International Journal of Engineering

Journal Homepage: www.ije.ir

Accelerating Legislation Processes through Semantic Similarity Analysis with BERT-based Deep Learning

J. Naseri[a], H. Hasanpour*[a], A. Ghanbari Sorkhi[b]

[a] *Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran*
[b] *Faculty of Electrical and Computer Engineering, University of Science and Technology of Mazandaran, Behshahr, Iran*

PAPER INFO

Paper history:
Received 22 November 2023
Received in revised form 06 January 2024
Accepted 20 Januray 2024

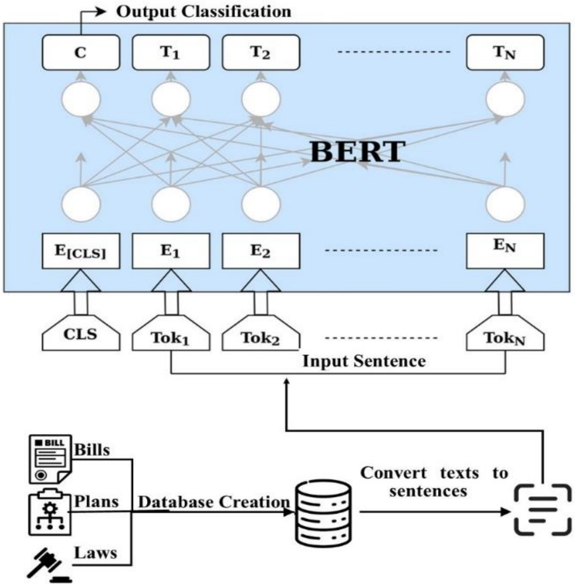
Keywords:
Text Mining
Neural Network
Semantic Search
Sentence Embedding in Vector Space
BERT Model

ABSTRACT

Countries are managed based on accurate and precise laws. Enacting appropriate and timely laws can cause national progress. Each law is a textual term that is added to the set of existing laws after passing a process with the approval of the assembly. In the review of each new law, the relevant laws are extracted and analyzed among the set of existing laws. This paper presents a new solution for extracting the relevant rules for a term from an existing set of rules using semantic similarity and deep learning techniques based on the BERT model. The proposed method encodes sentences or paragraphs of text in a fixed-length vector (dense vector space). Thereafter, the vectors are utilized to evaluate and score the semantic similarity of the sentences with the cosine distance measurement scale. In the proposed method, the machine can understand the meaning and concept of the sentences by using the BERT model coding method. The BERT model considers the position of the entities in the sentences. Then the semantic similarities of documents, calculating the degree of similarity between their documents with a subject, and detecting their semantic similarity are done. The results obtained from the test dataset indicated the precision and accuracy of the method in detecting semantic similarities of legal documents related to the Islamic Consultative Assembly of Iran, as well as the precision and accuracy of performance above 90%.

doi: *10.5829/ije.2024.37.06c.01*

Graphical Abstract



*Corresponding Author Email: h.hassanpour@shahroodut.ac.ir (H. Hassanpour)

Please cite this article as: Naseri J, Hasanpour H, Ghanbari Sorkhi A. Accelerating Legislation Processes through Semantic Similarity Analysis with BERT-based Deep Learning. International Journal of Engineering, Transactions C: Aspects. 2024;37(06):1050-8.

1. INTRODUCTION

In a legal system, legislation plays a critical role in shaping the policies and regulations that guide a nation's development. Therefore, the process of enacting laws must be approached with a high degree of precision and deliberation to ensure that they are beneficial for the country's progress. To this end, a plan or bill sent to the assembly for approval is first evaluated and analyzed by experts from different aspects. However, human mistakes in data collection and evaluation of legal frameworks in the review of plans and bills can challenge this legislation. Artificial Intelligence (AI)-based software can play the role of an intelligent assistant in law and legislative affairs, and help representatives and experts prepare plans and bills, evaluate them, control documents regarding solecism, and identify and extract relevant laws. This capability accelerates the legislative process, decreases costs, and improves the quality of decrees (1, 2).

The legislative process is time-consuming, involving lengthy discussions and multiple levels of expertise (3). Hence, it is required to analyze a significant volume of textual data in a short timeframe, making it challenging to accurately assess the relationships between proposed bills and existing laws. The use of AI-powered software can significantly reduce the time and effort required for this analysis, allowing legislators to focus on critical decision-making and debate. Given the recent development in text processing, deep learning techniques based on BERT model (4) can be utilized for the semantic search of different parts of a plan and bill in the existing set of laws, and then they can be carefully evaluated by experts and legislators (5).

In this method, the text of a plan and bill, which is submitted to the assembly, is first pre-processed. Therefore, spelling errors are corrected, the text is converted into sentences, the word roots in the sentence are found, and the roles of words in the sentence are identified. Thereafter, the prepositions are removed and the text sentences are converted into numerical vectors (6). Further, the semantic similarities (7-9) of the text sentences with the relevant laws are extracted in such a way that the sentences or paragraphs of the texts are encoded in a vector space using BERT model based on deep learning. Then the scores of semantic similarities of the sentences are calculated using a similarity criterion. Additionally, the sentences with the same or close meanings are extracted. Therefore, the important semantic words are discovered from documents and the model identifies which sentences are semantically similar by considering the position of their constituents.

The comparison and evaluation of the proposed method with the existing methods for searching in the text indicated that this method can provide better results in the semantic search of documents due to recognizing and understanding the relationships between words, and ultimately the meaning of the sentence.

The initial methods for text search, which perform the search only based on the matching of the query string, cannot work properly if the words are embedded in the sentence or the word is used in an expanded way in the sentence (6, 10).

In the second text search methods, which use Term Frequency - Inverse Document Frequency (TF-IDF) and BM25 techniques, the keywords are first extracted, and then the topics and meanings of the sentences are detected using these keywords. In this method, keywords are specified and weighed for the semantic analysis of sentences. Keywords represent the entire content of the text, and thus they are weighed based on the number of repetitions and their importance. However, this method cannot recognize synonyms and more complex search terms in a paragraph because TF-IDF and BM25 techniques work based on a "bag of words" simplifying principle. In other words, the text is classified into a set of words from which numerical vectors are generated, but the order and positions of the words in the text are not observed. Therefore, this method cannot accurately understand and extract semantic similarities of documents.

The third methods (such as the proposed method) are designed based on the neural network. In semantic search with the neural network, the neural model learns to encode a query as a vector to better understand the meaning or semantic value of a query and calculate the association between sentences by placing sentences in a vector space. This technique seeks to encode a sentence or paragraphs of short text into a fixed-length vector (dense vector space), and then use the vector to evaluate to some extent their similarities reflect human semantic judgments. In this method, the neural language model is designed based on the transformer architecture and it allows detecting the relationships between words easily, and then correctly extracting the semantic relationship of documents (11, 12).

The remainder of this paper is as follows. Section 2 presents a review of the literature on text processing and the use of artificial intelligence in legislation in recent years. Section 3 elaborates the proposed method and the different steps of this method, in addition to some basic concepts to identify the semantic similarities. Section 4 discusses the evaluation criteria, result analysis, and performance appraisal. Finally, the conclusion is presented in section 5.

2. LITERATURE REVIEW AND RESEARCH BACKGROUND

The Islamic Parliament Research Center of Iran has recognized the potential of AI in enhancing the policymaking and planning processes, and has emphasized the need for skilled human resources in AI and law to effectively implement AI technology in the legislative process. This recognition highlights the importance of investing in AI-related training and education to support the effective use of AI in lawmaking (13).

The research conducted by Leskovec et al. (14) focused on the application of data mining techniques for analyzing textual data, specifically in terms of terminology extraction and document similarity assessment. The paper discussed various segmentation and distance measurement methods used in text analysis, including Euclidean, Jaccard, and cosine distances.

Researchers proposed a method that leverages semantic similarities in thesauri to identify key terms in textual data (15, 16). After pre-processing the text, significant words are initially identified using statistical techniques. Next, secondary key terms are generated by analyzing the embedded terms derived from the significant words and the thesaurus, with the TF-IDF weighting scheme. Finally, the ultimate key terms are selected based on the hierarchical and equivalent relations present in the thesaurus, utilizing clustering techniques (17).

Within the field of AI, textual databases continue to grow, encompassing an extensive range of documents such as new articles, books, and web pages. This rapid expansion highlights the pressing need for automated evaluation tools to assess text sources efficiently. One such solution is automatic text summarization, which utilizes text processing methods to extract precise summaries and employs scoring criteria and part-of-speech tagging to assign importance weights to words in a sentence (18).

Earlier research has utilized deep learning techniques to evaluate the semantic similarity of textual sentences by mapping them into a vector space. This approach involves embedding sentences and paragraphs of documents within a vector space and identifying the most similar embeddings from the collection, enabling the extraction of semantically similar documents that share a high degree of overlap (19).

Numerous researchers emphasized the potency of text mining in uncovering critical patterns within big data by analyzing unstructured text. The assessment of word or document similarities is a critical component of natural language processing and information retrieval, as these similarities can be identified through statistical calculations and word similarity measurements (20).

Jiang et al. (21) explored the extraction of semantic relationships between documents and text similarities, building a concept space for each term based on the Wikipedia knowledge base. This concept space is constructed by weighting various parts of the Wikipedia website, enabling the calculation of semantic similarities between two terms. Moreover, two texts can be compared using the weighted concept space to evaluate their level of similarity.

Karaa (22) proposed a method for performing stemming, retrieving information, and locating documents relevant to users' information needs. Stemming is a pre-processing technique in text mining and a key requirement in natural language processing-related applications. Therefore, it plays a vital role in information retrieval systems, contributing to their accuracy and effectiveness.

Kamyar et al. (23) highlighted the significance of weighting words, a crucial step in language processing that influences the precision of text classification. They also presented a novel approach to weighting words, building upon the statistical TF-IDF weighting method. This method modifies the TF parameter, which measures word frequency in a text, by incorporating additional linguistic features that enhance its accuracy.

Reimers and Gurevych (24) evaluated the performance of BERT model and established that it displays exceptional performance in sentence classification and sentence pair regression. This model leverages the power of transformer networks and mutual encoding to encode two sentences as input, subsequently predicting a target value and mapping the text to a vector space, resulting in the extraction of semantic similarities between text sentences (25).

Before the advent of neural search, the initial approaches to text retrieval were restricted to basic string matching, which involved searching through databases or text files to identify instances of the query string. These methods were not sufficiently sophisticated to satisfy the demands of semantic search in large text corpora, resulting in limited accuracy and relevance of search results.

The second generation of text retrieval relied on algorithms such as TF-IDF and BM25 to identify crucial keywords in the text and extract semantic topics based on keyword weighting. Despite this advancement, these algorithms still faced challenges in identifying broader search queries within paragraphs and detecting synonyms, as they operated on the bag-of-words principle, converting text into numerical vectors without preserving the word order and position in the document.

The third generation of text retrieval, which encompasses the proposed method, has leveraged neural networks to enhance semantic search. This approach trains a model to encode queries as vectors, thereby enhancing the understanding and semantic value capture of the query. Moreover, it can establish relationships between sentences by placing them in a vector space, facilitating more sophisticated and accurate retrieval of relevant information.

3. THE PROPOSED METHOD

This section proposes a method for processing text and discovering semantic similarities in plans and bills. The overall structure of the proposed method is given in Figure 1.

The proposed method involves pre-processing textual data, including plans, bills, and laws, into a structured database (26). The BERT model, which has been pre-trained on large text datasets, is used to identify semantic similarities between sentences in the documents. The sequential neural network and transformer architecture enable the encoding of each text into a fixed-length vector, which is then used to score the semantic similarity of sentences based on cosine distance. This approach can effectively handle documents with multiple paragraphs, accurately extracting semantic similarities between sentences in plans, bills, and laws.

After the semantic similarity score is computed, the model extracts a semantic classification, identifying words with similar semantic values. Furthermore, it evaluates each target sentence in the text based on its context, considering the preceding and succeeding sentences, and measuring the score based on the position of the sentence. This classification of semantic similarities between plans and bills provides a valuable tool for legislative experts, enabling them to quickly identify relationships between laws and accurately detect semantic similarities between plans and bills (27).

The proposed method involves the following steps:

- Pre-processing: All plans, bills, and laws are entered into a structured database after text pre-processing.
- Sentence extraction: Sentences are extracted from the texts of plans, bills, and laws.
- Semantic similarity analysis: Deep learning and the BERT model are used to identify semantic similarities between sentences.
- Score calculation: The semantic similarity score of each document (plan or bill) is computed.
- Semantic classification: The semantic classification of documents is extracted based on the similarity score.

These sections are explained in detail as follows.

3. 1. Pre-processing The first stage of the proposed method involves text pre-processing and preparation, including concept identification and extraction. Next, text enrichment is performed, which involves labeling the semantic roles of word components in the sentences, eliminating stop words such as "from" and "with," stemming the text words, and extracting and identifying keywords and nominal entities.

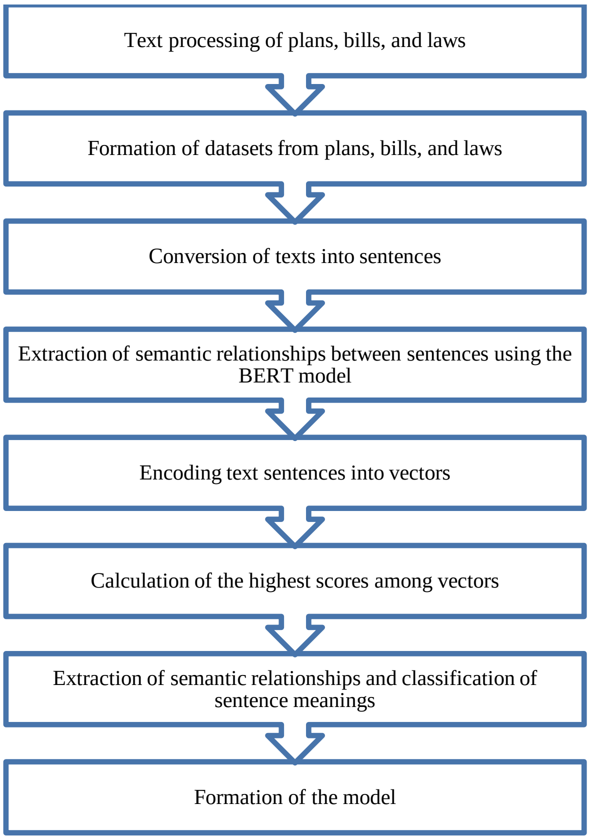

Text processing of plans, bills, and laws
Formation of datasets from plans, bills, and laws
Conversion of texts into sentences
Extraction of semantic relationships between sentences using the BERT model
Encoding text sentences into vectors
Calculation of the highest scores among vectors
Extraction of semantic relationships and classification of sentence meanings
Formation of the model

Figure 1. The overall structure of the proposed method

3. 2. Using BERT Model To Identify Semantic Similarity BERT is a deep learning algorithm that focuses on natural language processing. It understands the meaning of the smallest units of spoken language, including the use of prepositions in text and expressions. By comprehending the structure and complexity of language, it grasps the relationships between words and their meanings. Additionally, unlike language models that can only read input text unidirectionally (from left to right or from right to left), BERT can process information bidirectionally simultaneously. This allows it to understand the meaning of words in context and within the text.

In the second stage of the semantic search, the semantic similarities between sentences are evaluated using the BERT model. Semantic search systems examine different criteria such as understanding the nature, meanings of words, and variety of words to find the concept of the search and use the meaning of words to provide more interactive results in a database and help the searcher to extract answers and results . Semantic search is performed by embedding words in the text and finding the closest embeddings to detect inputs with the same meanings as the query (27).

Using the BERT model, which is a pre-trained model on large texts, this paper seeks to identify semantic similarities between sentences. BERT model can perceive the meaning of sentences, using sequential neural networks and a technical architecture called transformer. Sequential neural networks and a technical architecture, called transformer are famous and advanced neural network architecture for text processing and machine translation. Transformer utilizes self-attention layers to model dependencies between words in sentences. The transformer can pay attention to the connections between each word and other words of the sentence, using the self-attention operations on different words of the sentence.

3. 2. 1. Transformer Model Architecture

A transformer model receives a sentence at the input and delivers its translation at the output.

As shown in Figure 2 this model is classified into groups, encoders and decoders, which are connected together.

In this structure, each encoder has two separate sub-layers, a self-attention layer, and a feedforward neural network. The encoder input first passes through a self-attention layer which helps the encoder focus on other words in the sentence during the word encoding process. The output of the self-attention layer enters a feedforward neural network layer.

Each decoder has two layers, self-attention and feedforward neural network, but in the decoders of the attention layer, there is another layer, called Encoder-Decoder Attention, which helps the decoder focus on relevant words. To the best of our knowledge, natural language processing first needs to convert the input words into vectors to be understandable for the machine, which is done by word embedding algorithms, and thus they embed words in the transformer model only in the first encoder (lowest encoder). When each word is converted into a vector, each of the two sub-layers passes in the encoders. Therefore, these words enter the encoder in parallel, and another vector is added to each input embedding vector in the transformer model; thereby helping the model to recognize the position of each word or the space between different words in the input sequence.

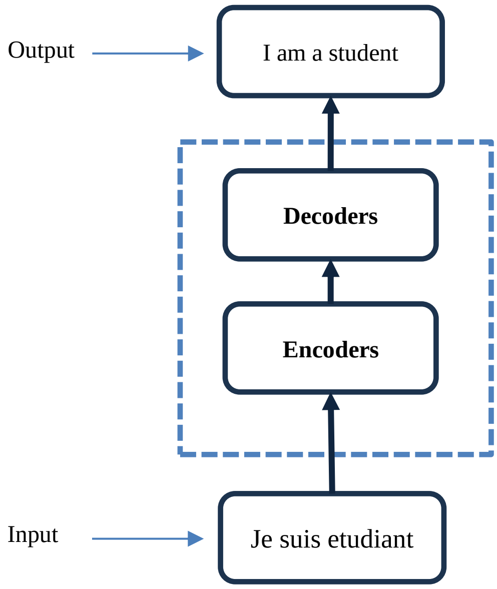


Figure 2. Transformer model architecture for text translation from French to English

The BERT model is beneficial for finding semantic similarities in documents since this model can collectively and simultaneously evaluate the semantic information of documents, analyze the connections between sentences and different stages of text processing, and thus easily identify the latent semantic similarities of the text (28-30).

Thereafter, the texts are converted into vectors and special features (the same features of an entity) are extracted. These special features are the criteria which are used to score the semantic similarities. Then, the semantic analysis of words is performed to find the semantic similarities between the text components. In the next stage, the feature selection is performed based on repetition and removal of irrelevant information, aiming to eliminate irrelevant and extra information from the text.

The most important feature is selected through the scores of words, and thus only important and relevant information with semantic similarities remains. Therefore, the semantic similarity score between the text components is calculated. This score is based on the sentence and criterion positions. At this stage, the semantic classification of documents is extracted after calculating the semantic similarity score (plan or bill). This classification helps legislative experts make decisions to find the relationships of laws and detect the rate of semantic similarity between plans and bills with laws (31, 32).

This manuscript utilized transformer-based models as the most advanced language models based on deep neural networks. These models are particularly effective for natural language processing problems (33). For example, assume the following sentence: "I like to read articles about artificial intelligence". To use the BERT model, first, we convert the sentence into numerical vectors. This stage is done as pre-processing, consisting of several stages. Tokenization is the first stage (at this stage, all punctuation marks are converted into a string of words by removing blank spaces and commas). Therefore, the sentence is divided into smaller tokens. For example, our sentence is tokenized as follows: "I, like, to, read, articles, about, artificial, intelligence". These tokens are displayed as a string of numeric numbers. In chipping, the second stage, the BERT model usually accepts only fixed-length inputs. Therefore, if the length of a sentence is longer than the limit of the model, we must decrease its components or make it a fixed length. At this stage, only selecting some tokens and removing the rest is considered as a common technique.

The third stage comprises the conversion of vectors. As the tokens are ready, the BERT model utilizes its transformer layers to assign vectors to each token. These vectors are usually vectors with large dimensions. Thus the connections between words are modeled in the sentence. Model training is the fourth stage where the models are trained using large sets of labeled data to learn linguistic features and relations. This training consists of optimization and adjustment of the weights related to the model to have the best performance in different problems. To this end, sentences and words in the text can be converted into numerical vectors through the transformer and BERT model. Furthermore, these vectors can be used for a large number of language-processing tasks such as text classification, entity recognition, and translation. This technique encodes the sentence in a fixed-length vector (dense vector space) and embeds each word in the vector space into two vectors, lexical and semantic (see Figure 3).

The similarity between these vectors is calculated at this stage after mapping the words in the vector space using the Cosine distance criterion. Therefore, the coded terms have a set of corresponding vectors in this vector space. If the text has similar words, the vectors are placed close to each other in the vector space, and if they have opposite meanings, the vectors move away from each other or their directions are opposite. Therefore, the semantic similarity score is obtained by calculating the differences.

4. IMPLEMENTATION AND ANALYSIS OF RESULTS

4. 1. Database This manuscript analyzes data from250 documents approved by the Islamic Consultative Assembly from 2006-2021, including 50 bills, 50 plans, and 150 laws. The documents analyzed include Law of Objectives, Duties, and Organizations of the Ministry of Science, Research, and Technology; the statute of the Islamic Republic of Iran Broadcasting (IRIB); the Law of the Iranian Tobacco Company; the Law of universities and higher education centers; and the plan for the conservation and management of rivers in Iran. Examples of bills analyzed include the annual budget bill and the bill to amend the statute of the cultural heritage organization (34).

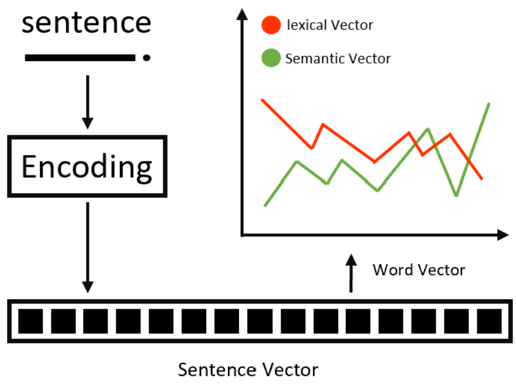


Figure 3. Comparing lexical vector and semantic vector of a sentence Using Bert Model

4. 2. Data Analysis A total of 100 plans and bills, and 150 laws were considered to analyze the results. The semantic similarity scores were measured to extract the semantic similarity of the plans and bills with the laws, using the model, and then compared them with experts' opinions.

The present model aims to present a system as an assistant for legislation experts in such a way that the model can quickly discover and recognize the semantic similarities of documents. In this method, the plans and bills related to the laws are scored, and a scoring threshold is obtained for the performance and accuracy of the results. In other words, the plan or bill is compared with the law sentence by sentence, the semantic score is calculated, and the sentences and paragraphs of the texts are encoded in a fixed-length vector (dense vector space). Furthermore, the vectors are used for evaluation and scoring the sentences, using cosine distance. Additionally, this method compares experts' previously measured semantic similarity with the semantic similarity score discovered by the model.

This research aimed to design and implement an AI model based on deep learning as an assistant to experts in finding laws related to plans and bills that are proposed to the Islamic Consultative Assembly. The AI model developed in this research should work in such a way that it can have the minimum error in finding the relevant semantic items, and the model can extract the maximum relevant semantic items.

To evaluate the performance of the AI model, we compared its results with the opinions of experts in the Islamic Consultative Assembly. This comparison of the model settings and parameters revealed that the error rate in detecting semantic similarities decreased. After conducting multiple tests and gathering extensive feedback and information from the assembly, we determined that a semantic similarity score of 34% or higher indicates a correct extraction of relevant semantic items.

The Receiver Operating Characteristic (ROC) curve of this model is shown in Figure 4 to evaluate the efficiency of data classification parameters, called True Positives (TP), False Positives (FP), True Negatives (TN), and False negatives (FN). The TP class occurs when the expert and the AI model report the positive semantic similarity between the two documents. The FP class happens when the expert reports a negative semantic similarity between the two documents, but the AI model reports a positive semantic similarity between the two documents. The TN class occurs when the expert and the AI model declare a negative semantic similarity between the two documents. Furthermore, the FN class happens when the expert declares a positive semantic similarity between the two documents, but the AI model declares a negative semantic similarity between the two documents. The ROC curve indicate the good performance and very high precision of the model in the evaluation. Implementation results of the proposed model on three sample plans and bills from the dataset are shown in Table 1.

Table 2 compares the results between experts and the AI model on the data provided in this research. The best result was found regarding the accuracy and precision of the model performance by considering a 34% threshold of the similarity score for documents. A precision of above 96% was obtained for the model on the evaluation data. Results provided in Table 2 indicatie the accuracy and precision of the AI model in detecting the semantic similarities of legal documents related to the Islamic Consultative Assembly. It also shows the achievement of accuracy and precision of performance above 90%.

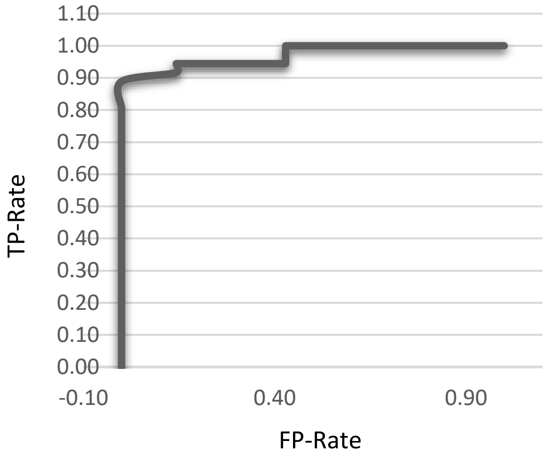


Figure 4. ROC curve of the method evaluation

TABLE 1. Dataset model, comparison of the results between experts and artificial intelligence model

Name of plan or bill	Name of relevant law	Semantic evaluation percentage of the expert	Semantic evaluation percentage of the model	Semantic similarity of the expert	Semantic similarity of the model	Confusion
1. The plan of regulating some financial, administrative, and support regulations of the Ministry of Education- 15/06/2010	The Law on the creation of education councils in provinces, counties, and districts- 18/05/2000	40	43	No	yes	FN
2. The plan of saving and revitalizing Iran's lakes and wetlands- 15/06/2010	The law on Iran's fifth development plan- 5/01/2011	40	27	No	No	TN
3.The plan of supporting handicraft masters and artists- 15/06/2010	The law on social insurance for carpet weavers, and handicraft workers with ID- 9/08/2009	20	29	No	No	TN

TABLE 2. Results of the proposed method evaluation with specific parameters for test dataset samples

The mean score of semantic similarities of sentences given by the expert	**59**
The mean score of semantic similarities of sentences (proposed method)	48

Performance statistics	TP=36, FP=3,
	FN=0, TN=4

Precision	0.92
Recall	1.0
Accuracy	0.93
Precision and recall outcome- F1-Score	0.96

5. CONCLUSION

The semantic searching method proposed in this paper was designed and implemented using a neural network, BERT model, and transformer architecture based on deep learning. This paper introduces a model that offers an efficient and useful solution for detecting laws related to plans and bills. The model can encode sentences or paragraphs into fixed-length vectors, using cosine distance to evaluate and score sentences. Furthermore, the proposed approach's consideration of text conceptual and semantic relationships suggests that it is well-suited for classifying complex texts and accurately extracting semantic relationships. The test dataset sample results demonstrate the method's high precision and accuracy in identifying semantic similarities between relevant legal documents on real datasets.

6. REFERENCES

1. National strategic plan for research and development of artificial intelligence and legislation in Iran. In: Center MR, editor.: Islamic Parliament Research Center of The Islamic Republic of IRAN; 2018.
2. Research in artificial intelligence and legislation and review of civil law in the field of robotics of the European Union Parliament. In: Center MR, editor.: Islamic Parliament Research Center of The Islamic Republic of Iran; 2019.
3. Burri T, Von Bothmer F. The new EU legislation on artificial intelligence: a primer. Available at SSRN 3831424. 2021. https://doi.org/10.2139/ssrn.3831424
4. Farhoodi M, Toloie Eshlaghy A, Motadel M. A Proposed Model for Persian Stance Detection on Social Media. International Journal of Engineering, Transactions C: Aspects. 2023;36(6):1048-59. https://doi.org/10.5829/IJE.2023.36.06C.03
5. Cath C. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2018;376(2133):20180080. https://doi.org/10.1098/rsta.2018.0080
6. Kornilova A, Eidelman V. BillSum: A corpus for automatic summarization of US legislation. arXiv preprint arXiv:191000523. 2019. https://doi.org/10.48550/arXiv.1910.00523
7. Saraswat N, Li C, Jiang M. Identifying the Question Similarity of Regulatory Documents in the Pharmaceutical Industry by Using the Recognizing Question Entailment System: Evaluation Study. JMIR AI. 2023;2(1):e43483. https://doi.org/10.2196/43483
8. Amur ZH, Kwang Hooi Y, Bhanbhro H, Dahri K, Soomro GM. Short-Text Semantic Similarity (STSS): Techniques, Challenges and Future Perspectives. Applied Sciences. 2023;13(6):3911. https://doi.org/10.3390/app13063911
9. Fradelos G, Perikos I, Hatzilygeroudis I, editors. Using Siamese BiLSTM Models for Identifying Text Semantic Similarity. IFIP International Conference on Artificial Intelligence Applications and Innovations; 2023: Springer.
10. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018. https://doi.org/10.48550/arXiv.1810.04805
11. Silic A, Saric F, Basic BD, Snajder J, editors. TMT: Object-oriented text classification library. 2007 29th International Conference on Information Technology Interfaces; 2007: IEEE.
12. Reimers N, Gurevych I. Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:200409813. 2020. https://doi.org/10.48550/arXiv.2004.09813
13. Artificial Intelligence and Legislation. In: Center MR, editor.: Islamic Parliament Research Center of The Islamic Republic of Iran; 2018.
14. Leskovec J, Rajaraman A, Ullman JD. Mining of massive data sets: Cambridge university press; 2020.
15. Sadjadi S, Mashayekhi H, Hassanpour H. A two-level semi-supervised clustering technique for news articles. International Journal of Engineering, Transactions C: Aspects. 2021;34(12):2648-57. https://doi.org/10.5829/IJE.2021.34.12C.10
16. Hassanpour H, AlyanNezhadi M, Mohammadi M. A signal processing method for text language identification. International Journal of Engineering, Transactions C: Aspects. 2021;34(6):1413-8. https://doi.org/10.5829/IJE.2021.34.06C.04
17. Rao KS, Murthy D, Kancherla GR. Semantic similarity based automatic document summarization method. International Journal of Engineering and Advanced Technology (IJEAT) ISSN.2249-8958. https://doi.org/10.35940/ijeat.F8566.088619
18. Hosseinikhah T, Ahmadi A, Mohebi A. A new Persian text summarization approach based on natural language processing and graph similarity. Iranian Journal of Information Processing and Management. 2018;33(2):885-914. https://doi.org/10.35050/JIPM010.2018.084
19. Wang B, Liu W, Lin Z, Hu X, Wei J, Liu C. Text clustering algorithm based on deep representation learning. The Journal of Engineering. 2018;2018(16):1407-14. https://doi.org/10.1049/joe.2018.8282
20. Dang S, Ahmad PH. Text mining: Techniques and its application. International Journal of Engineering & Technology Innovations. 2014;1(4):22-5.
21. Jiang Y, Zhang X, Tang Y, Nie R. Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. Information Processing & Management. 2015;51(3):215-34. https://doi.org/10.1016/j.ipm.2015.01.001
22. Karaa WBA. A new stemmer to improve information retrieval. International Journal of Network Security & Its Applications. 2013;5(4):143. https://doi.org/10.5121/ijnsa.2013.5411
23. Kamyar H, Kahani M, Kamyar M, Poormasoomi A, editors. An automatic linguistics approach for persian document summarization. 2011 International Conference on Asian Language Processing; 2011: IEEE.
24. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:190810084. 2019. https://doi.org/10.48550/arXiv.1908.10084
25. Le Q, Mikolov T, editors. Distributed representations of sentences and documents. International conference on machine learning; 2014: PMLR.
26. Haveliwala TH, Gionis A, Klein D, Indyk P, editors. Evaluating strategies for similarity search on the web. Proceedings of the 11th international conference on World Wide Web; 2002.
27. Pennington J, Socher R, Manning CD, editors. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014.
28. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Advances in neural information processing systems. 2014;27. https://doi.org/10.48550/arXiv.1409.3215
29. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013. https://doi.org/10.48550/arXiv.1301.3781
30. Khosrovian K, Pfahl D, Garousi V, editors. Gensim 2.0: a customizable process simulation model for software process evaluation. International conference on software process; 2008: Springer.
31. Hossain MZ, Akhtar MN, Ahmad RB, Rahman M. A dynamic K-means clustering for data mining. Indonesian Journal of Electrical engineering and computer science. 2019;13(2):521-6. https://doi.org/10.11591/ijeecs.v13.i2.pp521-526
32. Yi J, Zhang Y, Zhao X, Wan J. A novel text clustering approach using deep-learning vocabulary network. Mathematical Problems in Engineering. 2017;2017. https://doi.org/10.1155/2017/8310934

33. Navigli R, Velardi P. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. IEEE transactions on pattern analysis and machine intelligence. 2005;27(7):1075-86. https://doi.org/10.1109/TPAMI.2005.149

34. Floridi L. The European Legislation on AI: A brief analysis of its philosophical approach. Philosophy & Technology. 2021;34(2):215-22. https://doi.org/10.1007/s13347-021-00460-9

Persian Abstract

چکیده

مدیریت و اداره کشورها بر پایه قوانین دقیق و صحیح پایه‌گذاری می‌شود. وضع قوانین مناسب و به هنگام می‌تواند سبب پیشرفت یک کشور شود. هر قانون یک عبارت متنی است که پس از طی فرایندی با تصویب مجلس به مجموعه قوانین موجود اضافه می‌شود. در بررسی هر قانون جدید، از میان مجموعه قوانین موجود، قوانین مرتبط استخراج و مورد تحلیل قرار می‌گیرد. در این مقاله، یک راهکار نوین برای استخراج قوانین مرتبط برای یک عبارت مورد بحث از میان مجموعه قوانین موجود با استفاده از تکنیک های ارتباط معنایی ویادگیری عمیق مبتنی بر مدل برت ارائه شده است. در روش پیشنهادی، جملات یا پاراگراف‌های متنی را در یک بردار با طول ثابت (فضای برداری متراکم) رمزگذاری کرده سپس از آن بردارها برای ارزیابی و امتیازدهی به ارتباط معنایی جملات با مقیاس اندازه گیری فاصله کسینوسی استفاده می شود. در این روش ، ماشین می‌تواند معنا و مفهوم جملات را با استفاده از روش کدگذاری مدل برت ،با در نظر گرفتن موقعیت موجودیت های به کار برده شده در جملات درک کند. و روابط معنایی اسناد را کشف و میزان ارتباط اسناد آنهارا با یک موضوع محاسبه و شباهت معنایی آن‌ها را تشخیص دهد. نتایج بدست آمده از نمونه کوچک تستی دیتاست نشان دهنده صحت و دقت روش پیشنهادی در تشخیص روابط معنایی اسناد قانونی مرتبط با مجلس شورای اسلامی می باشدودستیابی به میزان صحت و دقت عملکرد بالای ۹۰ درصد را نشان می‌دهد.