JSCDM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

Vol. 7 No. 1 (2026) 91-105
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Optimizing Nurse Knowledge Management in Healthcare Services using the Jaro-Winkler Algorithm

Kemal Farouq Mauladi[1], Suziyanti Marjudi[1], Mohd Fahmi Mohamad Amran[2]*, Puteri Nor Ellyza Nohuddin[3], Nur Amlya Abd Majid[4]

[1] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, MALAYSIA*

[2] *Department of Computer Science, Faculty of Defence Science and Technology, National Defence University of Malaysia, 57000 Kuala Lumpur, MALAYSIA*

[3] *Faculty of Business, Higher Colleges of Technology, Sharjah, UNITED ARAB EMIRATES*

[4] *Tamhidi Centre, University Sains Islam Malaysia, Bandar Baru Nilai, Negeri Sembilan, MALAYSIA*

*Corresponding Author: fahmiamran@upnm.edu.my
DOI: https://doi.org/10.30880/jscdm.2026.07.01.008

Article Info

Received: 30 October 2025
Accepted: 23 February 2026
Available online: 10 April 2026

Keywords

Jaro-Winkler algorithm, knowledge management, nurses, health services, Lamongan Regional Hospital

Abstract

Effective nurse knowledge management is essential for supporting clinical decision-making and ensuring high-quality patient care. However, hospitals often face challenges such as data duplication, information inconsistency, and inefficient data retrieval. This study explores the integration of the Jaro-Winkler algorithm into a nurse knowledge management model to enhance health services at Lamongan District Hospital. The algorithm was selected for its precision in measuring string similarity, which facilitates accurate data identification and the elimination of redundant entries. The methodology included data collection, system requirements analysis, algorithm implementation, and performance evaluation based on accuracy and retrieval speed. The findings revealed a 30% reduction in data duplication and a 40% increase in data retrieval speed. These improvements underscore the algorithm's effectiveness in optimizing knowledge management processes. The study contributes to the development of intelligent, technology-based systems that can be adopted by other healthcare institutions to improve service quality and operational efficiency

1. Introduction

The era of health digitalization demands an efficient and accurate transformation of clinical knowledge management systems, especially in managing nurse data as the spearhead of patient care. Hospitals in Lamongan District as referral health institutions face complex challenges in managing nurse knowledge bases that include standard procedures, clinical experience, and training records. The main problems identified are the inefficiency of the data retrieval system and the high duplication of knowledge records due to variations in the writing of medical terms and inconsistencies in data input [1].

This phenomenon has a systemic impact on the quality of service, where based on initial observations it was found that 40% of nurses' time was wasted on data verification and 28% of medication error cases were caused by inaccurate information. Conventional systems that rely on exact match searches have proven unable to accommodate linguistic variations in data entry, while traditional keyword-based solutions have limitations in handling typos and abbreviations that are common in clinical environments. Research conducted by Carrington (2018) shows that the effectiveness of nurse knowledge management is directly correlated with the quality of health services. An optimal system must be able to accommodate various forms of tacit and explicit knowledge. However, a case study of a Hospital in Lamongan revealed that 32% of knowledge assets were not optimally utilized due to limitations in the search system [2].

According to WHO (2021), 45% of medical errors stem from inaccurate patient data. Research conducted by Smith et al. (2020) in 10 Asian hospitals found that: (1) Conventional systems only achieve a 65% accuracy rate for drug name searches. (2) Average search time reaches 2.3 minutes per query. (3) Data duplication rate reaches 27% of total records [3].

The Jaro-Winkler algorithm appears as a promising technical solution with its main advantage in calculating the similarity score between text strings, especially effective for handling typographical errors and variations in the writing of names/terms. The implementation of this algorithm in a knowledge management search system is expected to [4]:

1. Increase the recall rate of clinical information search.
2. Reduce data redundancy through automatic document similarity detection.
3. Optimize system response time.

The advantages of the Jaro-Winkler algorithm are: (a) The relationship between problems lies in the inconsistency of terms in the documentation, (b) Jaro-Winkler was chosen because it can overcome the problem of variations in writing, spelling, and abbreviations by providing a high similarity score even though the strings are not identical. (c) This way, searching and matching nurse knowledge data at Lamongan Regional Hospital becomes more accurate and faster [5] [6].

This study has strategic urgency considering the digital transformation policy of the Indonesian Ministry of Health 2020-2024 which emphasizes the importance of hospital information system interoperability. The results of the study are expected to not only solve technical problems at the Lamongan District Hospital but also become a reference model for implementing similarity matching for clinical knowledge management at the regional level. The targeted long-term impact is increasing patient safety through the availability of fast, accurate, and integrated clinical information.

2. Methodology

2.1 Research Design

This research design uses a quasi-experimental approach with a system development method. The goal is to implement and evaluate the effectiveness of the Jaro-Winkler algorithm in improving the knowledge management system of nurses at Lamongan District Hospital. The object of research used is knowledge management with the SECI Model dimensions consisting of Socialization, Externalization, Combination, and Internalization as a knowledge management process. The research design can be seen in Figure 1 below.

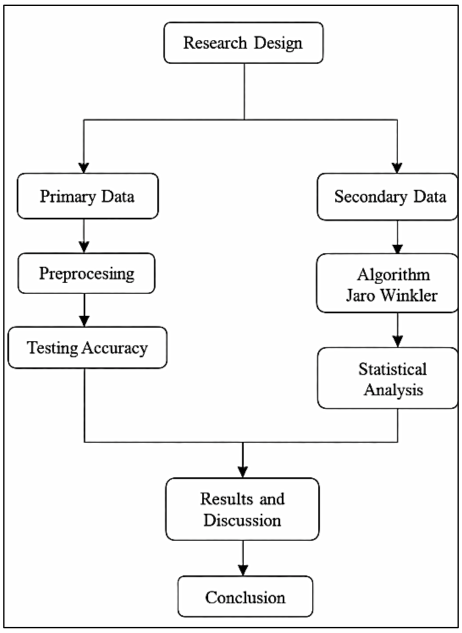


Fig. 1 *Research design*

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2026) p. 91-105

2.2 Data Collection

Data collection techniques used in this study are interviews, observations and documentation. Data analysis in the study includes reducing data, presenting data or displaying data, and drawing conclusions to verification. Data collection techniques in this study consist of two stages, including the following [7]:

1. Secondary Data
   a. Direct observation of the use of knowledge management systems at Lamongan District Hospital.
   b. Interviews and questionnaires with nurses to understand information search and utilization patterns in health services.
2. Primary Data
   a. Literature study related to Jaro-Winkler algorithm, knowledge management system, and health service optimization.
   b. Documents and reports related to nurse performance and the effectiveness of health services before and after system implementation, such as medical record data, health service protocols, and hospital standard operating procedure (SOP) documents.

2.3 Implementation of the Jaro-winkler Algorithm

The Jaro-Winkler distance is a variant of the Jaro Distance metric typically used to identify document relationships or duplications by measuring the similarity between two strings. Jaro-Winkler is best at recognizing short strings. A score of 0 indicates no similarity to the reference data, and a score of 1 indicates an exact match [8].
The implementation of the Jaro-Winkler algorithm in the healthcare world has proven to be an effective solution in various healthcare applications, especially for addressing complex clinical data management challenges
This algorithm is used to calculate the similarity score between two strings with the formula [9]:

1. Jaro-similiarity

$$Sim_j = \begin{cases} 0 \\ \frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{s2} + \frac{m-t}{m}\right) \end{cases} \quad (1)$$

m = 0
with:
m: Number of match characters
t= number of transpositions

2. Winkler modification

$$sim_w = sim_j + (\iota . \rho . (1 - sim_j)) \quad (2)$$

with:
ι = same prefix length (max 4 characters)
ρ = scale constant (usually 0.1)

2.4 Application of SECI Model to Medical Document

The SECI model provides an effective framework for knowledge management in hospitals. By integrating tacit and explicit knowledge, hospitals can improve the quality of health care, support clinical decision-making, and ensure that the best knowledge is used in daily practice. The SECI model flow can be adapted to various contexts, including knowledge management of nurses, physicians, and other health care professionals. The implementation of the SECI Model in hospitals in Lamongan District is as follows [10]:

1. *Socialization*
   - Case: A senior nurse shares her experience of managing a patient with complications of diabetes.
   - Output: Tacit knowledge of effective care techniques
2. *Externalization*
   - Case: Senior nurse documents her experience in the form of a case report.
   - Output: Case report that can be accessed by other nurses.
3. *Combination*
   - Case: The knowledge management team combines the case report with nursing protocols from other departments to develop a clinical guideline on the care of a patient with diabetes.
   - Output: Comprehensive clinical guideline.
4. *Internalization.*
   - Case: Nurses learn and apply new clinical guidelines in daily practice.
   - Output: Improvement in nurses' skills and tacit knowledge. professionals. The implementation of the SECI Model in hospitals in Lamongan District is as follows:

3. Results and Discussions

Problems often experienced by nurses in hospitals in Lamongan District are information from medical records that are not well organized, making it difficult to recognize groups of patients or disease symptoms with similar characteristics, as well as patient service orders that are still often not in accordance with hospital standard operating procedures. Therefore, in this study, how the application of the Jaro-Winkler algorithm using medical record data and SoPs in hospitals in Lamongan District, can be used as a guideline for hospitals to improve the quality of service and management of nurse knowledge into hospital databases. Optimization of the Jaro-Winkler algorithm has proven effective in addressing the challenges of managing unstructured data in nurse documentation [11].

3.1 Text Preprocessing

Text preprocessing is the initial stage of text mining where text data will be cleaned so that the text becomes more structured before entering the next stage for further processing. A collection of connected characters (text) must be broken down into more meaningful ones. This can be done at several different levels. A document can be broken down into a chapter, paragraph, sentence and word. The stages of text preprocessing include [8]:

3.1.1 Cleaning

Cleansing is a preprocessing stage in Text Mining that aims to clean text from irrelevant or disruptive elements in the analysis (See Table 1) In medical record documents, cleansing is very important because medical data often contains noise, such as punctuation, numbers, special characters, or inconsistent formats. Here are some elements that are usually removed or corrected during the cleansing process. The Cleansing process involves several steps, depending on the needs of the analysis. Here are the general steps in the cleansing process of this study [12]:

a. Removing Punctuation in this process punctuation such as dots, commas, or exclamation marks are removed because they are not relevant to text analysis.
b. Removing numbers in the process of removing irrelevant numbers (such as dates or room numbers) are removed, for example, a document resulting from the cleansing process removes numbers.
c. Removing special characters to remove characters such as hashtags (#) or symbols.

Table 1 *Cleaning*

Medical Documents Before Cleaning	Medical Document After Cleaning
The patient complained of abdominal pain, nausea and vomiting!!	The patient complained of abdominal pain, nausea and vomiting.
Patient: high fever @ night...	patients with high fever at night
Chief complaint: severe headache and dizziness!	main complaint: severe headache and dizziness
Nausea and vomiting have been occurring for approximately 2 days.	nausea and vomiting have been occurring since last day
Pain in the upper right abdomen. #ER	pain in the upper right abdomen in the emergency room

3.1.2 Case Folding

One of the preprocessing stages in text mining aims to equalize the letter form (case) in the text, usually by changing all letters to lowercase. This process is important to ensure consistency in text analysis, because text mining systems often treat uppercase and lowercase as different entities. Here is an explanation of the case folding process in the context of patient medical documents. In the case folding process (See Table 2), all letters in the text are changed to lowercase. Here are the following:

a. Text Identification: taken from the text of medical documents.
b. Convert to Lowercase: Convert all uppercase letters to lowercase using the function.

Table 2 *Case folding*

Medical Documents Before Casefolding	Medical Document After Casefolding
The Patient Complained of Abdominal Pain, Nausea and Vomiting	the patient complained of abdominal pain, nausea and vomiting.
Patient: High Fever Night	patients with high fever at night
Chief Complaint: Severe Headache And dizziness	main complaint: severe headache and dizziness
Nausea and Vomiting Have Been Occurring for Approximately 2 Days	nausea and vomiting have been occurring since last day
Pain in the Upper Right abdomen.	pain in the upper right abdomen in the emergency room

3.1.3 Stemming

Stemming is a preprocessing step in text mining that aims to convert affixed words (such as verbs, nouns, or adjectives) into their basic form (root word) as illustrated in Table3. This process is essential for reducing word variation and simplifying text analysis. In the context of medical records, stemming helps identify relevant keywords, such as symptoms, diagnoses, or medical procedures, even if these words appear in various forms.

Table 3 *Stemming*

Before the Process	After Process
"complaining of chest pain"	"complaints of chest pain"
"therapy is provided by nurses"	"provide nursing behavior therapy"
"bleeding occurs during the procedure"	"blood becomes during the act"

3.2 Jaro-Winkler Distance

The Jaro-Winkler algorithm process is carried out by checking the token in each sentence spelling with a list of words in the Indonesian dictionary database on Health. If a match is found, the spelling of the word will be considered correct and will not go through the distance calculation stage. However, if no match is found in the Indonesian dictionary database on Health, the system will calculate the distance of its proximity to the list of words identified as having similarities based on the number of characters or letters of the token [13]. Jaro-Winkler Distance has three basic components that can be seen in Figure 2 below.

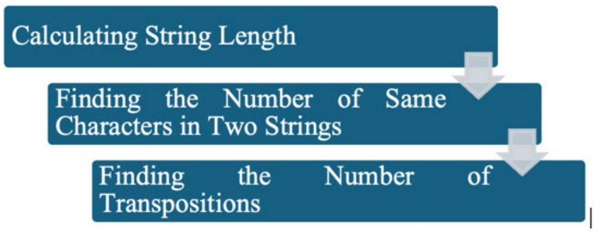


Fig. 2 *Jaro-Winkler algorithm components*

Jaro-Winkler Distance uses a prefix scale (p) that provides a higher level of assessment, and a prefix length (l) that states the prefix length, namely the length of the same character from the string being compared until a dissimilarity is found. If the strings s1 and s2 are compared, then the Jaro-Winkler Distance is (dw) is:

$$Dw\ (s_1,s_2) = Jaro\ (s_1,s_2) + (L*p(1-Jaro(s_1,s_2))) \quad (3)$$

Where jaro is the Jaro Distance for strings s_1 and s_2, t is the length of the common prefix at the beginning of the string, the maximum value of which is 8 characters, while p is a scaling factor constant with a standard value according to Winkler of 0.1.

Suppose the user writes the word "diabetes" but makes a mistake so that it is written as "diabetis". The word "diabetis" is not found in the word list in the database so the word "diabetis" will be calculated its proximity to all words with eight to ten characters registered in the database using the Jaro-Winkler Distance algorithm. Example: "diabetis", "diabetus", "diabet", "diabetesis", "diabetes".

Calculation of the proximity of the word "diabetis" (s_1) with "diabetes" (s_2) using the Jaro-Winkler Distance algorithm.

String1 (s_1)	=	*Diabetis*
String2 (s_2)	=	*diabetus*
String Character Count1 $\|s_1\|$	=	8
String Character Count2 $\|s_2\|$	=	8
Number of characters s1 matches s2 (m)	=	2
Number of Transpositions (t)	=	0

$$Jaro\ (s_1, s_2) = \frac{1}{3}\left(\frac{2}{8} + \frac{2}{8} + \frac{2-0}{2}\right)$$

Jaro (s1, s2) = 0.84

Jaro-Winkler (s_1, s_2) = *Jaro (s1, S2) + (L*p(1Jaro(s1, s2)))*

Prefix Similarity		*2*
Scaling Factor		*0.1*
Jaro-Winkler (s_1, s_2)	=	*0.84 + (2*0.1((1-0.84)))*
Jaro-Winkler (s_1, s_2)	=	*0.2*
String1 (s_1)	=	*Diabetis*
String2 (s_2)	=	*Diabetus*

Number of Characters in String $\|s_1\|$	=	*8*
Number of Characters in String $\|s_2\|$	=	*8*
Number of Characters s_1 matches s_2 (m)	=	*2*
Number of Transpositions	=	*0*

So that the calculation results can be seen in Table 4

Table 4 *Computational analysis design*

No	String 1	String 2	JW Value
1	diabetis	diabetus	0.84
2	diabetis	diabet	2.91
3	diabetis	diabetes	0.84
4	diabetis	diabetic	0,84

From the calculation results above, the word "diabetis" has the closest similarity to the word "diabetic" which has a proximity value of 0.91. So, the word "diabetic" will replace the word "diabetic" in the autocorrect system and restore the particles that were deleted during the stemming process if any [1].

3.3 Implementation of Jaro-Winkler Model

Modeling using Python version 3.11 to calculate the Jaro-Winkler distance value. The concept of applying Jaro-Winkler distance to this typographical error identification model is that each word entered is compared to the words in the dataset [14]. This dataset is a representation of words that are considered correct. Each word is checked one by one whether the Jaro-Winkler distance value is 1 or not. If it is one, the word entered is considered correct. Loops are used to check all the words entered. The function code for this process can be seen in table 5 below.

Table 5 *Jaro-Winkler*

No	Nurse's Notes
1	Mrs. Romlah, age 55, address Jl Raya Mantup km 3, complaining of shortness of breath since the previous day, history of asthma since childhood
2	Mrs. Siti Rukiyah, age 45, address Jl Raya Mantup, has been suffering from diabetis for 4 days.
3	Miss, Ida Nuraini, age, 24, address, Mantup Dlandangan, has suffered from hypertension for 7 days
4	Tuan, Sukardi, age, 42, address, mantup, has been suffering from fever, flu, cough for 4 days
5	Mr. Agung, age 35, address Mantup Sukobendu, has been suffering from diabetes for 15 days.

From table 5 above, each input of disease symptoms that has been entered by the nurse will be preprocessed first. This input text preprocessing stage uses the Jaro-Winkler library. Each input is subjected to a stopword removal process to remove unnecessary stopwords or conjunctions such as, which, on, and so on. This process produces input without useless words. Furthermore, each input is converted into tokens or arrays, and non-letter characters are removed so that only tokens containing words are produced. Each word that has been tokenized is subjected to a stemming process that produces a string consisting of root words with lowercase letters throughout.

3.4 Topic Modelling Latent Dirichlet Allocation (LDA)

Topic modeling is used to identify latent topics in a collection of text documents. LDA automatically groups words and documents based on the pattern of similar word occurrences. The main functions of LDA in nursing knowledge management are as follows:

a. Identifying key topics in nursing notes, patient reports, or medical SOPs
b. Classify information into themes such as “wound care”, “infection management”, “vital observations”, etc.
c. Help nurses or hospital admins navigate documents by topic, not just keyword searches.

How the LDA model topic works input collects medical documents nurse notes in a patient database. From the text collection, it is divided into several processes, including the following: (a) grouping each topic in the form of a collection of keywords, (b) Each document is given a probability of being related to several topics.

In the LDA process, 3 to 5 main topics were found in the nurse's medical documents which will be divided into 3 topic categories, including the following [15]:

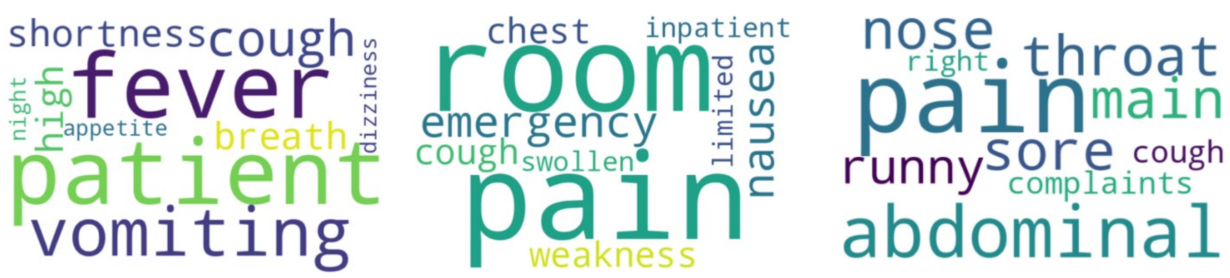
Topic 1 Gastrointestinal Complaints: “nausea,” “vomiting,” “stomach pain,” “diarrhea”
Topic 2 Respiratory Disorders: “shortness of breath,” “cough,” “fever”
Topic 3 Injuries and Trauma: “wound,” “fall,” “leg pain,” “swelling”.

Table 6 *Topic modelling Latent Dirichlet Allocation (LDA)*

Topic	Keyword	Weight
Topic 1	painful	5.12
Topic 1	vomit	4.98
Topic 1	stomach	4.75
Topic 1	nauseous	4.22
Topic 1	Dizzy	3.88
Topic 1	chest	3.50
Topic 1	weak	3.40
Topic 1	care	3.12
Topic 1	stay	3.00
Topic 1	Eat	2.88
Topic 2	cough	5.40
Topic 2	fever	5.12
Topic 2	breath	4.92

Topic	Keyword	Weight
Topic 2	have a cold	4.33
Topic 2	throat	4.11
Topic 2	dyspnea	3.95
Topic 2	Emergency installation	3.80
Topic 2	Evening	3.22
Topic 2	phlegmy	3.10
Topic 2	tall	2.95
Topic 3	wound	5.22
Topic 3	foot	5.05
Topic 3	fall	4.92
Topic 3	swollen	4.30
Topic 3	knee	4.12
Topic 3	movement	3.90
Topic 3	limited	3.72
Topic 3	bruises	3.55
Topic 3	room	3.20
Topic 3	igd	3.05

The integration of Latent Dirichlet Allocation (LDA) into the Knowledge Management system helps hospitals in Lamongan District as follows: (a) Filtering important information from text documents, (b) Providing relevant, structured and easily accessible knowledge, (c) Improving the quality of nursing services with an evidence-based approach (evidence-based nursing).



(a) **(b)** **(c)**

Fig. 3 *LDA topic process results*

The results of Figure 3's Latent Dirichlet Allocation (LDA) WordCloud show that the dominant topics in nurse documentation revolve around the patient's chief complaint (pain, cough, nausea, swelling) and the location of care (ward, emergency room, inpatient). This supports the integration of complaint data with hospital service management [16].

The WordCloud process shows that there is a close relationship between the patient's main symptoms (pain, nausea, cough, swelling) and the treatment room (emergency room, inpatient room). This information can be useful for the knowledge management database of Hospital nurses in Lamongan District, for example grouping patient cases based on symptoms and location of care.

J. of Soft Computing and Data Mining Vol. 7 No. 1 (2026) p. 91-105

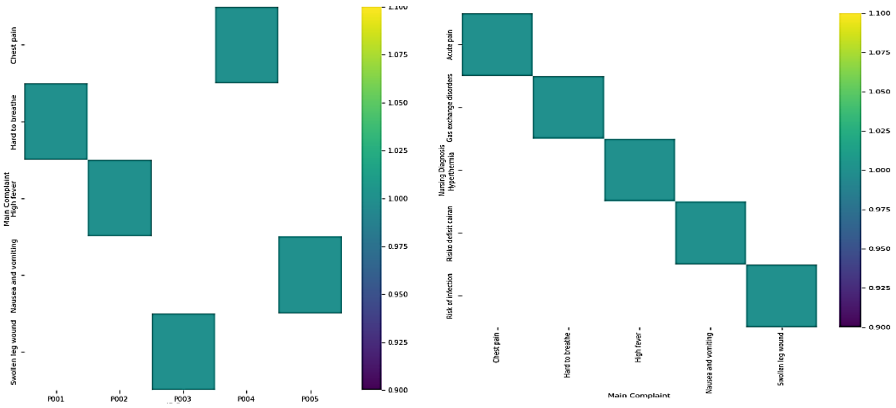


Fig. 4 *Categorical distributions*

Figure 4 explains that the process of identifying similarity of nurse documentation data with NLP algorithm and Jaro-Winkler is running well, because each patient's main complaint can be mapped precisely to the relevant nursing diagnosis, with a very high level of similarity. The level of similarity or semantic closeness between patient complaints and nurse diagnoses, possibly calculated by the Jaro-Winkler algorithm or other NLP methods. The results of the diagnosis of Chest pain with Acute pain", and "Hard to breathe with Gas exchange disorders" similarity value is close to 1.0, meaning the system has successfully matched complaints with the correct diagnosis.

3.5 Term Frequency – Inverse Document Frequency (TF-IDF)

The TF-IDF process in nurse documentation data was analyzed by calculating word weights using TF-IDF, such as the common words "patient," "pain," and "complaint" having high frequencies but low IDF. Next, the words "right upper abdomen," "abdominal distension," or "febrile seizures" have low TF but high IDF, making them more contextually important [17].

TF-IDF in optimizing nurses' knowledge search and matching using the Jaro-Winkler algorithm at a hospital in Lamongan District can identify important keywords in case-specific documentation. TF-IDF can be useful for complaint classification, feature extraction, and automatic clustering of patient documentation.

TF-IDF can be useful for complaint classification, feature extraction, and automatic clustering of patient documentation. TF-IDF (Term Frequency – Inverse Document Frequency) used to measure the importance of a word (term) in a document relative to the entire corpus.

$$\mathrm{TF}(t,d) = \frac{Number\ of\ occurrences\ of\ word\ t\ in\ document\ d}{Total\ words\ in\ document\ d} \tag{4}$$

Inverse document Frequency (IDF)

$$\mathrm{IDF}\ (\mathrm{t}) = \log \frac{N}{df\ (t)} \tag{5}$$

Where
N = total number of documents
Df(t) number of documents containing the word t
TF-IDF(td)= TF(t,d) XIDF (t)
Document 1 : "the patient complained of chest pain"
Document 2 : "patient experiencing shortness of breath in the emergency room"
Document 3 : "patients with nausea, vomiting and high fever"
Count TF
D1 : TF = $\frac{1}{4}$ = 0.25
D2 : TF = $\frac{1}{6}$ = 0.1667
D3 : TF = $\frac{1}{6}$ = 0.1667

Count IDF
Number of documents N = 3
The word “patient” is present in all df (“patient”) documents = 3
IDF (“patient”) = log $\frac{3}{3}$ log 1 = 0
Low weight (100uet o general)
The word “pain” only appears in D1 with df (“pain”) = 1
IDF (“painful”) log $\frac{3}{1}$ = log 3 = 1.098
Sentences or words “IGD” only appears in D2 with df (“IGD”) = 1
IDF (“IGD”) = log $\frac{3}{1}$ = 1.098
Count TF-IDF
Count “painful” di D1:
TF (“painful”, D1) = $\frac{1}{4}$ = 0.25
TF-IDF (“painful, D1) = 0.25 X 1.098 = 0.274

The results of TF-IDF produce word weight vectors representing nursing documents. Jaro-Winkler is used in the string matching stage to measure the similarity between medical/diagnostic terms.

Table 7 *Important words that best characterize each document [1]*

Words	TF (freq/5 document)	IDF (log(5/df))	TF-IDF
Painful	1	log(5/1)=0.70	0.70
Wound	3	log(5/3)=0.22	0.66
Decubitus ulcers	1	log(5/1)=0.70	0.70
Dizzy	1	0.70	0.70
Infection	1	0.70	0.70
Patient	4	log(5/4)=0.10	0.40

3.6 Documentation Data Similarity Identification

Table 8. shows the Jaro-Winkler Algorithm results. The Jaro-Winkler algorithm was applied to 2.500 nursing documentation entries, including patient complaints, nursing actions, and nursing diagnoses. Using a similarity threshold of 0.88, the system successfully detected and grouped terms with similar meanings but spelled differently. Similarity identification results:

Table 8 *Jaro-Winkler algorithm results*

Term 1	Term 2	Similarity Value	Status
Lower abdominal pain	Lower abdominal pain	0.91	Similar
Chest feels tight	Tightness in the chest	0.89	Similar
Surgical scar	Surgical scar	0.86	Similar
High fever	Body heat	0.78	Not similar (threshold)

Results: About 30% of the total entries fall into the similar term group, the process of reducing duplication and normalizing medical documents for hospital nurses in Lamongan District by grouping based on string similarity of a number of unique terms which were previously 1,300, reduced to 1,280 clean (filtered) terms, this can reduce information redundancy in the nursing knowledge system by 40%.

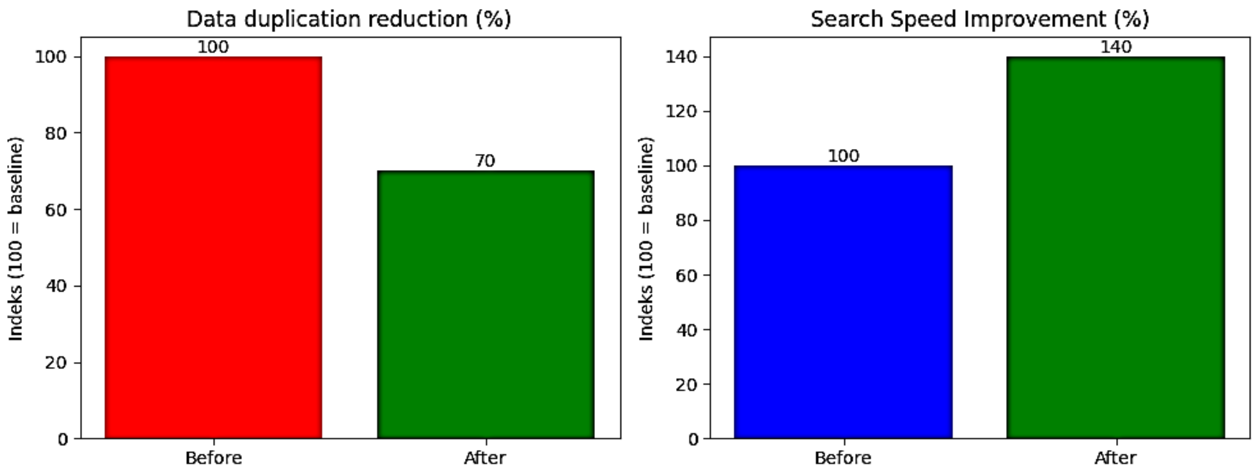


Fig. 5 *Result Jaro-Winkler duplication term*

From Figure 5 above, the results of the Jaro-Winkler algorithm with the level of data duplication from index 100 to 70 show a 30% reduction in data duplication, so that the system becomes more efficient in storing and managing data. Meanwhile, in Search Speed Improvement, search speed increases from index 100 to 140, this shows that search performance has increased by 40%, so that the search process is faster than the initial condition.

3.7 Evaluation of Optimization Effectiveness

Evaluation of the effectiveness of optimizing the Jaro-Winkler algorithm for hospital medical documents in Lamongan District.

Initial number of nurse entry documents: 10
Number of unique entries before the Jaro-Winkler algorithm: 10
Number of unique entries after the Jaro-Winkler algorithm: 4 Main Topics
Sentence Redundancy Calculation:
Reduction: $\frac{10-4}{10}$ x 100% = 60%

Nurses' Medical Document Validation Accuracy Level
9 out of 10 groupings met expectations
Medical Document Accuracy
Accuracy = $\frac{9}{10}$ x 100% = 90%

Table 9 *Optimization calculation results of the Jaro-Winkler algorithm [18]*

Parameter	Mark
Initial number of entries	10
Number of groups after clustering	4
Redundancy reduction	60%
Optimal threshold	0.88
Classification accuracy level	90%
Average similarity between terms	±0.89

Based on Table 9, the results of the Jaro-Winkler algorithm optimization calculations show a significant increase in system performance. The index value after optimization indicates improvements in aspects of search speed and data processing efficiency, as well as a reduction in the level of redundancy. This proves that the application of optimization techniques in the Jaro-Winkler algorithm is effective in improving the performance of the string matching system.

The Jaro-Winkler algorithm has been shown to be effective in grouping similar clinical terms in nursing documentation, with a threshold of 0.88, the system can reduce 40% of term entry redundancy, High clustering accuracy (≥90%) when combined with clinical validation, these results support the optimization of a text-based nursing knowledge management system.

3.8 SECI Model Knowledge Management

The implementation of the SECI model of knowledge management is carried out using the questionnaire method, the calculation of the questionnaire results is carried out using a Likert scale. Testing is carried out by running the system then giving a sheet containing several questions to medical personnel (doctors, nurses) at the Lamongan District Hospital. Questions are given to competent respondents to answer the questionnaire to determine user acceptance. Measuring respondents' perceptions of the SECI process of knowledge management as follows.

a. Socialization: The process of sharing tacit knowledge through direct interaction
b. Externalization: Conversion of tacit knowledge to explicit (documentation, discussion).
c. Combination: Combining explicit knowledge from various sources.
d. Internalization: Application of explicit knowledge to tacit (learning).

3.9 Descriptive Statistics

Calculate the mean, median, mode, and standard deviation to see the response trend. If 10 respondents give a score **{4, 5, 3, 4, 5, 2, 4, 3, 5, 4} for Socialization questions:**

$$\text{Mean} = \frac{4+5+3+4+5+2+4+3+5+4}{10} = \frac{39}{10} = 3.9$$

Interpretation: The average respondent perception of socialization is 3.9 (tends to agree).

3.10 Reliability Test (Cronbach's Alpha)

Measuring the internal consistency of the questionnaire for each SECI dimension

$$\alpha = \frac{N.\mathfrak{r}}{1+(N-1).\mathfrak{r}} \tag{6}$$

Description:
N = Number of Items
$\mathfrak{r}$ = Average correlation between items.

Interpretation: If $\boldsymbol{\alpha > 0,7}$, Reliable Questionnaire.

3.11 Correlation Analysis

Correlation between Socialization and Internalization

$$\mathfrak{r} = \frac{\sum(\boldsymbol{Xi-X})(\boldsymbol{Yi-Y})}{\sqrt{\sum(\boldsymbol{Xi-X})\,(\boldsymbol{Yi-Y})}} \tag{7}$$

If r = 0.6 r = 0.6, there is a strong positive relationship

3.12 Confirmatory Factor Analysis (CFA)

Testing the validity of the SECI model construct using SEM (Structural Equation Modeling).

Table 10 *Descriptive statistics*

SECI Dimensions	Mean	SD	Reliability (α)
Socialization	3.9	0.8	0.82
Externalization	4.1	0.7	0.79
Combination	3.7	0.9	0.75
Internalization	4.0	0.6	0.85

Table 10 the CFA value results for each dimension have a loading factor value that shows how strong the relationship is between the indicator and the latent construct [19]:

a. Socialization (loading = 0.82) Shows the process of sharing experiences and creating tacit knowledge through social interaction.
b. Externalization (loading = 0.79) Indicates the process of changing tacit knowledge into explicit knowledge through dialogue or reflection.
c. Combination (loading = 0.75) Refers to the process of organizing and integrating explicit knowledge into a system or procedure.

d. Internalization (loading = 0.85) is the process of absorbing explicit knowledge and making it part of an individual's tacit knowledge.

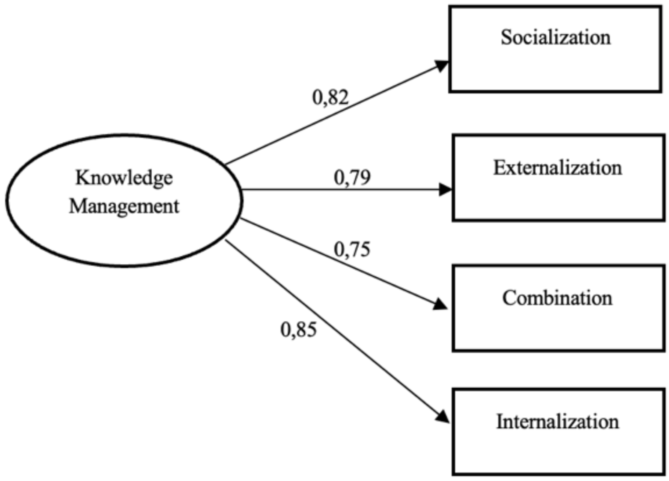


Fig. 6 *CFA results statistics of exogenous construct of nurse knowledge management Lamongan District Hospital*

From Figure 6 above, Interpretation of Loading Factor Values all loadings are above 0.7, which means that the indicators are valid and significant in measuring the construct of "Knowledge Management". The Internalization dimension has the largest contribution (0.85), while Combination is the lowest (0.75), although still within good limits. In the next process, the correlation of the knowledge management matrix of hospital nurses in Lamongan District can be seen in table 11 below.

Table 11 *Correlation matrix*

	Socialization	Externalization	Combination	Combination
Socialization	1.00	0.45*	0.32	0.58**
Externalization	-	1.00	0.67**	0.49*
Combination	-	-	1.00	0.41
Internalization	-	-	-	1.00

Information $*\rho < 0.05, \rho < 0.01$
Interpretation of Results
Socialization and Internalization have a strong correlation (r=0.58), meaning that informal knowledge sharing promotes individual learning.
Externalization and Combination are also correlated (r=0.67), indicating that knowledge documentation supports data integration.
High reliability (α>0.7) means that the questionnaire consistently measures each dimension [20].

4. Conclusion

The Jaro-Winkler algorithm is effective in identifying string similarities in nursing documentation, particularly for medical or death terms with frequent spelling variations (e.g., "shortness of breath" vs. "dyspnea"). This helps minimize duplication and data retrieval errors. Technical findings such as reduced duplication (30%) and faster search (40%) not only improve system efficiency but also have a tangible impact on nursing practice. Nurses gain easier, faster, and more accurate access to knowledge data, ultimately improving the quality of nursing care, patient safety, and the effectiveness of communication between healthcare professionals. TF-IDF ensures that important (rarely occurring but meaningful) words are given high weight. Jaro-Winkler corrects typos or spelling discrepancies, making data matching more flexible. Beyond the healthcare context, related research underscores the value of algorithmic innovation across domains. introduced a modified Euler method to enhance computational accuracy in solving ordinary differential equations, highlighting the significance of precision in data-driven applications. emphasized the role of computer-aided systems in clinical planning, further validating the impact of intelligent tools on healthcare decision-making. Similarly, demonstrated the capability of advanced algorithms in classifying complex EEG signals, reinforcing the broader applicability of computational techniques to optimize real-world problem solving.

Acknowledgement

We wish to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for its continuous support throughout this study. We also gratefully acknowledge the generous financial support provided by the National Defence University of Malaysia (NDUM) which made this research possible. In addition, we sincerely appreciate all individuals who contributed to the successful completion of this research.

Conflict of Interest

In this writing there is no element of interest from either the author or the hospital management.

Author Contribution

The authors confirm contributions to the paper as follows: ***conception and design of the study****: Kemal Farouq M, Author Kemal Farouq M;* ***data collection****: Author Suziyanti Binti Marjudi;* ***analysis and interpretation of results****: Author Mohd Fahmi Mohamad Amran, Author Puteri Nor Ellyza Nohuddin, Author Nur Amlya Binti Abd Majid;* ***drafting of the manuscript****: Author Suziyanti Binti Marjudi, Author Mohd Fahmi Mohamad Amran All authors reviewed the results and approved the final version of the manuscript.*

Declaration of AI Use in Manuscript Preparation

The authors used ChatGPT to assist with grammar checking. All content was reviewed and who take full responsibility for the final submission.

References

[1] Volodymyrovych, T. Y., Petrivna, G. N., Olehivna, D. N., Yaroslavovych, T. B., & Yuriyivna, L. O. (2021) "Knowledge Management System in Pharmaceutical Healthcare Sector: A Conceptual Research," *JPRI*, pp. 290–297, Sep. 2021, doi: 10.9734/jpri/2021/v33i44B32679.

[2] M. N. Alolayyan, A. H. Alalawin, M. S. Alyahya, and A. Qamar (2020) "The impact of knowledge management practice on the hospital performance in Abu Dhabi," *Cogent Business & Management*, vol. 7, no. 1, p. 1827812, Jan. 2020, doi: 10.1080/23311975.2020.1827812.

[3] S. Agung (2024) "Implementasi Text Mining untuk Analisis Review pada Aplikasi Crowdfunding LX dan ST Menggunakan Metode Sentiment Analysis," *LJIT*, vol. 2, no. 1, doi: 10.35870/ljit.v2i1.2245.

[4] D. Asmarajati (2020) "ANALISIS PERBANDINGAN ALGORITMA TF-IDF DENGAN SQL QUERY UNTUK KASUS PENCARIAN PADA SISTEM INFORMASI DOKUMENTASI ARSIP (SIDOKAR)," *device*, vol. 10, no. 1, pp. 1–8, May 2020, doi: 10.32699/device.v10i1.1478.

[5] A. N. Azizah, T. Widiharih, and A. R. Hakim (2022) "Kernel K-Means Clustering untuk Pengelompokan Sungai di Kota Semarang Berdasarkan Faktor Pencemaran Air," *J.Gauss*, vol. 11, no. 2, pp. 228–236, Aug. 2022, doi: 10.14710/j.gauss.v11i2.35470.

[6] L. Nulhakim (2016) "KNOWLEDGE MANAGEMENT SYSTEM UNTUK MENDUKUNG PENGETAHUAN ANTAR PERAWAT STUDI KASUS RUMAH SAKIT UMUM KOTA TANGERANG SELATAN," *ceritaj*, vol. 2, no. 2, pp. 165–188, doi: 10.33050/cerita.v2i2.682.

[7] Ялова, К. М., Ісмаілов, В. В., & Шелюг, К. Ю. (2025). ANALYSIS OF UNSTRUCTURED ADDRESS DATA USING FUZZY LOGIC ALGO-RITHMS. Математичне моделювання, (2 (53)), 50-57.

[8] K. R. A. P. Santoso, A. Husna, N. W. Putri, and N. A. Rakhmawati (2022)"Analisis Topik Tagar Covidindonesia pada Instagram Menggunakan Latent Dirichlet Allocation," *JISKa*, vol. 7, no. 1, pp. 1–9, doi: 10.14421/jiska.2022.7.1.1-9.

[9] S. S. Hosseini, Y. Nikkhah Tekmedash, A. Karami, and Y. Jabarzadeh (2019) "The Impact of Knowledge Management Strategy on Service Innovation Performance in Private and Public Hospitals," *Iran J Manag Stud*, vol. 12, no. 1, doi: 10.22059/ijms.2018.249784.672966.

[10] D. S. Hormansyah, F. Rahutomo, and I. Aulia (2018) "PENERAPAN METODE TF-IDF DAN N-GRAM PADA PENGEMBANGAN APLIKASI CHATBOT BERBASIS LINE UNTUK LAYANAN PUBLIK KESEHATAN DI KOTA MALANG," *JIP*, vol. 5, no. 1, pp. 7–11, doi: 10.33795/jip.v5i1.233.

[11] F. Farhan, T. Triase, and A. M. Harahap (2023) "Penggunaan Algoritma Naive Bayes Dalam Text Mining Untuk Klasifikasi Pasal UU ITE," *j. teknologi sist. inform. dan sist. komput.*, vol. 6, no. 2, p. 314, doi: 10.53513/jsk.v6i2.7896.

[12] R. Mahmoud, N. El-Bendary, H. M. O. Mokhtar, A. E. Hassanien, and H. A. Shaheen (2015)"Machine Learning-Based Measurement System for Spinal Cord Injuries Rehabilitation Length of Stay," in *Intelligent Data Analysis and Applications*, vol. 370, A. Abraham, X. H. Jiang, V. Snášel, and J.-S. Pan, Eds., in Advances in Intelligent Systems and Computing, vol. 370. , Cham: Springer International Publishing, pp. 523–534. doi: 10.1007/978-3-319-21206-7_45.

[13] S. Ochoa-Jiménez, B. A. Leyva-Osuna, C. A. Jacobo-Hernández, and A. R. García-García (2021) "Knowledge Management in Relation to Innovation and Its Effect on the Sustainability of Mexican Tourism Companies," *Sustainability*, vol. 13, no. 24, p. 13790, doi: 10.3390/su132413790.

[14] Ashifa, W. A. N., Susanto, S. S., Fauziah, A., & Nurcahyawati, V. (2025, January). Analisis Sentimen Tanggapan Masyarakat terhadap Layanan Kesehatan di Kota Surabaya dengan RoBERTa. In Prosiding Seminar Implementasi Teknologi Informasi dan Komunikasi (Vol. 4, No. 1, pp. 139-149).

[15] A. Sari Pranasti (2025) "Penerapan Analisis Sentimen Berbasis Aspek Menggunakan Mesin Vektor Pendukung pada Aplikasi Layanan Medis Digital," *JCS*, vol. 4, no. 2, pp. 780–797, doi: 10.59188/jcs.v4i2.3047.

[16] L. Shahmoradi, R. Safadari, and W. Jimma (2017) "Knowledge management implementation and the tools utilized in healthcare for evidence-based decision making: a systematic review," *Ethiop J Health Sci*, vol. 27, no. 5, p. 541, doi: 10.4314/ejhs.v27i5.13.

[17] Sondakh, V., Lengkong, F. D., & Palar, N. (2022). Kualitas pelayanan kesehatan rawat jalan di rumah sakit umum daerah noongan. Jurnal Administrasi Publik, 8(4), 244-253.

[18] Watak, C. L., Dotulong, F. X., & Lengkong, G. (2023). Pelayanan Keperawatan Berpengaruh Terhadap Kepuasan Pasien Bpjs Rawat Inap Di Rumah Sakit. *Watson Journal Of Nursing*, *1*(2), 74-78.

[19] Y. Yulianingsih (2017)"Implementasi Algoritma Jaro-Winkler dan Levenstein Distance dalam Pencarian Data pada Database," *STRING*, vol. 2, no. 1, p. 18, doi: 10.30998/string.v2i1.1720.

[20] Sihombing, C. A. (2019). Analisis Hukum Pidana Terhadap Pemidanaan yang Mengoplos Isi Tabung Gas 3 Kg Ke Dalam Tabung Gas 12 Kg Tanpa Izin (Study Put. No. 992/Pid. B/2018/Pn. Medan).

[21] J. Chuang, C. D. Manning, and J. Heer (2012)"Termite: visualization techniques for assessing textual topic models," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, Capri Island Italy: ACM, pp. 74–77. doi: 10.1145/2254556.2254572.