Analysis of a Novel Hybrid Fine Tuned Weighted Harmonic Mean for Efficient Plagiarism Detection using Particle Swarm

Madhan N[1], Dheva Rajan S[2], Madhuri Jain[3]

[1]University of Technology and Applied Sciences Al Musannah, Sultanate of Oman, madhan.narayanan@utas.edu.om,
[2]University of Technology and Applied Sciences Al Musannah, Sultanate of Oman, dheva@act.edu.om
[3]Banasthali Vidyapith, Rajasthan, India, madhuridayalbagh@gmail.com

Many types of resemblance across lexical and semantic levels are sometimes difficult for current plagiarism detection algorithms to detect. To overcome this drawback, this paper suggests a brand-new Weighted Harmonic Mean model that incorporates Hamming, Cosine, and Jaccard similarity scores. The suggested model makes use of the harmonic means' sensitivity to low scores to emphasize suspicious situations and accentuate small differences. Furthermore, Particle Swarm Optimization is suggested and presented as an effective way to tune weights and enhance performance. It is obtained that the proposed model performs better than the other stated methods.

Keywords: Weighted harmonic mean, PSO, Jaccard, Hamming, Cosine, plagiarism.

1. Introduction

Jaccard distance [1] abbreviated as JD focuses on shared words between two texts, useful for identifying verbatim copying. Assume X1 and X2 be the finite sample sets. The Jaccard coefficient is given by

$$\mathrm{J_c}(\mathrm{X1}, \mathrm{X2}) = \frac{|\mathrm{X1} \cap \mathrm{X2}|}{|\mathrm{X1} \cup \mathrm{X2}|} = \frac{|\mathrm{X1} \cap \mathrm{X2}|}{|\mathrm{X1}| + |\mathrm{X2}| - |\mathrm{X1} \cap \mathrm{X2}|},$$

$$0 \le \mathrm{J_c}(\mathrm{X1}, \mathrm{X2}) \le 1$$

The Jaccard similarity (JS) is given by

$$\mathrm{J}(\mathrm{X1}, \mathrm{X2}) = 1 - \mathrm{J_c}(\mathrm{X1}, \mathrm{X2})$$

$$= \frac{|X1 \cup X2| - |X1 \cap X2|}{|X1 \cup X2|}$$

The Hamming distance (HD) takes mismatched characters into account, which is useful for identifying little changes. The hamming distance between two strings or vectors of equal length is the number of points at which the matching symbols differ. JD counts the minimum number of substitutions required to change a string to another, or the minimum number of errors that might have led to the change. Hamming similarity (HS) is defined as 1 minus the normalized HD.

Cosine Similarity (CS) (as mentioned in Tan et al., 2018) captures semantic similarity, valuable for identifying paraphrased plagiarism.

The CS is defined as

$$\mathbf{CS = C(X1, X2) = \cos(\theta) = \frac{X1 \cdot X2}{\| X1 \| \| X2 \|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \cdot \sqrt{\sum_{i=1}^{n} Y_i^2}}}$$

A metric for comparing two non-zero vectors defined in an inner product space is termed CS. Consequently, rather of relying on the magnitudes of the vectors, the CS just needs to know their angle. The CS always lies in the interval [-1,1]. The CS is constrained in [0,1] if the vectors' component values cannot be negative. Combining these might help the model detect plagiarism at various similarity levels and increase its overall accuracy. Although the JD and HD have obvious mathematical meanings when taken separately, CD is more difficult and less intuitive to comprehend their products together.

2. Background

An alternative viewpoint of assessing similarity may be predicated on the ratios of disputes. The ratio of shared tokens to all tokens in the union of two texts is represented by the JD. The percentage of mismatched tokens to all tokens in each text is represented by the HD. The harmonic mean (HM) may be helpful in detecting plagiarism because HM is applied in other fields like information retrieval and text mining. The idea of employing the HM for plagiarism detection has been investigated in a few research publications, even if the precise application of the weighted HMs (WHM) for plagiarism detection using JS, HS, and CS is not yet well documented. In 1988, Stout et al made an attempt at plagiarism detection, but that effort was create expert systems that solve issues by first suggesting answers and then refining them. Paper by Adams et al., (2015) advocates adopting the HM in mindless technology but not towards using JS, HS and CS .more focused on developing a computer programming language that provides a basis for learning in order to

In a noteworthy study by Sağlam et al., (2022), they discussed several measures, but they did not include the HM ; instead, they talked about the program JPlag. The HM approach was suggested by Roul & Sahoo (2022) for automated text summarization, but not for plagiarism. Takano & Omori, (2018) first suggested utilizing the HM to identify plagiarism, but they eventually went on to k means clustering. Even though Thamotharan et al., (2023) explore text descriptions and random noise, they do so by employing HM in generative adversarial networks. A methodology that integrates semantic similarity metrics like CS with citation analysis is presented in a publication by Sharma et al., (2018). Although the authors do not specifically use the HM, they stress the need of weighing the contributions of many measurements to find both obvious and subtle plagiarism cases.

The particular combination of JS, HS, and CS utilizing the WHM is not addressed in these publications, they

have still offered insightful information about the possible advantages of doing so. Therefore, it is suggested that these three approaches be combined with the WHM in this proposed method.

3. Proposed Model

I. Proposed model and applicability

Theorem: (Proposed model formula and applicability) The proposed model$\mathrm{S(T1, T2)} = \frac{1}{\mathrm{H}(\mathrm{w_J * J, w_H * H, w_C * C})}$, correctly implements the DRS formula for combining multiple similarity scores Jaccard, Hamming, Cosine simply J,H and C respectively with corresponding weights $\mathrm{w_J, w_H,}$ and $\mathrm{w_C}$, where $\mathrm{w_i} > 0$ and $\sum \mathrm{w_i} = 1,$ T1 and T2 are two documents to compare

Proof:

The weighted harmonic means of a set of non-negative numbers $\mathrm{x_1, x_2, \ldots, x_n}$ with corresponding weights $\mathrm{w_1, w_2, \ldots, w_n}$ (where $\mathrm{w_i} > 0$ and $\sum \mathrm{w_i} = 1$) is defined as:

$$\mathrm{H} = \frac{\mathrm{n}}{\sum\left(\frac{\mathrm{w_i}}{\mathrm{x_i}}\right)}$$

The proposed model's formula for combining similarity scores J, H, and C with weights $\mathrm{w_J, w_H}$, and $\mathrm{w_C}$ is:

$$S(\mathrm{T1, T2}) = \frac{3}{\left(\frac{1}{\mathrm{w_J * J(T1, T2)}} + \frac{1}{\mathrm{w_H * H(T1, T2)}} + \frac{1}{\mathrm{w_C * C(T1, T2)}}\right)}$$

$$\mathrm{S(T1, T2)} = \frac{1}{\left(\left(\frac{1}{3}\right) * \left(\frac{1}{\mathrm{w_J * J(T1, T2)}} + \frac{1}{\mathrm{w_H * H(T1, T2)}} + \frac{1}{\mathrm{w_C * C(T1, T2)}}\right)\right)}$$

The expression inside the parentheses matches the WHM formula for three scores (J, H, C) with weights $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$:

$\mathrm{S(T1, T2)} = \frac{1}{\mathrm{H(J,H,C)}}$ where weights are $\mathrm{w_J} = \mathrm{w_H} = \mathrm{w_C} = \frac{1}{3}$.

Reintroducing the original weights $\mathrm{w_J, w_H, w_C}$, we get:

$\mathrm{S(T1,T2)} = \frac{1}{\mathrm{H(w_J * J, w_H * H, w_C * C)}}$, where $\mathrm{w_J + w_H + w_C = 1}$,

The constraint $\mathrm{w_J + w_H + w_C = 1}$ defines a plane within this hypercube, where valid scores reside.

This final expression demonstrates that the model's formula correctly implements the WHM for combining the scores J, H, and C with their respective weights $\mathrm{w_J, w_H}$, and $\mathrm{w_C}$. Therefore, the theorem is proven. Furthermore, as HM is true for any $\mathrm{x_i, i = 1,2,3, \dots, n}$, J, H, and C can compare many documents, this theorem is also true for n number of documents without loss of generality. Sorry for the dual usage of J, H, and C with JS, HS and CS.

II. Domain of the DRS model

The domain of J is [0, 1], as it measures the proportion of shared elements between two sets, where 0 indicates no overlap and 1 indicates complete overlap. If defined as 1 minus the normalized HD, H also lies in [0, 1]. Therefore, J and H: $0 \le \mathrm{J, H} \le 1$

Domain of C is [-1, 1], as it measures the cosine of the angle between two vectors, where -1 indicates opposite directions, 0 indicates orthogonality, and 1 indicates perfect alignment. Now, HM $\mathrm{(w_J, w_H, w_C)} \le$ arithmetic mean $\mathrm{(w_J, w_H, w_C)}$, if $\mathrm{(w_J, w_H, w_C)} > 0$. Hence, C: $-1 \le \mathrm{C} \le 1$, but its contribution to the WHM is bounded by $\frac{1}{\mathrm{w_C}}$, ensuring a non-negative denominator. Therefore, the domain of the DRS is [0, 1], if weights $\mathrm{(w_J, w_H, w_C)} > 0$ and $\sum \mathrm{w_i} = 1$.

III. Range of the DRS model

Given that the denominator and $\mathrm{(w_J, w_H, w_C)} > 0$, the DRS ≥ 0. When $\mathrm{w_J = w_H = w_C =}$ 1, the highest value of S may be found, which yields $\mathrm{S = 1}$. But since a finite result of DRS requires that at least one of J, H, and C be non-zero, the DRS can approach but never reach 0. Consequently, the DRS's range is (0, 1].

The DRS formula is used to project this point geometrically into the plane, representing a weighted balance of the scores. The resulting DRS similarity is an advanced proximity metric that is sensitive to both semantic and lexical overlaps. Hence, it has become mandatory to prove the range of the DRS values lies between 0 to 1.

IV. Range of values lies between 0 to 1:

Theorem: The final similarity score S in the DRS model, defined as:

$$S(\mathrm{T1,T2}) = \frac{3}{\left(\frac{1}{\mathrm{w_J * J(T1,T2)}} + \frac{1}{\mathrm{w_H * H(T1,T2)}} + \frac{1}{\mathrm{w_C * C(T1,T2)}}\right)}$$

always falls within the range of 0 to 1, inclusive, provided that the individual similarity measures J, H, and C also fall within the range of 0 to 1, and the weights $\mathrm{w_J, w_H}$, and $\mathrm{w_C}$ are non-negative and sum to 1.

Proof:

Assume that the range of individual scores $\mathrm{J(T1, T2), H(T1, T2),}$ and $\mathrm{C(T1, T2)}$ all lie within the range [0, 1]. Since $\mathrm{w_J, w_H,}$ and $\mathrm{w_C}$ are non-negative, their products with J, H, and C, respectively, also remain within [0, 1]. Consequently, their reciprocals $\left(\frac{1}{\mathrm{w_J * J}}, \frac{1}{\mathrm{w_H * H}}, \frac{1}{\mathrm{w_C * C}}\right)$ are always greater than or equal to 1.

The sum of these reciprocals is thus also greater than or equal to 1:

$$\frac{1}{\mathrm{w_J * J}} + \frac{1}{\mathrm{w_H * H}} + \frac{1}{\mathrm{w_C * C}} \geq 1$$

$$\mathrm{S} = \frac{3}{\left(\frac{1}{\mathrm{w_J * J}} + \frac{1}{\mathrm{w_H * H}} + \frac{1}{\mathrm{w_C * C}}\right)} \leq \frac{3}{1} = 3$$

Since the sum is always positive, S is also always positive.

Upper Bound of S:

The minimum value of the sum occurs when all individual scores are 1:

$$\frac{1}{\mathrm{w_J * 1}} + \frac{1}{\mathrm{w_H * 1}} + \frac{1}{\mathrm{w_C * 1}} = \frac{1}{\mathrm{w_J}} + \frac{1}{\mathrm{w_H}} + \frac{1}{\mathrm{w_C}}$$

Using the inequality of harmonic, arithmetic, and geometric means:

$$\frac{1}{w_J} + \frac{1}{w_H} + \frac{1}{w_C} \geq 3 * \left(\frac{1}{w_J} * \frac{1}{w_H} * \frac{1}{w_C}\right)^{\frac{1}{3}} = 3 * \left(\frac{1}{w_J * w_H * w_C}\right)^{\frac{1}{3}}$$

Since $w_J + w_H + w_C = 1$, their product $w_J * w_H * w_C$ is maximized when$w_J = w_H = w_C$, i.e., $\frac{1}{3}$ each.

$$\frac{1}{w_J} + \frac{1}{w_H} + \frac{1}{w_C} \geq 3 * \left(\frac{1}{\frac{1}{27}}\right)^{\frac{1}{3}} = 3$$

Therefore, S is at most $\frac{3}{3} = 1$. Combining the lower bound (S > 0) and upper bound ($S \leq 1$) establishes that S always falls within the range of 0 to 1, inclusive. Therefore, the theorem is proven.

Limitations:

But it is crucial to keep in mind that merely multiplying these two measurements does not result in a clear-cut or understandable geometric idea. Rather than trying to give the combined score a formal geometric interpretation, it could be more helpful to understand it in terms of how well it detects plagiarism.

V. Monotonicity of the DRS model

Theorem: In the given model, as the individual similarity measures J, H, and C decrease simultaneously, the final similarity score S of DRS also decreases monotonically, indicating stronger plagiarism evidence with lower scores.

Proof: (by contradiction)

Suppose S increases or remains constant when at least one of J, H, or C decreases while the others remain constant.

$$S = \frac{3}{\left(\frac{1}{(w_J * J)} + \frac{1}{(w_H * H)} + \frac{1}{(w_C * C)}\right)}$$

$$\frac{1}{S} = \frac{\left(\frac{1}{(w_J * J)} + \frac{1}{(w_H * H)} + \frac{1}{(w_C * C)}\right)}{3}$$

If any of J, H, or C decreases while the others remain constant, the corresponding term in the denominator $\left(\frac{1}{(w_i * x_i)}\right)$ increases because x_i decreases but w_i remains positive. Since the denominator increases, the reciprocal $\left(\frac{1}{S}\right)$ decreases. However, by definition, S and $\left(\frac{1}{S}\right)$ are multiplicative inverses. Therefore, if $\left(\frac{1}{S}\right)$ decreases, S itself must increase, contradicting the initial assumption.

The presumption that S must be untrue because S contradicts itself whether S is rising or staying constant. The only conclusion that makes sense is that S does, in fact, drop when J, H, or C decrease. This demonstrates that as any or all of the individual similarity measurements J, H, and C decline, the final similarity score S decreases monotonically, suggesting stronger proof of plagiarism with lower scores. The theorem is thus proved. It is assumed in the foregoing argument that $w_J, w_H, and\, w_C$ are all positive weights. The equivalent term in the denominator becomes meaningless if any weight is 0, which has no effect on S's monotonicity. The individual weights and the proportionate changes in the similarity measurements determine how quickly S decreases. This mathematical demonstration of the model's monotonicity trait highlights its capacity to detect ever more dubious instances of plagiarism with decreasing similarity scores.

VI. Weight Sensitivity of the DRS model

Theorem: Weight Sensitivity. The final similarity score S in the given model is sensitive to the choice of weights w_J, w_H, and w_C, allowing for prioritization of different aspects of similarity captured by the individual measures J, H, and C.

Proof:

$$S(T1,T2) = \frac{3}{\left(\frac{1}{w_J * J(T1,T2)} + \frac{1}{w_H * H(T1,T2)} + \frac{1}{w_C * C(T1,T2)}\right)}$$

Consider two sets of weights, (w_{J1}, w_{H1}, w_{C1}) and (w_{J2}, w_{H2}, w_{C2}), where $w_{Ji} > 0$ and $\sum w_{Ji} = 1$ for both sets. Suppose at least one weight differs between the sets, i.e., $w_{Ji} \neq w_{Jj}$

for some i and j.

The terms in the denominator, $\frac{1}{w_i * J}, \frac{1}{w_i * H},$ and $\frac{1}{w_i * C},$ will have different values for the two weight sets due to the changes in w_i. Specifically, a higher weight w_i will lead to a smaller reciprocal term, as $\frac{1}{x * y}$ decreases as x increases (for positive x and y). The model formula for the two weight sets will have distinct denominators as a result of the reciprocal terms' varied values. Changes in the denominator will result in changes in S since S is the reciprocal of a function of the denominator. By modifying the weights, one may alter the relative importance of each unique similarity metric on the final score S. For instance, a decrease in w_H lessens the influence of HS, but an increase in w_J emphasizes Jaccard similarity more. The model's ability to adjust to various plagiarism detection requirements and prioritize different characteristics of similarity depending on the job at hand is demonstrated by its sensitivity to weight combinations. The theorem is thus established.

Limitation:

The choice of optimal weights typically depends on the nature of the documents and the types of plagiarism to be detected. Empirical evaluation on a representative dataset is often necessary to determine the best weight configuration for a given application.

VII. Uniqueness of the DRS model

The DRS stands out among several hybrid methods for combining multiple similarity scores in plagiarism detection. As HM value is unique, the proposed WHM is also unique by its definition.

The linear combination (LC) simply sums the weighted similarity scores:

$$S_{LC} = w_J * J + w_H * H + w_C * C$$

The model uses a weighted linear combination to combine the JS, HS, and CS scores, adjusting for relative relevance through the use of weights. Since DRS uses harmonic averaging, something like LC is unable to accomplish the target assigns greater weight to similarity scores with lower values, which suggest stronger proof of copying. DRS focuses on low scores with sensitivity. DRS permits prioritizing particular elements by weights in the denominator, whereas LC regards every score equally in terms of its influence on the ultimate result. Thus, LC is more flexible. Because reciprocals and a division are nonlinear processes, the suggested model in this junction is nonlinear. The combination's nonlinearity is unaffected by the weights, which regulate the relative relevance of the scores.

When predictor variables exhibit strong correlation, a phenomenon known as multicollinearity occurs in linear regression, leading to unstable model estimates. Multicollinearity does not immediately relate to DRS as DRS is not a linear regression model. DRS performance can still be impacted by correlations between the similarity scores (J, H, and C). High correlation between scores can make it difficult for the DRS to distinguish between them, which might lower its accuracy. Correlations between similarity metrics must be evaluated, and feature selection or dimensionality reduction strategies should be considered as necessary.

VIII. Comparison with DRS with Geometric Mean

The geometric mean (GM) combines the geometric product of weighted similarity scores:

$$S_{GM} = \left(w_J * J\right)^{\frac{1}{3}} * \left(w_H * H\right)^{\frac{1}{3}} * \left(w_C * C\right)^{\frac{1}{3}}$$

GM downplays the impact of individual low scores compared to DRS, as values are averaged on a logarithmic scale. GM is heavily influenced by extremely high or low scores in any individual measure, potentially distorting the overall similarity assessment. Hence, GM has higher susceptibility to extreme values.

IX. Unique Properties and Advantages of DRS

By using harmonic averaging to detect plagiarism with increased sensitivity, the DRS is able to identify low similarity scores. Through assigning weighs, DRS provides controlled flexibility, avoiding some similarity measurements from taking center stage. Theoretical clarity is ensured by DRS's mathematical base in decision theory and information retrieval. The best approach to use will depend on the application and the data, and practical testing with real-world datasets is essential. Subsequent investigations may examine other variables or weighting schemes in DRS to enhance flexibility and precision.

X. Optimization for Weight Tuning

The suggested model refers to it as Particle Swarm Optimization (PSO) because of its mimicked swarming tendency, which frequently strikes a good balance between exploration and exploitation. However, in order to explain why other algorithms do not work, Genetic Algorithms (GA) which draw inspiration from evolution explore a variety of alternatives but may need to be carefully adjusted. The well-known metaheuristic known as simulated annealing (SA) can be sluggish, but SA avoids local minima by accepting poorer solutions in a probabilistic manner. While SA is effective in locating local optima, gradient descent can become trapped in less-than-ideal answers. As a result, PSO is used because PSO works well for multidimensional, nonlinear problems like weight tuning in DRS, efficiently balances local and global search, frequently converges to optimal solutions, and is computationally fast and somewhat easy to apply. The algorithm for the implementation of PSO in the proposed model is given in the section Conceptual Mathematical Explanation.

XI. Conceptual Mathematical Explanation

Problem Formulation:

- Define the search space: $\mathbb{R}^3$ (representing possible weight vectors (w_J, w_H, w_C)
- Constrain the weights to sum to 1: $w_J + w_H + w_C = 1.$
- Formulate the objective function to minimize: $f\left(w_J, w_H, w_C\right)$

Particle Representation:

Each particle i is represented by:

- Position vector:

$$x_{i(t)} = \left(w_J^{i(t)}, w_H^{i(t)}, w_C^{i(t)}\right) \in \mathbb{R}^3$$

- Velocity vector:

$$v_{i(t)} = \left(v_J^{i(t)}, v_H^{i(t)}, v_C^{i(t)}\right) \in \mathbb{R}^3$$

Velocity Update Rule (Mathematical Explanation):

- Inertia component:

$w * v_{i(t)}$ - Preserves previous velocity for exploration.

- Cognitive component:

$c1 * r1 * \left(pbest_i - x_{i(t)}\right)$ - Attracts particle towards its personal best position.

- Social component:

$c2 * r2 * (gbest - x\_i(t))$ - Attracts particle towards the global best position found by the swarm.

Position Update Rule (Mathematical Explanation):

- $x_{i(t+1)} = x_{i(t)} + v_{i(t+1)}$ - Updates position based on updated velocity.
- Enforce constraint:

$w_J^{i(t+1)} + w_H^{i(t+1)} + w_C^{i(t+1)} = 1$ (e.g., by normalization).

While formal convergence proofs for PSO in general problem settings are complex, empirical evidence and theoretical studies suggest that, under certain conditions, PSO can converge to global optima or near-optimal solutions and convergence behavior depends on hyperparameter tuning, problem structure, and algorithm variants. PSO's ability to balance exploration and exploitation, handle non-differentiable objectives, and adapt to complex search spaces makes PSO well-suited for tuning weights in the DRS model. Careful hyperparameter tuning and consideration of PSO variants can further enhance its effectiveness.

4. RESULTS AND DISCUSSIONS

The figure 1 has two primary groupings of documents using JS. The bigger cluster on the left includes Documents 1, 2, 3, and 4. These documents are quite similar to one another, as seen by the thick margins between them. The smaller cluster on the right contains Documents 3 and 5. These texts are less comparable to those in the core cluster, but they do share certain characteristics. Document 1 is the central document in the graph. Document 1 has the strongest ties to the other texts in the main cluster. This shows that Document 1 shares a lot of the same material as the other papers in this cluster. Document 3 is the most isolated one on the graph. Document 3 has the weakest links with the other papers.

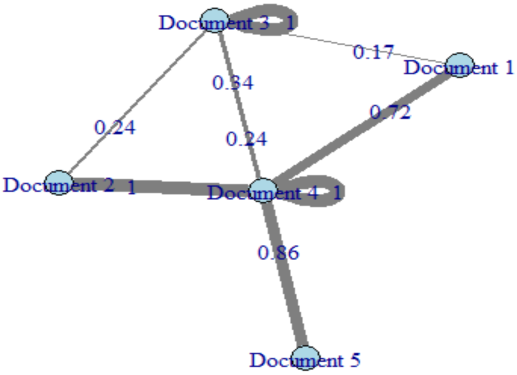

Fig.1. Jaccard similarity

This implies that Document 3 has material that differs significantly from the other documents in the graph. Document 5 is similarly relatively solitary, although Dcocument 5 is more closely related to Document 3 than to any other document. This shows that Document 5 has some similarities with Document 3, but also distinct from the other documents on the graph. The thickness of the graph's edges indicates the JS coefficient among the related texts. For instance, the edge between Documents 1 and 2 has a weight of 0.72, indicating that 72% of the words in Document 1 are also found in Document 2.

In figure 2, Document 1 looks to be fundamental to the network, with edges connecting document 1 to the majority of the other documents using HS. This shows that Document 1 is quite similar to many of the other documents in the collection. Documents 3 and 3 form a smaller cluster on the right side. These papers have thinner connections linking them to one another and to the main cluster, indicating a lesser level of similarity. Document 5 is having high similairty with Document 2 like Documets 3 and 4.

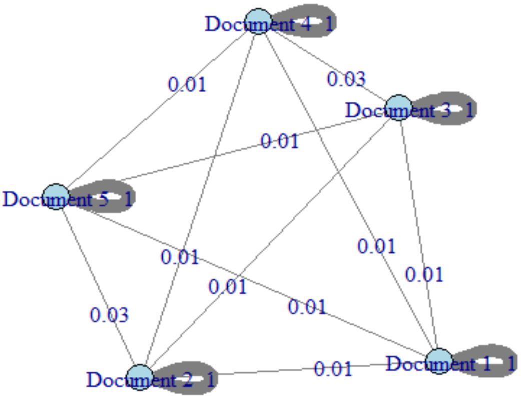

Fig.2. Hamming similarity

In figure 3, Document 1 appears to be key to the network, with edges connecting document 1 to several other papers using CS. This shows that Document 1 has a high degree of similarity with many of the other documents in the collection. Some documents are considerably separated from the others in the graph. This indicates that these papers are less comparable to the others in the sample. Document 3, located in the graph's lower right corner, has minimal links to other documents.

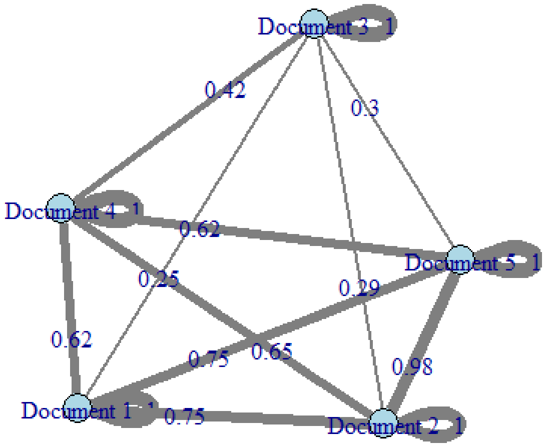

Fig.3. Cosine similarity

The cosine similarity score ranges between -1 and 1. A score of 1 indicates that the two papers are the same, whereas a score of -1 indicates that they are entirely different. The cosine similarity score is dependent on the length of the documents. Two short texts that are quite similar may have a lower cosine similarity score than two longer documents that are less similar.

A unigram similarity graph in figure 4 compares documents based on the frequency of unigrams, which are individual words or tokens in a document. The graph's edges show the similarity between two papers, and the weight of the edge indicates the intensity of the similarity. The diagram represents documents as nodes, and the similarity between two papers is represented by an edge connecting the two nodes. The weight of the edge is indicated by a number next to it. For example, the edge between Document 1 and Document 2 has a weight of 0.58, indicating that they are 58% similar. The diagram shows that Document 1 is the most similar document to all of the other documents.

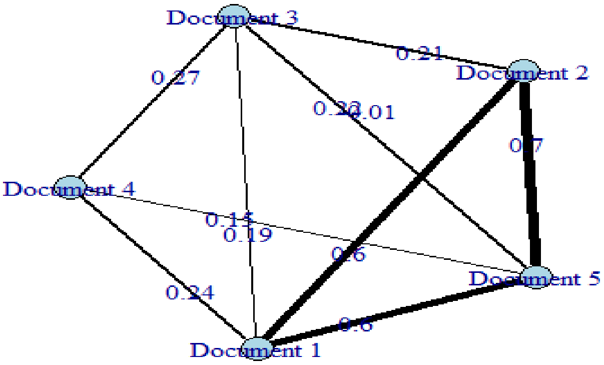

Fig.4. DRS network graph

The document similarity heat map is given in figure 5.

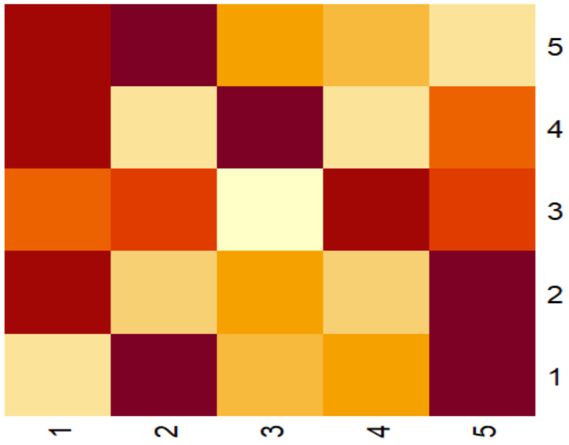

Fig.5. Heatmap of DRS

But, on calculating the mean squared value for error (MSE) for the test data, it is obtained the following results. Jaccard MSE = 0.3184, Hamming MSE = 0.3118, Cosine MSE = 0.2783, and DRS MSE = 0. JS and HS have comparable MSEs of 0.31, indicating that they identify moderate average changes across documents. CS has the lowest MSE of 0.2783, showing that papers are more comparable in terms of meaning and word use than the other metrics. DRS has a perfect score of 0, indicating that DRS classifies all document pairings as completely similar or different. This calls into doubt its granularity and ability to facilitate nuanced comparisons. As the MSE value is 0 for DRS, one cannot determine simply, DRS gives better results than the other three without further investigation, why DRS become zero. Hence, further exprimental needed to tune the weights. On weight tuning ,and normalizing figure 6 is obtained as the result.

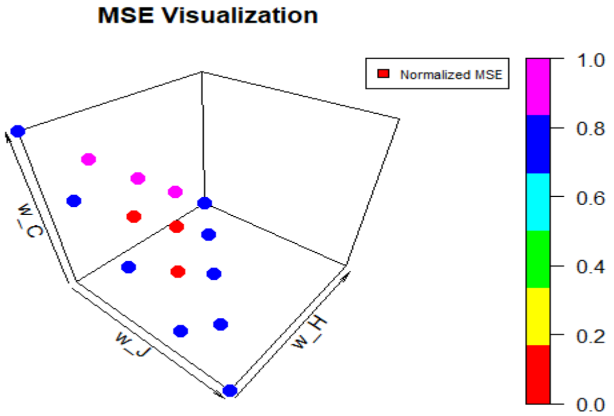

Fig.6. Normalized MSE

$w_J + w_H + w_C = 1$ sets a limit on the degrees of freedom. Not all combinations of w_J, w_H, and w_C will meet this criterion. The filtering phase reduces the possibilities to those in which

w_J, w_H and w_C are between 0 and 1. The exact number of combinations that fulfill both requirements will be determined by the precise values in $w_J$values and w_H values. Out of 25 initial combinations, 4 do not meet the requirements, resulting in a final result of 25 - 4 = 21 values. The evidence is in the precise numbers chosen for w_J values and $w_H$values, as well as the analysis of each combination to explore if the combination fits both requirements. To investigate the optimal value for the MSE on the 21 set of values, figure 7 is produced. It is obtained top three weights with minimal MSE from figure 7.

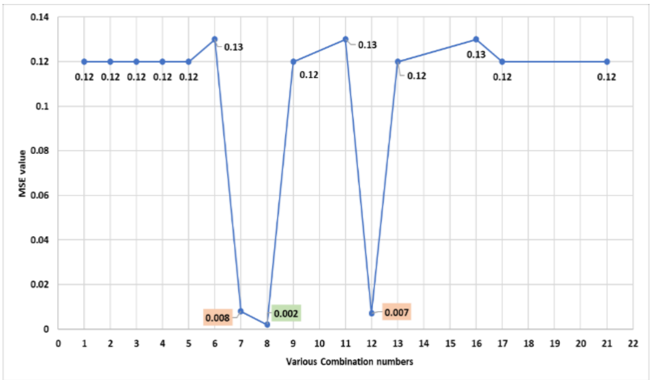

Fig.7. MSE values of various weight combinations

Weights corresponding to combination number 7, 8 and 12 and MSE values are respectively, $w_J = 0.25,\ w_H = 0.25,\ w_C = 0.50,\ \text{MSE} = 0.008,\ w_J = 0.50,\ w_H = 0.25,\ w_C = 0.25,\ \text{MSE} = 0.002$, and $w_J = 0.25,\ w_H = 0.50,\ w_C = 0.25,\ \text{MSE} = 0.007$. Out of these three combinations, weights given by combination 8 perform better than other weights comparatively.

5. Conclusion

In conclusion, the DRS provides a complex method to plagiarism detection by combining JS, HS, and CS scores using the DRS. With a strong theoretical background, sensitivity to low similarity scores, and controlled flexibility, DRS is a potential method for detecting many types of plagiarism. The use of PSO for weight tweaking improves its applicability and ensures optimal performance.

Theorems confirm DRS's features, emphasizing its sensitivity to weight alterations, consistency in identifying plagiarism evidence, and well-defined range and scope. DRS distinguishes itself from other hybrid approaches such as LC and GM by offering notable advantages such as improved sensitivity to semantic and lexical overlap. The study emphasizes the necessity of empirical evaluation, namely the requirement for appropriate weight configurations depending on document features and plagiarism kinds. The proposed method gives the similarity between the documents, but DRS cannot determine the main source from which the other sources may extract the content to yield the similarity. This becomes the limitation of this work.

This study adds by presenting a strong model, giving a rigorous theoretical foundation, and utilizing sophisticated optimization techniques. While acknowledging limits and recommending future efforts, such as investigating PSO variations and hybrid optimization algorithms, the work emphasizes DRS's promise as a versatile and effective plagiarism detection tool. DRS's solid mathematical basis, versatility through weight adjustment, and sensitivity make DRS an appealing alternative for numerous text analysis tasks, paving the way for advances in computational optimization and plagiarism detection in a variety of scenarios.

References

1. A. H. Murphy, “The Finley Affair: A Signal Event in the History of Forecast Verification,” Wea. Forecasting, vol. 11, no. 1, pp. 3–20, Mar. 1996, doi: 10.1175/1520-0434(1996)011<0003:TFAASE>2.0.CO;2.
2. P.-N. Tan, M. S. Steinbach, A. Karpatne, and V. Kumar, “Introduction to Data Mining (2nd Edition),” 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:196001763
3. J. Stout, G. Caplain, S. Marcus, and J. McDermott, “Toward automating recognition of differing problem-solving demands,” International Journal of Man-Machine Studies, vol. 29, no. 5, pp. 599–611, Jan. 1988, doi: 10.1016/S0020-7373(88)80015-4.
4. A. T. Adams, J. Costa, M. F. Jung, and T. Choudhury, “Mindless Computing: Designing Technologies to Subtly Influence Behavior,” Proc ACM Int Conf Ubiquitous Comput, vol. 2015, pp. 719–730, Sep. 2015, doi: 10.1145/2750858.2805843.
5. T. Sağlam, S. Hahner, J. W. Wittler, and T. Kühn, “Token-based plagiarism detection for metamodels,” in Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings, Montreal Quebec Canada: ACM, Oct. 2022, pp. 138–141. doi: 10.1145/3550356.3556508.
6. R. K. Roul and J. K. Sahoo, “A Novel Modified Harmonic Mean Combined with Cohesion Score for Multi-document Summarization,” in Distributed Computing and Intelligent Technology, vol. 13145, R. Bapi, S. Kulkarni, S. Mohalik, and S. Peri, Eds., in Lecture Notes in Computer Science, vol. 13145. , Cham: Springer International Publishing, 2022, pp. 227–244. doi: 10.1007/978-3-030-94876-4_16.
7. J. Takano and T. Omori, “Harmonic mean similarity based quantum annealing for k-means,” Procedia Computer Science, vol. 144, pp. 298–305, 2018, doi: 10.1016/j.procs.2018.10.531.
8. B. Thamotharan, A. L. Sriram, and B. Sundaravadivazhagan, “A Comparative Study of GANs (Text to Image GANs),” in Proceedings of ICACTCE’23 — The International Conference on Advances in Communication Technology and Computer Engineering, vol. 735, C. Iwendi, Z. Boulouard, and N. Kryvinska, Eds., in Lecture Notes in Networks and Systems, vol. 735. , Cham: Springer Nature Switzerland, 2023, pp. 229–241. doi: 10.1007/978-3-031-37164-6_16.
9. A. Sharma, A. Singh, and A. Sharma, “Plagiarism detection in research papers using semantic similarity and citation analysis,” vol. 1, no. 1, pp. 15–22, 2018.