ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20231011001013
Original scientific paper

An Empirical Study on Document Similarity Comparison Evaluation Between Machine Learning Techniques and Human Experts

Won-Jung JANG

Abstract: Current machine-learning training focuses solely on accuracy. In this study, the weights of other dimensions were examined rather than measuring only the accuracy of machine learning. By comparatively analyzing the decision-making of machine learning and humans in various fields, this study examines how well organizational vision is propagated to lower levels of the organization. Also, the results evaluated by humans and machine learning models were comparatively analyzed from multiple perspectives. As numerical representation methods of words, count-based models (Bag of Words, *TF-IDF*), artificial neural network (ANN) models (Word2Vec, GloVe), and a vision propagation measurement (VPMS) model combining two methods were used to calculate the similarity between documents, which are comparatively analyzed with the actual results measured by an expert group. The findings of this study can be used as an evaluation metric for how effectively the vision of the upper organization is being disseminated to the lower-level organizations. Additionally, it could be utilized in developing algorithms such as customer segmentation for target marketing using text data. The study makes two key contributions - (i) providing an extensive empirical comparison of document similarity analysis by different ML techniques versus human experts, and (ii) proposing a new VPMS model that outperforms existing methods.

Keywords: ANN model; count-based model; document similarity; ensemble learning model; machine learning

1 INTRODUCTION

In the era of the 4IR (Fourth Industrial Revolution), personalized customer experience through big data-based value enhancement can have a significant impact on corporate profits. Data generation is expected to increase significantly owing to advancements in communication technology, such as the Internet and smartphones. According to an IDC report, 16.1 ZB (zettabyte) of digital data generated in 2016 will increase by 10 times in less than 10 years, while the rate of increase will also go up [9]. The Ministry of Science and ICT of South Korea reported that the data market size was 14 trillion won in 2018 and will expand to 30 trillion won by 2023 based on relevant policies and programs that foster global competitiveness in related industries, such as data and artificial intelligence (AI). The enormous net worth of Facebook is attributable to the large and unique data assets of the company [29]. Most companies are focusing on strengthening their AI-based products and service competitiveness, and securing big data based on AI has become crucial to improving competitiveness. In particular, increasing the competitiveness of a company using unstructured data is important because of the large volumes of data generated. From the technological advances in personal computers (PCs) in the 1990s to those in the Internet (2000) and smartphones (2007), digitized information and objects have become interconnected. In the digital era, the success of a company is determined by the efficiency of these connections. Companies such as Yahoo and Google are typical examples of these success stories. Therefore, searching for accurate information easily is critical for the survival of companies in the digital era. Finding an entity that quickly and easily provides information with a high level of similarity is an important issue. Because individuals communicate and connect with each other using documents, which are typical text data and may also be used for important decision-making, research based on documents is emerging as a significant research topic. Studies on developing algorithms to quickly and easily find accurate documents have been actively conducted, where document similarity has been used as a crucial indicator to find accurate documents. It has been confirmed that the automated approach of text similarity search can improve the search results for the previous patent technology related to quality, in addition to accelerating the search process [7]. Carvallo et al. 2020 [3], in their study on the automatic document screening of medical literature, demonstrated that the performance of an artificial neural network (ANN) (BioBERT, BERT, Word2Vec, and GloVe) models is better than that of the term frequency-inverse document frequency (*TF-IDF*) model. Mell et al. 2021 [24] used an expert model and machine learning hybrid approach to predict the human-computer negotiation results using various data; their approach is expected to benefit people who want to combine broad domain knowledge with a more automated approach in human-computer negotiation. In the era of the 4IR (Fourth Industrial Revolution), companies can grow sustainably if they continuously provide new experiences to customers with personalized products and services [11]. Companies should not merely pursue accuracy when providing personalized services. The coexistence of humans and AI needs to be considered, while the similarity between the decisions of the two entities needs to be understood in the context of cooperation between humans and AI. However, current training in machine learning focuses solely on accuracy. This study examined the weights of other dimensions rather than merely pursuing the accuracy of machine learning. In other words, it aims to examine how similar the decision-making results of humans and machines are so as to foster better cooperation and understanding between them. To this end, an empirical study using Korean text processing was conducted to examine how well OO University's founding ethos, educational goals, and desired human character in the 2018 Curriculum Handbook are reflected in the educational goals and desired human character of 44 academic departments. As the count-based numerical representation of words, a bag-of-words (BoW) model, which is a data quantification method of word frequencies, and a *TF-IDF* weight model, which uses weights to indicate the importance of a certain word in the

document were used. As ANN-based numerical representation methods of words, we have used a Word2Vec model that can reflect the meaning of words in a given window and a GloVe model that is trained based on the entire corpus. Count-based numerical representation methods of words cannot reflect the meaning of surrounding words, and ANN-based methods are not suitable for determining the importance of a certain word in a document. The VPMS model mentions keywords and reflects the meanings of words around the keywords. The machine learning method uses the BoW, *TF-IDF*, Word2Vec, GloVe, and VPMS models to calculate document similarity. Furthermore, a group of experts was put in place by selecting 12 persons among the full-time professors of Catholic Kwandong University; these experts performed a comparative evaluation of the vision document of the upper organization and the execution documents of the lower organizations. For the evaluation of 44 departments, four teams with three persons per team were formed to conduct the actual measurements. In this study, the document similarity calculated using machine learning and expert evaluation results were comparatively analyzed. Based on this, an analysis to determine the similarity between the results of machine learning and the results of human judgments when the accuracy pursued in conventional studies is high.

1. Whether the document similarities computed by trained models agree with the expert evaluation results when the former is high was investigated. In other words, the correspondence of high accuracy to human judgments was evaluated.
2. The model that shows the closest agreement with humans by comparing the document similarities computed by different models with expert evaluation results was selected.
3. The evaluation results of both the models and the experts for each of the evaluation classes - excellent, good, fair, and poor, was cross-validated.

Furthermore, the correspondence between trained models and expert evaluation was examined from multiple perspectives, as explained above, and the validity of the models was assessed.

2 BACKGROUND AND RELATED WORK

2.1 Similarity Distance Functions

The distance between documents is measured in reduced dimensions, and the cosine similarity method is often used when reviewing the similarity between documents [29]. The cosine similarity can be computed as shown in Eq. (1).

$$Cosine = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{1}$$

Cosine distance can be calculated using the equation, 1 - cosine similarity.

$$d_{cosine} = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{2}$$

The Euclidean distance is calculated using the square of the distance in each dimension.

$$d_{Euclidean} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

2.2 Work Representation Classification

Word representation methods can be classified into discrete and distributed representations. Discrete representation, also referred to as local representation, is a word representation method in which word frequency is counted for quantification, whereas distributed representation, also referred to as continuous representation, quantifies word representation by referring to the context words of the center word [10]. Word representation, the mathematical object of words, is represented using vectors [33]. A one-hot vector and an embedding vector are used to represent words in quantities. A one-hot vector was generated through a one-hot encoding process. An embedding vector represents words based on the value set by the user and is also referred to as a dense vector because the vector dimension becomes dense. Word embedding refers to the use of dense vectors in the word representation method.

2.3 Count Based Word Representation

There are different ways to represent unstructured text data in natural language processing (NLP). Count-based word representation methods include the BoW model, one-hot encoding, and the *TF-IDF* weight model. These representation methods have the disadvantages of an increased vector dimension as the number of words increases, in addition to the inability to calculate or represent the similarity between the center word and context words. One-hot encoding is a type of vector representation in which 1 is assigned to the index of the words being represented and 0 is assigned to all other indices, and the size of the word set is set as the size of the vector. These representation vectors are called one-hot vectors. The process of one-hot encoding involves assigning a unique index to the words in the document and assigning 1 to the index position of words being represented or 0 to the index position of other words. The BoW model is a quantification method of text data that focuses on the frequency of words regardless of the order of words, and is the most common model for the mathematical representation of documents [6]. The BoW method was proposed in the problem of text search domain for text document analysis, and later applied in computer vision application programs [2]. The relationship between query and document can be identified through the frequency of specific words in a document based on the BoW representation [34]. The set of paired words in which the words in the document are matched with the frequency is defined as the representation of the document. Each vocabulary representing a document is a feature of the document, and the frequency of the vocabulary becomes the value of the feature. The order of words appearing in a document does not preserve the information; hence, two documents are assumed to match based on the same vocabulary and frequency distribution. Chayangkoon and Srivihok 2021 extracted text features using a combination of BoW and Word2Vec models to supplement the drawbacks of the BoW model in a text classification model.

In a study on motion recognition algorithms, Huang et al. 2021 [8] confirmed that the BoW model's recognition rate and robustness can outperform conventional motion recognition algorithms. The BoW process involves assigning a unique index to each word and creating vectors in which the frequency of a token appearing in each index position is recorded. *TF-IDF* shows the importance of specific words in a document when there are corpora consisting of multiple documents in text mining [20, 30]. *TF-IDF* has been used in theoretical fields for the automation of information retrieval and automatic classification of documents [20]. The *TF-IDF* weight can be used to extract keywords from a document, determine the ranking of search results, and measure the similarity of documents [36]. Term frequency (*TF*) is a value that indicates how frequently a word appears in a document; thus, a word is considered more important if it has a higher *TF* value. A *TF* value needs to be normalized to compensate for the difference in document lengths, where the frequency of a word appearing in the document is divided by the total number of words in the document [19, 29]. Tab. 1 shows the calculation formula for the *TF-IDF* weight model.

Table 1 TF-*IDF* weight model

TF value	$TF_{(t,d)} = 0.5 \cdot \frac{0.5 \cdot f(t,d)}{\max f(w,d) : w \in d}$
IDF value	$IDF(t,D) = \log \frac{\lvert D \rvert}{\lvert d \in D : t \in d \rvert}$ $\lvert D \rvert$: Total number of documents included in the set of documents $\lvert d \in D : t \in d \rvert$: Number of documents containing letter 't'
TF-IDF weight	$TFIDF(t,d,D) = TF_{(t,d)} \cdot IDF_{(t,D)}$

The *IDF* (inverse document frequency) value indicates that the lower the number of documents in which a word appears out of the total documents, the more significant its meaning is [31]. The *TF* value represents the word frequency in a specific document, while the *IDF* value represents the number of documents in which the word occurs. The *TF-IDF* value is the product of the *TF* and *IDF* values. Based on the *TF-IDF* weight, a word that appears frequently in a document and is present in a fewer number of documents among the entire document set can be evaluated as an important word [19].

2.4 Word Embedding Based Word Representation

A neural network corresponds to a distributed representation. A distributed representation method starts with a premise of the distributional hypothesis [17, 35]. Word embedding-based word representation methods include the Word2vec and GloVe models. The Word2vec model reflects the meaning of the context words of the center word, whereas the GloVe model reflects the meaning between words using the entire corpora. However, this may be insufficient for identifying the extent of the importance of certain words in a document. Moreover, a specific model is required to extract keywords from a document and reflect the meaning of the context words of keywords. A distributed representation method was created based on the assumption of a distributional hypothesis, where the words appearing around the center word are assumed to have similar meanings. Word2vec is an ANN algorithm that utilizes the distributional hypothesis to represent words as vectors in a vector space, where vectors of words with similar meanings are expected to be located closer to each other [25]. Furthermore, Word2vec is a neural network-based algorithm that projects a word into a characteristic vector of m-dimension [25]. Skip-gram is a model that trains the word vector to maximize the probability of combining the context word with an m-dimensional vector through the window size set based on the center word, which is the input value. Word2vec, a model for NLP, was developed by introducing an ANN using a recurrent neural network (RNN) [26]. Word2vec is a neural network-based algorithm that can be implemented through a continuous bag of words (CBOW) that predicts words belonging to empty spots based on context words or skip-gram, which finds the words to be input in the surroundings based on the given words [27]. Word2vec is trained through the text that was input into two layers, and the neural network was trained with other words appearing around a specific word. The Skip-gram model aims to predict surrounding words given a center word, while the CBOW model aims to predict the missing word given surrounding words as the training goal [27]. With regard to the m-dimensional vector of Word2vec, which is used to identify the correlation between words in a sentence, users decide the extent to which neighboring words are considered in the process of identifying the correlation between words, which is the window size. The global vector (GloVe) model is an algorithm trained based on the entire corpora, rather than training only within the designated window, such as Word2vec [28]. The GloVe model ensures that the inner product of two words embedded in a specific context is equal to the ratio of their co-occurrence probabilities. It is a log-bilinear model that includes a weighted least square number. The basic hypothesis of the GloVe model is based on the observation that the ratio of co-occurrence probability of two words can encode the meaning of a certain form. Fig. 1 shows an example of an inner product value between 'ice' and 'steam' vectors when the context word 'solid' is provided. Here, when k = solid is provided, the probability of the word ice appearing is 0.00019, and that of the steam is 0.000022 with regard to the entire corpora. The ratio ($P(k|\text{ice})/P(k|\text{steam})$) of these two cases was 8.9. The word 'solid' means being hard and therefore is assumed to appear more frequently with 'ice' than with 'steam'.

Probability and Ration	k=solid	k=gas	k=water	k=fashion
P(k|ice)	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
P(k|steam)	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
P(k|ice) / P(k|steam)	8.9	8.5×10^{-2}	1.36	0.96

Figure 1 Example of inner product value between ice and steam vectors of the context word 'solid'

3 MATERIALS AND METHODS

3.1 Machine Learning Analysis

Prior to the era of big data, data analysis generally involved a statistical or mathematical algorithm-based

approach. These analytical methods use samples and statistics owing to the limitation of information systems that process populations to identify parameters, fundamental constraints of populations, and economic efficiency [36]. Because a large amount of data or big data can now be analyzed in real-time, which was not previously possible, processing technology and securing high-quality big data for training AI have become important [12, 16]. Big data analysis can analyze both structured and unstructured data using advanced analysis methods and support for information systems that target entire populations. A company's data analysis skills are further improved as the processes of planning big data, building infrastructures, performing analytical projects, and promoting analytical skills across the company are repeated [11]. The amount of digital data, both structured and unstructured, such as images, voice, and text, is increasing at an exponential rate [40]. Unstructured data accounts for 92% of the entire data, whereas structured data accounts for only 8%; thus, data analysis using unstructured data is becoming increasingly important [9]. Machine learning is a technology used to find optimized parameters through data learning and is one of the ways to implement AI in commercial applications. Machine learning classifies learning methods based on the method of obtaining learning data [18, 38, 39]. Supervised learning ensures that the input data is as close to the expected output data as possible, and has output values that correspond to the input values of the data. Unsupervised learning analyzes the inherent meaning of input data and has only input values as the data. Semi-supervised learning constitutes learning data by training unprocessed data to supplement output values to correspond to input values when output values are insufficient in supervised learning. In reinforcement learning, behaviors are decided and executed according to the characteristics of input data, execution results are evaluated, and rewards are given based on the evaluation. The algorithm trained by inputting the training data evaluated the performance and overfitting of the model using the validity data. Then, the final algorithm was evaluated using the test data and generalized to develop the model. McCulloch and Pitts [23] proposed the theory of an ANN that can implement arithmetic, sign, and logical operation functions by simplifying the theory of a biological neural network, and further defined it as a threshold logic unit (TLU). Rosenblatt proposed the perceptron theory based on the TLU theory of McCulloch and Pitts and Hebb's study on the neurobehavioral model [32]. The perceptron is the first algorithm to explain the theory of an ANN [15]. The perceptron model consists of a single-layer perceptron (input layer, output layer) and a multi-layer perceptron (input layer, one or more hidden layers, output layer). If there are two hidden layers, it can be referred to as a two-layer perceptron. In general, an ANN is considered a multi-layer perceptron in which information flows from an input layer to a hidden layer and then to an output layer. Furthermore, an ANN is called a feedforward neural network when information is delivered in the forward direction, resembling a biological nervous system [15]. A hidden layer linearly sums the data delivered from the previous layer considering the weight, and then performs feedforwarding so that the summed value is applied to the activation function to be delivered as a neuron of the next layer. The activation function used in the hidden layer serves the purpose of transforming the aggregated signals from neurons into a more discriminative state for the next layer [15]. An RNN is a type of machine learning that has a recursive structure or neural network-based deep learning. Deep learning differs from conventional machine learning in two major aspects: learning is performed by each layer, and an autoencoder, which is an information compressor, is used [22]. An RNN uses the weight of a neural network as the set of inputs in the subsequent neural network, using the output values that have time continuity. A long short-term memory (LSTM) model overcomes the drawback of an RNN, wherein learning becomes difficult because of an increased vanishing gradient as the hidden layer becomes deeper. In the LSTM model, each unit of the middle layer neural network is substituted with memory and consists of three gates: input, forget, and output gates. In a study to predict the average speed of vehicles in Beijing, Ma et al. (2015) [21] proved that the LSTM model produced better results than the RNN model when extracting the modeling information of long-time-step. In a study on autonomous driving through driving mode recognition, Kim and Oh (2017) proved that the LSTM model exhibits a higher performance than the RNN model in mode recognition of long-time-steps including lateral-direction driving [14]. Alami et al. (2021) demonstrated that when an ensemble model is used in Arabic text processing, the model trained for theme representation improves the text summarizing performance significantly [1]. In this study, machine learning methods were applied in three aspects to analyze the similarity between documents. The BoW and *TF-IDF* models were used as count-based numerical representation methods of words, and as ANN-based methods, Word2Vec and GloVe models, were used. Furthermore, the VPMS model, an ensemble learning model, was used to calculate the similarity between documents.

3.2 Model Evaluation

Model evaluation in machine learning is important not only as a means of achieving quality goals, but also for tuning hyperparameters and selecting an algorithm for the model development process[36]. It is difficult to find the optimum value in machine learning, where the prediction error value becomes minimum, and interpretation varies based on the evaluation criteria, evaluation data, and domain characteristics. The following are the preconditions to evaluate classification and regression models.

1. Data are classified into training, validation, and testing data, and each data type is assumed to be independent and identically distributed (IID).
2. The prediction accuracy is calculated as 1 - error, where the error is the expected value of the deviation.

$$predicted\ \ accuracy = 1 - error \tag{4}$$

$$error = B\left(L\left(y - \hat{y}\right)\right) \tag{5}$$

3. Bias is the difference between the expected value of the estimator and the actual value, and variance is the expected value of the squared deviation.

$$\text{bias} = B(\hat{y}) - y \tag{6}$$

$$variance = B\left(\left(\hat{y} - B(\hat{y})\right)^2\right) \tag{7}$$

The regression model is calculated using the expected value and deviation to determine *RMSE* (Rooted Mean Squared Error), *MAPE* (Mean Absolute Percent Error), *AIC* (Akaike Information Criteria), *BIC* (Bayesian Information Criteria), and R^2 (R-squared).

$$RMSE = \sqrt{\frac{1}{n}\Sigma\left(y - \hat{y}\right)^2} \tag{8}$$

$$MAPE = \frac{100}{n}\Sigma\left|\frac{y - \hat{y}}{y}\right| \tag{9}$$

$$AIC = n\,ln\left(\frac{SSE}{n}\right) + 2p \tag{10}$$

$$BIC = n\,ln\left(\frac{SSE}{n}\right) + p\,ln(n) \tag{11}$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{12}$$

A classification model was used to write a confusion matrix using the results of a model evaluation and to calculate the accuracy, precision, sensitivity, and specificity [5]. A confusion matrix can be used by selecting a cut-off value appropriate for the domain characteristics. Tab. 2 lists the confusion matrix for the model evaluation.

Table 2 Confusion matrix for model evaluation

Confusion matrix		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

The classification rate (%) refers to the proportion of accurately classified documents [13].

$$Classification\ rate\ (\%) = \\ = \frac{No.\ of\ currently\ classified\ document}{Total\ no.\ of\ document\ used} \cdot 100 \tag{14}$$

The classification rate (%), which was used as the performance evaluation metric in this study, refers to the proportion of documents that are matched when the document similarity between machine learning and experts is classified per layer.

3.3 Data Collection and Preprocessing

Unstructured text data were collected from the founding ethos, educational goals, and desirable human character of OO University, as well as the educational goals and desirable human character of 44 departments in this study for forming 45 corpora. The data were preprocessed by various methods, including tokenization, cleaning, normalization, stop word removal, stemming, and keyword extraction. Fig. 2 illustrates the data collection and preprocessing processes. Fig. 3 shows the text data composition. Document doc00 is a document in which Korean text data have been collected for the university's founding ethos, educational goals, and desired human character. In contrast, documents from doc01 to doc44 are documents in which the Korean text data have been collected for the educational goals and desired human character of 44 departments. The nouns were extracted from the collected 45 documents and used in the empirical study.

3.4 Document Similarity Computation Process

The characteristics of the preprocessed textual data were extracted and transformed into 2 - dimensional matrices. This was done to apply a machine-learning technique to unstructured textual data based on mathematical algorithms. The representation methods of textual data can be largely classified into local and distributed representations [10]. The local representation maps specific values with words, whereas the distributed representation expresses a word based on its neighbours [10]. A word representation is a mathematical object related to a word and uses vectors to describe the word. The representative examples of the mainly used vectors include one-hot vectors and embedding vectors. In this process, words are represented in forms appropriate for machine learning techniques (bag of words, *TF-IDF*, Word2vec, GloVe, VPMS), and the document similarities are computed via machine learning analysis.

4 COMPARISON MODEL FOR DOCUMENT SIMILARITY

4.1 Research Model

This study aimed to investigate the similarity between results obtained by a machine learning technique and those obtained through actual measurements by experts through an empirical study. Fig. 4 shows the research model adopted for a comparative study of document similarity. Document similarity was computed using a machine learning technique, and the results were compared with the measurement results of experts. A research model is proposed in which a founding ethos, educational goal, and desirable human character of OO University as well as the educational goals and desirable human character of 44 departments at OO University are taken from the 2018 curriculum handbook to perform an empirical study. The distinction drawn from the previous research is that both the model accuracy and expert evaluation results are validated from various perspectives, instead of only enhancing the model accuracy. Hence, this study aimed to examine the similarity of the judgments by humans and

machines in various aspects to promote the understanding and cooperation between each other.

4.2 Vision Propagation Measurement (Vpms) Model

To determine how well the founding ethos, educational goals, and desirable human character of OO University are reflected in the educational goals and desirable human character of 44 departments at the university, weights were applied to context words based on the assumption that keywords are mentioned in the document and that the words within the designated window of the center word have similar meanings. In particular, this model is proposed to measure the vision propagation of an organization. The unstructured text data contained in the 2018 curriculum handbook of OO University were used for an empirical comparison of document similarity between machine learning and experts. The VPMS model can be used to measure the vision propagation of other organizations. To extract keywords, the *TF-IDF* and Word2vec models were used to reflect the meaning of the words around the center word. Eq. (18) was used to calculate the word weight matrix. Fig. 5 shows the VPMS (Vision Propagation Measurement) model.

$$W_{ij} = exp\left(-\frac{d\left(x_i, y_j\right)^2}{2}\right) \tag{15}$$

The document similarity score based on the VPMS model can be calculated using Eq. (19).

$$Document\ similarity\ score_i = \frac{\sum_{j=1}^{n} Keyword\ Score_j}{N} \tag{16}$$

where document similarity $Score_i$ represents the *i*-th document similarity score, keyword $Score_j$ represents the *j*-th keyword's document similarity, and N denotes the number of keywords.

The document similarity score was calculated using a machine learning technique, and the classes were categorized as excellent, good, average, and poor based on the inverse order of the scores. Tab. 3 lists the document similarity classification criteria.

Table 3 Document similarity classification criteria

Class	Score	Rank	Application criteria
Excellent	Document similarity Score	1 - 44	rank ≥ 25%
Good			25% < rank ≤ 50%
Average			50% < rank ≤ 75%
Poor			75% < rank ≤ 100%

5 RESULTS AND DISCUSSION

5.1 Text Data Preprocessing Results

The Korean text was preprocessed and analyzed using the data analysis tool R for an empirical study on the comparison of the document similarity computed by machine learning and the document similarity determined by the experts. The packages used in the process included stringr, reshape2, ggplot2, dplyr, tm, wordcloud, RColorBrewer, plyr, KoNLP, NIADic, tidytext, devtools, wordVectors, tsne, readr, MASS, quanteda, text2vec, and lsa. Nouns were extracted from the 45 collected corpora, and word frequency DTM and one-hot encoding DTM were generated. A dendrogram based on the Euclidean distance can be used to identify the essential relationship between documents [9]. Considering how a close distance in the dendrogram analysis implies that although similarity exists, doc34, doc4, doc24, doc1, and doc16 have a relatively low level of similarity with other documents. Fig. 6 shows the results of the dendrogram analysis using 45 corpora.

5.2 Word Embedding Modeling Results

The Word2vec model is an embedding vector that has been trained using the distributional hypothesis, in which the words around the center word have similar meanings and are more likely to appear in similar locations within a document. To generate a word-embedding vector using the collected corpora, the skip-gram model was trained with a window size of two and 100 dimensions. Multidimensional scaling (MDS) is a technique for highlighting a low-dimensional space in which similarity or dissimilarity between number of objects is observed [37]. The MDS function in the MASS packing of R was used to reduce the result of word embedding trained with 100 dimensions into two dimensions. Fig. 7 shows the result of reducing the data trained with the Word2vec model into two-dimensional data. The word weight matrix calculation in Eq. (18) was applied to the word vector that has been reduced to two dimensions to generate a word similarity matrix. The GloVe model was trained with 100 dimensions using the collected corpora to generate a word-embedding vector. Fig. 8 shows the result of reducing the data trained with the GloVe model into two-dimensional data. The word similarity matrix was generated using the word vector reduced to two dimensions. Similar to the result of v(China) – v(Beijing) = v(Russia) – v(Moscow) in a study by Mikolov et al., the coordinates of each word are assigned in a high-dimensional space in a semantic space between the words trained with 'deep learning + NLP' [27]. Two words with similar vector coordinates in a two-dimensional coordinate system have similar impressions.

5.3 Extraction Results of Keywords

Seven keywords were extracted from the 45 collected corpora based on the *TF-IDF* weight: nature, world, great spirit, Catholic, warm, values, and humans. Tab. 4 lists the *TF-IDF* weight values of the seven keywords.

Table 4 *TF-IDF* weight values of seven keywords

No	Keywords	*TF-IDF*
1	Nature	0.2707
2	World	0.2253
3	Great spirit	0.2105
4	Catholic	0.2099
5	Warm	0.1911
6	Values	0.1911
7	Humans	0.1838

5.4 Expert Measurement Result

Twelve full-time professors from OO University performed actual measurements on vision documents of an upper level of the organization and action documents of a lower level of the organization. To evaluate 44 departments, 12 professors were divided into four teams of three professors, who evaluated 11 departments each. Tab. 5 shows the classification results based on expert measurements.

Table 5 Classification results based on expert measurements

Class	Classification results based on expert measurements	
	Document	No
Excellent	doc02, doc25, doc01, doc14, doc39, doc15, doc32, doc30, doc16, doc07, doc11	11
Good	doc37, doc40, doc12, doc17, doc05, doc22, doc04, doc38, doc44, doc31, doc19	11
Average	doc28, doc36, doc03, doc41, doc26, doc24, doc27, doc33, doc34, doc13, doc29	11
Poor	doc06, doc09, doc42, doc20, doc10, doc43, doc23, doc08, doc18, doc35, doc21	11

The top 10 documents with a high evaluation score (document similarity score) based on expert measurements are doc02, doc25, doc01, doc14, doc39, doc15, doc32, doc30, doc16, and doc07. The bottom 10 documents with a low evaluation score (document similarity score) based on expert measurements are doc09, doc42, doc20, doc10, doc43, doc23, doc08, doc18, doc35, and doc21.

5.5 Machine Learning Analysis Results

Words were quantified using a machine learning technique, and document similarity was computed using the trained data. Furthermore, the empirical results of the BoW, *TF-IDF*, Word2vec, GloVe, and VPMS models were analyzed. Fig. 9 shows the cosine similarity line graph of each model. The line graph of cosine similarity for each model showed that the Word2vec and GloVe models, which are ANN models, have a larger cosine similarity value compared with the BoW, *TF-IDF*, and VPMS models. However, the cosine similarity value of the GloVe model also exhibits a significant deviation in the document similarity value compared to the other models.

5.6 Empirical Comparison Results of Document Similarity

Tab. 6 shows the results of an empirical comparison of document similarities between machine learning and experts. Fig. 10 shows the line graph analysis of the classification (%) results of each model. The line graph analysis results showed that the classification rates of the BoW, *TF-IDF*, and Word2vec models have a relatively greater deviation, whereas the classification rates of the GloVe and VPMS models have a relatively smaller deviation. The VPMS model exhibits the highest classification rate (%) in the excellent class. On the other hand, the GloVe model demonstrates the highest classification rate (%) in the poor class. Tab. 7 shows a comparison of the overall document similarity classification rate (%) of each word representation method. The average classification rate (%) of each word representation method showed that the combined method had the highest rate of 43.18%, followed by the count-based word representation method at 38.64% and the ANN-based word representation method at 32.95%. It can be inferred that keywords mentioned in a document and the appearance of context words around the keywords have great significance for decision-making when determining document similarity. Furthermore, keywords mentioned in a document have a greater influence on decision-making than the meaning of the overall context. In count-based word representation methods, keywords mentioned in a document have a relatively greater effect on decision-making than determining the type of document based on the frequency of word appearance. In ANN-based word representation methods, the meaning of the overall context of a document has a greater effect on decision-making than the meaning of context words around keywords.

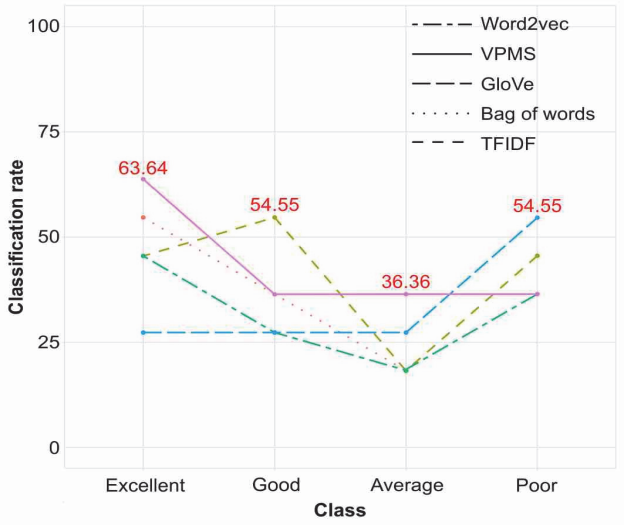

Figure 10 Line graph of classification result of each model

Fig. 11 shows a comparison of the average classification rate (%) per class. In terms of the average classification rate (%), the classification rate of excellent class was the highest at 47.27%, followed by poor class at 41.82%, good class at 36.36%, and average class at 23.73%. The classification rates (%) of the excellent and poor classes were relatively higher than those of the good and average classes. The average classification rate (%) per class showed that the documents classified as excellent or poor class with respect to document similarity had a relatively higher classification rate (%), which indicates that the decisions by a machine and humans are fairly similar. The documents classified as good or average class with respect to document similarity have a relatively lower classification rate (%), which indicates that the decisions made by a machine and humans vary considerably. It cannot be determined whether the decision made by a machine is more accurate (correct) than the decision made by humans, and the lowest classification rate (%) shown in the average class implies that the documents that are difficult to classify for humans are also difficult to classify for machines. Moreover, the decisions made by a machine and those by humans vary substantially for documents in the average class.

Figure 11 Comparison of average classification rate per class

6 CONCLUSIONS AND FURTHER WORK

This study presented an emprical analysis comparing document similarity assessments using five machine learning models with those of human expert evaluations. The ML models implemented include count-based BoW and *TF-IDF*, neural Word2Vec and GloVe, and a proposed VPMS model combining count and neural methods. Several key insights emerge from the analysis: The combined VPMS model shows the highest overall classification match rate of 43% when compared to human judgments, indicating that using keywords and their local contexts is beneficial. Documents rated by humans as having either excellent or poor similarity show higher agreement rates with the ML models, indicating a greater alignment between human and machine evaluations in cases of extreme similarities or dissimilarities. Conversely, documents considered average or challenging to classify by humans exhibit the lowest match rate of 24%, underlining the greatest disparity in human and machine assessments. The study makes two key contributions - (i) providing an extensive empirical comparison of document similarity analysis by different ML techniques versus human experts, and (ii) proposing a new VPMS model that outperforms existing methods. The limitations of this study include a limited dataset size and absence of extensive model validation. Future research can involve assessing the VPMS model using larger and more varied document datasets. Additionally, a broader comparative analysis incorporating more ML techniques can be conducted. Overall, this research offers valuable insights that can guide the development of ML techniques more closely aligned with human evaluations in text analysis tasks.

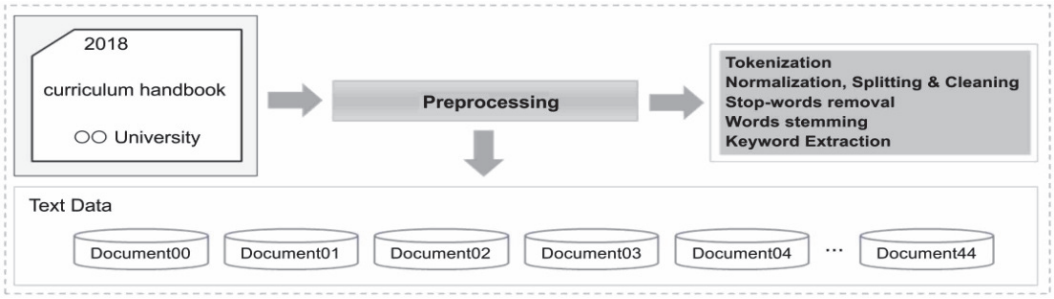

Figure 2 Data collection and preprocessing process

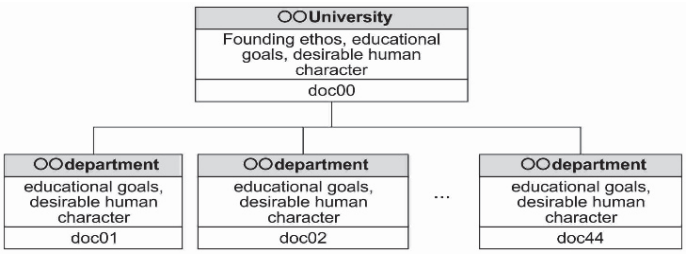

Figure 3 Text data composition

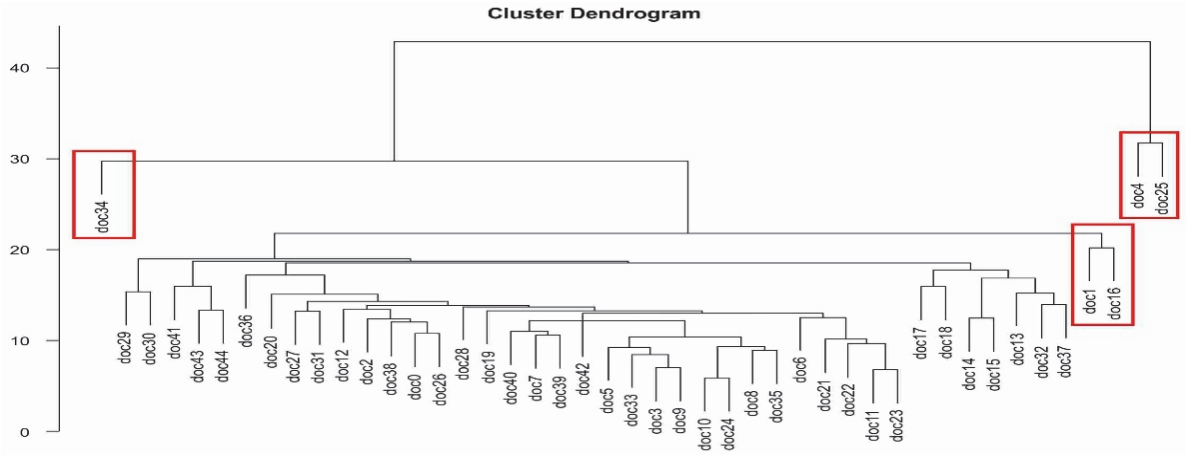

Figure 6 Results of dendrogram analysis using 45 corpora

Table 7 Comparison of the overall document similarity classification rate of word representation methods

Word representation method	Overall document similarity per model		Word representation method Average classification rate / %
	Model	Classification rate / %	
Combined method	VPMS	43.18	43.18
Count-based	*TF-IDF*	40.91	38.64
	BoW	36.36	
ANN-based	GloVe	34.09	32.95
	Word2vec	31.82	

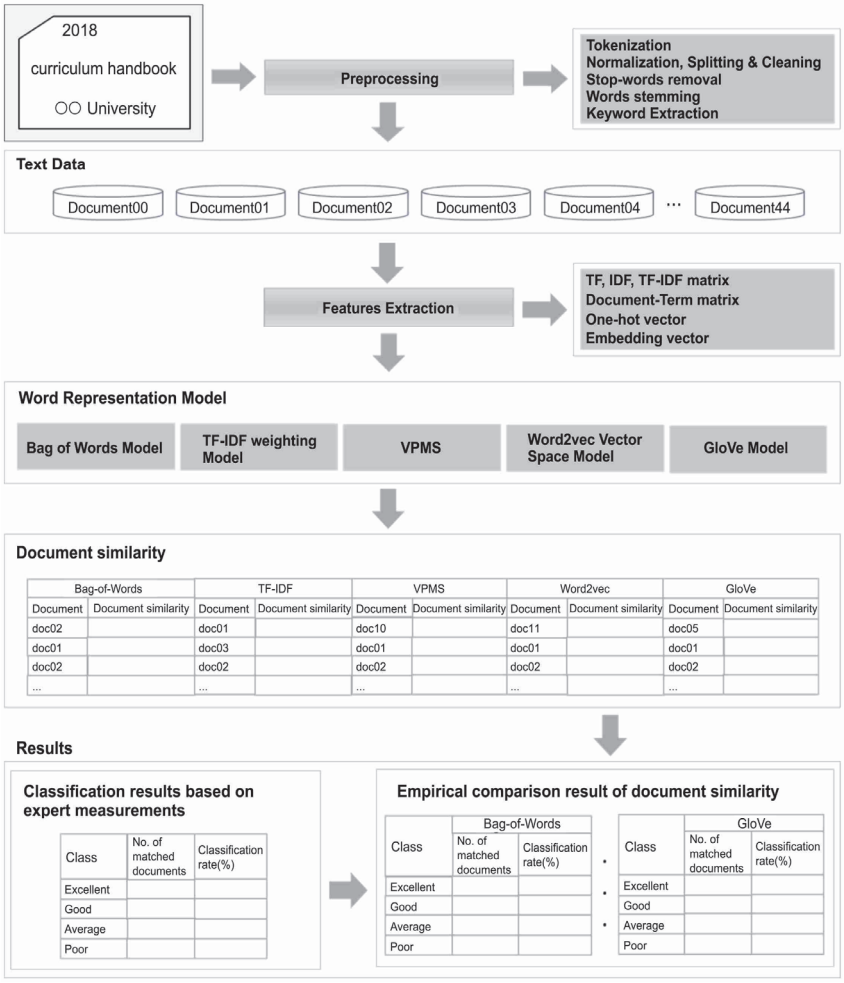

Figure 4 Comparison study model for document similarity

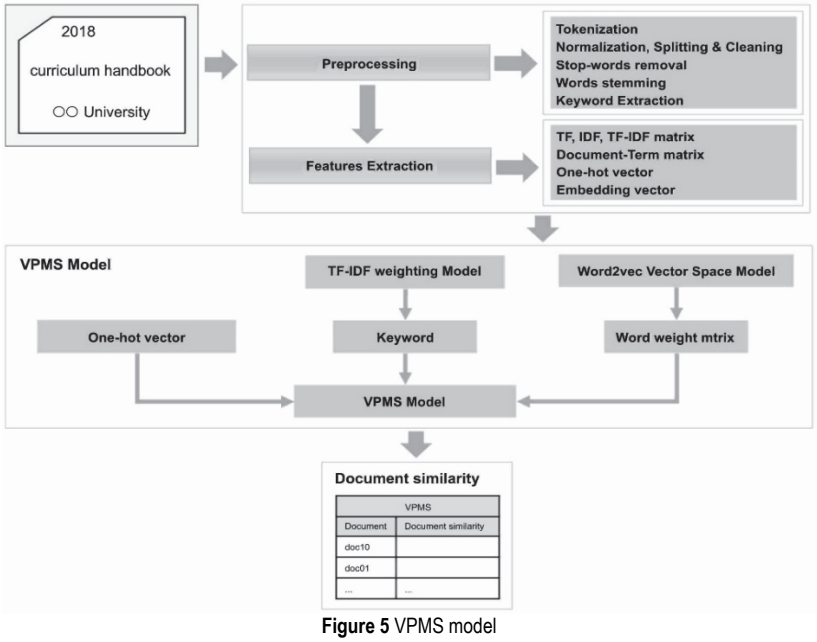

Figure 5 VPMS model

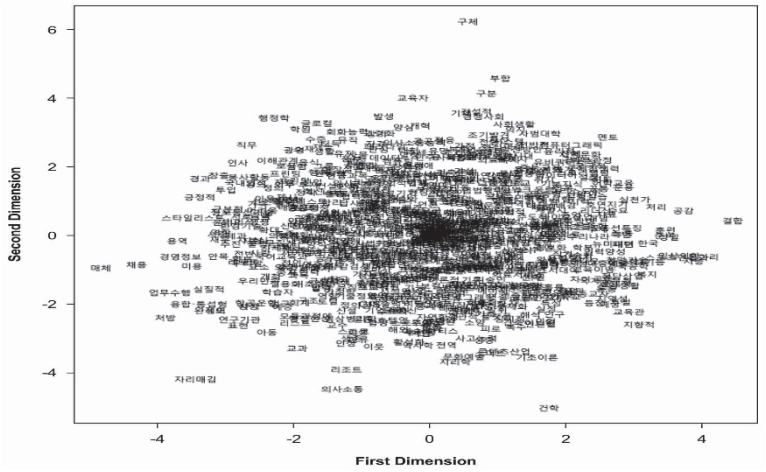

Figure 7 Word vector result reduced to two dimensions (Word2vec)

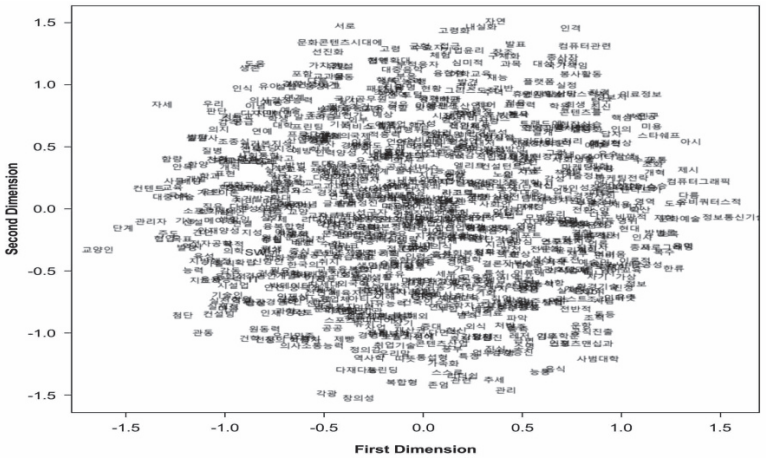

Figure 8 Word vector result reduced to two dimensions (GloVe)

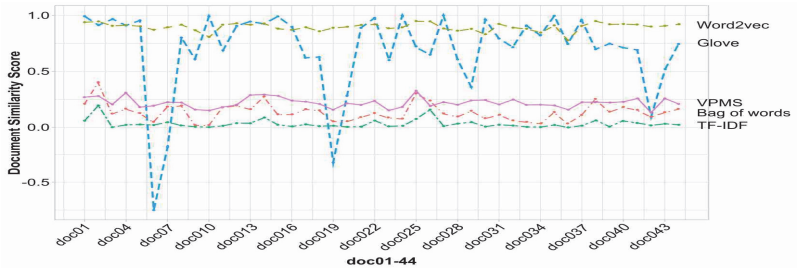

Figure 9 Cosine similarity line graph of each model

Table 6 Empirical comparison result of document similarity

Class	Count-based word representation method				ANN-based word representation method				Combined method	
	BoW		*TF-IDF*		Word2vec		GloVe		VPMS	
	No. of matched documents	Classification rate / %	No. of matched documents	Classification rate / %	No. of matched documents	Classification rate / %	No. of matched documents	Classification rate / %	No. of matched documents	Classification rate / %
Excellent	6	54.55	5	45.45	5	45.45	3	27.27	7	63.64
Good	4	36.36	6	54.55	3	27.27	3	27.27	4	36.36
Average	2	18.18	2	18.18	2	18.18	3	27.27	4	36.36
Poor	4	36.36	5	45.45	4	36.36	6	54.55	4	36.36
Total	16	36.36	18	40.91	14	31.82	15	34.09	19	43.18

7 REFERENCES

[1] Alami, N., Meknassi, M., Noureddine, E. N., Yassine, E. A., & Ammor, O. (2021). Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling. *Expert Systems with Applications,* 172. https://doi.org/10.1016/j.eswa.2021.114652

[2] Bosch, A., Muñoz, X., & Martí, R. (2007). Which is the best way to organize/classify images by content?. *Image and Vision Computing, 25*(6), 778-791. https://doi.org/10.1016/j.imavis.2006.07.015

[3] Carvallo, A., Parra, D., Lobel, H., & Soto, A. (2020). Automatic document screening of medical literature using word and text embeddings in an active learning setting. *Scientometrics, 125*(3), 3047-3084. https://doi.org/10.1007/s11192-020-03648-6

[4] Chayangkoon, N. & Srivihok, A. (2021). Text classification model for methamphetamine-related tweets in Southeast Asia using dual data preprocessing techniques. *Int. J. Electrical and Computer Engineering, 11*(4), 3617-3628. https://doi.org/10.11591/ijece.v11i4.pp3617-3628

[5] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[6] Harris, Z. S. (1954). Distributional structure. *Word, 10*, 146-162. https://doi.org/10.1080/00437956.1954.11659520

[7] Helmers, L., Horn, F., Biegler, F., Oppermann, T., & Muller, K. R. (2019). Automating the search for a patent's prior art with a full text similarity search. *PLoS One, 14*(3). https://doi.org/10.1371/journal.pone.0212103

[8] Huang, T., Ru, S. R., Zeng, Z. H., & Zhang, L. (2021). Research on motion recognition algorithm based on bag-of-words model. *Microsystem Technologies, 27*, 1647-1654. https://doi.org/10.1007/s00542-019-04462-8

[9] Jang, W. J. (2019). *An empirical study on comparative evaluation of document similarity between machine learning and an expert*. Doctoral dissertation. Soongsil University, Seoul, Korea.

[10] Jang, W. J. (2020). An empirical study on comparison of tourist destination preference using word embedding: focusing on SNS users on the Gangneung tourism area. *Global Business Administartion Revview, 17*(6), 54-71. https://doi.org/10.38115/asgba.2020.17.6.54

[11] Jang, W. J., Cho, S. In., Kim, S. S., & Gim, G. Y. (2018). A study on the implementation of big data infrastructure in smart factory. *Asia-pacific J. Multimedia Serv. Convergent Art Humanities Sociology, 8*(10), 11-23.

[12] Jang, W. J, Kim, J. Y., Lim, B. T., & Gim, G. Y. (2018). A study on data profiling based on the statistical analysis for big data quality diagnosis. *Int. J. Adv. Sci. Technology, 117*, 77-88. https://doi.org/10.14257/ijast.2018.117.07

[13] Joo, K. H., Shin, E. Y., Lee, J. I., & Lee, W. S. (2011). Hierarchical automatic classification of news articles based on association rules. *J. Korea Multimedia Society, 14*(6), 730-741. https://doi.org/10.9717/kmms.2011.14.6.730

[14] Kim, E. H. & Oh, A. (2017). Automated vehicle research by recognizing maneuvering modes using LSTM model. *J. Korea Inst. Intell. Transport. Sys., 16*(4), 153-163. https://doi.org/10.12815/kits.2017.16.4.153

[15] Kim, Y. J. (2016). *In Introduction to artificial intelligence, machine learning, and deep learning*. Park, C. K. & Lee, D. Y(Eds). Artificial neural network. Seoul: Wikibook.

[16] Kim, H. C. & Gim, G. Y. (2015). A study on public data quality factors affecting the confidence of the public data open policy. *Journal of Korea Society of IT Services, 14*(1), 53-68. https://doi.org/10.9716/KITS.2015.14.1.053

[17] Le, Q. & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning, PMLR, 32*(2), 1188-1196.

[18] Lee, H. H., Chung, S. H., & Choi, E. J. (2016). A case tudy on machine learning applications and performance improvement in learning algorithm. *Journal of Digital Convergence, 14*(2), 245.258. https://doi.org/10.14400/JDC.2016.14.2.245

[19] Lee, S. J. & Kim, H. J. (2009). Keyword extraction from news corpus using modified *TF-IDF*. *J. Society for e-Business Studies, 14*(4), 59-73.

[20] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM J. Ressearch and Development, 1*(4), 309-317. https://doi.org/10.1147/rd.14.0309

[21] Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transport Res. C: Emerg. Technologies, 54*, 187-197. https://doi.org/10.1016/j.trc.2015.03.014

[22] Matsuo, Y. (2015). *Artificial Intelligence and Deep Learning*. Park, J. E. & Park, H. J (Eds). What is artificial intelligence? Seoul: Dong-A M & B.

[23] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bull. *Math. Biophysis, 5*,115-133. https://doi.org/10.1007/BF02478259

[24] Mell, J., Beissinger, M., & Gratch, J. (2021). An expert-model and machine learning hybrid approach to predicting human-agent negotiation outcomes in varied data. *J. Multimodal User Interfaces, 15*, 215-227. https://doi.org/10.1007/s12193-021-00368-w

[25] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Int. Conference on Learning Representation*.

[26] Mikolov, T., Karafiát, M., Burget, L., Jan, C., & Khudanpur, S. (2010). Recurrent neural network based language model. *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 1045-1048. https://doi.org/10.21437/Interspeech.2010-343

[27] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems, 2*, 3111-3119.

[28] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. https://doi.org/10.3115/v1/D14-1162

[29] Provost, F. & Fawcett, T. (2014). *Data Science for Business*. Kim, S. S(Eds). Data analytical thinking. Seoul: HANBIT Media.

[30] Ramos, J. (2003). Using *TF-IDF* to Determine ord relevance in document queries. *Proceedings of the twenty-first international conference on Machine learning, 242*, 29-48.

[31] Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for *IDF*. *J Documentation, 60*, 503-520.

[32] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Revview, 65*(6), 386-408. https://doi.org/10.1037/h0042519

[33] Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384-394.

[34] Turney, P. D. & Pantel, P. (2010). From frequency to eaning: vector space models of semantics. *J.Artificial Intell. Res, 37*, 141-188. https://doi.org/10.1613/jair.2934

[35] Wolf, L., Hanani, Y., Bar, K., & Dershowitz, N. (2014). Joint word2vec networks for bilingual semantic representations. *Int. J. Computational Linguistics and Applications, 5*(1), 27-44.

[36] Yun, S. Y. (2018). *A study on patent data analysis and competitive advantage strategy using machine learning & neural network* .Doctoral dissertation. Soongsil University, Seoul, South Korea.

[37] Huh, M. H. & Lee, Y. G. (2012). Multidimensional Scaling of Asymmetric Distance Matrices. *The Korean Journal of Applied Statistics, 25*(4), 613-620. https://doi.org/10.5351/KJAS.2012.25.4.613

[38] Al-akashi, F. (2021). Improving Learning Performance in Neural Networks. *International Journal of Hybrid Innovation Technologies, 1*(2), 27-42, https://doi.org/10.21742/ijhit.2653-309X.2021.1.2.02

[39] Bellotti, W., Davies, D. N., & Wang, Y. H. (2022). Development and Replication of Improved Recommendation Algorithm for Network Representation Learning. *Journal of Science and Engineering Research, 1*(2), 61-75, https://doi.org/10.56828/jser.2022.1.2.5

[40] Zhang, T. (2015). An Automatic Software Requirement Analysis Model based on Planning and Machine Learning Techniques. *International Journal of Future Generation Communication and Networking, 8*(5), 177-188, https://doi.org/10.14257/ijfgcn.2015.8.5.18

Contact information:

Won-Jung JANG
Catholic Kwandong University,
25601 #502, The Mary Hall, 24, Beomil-ro 576, Gangneung-si, Gangwon-do,
South Korea
E-mail: wjjang@cku.ac.kr