INVOTEK

Jurnal Inovasi Vokasional dan Teknologi

http://invotek.ppj.unp.ac.id/index.php/invotek

Vol. 25, No. 3, Oct 2025

ISSN: 1411 – 3411 (p)
ISSN: 2549 – 9815 (e)

Grouping of Research Interests of Final Project of Computer Science Students at UINSU Using the Fuzzy C-Means Approach

Harun Al Rasyid[1] *, Muhammad Ikhsan[1]

[1] Department of Computer Science, State Islamic University of North Sumatra
Jl. Willem Iskandar, Pasar V, Medan Estate, Deli Serdang, North Sumatra, Indonesia-20371

Corresponding author: haarnst@gmail.com

Doi: https://doi.org/10.24036/invotek.v25i3.1332

Abstract

Determining research topics for final projects that align with students' competencies and interests remains a challenge in academic management because the process is often influenced by subjective factors. This situation can lead to various impacts, such as inappropriate research topics, delays in study completion, and imbalances in the guidance workload among supervisors. This study aims to cluster the research interests of students preparing their final projects in the Computer Science Study Program at the State Islamic University of North Sumatra (UINSU) using the Fuzzy C-Means (FCM) algorithm. The data used were obtained from a student interest questionnaire from the 2020–2022 intake and historical data on final project titles. Furthermore, the text data underwent preprocessing and was converted into numerical form using the Term Frequency–Inverse Document Frequency (TF-IDF) method. The FCM algorithm was then used to form research interest clusters with fuzzy membership degrees. Based on the results of the cluster quality evaluation, it was found that the most optimal number of clusters was six, with a Silhouette Index value of 0.6311 and a Davies–Bouldin Index of 0.5505, which indicates that the cluster structure formed is classified as good. The clustering results indicated that student interests were dominated by Software Engineering and Artificial Intelligence, with a fairly high degree of overlap. This study combines student interest questionnaire data and historical final project title data, represented using TF-IDF and clustered using the Fuzzy C-Means algorithm to map multidimensional research interests. The results suggest that this approach provides a more objective basis for identifying students' research tendencies and can support topic recommendation systems and academic supervision planning.

Keywords: Fuzzy C-Means, Clustering, Research Interests, Final Projects, Computer Science.

1. Introduction

Determining a final project research topic according to a student's interests and abilities often presents obstacles due to a lack of objective guidance. A process that still relies on subjective assessments by both students and supervisors has the potential to lead to repeated topic changes, delays in graduation, and an uneven distribution of the supervision load [1]. Bibliometric studies confirm that the suitability of areas of interest and the effectiveness of academic management have become a major focus in the Southeast Asian region in the last two decades [2], [3].

To overcome this problem, a classification approach can be used with an unsupervised learning algorithm such as clustering [4]. Three commonly used methods are K-Means, Hierarchical Clustering, and Fuzzy C-Means (FCM). K-Means excels in speed, but can only group data rigidly. Hierarchical Clustering is suitable for visualization, but is less efficient on large data. Meanwhile, FCM is more flexible because it allows one data to enter several groups at once, according to the complexity of student interests. Although sensitive to initial parameters, FCM is considered more appropriate for representing non-singular interest tendencies [5].

The main problem underlying this research is the lack of an automatic data-based classification system to group the research interests of UINSU Computer Science students in their final projects. Currently, topic selection is still done manually based on personal perceptions, without the support of

Received date 10/02/2026, Revised date 14/04/2026, Accepted date 22/04/2026

objective data mapping. As a result, there is often a mismatch between student interests, topics, and competencies, which results in delays in the completion of final projects and increases the burden on supervisors [1]. The absence of a computational approach such as clustering to analyze these interest trends is one of the obstacles to more effective academic management. As an effort to address this problem, this study proposes the use of the Fuzzy C-Means Clustering algorithm to group the research interests of UINSU Computer Science students based on historical data, such as previous final projects topics, interest questionnaire results, or keywords from research titles [6]. This approach was chosen because FCM is able to capture the complexity of student interests which are often not limited to a single field [7].

Several recent studies have advanced the application of clustering and text mining in academic contexts. Damayanti et al. [1] developed a content-based final project supervisor recommendation system using TF-IDF and cosine similarity. Hill et al. [2] employed BERTopic for interest mining to recommend academic programs. Arrisalah et al. [8] designed a final project topic recommendation system using TF-IDF and cosine similarity. Urva et al. [9] applied Fuzzy C-Means for student academic interest clustering based on course selection data. Ali et al. [10] used FCM to cluster students based on engagement levels in educational contexts. Earlier studies have also applied FCM to group student interests, such as at Samarinda State Polytechnic (four clusters, MAPE 27.07%) and at Muhammadiyah University of Gresik (three clusters, silhouette 0.532) [5], [11]. However, these earlier studies generally used only one type of data source and did not conduct comprehensive FCM parameter experiments.

Despite these contributions, several gaps remain. First, most previous studies rely on a single data source either questionnaires [5], course selection data [9], or final project titles only [1], [8] without integrating multiple data types. Second, while FCM has been applied in academic contexts [5], [6], [9], [10], none have combined questionnaire data with historical final project title data using TF-IDF representation and validated the optimal number of clusters comprehensively using both Silhouette Index and Davies-Bouldin Index across a systematic range of K values (2 to 10). Third, existing FCM-based studies on student interests [5], [7] typically assume a fixed number of clusters (e.g., K=3) without empirical validation, potentially overlooking the multidimensional nature of student research preferences.

The novelty of this research lies in the application of the Fuzzy C-Means (FCM) algorithm to group the research interests of UINSU Computer Science students' final projects by utilizing a combination of interest questionnaire data and historical data of final project titles represented using TF-IDF and validated through the Silhouette Index and Davies–Bouldin Index [12]. This study contributes by integrating questionnaire-based student interest data with historical final project title data, represented using TF-IDF and analyzed using Fuzzy C-Means clustering. Unlike previous studies that rely on a single data source, this approach captures multidimensional and overlapping research interests more effectively. Therefore, the results provide a more objective and data-driven foundation for mapping student interests and supporting academic decision-making [13].

2. Method

2.1 Data Collection Techniques

This study involved 167 respondents from students of the Computer Science Study Program at Universitas Islam Negeri Sumatera Utara (UINSU), batch 2020–2022. Data collection was conducted using a questionnaire and supported by 12 historical final project titles**,** which were treated as text documents for analysis. The questionnaire consists of 9 items, divided into three main domains: Artificial Intelligence, Software Engineering, and Computer Networks, with 3 questions in each domain. These items are designed to measure students' level of interest in each research area. All questionnaire items were measured using a Likert scale ranging from 1 to 5, where 1 indicates very low interest and 5 indicates very high interest. The scores for each domain were then aggregated to obtain a total interest score for each respondent.

In addition to the questionnaire, respondents were also asked to select one preferred final project title/topic. The selected titles, along with historical final project title data, were treated as text documents and analyzed using the Term Frequency–Inverse Document Frequency (TF-IDF) method to extract important keywords representing students' research tendencies. Thus, this study combines numerical data from Likert-scale questionnaires and textual data from final project titles, providing a more

P-ISSN: 1411-3414
E-ISSN: 2549-9815

comprehensive representation of students' research interests. The collected data was then used as input in the clustering process using the Fuzzy C-Means (FCM) algorithm, which allows flexible clustering by considering the possibility of overlapping student interests across multiple research areas.

2.2 Design

The flowchart that has been designed is shown below:

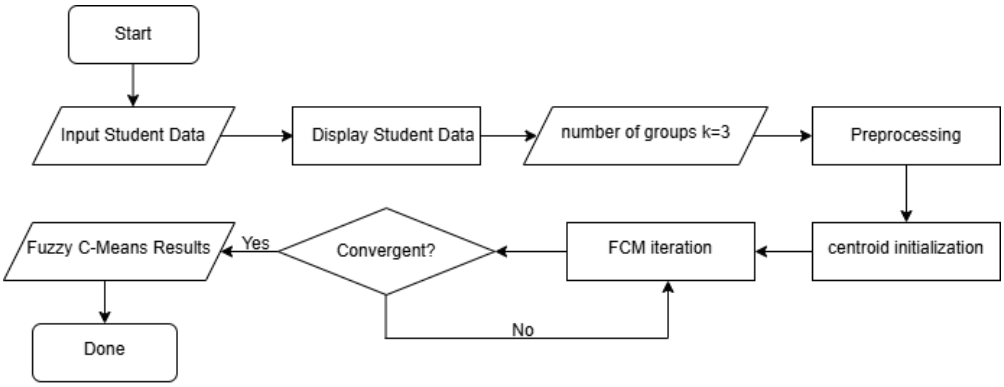

Figure 1. Research Workflow of FCM-Based Clustering Integrating Questionnaire Responses and TF-IDF Weighted Text Data

Figure 1 illustrates the research workflow, which begins with data collection in the form of a student interest questionnaire and historical data on final project titles. Data preprocessing was then performed to ensure that the data was clean and ready for use in the processing stage. Furthermore, weighting was performed using the TF-IDF method to convert the text data into numeric representations. The next stage was the application of the Fuzzy C-Means (FCM) algorithm to group student interests into several clusters based on the data patterns found. After the clustering process was complete, the quality of the resulting clusters was evaluated using internal validation metrics, namely the Silhouette Index and the Davies–Bouldin Index. The flowchart shows that the research was carried out in a structured and systematic manner, starting from the data input stage to the evaluation process of the obtained cluster results [14].

2.3 Data Preprocessing

The data processing in this study involves two types of data, namely numerical data from questionnaires and textual data from final project titles. For the numerical data, the questionnaire responses measured using a Likert scale (1–5) were normalized using StandardScaler from the Scikit-learn library to ensure uniform feature scaling and to prevent bias during the clustering process [15]. For the textual data, preprocessing was conducted through several stages, including case folding, tokenization, stopword removal, and stemming to normalize the text into its root form. The processed text was then transformed into numerical features using the Term Frequency–Inverse Document Frequency (TF-IDF) method implemented with Scikit-learn. In this study, TF-IDF was applied using a unigram representation to convert the processed text into numerical feature vectors. Furthermore, Uniform Manifold Approximation and Projection (UMAP) was applied to reduce the dimensionality of the TF-IDF feature space while preserving the data structure [16].

Furthermore, dimensionality reduction was applied using the Uniform Manifold Approximation and Projection (UMAP) algorithm to reduce the high-dimensional TF-IDF feature space while preserving the intrinsic structure of the data. In this study, UMAP was configured with n_neighbors = 15 and min_dist = 0.1.

Afterward, the normalized numerical data and the reduced textual features were combined into a single feature set using feature concatenation. The combined data were then re-normalized using StandardScaler to ensure consistency across all features before being used as input for the clustering process. Finally, clustering was performed using the Fuzzy C-Means (FCM) algorithm, and the results were evaluated using the Silhouette Index (SI) and Davies–Bouldin Index (DBI) to determine the optimal clustering performance.

P-ISSN: 1411-3414
E-ISSN: 2549-9815

2.4 Fuzzy C-Means (FCM) Method

Figure 2 shows an illustration of the clustering concept using the Fuzzy C-Means method, where each data object is not only absolutely included in one cluster, but has a degree of membership in several clusters at once. This illustration emphasizes the main characteristics of FCM as a soft clustering method that is able to represent data conditions that have relationships between groups. In the context of this research, this approach is relevant because student interests are often not limited to one particular field, but can overlap between Artificial Intelligence, Software Engineering, and Computer Networks [10].

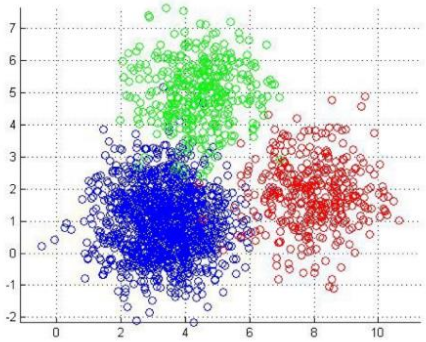

Figure 2. Illustration of Clustering Results Using the Fuzzy C-Means Method

In this study, the fuzziness coefficient was set to m = 2, the maximum number of iterations was set to 1000, and the stopping criterion (error tolerance) was set to 0.005. The initial membership matrix was generated randomly. The algorithm iteratively updates cluster centers and membership values until convergence is achieved [10].

To determine the optimal number of clusters, the number of clusters (K) was varied from 2 to 10. The range K=2 to K=10 was selected based on the following considerations: K=1 is trivial and meaningless for clustering, while K values greater than 10 would produce too many clusters relative to the total number of respondents (n=167) and increase the risk of over-segmentation. This range also aligns with previous clustering studies on student interest mapping and allows sufficient exploration from very coarse (K=2) to relatively detailed (K=10) groupings. For each value of K, clustering was performed using the Fuzzy C-Means algorithm and evaluated using the Silhouette Index (SI) and Davies-Bouldin Index (DBI). The optimal number of clusters was selected based on the highest SI value and the lowest DBI value.

2.5 Term Weighting

Term weighting is a text representation technique that assigns a weight to each term in a document based on its importance within the context of a single document relative to the entire document collection [14]. This technique aims to separate informative terms (high-value terms) from general terms (widely distributed ones), allowing term weighting to form more effective document vectors for classification or clustering tasks, where high-weighted terms reflect relevance to a particular topic, while low-weighted terms are considered less informative. Approaches such as TF-IDF are examples of practical applications of this technique in various modern document processing systems [16].

In text mining or text classification practices, term weighting is commonly integrated at the document representation stage before applying the learning algorithm. The application of these weights allows the model to focus on keywords that distinguish one document from another. Various term weighting methods (e.g., TF-IDF, TF-IDF-ICF, and other variants) have been compared to determine their impact on classification or clustering performance. Research has shown that selecting an appropriate weighting scheme can improve modeling accuracy and quality, especially for datasets with multiple classes or diverse topics [17].

After preprocessing, text data needs to be represented numerically so it can be processed by the FCM algorithm. One technique used is TF-IDF (Term Frequency–Inverse Document Frequency), which calculates the weight of each word based on its frequency of occurrence and distribution across documents [16].

P-ISSN: 1411-3414
E-ISSN: 2549-9815

$$TF - IDF(d,t) = TF(d,t) * IDF(t) \quad (1)$$

$$TF(d,t) = \frac{f_{t,d}}{\Sigma_{t\prime \epsilon d} f_{t',d}} \quad (2)$$

$$IDF(t) = \log(\frac{N}{DF(t)}) \quad (3)$$

where:

$f_{t,d}$: Number of occurrences of term t in document d
$\Sigma_{t\prime \epsilon d} f_{t\prime,d}$: Total number of all terms in document d
N : Total number of documents in the collection
$DF(t)$: Number of documents containing term t

Table 1 shows an example of preprocessing results in the form of a list of key words from several final project title documents. Each document is represented as a collection of words that have undergone text cleaning processes such as tokenization, stopword removal, and word normalization. This table shows that the words that appear are dominated by research-relevant terms such as “analysis”, “artificial intelligence”, “iot”, “network”, and “monitoring”.

Table 1. Extracted Keywords from Five Sample Final Project Titles After Preprocessing (Case Folding, Tokenization, Stopword Removal, and Stemming)

D1	['analysis', 'applied', 'artificial', 'intelligence', 'prediction', 'performance', 'academic', 'student', 'ilkomp', 'uinsu']
D2	['study', 'characteristics', 'data', 'sensor', 'iot', 'smart', 'agriculture', 'uinsu']
D3	['exploration', 'data', 'science', 'analysis', 'behavior', 'for', 'app', 'campus']
D4	['map', 'pattern', 'interaction', 'student', 'chatbot', 'ai', 'ilkomp', 'uinsu']
D5	['evaluation', 'mainstay', 'network', 'iot', 'monitoring', 'environment', 'campus']

These results serve as the basis for the TF-IDF weighting process, as these key words will then be calculated for their frequency of occurrence and importance within the entire document.

$$TF('study, D1) = \frac{1}{8} = 0.125 \quad (4)$$

The word ‘study’ appears once in D2, and 8 is the sum of all terms in D2. After performing all calculations, the term frequency values for the five main words are displayed in Table 2.

Table 2. Term Frequency (TF) Values for Selected Keywords Across Five Documents, Showing Word Dominance Patterns

Term	D1	D2	D3	D4	D5
Analysis	0.1	0	0.125	0	0
Data	0	0.125	0.125	0	0
Study	0	0.125	0	0	0
Iot	0	0.125	0	0	0.143
Uinsu	0.1	0.125	0	0.125	0

Table 2 displays the term frequency (TF) values for several keywords in each document. Term frequency indicates how frequently a word appears in a document compared to the total number of words in the document. The higher the TF value, the more dominant the word's occurrence in the document. The table shows that some words, such as "uinsu" and "iot," appear more frequently in certain documents, reflecting the differences in thematic focus of each final project title. These TF values are then used as the initial component in the TF-IDF weighting process.

$$IDF('study') = log\frac{5}{1} = 0.699 \quad (5)$$

P-ISSN: 1411-3414
E-ISSN: 2549-9815

The IDF value is calculated by dividing the logarithm of the number of documents by the number of occurrences of the term. Here, 5 is the total number of documents and 1 is the number of occurrences of the word "study."

Table 3. Document Frequency (DF) and Inverse Document Frequency (IDF) Values, Where Lower DF Produces Higher IDF Indicating Term Specificity

Term	DF	IDF
Analysis	1	0.699
Data	2	0.398
Study	1	0.699
Iot	2	0.398
Uinsu	3	0.222

Table 3 shows the document frequency (DF) and inverse document frequency (IDF) values for each keyword. The DF value describes the number of documents containing a particular word, while the IDF indicates the word's rarity within the entire document collection. Words with a lower DF value will have a higher IDF value, making them more specific and providing more valuable information. Conversely, words that appear frequently in many documents will have a lower IDF because they are considered less able to distinguish one document from another. This table shows that words that appear in only a few documents have a higher IDF weight and are therefore more important in the weighting process.

$$TF - IDF('study', D3) = TF('study, D3) * IDF('study') = 0.125 * 0.699 = 0.0874 \quad (6)$$

TF-IDF is obtained by multiplying DF (tf) and IDF(t). Here, we can see that 0.125 is the DF value of the study term in document 3, and 0.699 is the IDF value of '*study'*.

Table 4. Final TF-IDF Weights for Selected Keywords, Representing Term Importance in Each Document Relative to the Entire Corpus

Term	D1	D2	D3	D4	D5
Analysis	0.0699	0	0.0874	0	0
Data	0	0.0498	0.0498	0	0
Study	0	0.0874	0	0	0
Iot	0	0.0498	0	0	0.0570
Uinsu	0.0222	0.0278	0	0.0278	0

Table 4 shows the final TF-IDF weighting results obtained by multiplying the TF and IDF values. The TF-IDF value describes the level of importance of a word in a particular document compared to the entire document. Words that have a high TF in a particular document but rarely appear in other documents will produce a large TF-IDF value. This table shows that certain words have a more dominant weight in certain documents, so it can be used as an effective numerical representation for the clustering process using Fuzzy C-Means. Thus, this table serves as the basis for transforming text data into a numeric vector form ready for analysis.

3. Result and Discussion

3.1 Descriptive Statistical Analysis of Respondent Data

Figure 3 shows the distribution of respondents based on the student cohorts studied, namely the 2020, 2021, and 2022 cohorts. This distribution illustrates that the research data encompasses several different cohort groups, so the analysis results obtained do not only describe the interests of one cohort, but are more representative of the UINSU Computer Science student population over a three-year period. This cohort diversity is also important because each cohort can have different interest tendencies according to technological trends and the curriculum they are pursuing.

P-ISSN: 1411-3414
E-ISSN: 2549-9815

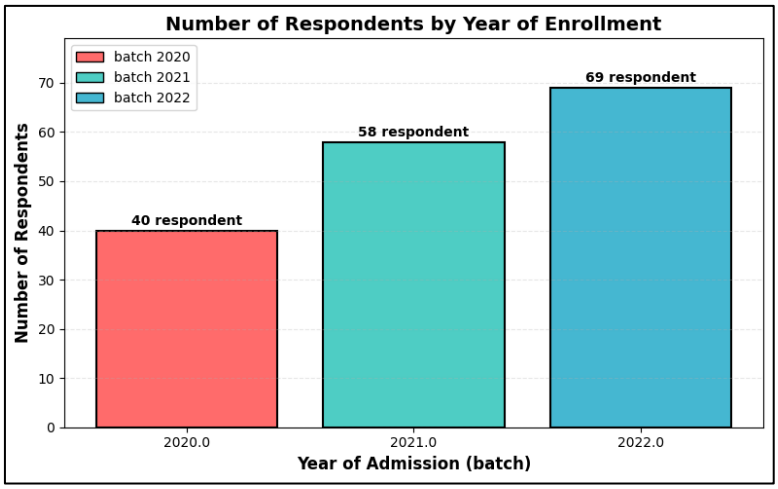

Figure 3. Respondent Distribution Across 2020, 2021, and 2022 Cohorts Showing Balanced Representation

3.2 Descriptive Statistical Analysis of Research Data

3.2.1 Categorization of Interest Questionnaire Answers

Each respondent answered 3 questions in each area of interest: Artificial Intelligence (AI), Software Engineering (SEE), and Computer Networks. All answers were then summed to obtain a total score for each area. For ease of understanding, the scores were further classified into three categories: Low, Medium, and High.

3.2.2 Distribution of Respondents' Interests Per Field

The dominance of Software Engineering and Artificial Intelligence observed in Figure 4 is not merely a descriptive finding, but reflects broader technological and academic trends. The rapid growth of digital industries, particularly in application development and intelligent systems, has increased student exposure and interest in these areas. In contrast, Computer Networks, while fundamental, tends to be perceived as less directly visible in terms of output, which may reduce its attractiveness among students.

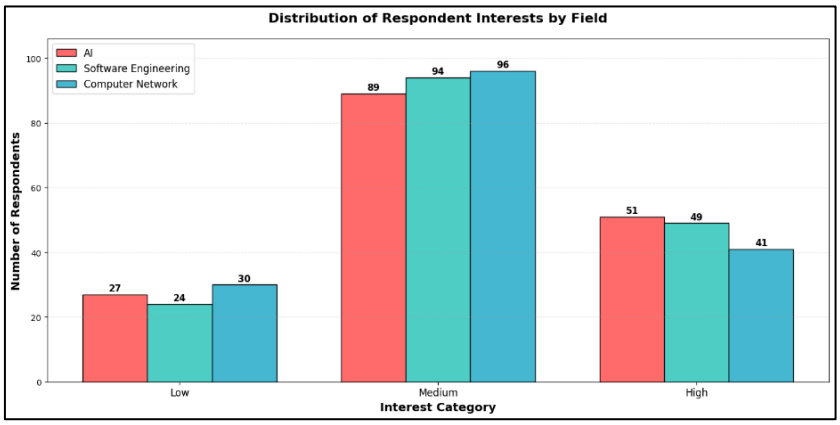

Figure 4. Distribution of Student Interests Per Domain: Software Engineering and AI Dominate Over Computer Networks

In addition, the overlap between interest areas indicates that student preferences are inherently multidimensional. Many students are not confined to a single domain but instead combine interests, such as developing intelligent applications that integrate Artificial Intelligence and Software Engineering. This phenomenon confirms the relevance of using Fuzzy C-Means, as it is capable of capturing the fuzzy boundaries between these domains.

The six-cluster structure further provides a more detailed representation of student interests compared to the initial three-domain framework. This structure allows academic stakeholders to identify

more specific subfields, which is particularly useful for designing targeted topic recommendations, improving supervisor allocation, and supporting adaptive curriculum planning.

3.2.3 Analysis of Answers Per Questionnaire Indicator

The analysis revealed that terms such as "machine learning" and AI-related topics frequently appeared in historical data. This finding aligns with questionnaire results, which demonstrated high student interest in the field of AI. This similar pattern suggests that student interest in the field is not simply spontaneous but also reflected in emerging research trends over the past few years.

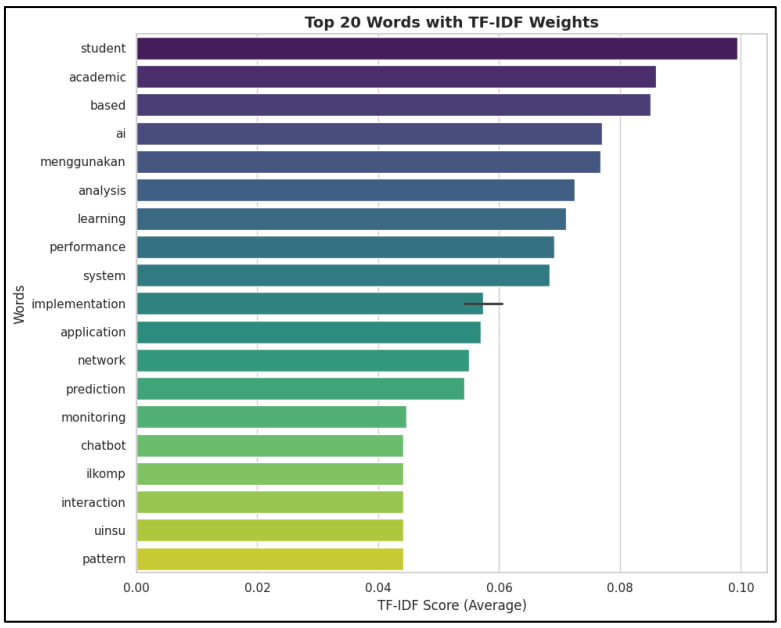

Figure 5. Top Keywords from Historical Final Project Titles Based on TF-IDF Weighting, with 'Analysis' and 'System' as Dominant Terms

Figure 5 shows the most highly weighted words based on the TF-IDF calculation of student final project titles analyzed using historical data. Highly weighted words indicate terms that appear frequently and have a high level of relevance in the collection of final project title documents. These results show that words such as "analysis," "system," "implementation," and other technology-related terms appear predominantly. This indicates that student final project research largely focuses on system development and the study of technology applications. Furthermore, the emergence of terms related to AI and machine learning also indicates an increasing research trend in the field of artificial intelligence, further confirming the relevance of this study in mapping student interests based on real trends found in final project title data. This pattern strengthens the argument that student interests are influenced not only by personal preference but also by exposure to trending research topics.

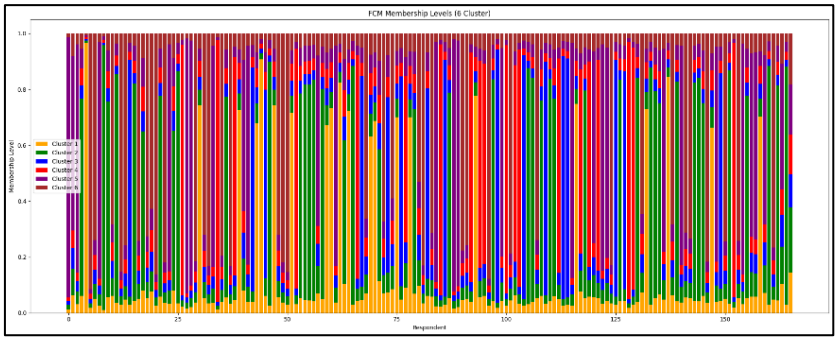

Figure 6. Membership Degree Matrix of 167 Respondents Across Six FCM Clusters, Showing Multidimensional Interest Patterns

Figure 6 shows the membership degree matrix generated by the Fuzzy C-Means (FCM) algorithm, which depicts each respondent's membership level in the six clusters formed. In the matrix, each row

P-ISSN: 1411-3414
E-ISSN: 2549-9815

represents one respondent, while each column indicates a specific cluster. The value or color intensity displayed reflects the respondent's membership level in each cluster. The analysis results indicate that many respondents not only have dominant membership in one cluster but also exhibit quite high membership levels in several other clusters. This condition indicates that students' research interests are multidimensional and often overlap. Therefore, the FCM method is considered appropriate for representing the characteristics of student interests, which are not always singular or rigid.

3.3 Inferential Statistical Analysis

3.3.1 Results of Fuzzy C-Means Clustering

Figure 7 shows the results of combining six clusters generated by the Fuzzy C-Means method into three main areas: Artificial Intelligence, Software Engineering, and Computer Networks. This visualization illustrates the tendency of respondents' membership levels in the three areas after remapping the clustering results. From these results, it can be seen that most respondents have a high membership level in the AI and Software Engineering fields, while the value in the Computer Networks field is relatively lower. This indicates that student interest is more directed towards software development and artificial intelligence, while the field of computer networks is not a top priority. This finding is also in line with the current development of the digital industry which places more emphasis on software-based technology and artificial intelligence.

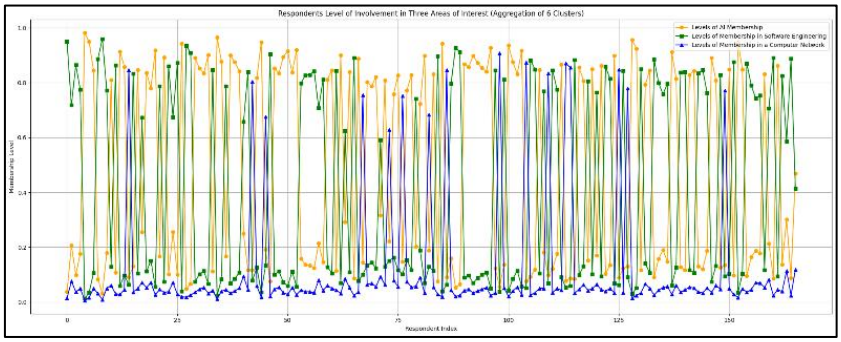

Figure 7. Aggregated Membership Degrees of Six Clusters into Three Domains, Confirming AI and Software Engineering Dominance

Although this study initially departed from three main domains (Artificial Intelligence, Software Engineering, and Computer Networks), the optimal number of clusters was not assumed to be three. The Fuzzy C-Means algorithm was applied with K values ranging from 2 to 10, and the best K was selected based on cluster quality evaluation using the Silhouette Index and Davies-Bouldin Index. The evaluation results showed that six clusters produced the best structure, not three.

Conceptually, these six clusters are not replacements for the three domains but rather sub-specializations that provide a more detailed view of student interests. For instance, the Artificial Intelligence domain splits into two distinct clusters: AI for prediction/analytics and AI for interactive systems. The Software Engineering domain divides into application development and software evaluation. The Computer Networks domain forms its own cluster, often combined with IoT topics, while an additional cluster captures overlapping interests between AI and Software Engineering.

Using six clusters instead of three provides a dual-layer perspective that would not be possible with only three clusters. The three-domain view offers a macro-level understanding for curriculum planning, while the six-cluster view reveals specific interest specializations that help in recommending precise final project topics and allocating supervisors more effectively. In contrast, forcing the data directly into three clusters would overlook the nuanced overlaps and multidimensional nature of student interests, such as students who have strong interests in both AI and Software Engineering simultaneously.

3.3.2 Cluster Quality Evaluation

To ensure that the clustering results have a clear and meaningful structure, an evaluation process was conducted using two internal validation metrics: the Silhouette Index (SI) and the Davies–Bouldin

Index (DBI). These two metrics were calculated by testing various cluster counts (K), ranging from 2 to 10.

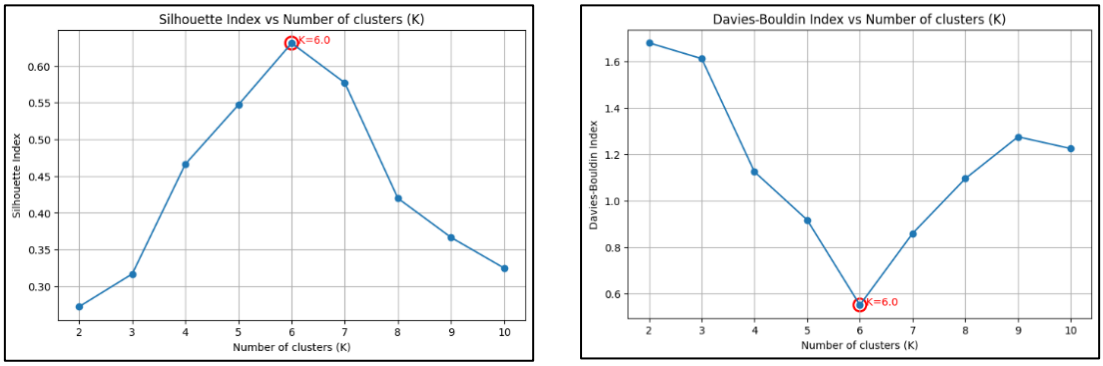

Figure 8. Silhouette Index and Davies-Bouldin Index values for K=2 to 10, with optimal performance at K=6

Figure 8 shows the results of cluster quality evaluation with varying number of clusters (K) from 2 to 10 using two internal validation metrics, namely the Silhouette Index (SI) and the Davies–Bouldin Index (DBI). The Silhouette Index serves to assess how well a data is located in its cluster compared to other clusters, where a higher value indicates a better quality cluster structure. Meanwhile, the Davies–Bouldin Index is used to measure the level of similarity between clusters, where a smaller value indicates a better and optimal cluster separation. Based on the graph shown, the highest Silhouette Index value and the lowest DBI value were obtained at K = 6. This indicates that the six-cluster configuration is the most optimal for the dataset in this study, so the number of clusters selected as the final result is K = 6.

Table 5. Cluster Quality Evaluation Results Showing Optimal Performance at K=6 with SI=0.6311 (Good Category) and Lowest DBI=0.5505

K	SI	DBI
2	0.2719	1.6089
3	0.3165	1.6128
4	0.4662	1.1250
5	0.5502	0.8939
6	0.6311	0.5505
7	0.5185	0.8636
8	0.4765	1.0935
9	0.3663	1.2758
10	0.3244	1.2251

Based on the Table 5 and graphs presented, it can be seen that K = 6 provides the best combination of cluster qualities. The Silhouette Index value reached 0.6311, which is included in the "good" category according to Rousseeuw's standard, where values above 0.5 indicate that the cluster structure is quite clearly separated. On the other hand, the Davies–Bouldin Index value of 0.5505 is the lowest value compared to all the configurations of the number of clusters tested. This indicates that the use of six clusters produces the most natural data structure, with a high degree of homogeneity within clusters and clear differences between clusters. These findings provide a strong basis for determining the final number of clusters in the FCM analysis, while also confirming that the complexity of student interests requires a more detailed representation than just the three main, classical categories.

3.4 Discussion

In this section, the research results are not only presented quantitatively but also interpreted by linking them to the theory and literature review discussed in Chapter II. The findings in this study not only reflect individual student preferences but also illustrate the dynamics of computer science development in the local academic environment. The high interest in Software Engineering and Artificial Intelligence (AI), and the relatively low interest in Computer Networks, is not a coincidental

P-ISSN: 1411-3414
E-ISSN: 2549-9815

phenomenon. This pattern aligns with the development of the digital industry, which increasingly emphasizes software-based solutions and artificial intelligence. Meanwhile, network infrastructure, while still playing a fundamental role, often serves only as a foundational layer that is less visible to end users. Therefore, students tend to connect more easily, emotionally and academically, with fields whose results can be directly realized in the form of applications, interfaces, or predictive models [18]. On the other hand, the soft clustering characteristics generated by FCM provide insight into the notion that academic interests are neither static nor dichotomous. A student does not always have to choose sharply between "becoming a software developer" or "becoming an AI researcher," as many students find themselves at the intersection of the two. This situation highlights the importance of a more flexible curriculum approach. One approach is to provide cross-disciplinary courses or collaborative projects that integrate network theory, software engineering, and artificial intelligence techniques in a single, integrated learning environment. With this approach, overlapping interests are no longer seen as obstacles but as valuable intellectual potential for students' academic development..

3.4.1 Interpretation of Interest Grouping Results

The results of the Fuzzy C-Means algorithm application show that student interests are not always rigidly divided into one particular field. Instead, there is overlapping interest that reflects multidimensional nature, for example, a student can have a strong interest in both Artificial Intelligence and Software Engineering. This condition is in line with the concept of soft clustering, which is an advantage of the FCM method compared to hard clustering methods such as K-Means. This flexibility or uncertainty of interests is normal in the academic environment, especially when students begin to determine the topic of their final project which often combines more than one scientific field. Therefore, the representation of interests using membership degrees can provide a more contextual and realistic picture of students' academic preferences [8].

The dominance of Software Engineering and Artificial Intelligence can be explained by two main factors. First, from an industry perspective, the current digital transformation has created high demand for software developers and AI specialists, making these fields more attractive to students who perceive better career prospects. Second, from a pedagogical perspective, courses in Software Engineering and AI often produce tangible outputs such as working applications or predictive models, which provide immediate satisfaction and a sense of achievement. In contrast, Computer Networks, while fundamental, is often perceived as more abstract and infrastructure-oriented, with less visible results, which may explain its lower popularity among students.

The overlapping interests observed in this study reflect the reality of modern computer science education. A student interested in developing an intelligent chatbot, for example, requires knowledge from both AI (natural language processing) and Software Engineering (application development). Similarly, implementing an IoT-based monitoring system combines Computer Networks with AI for data analysis. This overlap is not a weakness but rather a natural consequence of how these disciplines have evolved to complement each other. The FCM algorithm's ability to capture this overlap through fuzzy membership degrees is therefore a strength, as it reflects real-world academic preferences more accurately than hard clustering methods.

The six-cluster structure provides a more granular understanding of student interests than the traditional three-domain classification. For academic practice, this means that study programs can move beyond broad categories like "AI student" and instead recognize specializations such as "AI for prediction and analytics" versus "AI for interactive systems." This distinction is practically useful because these sub-fields require different skill sets, different supervision expertise, and often lead to different types of final projects. For instance, a student in the "AI for prediction" cluster would benefit more from training in regression models and data visualization, while a student in "AI for interactive systems" would need more exposure to user interface design and real-time processing.

The clustering results have three practical applications. First, for topic recommendation, the membership degree matrix can be used to suggest final project topics that align with a student's dominant cluster, or to recommend cross-disciplinary topics for students with overlapping interests. Second, for supervision allocation, study programs can match students with supervisors whose expertise lies in the corresponding sub-specialization, reducing mismatches and improving guidance quality. Third, for curriculum planning, the distribution of students across the six clusters can inform decisions about which

elective courses to offer. For example, if many students fall into the "AI for interactive systems" cluster, the program might consider offering specialized courses in conversational AI or chatbot development.

3.4.2 Recommended Final Project Topics

Based on the membership degree matrix generated by the FCM, a final project topic recommendation scheme can be designed that is tailored to each student's interest profile. For example, students with a dominant membership level in the Artificial Intelligence cluster can be directed to research topics related to machine learning or intelligent data processing. Meanwhile, for students with nearly equal membership values between the AI and Software Engineering fields, topic recommendations can focus on integrating the two fields, for example through the development of artificial intelligence-based applications or predictive systems implemented in web or mobile platforms. This kind of approach not only helps students find a suitable research direction, but also provides an opportunity for supervisors to align guidance with students' actual interests.

3.4.3 Research Limitations and Implications

Despite the promising results, this study has several limitations that should be considered when interpreting the findings. First, the questionnaire data are based on self-reported responses, which may introduce subjective bias in reflecting students' actual research interests. Second, the historical dataset of final project titles is relatively limited in both quantity (only 12 titles) and topic diversity, which may affect the representativeness of the extracted text features. Third, this study focuses solely on the Fuzzy C-Means (FCM) algorithm without comparing it to other clustering methods, such as K-Means or Hierarchical Clustering, which could provide additional insights into clustering performance. Finally, although the results suggest the potential for supporting topic recommendation and supervision planning, the proposed approach has not yet been implemented as a real recommendation system or prototype. Therefore, further development and validation are needed to confirm its practical applicability [9].

4. Conclusion

This study demonstrates that the Fuzzy C-Means (FCM) algorithm is able to systematically group the research interests of final project students in the Computer Science Study Program at the State Islamic University of North Sumatra using a data-driven approach. The data processing stage begins with preprocessing, followed by term weighting using the TF-IDF method to convert textual data into numerical representations suitable for clustering. The evaluation of cluster quality using the Silhouette Index and Davies-Bouldin Index indicates that the most optimal number of clusters is six. The findings suggest that student research interests are multidimensional and not always confined to a single field. Therefore, the FCM method appears to be effective in mapping student interest trends, particularly in capturing overlapping interests across domains. The results also indicate that the cluster mapping can support study programs in developing more targeted academic guidance strategies, ensuring a more balanced distribution of supervisors, and providing a more objective basis for final project topic recommendations. In addition, the clustering results may help study programs identify dominant research trends, which can be used as a consideration in curriculum development and competency alignment with evolving technological needs. However, this study has several limitations. First, the questionnaire data are based on self-reported responses, which may introduce subjective bias. Second, the historical dataset of final project titles is limited to only 12 titles, which may affect the representativeness of the extracted text features. Third, this study focuses solely on the FCM algorithm without comparing it to other clustering methods. For future work, several directions are recommended. First, the proposed approach should be validated using a larger and more diverse dataset, including at least 50 historical final project titles and more than 500 respondents from multiple cohorts or different universities. Second, a comparative study between FCM and other clustering algorithms such as K-Means and Hierarchical Clustering should be conducted to evaluate the relative performance of each method in the context of academic interest mapping. Third, a functional prototype of a recommendation system should be developed and tested with real users to assess its usability and effectiveness in supporting topic selection and supervisor allocation. Finally, future research could explore the use of

more advanced text representation techniques, such as word embeddings (Word2Vec, GloVe) or transformer-based models (BERT), to further improve the quality of interest clustering.

References

[1] A. Damayanti, F. Purwani, and M. Kadafi, "A Content-Based Thesis Supervisor Recommendation System Based on Research Interest Clustering and Cosine Similarity," *JUSIFO (Jurnal Sistem Informasi)*, vol. 11, no. 2, pp. 111–120, Dec. 2025, doi: 10.19109/jusifo.v11i2.27605.

[2] A. Hill, K. Goo, and P. Agarwal, "Recommending the right academic programs: an interest mining approach using BERTopic," *Data Mining and Knowledge Discovery*, vol. 39, no. 20, 2025, doi: 10.1007/s10618-024-01087-y.

[3] J. S. Barrot, "Research on education in Southeast Asia (1996–2019): a bibliometric review," *Educational Review*, vol. 75, no. 2, pp. 348–368, 2023, doi: 10.1080/00131911.2021.1907313.

[4] B. Suprapty and Fariyanti, "Klasifikasi Minat Siswa Untuk Program Studi Jurusan Teknologi Informasi - Politeknik Negeri Samarinda Menggunakan Metode Fuzzy C-Means Clustering," *Jurnal Komputer dan Informatika*, vol. 8, no. 1, pp. 53–62, Mar. 2020, doi: 10.35508/jicon.v8i1.2184.

[5] R. D. Abdika, "Pemetaan Bidang Keilmuan Mahasiswa Dengan Menggunakan Metode Fuzzy C-Means (Studi Kasus: Program Studi Teknik Informatika Universitas Muhammadiyah Gresik)," *Indexia*, vol. 4, no. 2, pp. 28–60, 2022, doi: 10.30587/indexia.v4i2.3639.

[6] I. Rahmatullah, G. S. Nugraha, and A. Aranta, "Feature Selection on Grouping Students Into Lab Specializations for the Final Project Using Fuzzy C-Means," *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, vol. 23, no. 1, pp. 143–154, Nov. 2023, doi: 10.30812/matrik.v23i1.3341.

[7] R. Ghaniy and F. Indriyaningsih, "Penerapan Metode Fuzzy C-Means dalam Pemilihan Program Studi Mahasiswa Baru di Perguruan Tinggi," *Jurnal Teknois*, vol. 10, no. 2, pp. 19–30, 2020, doi: 10.36350/jbs.v10i2.84.

[8] M. B. Arrisalah, M. M. A. Haromainy, and A. Junaidi, "Design of Thesis Topic Recommendation System Using TF-IDF and Cosine Similarity," *bit-Tech*, vol. 8, no. 3, pp. 3553–3564, Apr. 2026, doi: 10.32877/bt.v8i3.3579.

[9] G. Urva and W. Desriyati, "Fuzzy C-Means Algorithm for Grouping Students Based on Preferences and Academic Potential," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 9, no. 1, pp. 366–373, Jan. 2025, doi: 10.33395/sinkron.v9i1.14369.

[10] S. Ali, B. Senapati, N. Mansurova, A. Nikulushkin, N. Muračova, and R. Tsarev, "Fuzzy C-Means: A Fuzzy Approach to Clustering in Educational Contexts," in *Software Engineering: Emerging Trends and Practices in System Development: Proceedings of 14th Computer Science On-line Conference 2025*, vol. 1558, Springer, 2025, pp. 278-289. doi: 10.1007/978-3-032-03406-9_18

[11] D. Rahmayanti, S. Sunardi, and T. Wahyuningrum, "Determine Majors in Vocational High Schools Based on Fuzzy C-Means Algorithm," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 30, no. 1, pp. 20–32, May 2024, doi: 10.21831/jptk.v30i1.59615.

[12] K. Ouassif, B. Ziani, J. Herrera-Tapia, and C. A. Kerrache, "Empowering Education: Leveraging Clustering and Recommendations for Enhanced Student Insights," *Educ. Sci.*, vol. 15, no. 7, p. 819, Jun. 2025, doi: 10.3390/educsci15070819.

[13] R. R. Az-zahra, "Analisis Perbandingan Metode Self Organizing Map Dan Metode Fuzzy C-Means Pada Pengelompokkan Pemintaan Jurusan Di Sekolah Menengah Kejuruan," *Multitek Indonesia: Jurnal Ilmiah*, vol. 16, no. 2, pp. 81–89, 2020, doi: 10.24269/mtkind.v16i2.5603

[14] A. M. Khafid, A. Alimudin, and A. Suyanto, "Enhancing K-Means Clustering for Journal Articles

Using TF-IDF and LDA Feature Extraction," *Brilliance: Research of Artificial Intelligence*, vol. 4, no. 2, pp. 592–601, Dec. 2024, doi: 10.47709/brilliance.v4i2.5547.

[15] C. Wongoutong, "The Impact of Neglecting Feature Scaling in K-Means Clustering," *PLOS ONE*, vol. 19, no. 12, p. e0310839, Dec. 2024, doi: 10.1371/journal.pone.0310839.

[16] D. A. Suryaningrum, R. Syaifudin, and H. R. P. Putra, "Integrasi Word Embeddings Dan Inverse Book Frequency Dalam Pembobotan Term Untuk Peningkatan Pencarian Dokumen," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 4, pp. 2529–2537, 2024, doi: 10.29100/jipi.v9i4.7557.

[17] E. W. Pratomo and E. Utami, "Hybrid TF-IDF and Embedding Model for Improving Similarity and Clustering Accuracy," *JOINTECS (Journal of Information Technology and Computer Science)*, vol. 10, no. 1, pp. 33–40, 2025, doi: 10.31328/jointecs.v10i1.7344.

[18] K. L. C. Tuluswati and D. Trisnawarman, "Hybrid Clustering-Classification Untuk Personalisasi Rekomendasi Unit Kegiatan Mahasiswa Baru," *Jurnal Ilmu Komputer dan Sistem Informasi*, vol. 14, no. 1, 2026, doi: 10.24912/664mgm62.