



Implementation Of Support Vector Machine (SVM) Algorithm With Term Frequency–Inverse Document Frequency (TF-IDF) Weighting For Web-Based Book Classification Using Dewey Decimal Classification (DDC)

Hikma^a, Sutardi^b, Muhammad Ihsan Sarita^c

^{a,b,c}Department of Informatics, Faculty of Engineering, Halu Oleo University, Kendari, Indonesia

Article Info

Keywords:

Dewey Decimal Classification
Book Classification
Support Vector Machine
TF-IDF
Cosine Similarity

ABSTRACT

Book classification is an important activity in library management because it determines the grouping of collections based on subject areas and facilitates information retrieval. Manual classification requires considerable time and accuracy from librarians, particularly as the number of collections increases and book subjects become more diverse. This study aims to implement the Support Vector Machine (SVM) algorithm with Term Frequency–Inverse Document Frequency (TF-IDF) weighting to assist book classification based on the Dewey Decimal Classification (DDC) and to develop a web-based book classification system as a decision-support tool for librarians. The research data consist of book titles, synopses, and DDC labels obtained from the National Library of the Republic of Indonesia (Perpusnas) catalog and other library OPAC catalogs available on the internet. The research stages include text preprocessing, vocabulary construction, TF-IDF weighting, Cosine Similarity measurement as a data-scope filtering process, and classification using SVM. The Cosine Similarity testing on 20 out-of-category data and 85 in-scope data showed that a threshold of 22% achieved an accuracy of 96.19%. Furthermore, SVM model testing on 235 testing data across 46 DDC classes achieved an accuracy of 96.17%, precision of 97.09%, recall of 95.90%, and F1-Score of 96.49%. These results indicate that the combination of TF-IDF and SVM provides good classification performance in determining the DDC class of books. The developed system also provides librarian validation, as well as model management and evaluation, allowing the classification results to be used as recommendations before being established as the final classification.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Hikma
Department of Informatics
University of Halu Oleo

Address campus Kendari
Email: hikmaaaa36@gmail.com

1. INTRODUCTION

Libraries are one of the sources of information that play an important role in supporting education, research, and knowledge development. In collection management, books need to be grouped according to their subjects or fields of knowledge so that collections can be systematically organized and information retrieval can be facilitated [1]. One of the widely used classification systems is the Dewey Decimal Classification (DDC), which organizes library materials hierarchically according to fields of knowledge [2]. The implementation of a classification system is also an important part of organizing and arranging library materials, particularly in supporting more structured collection management [3]. With a classification system, collections with similar subjects or characteristics can be placed into appropriate groups, allowing collection management and information retrieval to be carried out more systematically.

In practice, the book classification process still requires librarians to carefully understand the content of books and determine the appropriate classification class. As the number of collections increases and book topics become more diverse, manual classification may require more time. Systematic classification and arrangement of library materials are therefore necessary to support more effective collection management [4]. This condition indicates that the classification process requires not only an understanding of the content of library materials but also accuracy in determining the appropriate classification class. Along with technological development, this condition creates opportunities to utilize machine learning and text mining technologies to assist in document processing and classification [5]. The use of text-based approaches allows information contained in documents, such as titles and synopses, to be processed into data that can be used for automated classification.

Book data in the form of titles and synopses are textual data that need to be processed before being used in the classification process. The preprocessing stage is performed to produce more structured text that can be used in the feature extraction process. One method that can be used to represent textual data in numerical form is Term Frequency–Inverse Document Frequency (TF-IDF). The TF-IDF method assigns weights based on the frequency of occurrence of a term within a document and its occurrence across the entire document collection, allowing important characteristics of a document to be represented numerically [6]. The application of TF-IDF has also been used in web-based public complaint document classification [7]. This weighting approach allows originally unstructured textual information to be transformed into feature representations that can be utilized by machine learning algorithms for classification.

Text representation using TF-IDF can be combined with Cosine Similarity to measure the degree of similarity between documents. The application of Cosine Similarity can be used to determine similarity based on document feature representations [8]. The combination of TF-IDF and Cosine Similarity has also been applied in retrieving relevant articles based on the degree of similarity between documents and queries [9]. Based on this approach, Cosine Similarity can be used to determine the degree of similarity between data based on their feature representations. In this study, Cosine Similarity is not used as the primary classification algorithm but as a filtering stage to determine whether new book data remain within the scope of the data learned before proceeding to the SVM classification process. This filtering stage is used to help distinguish data that remain within the classification scope of the study from data that fall outside the defined scope.

After the weighting and filtering stages, the classification process is performed using the Support Vector Machine (SVM) algorithm. SVM is suitable for text classification because it can handle data with a high number of features. The application of SVM with TF-IDF in text classification has demonstrated that the combination of feature representation and classification algorithms can be used to determine document classes [10]. In DDC-based book classification, the use of TF-IDF and SVM achieved the highest accuracy of 80.55% using the Gaussian RBF kernel [11]. This result indicates that SVM can be used to classify books based on textual characteristics contained in documents. Other studies have also demonstrated the application of SVM to book data. The combination of TF-IDF and Word2Vec with SVM has been used for school library book classification [12]. Meanwhile, book titles and synopses have been utilized as features for book genre classification using SVM [13]. In DDC-based library collection classification, the TF-RF and K-Nearest Neighbor approaches have also been used to classify collections based on textual characteristics [14].

These previous studies demonstrate that textual information from books, particularly titles and synopses, can be utilized as a source of features in the classification process. However, the combination of TF-IDF as a weighting method, Cosine Similarity as a scope-filtering stage, and SVM as the DDC classification algorithm can still be further developed to assist the book classification process. Based on this condition, this study implements the Support Vector Machine (SVM) algorithm with Term Frequency–Inverse Document Frequency (TF-IDF) weighting for book classification based on DDC. The study uses

book titles and synopses as textual data and Cosine Similarity as a filtering stage before the SVM classification process. The classification scope focuses on engineering disciplines, namely Informatics Engineering/Computer Science, Civil Engineering, Mechanical Engineering, Architectural Engineering, and Electrical Engineering.

This study aims to develop a web-based book classification system capable of automatically providing DDC class recommendations based on book titles and synopses. The system is intended to assist librarians in the classification process by helping accelerate class determination and providing consistent classification recommendations. Although the prediction process is performed automatically using the SVM model, the classification results still need to be reviewed and validated by librarians before further use. Thus, the developed system does not replace the role of librarians but serves as a supporting tool for the classification decision-making process.

2. RESEARCH METHOD

This study was conducted through the stages of data collection, method analysis, system development, and testing. These stages were carried out to develop a book classification model based on the Dewey Decimal Classification (DDC) and implement it into a web-based system.

3.1 Data Collection

Data collection was conducted through literature study, interviews, and book data collection. The literature study was carried out by reviewing various sources related to book classification, text mining, Term Frequency–Inverse Document Frequency (TF-IDF) weighting, Cosine Similarity, and the Support Vector Machine (SVM) algorithm. Interviews were conducted with librarians to obtain information regarding the book classification process, the problems encountered, and the system requirements.

The book data used consisted of titles, synopses, and DDC labels. The data were obtained primarily from the catalog of the National Library of the Republic of Indonesia (Perpusnas) and partially from OPAC catalogs of other libraries available on the internet. The data were then selected based on the scope of the study, which included Informatics Engineering/Computer Science, Civil Engineering, Mechanical Engineering, Architectural Engineering, and Electrical Engineering. Data that met the specified criteria were subsequently used as training and testing data.

3.2 Analysis of the Method

The book title and synopsis data first undergo a preprocessing stage consisting of case folding, cleaning, stopword removal, and stemming. The results of the preprocessing stage are used to construct the vocabulary and transform the textual data into numerical representations using TF-IDF weighting. The resulting TF-IDF representations are then used to calculate Cosine Similarity between the book data and the class centroids. This stage serves as a filtering process to determine whether the book data remain within the scope of the learned classes. Data that meet the specified threshold value are subsequently processed through the SVM classification stage. SVM is used to determine the DDC class based on the features generated from TF-IDF weighting. The One-vs-Rest strategy is employed to handle multiclass classification. The classification results consist of DDC class recommendations based on the title and synopsis information of the books.

3.3 System Development

The system was developed using the Rational Unified Process (RUP), which consists of four stages: inception, elaboration, construction, and transition.

1. Inception

The inception stage was conducted to identify the problems, requirements, objectives, and scope of the book classification system to be developed. This stage produced an initial overview of the system requirements and the features needed.

2. Elaboration

The elaboration stage was conducted to design the system in greater detail. Activities at this stage included designing the system architecture, database, and user interface, as well as modeling the system using UML. In addition, the process flow for TF-IDF and SVM to be implemented in the system was designed.

3. Construction

The construction stage involved implementing the system based on the design that had been prepared. At this stage, program code development, implementation of the TF-IDF and SVM processes, data management features, and development of the web-based system interface were carried out. Initial testing was also conducted to ensure that each system function operated according to the design.

4. Transition

The transition stage involved implementing the system in the user environment. The system was directly tested to ensure that the book classification process operated according to the requirements. Improvements and refinements were made when deficiencies were identified during the testing process before the system was put into use.

3.4 Testing

Testing was conducted to evaluate the ability of the proposed method and system to classify books. *Cosine Similarity* testing was performed using several *threshold* values to determine the boundary of the data scope. Subsequently, SVM performance was evaluated using a *confusion matrix* with *accuracy*, *precision*, *recall*, and *F1-Score* metrics. User acceptance testing was conducted using the *User Acceptance Test (UAT)* to determine the level of user acceptance of the developed system.

3. RESULTS AND DISCUSSION

3.1. System Design

The system design was conducted using *Unified Modeling Language (UML)*. UML was used to describe the structure and behavior of the system so that system requirements and relationships between system components could be represented in a more structured manner. The use of UML in information system development can also support the analysis and documentation of the system[15]

3.1.1 Use Case Diagram

The *Use Case Diagram* is used to illustrate the interactions between actors and the main functions available in the system, as well as to represent the functional requirements from the users' perspective [16]. Based on the diagram, the system has two main actors, namely Admin and Librarian. The Admin is responsible for managing users, training data, model training, and model evaluation, while the Librarian manages book data and performs the classification process. During the classification process, the system internally performs *preprocessing*, TF-IDF weighting, and classification using SVM. The classification results are then displayed and stored in the system.

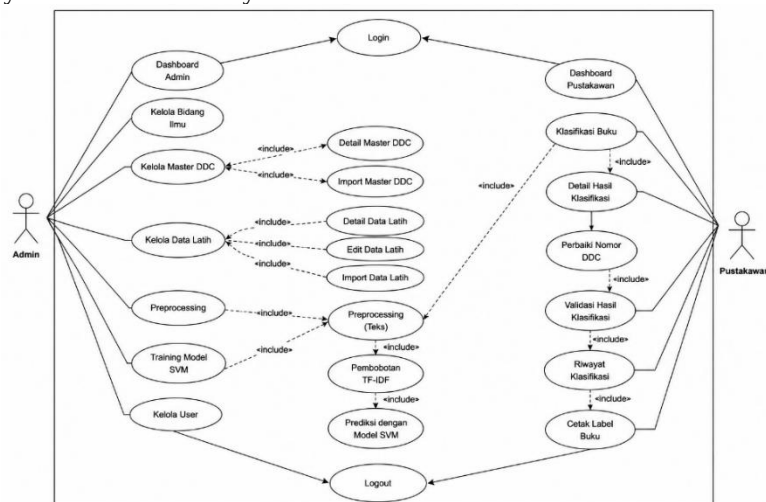


Figure 1. Use Case Diagram

3.1.2 Activity Diagram

The *Activity Diagram* for the Book Classification and Validation process was selected because it most directly represents the main process of the study. The *Activity Diagram* is used to illustrate the sequence of activities, decision-making processes, and workflow from beginning to end [17]. The diagram shows the interaction between the Librarian and the system, starting from entering book data, followed by *preprocessing*, TF-IDF weighting, SVM classification, and validation and storage of the classification results. This diagram was selected because it clearly illustrates the direct relationship between the research methods and their implementation in the system, allowing the application of TF-IDF and SVM within the system to be clearly demonstrated.

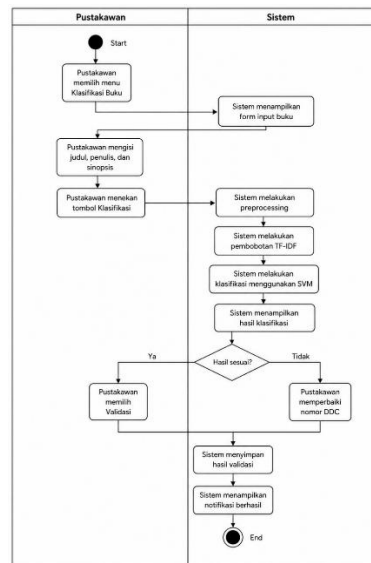


Figure 2. Activity Diagram

3.1.3 Class Diagram

The *Class Diagram* is used to illustrate the static structure of the system through classes, attributes, methods, and relationships between classes [18]. In the book classification system, the *Class Diagram* represents data structures such as users, books, training data, DDC, classification processes, and classification results, which are interconnected. This structure serves as the basis for database implementation and system functionality, ensuring that data management and the classification process can be integrated effectively.

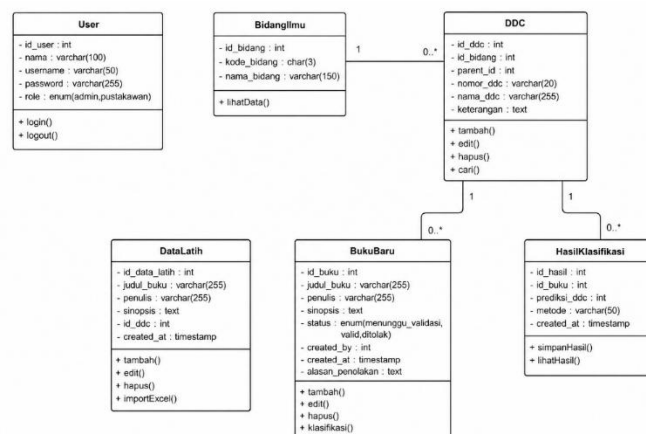


Figure 3. Class Diagram

3.1.4 System Interface

The book classification system is implemented as a web-based application designed to support the book classification and validation processes by users.

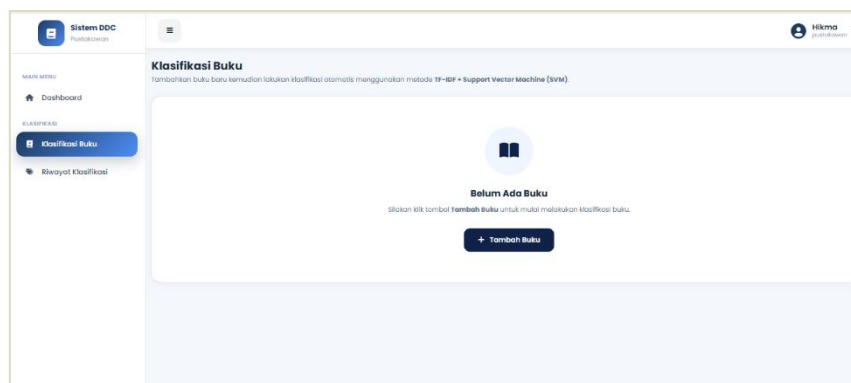


Figure 4. Halaman Klasifikasi Buku

The classification page is used to enter the title and synopsis of the book to be classified. The entered data are then processed by the system to generate a recommended DDC class..

Figure 5. Book Classification Results — Part 1

Figure 6. Book Classification Results — Part 2

The classification results page displays the DDC number recommendation generated by the system. The result can then be reviewed and validated by the librarian before being established as the final classification result.

NO	JUDUL BUKU	PENULIS	NOMOR DDC	AKSI
1	Multimedia Immersive Virtual Tour 3D Panorama 360: Publish Platform WEB, WIN, dan MAC	Zayid Muslofa	006.69	[Edit] [Hapus] [Tambah]
2	Multimedia Interaktif! Menggunakan Unity 2D	Suhendi	006.69	[Edit] [Hapus] [Tambah]
3	Kreasi Animasi Kartun dengan Adobe Flash	andi	006.69	[Edit] [Hapus] [Tambah]
4	Animasi Kartun dengan Flash MX	Ikaba	006.69	[Edit] [Hapus] [Tambah]

Figure 7. Training Data Page

The training data page is used to manage the data used in the formation and training of the classification model.

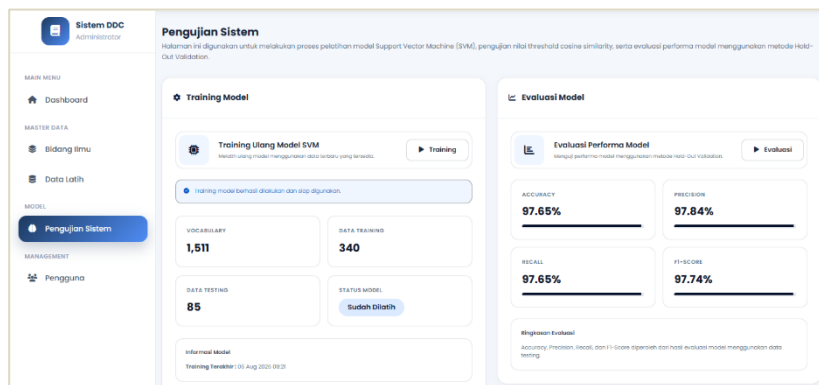


Figure 8. Model Evaluation Page

3.2. Results of Cosine Similarity Testing

The Cosine Similarity test was conducted to determine the threshold value used as the filtering boundary between data that remain within the scope (in-scope) and data outside the scope (out of category). The test used 20 out of category data and 85 in-scope data with eight threshold values, namely 15%, 18%, 20%, 22%, 25%, 30%, 35%, and 40%.

Table 1. Results of Cosine Similarity Threshold Testing

Threshold	Correctly Rejected OOC	Correctly Classified In-Scope	Total Correct	Overall Accuracy
15%	18/20	81/85	99/105	94,29%
18%	19/20	81/85	100/105	95,24%
20%	19/20	81/85	100/105	95,24%
22%	20/20	81/85	101/105	96,19%
25%	20/20	81/85	101/105	96,19%
30%	20/20	79/85	99/105	94,29%
35%	20/20	77/85	97/105	92,38%
40%	20/20	75/85	95/105	90,48%

Based on Table 1, the *threshold* values of 22% and 25% produced the highest accuracy of 96.19%. A *threshold* of 22% was selected for the system because it represents the lower threshold with the same accuracy, allowing more data that remain within the scope to proceed to the classification process. At a 22% threshold, all 20 OOC data were correctly rejected, while 81 of the 85 *in-scope* data were correctly classified. These results indicate that *Cosine Similarity* can assist in filtering data before the SVM classification process. When the threshold was increased to 30% or higher, the accuracy decreased because more *in-scope* data failed to meet the similarity threshold.

3.3. Classification Results

After passing the *Cosine Similarity* filtering stage, data that met the threshold were subsequently classified using the *Support Vector Machine* (SVM) algorithm. The evaluation was conducted using 235 testing data covering 46 DDC classes.

Table 2. Confusion Matrix Testing Results

Metrik Evaluasi	Value
Accuracy	96,17%
Precision	97,09%
Recall	95,90%
F1-Score	96,49%

Source: Research testing results.

Based on Table 2, the SVM model achieved an accuracy of 96.17%, with a precision of 97.09%, recall of 95.90%, and F1-Score of 96.49%. Of the 235 testing data, 226 were correctly classified, while 9 data were misclassified. These results indicate that the model was able to provide DDC classification with a high level of accuracy on the testing data. The F1-Score of 96.49% also indicates a good balance between the model's ability to produce accurate predictions and correctly identify data belonging to the appropriate classes.

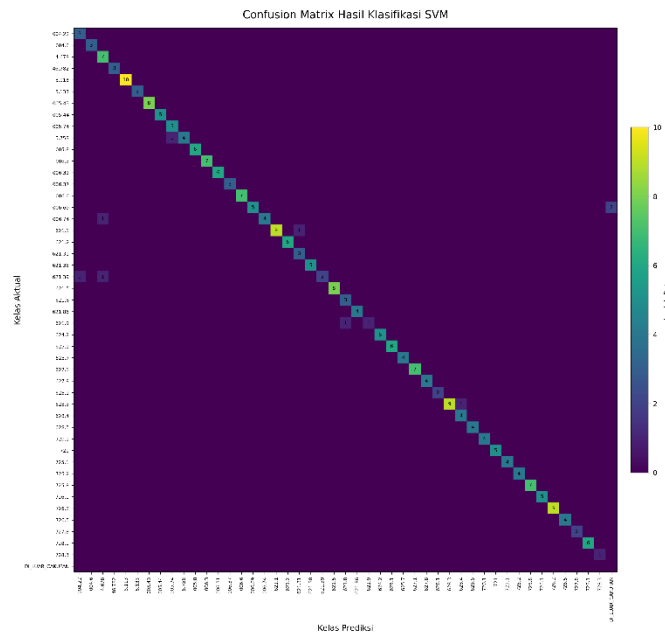


Figure 4. Confusion Matrix

The confusion matrix shows the comparison between the actual classes and the predicted classes. The results indicate that most of the data were classified into their corresponding classes, while 9 data were misclassified. These classification errors indicate that some books still have textual characteristics with similarities between different DDC classes.

3.4 User Acceptance Test (UAT)

The *User Acceptance Test* (UAT) was conducted to determine the level of user acceptance of the developed book classification system. The test involved the Admin and Librarian as system users.

Table 3. User Acceptance Test (UAT) Results

User	Percentage	Category
Admin	86,66 %	Very Feasible
Librarian	86,66%	Very Feasible
Average	86,66%	Very Feasible

Source: System testing results.

Based on Table 3, the UAT results obtained a percentage of 86.66% for both the Admin and Librarian, with an overall average of 86.66%. Based on the assessment category used, these results fall into the Very Feasible category. This result indicates that the system is acceptable and can be used to assist the book classification process.

4. CONCLUSION

Based on the research results, the implementation of the *Support Vector Machine* (SVM) algorithm with *Term Frequency-Inverse Document Frequency* (TF-IDF) weighting can be used to assist the book classification process based on *Dewey Decimal Classification* (DDC). The application of *Cosine Similarity* as a filtering stage resulted in a threshold of 22% with an accuracy of 96.19%. Furthermore, evaluation of the SVM model using 235 testing data across 46 DDC classes resulted in an accuracy of 96.17%, precision of 97.09%, recall of 95.90%, and F1-Score of 96.49%.

The developed web-based classification system also achieved a User Acceptance Test (UAT) score of 86.66%, which falls into the Very Feasible category. Therefore, the developed system can be used as a supporting tool in the book classification process, while the prediction results can still be validated by librarians before being established as the final classification.

REFERENCES

- [1] M. el-K. Kesuma, irva Yunita, and M. C. Putri, "Penggunaan Sistem Klasifikasi Di Perpustakaan Daerah Provinsi Lampung Sebagai Bentuk Peningkatan Pengelolaan Perpustakaan," *Jurnal Ilmu Perpustakaan dan Informasi*, vol. 5, pp. 85–96, 2021.
- [2] S. Dethan and N. Mayesti, "Penggunaan Dewey Decimal Classification Oleh Perpustakaan Akademik Di Indonesia," 2022. [Online]. Available: <http://journal.ummat.ac.id/index.php/JIPER/index>

- [3] T. Anggraini, S. T. Sumantri, and Khoirul, "Implementasi Pengklasifikasian Dan Penataan Bahan," *Jurnal Keislaman dan Kemsyarakatan*, vol. 8, no. 2, pp. 491–516, 2024.
- [4] F. A. R. Shalaby, Misroni, and ahmad Wahidi, "Pengklasifikasian Dan Penataan Ulang Buku Di Perpustakaan Sdn 15 Gelumbang," 2019.
- [5] E. Febriyanti, E. Noviyantono, and D. M. Praseptian, "Rekayasa Aplikasi Rekomendasi Pencarian Lokasi Dan Analisis Sentimen Menggunakan Penambangan Teks," *Journal of Big Data Analytic and Artificial Intelligence*, vol. 7, no. 2, pp. 69–74, 2024.
- [6] M. Das, S. Kamalanathan, and P. Alphonse, "A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset," 2020.
- [7] R. R. Anugrah, "Penerapan Cosine Similarity Dan Pembobotan Tf-Idf Untuk Klasifikasi Pengaduan Masyarakat Berbasis Web (Studi Kasus : Bagwassidik Ditreskrimum Polda Kalbar)," *Coding Jurnal Komputer dan Aplikasi*, vol. 11, no. 1, p. 100, May 2023, doi: 10.26418/coding.v11i1.55598.
- [8] R. F. Kurniawan, "Implementasi Text Mining Menggunakan Metode Cosine Similarity Untuk Klasifikasi Konten Berita Di Postingan Grup Facebook Info Lantas Dan Kriminal Pasuruan," *JAMI: Jurnal Ahli Muda Indonesia*, vol. 3, no. 1, pp. 9–17, Jun. 2022, doi: 10.46510/jami.v3i1.41.
- [9] S. Madoa, I. Mulyadi, and Darniati, "Implementasi Metode TF-IDF dan Cosine Similarity pada Sistem Pencarian Artikel yang Relevan," *Journal of Muhammadiyah's Application Technology*, vol. 5, no. 2, pp. 237–244, 2026, doi: 10.26618/a6wkdt68.
- [10] N. Arifin, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (Svm) Dengan Tf-Idf N-Gram Untuk Text Classification," 2021.
- [11] D. H. Amalia and W. Yustanti, "Klasifikasi Buku Menggunakan Metode Support Vector Machine pada Digital Library," *Journal of Informatics and Computer Science*, vol. 03, no. 01, pp. 55–61, 2021, [Online]. Available: <https://opac.unesa.ac.id/>,
- [12] S. N. Cahyani and G. W. Saraswati, "Implementation Of Support Vector Machine Method In Classifying School Library Books With Combination Of Tf-Idf And Word2vec," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 6, pp. 1555–1566, Dec. 2023, doi: 10.52436/1.jutif.2023.4.6.1536.
- [13] N. W. E. R. Dewi, I. M. S. Putra, and D. P. Githa, "Klasifikasi Genre Buku Berbasis Judul dan Sinopsis Menggunakan Metode Support Vector Machine," 2025.
- [14] A. Sulala et al., "Klasifikasi Koleksi Perpustakaan Berbasis Dewey Decimal Classification Menggunakan Algoritma Tf-Rf Dan K-Nearest Neighbor," *Jurnal Informatika Polinema*, vol. 12, no. 3, pp. 465–474, 2026.
- [15] S. W. Ramdany, S. A. Kaidar, B. Aguchino, C. A. A. Putri, and R. Anggie, "Penerapan UML Class Diagram dalam Perancangan Sistem Informasi Perpustakaan Berbasis Web," 2024.
- [16] Ichсандi, W. Yanto, H. Alhaq, R. S. Sari, and M. Juanda, "Implementasi UML dalam Desain Sistem Informasi Program Studi SI di Universitas Merangin," 2025.
- [17] S. Narulita, A. Nugroho, and M. Z. Abdillah, "Diagram Unified Modelling Language (UML) untuk Perancangan Sistem Informasi Manajemen Penelitian dan Pengabdian Masyarakat (SIMLITABMAS)," *Jurnal publikasi Sistem Informasi dan Telekomunikasi*, vol. 2, no. 3, pp. 244–256, Aug. 2024, doi: 10.62951/bridge.v2i3.174.
- [18] I. D. Perwitasari, J. Hendrawan, F. Y. Panggabean, and M. Raihansyah, "Model UML Aplikasi Augmented Reality Pengenalan Desa Pertumbuhan," *Jurnal Minfo Polgan*, vol. 13, no. 2, pp. 1887–1896, Dec. 2024, doi: 10.33395/jmp.v13i2.14301.