ICIC Express Letters
Part B: Applications
Volume **15**, Number **6**, June **2024**

ICIC International ©2024 ISSN 2185-2766
pp. **601–610**

DEVELOPMENT OF SECURITY POLICIES ASSESSMENT TOOLS FOR DATA COMMUNICATION IN ACCORDANCE WITH THE INTERNATIONAL STANDARD ON INFORMATION SECURITY MANAGEMENT ISO 27001:2013 USING ONTOLOGICAL CONCEPTS AND TEXT MINING METHODS

Pongsak Sirisom[1], Winai Wongthai[1,2,*] and Janjira Payakpate[1,2]

[1]Department of Computer Science and Information Technology
[2]Research Center for Academic Excellence in Nonlinear Analysis and Optimization
Faculty of Science
Naresuan University
99 Moo 9, Thapo Sub-District, Muang District, Phitsanulok 65000, Thailand
{ pongsaks58; janjirap }@nu.ac.th; *Corresponding author: winaiw@nu.ac.th

Received October 2023; accepted December 2023

Abstract. *This research study presents the development of tools for assessing the similarity of information security management policies of organizations that aim to compare them with the ISO 27001:2013 international standard for information security. The scope of this research focuses on communication security and applies the concepts of ontology and text-mining techniques. The measurements of similarity values within sentences were achieved by assessing the similarity based on word weights, word relationships, and semantic similarity. These techniques were integrated and implemented through the design and testing of a model using RapidMiner Studio software. To measure the effectiveness of the developed model in assessing text similarity, the researchers used a dataset of 300 test sets from the Microsoft Research Paraphrase Corpus Dataset. This dataset was publicly released to support research related to text extraction and semantic similarity evaluation. The evaluation of the developed model included precision, recall, and the F-measure, which yielded a high level.*

Keywords: ISO 27001:2013, Ontology, Similarity comparison on ontology, Text mining

1. **Introduction.** ISO 27001 [1] is an international standard for information security management, focusing on the importance of a management system within organizations. It sets out various requirements that organizations should follow to maintain the security of data, protect business processes, and safeguard critical information assets from threats and risks in various forms. In addition, emergency response plans that may arise are required to reduce losses and maintain the ability to operate the business continuously. It is a security standard that every organization has to uphold [2]. The standard is widely used around the world, such as in Japan [3], European countries [4] and the Middle East [5]. It is the most famous for data security [6,7]. To ensure that organizations achieve the control measures specified in the ISO 27001:2013 standard, it is necessary to engage consultants for the assessment and verification of their own policies and operational measures to determine their compliance with international security control standards. However, this process requires a significant budget [8], which can impose a burden on small to medium-sized organizations in preparing the necessary funds to achieve security standards compliance [4]. With the advancements in computer technology, researchers have been utilizing the framework of ontological engineering, a novel language development technology [9], and text mining [10] technique together to create tools for evaluating data communication security policies according to international standards. Nevertheless, most of the current

DOI: 10.24507/icicelb.15.06.601

research in security ontology is not publicly accessible or available for widespread use. The ability to search and reuse these existing resources is challenging [11]. In this study, the researchers chose a case study within the scope of communications security, as it is a crucial core aspect in business operations and can be considered a fundamental framework that every organization must implement in the present time. The significance of this research is to enable small and medium-sized organizations to easily self-assess their compliance within the desired scope without the need for extensive budget allocation for consultancy services. This flexibility allows for adjustments to be made to control measures based on potential changes in future versions of international standards, without complexity.

This paper is organized into several sections. Section 2 describes related theories. Section 3 describes the architectural design of the tool. Section 4 reports the testing of the tool. Finally, Section 5 concludes the paper and gives recommendations for future research.

2. Background.

In this study, the researcher presented the concept and relevant theory. These are described below.

2.1. International standard for information security management ISO 27001: 2013.

ISO 27001 [1], or the full name ISO/IEC 27001 (Information Security Management System: ISMS), is a standard for managing and maintaining information security. Its objective is to ensure the continuous operation of business processes through various specified requirements. This standard is established by ISO (The International Organization for Standardization) and IEC (The International Electrotechnical Commission). ISO 27001 Application Corporate information from threat risks is an important standard that every organization should have [2]. ISO/IEC27001:2013 is a dynamic system based on the PDCA Model (Plan Do Check Action). It is a universal administrative structure that is used all over the world.

2.2. Ontology.

Ontology [9] is a concept that describes and represents the knowledge domain of interest in a structured and relational manner. It involves defining and formalizing terms and their meanings to describe the desired knowledge domain. Such relational structure can be understood and interpreted by computers through the use of classes, relationships between classes, including class hierarchies, and properties [12]. In this research, the principles of ontology have been applied to the design of the model and the storage of data related to the international standard for information security management, ISO 27001:2013, specifically within the scope of communications security.

2.3. Ontology Web Language (OWL).

OWL is one of the language groups used to represent knowledge, developed by the W3C (World Wide Web Consortium) [13,14], with the aim of facilitating the management and presentation of data in a format of knowledge that is easily understandable by humans. OWL consists of classes that define things to be described. In this research, the OWL language is used to describe ontology and store ontology in the .owl file format.

2.4. Text mining.

Text mining is considered as a part of Artificial Intelligence (AI) and is a technique used to discover patterns in a large amount of text or documents. It is used to represent documents and find values using statistical methods based on predefined rules or relationships. It can be employed to analyze and find similarities among texts within document files. In this research work, the researchers have applied it to analyzing similarities using text similarity and semantic similarity methods [15].

2.5. **Cosine similarity.** Cosine similarity is a method of measuring the similarity between two vectors, determining whether they are pointing in the same direction or not. It is a technique used to compare the similarity of two documents. Each document is represented by an N-dimensional vector, which stores the weight values of each word in the document. The comparison of document similarity is done by examining the cosine angle between the two document vectors. If the two documents are very similar, the vectors of the two documents overlap almost completely at corners, so the angle is small and the resulting cosine coefficients are large [16,17].

2.6. **N-gram.** The concept of n-gram is an idea that applies probabilistic knowledge to creating a model for predicting the next unit based on the previous $n-1$ units. It is widely used in natural language processing and is applied to represent language patterns at various levels, including character level, word level, or sequence of words [17]. An n-gram refers to an n-token. The concept of n-gram can be applied in analyzing and determining the weight values of words appearing in sequence within a text to find the similarity of texts.

2.7. **WordNet.** The WordNet database is a network of words or a database that collects English vocabulary. It consists of words that are grouped together based on their meanings, known as synsets. There are various semantic relationships between synsets, such as synonyms (words with similar meanings) and antonyms (words with opposite meanings). The type of relationship depends on the type of word, including nouns, verbs, adjectives, and adverbs. These components are categorized to accommodate the addition of new words [18].

3. **Architectural Design.** Architectural design involves various steps and methods of operation, as follows.

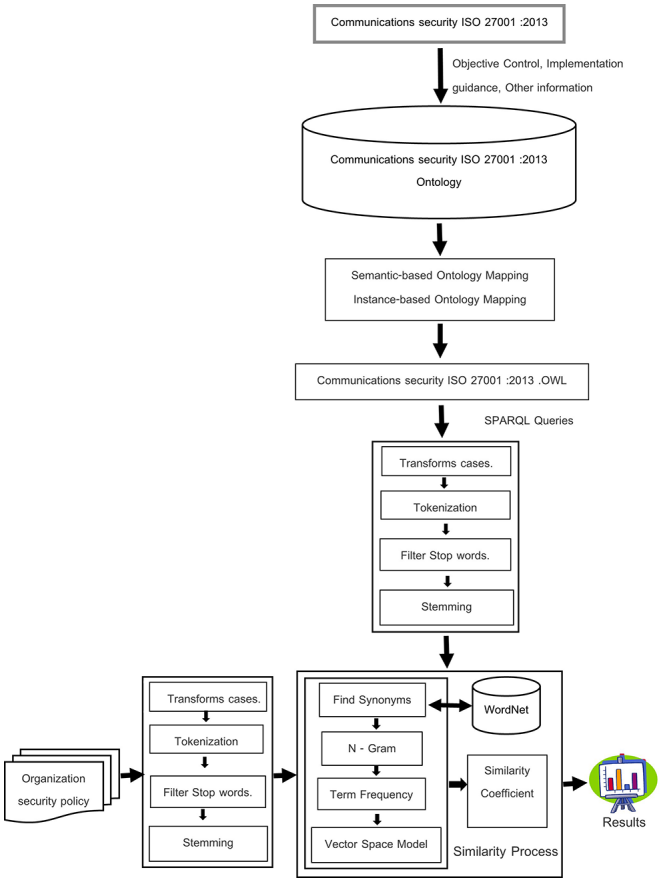

FIGURE 1. Architectural design of the tool

From Figure 1, the architectural design of the tool can be described as follows.

3.1. **Ontology construction.** The ISO 27001:2013 ontology control framework for communications security, which is part of the international standard for Information Security Management System (ISMS) ISO 27001:2013, was constructed (Figure 2).

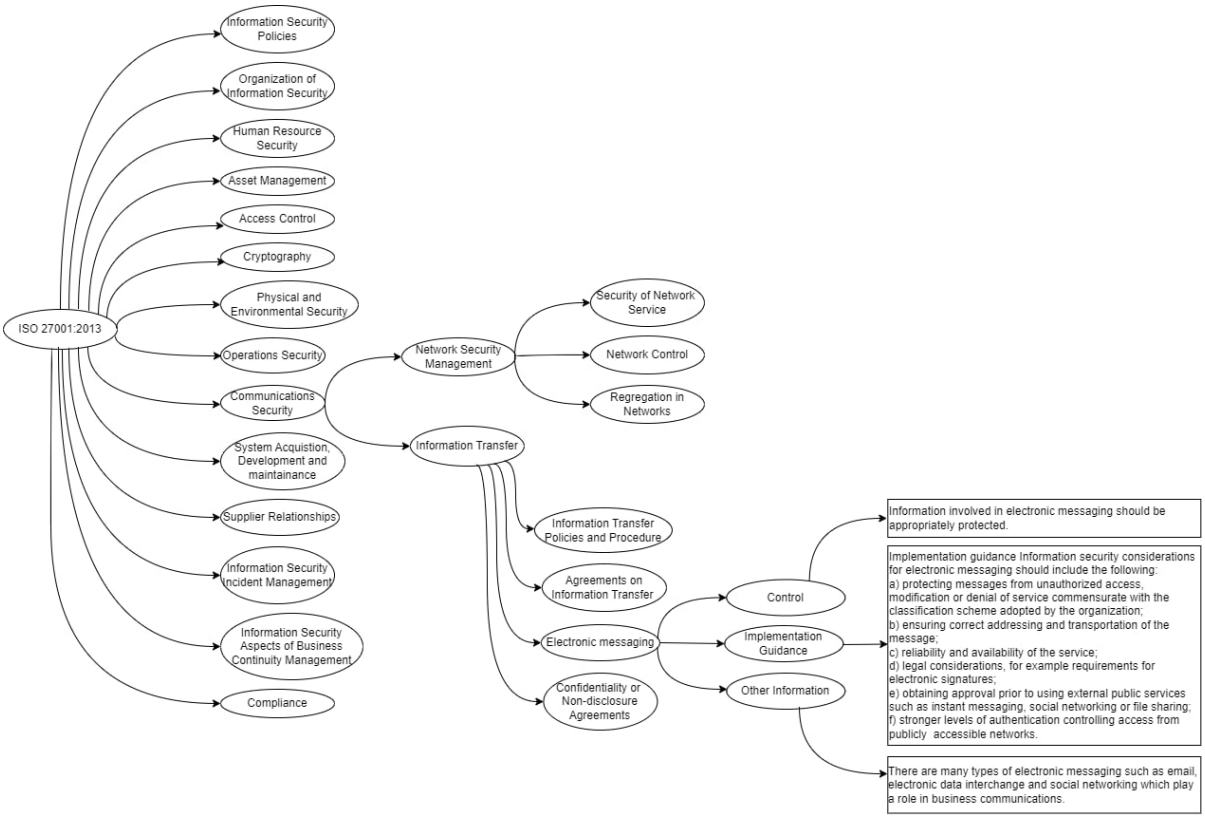

FIGURE 2. The ontology construction

From Figure 2, the Network Security Management class and the Information Transfer class consist of properties that are linked to the properties of various subclasses, as shown in Table 1.

TABLE 1. Properties linked to properties of subclasses

Class name	**Property name**	**Property type**	**Description**
Network Security Management	Has Security of Network Service	Object property	Network service stability
	Has Network Controls	Object property	With network control
	Has Segregation in Networks	Object property	There is network separation.
Information Transfer	Has Information Transfer Policies and Procedures	Object property	There are policies and procedures for data transfer.
	Has Agreements on Information Transfer	Object property	There are policies and procedures for data transfer.
	Has Electronic Messaging	Object property	Have electronic messages
	Has Confidentiality or Non-disclosure Agreements	Object property	Have a confidentiality or non-disclosure agreement

From the ontology diagram, the researchers created an ontology knowledge base using the HOZO Ontology Editor software [19] for design and conversion into an OWL file format.

3.2. **Pre-processing: preparing the text before processing.** This step involves preparing the text before it is analyzed and processed using text mining techniques. The process includes the following steps.

- Transform Case: convert English text in sentences to all lowercase.
- Tokenize: divide the text of a sentence into words. Delete numbers and special characters or punctuation marks that appear in a block of text.
- Filter Stop Words (English): filter stop words that are not important in expressing the meaning in a sentence, such as a, an, the, also, just, quite, unless, and in.
- Stem (Porter): perform word stemming to reduce words to their root form. In this research, the researchers chose to use Porter Stemming, which is a popular algorithm for reducing word variations and transforming them into their root form [20].

3.3. **Processing text similarity.** To process text similarity, the steps are as follows.

3.3.1. *Semantic similarity measurement.* Semantic similarity measurement is a method of quantifying the resemblance between words by leveraging the semantic relationships obtained from the WordNet lexical database. It aimed to determine the degree of semantic similarity between sentences by using semantic vectors generated from the similarity values of individual words within the sentence. In this research, the focus was on words that differed but have similar meanings (synonyms). The process involved creating vectors for words with similar meanings, representing the word weights in vector form, and then calculating the similarity value. The maximum recursion depth parameter was set to 5, as selecting a large number of words would diminish the weight of the key words used to construct the matrix for comparison.

3.3.2. *Measuring similarity from the relationship of words in sentences.* Similarity measurement utilizes the relationships between words within a sentence, such as the position or order of the words. The method involves transforming the sentence to be measured into a vector representation based on the word positions or order. These vectors are then compared against a set of combined word vectors generated from the two sentences. In this research, the words are ordered using the N-gram approach, which is widely used in natural language processing and employed to represent language patterns at various levels, including character level, word level, or word sequence level. Specifically, the words are ordered using a 2-gram approach [21].

3.3.3. *Word similarity measurement.* Measuring the similarity of words is a fundamental method widely used for assessing text similarity. In this research, the word similarity is measured using a word overlap similarity measure, which calculates and compares the number of shared words between two texts with the total number of words in both texts [20].

3.3.4. *Measurement of similarity from the weight of words.* Assigning weights to words in a text is based on the fundamental concept that not all words in a text are equally important. Words with higher frequencies appearing in relevant documents are considered more significant than those with lower frequencies. In designing and constructing this model, the researcher chose to use the Term Frequency (TF) as a measure for calculating the weights [20]. This is because TF compares the occurrence of the term of interest in the document. A higher TF value indicates that the term of interest appears more frequently in the document. The TF value can be calculated using Equation (1) [20].

$$TF = \frac{number\ of\ words\ that\ appear\ in\ the\ document}{total\ number\ of\ words\ in\ the\ document} \tag{1}$$

The weight of the word is then represented in vector form, and the similarity value is calculated with the equation for finding the coefficient of similarity.

3.3.5. *Coefficient of similarity.* In this research, we determined the similarity of text with the cosine similarity coefficient, calculated from Equation (2) [22,23].

$$\text{Similarity}_{\text{cosine}} = \cos\theta = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

Examples: the following 2 messages.

S1: Information involved in electronic messaging should be appropriately protected.

S2: Electronic messaging information should be adequately protected.

From text sentence examples as the text was prepared before processing, it was found that there were 4 out of 7 words that appeared in both documents and could be converted into vector space as follows.

TABLE 2. Text space vector before processing

Sentence	Adequ	Appropri	Electron	Inform	Involv	Messag	Protect
S1	0	1	1	1	1	1	1
S2	1	0	1	1	0	1	1

From the example vector space above, the cosine similarity coefficient could be calculated as follows:

$$\text{Similarity}_{\text{cosine}} = \frac{4}{\sqrt{6} \times \sqrt{5}} = \frac{4}{2.449 \times 2.236} = 0.730$$

4. **Model Testing.** From the architectural design discussed earlier, the researcher has created a model by combining various similarity measurement techniques mentioned above. The RapidMiner Studio software [24] was used for this purpose. An example of the program usage and the resulting outputs are shown in Figures 3 and 4, respectively.

The performance of the model could be evaluated by setting baseline values for assessing its effectiveness. This involved comparing the calculated similarity scores from the

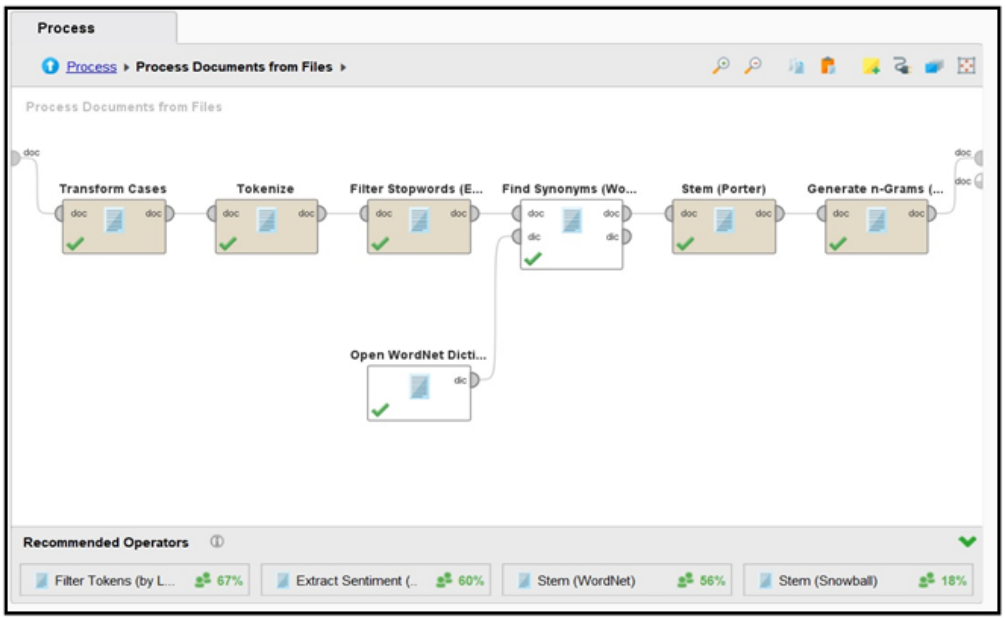

FIGURE 3. Modeled with RapidMiner Studio

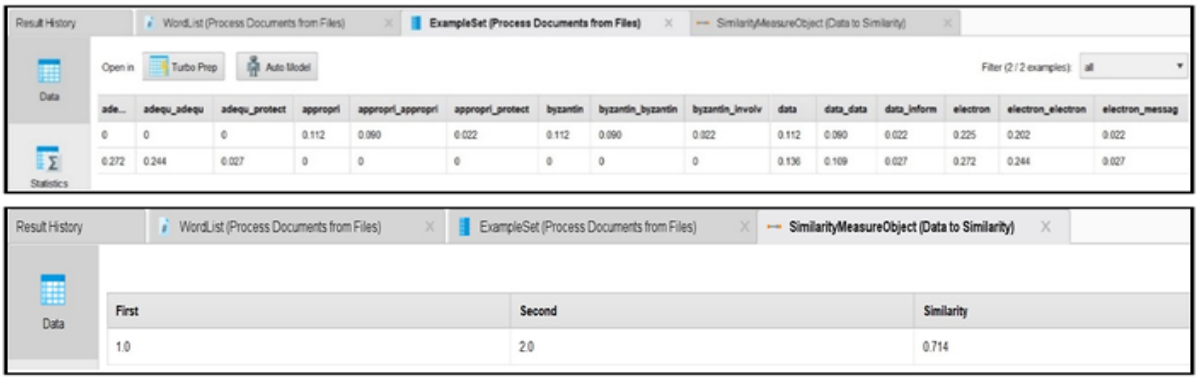

FIGURE 4. Model results

developed tool with the predefined similarity scores of the test texts, which were known to be similar. The indicators were the following:

1) True Positive (TP): The correct number of samples of text that the model could process as similar;

2) False Positive (FP): The number of samples of text that the model processes as similar, but not similar;

3) True Negative (TN): The correct number of samples of text that the model processes as not having similarities;

4) False Negative (FN): The number of samples of text that the model processed as having no resemblance, but rather text with similarity.

The above four indicators could be used to calculate the criteria for evaluating the system's performance as follows [25].

Precision: The ratio between the text similarity obtained from the developed tool and the processed similarity value would represent the accuracy of the similarity calculation. This can be calculated using Equation (3).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Recall: The ratio between the correctly measured text similarity and the total correctly measured similarity would represent the coverage of the similarity calculation. This can be expressed as Equation (4).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

F-measure: A method for calculating the accuracy and completeness values by using the weighted harmonic mean of accuracy and completeness. It can be calculated using Equation (5).

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Then, the researchers conducted model testing using a dataset of 300 pairs of sentences from the file "msr_paraphrase_test.txt". This dataset is part of the research work called Microsoft Research Paraphrase Corpus [26] which is a collection of paraphrased sentences. The conformity between the test data and the model's test results was determined as follows:

1) The test data set had a quality value of 0 and the test results from the model of < 0.5 are considered similar (assign a similarity value = 1);

2) The test data set had a quality value of 1, and the test results from the model are < 0.5, assume no similarity (assign a similarity value = 0);

3) The test data set had a quality value of 0, and the test results from the model had a value of ≥ 0.5, assume no similarity (assign a similarity value = 0);

4) The test data set had a quality value of 1, and the test results from the model had a value of ≥ 0.5, assume similarity (assign a similarity value = 1).

Examples of model testing with the test dataset are shown in Table 3.

Table 3. Samples of test data and test results

Sentence	Quality	Results	Similarity
S1: The company did not detail the costs of the replacement and repairs. S2: But company officials expect the costs of the replacement work to run into the millions of dollars.	0	0.338	1
S1: I am proud that I stood against Richard Nixon, not with him, Kerry said. S2: I marched in the streets against Richard Nixon and the Vietnam War, she said.	0	0.571	0
S1: Ballmer has been vocal in the past warning that Linux is a threat to Microsoft. S2: In the memo, Ballmer reiterated the open-source threat to Microsoft.	0	0.236	1

The results of the model testing can be summarized as follows.

1) The number of test text samples correctly identified as having similarity (True Positive, TP) by the model was 246.

2) The number of test text samples incorrectly identified as having similarity (False Positive, FP) by the model was 39.

3) The number of test text samples correctly identified as not having similarity (True Negative, TN) by the model was 118.

4) The number of test text examples incorrectly identified as not having similarity but actually having similarity (False Negative, FN) by the model was 15.

When the data obtained from the test was used to calculate the F-measure, the following results were obtained:

$$\text{Precision} = \frac{246}{246+39} = 0.863, \quad \text{Recall} = \frac{246}{246+15} = 0.943$$

$$\text{F-measure} = 2 \times \left(\frac{0.863 \times 0.943}{0.863 + 0.943} \right) = 0.901$$

From the F-measure, it shows that the developed model is efficient.

5. **Conclusion.** This research applies the framework of ontology-based technology and text mining methods, utilizing techniques such as measuring word similarity, measuring similarity based on word weights, measuring similarity based on word order in a sentence, and measuring semantic similarity. These techniques are integrated and used together in the research. The researchers designed the model using RapidMiner Studio, a popular tool for text mining. Cosine similarity coefficients were determined to assess the similarity between texts. The measurement of similarity was based on the word weights, where the researchers used the Term Frequency (TF) to calculate the weights. Since this involved comparing the text of two sentences while creating an index for comparison based on the concept of a bag of words disregarding the word order in the text, it may result in different meanings for the sentences even if they contain the same words. Therefore, one important consideration was the size of the n-grams that should be used to ensure appropriateness. The researchers developed a program to determine the size of the n-grams that could be matched together and found that the 1-gram and 2-gram word sizes had the highest percentage of matches. As for the use of different but semantically similar

words (synonyms) from WordNet to capture similar meanings, it was found that selecting a large number of words might result in a reduction of the weight assigned to the main word used for constructing the metric for comparison. Therefore, we assigned a maximum depth of only 5 levels. Then, a test dataset consisting of 300 pairs of sentences was used from the Microsoft Research Paraphrase Corpus Dataset, which has been publicly released to support research purposes related to text understanding and semantic similarity evaluation. The testing process to find precision, recall and F-measure value found that they were at a high level. If there is a filtering process that focuses on words with similar meanings within the same category or the creation of a specialized vocabulary database for comparative purposes, in addition to using the WordNet database alone, it will greatly enhance the performance of semantic comparisons. This serves as a direction for future research endeavors.

REFERENCES

[1] *An Introduction to ISO 27001 (ISO27001)*, http://www.27000.org/iso-27001.htm, Accessed in May, 2023.

[2] Z. Lovric, Model of simplified implementation of PCI DSS by using ISO 27001 standard, *Central European Conference on Information and Intelligent Systems*, 2012.

[3] K. Yasuko, *Information Security Measures Benchmark (ISM-Benchmark)*, IT Security Center, Information-technology Promotion Agency (IPA), Japan, 2007.

[4] N. K. Sharma and K. D. Prabir, Effectiveness of ISO 27001, as an information security management system: An analytical study of financial aspects, *Far East Journal of Psychology and Business*, vol.9, no.3, 2012.

[5] B. AbuSaad, F. Saeed, K. S. Alghathbar and B. Khan, Implementation of ISO 27001 in Saudi Arabia – Obstacles, motivations, outcomes, and lessons learned, *The 9th Australian Information Security Management Conference*, 2011.

[6] C. Giovanna, N. Guido, P. Matteo and S. Marco, The ISO/IEC 27001 information security management standard: Literature review and theory-based research agenda, *The TQM Journal*, vol.33, no.7, pp.76-105, 2021.

[7] G. Daniel, K. Christos, M. Haralambos and M. Saeed, Approaches to develop and implement ISO/IEC 27001 standard – Information security management systems: A systematic literature review, *International Journal on Advances in Software*, vol.12, nos.3&4, 2019.

[8] H. Carol, W. Tawei and L. Ang, The impact of ISO 27001 certification on firm performance, *Hawaii International Conference on System Sciences (HICSS)*, 2016.

[9] M. Uschold and M. Gruninger, Ontologies: Principles, methods and applications, *The Knowledge Engineering Review*, vol.11, no.2, 1996.

[10] B. Richardson and A. Wicaksana, Comparison of IndoBERT-lite and RoBERTa in text mining for Indonesian language question answering application, *International Journal of Innovative Computing, Information and Control*, vol.18, no.6, pp.1719-1734, 2022.

[11] Í. Oliveira, M. Fumagalli, T. P. Sales and G. Guizzardi, How FAIR are security core ontologies? A systematic mapping study, in *Research Challenges in Information Science. RCIS 2021. Lecture Notes in Business Information Processing*, S. Cherfi, A. Perini and S. Nurcan (eds.), Cham, Springer, 2021.

[12] G.-Q. Zhang, S. S. Sahoo and S. D. Lhatoo, From classification to epilepsy ontology and informatics, *Epilepsia*, vol.53, no.2, pp.28-32, DOI: 10.1111/j.1528-1167.2012.03556.x, 2012.

[13] *OWL Web Ontology Language Overview*, https://www.w3.org/TR/owl-features, Accessed in May, 2023.

[14] P. Jorge, A. Marcelo and G. Claudio, Semantics and complexity of SPARQL, *ACM Transactions on Database Systems (TODS)*, vol.34, no.3, 2009.

[15] Y. Li, Z. Bandar, D. McLean and J. O'Shea, A method for measuring sentence similarity and its application to conversational agents, *The 17th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*, FL, USA, 2004.

[16] Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea and K. Crockett, Sentence similarity based on semantic nets and corpus statistics, *IEEE Transactions on Knowledge and Data Engineering*, vol.18, 2006.

[17] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall PTR, Upper Saddle River, NJ, United States, 2009.

[18] M. George, A. B. Richard, F. Christiane, G. Derek and M. Katherine J, Introduction to WordNet: An on-line lexical database, *International Journal of Lexicography*, vol.3, pp.235-244, 1995.
[19] *Hozo – Ontology Editor*, https://www.hozo.jp, Accessed in May, 2023.
[20] A. Piyadanai and P. W. Alex, Seq3uential model-based optimization for natural language processing data pipeline selection and optimization, *Intelligent Information and Database Systems*, pp.303-313, 2021.
[21] G. Aditya, D. Vikrant, K. Shrikant and B. Laxmi, Detecting hate speech and offensive language on Twitter using machine learning: An n-gram and TFIDF based approach, *IEEE International Advance Computing Conference*, 2018.
[22] M. Gawich, A. Badr, H. Ismael and A. Hegazy, Alternative approaches for ontology matching, *International Journal of Computer Applications*, vol.49, no.18, pp.29-37, 2012.
[23] M. K. Vijaymeena and K. Kavitha, A survey on similarity measures in text mining, *Machine Learning and Applications: An International Journal (MLAIJ)*, vol.3, no.1, 2016.
[24] *RapidMiner Documentation*, https://docs.rapidminer.com/10.0/studio/, Accessed in May, 2023.
[25] F. Samuel and S. Mark, *A Semantic Similarity Approach to Paraphrase Detection*, Department of Computer Science, University of Sheffield, S1 4DP, UK, 2008.
[26] B. William, D. Bill, B. Chris and Q. Chris, *Microsoft Research Paraphrase Corpus*, Microsoft Research March 2, 2005.