JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 June 2026
Page: 393-400
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1026

Structural Classification of Indonesian Arithmetic Word Problems Using Hierarchical Agglomerative Clustering

Mochammad Dhani Aprianto[1]*, Agung Prasetya[2]

[1,2]Informatics, Faculty of Science and Technology, Universitas Bhinneka PGRI Tulungagung, Tulungagung, Indonesia
Email: [1]daniaprianto84@gmail.com, [2]*agung@ubhi.ac.id
*(*Email Corresponding Author: agung@ubhi.ac.id)*
Received: May 9, 2026 | Revision: May 13, 2026 | Accepted: May 16, 2026

Abstract

Arithmetic word problems (MWP) are a fundamental component of elementary mathematics education that integrate linguistic comprehension with quantitative reasoning. In practice, collections of MWPs are commonly organized based on teacher intuition or broad curriculum categories, which are inherently subjective and often fail to reflect the true mathematical similarity between problems. This study aims to classify Indonesian arithmetic word problems based on their underlying relational structures using Hierarchical Agglomerative Clustering (HAC). The dataset consists of 897 elementary-level arithmetic word problems represented through 143 binary features encoding five relational dimensions, namely combine, change, compare, equal groups, and fair division. Hamming Distance is employed as the dissimilarity metric, and clustering is performed using the complete linkage method. The optimal number of clusters is determined using three internal validity indices: the Calinski–Harabasz Index, Silhouette Score, and Davies–Bouldin Index. Although statistical indices favor smaller cluster configurations, four clusters are selected as the optimal number based on domain-specific interpretability, as they align with established theoretical categories of arithmetic relational structures. This approach effectively identifies latent structural patterns within the dataset and demonstrates the potential of feature-based binary representation combined with HAC for systematic MWP classification. The findings offer practical support for adaptive problem bank development, automated curriculum analysis, and intelligent tutoring system design.

Keywords **:** Arithmetic Word Problems, Hierarchical Agglomerative Clustering, Hamming Distance, Educational Data Mining, Cluster Validity Index

1. INTRODUCTION

Mathematics education at the elementary school level plays a foundational role in developing students' quantitative reasoning and problem-solving abilities. Among the various forms of mathematical tasks used in elementary education, arithmetic word problems (MWP) occupy a particularly important position, as they integrate linguistic comprehension with numerical reasoning. Unlike purely computational exercises, word problems require students to interpret contextual narratives, identify relevant quantities, and construct appropriate mathematical models to arrive at a solution [1]. This dual demand on both language and mathematics makes word problems a critical yet complex component of the elementary mathematics curriculum.

In practice, collections of arithmetic word problems used in Indonesian elementary schools are typically organized based on teacher intuition or broad curriculum categories such as addition, subtraction, multiplication, and division. While this approach is practical, it is inherently subjective and may not reflect the deeper relational structures that underlie different problem types. Research in mathematics education has established that arithmetic word problems can be systematically categorized based on their underlying mathematical relations, including combine (part–whole relationships), change (quantitative transformation over time), compare (relationships between quantities), equal groups (repeated identical units), and fair division (equal distribution) [1]. These structural categories reflect stable relational patterns that persist across variations in surface-level linguistic expression and narrative context.

The growing availability of digital problem banks and learning management systems has created new opportunities for the systematic analysis of large collections of educational content. Educational Data Mining (EDM) has emerged as a powerful methodological framework for extracting meaningful patterns from educational datasets using computational techniques [2]. Within the EDM paradigm, clustering-based approaches have been widely applied to identify latent groupings in student performance data, course content, and learning behaviors [3]. However, despite the significant potential of clustering methods for the structural classification of mathematical problem collections, this line of research remains underexplored, particularly for Indonesian-language datasets.

Several related studies have investigated the application of clustering and data mining techniques in educational contexts. Ersozlu et al. reviewed machine learning methods applied to educational data, establishing the theoretical basis for applying unsupervised learning techniques to educational datasets, which directly informs the methodological approach of the present study [2]. Aldisa applied Agglomerative Hierarchical Clustering (AHC) to determine student major placement, demonstrating the practical utility of HAC in Indonesian educational data mining contexts [4]. Andre

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 June 2026
Page: 393-400
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1026

et al. extended this line of work by applying clustering approaches to classify undergraduate student thesis topics using educational data mining, highlighting the effectiveness of clustering in organizing academic content at scale [5]. Zhang et al. conducted a systematic review of EDM techniques for student performance prediction, providing a comprehensive overview of machine learning methods applicable to educational datasets and emphasizing the importance of unsupervised approaches for exploratory analysis [3]. Murtagh and Contreras provided a comprehensive algorithmic overview of hierarchical clustering methods, including agglomerative strategies and linkage criteria, which serves as a key technical reference for the HAC implementation in this study [6].

Despite the contributions of these related works, a significant research gap remains. Existing studies on arithmetic word problems predominantly focus on automatic problem solving or equation generation rather than structural classification [1]. Furthermore, research on the systematic classification of MWPs based on their relational structures is scarce, particularly for Indonesian-language datasets. Most clustering-based EDM studies target student performance or behavioral data rather than the intrinsic structural properties of problem content itself [3], [5]. Additionally, prior work has not explored the use of binary feature representation combined with Hamming Distance-based HAC as a methodological framework for MWP classification [6]. These gaps collectively motivate the present study.

To address these limitations, this study proposes a feature-based relational representation framework combined with Hierarchical Agglomerative Clustering to systematically classify Indonesian arithmetic word problems. Each problem is encoded as a binary vector of 143 features capturing relational dimensions such as combine, change, compare, equal groups, and fair division. Hamming Distance is employed as the dissimilarity metric due to its suitability for binary feature spaces [7], and complete linkage is adopted as the clustering strategy to produce compact and well-separated clusters [8]. The optimal number of clusters is determined through a combination of three internal cluster validity indices: the Calinski–Harabasz Index, Silhouette Score, and Davies–Bouldin Index [9], [10].

The primary objective of this study is to analyze the effectiveness of HAC in identifying latent structural patterns within a dataset of 897 Indonesian arithmetic word problems. The expected outcomes include the identification of meaningful cluster configurations that align with established theoretical categories in mathematics education, as well as the development of a replicable and scalable classification framework applicable to adaptive problem bank systems and curriculum analysis. By bridging the fields of educational data mining and mathematics education research, this study aims to contribute both methodological and practical insights to the systematic organization of arithmetic problem collections in Indonesian elementary education.

2. RESEARCH METHODOLOGY

2.1 Research Design

This study adopts a quantitative exploratory approach to identify latent clustering structures within a collection of Indonesian arithmetic word problems (MWP). The method employed is Hierarchical Agglomerative Clustering (HAC), which belongs to the unsupervised learning paradigm, as it does not rely on predefined class labels [11]. This approach is particularly suitable for exploratory structural analysis where the number of natural groupings within the data is not known in advance, making it an effective tool for discovering hidden structural patterns in educational datasets [11]. The overall research process flow is illustrated in Figure 1.

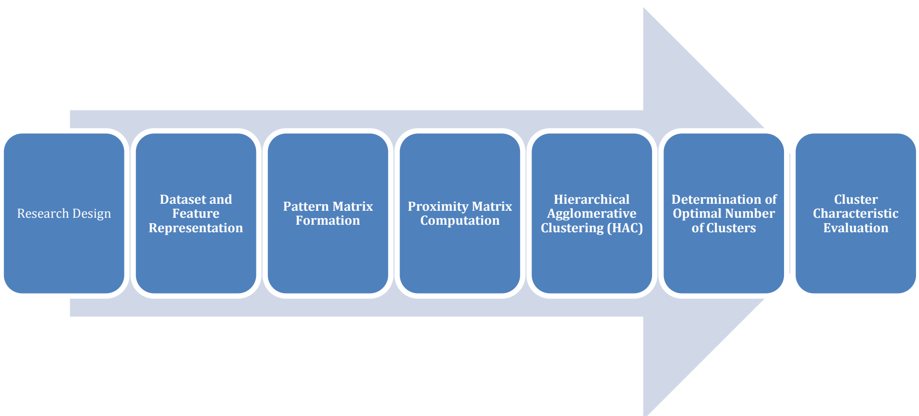

Figure 1. Research Process Flowchart

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 June 2026
Page: 393-400
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1026

2.2 Dataset and Feature Representation

2.2.1 Research Object

The objects of this study are arithmetic word problems at the elementary school level [1]. Each problem must satisfy the following criteria: (1) contains one or two basic arithmetic operations, namely addition, subtraction, multiplication, or division; (2) does not involve algebraic symbols; (3) can be modeled using simple arithmetic equations; and (4) is written in the Indonesian language [1]. These criteria ensure that the selected problems reflect a consistent and well-defined scope of elementary arithmetic reasoning, enabling reliable feature extraction and structured clustering analysis.

Each word problem is treated as an individual object x_i. If the total number of problems is n, the dataset is formally defined as:

$$X = \{x_1, x_2, x_3, \ldots, x_n\} \tag{1}$$

To illustrate the structure and characteristics of the dataset, several representative examples are presented in Table 1.

Table 1. Sample of Arithmetic Word Problems

Word Problem
Ani has 5 apples and buys 3 more apples. How many apples does she have now?
Budi had 10 candies and gave 4 to his friend. How many candies remain?
There are 6 bags with 5 oranges each. How many oranges are there in total?

2.2.2 Relational Feature Dimensions

Each problem is represented using a set of binary features that capture its underlying mathematical relational structure [12]. The main relational dimensions include: (1) *Combine*, representing part–whole relationships; (2) *Change*, representing quantitative transformation over time; (3) *Compare*, representing the difference between quantities; (4) *Equal Groups*, representing repetition of identical units; and (5) *Fair Division*, representing equal distribution across entities.

Each dimension consists of several sub-features, including the presence of an initial quantity, presence of a final quantity, explicit change indicator, difference between entities, repetition of identical units, and distribution across recipients. If the total number of features is m, then each problem object is represented as a binary vector:

$$x_i = (f_{i1}, f_{i2}, f_{i3}, \ldots, f_{im}),\ f_{ij} \in \{0, 1\} \tag{2}$$

where a value of 1 indicates the presence of a feature, and 0 indicates its absence.

2.3 Pattern Matrix Formation

The entire dataset is represented as a pattern matrix of dimension $n \times m$, where each row represents a word problem and each column corresponds to a specific binary feature[13]. The pattern matrix is formally defined as:

$$X = [f_{ij}]_{nxm},\ \text{where } f_{ij} \in \{0, 1\} \tag{3}$$

Since all features are binary in nature, the dataset resides in a nominal feature space. This representation enables a structured and quantitative analysis of heterogeneous problem types across the dataset, ensuring that structural similarities between problems are captured objectively rather than relying on subjective manual categorization [13].

2.4 Proximity Matrix Computation

To perform clustering, a distance measure between objects is required to quantify dissimilarity. Given the binary nature of all features in this study, the Minkowski distance with parameter $p = 1$ is used, which is mathematically equivalent to the Hamming Distance [7]. The Hamming Distance is particularly well-suited for binary and categorical feature spaces, as it directly counts the number of feature disagreements between any two binary vectors without being affected by feature scaling issues that commonly arise with continuous distance metrics such as Euclidean distance [7]. The distance between two objects x_i and x_j is defined as:

$$d(x_i, x_j) = \Sigma_{k=1}^{m} |f_{ik} - f_{jk}| \tag{4}$$

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 June 2026
Page: 393-400
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1026

This distance measure satisfies the following fundamental properties: (1) non-negativity, where $d(x_i, x_j) \geq 0$; (2) identity, where $d(x_i, x_j) = 0$ if and only if $x_i = x_j$; and (3) symmetry, where $d(x_i, x_j) = d(x_j, x_i)$. All pairwise distances between the *n* objects are compiled into a proximity matrix defined as:

$$D = [d(x_i, x_j)]_{nxn} \tag{5}$$

2.5 Hierarchical Agglomerative Clustering (HAC)
2.5.1 Formal Definition

Hierarchical Agglomerative Clustering is a bottom-up clustering method that iteratively merges the closest pair of clusters until a single cluster remains[14]. Let $X = \{x_1, x_2, \dots, x_n\}$ be a set of objects. A clustering partition divides X into *k* clusters $C = \{C_1, C_2, \dots, C_k\}$, such that [14]:

a) $C_i \cap C_j = \emptyset$, for $i \neq j$ (non-overlapping)
b) $\cup_{i=1}^{k} C_i = X$ (complete coverage)

The HAC algorithm proceeds as follows: initially, each object forms its own singleton cluster; subsequently, the two closest clusters are identified and merged based on the defined distance metric; this iterative merging process continues until all objects belong to a single cluster [6]. The resulting hierarchical structure is represented as a dendrogram, which enables flexible cluster extraction at different levels of granularity.

2.5.2 Linkage Method

This study employs the complete linkage method, where the distance between two clusters C_a and C_b is defined as the maximum pairwise distance between their respective members [8]:

$$d(C_a, C\_b) = \max \{ d(x_i, x_j) \mid x_i \in C_a, x_j \in C\_b \} \tag{6}$$

Complete linkage is selected because it tends to produce compact and well-separated clusters by considering the worst-case distance between cluster members [8]. This property is particularly desirable in the context of arithmetic word problem classification, where clear structural boundaries between problem categories are expected based on their relational dimensions.

2.6 Determination of Optimal Number of Clusters

To determine the optimal number of clusters, three complementary cluster validity indices (CVI) are employed [9], [10]. The use of multiple indices is recommended in the literature to ensure robust and reliable cluster selection, as different indices may favor different cluster configurations depending on the data structure [10].
The primary index used is the Calinski–Harabasz (CH) Index, defined as:

$$CH(k) = [SSB / (k - 1)] / [SSW / (n - k)] \tag{7}$$

where *n* is the total number of objects, *k* is the number of clusters, SSB is the between-cluster variance, and SSW is the within-cluster variance. A higher CH Index value indicates better-defined clusters, and the optimal number of clusters corresponds to the configuration that maximizes this value [10].
As additional validation indices, the Silhouette Score and Davies–Bouldin Index (DBI) are also computed [9], [15]. The Silhouette Score measures how well each object fits within its assigned cluster relative to neighboring clusters, with values ranging from −1 to +1 where higher values indicate better assignment [9]. The Davies–Bouldin Index measures the average similarity between each cluster and its most similar counterpart, where lower values indicate better cluster separation [15]. The combined use of these three indices provides a comprehensive and multi-perspective evaluation of clustering quality [9].

2.7 Cluster Characteristic Evaluation

To interpret the characteristics of each resulting cluster, two complementary measures are applied following the approach adopted in related clustering-based EDM studies [4].

2.7.1 Proportion (P)

Proportion measures the dominance of a specific feature within a given cluster [4]. It is defined as the ratio of the number of objects in cluster C_c that possess feature *j* to the total number of objects in that cluster:

$$P_{jc} = \Sigma\_\{x_i \in C_c\}\ f_{ij} / |C_c| \tag{8}$$

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 June 2026
Page: 393-400
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1026

A higher proportion value indicates that the feature is more dominant and characteristic within the cluster.

2.7.2 Significance (S)

Significance is defined using a probabilistic approach that reflects the conditional likelihood of a feature appearing within a specific cluster [4]:

$$S_{jc} = P(f_j \mid C_c) \tag{9}$$

This measure indicates how strongly a feature is associated with a particular cluster, enabling meaningful differentiation between clusters based on their dominant relational structures. Together, Proportion and Significance provide a comprehensive framework for the interpretation of cluster characteristics in the context of arithmetic word problem classification [5].

3. RESULTS AND DISCUSSION

3.1 Dataset Characteristics

The dataset used in this study consists of 897 Indonesian arithmetic word problems collected at the elementary school level. Each problem is represented using 143 binary features that encode relational structures and contextual attributes based on five main relational dimensions: combine, change, compare, equal groups, and fair division. Formally, the dataset is represented as a binary matrix:

$$X \in \{0, 1\}^{897x143} \tag{10}$$

Each row corresponds to an individual word problem, while each column represents a specific relational feature. Due to the binary nature of the data, the feature space is nominal and particularly suitable for distance-based clustering using the Hamming Distance metric. This representation enables a structured and quantitative analysis of heterogeneous problem types across the dataset, ensuring that structural similarities between problems can be captured objectively rather than relying on subjective manual categorization.

The dataset exhibits a diverse distribution of relational structures, reflecting the variety of arithmetic problem types commonly found in Indonesian elementary school curricula. This diversity is essential for validating the effectiveness of the clustering approach in distinguishing structurally distinct problem categories.

3.2 Proximity Matrix Construction

Based on the pattern matrix $X \in \{0, 1\}^{897x143}$, pairwise distances between all 897 objects are computed using the Hamming Distance metric as defined in the methodology. The resulting proximity matrix is defined as:

$$D \in \mathbb{R}^{897x897} \tag{11}$$

This matrix satisfies the three fundamental properties required for a valid distance measure: non-negativity, symmetry, and a zero diagonal. A smaller distance value between two problems indicates a higher degree of structural similarity in terms of their relational features, while a larger value indicates greater structural dissimilarity.

The proximity matrix serves as the foundation for the HAC algorithm, guiding the iterative merging process based on complete linkage. Given the high dimensionality of the feature space (143 features), the Hamming Distance proves to be an effective and computationally appropriate metric, as it directly measures the number of feature disagreements between any two binary vectors without being sensitive to feature scaling issues that affect continuous distance metrics such as Euclidean distance.

3.3 Determination of Optimal Number of Clusters

To determine the optimal number of clusters, experiments were systematically conducted by varying the number of clusters *k* from 2 to 6. Three cluster validity indices were employed to evaluate each configuration: the Calinski–

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 June 2026
Page: 393-400
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1026

Harabasz (CH) Index, the Silhouette Score, and the Davies–Bouldin Index (DBI). The complete evaluation results are presented in Table 2.

Table 2. Cluster Evaluation Results [Source: Research Data]

k	CH Index	Silhouette Score	DBI
2	164.48	0.2632	2.2919
3	125.99	0.2680	2.4295
4	99.31	0.2576	2.3484
5	85.29	0.2184	2.5902
6	74.58	0.2150	2.5071

The evaluation results presented in Table 2 reveal distinct patterns across the three indices. The Calinski–Harabasz Index reaches its maximum value at k = 2 (CH = 164.48), and decreases consistently as k increases, suggesting that a two-cluster solution provides the highest between-cluster to within-cluster variance ratio from a purely statistical standpoint. The Silhouette Score achieves its highest value at k = 3 (Score = 0.2680), indicating that a three-cluster configuration yields the best average object-to-cluster assignment quality. Meanwhile, the Davies–Bouldin Index is lowest at k = 2 (DBI = 2.2919), further reinforcing the statistical preference for a smaller number of clusters.

However, it is important to note that statistical indices alone do not always yield the most practically meaningful clustering solution, particularly in domain-specific applications such as educational data mining. While smaller cluster configurations demonstrate superior numerical performance, the selection of k = 4 is justified based on domain knowledge and interpretability considerations. The four-cluster solution aligns directly with well-established theoretical categories in arithmetic problem research, namely: combine, change, compare, and equal groups/fair division. This alignment ensures that the resulting clusters are not only statistically valid but also interpretable and actionable in real educational contexts.

This finding highlights a critical methodological insight: in educational data mining, the integration of domain expertise into the cluster selection process is as important as quantitative validation. Relying solely on statistical indices may lead to oversimplified representations that fail to capture the nuanced structural diversity of arithmetic word problems.

3.4 Hierarchical Structure and Dendrogram Analysis

The application of Hierarchical Agglomerative Clustering with complete linkage produces a hierarchical structure represented as a dendrogram.
At the initial level, the dataset is divided into two major groups based on dominant relational differences. As the clustering process progresses, finer-grained structures emerge.
By applying an appropriate cut-off threshold to the dendrogram, four stable and well-separated clusters are obtained. This hierarchical representation allows flexible interpretation at different levels of granularity.

3.5 Cluster Characteristics

Cluster 1 predominantly exhibits combine structure features, where the majority of combine-related features are active while no explicit state changes or comparison relationships are present. This cluster represents problems with a static part–whole structure, typically involving summation without temporal transitions. Problems in this cluster are characterized by a straightforward aggregation pattern, where two or more quantities are joined to form a whole, without any indication of change over time or relational comparison between entities.

Cluster 2 is dominated by change structure features, characterized by the presence of both initial and final quantities along with explicit change indicators. Problems in this cluster follow a temporal transformation pattern, where a quantity undergoes a transition from an initial state to a final state, commonly solved through addition or subtraction. The explicit presence of change indicators distinguishes this cluster from the combine cluster, as the problems inherently describe a dynamic process rather than a static aggregation.

Cluster 3 is characterized by active difference-related features and involves direct comparison between two or more quantities without necessarily focusing on total aggregation. Problems in this cluster emphasize relational differences, where the primary mathematical operation involves determining how much more or less one quantity is

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 June 2026
Page: 393-400
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1026

relative to another. This comparative structure makes Cluster 3 structurally distinct from both the combine and change clusters, as the focus lies on the relationship between quantities rather than their combination or transformation.

Cluster 4 is defined by the presence of repeated identical units and a distribution-oriented structure, typically solved through multiplication or division. Problems in this cluster are based on grouping and equal distribution patterns, where a quantity is either repeatedly added across identical groups or evenly distributed among a set of recipients. The structural dominance of equal groups and fair division features in this cluster reflects a fundamentally different relational pattern compared to the other three clusters, as the problems inherently involve proportional reasoning rather than simple aggregation, transformation, or comparison.

3.6 Cluster Distribution Analysis

The distribution of problems across the four clusters is relatively balanced, indicating that the dataset does not exhibit extreme dominance by any single relational structure. This balance is a positive indicator of the quality and diversity of the dataset, as it suggests that the 897 collected problems represent a broad and varied range of arithmetic relational structures commonly found in Indonesian elementary school curricula. A highly imbalanced distribution would have indicated either a bias in the problem collection process or an inadequate representation of certain relational categories, both of which would have limited the generalizability of the clustering results.

Furthermore, the successful separation between clusters demonstrates that the selected 143 binary features and the Hamming Distance metric are effective in distinguishing structurally distinct problem types. The clear inter-cluster boundaries confirm that the feature representation framework is capable of capturing the underlying mathematical differences between combine, change, compare, and equal groups/fair division problem categories in a quantitatively meaningful way. This finding supports the validity of the proposed classification framework as a reliable and scalable approach for the structural organization of arithmetic word problem collections in educational data mining applications.
.

4. CONCLUSION

This study successfully classified 897 Indonesian elementary arithmetic word problems based on their underlying mathematical relational structures using Hierarchical Agglomerative Clustering (HAC) with complete linkage and Hamming Distance as the dissimilarity metric. Each problem was represented through 143 binary features capturing five relational dimensions, namely combine, change, compare, equal groups, and fair division, enabling a structured and quantitative analysis of structural similarities across the dataset. Although internal validity metrics — including the Calinski–Harabasz Index, Silhouette Score, and Davies–Bouldin Index — statistically favored smaller cluster configurations, four clusters were selected as the optimal number based on domain-specific interpretability, as the resulting clusters align meaningfully with established theoretical categories in arithmetic problem research. These findings demonstrate that feature-based relational representation combined with HAC is effective in identifying latent structural patterns in arithmetic word problems, providing a systematic and replicable classification framework with practical implications for adaptive problem bank development, automated curriculum analysis, and intelligent tutoring system design. However, this study has several limitations, as the feature representation relies on manual annotation which may introduce subjectivity, and the evaluation is confined to internal clustering metrics without external expert validation. Future research may therefore explore automated feature extraction using natural language processing techniques, incorporate expert-based evaluation for external validation, and extend the approach to more complex multi-step or algebraic word problems.

REFERENCES

[1] S. Acharya, R. Basak, and S. Mandal, “Solving Arithmetic Word Problems Using Natural Language Processing and Rule-Based Classification,” *Int. J. Intell. Syst. Appl. Eng.*, vol. 10, no. 1, pp. 87–97, Mar. 2022, doi: 10.18201/ijisae.2022.271.

[2] Z. Ersozlu, S. Taheri, and I. Koch, “A review of machine learning methods used for educational data,” *Educ. Inf. Technol.*, vol. 29, no. 16, pp. 22125–22145, Nov. 2024, doi: 10.1007/s10639-024-12704-0.

[3] Y. Zhang, Y. Yun, R. An, J. Cui, H. Dai, and X. Shang, “Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis,” *Front. Psychol.*, vol. 12, Dec. 2021, doi: 10.3389/fpsyg.2021.698490.

[4] R. T. Aldisa, “Data Mining Penentuan Jurusan Siswa Menggunakan Metode Agglomerative Hierarchical Clustering (AHC,” *J. MEDIA Inform. BUDIDARMA*, vol. 7, no. 2, p. 873, Apr. 2023, doi: 10.30865/mib.v7i2.6092.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 June 2026
Page: 393-400
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1026

[5] Andre, N. Suciati, H. Fabroyir, and E. Pardede, "Educational Data Mining Clustering Approach: Case Study of Undergraduate Student Thesis Topic," *IEEE Access*, vol. 11, pp. 130072–130088, 2023, doi: 10.1109/ACCESS.2023.3332818.

[6] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *WIREs Data Min. Knowl. Discov.*, vol. 2, no. 1, pp. 86–97, Jan. 2012, doi: 10.1002/widm.53.

[7] R. Argiento, E. Filippi-Mazzola, and L. Paci, "Model-Based Clustering of Categorical Data Based on the Hamming Distance," *J. Am. Stat. Assoc.*, vol. 120, no. 550, pp. 1178–1188, Apr. 2025, doi: 10.1080/01621459.2024.2402568.

[8] T. Märzinger, J. Kotík, and C. Pfeifer, "Application of Hierarchical Agglomerative Clustering (HAC) for Systemic Classification of Pop-Up Housing (PUH) Environments," *Appl. Sci.*, vol. 11, no. 23, p. 11122, Nov. 2021, doi: 10.3390/app112311122.

[9] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Cluster validity indices for automatic clustering: A comprehensive review," *Heliyon*, vol. 11, no. 2, p. e41953, Jan. 2025, doi: 10.1016/j.heliyon.2025.e41953.

[10] B. A. Hassan, N. B. Tayfor, A. A. Hassan, A. M. Ahmed, T. A. Rashid, and N. N. Abdalla, "From A-to-Z review of clustering validation indices," *Neurocomputing*, vol. 601, p. 128198, Oct. 2024, doi: 10.1016/j.neucom.2024.128198.

[11] A. A. Wani, "Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions," *PeerJ Comput. Sci.*, vol. 10, p. e2286, Aug. 2024, doi: 10.7717/peerj-cs.2286.

[12] *Pattern Recognition and Machine Learning*. Springer New York, 2006. doi: 10.1007/978-0-387-45528-0.

[13] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, Jun. 2015, doi: 10.1007/s40745-015-0040-1.

[14] E. K. Tokuda, C. H. Comin, and L. da F. Costa, "Revisiting agglomerative clustering," *Phys. A Stat. Mech. its Appl.*, vol. 585, p. 126433, Jan. 2022, doi: 10.1016/j.physa.2021.126433.

[15] F. Ros, R. Riad, and S. Guillaume, "PDBI: A partitioning Davies-Bouldin index for clustering evaluation," *Neurocomputing*, vol. 528, pp. 178–199, Apr. 2023, doi: 10.1016/j.neucom.2023.01.043.