VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7663

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

UTILIZING TEXT MINING FOR ENCRYPTION ALGORITHM RECOMMENDATION USING CONTENT-BASED FILTERING

Mukti Qamal[1*]; M. Iqbal[1]; Yesy Afrillia[1]

Informatics Engineering Program[1]
Malikussaleh University Lhokseumawe, Aceh, Indonesia[1]
https://unimal.ac.id[1]
mukti.qamal@unimal.ac.id*, iqbal.210170038@mhs.unimal.ac.id, yesyafrillia@unimal.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *The selection of an appropriate encryption algorithm is crucial in ensuring data security, as each algorithm has distinct advantages and disadvantages in terms of speed, efficiency, and security level. Many users struggle to determine the most suitable algorithm due to limited technical knowledge and the vast amount of literature that must be reviewed. Therefore, this study proposes a recommendation system based on Content-Based Filtering (CBF) integrated with text mining to facilitate faster, more accurate, and data-driven algorithm selection. The objective of this research is to develop a recommendation system capable of analyzing the technical characteristics of encryption algorithms from scientific literature and providing relevant suggestions according to user needs. The methodology includes collecting 300 articles from the Garuda Kemdikbud portal using web scraping, performing data preprocessing such as tokenization, stop word removal, and case folding, representing text with TF-IDF, and calculating similarity using Cosine Similarity. The results indicate that the most frequently discussed algorithms are RSA (52 articles), AES (40 articles), and RC4 (25 articles), reflecting research trends focusing on modern public-key and symmetric cryptography. The evaluation results show that the system achieved Precision@3 of 1.0000 and Average Precision (AP) of 0.0583, indicating that the top recommendations generated are highly relevant to user needs. The developed system successfully generated recommendations tailored to specific needs, such as suggesting AES as the primary choice for "fast encryption of sensitive data." This study demonstrates that combining text mining and CBF is effective in assisting the selection of encryption algorithms through literature-based analysis.*

Keywords: *Content-Based Filtering, Cosine Similarity, Data Security, Encryption Algorithm, Text Mining.*

Intisari—*Pemilihan algoritma enkripsi yang tepat merupakan tantangan penting dalam menjaga keamanan data, mengingat setiap algoritma memiliki kelebihan dan kekurangan yang berbeda dari segi kecepatan, efisiensi, maupun tingkat keamanan. Banyak pengguna, baik individu maupun organisasi, kesulitan menentukan algoritma terbaik karena keterbatasan pengetahuan teknis serta banyaknya literatur yang harus dianalisis. Oleh karena itu, penelitian ini mengusulkan sistem rekomendasi berbasis Content-Based Filtering (CBF) yang terintegrasi dengan text mining untuk membantu proses pemilihan algoritma enkripsi secara lebih cepat, tepat, dan berbasis data. Tujuan penelitian ini adalah mengembangkan sistem rekomendasi yang mampu menganalisis karakteristik teknis algoritma enkripsi dari literatur ilmiah sehingga dapat memberikan saran yang relevan sesuai kebutuhan pengguna. Metode yang digunakan meliputi pengumpulan 300 artikel dari portal Garuda Kemdikbud melalui web scraping, preprocessing data menggunakan tokenisasi, stopword removal, dan case folding, representasi teks dengan TF-IDF, serta perhitungan kemiripan menggunakan Cosine Similarity. Hasil penelitian menunjukkan distribusi algoritma enkripsi yang paling banyak dibahas adalah RSA (52 artikel), AES (40 artikel), dan RC4 (25 artikel), yang mencerminkan fokus penelitian pada algoritma modern berbasis kunci publik dan simetris. Hasil evaluasi menunjukkan bahwa sistem mencapai Precision@3 sebesar 1,0000 dan Average Precision (AP) sebesar 0,0583, yang mengindikasikan bahwa rekomendasi teratas yang dihasilkan sangat relevan dengan kebutuhan pengguna. Sistem rekomendasi yang dibangun berhasil memberikan rekomendasi sesuai kebutuhan, misalnya untuk*

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7663

"algoritma cepat untuk data sensitif" direkomendasikan AES sebagai algoritma utama. Penelitian ini membuktikan bahwa kombinasi text mining dan CBF efektif dalam membantu pemilihan algoritma enkripsi berbasis analisis literatur.

***Kata Kunci**: Algoritma Enkripsi, Data Security, Content-Based Filtering, Cosine Similarity, Text Mining.*

INTRODUCTION

Choosing the right encryption algorithm is crucial in maintaining data efficiency and security. Each algorithm has different characteristics in terms of security, speed, and efficiency, so choosing the wrong one can pose serious risks. The main challenge arises when users do not have a deep enough technical understanding of the advantages and disadvantages of each algorithm, so that decisions tend to be based on popularity or general recommendations rather than specific needs. This condition emphasizes the importance of a data-driven approach that can help the algorithm selection process to be more accurate and measurable.

One potential solution is the implementation of a content-based filtering recommendation system combined with text mining. A recommendation system is a technique or method designed to provide suggestions or recommendations related to a product, service, or content to users based on their preferences and behavior [1]. Recommendation systems are developed to provide personalized information and learning experiences by analyzing user interaction patterns and preferences [2].

The most common approaches to implementing recommendation systems are Content-based Filtering, Collaborative Filtering, and Hybrid Filtering [3]. Content-based filtering is a recommendation system method based on the attributes, characteristics, or features of an item [4]. This method allows for the analysis of scientific literature to extract information related to encryption algorithms, then map their technical characteristics to compare them with user needs. With this approach, the recommendation system does not only rely on the opinions or preferences of other users, but also generates recommendations based on objective parameters. This advantage makes the content-based filtering method superior to collaborative filtering, especially for specific domains such as encryption, where the needs of each user vary greatly [5].

In its implementation, this recommendation system utilizes the TF-IDF (Term Frequency–Inverse Document Frequency) method to measure how important a word is in a document relative to a collection of other documents [6]. The general TF-IDF formula is written as follows.

$$W_{i,j} = TF_{\mathrm{i,j}}\ x\ IDF_{\mathrm{j}} \quad (1)$$

$$W_{i,j} = TF_{\mathrm{i,j}}\ x\ \left(in\left(\frac{D}{df_j}\right)+1\right) \quad (2)$$

Where is the TF-IDF weight (word i in document j), is the term frequency (word i in document i), D is the number of documents, and is the number of times the term (word j) appears in the documents. The vector representation results of TF-IDF are then compared using Cosine Similarity, which is a method that calculates the similarity between documents based on the cosine angle value between the vectors [7]. The Cosine Similarity formula is

$$\cos\alpha = \frac{A.B}{|A|B|} = \frac{\sum_{i=1}^{n} A_i\ x\ B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2\ x\ \sum_{i=1}^{n}(B_i)^2}} \quad (3)$$

Where A and B are the representation vectors of the two documents being compared. The calculated value ranges from 0 to 1, where a value close to 1 indicates a high degree of similarity. The combination of these two methods allows the system to find articles or algorithms with characteristics that are most relevant to user needs in a mathematical and objective manner.

Furthermore, text mining plays an important role in this process because it is capable of analyzing large amounts of text, identifying patterns, and finding the latest research trends [8], [9]. A number of previous studies have also confirmed that the integration of text mining with recommendation systems can improve the relevance, accuracy, and speed of data analysis [10], [11]. In addition, recommendation personalization can be done by matching the unique needs of users with the features of the encryption algorithm, so that the results provided are more targeted [12]. This is very useful in the context of increasing demand for data security in the digital era, where cyber-attacks and data leaks are increasingly frequent.

Different from previous studies, which generally focus on performance comparison or manual evaluation of encryption algorithms, this study introduces a literature-driven recommendation approach by integrating text mining on a large-scale national scientific corpus (300 articles from the Garuda Kemdikbud portal) with a Content-Based Filtering method. This approach enables the extraction of technical characteristics of encryption algorithms directly from scientific publications and transforms them into personalized recommendations based on user needs. To the best of our knowledge, such an integration has not been widely explored in the context of encryption algorithm selection.

Based on this background, this study was conducted to develop a recommendation system for selecting encryption algorithms based on text mining and content-based filtering. This research is expected to contribute practical solutions for individual users and organizations that need fast, accurate, and valid data-based encryption algorithm recommendations. In addition, this research also plays a role in strengthening the literature on the use of text mining in recommendation systems, while promoting awareness of the importance of choosing the appropriate encryption algorithm to improve information security in various sectors.

MATERIALS AND METHODS

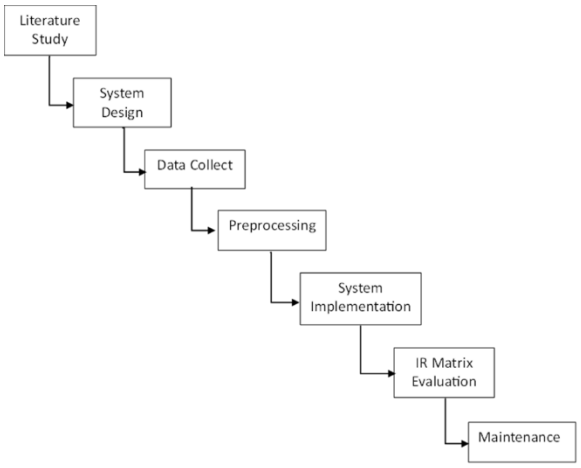
Data Sources

The data for this study was sourced from secondary data obtained from journals, scientific articles, and other reliable sources discussing various encryption algorithms. This data was used to analyze the characteristics, advantages, and disadvantages of each encryption algorithm, thereby determining the most optimal algorithm for protecting sensitive data. The selection of data sources was based on the accuracy, relevance, and credibility of the available information, thereby providing a solid basis for comparing encryption algorithms and ensuring that the recommendations produced have a high level of validity.

Data Collection Techniques

The data collection technique in this study was conducted through web scraping, which is an automated method of retrieving data from various websites that provide information related to encryption algorithms [13] or it could also be a technique for obtaining information from a particular website so that data can be retrieved either manually or automatically [14]. Scraping was carried out on reliable sources that published research, technical documentation, and comparisons of encryption algorithms based on security, speed, and efficiency. The data obtained through scraping will be further processed using TF-IDF and Cosine Similarity to analyze the similarities in the characteristics of encryption algorithms and determine the most optimal algorithm for protecting sensitive data.

Research Work Scheme



Source: (Research Results, 2025)

Figure 1. Research Work Scheme

This research stage shows the research process that will be carried out and describes the research as a whole. The stages of this research are:

a) Literature Review: This stage focuses on literature studies covering theories, methods, and previous research related to data encryption and Content-Based Filtering-based recommendation systems. The researcher also reviews various encryption algorithms, TF-IDF calculation techniques, Cosine Similarity, and previously implemented recommendation system implementations.
b) System Design: At this stage, the architecture of the recommendation system was designed, including how user input would be processed to generate encryption algorithm recommendations. The system was designed to be able to accept criteria inputted by users and process them using the TF-IDF, Cosine Similarity, and Content-Based Filtering methods.
c) Data Collection: At this stage, data is collected through web scraping techniques from various journals and research articles discussing encryption algorithms. The Newspaper library in Python is used to retrieve text from relevant sources, which will then be analyzed to determine the most optimal encryption

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

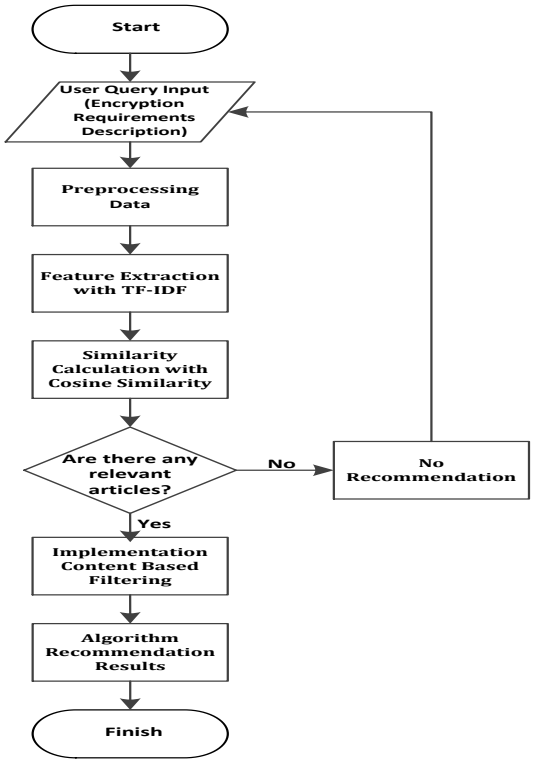
VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7663

algorithm [15]. The data collected includes the characteristics of encryption algorithms, their advantages, disadvantages, and performance in various scenarios. The dataset is explicitly described as 300 national scientific articles collected from the Garuda Kemdikbud portal using web scraping techniques with the keyword "security algorithm". Each article not only includes basic metadata such as title, author, journal, and abstract, but also is supplemented with manual annotations in the form of encryption algorithm type, data type, advantages, and disadvantages of the algorithm. The structure of this dataset ensures that each document has a rich and relevant textual representation for the TF-IDF-based feature extraction process.

d) Data Preprocessing: Before being used in the recommendation system, the collected data must be cleaned and further processed. The stages of preprocessing include: Case Folding, which converts all text to lowercase for uniformity [16]. Tokenizing, which separates text into smaller words or tokens [17]. Filtering, which removes words that have no significant meaning, such as conjunctions or stopwords. Stemming, which converts words to their root form to standardize variations in the text [18].
e) System Implementation: The designed system is implemented using the TF-IDF weighting method and Cosine Similarity to determine the similarity between words [19]. Then, Content Base Filtering will be applied to obtain recommendations from the analysis results.
f) Evaluation: The evaluation aims to assess how well the system works by measuring the precision value produced. The system was evaluated using an Information Retrieval approach with Average Precision (AP) and Precision@k metrics, which are suitable for ranking-based recommendation systems. The evaluation was conducted using real query scenarios, such as "I want to secure bank data," with a ground truth of 103 relevant documents that had been identified in the dataset. The Precision@3 value of 1.0000 indicates that all top recommendations provided by the system are relevant to user needs, while the Average Precision of 0.0583 provides an overview of the system's overall performance in ranking relevant documents. By explaining in detail the dataset, the curation process, and the evaluation scheme and metrics used, this study provides adequate transparency on the research steps and allows for validation and further development in subsequent studies.

System Work Scheme

The following figure is a flowchart of how the recommendation system will work.



Source: (Research Results, 2025)

Figure 2. System Workflow Diagram

The following is an explanation of the stages in the workflow diagram for the encryption algorithm selection recommendation system:

a) Start: The system is activated and ready to receive input from users. At this stage, the system is still waiting for queries from users to be processed further.
b) User Query Input: The user enters a description of the encryption requirements they want to use.
c) Data Preprocessing: The system performs text processing on the user query and available research journals.
d) TF-IDF: The system applies the TF-IDF (Term Frequency-Inverse Document Frequency) method to extract features from the processed text [20].
e) Cosine Similarity: This technique compares the text vector of the query with the text vector of the journal to determine the level of similarity between the two [21].

f) Are there any relevant articles?: The system evaluates the Cosine Similarity calculation results to determine whether there are articles that are relevant enough to the user's query.
g) Content Based Filtering: The system applies the Content-Based Filtering technique to filter and select articles that best suit the user's needs based on content similarity.
h) Recommendation Results: The system displays recommendation results in the form of appropriate encryption algorithms based on the analyzed research articles.
i) Finish: After the recommendation results or message are displayed, the system completes the process.

RESULTS AND DISCUSSION

Data Collection

The data used in this study was obtained from the Garuda Kemdikbud portal, one of the official databases containing national scientific publications. The data collection process was carried out using a web scraping technique designed through the Python programming language, utilizing the Requests and Beautiful Soup libraries. The search keyword used was “security algorithm” so that the articles that appeared were relevant to the research topic. Each article that was successfully retrieved had its metadata extracted, such as the title, author, journal name, publisher, abstract, and document link. This data was then stored in JSON format to facilitate the next preprocessing stage. This approach was taken to ensure that the resulting dataset truly met the needs of the recommendation system. This stage also served as data curation, as not all articles that were scraped were relevant to the focus of the research. Articles that did not directly discuss data security algorithms were eliminated, while relevant articles were retained to form the final dataset. This dataset consisted of several core columns, such as title, author, journal, abstract, publisher, and additional columns resulting from manual annotation in the form of algorithms, data types, advantages, and disadvantages. Overall, the amount of data successfully extracted was 300 articles. Thus, the collected data is ready to be used in the preprocessing, feature engineering, and development stages of a Content-Based Filtering (CBF) recommendation system.

Data Preprocessing

In this study, preprocessing was carried out as an important stage to clean and prepare the data before further analysis. This process includes four main stages, namely case folding, which standardizes word representation by converting all letters to lowercase for consistency; stemming, which removes irrelevant punctuation marks to clean up the text; tokenization, which breaks sentences into words or tokens for easy analysis; and stop word removal, which removes common words that have no significant meaning in the context of data security. These four stages are carried out sequentially so that the scraped data becomes more structured, relevant, and ready to be used in the analysis process and the development of an encryption algorithm recommendation system. After the preprocessing process is complete, the next step is to combine features from several important columns in the dataset. This combination aims to form a more comprehensive text representation, so that each article is not only represented by one aspect, but also includes core information from various attributes. In this study, the features used are title, abstract, algorithm, and data type. The title provides an overview of the article's content, the abstract presents a more in-depth summary, the algorithm reflects the method used, while the data type confirms the context of the algorithm's application.

TF-IDF Weighting

Term Frequency (TF) is the number of times each word appears in a document, while Inverse Document Frequency (IDF) adjusts the weight of words that appear frequently, increasing the importance of less common words in the document [22]. TF-IDF combines the concepts of TF (Term Frequency) and IDF (Inverse Document Frequency) to give a more balanced weight to words in a document. Words that appear frequently in one document but rarely appear in other documents will have a higher weight. To build a TF-IDF vector, the vectorizer object created from the TfidfVectorizer class is used to process text data and convert it into a TF-IDF-based numerical representation. The fit_transform() function calculates the TF-IDF matrix from the processed_text column in the df dataset, where df[‘processed_text’] contains the preprocessed text (title, abstract, and article algorithm). This process calculates TF (word frequency in a document) and IDF (word rarity across all documents), then multiplies them according to the formula (1).

$W_{i,j} = TF_{i,j}\ x\ IDF_j$ 13 x 1.218173 = 0.315661

Table 1. Sampel TF-IDF

TERM	TF – IDF
data	0.315661
bit	0.298348

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7663

TERM	TF – IDF
vbnet	0.212117
process	0.198193
2008	0.195953
integrity	0.189808
128	0.162175
key	0.159991
symmetric	0.150001
needs	0.147956
receiver	0.133144
interested	0.133144
sent	0.125647
security	0.119995

Source: (Research Results, 2025)

Cosine Similarity

In this study, Cosine Similarity is implemented after the text goes through preprocessing stages (text cleaning, tokenization, stopword removal, and stemming) and TF-IDF transformation. TF-IDF representation produces numerical vectors for each document, facilitating similarity calculations. Thus, when a query (e.g., a request for an algorithm suitable for banking data security) is submitted, the system calculates the Cosine Similarity value between the query and each document in the dataset. The document with the highest similarity value is then recommended to the user.

In its implementation, the cosine similarity value ranges from 0 to 1. If the value is close to 1, it means that the query has a high level of similarity with the document, while if the value is close to 0, it means that the query does not have significant similarities. Furthermore, the cosine similarity calculation is done by calculating the dot product between the query vector and the vector of document A and document B, then dividing it by the product of the lengths (norm) of both vectors. The result obtained is a similarity value between the query and document A and between the query and document B, where a higher value indicates that the document is more relevant to the query. For more details, please refer to Table 4 below.

Table 2. Cosine Similarity

TERM	T (A)	T(B)
algoritthm	3	3
needs	2	0
sent	2	2
implementation	2	2
encryption	1	1
application	2	2

Source: (Research Results, 2025)

The vector A and vector B in the table above represent each term from document A and document B to see how many values are obtained from each unique word that appears in the document. Based on Table 4, the vector representation can be written as follows:

1) Vector A = (3, 2, 2, 2, 1, 2, 1)
2) Vector B = (3, 0, 2, 2, 1, 2, 1)

Suppose that the similarity value between term A and term B in the table above is calculated using equation (2). Based on the cosine similarity calculation, the similarity value obtained between terms A and B is 0.92. After successfully performing calculations using cosine similarity, the result obtained for the degree of similarity between term A and term B in the vector space was 0.92. This result shows that document A and document B have a very high degree of similarity because the value is close to 1.

Evaluation

After going through the preprocessing stage, weighting with TF-IDF, and measuring similarity using cosine similarity, the system needs to be tested to measure the accuracy and relevance of the results provided. The evaluation is carried out so that the system not only produces output, but its quality can also be assessed based on certain measures. This evaluation aims to assess the quality of the system using Information Retrieval (IR) metrics such as Average Precision (AP) and Precision@k. The evaluation results for the query "I want to secure bank data" can be seen in Table 5 below.

Table 3. Evaluation Metrics

Avarage Precision (AP)	Precision@3
0.0583	1.0000

Source: (Research Results, 2025)

Table 5 above shows the results of evaluating the algorithm recommendation system based on the query "I want to secure bank data." In these results, the number of relevant articles recorded in the ground truth for this query is 103 articles. This means that in the database, there are 103 documents that are considered truly relevant to user needs. The evaluation was conducted using Information Retrieval (IR) metrics, namely Average Precision (AP) and Precision@3. The Average Precision (AP) value obtained was 0.0583, which indicates the average accuracy of the system in placing relevant documents in the search results.

This relatively low AP value indicates that although the system successfully found relevant documents, there are still many other results that are not yet optimal in their ranking. Meanwhile, the Precision@3 value obtained is 1.0000. This means that of the top 3 recommendations provided by the system, all are relevant to the user's query. This

shows that the system is quite reliable in providing initial recommendations (top-N recommendation), although its overall performance (AP) can still be improved. Thus, these evaluation results illustrate that the system is capable of providing accurate recommendations at the top of the rankings, but further optimization is still needed so that the overall results can be more accurate and comprehensive for all relevant articles in the dataset.

The results show that RSA, AES, and RC4 are the most frequently discussed encryption algorithms in national scientific literature. This finding not only reflects current research trends but also indicates a strong preference among researchers and practitioners for well-established and widely adopted cryptographic algorithms. The dominance of these algorithms suggests that practical reliability and proven security remain key considerations in real-world data protection, which strengthens the relevance of using literature-driven analysis as a knowledge source for recommendation systems.

The system's Precision@3 value of 1.0000 indicates that all top-ranked recommendations are highly relevant to user needs. This result implies that the proposed system is particularly effective for early-stage decision support, especially for users with limited cryptographic expertise. In practical terms, the system can significantly reduce the time and effort required to identify suitable encryption algorithms, serving as an efficient preliminary selection tool before more detailed technical evaluations are conducted.

However, the relatively low Average Precision (AP) value of 0.0583 suggests that while the top recommendations are accurate, the overall ranking of relevant documents can still be improved. This implies that the TF-IDF and Cosine Similarity approach is effective in capturing surface-level relevance based on keyword similarity, but may not fully represent deeper semantic relationships between documents. Consequently, this finding highlights opportunities for future enhancement through the integration of semantic-based models, such as word embeddings or deep learning–based representations, to improve ranking performance across the entire recommendation list.

Furthermore, the successful recommendation of AES for the query "fast encryption for sensitive data" demonstrates the system's ability to extract implicit knowledge from scientific literature and translate it into context-aware recommendations. This indicates that scientific publications can function not only as passive reference materials, but also as active and dynamic knowledge sources for decision-support systems in the field of information security.

Overall, these findings suggest that the integration of text mining and Content-Based Filtering provides a promising and scalable approach for encryption algorithm selection based on empirical evidence from scientific literature. The approach is beneficial not only for researchers but also for practitioners and organizations seeking objective, data-driven guidance in selecting encryption algorithms tailored to their specific security requirements.

System Implementation and Testing

The developed algorithm recommendation system was then implemented in the form of a web-based Graphical User Interface (GUI). This implementation aims to enable users to easily enter queries and immediately obtain relevant algorithm recommendations interactively. The data used in the system comes from JSON files that have been processed previously, which have gone through preprocessing and representation stages using the TF-IDF method. Thus, the code applied to the website only focuses on implementing the TF-IDF and Cosine Similarity calculations that have been designed, so that the system can run practically without the need for reprocessing from scratch.

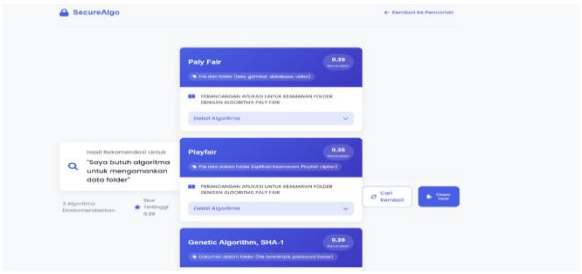
Source: (Research Results, 2025)

Figure 5. Input Page

On the home page, users will see a simple display with the tagline "Find the Perfect Security Algorithm for Your Needs." This page features a search field where users can enter specific requirements, such as "I need an algorithm to secure folders." After entering these requirements, the system will process the query with preprocessing, TF-IDF calculation, and calculate the suitability between the query and the dataset using Cosine Similarity. The "Get Recommendations" button triggers the search process, while users can also press the Enter key to speed up the interaction.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7663



Source: (Research Results, 2025)

Figure 6. Recommendation Results Page

On the recommendation results page, the system displays a list of algorithms relevant to the user's query. Each algorithm is displayed in the form of a recommendation card containing the algorithm name, similarity score, and a brief description of the algorithm's use. Brief description of the algorithm's use. The advantages and disadvantages of the recommended algorithms, so that users can directly compare them and adjust them to their specific needs. For example, for the query "I need an algorithm to secure folder data," the system recommends algorithms such as Play Fair, and Genetic Algorithm with SHA-1, with matching scores of 0.39, 0.38, and 0.36, respectively.

CONCLUSION

This study successfully collected 300 articles discussing data security algorithms, then performed preprocessing and distribution analysis, which showed that the RSA (52 articles), AES (40 articles), and RC4 (25 articles) algorithms were the most dominant, providing an overview that public key and symmetric cryptography algorithms are the main focus in the literature. The recommendation system built using the Content-Based Filtering (CBF) approach with TF-IDF representation and Cosine Similarity calculation proved capable of providing algorithm recommendations as needed. For example, in the case of "fast encryption algorithms for sensitive data," the system recommended AES as the main algorithm with the highest score. These results indicate that the proposed method can assist researchers or practitioners in selecting relevant data security algorithms. However, this study is limited by the use of a single national literature source and a purely content-based approach, which does not yet consider user preferences or real-world performance metrics of encryption algorithms. Future research will explicitly address these limitations by expanding the dataset with international publications from databases such as Scopus and Google Scholar, conducting comparative experiments using alternative text representation methods including Word2Vec, FastText, and transformer-based models (e.g., BERT), and evaluating system performance using additional metrics such as Recall@k, F1-score, and NDCG. Furthermore, hybrid recommendation approaches that combine content-based and user-based information will be explored to improve recommendation accuracy, ranking quality, and adaptability to diverse user needs.

REFERENCE

[1] M. Qamal, F. Fajriana, and M. Mardhatillah, "Metode Naive Bayes untuk Menentukan Rekomendsi Tempat Wisata Terbaik di Aceh," *TECHSI - J. Tek. Inform.*, vol. 13, no. 1, pp. 81–91, Apr. 2021, doi: https://doi.org/10.29103/techsi.v13i1.3132.

[2] X. Wei and S. Sun, "Personalized online learning resource recommendation based on artificial intelligence and educational psychology," *Frontiers in Psychology*, vol. 12, p. 767837, 2021, doi: 10.3389/fpsyg.2021.767837

[3] S. Sharma, V. Rana, and M. Malhotra, "Automatic recommendation system based on hybrid filtering algorithm," *Educ. Inf. Technol.*, vol. 27, no. 2, pp. 1523–1538, 2022, doi: httpS://doi.org/10.1007/s10639-021-10643-8.

[4] M. Zahrawi and A. Mohammad, "Implementing Recommender Systems using Machine Learning and Knowledge Discovery Tools," *Knowledge-Based Engineering and Sciences*, vol. 2, no. 2, pp. 44–53, 2021, doi: 10.51526/kbes.2021.2.2.44-53

[5] U. Javed, K. Shaukat, I. A. Hameed, F. Iqbal, T. M. Alam, and S. Luo, "A Review of Content-Based and Context-Based Recommendation Systems," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 3, pp. 274–306, 2021, doi: 10.3991/ijet.v16i03.18851

[6] R. Madhumala, B. Vineetha, M. R. Shree, R. Sanjesh, and B. Charanraj, "A semantic driven model for extraction of text using TF - IDF , SFLA and XGBoost," *Int. J. Inf. Technol.*, vol. 17, no. 6, pp. 3659–3664, 2025, doi: https://doi.org/10.1007/s41870-025-02549-2.

[7] N.-T.-K. Ho, H. Ho-Dac, and T.-A. Le, "Job Recommendation: An Approach to Match Job-Seeker's Interest with Enterprise's

Requirement," in *Research in Intelligent and Computing in Engineering*, 2021, pp. 361–367. doi: https://doi.org/10.1007/978-981-15-7527-3_35.

[8] Y. Afrilia, L. Rosnita, D. Siska, M. Rigayatsyah, and N. Nurqamarina, "Analisis Sentimen Ciutan Twitter Terkait Penerapan Permendikbudristek Nomor 30 Tahun 2021 Menggunakan TextBlob dan Support Vector Machine," *G-Tech J. Teknol. Terap.*, vol. 6, no. 2, pp. 387–394, 2022, doi: https://doi.org/10.33379/gtech.v6i2.1778.

[9] T. Ridwansyah, B. Subartini, and S. Sylviani, "Penerapan Metode Content-Based Filtering pada Sistem Rekomendasi," *Math. Sci. Appl. J.*, vol. 4, no. 2, pp. 70–77, 2024, doi: https://doi.org/10.22437/msa.v4i2.32136.

[10] M. Ikhsan and R. Kurniawan, "Penerapan Text Mining pada Sistem Rekomendasi Pembimbing Skripsi Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Indones. J. Comput. Sci.*, vol. 12, no. 1, pp. 4196–4207, 2023, doi: https://doi.org/10.33022/ijcs.v12i6.3500.

[11] W. Lemus Leiva, M.-L. Li, and C.-Y. Tsai, "A Two-Phase Deep Learning-Based Recommender System: Enhanced by a Data Quality Inspector," *Applied Sciences*, vol. 11, no. 20, p. 9667, 2021, doi: 10.3390/app11209667

[12] I. M. A. Bhaskara, M. Pasek, M. P. A. Ariawan, I. B. A. Peling, and I. P. A. Prayudha, "Studi Literatur : Analisa Perbandingan Teori Tentang Tingkat Keamanan Antar Algoritma Simetris," *J. Bangkit Indones.*, vol. 13, no. 01, pp. 40–45, 2024, doi: https://doi.org/10.52771/bangkitindonesia.v13i1.278.

[13] M. A. Khder, "Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application," *Int. J. Adv. Soft Comput. its Appl.*, vol. 13, no. 3, pp. 145–68, 2021, doi: https://doi.org/10.15849/IJASCA.211128.11.

[14] V. Krotov and M. Tennyson, "Web Scraping in the R Language: A Tutorial," *Journal of the Midwest Association for Information Systems (JMWAIS)*, vol. 2021, no. 1, Art. no. 5, 2021, doi: 10.17705/3jmwa.000066

[15] A. Weichselbraun, "Inscriptis - A Python-based HTML to text conversion library optimized for knowledge extraction from the Web," *Journal of Open Source Software*, vol. 6, no. 66, p. 3557, 2021, doi: 10.21105/joss.03557

[16] M. U. Albab, Y. Karuniyawati, and M. N. Fawaiq, "Optimization of the Stemming Technique on Text preprocessing President 3 Periods Topic," *J. Transform.*, vol. 20, no. 2, pp. 1–10, 2023, doi: https://doi.org/10.26623/transformatika.v20i2.5374.

[17] H. Yan, Y. Sun, X. Li, and X. Qiu, "An Embarrassingly Easy but Strong Baseline for Nested Named Entity Recognition," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Jul. 2023, pp. 1442–1452. doi: https://doi.org/10.18653/v1/2023.acl-short.123.

[18] Z. Abidin, A. Junaidi, and W. Wamiliana, "Text Stemming and Lemmatization of Regional Languages in Indonesia : A Systematic Literature Review," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 10, no. 2, pp. 217–231, 2024, doi: https://doi.org/10.20473/jisebi.10.2.217-231.

[19] N. Nilawati, H. Husaini, and J. Salat, "Penggunaan Metode Cosine Similarity dan TF-IDF untuk Klasifikasi Judul Seminar Proposal pada Fakultas Teknik Universitas Jabal Ghafur," *Sagita Acad. J.*, vol. 2, no. 1, pp. 72–79, 2024, doi: https://doi.org/10.61579/sagita.v2i1.60 p-ISSN:

[20] S. Jain, S. K. Jain, and S. Vasal, "An Effective TF-IDF Model to Improve the Text Classification Performance," in *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, 2024, pp. 1–4. doi: https://doi.org/10.1109/CSNT60213.2024.10545818.

[21] R. Oktavian and F. Amin, "Perancangan Sistem Rekomendasi Laptop dengan Model Prototyping dan Penerapan Content-Based Filtering Approach pada ELS Computer Shop Semarang," *G-Tech J. Teknol. Terap.*, vol. 8, no. 1, pp. 66–80, 2024, doi: https://doi.org/10.33379/gtech.v8i1.3490.

[22] H. Singhdev, S. Gupta, V. Srivastava, and A. Saxena, "Text recognition using improved dual attention based on textual double embedding network with aquila optimization algorithm," *Int. J. Inf. Technol.*, 2024, doi: https://doi.org/10.1007/s41870-024-01984-x.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri