Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD
Volume 8 ; Nomor 1 ; Januari 2025 ; Page 18-24
E-ISSN : 2615-5133 **; P-ISSN** : 2621-8976
https://ojs.trigunadharma.ac.id/index.php/jsk/index

Combination of TF-IDF and Rabin-Karp for Detecting Document Similarity in Student Thesis Abstracts

Pujo Hari Saputro1*, Fransisca Joanet Pontoh[2], Olivia Maria Tumurang[3]
[1,2] Informatic Engineering, Sam ratulangi University
[3]Civil Engineering, Sam ratulangi University
Email: [1] pujoharisaputro@unsrat.ac.id, [2]fransisca@unsrat.ac.id, [3]oliviatumurang@unsrat.ac.id
Email Penulis Korespondensi: pujoharisaputro@unsrat.ac.id

Article History:
Received Dec 12th, 2024
Revised Jan 9th, 2025
Accepted Jan 25th, 2025

Abstrak

Mahasiswa semester akhir diwajibkan untuk menyelesaikan tugas akhir dalam bentuk penelitian yang relevan dengan bidang studi masing-masing, untuk menemukan solusi inovatif dan mengembangkan kemampuan berpikir kritis. Namun, plagiarisme sering kali menjadi masalah yang muncul. Plagiarisme didefinisikan sebagai tindakan mengambil karya orang lain, termasuk pendapat, dan mengklaimnya sebagai milik sendiri. Oleh karena itu, teknologi dapat digunakan untuk mendeteksi kesamaan pada abstrak manuskrip mahasiswa yang diajukan saat pengajuan judul skripsi, sehingga memungkinkan deteksi dini terhadap plagiarisme. Korpus yang digunakan diambil dari direktori tugas akhir Program Studi Teknik Komputer, yang terdiri dari 98 data, dan Program Studi Teknik Sipil, yang terdiri dari 40 data. Dalam penelitian ini, dengan memanfaatkan algoritma TF-IDF dan Rabin-Karp, ditemukan bahwa TF-IDF mampu mendeteksi pentingnya suatu kata dalam dokumen relatif terhadap keseluruhan korpus. Algoritma Rabin-Karp juga terbukti efektif dalam mendeteksi pola yang cocok pada beberapa korpus, dengan tingkat akurasi pola yang cocok sebesar 70%.
Kata Kunci : abstraksi skripsi, plagiatisme, TF-IDF, Rabin Karp

Abstract

Final semester students are required to complete a final project in the form of research relevant to their respective fields of study, to find innovative solutions, and to develop critical thinking skills. However, plagiarism is a common problem that often arises. Plagiarism is defined as the act of taking someone else's work, including opinions, and claiming it as one's own. Therefore, technology can be used to detect similarities in the abstracts of student manuscripts submitted during thesis title submissions, allowing for early detection of plagiarism. The corpus used was taken from the directory of final projects from the Computer Engineering Study Program, consisting of 98 data points, and from the Civil Engineering Study Program, consisting of 40 data points. In this study, utilizing the TF-IDF and Rabin-Karp algorithms, it was found that TF-IDF is capable of detecting the importance of a word in a document relative to the entire corpus. Rabin-Karp has also proven effective in detecting matching patterns in several corpuses, with a known pattern matching accuracy of 70%.
Keyword : *abstact thesis, plagiarism, TF-IDF, Rabin-Karp.*

1. INTRODUCTION

A student's final project is an integral part of the higher education curriculum, aimed at testing the student's understanding and competence in their chosen field of study. This process involves the planning, execution, and analysis of research or a project independently conducted by the student under the guidance of an academic supervisor [1], [2]. In the final project, students are expected to identify relevant problems, formulate clear objectives, systematically collect data, and produce a comprehensive and accurate final report. Besides being an opportunity to apply the theories learned, the final project also serves as a platform to develop analytical skills, critical thinking, and communication abilities as students present and defend their work [3], [4].

In maintaining the originality and creativity of students in conducting research and preparing their final project manuscripts, rules and ethical codes related to these matters are established. One of the key issues is plagiarism. In Indonesia's Constitution No. 20 of 2003, Article 25, paragraph 2, and Article 70, the consequences of committing plagiarism are outlined. Furthermore, in the Regulation of the Minister of Education, Culture, Research, and Technology of the Republic of Indonesia No. 1 of 2023, it is explained that individuals involved in plagiarism cannot be appointed to

Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD
Volume 8 ; Nomor 1 ; Januari 2025 ; Page 18-24
E-ISSN : 2615-5133 **; P-ISSN** : 2621-8976
https://ojs.trigunadharma.ac.id/index.php/jsk/index

certain positions in government institutions. Sam Ratulangi University, as an educational institution with integrity that strives to create excellent education and research, also supports the implementation of anti-plagiarism measures in both faculty and student research. This is reflected in the Rector's Regulation of Sam Ratulangi University No. 02/UN12/KP/2016 regarding the code of ethics. To achieve this, further actions are required to promote creative, innovative, and plagiarism-free research.

The approach attempted by the author is a data processing or data mining approach, aimed at providing useful information as a reference for the preparation of future final projects [5],[6]. The data processing approach is expected to provide information on the level of similarity of final projects previously completed by students of the Faculty of Engineering at Sam Ratulangi University. This information will be beneficial for the study programs as a reference in approving students' future final project titles. Data processing or data mining is a field of study that examines patterns in existing data to obtain desired information according to the algorithm used [7], [8]. In this research, the author uses a combination of the Rabin-Karp and TF-IDF algorithms. The TF-IDF algorithm is used to represent documents with certain weights [9], [10], [11], while the Rabin-Karp algorithm is used for pattern searching and similarity matching [12], [13], [14].

2. METHODOLOGY

The method applied in this research uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) approach. CRISP-DM is a widely recognized framework that has been established as a standard methodology for conducting data mining processes. Although there are various approaches that can be chosen for data mining, such as Domain Specific Methodology, Knowledge Discovery in Databases (KDD), and Sample, Explore, Modify, Model, and Assess (SEMMA), CRISP-DM is considered one of the most effective models.

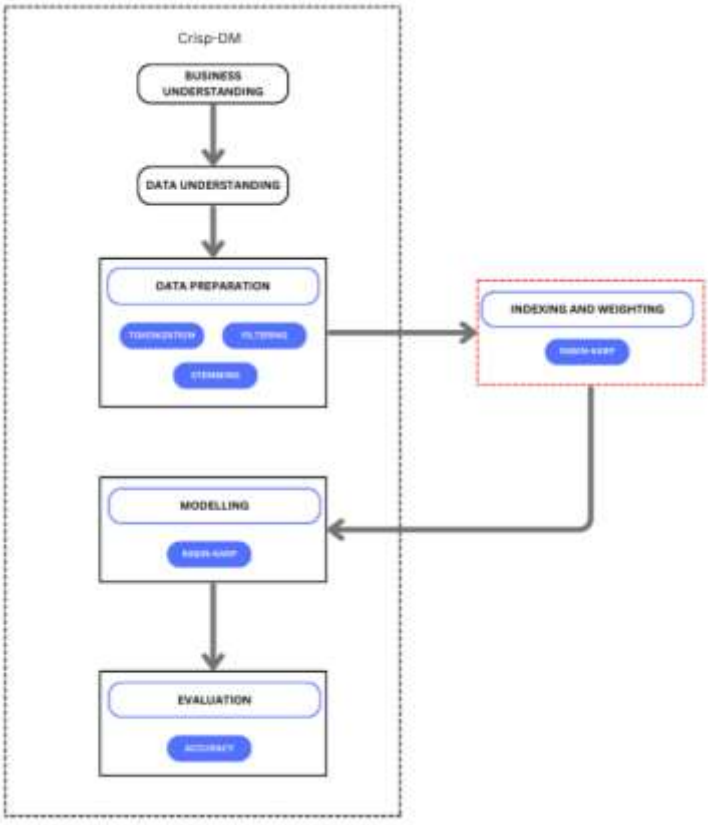

Figure 1. Methodology

The detailed methodology flow is discussed in the following section.

3. RESULT AND DISCUSSION

3.1 Business Understanding

This process is conducted to gain a thorough understanding of previous studies, which serves as a foundation for identifying knowledge gaps, refining research questions, and guiding the methodology. By analyzing and synthesizing existing literature, researchers can build upon established findings, address unresolved issues, and adapt relevant theories or techniques. This comprehensive comprehension not only ensures continuity in the research field but also facilitates the

integration of past insights into the current study, enabling meaningful advancements and practical applications in the ongoing research phase..

3.2 Data Understanding

The data for this study was sourced from the repository unsrat.ac.id, specifically from the student thesis repository. This repository serves as a comprehensive database of academic works, providing valuable insights into research trends and patterns over the years. The data spans a four-year period and focuses on theses from two distinct academic programs: the Informatics Engineering Program and the Civil Engineering Program. These programs were selected to represent diverse fields of study, offering a broad perspective on the academic contributions of students.

In total, 98 data points were collected from the Informatics Engineering Program, highlighting the scope and focus of research within this discipline. Similarly, 30 data points were gathered from the Civil Engineering Program, reflecting the specific interests and challenges unique to this field. The curated dataset serves as a strong foundation for analyzing academic productivity and thematic trends across these programs, supporting a deeper understanding of the contributions and progress in each domain..

3.3 Data Preparation

The collected data will then undergo processing during the preparation stage to ensure it is organized and ready for analysis. This stage involves various steps, including cleaning, structuring, and transforming the raw data into a usable format. Below is an example of the data preparation process applied to one of the data corpora, specifically focusing on the tasks of **indexing and weighting**. Indexing involves organizing the data into a searchable structure, making it easier to retrieve relevant information. Weighting, on the other hand, assigns importance to specific elements within the data, which is crucial for enhancing the accuracy and relevance of subsequent analyses or algorithms..

Table 1. Raw Data

Corpus Document
Penggunaan plastik kemasan yang semakin luas menyebabkan peningkatan jumlah sampah plastik yang dibuang secara tidak teratur dan mengancam lingkungan. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan sebuah aplikasi android yang menggunakan teknologi quick respons code untuk memonitoring pembuangan sampah plastik kemasan. Metode penelitian yang digunakan meliputi studi literatur tentang pengelolaan sampah plastik, analisis kebutuhan pengguna melalui survei dan wawancara, perancangan dan pengembangan aplikasi android, serta pengujian aplikasi melalui pengujian fungsionalitas dan pengujian pengguna. Aplikasi ini memungkinkan pengguna untuk memindai quick respons code yang terpasang pada produk kemasan dan tempat pembuangan sampah. Melalui pemindaian quick respons code, pengguna dapat mencatat lokasi dan jenis sampah plastik yang dibuang dengan lebih akurat. Aplikasi ini juga menyediakan fitur pengumpulan data tentang jumlah sampah plastik yang terkumpul di setiap lokasi, serta memberikan informasi edukatif tentang pentingnya pengelolaan sampah plastik. Evaluasi dan analisis data dilakukan untuk mengukur efektivitas dan kepuasan pengguna terhadap aplikasi. Penelitian ini memberikan kontribusi dalam pengembangan teknologi yang dapat memfasilitasi pemantauan dan pengolaan sampah plastik kemasan. Diharapkan bahwa aplikasi ini dapat digunakan oleh masyarakat dan pihak terkait dalam upaya pengurangan sampah plastik dan pelestarian lingkungan. Dimasa depan, aplikasi ini memiliki potensi untuk diperluas dan ditingkatkan sesuai dengan perkembangan teknologi dan kebutuhan pengguna

Source : (Repository unsrat.ac.id)

a. Tokenization

Tokenization is the process of converting a description, originally in the form of a sentence, into individual words or tokens. This process breaks down complex text into smaller, more manageable units that can be more easily analyzed and processed. By converting a sentence into tokens, tokenization enables each word to be treated independently, which is a crucial step in various text processing tasks such as text analysis, machine learning, and natural language understanding..

Table 2. Tokenization

Corpus Document
'Penggunaan', 'plastik', 'kemasan', 'yang', 'semakin', 'luas', 'menyebabkan', 'peningkatan', 'jumlah', 'sampah', 'plastik', 'yang', 'dibuang', 'secara', 'tidak', 'teratur', 'dan', 'mengancam', 'lingkungan', 'Oleh', 'karena', 'itu', 'penelitian', 'ini', 'bertujuan', 'untuk', 'mengembangkan', 'sebuah', 'aplikasi', 'android', 'yang', 'menggunakan', 'teknologi', 'quick', 'respons', 'code', 'untuk', 'memonitoring', 'pembuangan', 'sampah', 'plastik', 'kemasan', 'Metode', 'penelitian', 'yang', 'digunakan', 'meliputi', 'studi', 'literatur', 'tentang', 'pengelolaan', 'sampah', 'plastik', 'analisis', 'kebutuhan',

'pengguna', 'melalui', 'survei', 'dan', 'wawancara', 'perancangan', 'dan', 'pengembangan', 'aplikasi', 'android', 'serta', 'pengujian', 'aplikasi', 'melalui', 'pengujian', 'fungsionalitas', 'dan', 'pengujian', 'pengguna', 'Aplikasi', 'ini', 'memungkinkan', 'pengguna', 'untuk', 'memindai', 'quick', 'respons', 'code', 'yang', 'terpasang', 'pada', 'produk', 'kemasan', 'dan', 'tempat', 'pembuangan', 'sampah', 'Melalui', 'pemindaian', 'quick', 'respons', 'code', 'pengguna', 'dapat', 'mencatat', 'lokasi', 'dan', 'jenis', 'sampah', 'plastik', 'yang', 'dibuang', 'dengan', 'lebih', 'akurat', 'Aplikasi', 'ini', 'juga', 'menyediakan', 'fitur', 'pengumpulan', 'data', 'tentang', 'jumlah', 'sampah', 'plastik', 'yang', 'terkumpul', 'di', 'setiap', 'lokasi', 'serta', 'memberikan', 'informasi', 'edukatif', 'tentang', 'pentingnya', 'pengelolaan', 'sampah', 'plastik', 'Evaluasi', 'dan', 'analisis', 'data', 'dilakukan', 'untuk', 'mengukur', 'efektivitas', 'dan', 'kepuasan', 'pengguna', 'terhadap', 'aplikasi', 'Penelitian', 'ini', 'memberikan', 'kontribusi', 'dalam', 'pengembangan', 'teknologi', 'yang', 'dapat', 'memfasilitasi', 'pemantauan', 'dan', 'pengolaan', 'sampah', 'plastik', 'kemasan', 'Diharapkan', 'bahwa', 'aplikasi', 'ini', 'dapat', 'digunakan', 'oleh', 'masyarakat', 'dan', 'pihak', 'terkait', 'dalam', 'upaya', 'pengurangan', 'sampah', 'plastik', 'dan', 'pelestarian', 'lingkungan', 'Dimasa', 'depan', 'aplikasi', 'ini', 'memiliki', 'potensi', 'untuk', 'diperluas', 'dan', 'ditingkatkan', 'sesuai', 'dengan', 'perkembangan', 'teknologi', 'dan', 'kebutuhan', 'pengguna',

b. Filtering

At this stage, a filtering process will be performed to remove irrelevant words, such as common conjunctions, prepositions, and other non-informative terms, which do not contribute to the meaningful analysis of the text. This step is crucial for improving the quality of the data, as these words may otherwise skew the results and decrease the accuracy of subsequent text processing tasks. By eliminating such words, the focus shifts to more relevant and informative terms that provide better insight into the content. The filtering process is typically based on predefined stopword lists or customized criteria tailored to the specific context of the study. A summary of the words identified for filtering in this stage is presented in Table 3, offering a clear view of the terms removed during this process..

Table 3. Filtering

Corpus Document
yang, dan, untuk, ini, sebuah, oleh, karena, itu, pada, dengan, dalam, tersebut, seperti, dari, sebagai, melalui, tentang, serta, bahwa, oleh, lebih, dapat, akan, setiap, di, ini, dilakukan, sebuah, sudah

c. Stemming

Stemming is the process of reducing words to their base or root form, which allows for the simplification of variations in word forms that convey the same underlying meaning. This technique is commonly used in text mining and natural language processing to standardize words, enabling more accurate analysis by treating different forms of a word as equivalent. For example, words such as "running", "runner", and "ran" would all be reduced to their root form "run". By applying stemming, the text is transformed into a more consistent structure, which improves the efficiency of subsequent tasks such as text classification, information retrieval, and keyword extraction. The results of the stemming process for the dataset used in this study are further presented in Table 4, highlighting the root words derived from the original terms.

Table 4. Stemming

Corpus Document
guna plastik kemas makin luas sebab tingkat jumlah sampah plastik buang cara tidak atur ancam lingkungan. teliti tuju kembang aplikasi android guna teknologi quick respons code monitor buang sampah plastik kemas. metode teliti guna liput studi literatur kelola sampah plastik, analisis butuh guna survei wawancara, rancang kembang aplikasi android, uji aplikasi uji fungsi uji guna. aplikasi mungkin guna pindai quick respons code pasang produk kemas tempat buang sampah. pindai quick respons code guna catat lokasi jenis sampah plastik buang akurat. aplikasi juga sedia fitur kumpul data jumlah sampah plastik kumpul lokasi, beri info edukatif penting kelola sampah plastik. evaluasi analisis data ukur efektif puas guna atas aplikasi. teliti beri kontribusi kembang teknologi fasilitasi pantau olah sampah plastik kemas. harap aplikasi guna masyarakat pihak kait upaya kurang sampah plastik lesta lingkungan. masa depan aplikasi milik potensi luas tingkat sesuai kembang teknologi butuh guna

3.4 Indexing and Weighting

The TF-IDF (Term Frequency-Inverse Document Frequency) technique is a widely used method in text mining and natural language processing for evaluating the significance of a word within a specific document in relation to its frequency across a larger set of documents, also known as a corpus. The technique helps to identify words that are important in a particular document, but not necessarily common across all documents, making it valuable for tasks like information retrieval, document classification, and keyword extraction. The TF-IDF calculation works by multiplying

two components: term frequency (TF), which measures how often a word appears in a document, and inverse document frequency (IDF), which assesses how rare or unique a word is across the entire corpus. By weighing words based on these factors, the technique highlights the most informative terms in a document. A summary of the TF-IDF calculation results for the dataset used in this study is presented in Table 5, providing an overview of the most significant words identified through this method..

Table 5. TF-IDF

Corpus Document	Invers Document Frequency
(0.300)	0.0539
(0.420)	0.0317
(0.352)	0.0540
(0.621)	0.0540
(0.112)	0.0390
(0.389)	0.0561
(0.301)	0.1231
(0.266)	0.0540
(0.39)	0.0531
(0.304)	0.0531

3.5 Modelling

The model implemented in this stage is the **Rabin-Karp algorithm**, which is used to efficiently find a specific substring or pattern within a larger text. The algorithm works by hashing both the pattern and substrings of the text into numerical values, allowing for rapid comparison between them. If the hash values match, a more detailed comparison is performed to verify the exact match. The Rabin-Karp algorithm is particularly effective when searching for multiple patterns simultaneously, as it enables the checking of several potential matches in parallel using a single hash function, thus improving efficiency.

This algorithm is designed to handle large datasets effectively by minimizing the number of direct comparisons needed. By quickly eliminating substrings that do not match the pattern through hash value comparisons, it speeds up the overall search process. While the Rabin-Karp algorithm performs well in many text processing tasks, such as pattern matching and text searching, its effectiveness can depend on factors like the choice of hash function and the size of the dataset. In this study, the algorithm is applied to search for specific patterns within the corpus, facilitating efficient data retrieval and analysis. The results of the pattern matching process are discussed in the following sections..

Table 6. Result in Rabin-Karp

Corpus	**Result**
Corpus 1	0.2453
Corpus 2	0.1134
Corpus 3	0.2231
Corpus 4	0.2119
Corpus 5	0.3423
Corpus 6	0.1132

Corpus 7	0.2132
Corpus 8	0.2452
Corpus 9	0.1324
Corpus 10	0.3341

From Table 6, it can be observed that the values vary from 0.1132 (the lowest) to 0.3423 (the highest), indicating how well the searched pattern aligns with the text in each corpus. Corpus 5 (0.3423) and Corpus 10 (0.3341) have the highest results, suggesting that the searched pattern has a significant match with the content of these two corpora. This could imply that the discovered pattern appears more frequently or is more relevant in these particular corpora.

On the other hand, Corpus 6 (0.1132) and Corpus 2 (0.1134) exhibit low match values, suggesting that the searched pattern is either less relevant or infrequently appears within the texts of these corpora. These low values indicate that the patterns sought after may not align closely with the content of these particular corpora, possibly due to the rarity of the terms or their limited presence in the text. As a result, the algorithm struggles to identify meaningful matches in these corpora, highlighting the need for further refinement of the search parameters or the exploration of alternative patterns that may be more representative of the content.

3.6 Evaluation

The evaluation of Rabin-Karp from a precision perspective indicates that the patterns identified by the algorithm are indeed relevant to the corpus being analyzed. In this case, precision can be calculated by examining the high result values (for example, values above a certain threshold, such as 0.3) and assessing how many of those results are truly relevant. Corpus 5 and Corpus 10 have high matches (over 0.3), suggesting that the algorithm may be quite accurate in detecting the correct patterns in these two corpora. Conversely, for corpora with low values (such as Corpus 2 and 6), precision may be lower because the algorithm fails to identify significant patterns.

From the data in Table 6, it can be observed that although some corpora show high results (Corpus 5 and 10), other corpora such as Corpus 6 and 2 have low results. This may indicate that the algorithm does not capture all the patterns that may exist in those corpora, demonstrating low recall in some cases.

Here's the corrected translation with improved grammar:

The overall accuracy of the algorithm can be calculated as the ratio of the number of patterns that truly match to the total number of patterns tested. By examining the data in Table 6, the average result falls within the range of 0.2, which can be used as a reference for determining the threshold. The corpora with values above 0.2 are Corpus 1, 3, 4, 5, 7, 8, and 10, while the corpora with results below 0.2 are Corpus 2, 6, and 9. From this comparison, 7 out of 10 corpora have accurate matches, and the accuracy can be calculated using the following formula :

$$Accuration = \frac{7}{10} = 0{,}7\ or\ 70\%$$

Based on the results, it can be interpreted that the Rabin-Karp algorithm achieves an accuracy level of approximately 70% in identifying matching patterns within the tested corpus. This indicates that the algorithm is relatively effective at detecting patterns, though there is room for improvement, particularly in more complex or varied datasets. The 70% accuracy suggests that while the algorithm is successful in many instances, it may still miss some patterns or require optimization to handle more intricate cases. Further refinements could potentially enhance its performance, ensuring more precise and reliable pattern detection in future applications.

4. CONCLUSION

The values in the Corpus Document range from 0.112 to 0.621, reflecting the varying frequencies of terms across the documents. Terms with higher values are those that appear more frequently within the documents, whereas those with lower values are used less often. This variation indicates the level of importance and relevance of different terms in the corpus, with higher values typically representing more significant words within specific documents. On the other hand, the **Inverse Document Frequency (IDF)** values range from 0.0317 to 0.1231, highlighting the rarity of certain terms across the entire corpus. Higher IDF values suggest that the corresponding terms are less common in the broader corpus, making them more distinctive and potentially more valuable in identifying specific patterns or topics.

The overall accuracy of the pattern matching process is around 70%, which suggests that the algorithm is performing reasonably well in identifying relevant patterns within the text. However, the recall value shows lower results for corpora with lower values, indicating that the model may struggle to detect all relevant patterns in these cases. This could be due to the fact that the searched patterns may not be fully captured, especially when they are rare or less prominent in the dataset. The lower recall for corpora with low values points to potential challenges in achieving comprehensive detection, which may require further refinement of the model or the inclusion of additional data for better pattern recognition..

ACKNOWLEDGMENTS

We would like to express our gratitude to Sam Ratulangi University, especially the Research and Community Service Institute (LPPM), for their support in conducting this research. Our thanks also go to the journal's editors and reviewers for their constructive feedback, as well as to all parties who have contributed to the completion of this study. We hope that the results will benefit the advancement of science and technology.

REFERENCE

[1] T. e. a. Nurhaeni, "Sistem Penilaian Sidang Komprehensif Tugas Akhir Skripsi dan Tesis Berbasis Yii Framework Menggunakan Business Intelligence Methodology," *Technomedia J.*, vol. 5, no. 1, pp. 82–94, 2021.

[2] Carlos Lage-Gomez, "On the interrelationships between diverse creativities in primary education STEAM projects," *Think. Ski. Creat.*, vol. 51, 2024.

[3] Syaharuddin, "ENELUSURAN REFERENSI BERBASIS DIGITAL SEBAGAI PENINGKATAN SOFT SKILL MAHASISWA DALAM MENYELESAIKAN TUGAS AKHIR," *J. Pengabdi. Masy. Berkemajuan 3.2*, vol. 3, no. 2, pp. 151–155, 2021.

[4] M. Zaeni, "Urgensi penelitian pengembangan dalam menggali keterampilan berpikir kritis," 2021.

[5] R. N. Sari, "Data Mining Peminatan Mata Kuliah Pilihan Mahasiswa Tingkat Akhir Jurusan Informatika Menerapkan Algoritma C4. 5," *Bull. Comput. Sci. Res. 3.3*, vol. 3, no. 3, pp. 263–269, 2023.

[6] A. Sanders, "The Implementation of Data Mining to Get The Pattern for Selecting Students' Thesis Title," *J. Komputer, Inf. dan Teknol.*, vol. 1, no. 2, pp. 165–173, 2021.

[7] Kaile Chen, "Process mining and data mining applications in the domain of chronic," *Artif. Intell. Med.*, vol. 35, no. 10, 2023.

[8] et al Olson, David L., *Descriptive data mining*. Singapore: Springer, 2019.

[9] Meidelfi, "TF-IDF Implementation for Similarity Checker on The Final Project Title," *Int. J. Adv. Sci. Comput. Eng.*, vol. 3, no. 1, pp. 40–52, 2021.

[10] M. Naf'an, "Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen," *J. Linguist. Komputasional*, vol. 2, no. 1, pp. 23–27, 2019.

[11] Z. Zhu, "Hot topic detection based on a refined TF-IDF algorithm," *IEEE access*, vol. 7, pp. 26996–27007, 2019.

[12] I Billhaqqi, "Comparison analysis of Rabin-Karp and Winnowing algorithms in automated essay answer assessment system," 2022.

[13] A. D. Hartanto, "Best parameter selection of rabin-Karp algorithm in detecting document similarity," 2019.

[14] M. A. Yulianto, "The hybrid of jaro-winkler and rabin-karp algorithm in detecting Indonesian text similarity," *J. Online Inform.*, vol. 6, no. 1, pp. 88–95, 2021.