Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 13, No. 1, 2025, 85-94.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

DOSTE: Document Similarity Matching considering Informative Name Entities

Milad Allahgholi, Hossein Rahmani*, Amirhossein Derakhshan and Saman Mohammadi Raouf

School of Computer engineering, Iran University of Science and Technology, Tehran, Iran.

Article Info

Article History:
Received 14 December 2024
Revised 17 January 2025
Accepted 30 January 2025

DOI:10.22044/jadm.2025.15383.2641

Keywords:
Document Similarity, Name Entitie, Entities Graph.

* *Corresponding author: h_rahmani@iust.ac.ir (H. Rahmani).*

Abstract

Document similarity matching is essential for efficient text retrieval, plagiarism detection, and content analysis. Existing studies in this field can be categorized into three approaches: statistical analysis, deep learning, and hybrid approaches. However, to the best of our knowledge, none have incorporated the importance of named entities into their methodologies. In this paper, we propose DOSTE, a method which first extracts named entities using pre-trained models. Second, it constructs a graph of the entities to assign scores to each named entity based on its type, the distance between entities, or a combination of these factors. Finally, the calculated scores are utilized in the document similarity matching task. Empirical results indicate that DOSTE achieves better results by emphasizing named entities, resulting in an average improvement of 9% in the average recall metric compared to baseline methods. Notably, this improvement in recall is achieved without significantly reducing precision, suggesting that DOSTE is particularly suitable for applications where the completeness of retrieved documents is critical. Also, DOSTE unlike LLM-based approaches, does not require extensive GPU resources. Additionally, non-empirical interpretations of the results indicate that DOSTE is particularly effective in identifying similarity in short documents and complex document comparisons.

1. Introduction

With the ongoing advancements in technology and the exponential growth of data generated on the internet, information retrieval and data analysis have emerged as considerable challenges [1]. This data spans various types, including text, images, audio, and video, with a substantial share consisting of textual data. Textual information is derived from diverse sources such as websites, emails, chats, social media, and customer feedback. However, extracting meaningful insights from text has become a complex and time intensive process, primarily due to challenges such as processing time and inherent complexity [2, 3].

Document similarity matching is a subfield of natural language processing (NLP). NLP is a subfield of computer science and artificial intelligence that utilizes various algorithms and methods to process and analyze natural language. During the initial attempts at document similarity analysis, methods for feature extraction such as Term Frequency-Inverse Document Frequency (TF-IDF) [6], bag-of-words [4] and n-grams [5] were integrated with machine learning models like naive Bayes [8], support vector machines (SVM) [7], and K-Nearest Neighbors (KNN) [9] to carry out document similarity tasks. Advanced NLP methods have been developed to extract richer features from text, enabling a deeper understanding of its semantic content. For instance, approaches like sentence embeddings [10], attention mechanisms [11, 12], and transformer-based architectures [13] create more refined text representations that effectively capture the context and meaning of the language [14].Despite significant advancements in document similarity matching, many existing approaches face limitations in effectively handling the nuances of short texts, complex linguistic structures, and domain specific contexts. Traditional statistical methods often struggle to capture semantic relationships, while deep learning-based models require extensive computational resources and large labeled datasets. Addressing these challenges is essential for improving accuracy and efficiency in tasks such as similarity detection and complex document comparisons.

A key task in natural language processing involves determining the similarity between documents within a large dataset or assessing how closely two documents correspond to each other. [15]. This process significantly contributes to the development and improvement of text processing and artificial intelligence systems, with applications in various fields, including information retrieval [1], plagiarism detection [16], automatic question answering [17], dialogue systems [18] and machine translation [19].

In this study, we address the limitations of statistical methods, such as assuming term independence and ignoring the semantic dimension of sentences, as well as the limitations of deep learning methods, including the need for large volumes of data and extensive computational resources. To this end, we propose a method called DOSTE, which leverages the high importance of named entities in document. DOSTE first extracts named entities and then determines their importance based on their type or their distance from other entities. Finally, it incorporates the scores of these entities into the similarity matching process.

This study begins with an overview of previous works in the field of document similarity detection, and then the proposed method in this research is presented. Finally, we evaluate the results and analyze the performance of the proposed approach.

2. Related Work

Previous work on document similarity analysis can be categorized into three main approaches: statistical analysis, deep learning, and hybrid methods. Figure 1 illustrates the categorization of these existing methods.

Statistical analysis methods determine similarity based solely on word distribution patterns within documents, without considering the semantic dimension of words or potential relationships between them [20]. Bafna et al. [21] proposed a multi-step approach for document clustering using TF-IDF and clustering algorithms.

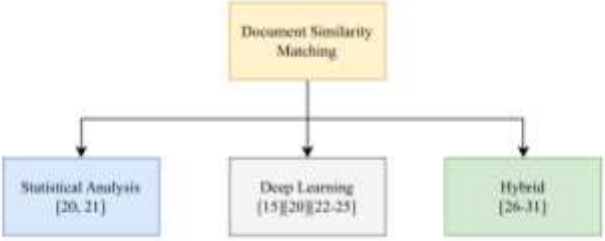
Document Similarity Matching
Statistical Analysis [20, 21]
Deep Learning [15][20][22-25]
Hybrid [26-31]

Figure 1. Document matching approaches can be categorized into three main types: statistical-based, deep learning-based, and hybrid-based approaches.

Their method involves preprocessing documents, creating a term-document matrix using TF-IDF, applying hierarchical agglomerative clustering and fuzzy K-means algorithms, and validating cluster quality using entropy and F-measure. Rashidi et al. [22] developed a fraud detection system based on feature selection and support vector machine (SVM) techniques. Their approach consists of three stages. In the first stage, preprocessing steps such as stop-word removal and lowercasing are performed. The second stage follows two paths: in the first path, paragraph comparisons are conducted using traditional methods, while the second path is based on calculating hyperplanes using SVM. Finally, in the third stage, it is determined which part of each document matches with sections of other documents. Moreover, other statistical methods have been used [20].

Deep learning approaches do account for the semantic aspects of words; however, they are often more costly than statistical methods and typically require large datasets, which may not be available in all languages [15]. Yang et al. [23] explored the task of long-to-long document matching. They introduced a model with a dual-stack network architecture, where each stack incorporates a multi-layer transformer based hierarchical encoder designed to capture document representations. Ostendorff et al. [20] framed the task of identifying relationships between two documents as a binary classification problem. Their experiments involved 32,168 pairs of Wikipedia articles. They selected nine key features from Wikidata for semantic document classification and tested six different approaches: AvgGlove, Doc2vec, Siamese BERT, Siamese XLNet, Vanilla BERT, and Vanilla XLNet. Ding et al. [24] proposed a new model, cogLTX, to solve the spatial and temporal complexity issue in BERT's self-attention mechanism for long documents. The model works based on the cognitive model of the human brain. It identifies key sentences by training a judgment model and combining them for reasoning, thus enabling multi-step reasoning. Wang et al. [15] found that existing methods for calculating semantic similarity in long texts lacked sufficient accuracy. To address this, they proposed an algorithm based on pre-trained deep learning models (BERT) for extracting semantic similarity in long texts, using deep learning to fine tune long text features. Anil et al. [25] proposed a knowledge-based retrieval system for engineering information, designed to overcome limitations of existing approaches, including challenges with similarity scoring in low-dimensional vector spaces, high execution times, and ineffective feature extraction. Richard [26] introduced the RecBERT recommendation system, which uses two main methods—domain adaptation BERT and fine-tuning for sentence modeling—to improve recommendation results. Using different domains and pre-trained base models, RecBERT aims to generate meaningful sentences for user reviews in online forums. Korade et al. [27] investigated sentence similarity using different embedding methods followed by various classifiers. Their experimental results indicate that sentence embedding with OpenAI embeddings achieved the best performance.

Hybrid methods incorporate both approaches, yet they still face challenges of high cost and the need for extensive data. Jha et al. [28] proposed the CoLDE method for long text document matching, addressing the challenge of interpretability by segmenting documents and using positional embeddings to capture structural information. Wang et al. [29] developed SECNN, a CNN-based approach for short-document classification. The method incorporates an attention mechanism to identify relevant words and leverages an external knowledge base to enrich semantic features. Finally, a classical CNN model is used to extract features and perform classification. Jha et al. [28] introduced a hierarchical recurrent neural network, based on a multi-depth attention mechanism, for semantic learning of long texts. This model, in addition to word information, uses the document structure to improve the representation of long documents. Farooq et al. [30] proposed a hybrid approach called HydMethod for measuring sentence similarity. This approach integrates various elements, including lexical databases, word embeddings, text set statistics, and word order information.

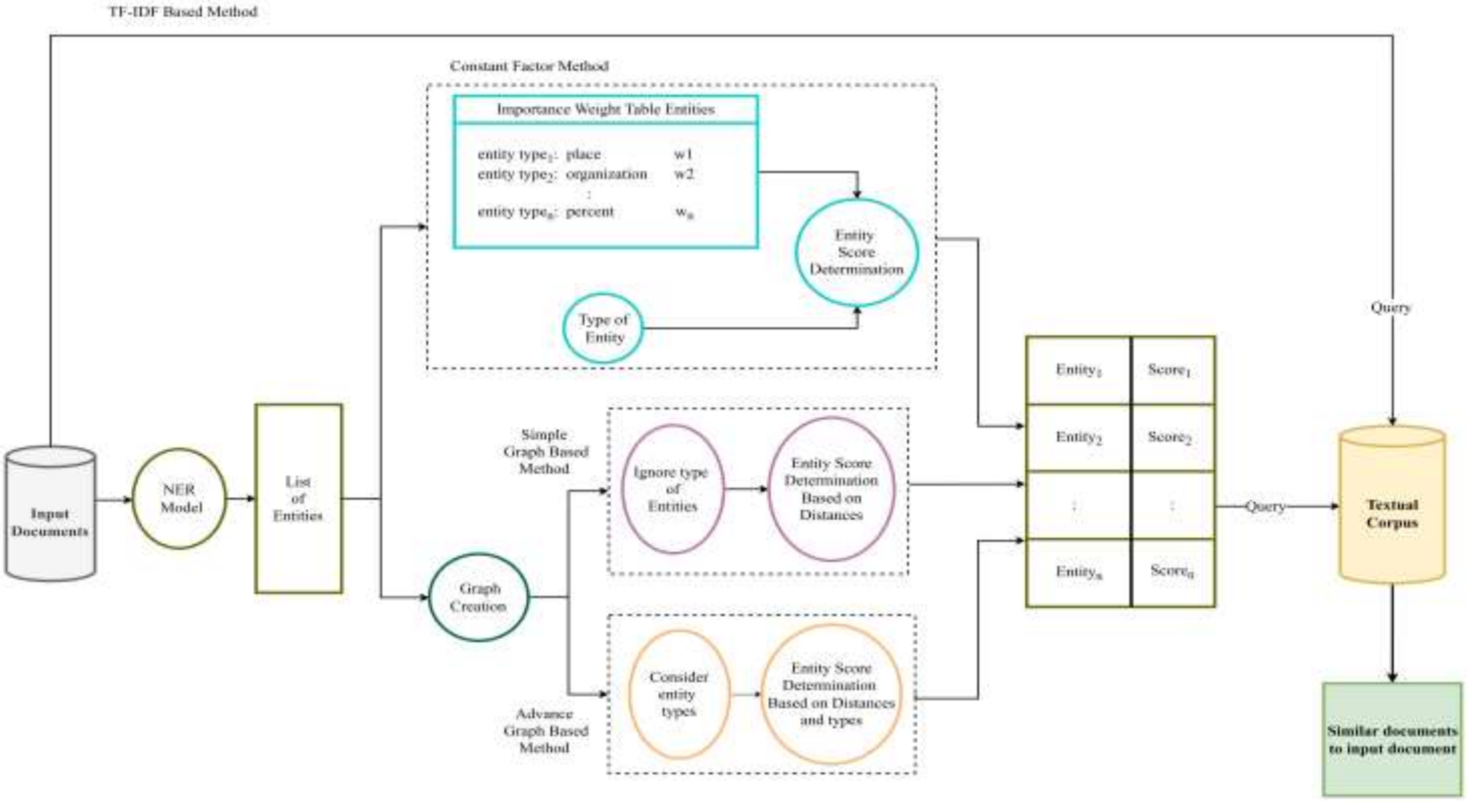
TF-IDF Based Method
Constant Factor Method
Importance Weight Table Entities
entity type$_1$: place w1
entity type$_2$: organization w2
entity type$_n$: percent w$_n$
Entity Score Determination
Type of Entity
Query
Input Documents
NER Model
List of Entities
Simple Graph Based Method
Ignore type of Entities
Entity Score Determination Based on Distances
Graph Creation
Advance Graph Based Method
Consider entity types
Entity Score Determination Based on Distances and types
Entity$_1$
Score$_1$
Entity$_2$
Score$_2$
Entity$_n$
Score$_n$
Query
Textual Corpus
Similar documents to input document

Figure 2. The DOSTE architecture incorporates three primary methods for document similarity matching, all centered on named entities within the text. These methods include named entity emphasis with fixed factors, a simple graph-based approach, and an advanced graph-based approach

It leverages common-sense knowledge encoded in lexical databases and is applicable to multiple text domains. Yu et al. [31] introduced the DDR-Match framework to enhance text matching performance, particularly for asymmetrical text pairs. This approach tackled the challenge of indistinguishable feature vectors in the semantic space, achieving improved matching results across various text domains. Viji and his colleague [32] found that traditional text similarity methods overlooked information about text meaning and the importance of word order and required large amounts of labeled data for training. They proposed a new method for improving semantic similarity detection, combining BERT feature extraction with Siamese bi-LSTM networks. Li et al. [33] introduced a new method that combines traditional retrieval approaches for lexical matching with newer transformer models like MPNet, Roberta, and MiniLM, which can capture stronger textual meanings.

Statistical approaches, such as TF-IDF, bag-of-words, or the use of handcrafted features, are employed to detect document similarity. Although these methods offer advantages such as simplicity, interpretability, and the ability to work with limited computational resources and small datasets, they have notable drawbacks. These include lower performance when detecting documents with complex patterns, the need for extensive feature engineering, and scalability issues, particularly when document representations require high-dimensional spaces. Deep learning methods, by employing more complex embedding techniques such as BERT, address issues like term independence and the need for precise feature engineering. However, the use of these models introduces challenges in terms of model interpretability. Additionally, training these models requires a large volume of data, and there is also the risk of overfitting during the training process. Hybrid methods, by combining statistical and deep learning approaches, are employed for document similarity detection. These methods have the potential to leverage the advantages of both statistical and deep learning techniques. However, achieving this goal comes with challenges such as the complexity of combining the methods, increased computational overhead, and the optimization challenges associated with using each approach effectively.

3. DOSTE: Our Proposed Method

This section proposes different methods used in DOSTE. To retrieve the most similar documents to a given document, four approaches were applied: (1) TF-IDF-based retrieval, (2) Named Entity Score Enhancement, (3) Simple Graph Construction, and (4) Advanced Graph Construction. A common feature across methods 2 to 4 is the extraction of named entities from the text and the boosting of their scores in the search process. However, each method differs in its approach to enhancing the score for each named entity. The architecture of the proposed method is illustrated in Figure 2.

3.1. TF-IDF Based Method

In this method, each word's score within the text is evaluated according to its TF-IDF (Term Frequency-Inverse Document Frequency) [21] score, without distinguishing between named entities and other words. This approach serves as a baseline method for comparison with other techniques.

3.2. The Named Entity Emphasis Method with Fixed Factors

In the first method of the DOSTE approach, named entities are extracted from the text, encompassing ten types: location, person, organization, date, time, money, percentage, facility, product, and event. For each entity type, DOSTE assigns an importance factor, which is configurable and can be adjusted using domain-specific knowledge to better suit specialized tasks. The score of each named entity is then multiplied by its corresponding importance factor based on its type. The importance factors for each type are listed in Table 1.

Table 1. Importance Factor for each Entity Type.

Entity Type	Importance Factors
Location	1.3
Person	1.5
Organization	2
Date	1.1
Time	1.1
Money	1.1
Percent	1.1
Facility	1.1
Product	1.2
Event	1.2

3.3. Simple Graph Based Method

In the second method of DOSTE, certain types of named entities are first extracted, and, similar to

Figure 3, a graph of the named entities in the text is constructed. In this graph, nodes represent the named entities, without distinguishing between different types. An edge is created between each pair of consecutive named entities in the text. The weight of each edge is then determined based on the distance between the named entities in terms of the number of words. The core idea of this method is that a shorter distance between two entities indicates a stronger relationship and higher significance. Therefore, as the distance between two entities decreases, their score should increase. The formulas for calculating the score of each entity in this method are presented in (1) and (2).

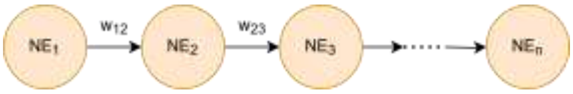

Figure 3. Named Entities Graph. entities are represented as nodes, with no distinction based on type, and the edge values correspond to the number of words separating them.

$$w'_{ij} = \frac{w_{ij} - \max(w)}{\min(w) - \max(w)}(\text{new\_max} - \text{new\_min}) + \text{new\_min} \qquad (1)$$

In (1), w_{ij} is the number of words between NE_i and NE_j , w shows the vector of all w_{ij} elements. So max(w) shows the largest distance between two sequential name entities in the text. Additionally new_min and new_max define minimum and maximum importance of entities. Using (1) the smallest w_{ij} is mapped to new_max and the largest w_{ij} is mapped to new_min.

$$Score\ \text{NE}_i = \max(w'_{i-1,i}, w'_{i,i+1}) \qquad (2)$$

In (2) $Score\ \text{NE}_i$ determines the score of i-th named entity, which is equal to the highest weight of the edge connected to that named entity after normalizing the weights using (1).

3.4. Advance Graph Based Method

In this method, similar to the previous approach, a graph of the named entities in the text is constructed. However, unlike the previous case, all types of entities are extracted, and the entity types also contribute to the scoring process. Figure 4 illustrates the graph structure based on entity types.

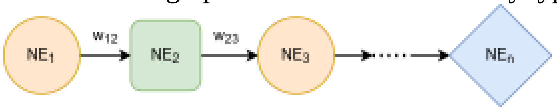

Figure 4. Advance Graph-Based method, Each figure type represents a distinct entity type, highlighting the differentiation between entity categories.

In this method, the score of each entity is influenced by all entities in the text, with shorter distances between entities leading to higher scores. However, this score increase also depends on the type of entities. When two entities are of the same type, the score boost is greater. Additionally, to prevent centrally located entities from gaining disproportionately high scores due to their proximity to others, the scores are gradually reduced as one progress through the named entities. The score for each entity in this method is calculated according to Equations 3 and 4. In (4), the DOSTE approach assigns a lower weight to pairs of entities of different types compared to pairs of entities of the same type (α <1).

$$Score\ \text{NE}_i = base\ \text{score} + \frac{1}{i}\sum_{j \neq i}^{N} \frac{f(type\ \text{NE}_i, type\ \text{NE}_j)}{\log(w_{ij}+1)} \qquad (3)$$

In (3), $Score\ \text{NE}_i$ represents the score of the i-th named entity. And base score defines the minimum score for each entity. We also use the logarithm of distances to prevent excessive reduction in scores.

$$f(type\ \text{NE}_i, type\ \text{NE}_j) = \begin{cases} 1 & \text{if type NE}_i = \text{type NE}_j \\ \alpha < 1 & \text{Others} \end{cases} \qquad (4)$$

In (4) the function f is introduced to adjust the influence of entities based on their types. This function takes the types of two entities as input and returns 1 if the types are identical; otherwise, it returns value α, which is less than 1.

4. Empirical Results

In this section, we evaluate DOSTE compared to the baseline TF-IDF approach. First, the dataset used will be described, followed by a presentation of the experimental results obtained.

4.1. Dataset

For the evaluation phase of the proposed document similarity methods, a range of datasets is commonly used, including plagiarism detection corpora, text similarity datasets, and QA datasets [16, 34-43]. In this study, the evaluation was conducted using the Persian plagiarism detection dataset, PAN2016 [44]. This dataset contains several textual documents designated as source documents and others labeled as suspicious documents potentially containing plagiarism. Plagiarism in this dataset can occur in three forms:

direct text substitution without modification, text substitution with minor changes—such as replacing certain words with synonyms—and text substitution with structural and more complex modifications intended to make plagiarism detection more difficult. For the evaluation of the proposed methods, the source documents in this dataset are considered as the original documents, while the documents containing plagiarism are treated as similar documents.

4.2. Comparing with base methods

Figure 5 shows the average recall metric for each method. For each method, the top n most similar documents to each source document were selected. The recall metric was then calculated by dividing the number of retrieved similar documents to the source document by the total number of similar documents to that source.

Figure 6 illustrates the results of various methods on the complex subset of the dataset where similarity has been created with structural and semantic modification.

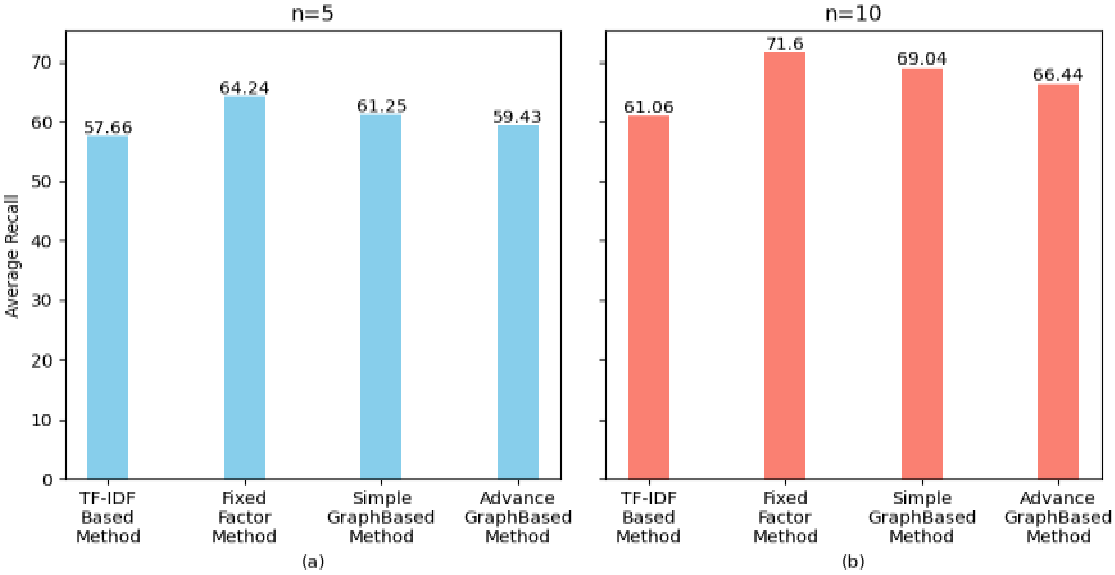

Figure 5. Average recall in the baseline method and proposed methods. (a) Average recall in the top 5 retrievals. (b) Average recall in the top 10 retrievals. The increase in the score of named entities with each proposed method indicates better performance than the baseline model.

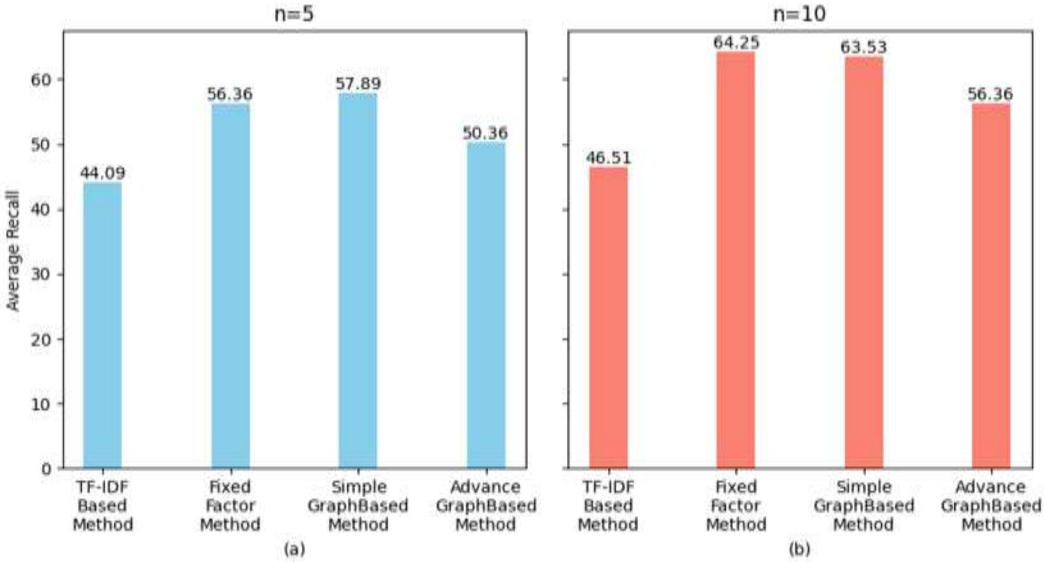

Figure 6. Average Recall Metric on the Complex Subset of the Dataset. (a) Average recall in the top 5 retrievals. (b) Average recall in the top 10 retrievals. The results show that the increase in the score of named entities leads to a smaller drop in recall for the proposed methods compared to the baseline method in the difficult portion of the dataset.

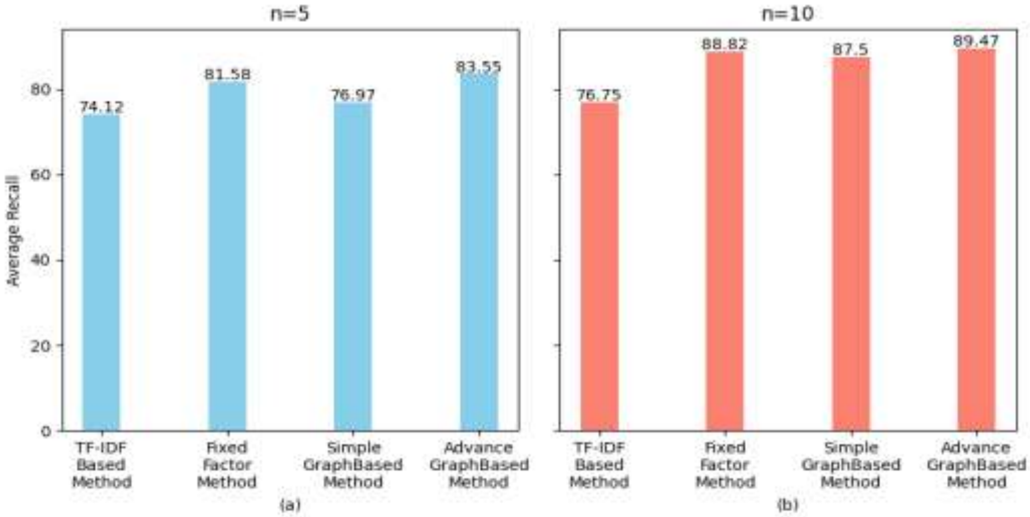

Figure 7. Average recall for 50 documents with shortest length as source document. (a) Average recall in the top 5 retrievals. (b) Average recall in the top 10 retrievals. The results show that, for short documents, the Advance Graph Based method yields the best result.

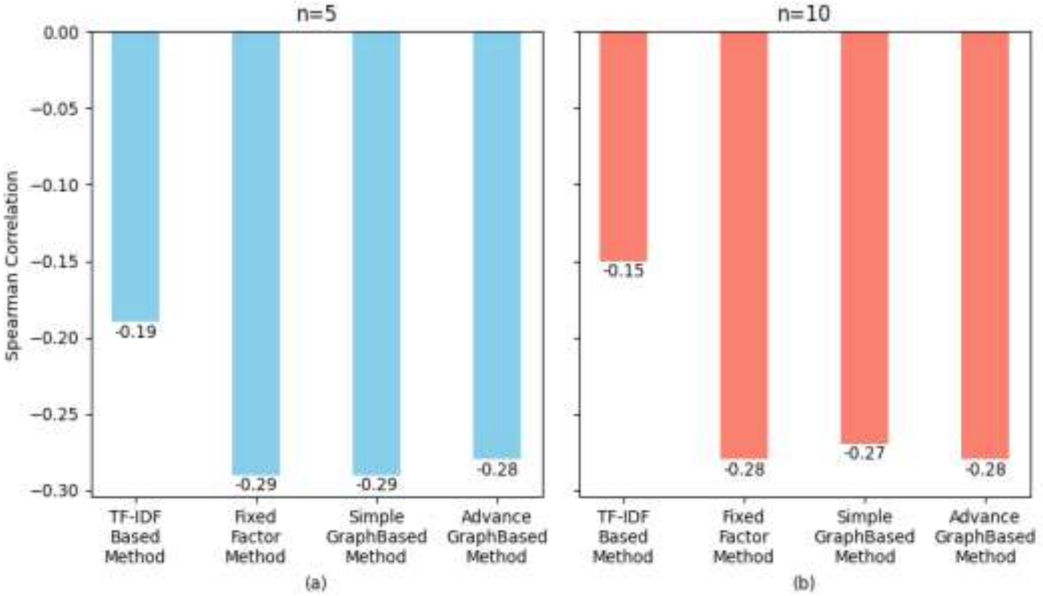

Figure 8. Spearman Correlation Between Document Length and Recall Metric Across Different Methods. (a) Correlation between recall and document length in the top 5 retrievals. (b) Correlation between recall and document length in the top 10 retrievals.

Given that in a lengthy document, substituting even a short sentence can constitute similarity and create a document similar to a source document, it is expected that shorter source texts will enhance the performance of graph-based models. Thus, in Figure 7, the average recall metric for the 50 documents with the fewest words in the dataset is presented. For a more precise evaluation of the dependency of each method's recall on document length, Figure 8 presents the Spearman correlation between the recall metric of each retrieval and the length of the corresponding document. As shown, in the graph-based methods, recall increases as document length decreases. Based on the obtained results, the performance of the graph construction method in DOSTE largely depends on the text length. When the text is longer, the number of named entities increases, which can result in the influence of irrelevant named entities on the similarity results. It was also observed that enhancing the scores of named entities improves the method's ability to detect complex similarities, leading to less accuracy decline compared to the baseline TF-IDF method. Additionally, in the equations introduced in Section 3, precise parameter tuning can significantly enhance the performance of each model.

5. Conclusions and Future Work

The detection of similar documents within extensive document collections remains a critical challenge in the field of document similarity detection, especially for tasks like plagiarism and academic fraud detection. This paper proposed DOSTE, a method that leverages the significance of named entities to enhance similarity assessment between documents. By extracting named entities and applying statistical and graph-based techniques to assign weights, DOSTE improves upon traditional methods. Our evaluations demonstrate that DOSTE not only achieves better recall—showing an average improvement of 9% over baseline methods—but also requires less computational time compared to large language model approaches like Siamese networks. Notably, the graph-based strategies within DOSTE are particularly effective for short texts and complex instances of academic fraud, underscoring the pivotal role of named entities in similarity detection.

Future research in document similarity detection can further enhance the DOSTE method by exploring the construction of entity graphs at different textual granularities, such as paragraphs or sentences, to capture more nuanced relationships within longer documents. Fine-tuning the weighting parameters for named entities using adaptive algorithms or machine learning techniques could improve the method's adaptability across diverse text types and domains. Integrating DOSTE with semantic embeddings from pre-trained language models may offer additional improvements in detection accuracy. Finally, incorporating domain knowledge into the weighting scheme could enhance similarity detection in specialized fields and domain specific tasks.

References

[1] P. Hambarde, "Information Retrieval: Recent Advances and Beyond," 2023.

[2] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning--based text classification: a comprehensive review," *ACM computing surveys (CSUR),* vol. 54, no. 3, pp. 1-40, 2021.

[3] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information,* vol. 10, no. 4, p. 150, 2019.

[4] M. W. Bilotti, P. Ogilvie, J. Callan, and E. Nyberg, "Structured retrieval for question answering," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 351-358.

[5] P. F. Brown, V. J. Della Pietra, P. V. Desouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," *Computational linguistics,* vol. 18, no. 4, pp. 467-480, 1992.

[6] C. Sammut and G. I. Webb, *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[7] S. Fatima and B. Srinivasu, "Text Document categorization using support vector machine," *International Research Journal of Engineering and Technology (IRJET),* vol. 4, no. 2, pp. 141-147, 2017.

[8] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some effective techniques for naive bayes text classification," *IEEE transactions on knowledge and data engineering,* vol. 18, no. 11, pp. 1457-1466, 2006.

[9] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved K-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications,* vol. 39, no. 1, pp. 1503-1509, 2012.

[10] N. Reimers, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *arXiv preprint arXiv:1908.10084*, 2019.

[11] C. Duan, L. Cui, X. Chen, F. Wei, C. Zhu, and T. Zhao, "Attention-Fused Deep Matching Network for Natural Language Inference," in *IJCAI*, 2018, pp. 4033-4040.

[12] C. Tan, F. Wei, W. Wang, W. Lv, and M. Zhou, "Multiway attention networks for modeling sentence pairs," in *IJCAI*, 2018, pp. 4411-4417.

[13] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021: IEEE, pp. 21-25.

[14] A. Fan, S. Wang, and Y. Wang, "Legal Document Similarity Matching Based on Ensemble Learning," *IEEE Access,* 2024.

[15] G. Wang, T. Zhang, G. Xu, Y. Zheng, Z. Du, and Q. Long, "A Deep Learning Based Method to Measure the Similarity of Long Text," in *2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, 2020: IEEE, pp. 173-178.

[16] F. Safi-Esfahani, S. Rakian, and M. Nadimi-Shahraki, "English-Persian Plagiarism Detection based on a Semantic Approach," Journal of AI and Data Mining, vol. 5, no. 2, pp. 275-284, 2017.

[17] N. Jiang and M.-C. de Marneffe, "Do you know that Florence is packed with visitors? Evaluating state-of-the-art models of speaker commitment," in Proceedings

of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4208-4213.

[18] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in Proceedings of the AAAI conference on artificial intelligence, 2016, vol. 30, no. 1.

[19] Q. Wang et al., "Learning deep transformer models for machine translation," arXiv preprint arXiv:1906.01787, 2019.

[20] M. Ostendorff, T. Ruas, M. Schubotz, G. Rehm, and B. Gipp, "Pairwise multi-class document classification for semantic relations between wikipedia articles," in Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, 2020, pp. 127-136.

[21] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016: IEEE, pp. 61-66.

[22] M. A. El-Rashidy, R. G. Mohamed, N. A. El-Fishawy, and M. A. Shouman, "An effective text plagiarism detection system based on feature selection and SVM techniques," Multimedia Tools and Applications, vol. 83, no. 1, pp. 2609-2646, 2024.

[23] L. Yang, M. Zhang, C. Li, M. Bendersky, and M. Najork, "Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching," in Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 1725-1734.

[24] M. Ding, C. Zhou, H. Yang, and J. Tang, "Cogltx: Applying bert to long texts," Advances in Neural Information Processing Systems, vol. 33, pp. 12792-12804, 2020.

[25] A. Sharma and S. Kumar, "Ontology-based semantic retrieval of documents using Word2vec model," Data & Knowledge Engineering, vol. 144, p. 102110, 2023.

[26] R. Wu, "RecBERT: Semantic recommendation engine with large language model enhanced query segmentation for k-nearest neighbors ranking retrieval," Intelligent and Converged Networks, 2024.

[27] N. B. Korade, M. B. Salunke, A. A. Bhosle, P. B. Kumbharkar, G. G. Asalkar, and R. G. Khedkar, "Strengthening Sentence Similarity Identification Through OpenAI Embeddings and Deep Learning," *International Journal of Advanced Computer Science & Applications,* vol. 15, no. 4, 2024.

[28] A. Jha, V. Rakesh, J. Chandrashekar, A. Samavedhi, and C. K. Reddy, "Supervised contrastive learning for interpretable long-form document matching," ACM Transactions on Knowledge Discovery from Data, vol. 17, no. 2, pp. 1-17, 2023.

[29] H. Wang, K. Tian, Z. Wu, and L. Wang, "A short text classification method based on convolutional neural network and semantic extension," International Journal of Computational Intelligence Systems, vol. 14, no. 1, pp. 367-375, 2021.

[30] F. Ahmad and M. Faisal, "A novel hybrid methodology for computing semantic similarity between sentences through various word senses," International Journal of Cognitive Computing in Engineering, vol. 3, pp. 58-77, 2022.

[31] W. Yu, C. Xu, J. Xu, L. Pang, and J.-R. Wen, "Distribution distance regularized sequence representation for text matching in asymmetrical domains," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 30, pp. 721-733, 2022.

[32] D. Viji and S. Revathy, "A hybrid approach of Weighted Fine-Tuned BERT extraction with deep Siamese Bi–LSTM model for semantic text similarity identification," *Multimedia tools and applications,* vol. 81, no. 5, pp. 6131-6157, 2022.

[33] P. Li, G.-J. Ren, A. L. Gentile, C. DeLuca, D. Tan, and S. Gopisetty, "Long-form information retrieval for enterprise matchmaking," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 3260-3264.

[34] F. Mashhadirajab, M. Shamsfard, R. Adelkhah, F. Shafiee, and C. Saedi, "A Text Alignment Corpus for Persian Plagiarism Detection," *FIRE (Working Notes),* vol. 1737, pp. 184-189, 2016.

[35] M. R. Sharifabadi and S. A. Eftekhari, "Mahak Samim: A Corpus of Persian Academic Texts for Evaluating Plagiarism Detection Systems," *FIRE (Working Notes),* vol. 1737, pp. 190-192, 2016.

[36] S. Abnar, M. Dehghani, H. Zamani, and A. Shakery, "Expanded n-grams for semantic text alignment," *Cappellato et al.[35],* 2014.

[37] K. Khoshnavataher, V. Zarrabi, S. Mohtaj, and H. Asghari, "Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation: Notebook for PAN at CLEF 2015," in *CLEF (Working Notes),* 2015.

[38] A. C. Marco, A. Myers, S. J. Graham, P. D'Agostino, and K. Apple, "The USPTO patent assignment dataset: Descriptions and analysis," 2015.

[39] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," *arXiv preprint arXiv:1705.03551,* 2017.

[40] A. Trischler *et al.*, "Newsqa: A machine comprehension dataset," *arXiv preprint arXiv:1611.09830,* 2016.

[41] Z. Yang *et al.*, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," *arXiv preprint arXiv:1809.09600,* 2018.

[42] D. D. Lewis, Y. Yang, T. Russell-Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *Journal of machine learning research,* vol. 5, no. Apr, pp. 361-397, 2004.

[43] D. D. Lewis, "text categorization test collection," ed: Tech. Rep., http://www. ics. uci. edu/~ kdd/databases/reuters21578 …, 2004.

[44] H. Asghari, S. Mohtaj, O. Fatemi, H. Faili, P. Rosso, and M. Potthast, "Algorithms and corpora for persian plagiarism detection: overview of PAN at FIRE 2016," in *Text Processing: FIRE 2016 International Workshop, Kolkata, India, December 7–10, 2016, Revised Selected Papers*, 2018: Springer, pp. 61-79.

دوستی: تطبیق شباهت اسناد با در نظر گرفتن موجودیت‌های نام‌دار حاوی اطلاعات مفید

میلاد اله‌قلی، حسین رحمانی*، امیرحسین درخشان و سامان محمدی رئوف

دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران.

ارسال ۲۰۲۴/۱۲/۱۴؛ بازنگری ۲۰۲۴/۰۱/۱۷؛ پذیرش ۲۰۲۴/۰۱/۳۰

چکیده:

تطبیق شباهت اسناد برای بازیابی کارامد اطلاعات، تشخیص سرقت ادبی و تحلیل محتوا ضروری است. مطالعات موجود در این زمینه را می‌توان به سه دسته تحلیل‌های آماری، روش‌های یادگیری عمیق و روش‌های ترکیبی تقسیم کرد. با این حال، با توجه به بهترین دانش ما هیچ کدام از مطالعات اهمیت موجودیت‌های نام‌دار را در روش‌های خود لحاظ نکرده‌اند. در این مقاله ما روش دو ستی را معرفی میکنیم. رو شی که ابتدا موجودیت‌های نام‌دار متن را با ا ستفاده از مدل‌های از پیش آموخته شده ا ستخراج می‌کند. سپس گرافی از موجودیت‌ها می سازد و به هر موجودیت بر ا ساس نوع آن، فا صله آن با دیگر موجودیت‌ها و یا ترکیب این دو امتیازی اختصاص می‌دهد. در نهایت از امتیازهای به دست آمده برای تطبیق شباهت اسناد استفاده می‌شود. نتایج تجربی نشان می‌دهد که دوستی با تاکید بر موجودیت‌های نام‌دار، باعث افزاش ۹٪ معیار میانگین فراخوانی بدون کاهش قابل توجه در معیار صحت می شود. این ویژگی نشان می‌دهد که دو ستی به ویژه در کاربردهایی که کامل بودن ا سناد بازیابی شده اهمیت دارد، منا سب است. علاوه براین برخلاف روش‌های مبتنی بر مدل‌های زبانی بزرگ دوستی نیازمند منابع پردازشی سنگین نیست. همچنین تحلیل کیفی نتایج نشان می‌دهد که دوستی در شناسایی شباهت اسناد کوتاه و مقایسه اسناد پیچیده عملکرد موثری دارد.

کلمات کلیدی: تطبیق شباهت اسناد، موجودیت‌های نام‌دار، گراف موجودیت‌ها.