), 99-103.

[8] Hariani.,P.P, Sari.,I.P, & Batubara., I.H. (2021). Implementasi e-Financial Report BUMDes. *IHSAN: JURNAL PENGABDIAN MASYARAKAT,* 3 (2), 169-177.

[9] Sari, I.P., Basri, Mhd., Ramadhani, F., & Manurung, A.A. (2023). Penerapan Palang Pintu Otomatis Jarak Jauh Berbasis RFID di Perumahan. *Blend Sains Jurnal Teknik*, 2(1), 16-25.
[10] Batubara, I.H., & Sari, I.P. (2021). Penggunaan software geogebra untuk meningkatkan kemampuan pemecahan masalah matematis mahasiswa. *Scenario (Seminar of Social Sciences Engineering and Humaniora),* 398-406
[11] Sari, I.P., & Batubara, I.H. (2020). Aplikasi Berbasis Teknologi Raspberry Pi Dalam Manajemen Kehadiran Siswa Berbasis Pengenalan Wajah. *JMP-DMT* 1(4), 6.
[12] Sari, I.P., Al-Khowarizmi, A.K., Ramadhani, F., & Sulaiman, O.K. (2023). Implementation of the Selection Sort Algorithm to Sort Data in PHP Programming Language. *Journal of Computer Science, Information Technology and Telecommunication Engineering*, 4(1).
[13] Batubara, I.H., Sari,I.P., Hariani, P.P., Saragih, M., Novita, A., Lubis, B.S., & Siregar, E.F.S. (2021). Pelatihan Software Geogebra untuk Meningkatkan Kualitas Pembelajaran Matematika SMP Free Methodist 2. Martabe: Jurnal Pengabdian Kepada Masyarakat, 4(3), 854-859.
[14] Sari., I.P, Batubara., I.P, Al-Khowarizmi., A, & PP Hariani. (2022). Perancangan Sistem Informasi Pengelolaan Arsip Digital Berbasis Web untuk Mengatur Sistem Kearsipan di SMK Tri Karya. Wahana Jurnal Pengabdian kepada Masyarakat 1 (1), 18-24.
[15] Batubara., I.H, Sari., I.P, EFS Siregar, & BS Lubis. (2021). Meningkatkan Kemampuan Penalaran Matematika Melalui Metode Penemuan Terpandu Berbantuan Software Autograph. Seminar Nasional Teknologi Edukasi Sosial dan Humaniora 1 (1), 699-705.
[16] Sari., I.P, A Syahputra, N Zaky, RU Sibuea, & Z Zakhir. (2022). Perancangan sistem aplikasi penjualan dan layanan jasa laundry sepatu berbasis website. Blend sains jurnal teknik 1 (1), 31-37.
[17] Sari., I.P, A Azzahrah, FQ Isnaini, L Nurkumala, & A Thamita. (2022). Perancangan sistem absensi pegawai kantoran secara online pada website berbasis HTML dan CSS. Blend sains jurnal teknik 1 (1), 8-15.
[18] Ramadhani., F, & Sari., I.P. (2021). Pemanfaatan Aplikasi Online dalam Digitalisasi Pasar Tradisional di Medan. Prosiding Seminar Nasional Kewirausahaan 2 (1), 806-811.
[19] Sari.,I.P, & Ramadhani., F. (2021). Pengaruh Teknologi Informasi Terhadap Kewirausahaan Pada Aplikasi Perancangan Jual Beli Jamu Berbasis WEB. Prosiding Seminar Nasional Kewirausahaan 2 (1), 874-878.
[20] Sari., I.P, A Jannah, AM Meuraxa, A Syahfitri, & R Omar. (2022). Perancangan Sistem Informasi Penginputan Database Mahasiswa Berbasis Web. Hello World Jurnal Ilmu Komputer 1 (2), 106-110.
[21] Hutasuhut, B.K., Sari, I.P., & Al-Khowarizmi, A (2023). Analysis the Effect of Digitalization and Technology on Web-Based Entrepreneurship. *Journal of Computer Science, Information Technology and Telecommunication Engineering* 4(1).
[22] Sari., I.P, & Batubara., I.H. (2021). Perancangan Sistem Informasi Laporan Keuangan Pada Apotek Menggunakan Algoritma K-NN. *Seminar Nasional Teknologi Edukasi dan Humaniora (SiNTESa)* 1 (2021 - ke 1.
[23] Ramadhani., F, A Satria, & Sari., I.P. (2022). Aplikasi Internet Berbasis Website sebagai E-Commerce Penjualan Komponen Sport Car. *Blend Sains Jurnal Teknik* 1 (2), 69-75.
[24] Sari., I.P, & Batubara., I.H. (2021). User Interface Information System for Using Account Services (Joint Account) WEB-Based. *International Journal of Economic, Technology and Social Sciences (Injects),* 462-469.
[25] PP Hariani, Sari., I.P, & Batubara., I.H. (2021). Android-Based Financial Statement Presentation Model. *JURNAL TARBIYAH* 28 (2), 1-16.
[26] Sari., I.P, Batubara., I.H, & M Basri. (2022). Implementasi Internet of Things Berbasis Website dalam Pemesanan Jasa Rumah Service Teknisi Komputer dan Jaringan Komputer. *Blend Sains Jurnal Teknik* 1 (2), 157-163.
[27] Ramadhani, F., Satria, A., & Sari, I.P (2023). Implementasi Metode Fuzzy K-Nearest Neighbor dalam Klasifikasi Penyakit Demam Berdarah. *Hello World Jurnal Ilmu Komputer* 2(2), 58-62.
[28] Sari., I.P, Al-Khowarizmi., A, & Batubara., I.H. (2021). Implementasi Aplikasi Mobile Learning Sistem Manajemen Soal dan Ujian Berbasis Web Pada Platform Android. *IHSAN: JURNAL PENGABDIAN MASYARAKAT* 3 (2), 178-183.
[29] Batubara, I.H., Saragih, S., Simamora, E., Napitupulu, E.E., Sari, I.P. (2022). Analysis of Student's Mathematical Communication Skills through Problem Based Learning Models Assisted by Augmented Reality. *Budapest International Research and Critics Institute-Journal (BIRCI-Journal)*, 5(1), 1024-1037.