Expert Systems With Applications 207 (2022) 118034

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Improving plagiarism detection in text document using hybrid weighted similarity

Hamed Arabi [a], Mehdi Akbari [a,b,*]

[a] Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran
[b] Big Data Research Center, Najafabad Branch, Islamic Azad University, Najafabad, Iran

ARTICLE INFO

Keywords:
Extrinsic plagiarism
Word Embedding Technique
Bag of Word Technique
Structural Similarity
FastText

ABSTRACT

Plagiarism is a misconduct, which refers to the use of scientific and literary content contained in other sources without reference to them. Today, the rise of plagiarism has become a serious problem for publishers and researchers. Many researchers have discussed this problem and tried to identify types of plagiarism; however, most of these methods are not effective in detecting intelligent plagiarism. In other words, most of these methods focus on direct copying. Therefore, in this study, two methods are proposed to identify Extrinsic plagiarism. In both methods, to limit the search space, two stages of filtering based on the bag of word (BoW) technique are used at the document level and at the sentence level, and plagiarism is investigated only in the outputs of these two stages. In the first method to detect similarities in suspicious documents and sentences, the combination of pre-trained network technique of words embedding FastText and TF-IDF weighting technique to form two structural and semantic matrices and in the second method to form the two matrices, WordNet ontology and weighting TF-IDF is used. After forming the above matrices and calculating the similarity between the pairs of matrices of each sentence, using the Dice similarity and the structural similarity of the weighted composition, two similarity values are calculated. By comparing the similarity of suspicious sentences with the minimum threshold, the document containing the suspicious sentence receives the label of plagiarism or non-plagiarism. Experimental results on the PAN-PC-11 database show that the first method has achieved 95.1% precision and the second method 93.8% precision, which shows that the use of word embedding network compared to WordNet ontology can be more successful in detecting Extrinsic plagiarism.

1. Introduction

Plagiarism occurs when a person uses the ideas, words and expressions without the required reference to them. Plagiarism is a common problem in academic fields, research articles, publications, inventions, etc. (Lovepreet & Kumar, 2019). Plagiarism is done in different ways; for instance, self-plagiarism (such as publishing an article in several magazines) or using other authors' texts which is included in plagiarism. Plagiarism can be observed in both academic and non-academic fields (Lovepreet & Kumar, 2019). Academic plagiarism is one of the most severe forms of research violations and affects the university and the public negatively. Research articles including plagiarism hinder the scientific process (Foltýnek, Meuschke, & Gipp, 2019). Wrong findings can expand and affect the next researches or scientific applications. For instance, in medicine or pharmacology, *meta*-studies are important tools for assessment of efficiency and safety of medicines and medical treatments. Plagiarized research articles can deviate *meta*-studies and consecutively endanger the patient's safety (Foltýnek et al., 2019). Furthermore, plagiarism wastes the scientific resources. Even in the best case, if the plagiarism was detected, review and punishment of plagiarized research articles and allowance requests would still cause much effort for jury, damaged institutes and sponsors (Foltýnek et al., 2019). If plagiarism was not detected, its negative effects would be even more severe. Plagiarists can receive unfair allowances and promotions as the sponsor agents may grant allowance for plagiarized ideas or accept the plagiarized research as research projects' results.

Motivation: The studies show that some plagiarized articles refer at least to the original version (Foltýnek et al., 2019). On the other hand, examining plagiarism in a suspicious document manually is an extremely difficult and time-consuming process for different source documents (Lovepreet & Kumar, 2019). Therefore, using computer systems that can do the process with minimum user interference is

* Corresponding author at: Big Data Research Center, Najafabad Branch, Islamic Azad University, Najafabad, Iran.
E-mail address: mehdi_akbari@pco.iaun.ac.ir (M. Akbari).

https://doi.org/10.1016/j.eswa.2022.118034
Received 2 February 2022; Received in revised form 15 May 2022; Accepted 30 June 2022
Available online 8 July 2022

considered as a significant step in this regard. The plagiarism detection tools which have been proposed so far have high capabilities in detection of different kinds of plagiarism; however, detection of existence of plagiarism in a text depends on human resource monitoring (Gillam & Vartapetiance, 2016).

Overall solution: Generally, the methods of plagiarism detection in monolingual model are divided into two categories of extrinsic (Mahdavi, Siadati, & Yaghmaee, 2014; Wang, Qi, Kong, & Nu, 2013) and intrinsic (Stein, Lipka, & Prettenhofer, 2011; Zu Eissen & Stein, 2006). In extrinsic methods, the suspicious document is compared to other documents and in intrinsic methods, it is examined alone (Elhadi & Al-Tobi, 2009; Stein, zu Eissen, & Potthast, 2007). One of the important problems in both methods is that the person expresses the plagiarized section using synonym words or changing sentence structures and words movement; in this case, plagiarism detection is not that easy (Joseph & Haroon, 2016).

Main contribution: Therefore, in this article, two methods of extrinsic plagiarism detection have been proposed in a way that it can remove the mentioned limitation. In line with this, to improve plagiarism detection in both proposed methods, formation of semantic vector of sentences based on tow techniques of FastText pre-trained words embedding network and WordNet ontology have been used so that the problem of structural change or using similar words in plagiarism can be solved to some extent. Moreover, structural matrix based on weighting criterion such as TF-IDF (Chen, Zhang, Long, & Zhang, 2016; Dogan & Uysal, 2019) has been used. In both of these methods, it is tried to limit the problem search space using two-stage filtering based on Bag of Words technique at document and sentence level and to examine the output of these two stages for plagiarism detection. In both methods, if each pair of suspicious document and source passed the document level filtering, they would be analyzed into sentences and in case the pair of their sentences passed the sentence level filtering, their matrices would be compared regarding common and non-common words and using several similarity criteria and finally, the similar sentences are determined as sentences with plagiarism. The main sections of the proposed methods in this article are summarized as follows:

- Proposing two systems of extrinsic plagiarism detection based on structural and semantic hybrid similarity of sentences and using words weighting technique to form sentence structural matrix.
- Using FastText pre-trained network with more than 16 billion tokens to form sentence structural matrix.
- Using WordNet ontology to form sentence semantic matrix
- Defining two filtering stages at document and sentence level using Bag of Words technique to limit the search space

In the following, in section 2, the theoretical background including the studies in this field and the structure of literary and scientific plagiarism systems have been addressed. In section 3, the proposed model has been introduced and in section 4, the results have been reported. In the final section, conclusion and further studies have been presented.

2. Theoretical background and related works

2.1. Plagiarism

Plagiarism with its present meaning was introduced in the early 1990s for the first time (Brin, Davis, & Garcia-Molina, 1995; Shivakumar & Garcia-Molina, 1995). Plagiarism refers to the use of other authors' document without changing them or referring to them. To be more precise, plagiarism means illegal adaptation of other documents and this also includes reuse of the same author's text (Ceska, Toman, & Jezek, 2008; Potthast, Stein, Barrón-Cedeño, & Rosso, 2010).

Plagiarism has different forms including exact copy of a scientific document without reference and exact translation of a text from another language which are both in the category of plagiarism in documents. Idea plagiarism and plagiarism in code are among other kinds of plagiarisms (Gruner & Naven, 2005; Lukashenko, Graudina, & Grundspenkis, 2007). Plagiarism is also very important in other fields such as music due to the huge amount of money that music produces; Which is not covered in this article (De Prisco, Esposito, et al., 2017; De Prisco, Malandrino, Zaccagnino, & Zaccagnino, 2017).

- **Plagiarism in documents**: most of the conducted studies and plagiarism systems proposed in documents have focused on scientific and academic plagiarism. These methods are used for academic documents. As shown in Fig. 1, plagiarism in documents is divided into two main categories of exact copy or literal and modified copy or intelligent.

Exact copy or literary plagiarism that is known as naive plagiarism refers to the case that the author inserts another author's text without reference. Naïve plagiarism is done in the minimum possible time by the profiteer (Alzahrani et al., 2012). In Fig. 2, a sample of exact copy plagiarism has been shown. In modified copy or intelligent plagiarism, the profiteer tries to do plagiarism by manipulating and modifying the initial source text. Modified plagiarism can be done for instance by replacing synonym similar words like the words of source text; therefore, detection of this method is more difficult than exact copy plagiarism.

Another kind of plagiarism that is included in intelligent plagiarism category is exact translation from document in another language without reference (Alzahrani et al., 2012). In Figs. 3 and 4, examples of intelligent plagiarism have been shown.

In Fig. 5, the techniques of plagiarism in documents detection are categorized from another viewpoint including two main categories of cross-lingual (Agarwal, 2019; Al-Suhaiqi, Hazaa, & Albared, 2018; Muneer & Nawab, 2021; Zubarev & Sochenkov, 2019) and monolingual (Ahnaf, Saha, & Hossain, 2020; Alvi, 2020).

The methods of detecting monolingual plagiarism refer to the models in which the suspicious and source documents belong to a language like Arabic, English, Persian, etc. (Elamine, Bougares, Mechti, & Belguith, 2019; Meuschke, Stange, Schubotz, & Gipp, 2018). Most of the methods proposed so far belong to this category. The models of monolingual plagiarism detection are in two forms of extrinsic (Belguith, 2021; Boukhalfa, Mostefai, & Chekkai, 2018; Muhammad, 2020) and intrinsic (Polydouri, Vathi, Siolas, & Stafylopatis, 2020; Potthast et al., 2012; Saini, Sri, & Thakur, 2021). The methods of extrinsic monolingual plagiarism detection compare the suspicious document to a set of candidate source documents in which there are probably some parts of the suspicious document for detecting the plagiarized parts (Stein et al., 2007). In the methods of intrinsic monolingual plagiarism detection, the suspicious document is compared to itself. In fact, the source documents are not in focus in these methods, but the focus is on the writing style of a document in order to detect plagiarism (Gunawan, Krisnawati, & Chrismanto, 2020). Cross-lingual plagiarism refers to the one in a translated text. In this kind of plagiarism, the source text and the plagiarized one belong both to one language; for instance, the source text can be in English and the plagiarized one in Arabic, Urdu, Turkish, Persian, etc. (Aravind, Shanmukh, Charan, & Nellikoppad, 2020; Barrón-Cedeno, Rosso, Pinto, & Juan, 2008; Pinto, Civera, Barrón-Cedeno, Juan, & Rosso, 2009; Potthast, Barrón-Cedeño, Stein, & Rosso, 2011).

2.2. The detection process of plagiarism in documents

Plagiarism detection is a two-stage process; in the first stage, the documents with plagiarism are detected and in the second step, the precise location of plagiarism in documents is determined. The plagiarism detection systems are precisely proposed in two sub-categories of candidate retrieval and text alignment. The plagiarism detection

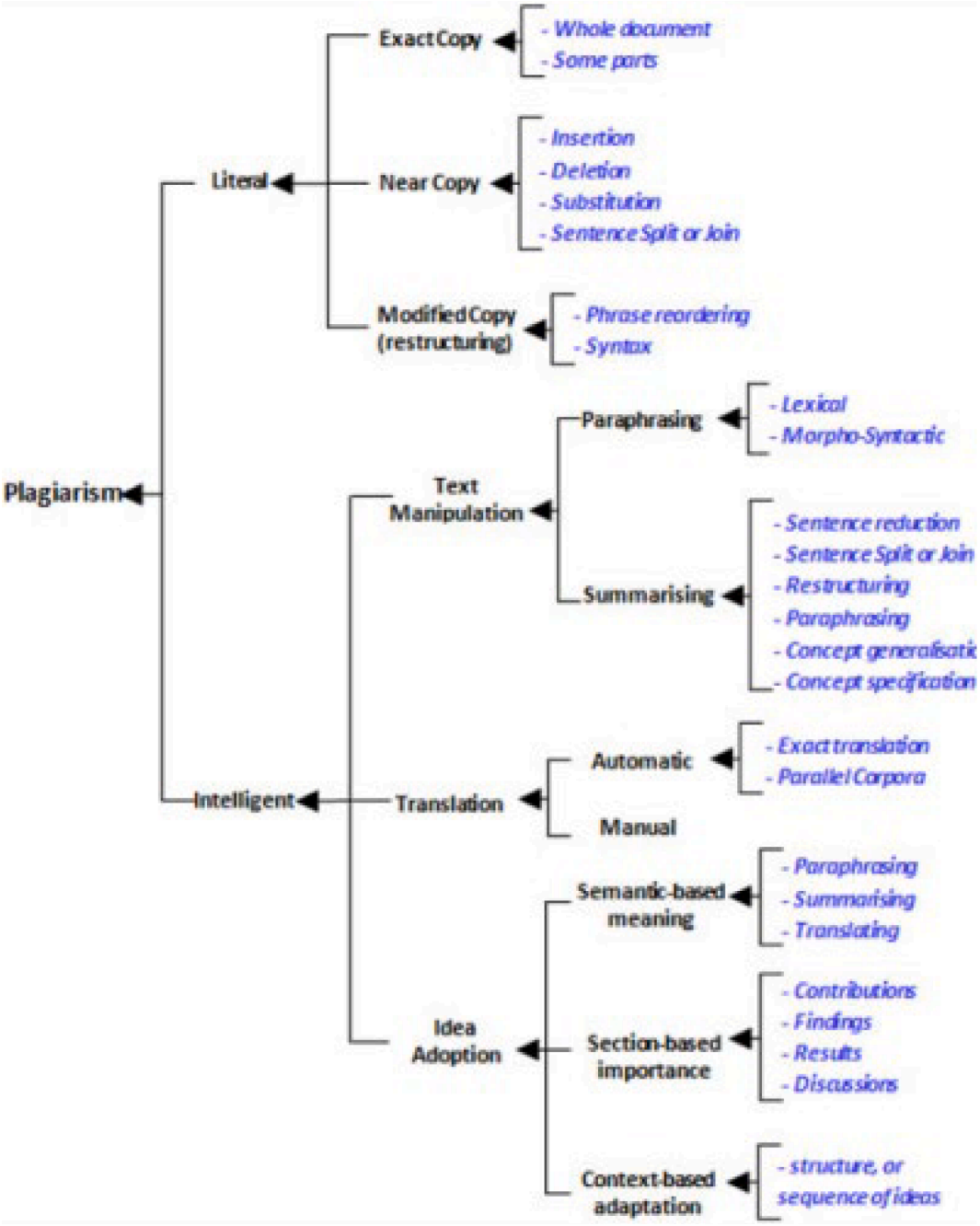

Fig. 1. Different kinds of plagiarism in documents (Alzahrani, Salim, & Abraham, 2012; Wakil, Ghafoor, Tariq, & Ahmad, 2017).

methods proposed in candidate retrieval phase have only the task to find similar documents to the suspicious one. The systems proposed in text alignment phase have the task to determine the precise location of plagiarism. As shown in Fig. 6, generally these systems receive as input a query document that is a suspicious file and they search in the dataset that can be local or web-based and they determine as output, the plagiarized parts of the suspicious document (Alzahrani et al., 2012).

2.3. WordNet ontology

Ontology which is called cosmology in some sources is an official structural display of knowledge in a given domain. In the structure of an ontology, the concepts of a domain and the relationship between the concepts are displayed in the best way (Fensel, 2001). To be precise, cosmology shows the words' meanings and also, their relationships and provides easier understandability (Staab & Studer, 2010).

WordNet is one of the practical general ontologies that is a large electronic words database for English and it was created in 1986 in Princeton University and it is still developed and maintained there (Fellbaum, 2010). This ontology is made of nouns, adjectives, adverbs and verbs. In WordNet ontology's structure, the synonym words are placed in groups under the name of Synset and each Synset group provides different concept. One of the main applications of WordNet is extracting similarity or dissimilarity of the words that is done using predefined relations. There are different relations among Synsets in the hierarchical structure of this ontology. The main relations in this ontology are Hyponym/Hypernym and Meronym/Holonym which are also known as "Is-A" and "Part-of" (Slimani, 2013). A sample of these two relations has been shown in the following. Regarding WordNet structure and the relationships between its identities, the calculation

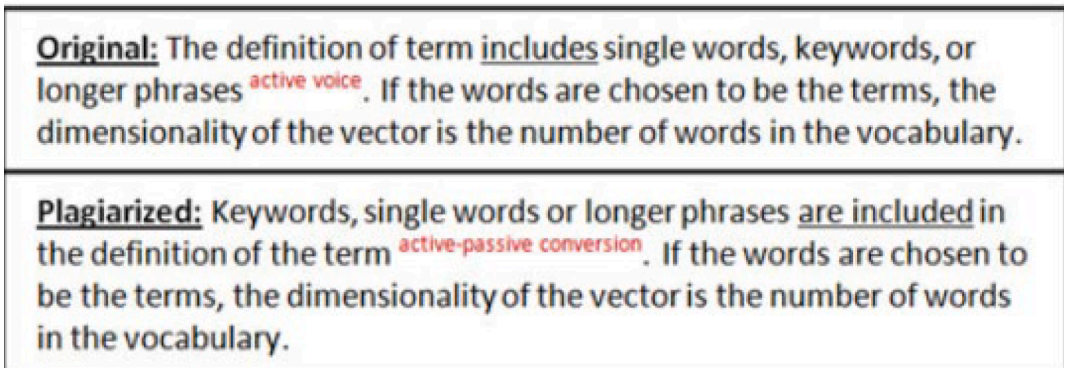

Fig. 2. An example of plagiarism with exact copy method (Alzahrani et al., 2012).

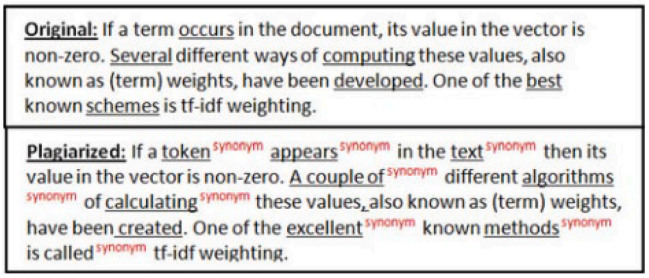

Fig. 3. Intelligent plagiarism with paraphrasing method (Alzahrani et al., 2012).

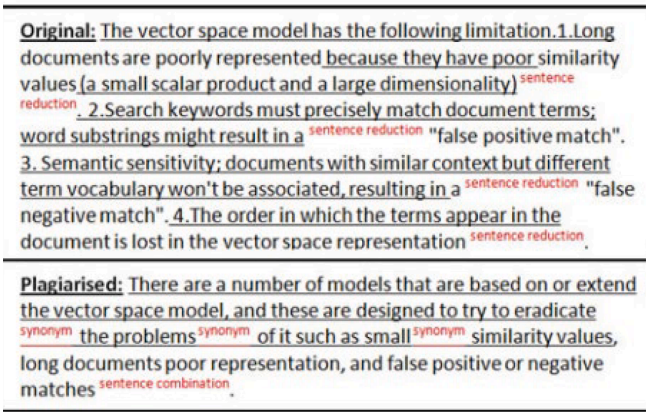

Fig. 4. Intelligent plagiarism with summarization method (Alzahrani et al., 2012).

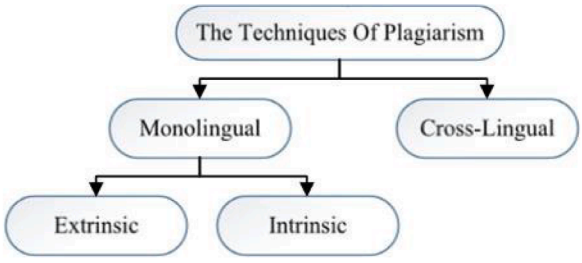

Fig. 5. Categorization of plagiarism in documents from language viewpoint.

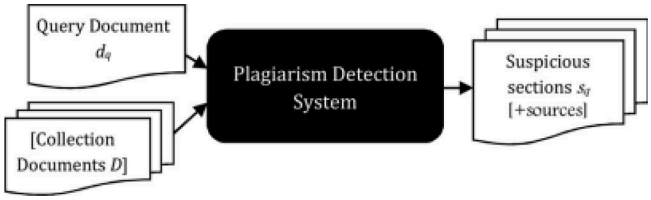

Fig. 6. A general system of plagiarism detection (Alzahrani et al., 2012).

criteria of similarity or dissimilarity of words are 4 main categories of structural, feature-based, hybrid and information-based (Meng, Huang, & Gu, 2013).

{X is a hypernym of Y if Y is a type of X}: Color is a hypernym of red.

{X is a hyponym of Y if X is a type of Y}: spoon is a hyponym of cutlery.

{Y is a meronym of X if Y is a part of X}: window is a meronym of building.

{Y is a holonym of X if X is a part of Y}: building is a holonym of window.

2.4. Vector space and TF-IDF weighting model

Vector space model is one of the most common information retrieval models (Salton, Wong, & Yang, 1975). In this method, to calculate the similarity or dissimilarity rate between the suspicious document and the other documents in database, first, all documents are mapped to an n-dimensional space and n shows the number of words in the whole database (Raghavan & Wong, 1986; Wong & Raghavan, 1984). In this model, the vector of each document is made of the words of that document (Raghavan & Wong, 1986; Wong & Raghavan, 1984). The amount and value of each word is defined using weighting criterion such as TF-IDF (Aizawa, 2003). To calculate the similarity among the documents, similarity or dissimilarity criteria such as Jaccard, Cosine and Euclidian are used (Huang, 2008). Among the similarity criteria, the cosine ones are more common in which the angle between document vectors are used to calculate similarity (Turney & Pantel, 2010). In TF-IDF, the words of more significance have more weights and vice versa. TF-IDF criterion is obtained from Eq. (1) (Ramos, 2003).

$$W_{tfidf_t} = Tf_t \times IDF_t where Tf_t = rawcount, IDF_t = \log\frac{N}{n_t} \quad (1)$$

In the equation above, N is the total number of documents in the database and n_t shows the number of documents including the word t. Moreover, in vector space model, Eq. (2) is used to calculate cosine similarity between two vectors (Boukhalfa et al., 2018).

$$Cosine_{similarty}(d,q) = Cos(\theta) = \frac{d \times q}{\|d\| \times \|q\|} = \frac{\sum_{i=1}^{n} d_i.q_i}{\sqrt{\sum_{i=1}^{n} d_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}} \quad (2)$$

In the equation above, the numerator is inner product of two vectors and denominator is multiplication of magnitudes. d_i and q_i are the components of two vectors of d and q. In this criterion, the angle will not be bigger than 90°. Therefore, the output is an amount between 0 and 1 and the nearer is cosine similarity to 1, it means the documents are more similar (Giller, 2012; Sidorov, Gelbukh, Gómez-Adorno, & Pinto, 2014).

2.5. Embedding words technique

Embedding words technique maps each word of the text to a vector of numbers; this vector shows the word concept based on its content. One of main strengths and advantages of this technique is that the formed vectors for words with nearer meanings or near in some sense are more similar and in contrast, the unrelated words have less similarity (Khattak et al., 2019). In Fig. 7, the performance of embedding words technique has been shown. In this figure, the input vector is one-hot and the output vector has been produced by embedding words. The process of word vector production in words embedding network is in a way that first, changing the text to one-hot encoding is done. In the vectors of one-hot encoding, each word is mapped in a list including V single words to a single index in the vector. Therefore, each word is displayed with a vector of 0 and 1 instead of the given words. Words embedding methods change the solitude vectors of one-hot encoding into dense display which have two main features: 1) the vectors' dimensions are smaller than the word's size and 2) the vectors' components embrace the latent semantic (Khattak et al., 2019; Kim, Park, & Lee, 2020).

One of main and most common word embedding technique is word2vec nervous system. This network changes the text into vectors including the meaning and the relationship among words. The performance of word2vec nervous system with low depth depends on two parameters of vector dimension and the maximum window size between the given word and the words around it in a sentence. Two techniques of skip-gram and Continuous Bag of Words are used for teaching and learning (configuring) this nervous system. The difference between these two educational methods is that skip-gram method uses reference word to predict the words around it; however, CBoW predicts the current or reference word using the words around (Kim et al., 2020). In Fig. 8, the difference between these two methods has been shown clearly. The main advantage of word2vec nervous system in comparison to other methods of changing text to number such as Bag of Words, vector space model based on TF-IDF and latent semantic analysis, etc. is its capability in detecting similar words (Yilmaz & Toklu, 2020).

2.6. Related works

In this section, several cases of studies in plagiarism filed will be examined.

Zubarev et al. in (Zubarev & Sochenkov, 2019) proposed a cross-lingual plagiarism detection for Russian and English languages. Their proposed method is a plagiarism detection method of text alignment which focuses on plagiarism detection based on translation. They have proposed three models; the first one is based on different text similarity score which uses word embedding. The second model expands the

$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

Fig. 7. Words embedding model.

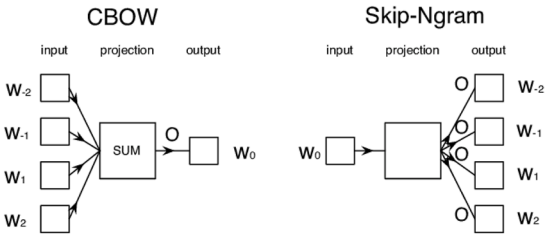

Fig. 8. Two educational methods of Word2vec words embedding network (Ling, Dyer, Black, & Trancoso, 2015).

previous one with the obtained features through nervous machine translation. The third model has been made of the fine tune of pre-trained model of Bert language display. Bert model shows high efficiency and works better than the other models. However, it needs more computing resources comparing easier models. Ezzikouri et al. in (H Ezzikouri, Oukessou, Erritali, & Madani, 2019) proposed a cross-lingual Fuzzy plagiarism detection for Arabic documents. In this method, Fuzzy semantic similarity approaches are used. Furthermore, to find semantic relationship among the concepts, lin, lch and wup similarities have been used. This method has been done by Apache Hadoop MapReduce in a parallel way. The proposed method extracts the semantic and syntactic information from text documents, then the semantic similarity among the words is calculated in the suspicious and source documents using path and depth similarity. Finally, dice similarity criterion is used as matrix similarity criterion to find semantic similarity among the sentences. This method was assessed on 600 collected documents by the authors in Arabic and English and the results show that their method could achieve more than 60% similarity. In another research, Norman et al. in (Meuschke, Stange, Schubotz, Kramer, & Gipp, 2019) proposed a plagiarism detection method in academic texts that uses current references and mathematical relations in the text to examine similarity. Norman has expanded his previous research in the field of mathematic content and academic documents in this article. In this method, a two-stage identification process has been proposed that is a combination of mathematical content similarity evaluation of academic and text documents. Moreover, in this method, a new similarity criterion was proposed which considers mathematical characteristic order. The results of this method shows that hybrid analysis result from mathematical content features and based on reference provides identification possibility of suspicious potential cases. In this method, 400 STEM documents have been assessed; the results show that the mathematical relations, references and text could achieve 70%, 90% and 90% recall respectively in plagiarism detection. Saleh et al. in (Altheneyan & Menai, 2020) proposed a method of obfuscation plagiarism detection. Obfuscated text is especially one of the most difficult kinds of plagiarism; the proposed method in this article has addressed the detection of this model of plagiarism. In this article, an automatic plagiarism detection system based on back-up vector machine classification has been proposed which uses a set of lexical, syntactic and semantic features. The proposed system performance in English and Arabic data, PAN 2012, PAN 2014, PAN 2013 and PAN@FIRE2015 has been examined. The results show that the proposed system has had the best performance in terms of criterion F in PAN 2012 and could achieve criterion F of 84.56%. Roostaee et al. in (Roostaee, Fakhrahmad, & Sadreddini, 2020) proposed a method of plagiarism detection. In this article, a two-level adaptation approach aiming at considering syntactic and semantic information to determine the exact location of plagiarism parts from the source and suspicious documents. At the first level, the document word vector is made using local weighting technique; this section is done aiming at limitation of number of pairs of suspicious and source documents. In this method, modeling the words and their relations, an acceptable increase is also observed in the system accuracy.

Experimental results in different datasets like PAN-PC-11, PAN-PC-12 and SemEval-2017 show that the cross-lingual text alignment approach in this article has significantly better performance than advanced models. This method could achieve the recall between 88.6% and 99.4% in different documents. This method had a better performance in Spanish and English data comparing German and English data. Ahanf et al. in (Ahnaf et al., 2020) believe that there are several methods for plagiarism detection in English and other Western languages. However, non-plagiarism detection tool is available to support plagiarism detection in Bangali language. However, Bangali language is one of the most popular languages of world and is used widely in internet. Therefore, they developed the first plagiarism tool for Bangali language. To reduce the domain complexity in this article, they decreased their domain to only educational books. Moreover, they proposed a set including National Curriculum and Text Books (NTBC) of Bangladesh from I to XII class. Their proposed method shows accuracy of 96.75%. Ahuja et al. in (Ahuja, Gupta, & Kumar, 2020) show that this method, forming two semantic and structural matrices tries to find similar sentences and detect plagiarism in the sentences with extrinsic method. In this method, the structural matrix of suspicious and source sentences is formed by words' location and word semantic matrix is defined regarding the depth and shortest path among the words. Therefore, this causes the hybrid proposed method by them to achieve the total rank of 87% in plagiarism detection in a very limited number of documents of PAN-PC11 database that is not a desirable result. Furthermore, according to the report of authors, this method does not have the capability of plagiarism detection in the manipulated texts. This method also does not include the candidate retrieval stage and search space limitation and this caused it not to be efficient in big databases. In 2015, Abdi et al. Used the WordNet ontology and the location of words to detect external plagiarism in English. In this method, two semantic and sequential vectors are formed for each sentence and plagiarism is discovered. One of the differences between this method and one of the proposed methods is the formation of structural vectors. Simply put, the proposed method uses the TF-IDF weight of the words instead of their location. On the other hand, the proposed method has two levels of filtering that uses the Bag of Words technique for this phase. Also, the second proposed method is based on the FastText pre-trained word embedding technique, which is another difference with the Abdi model (Abdi, Idris, Alguliyev, & Aliguliyev, 2015). The above model is one of the models that are compared with the proposed method in the results section. Roostaee et al. in (Roostaee, Sadreddini, & Fakhrahmad, 2020) proposed a plagiarism detection and cross-lingual candidate document retrieval method for Spanish, English and German-English. Regarding the significance of document retrieval stage in plagiarism detection, their study focuses on the task of candidate document retrieval and its purpose is extracting at least the potential source document set accurately. This article has proposed a combination of retrieval models based on concept and keyword for this purpose. A dynamic interpolation factor in the proposed method has been used to combine the results of conceptual models and a set of words. The efficiency of the proposed model has been compared on German-English and Spanish-English partitions. The results show the proposed candidate retrieval model acts better than advanced models and can be considered as an appropriate choice to embed in cross-lingual plagiarism detection systems. This model was examined in three databases of PAN-, RC-Acquis, Wikipedia and PC11 and has achieved in the best state F1 criterion of 76.01%. Gharavi et al. in (Gharavi, Veisi, & Rosso, 2020) used text embedding vectors to compare the similarity of documents in order to detect plagiarism. The word vectors are combined with a simple accumulation function to show a text document. This display includes the semantic and syntactic information of the text and results in efficient text alignment among suspicious and original documents. Comparing the sentence representation of source and suspicious documents, the pair sentences are considered as plagiarism candidate with maximum similarity. They applied their proposed method on datasets of PersianPlagDet, PAN-PC2014 and in English, Persian and Arabic in text alignment section to evaluate the robustness of the proposed methods from language viewpoint. They achieved the total rank of 89.67% in English and 97.11% in Persian and 87.90% in Arabic. Lazemi et al. in (Lazemi & Ebrahimpour-Komleh, 2020) proposed a plagiarism detection method in Persian language called ParsiPayesh. The main purpose of this article is detecting plagiarism and exact copy in Persian scientific texts. In their proposed method, after retrieval of candidate documents based on statistical features, structural and semantic analysis has been done in text alignment stage to identify exact copy plagiarism. Structural similarity rate of the phrase has been evaluated through dependency tree analysis. In this article, the semantic role labeling obtained from deep learning model has been used to examine semantic similarity. The test on the prepared body in AAIC2015 competitions and the body of PAN2015 competitions have been done and have achieved F1 criterion of 84.14%. The results show that the structural and semantic information improve the performance of their method. Vaz et al. in (Vaz, 2021) proposed a cross-lingual plagiarism detection method using cross-lingual pre-trained contextualized embeddings network of Bert. Contextualized embedding can help examine the fundamental features of language such as polysemy and synonymy considering the context in which a word occurs. The multilingual pre-trained models have shown an excellent performance in understanding natural language such as detection of sentence similarity and prediction of the next sentence. Regarding the promising results of these techniques, they proposed a new method to detect multi-lingual plagiarism detection using contextualized embedding pre-trained multi-lingual models. The tests conducted on different datasets such as PAN-PC-12 show that their method for pair language of English and German has achieved 72.51%. Alvi et al. in (Alvi, Stevenson, & Clough, 2021) in 2021 proposed some methods to identify two kinds of significant plagiarism including synonym replacement and word different order. They proposed a three-stage approach which uses content adaptation and pre-trained word embedding of Concept Net to identify synonym replacement and word different order. Their proposed approach shows that using Smith Watermen Algorithm to detect plagiarism and pre-trained word embedding of Concept Net Number batch has the best performance in terms of F1 scores; they could achieve F1 criterion of 80.8% in database of Plagiarized Short Answers. A summary of the above articles is reported in Table 1.

3. The proposed extrinsic monolingual plagiarism detection method

Two proposed models in this research are two extrinsic monolingual plagiarism detection models i.e., to detect plagiarism, the suspicious and source documents are compared and the domain of these two methods is English language. In these two methods, WordNet Ontology and pre-trained word embedding network of FastText are used to form semantic matrix and weighting method of TF-IDF is used to form structural matrix. The diagram of two proposed models has been shown in Figs. 9 and 10. In the first diagram, the general stages of two models and in the second one, the detailed stages of them have been displayed.

3.1. Pre-processing

The first stage of two proposed model is preparation and pre-processing of text. This stage includes different sub-steps. The following stages are done to pre-process and prepare the suspicious and source documents' text:

1) First, the suspicious and source documents are analyzed into their sentences. It should be noted that in the first step of text pre-process, the sentences with length of lower than 50 are omitted considering the spaces and punctuations.

Table 1
Summarizes related studies.

Reference	Method	Database	Advantages and Disadvantages	Result
(Zubarev & Sochenkov, 2019)	interlanguage for Russian and English languages	Collected Russian English	+ Provide a new database to detect plagiarism between English-Russian for text alignment	–
(Hanane Ezzikouri, Madani, Erritali, & Oukessou, 2019)	Fuzzy interlanguage plagiarism	Arabic-English documents collected	+ Consider semantic similarity of words using WordNet	Accuracy:60%
(Meuschke et al., 2019)	A combination of assessing the similarity of the mathematical content of academic documents and the text	STEM Doc.	+ Consider document references to detect plagiarism	Recall:90%
(Altheneyan & Menai, 2020)	Ambiguous plagiarism based on support vector machine classification	PAN 2012, PAN 2013, PAN 2014 and PAN@FIRE2015	+ Use of a set of lexical, syntactic and semantic features, English Arabic documents	Recall:84.56%
(Roostaee, Fakhrahmad, et al., 2020)	Detection of interlinguistic plagiarism with a two-level matching approach	PAN-PC-11 ، PAN-PC-12 و SemEval-2017	+ Review in German-English and Spanish-English	Recall:97%
(Ahnaf et al., 2020)	They developed the first plagiarism detection tool for the Bangladeshi language	educational books	- Semantic similarity is not considered. + Focused on Bengali language.	Precision:96.75%
(Roostaee, Sadreddini, et al., 2020)	Retrieve interlanguage candidate documents for German-English and Spanish-English	RC-Acquis PAN-PC11 Wikipedia	+ Considers semantic information and keywords.	F1: 76.01 %
(Gharavi et al., 2020)	Use text embedding vectors to compare similarities between documents	, PersianPlagDet PAN-PC2014	+ Consider the structure and concept simultaneously using the word embedding technique.	Plagdet:89.67%
(Lazemi & Ebrahimpour-Komleh, 2020)	Discovery of plagiarism in Persian with structural analysis and semantic analysis	PAN2015 و AAIC2015	+ Consider structural and semantic similarity	F1: 84.14 %
(Ahuja et al., 2020)	Two semantic and structural matrices	PAN-PC2011	- Lack of production of candidate documents and time consuming and considering the structure of word substitution. + Consider the semantic similarity of words	Plagdet:87%
(Abdi et al., 2015)	Use order of words and WordNet	PAN-PC2011	Lack of production of candidate documents and time consuming and considering the structure of word substitution.	Plagdet: 78.9%
(Vaz, 2021)	Interlanguage plagiarism using text embedding network	PAN-PC-12	+ Provide an interlinguistic approach by considering the structure and similarity of words	PlagDet:72.51%
(Alvi et al., 2021)	Two important types of plagiarism include the synonymous substitution and different order of words by combining the Smith Waterman algorithm and embedding pre-trained words.	Plagiarized Short Answers	+ Consider ambiguous plagiarism	F1: 80.8 %

2) In the second step of text documents pre-processing stage, tokenization of documents' sentences is done. In this stage, the sentences are analyzed into their words.
3) In the following, the pre-processing of tokenized sentences is done. In this step, all punctuations are omitted from the sentences. These punctuations include email and website addresses, hashtag (#), emojis, @ sign, punctuations like?,!,.,;, etc. Then, the useless words are omitted from the sentences. These words include possessive pronouns, connective verbs, prepositions, etc. In normalization, the words are lemmatized. The pre-processing stages of a sentence from a suspicious document in Fig. 11 have been shown. In this figure, Image (1) shows the sentence before pre-processing stages. Image (2) shows the sentence after punctuation omission in which (:) was omitted and the number of tokens has reduced to 9. Image (3) shows the sentence status after omitting the useless words. In this figure, it, was, a and from have been omitted and the number of tokens has reduced to 5. Image (4) shows the word normalization process in which evoked changed into evoke and the initial letters of words have become small.

3.2. Filtering at document level

In two proposed models, two filtering stages have been done to limit search space and reaching the candidate sub-category of documents. In the first stage, filtering is done at document level to limit search space. In this step, after pre-processing of two pairs of suspicious and source documents, bag of words related to two documents is formed using Bag of Words technique. In Fig. 12, the characteristics of bag of words in a suspicious document have been shown. As observed in this suspicious document, it has 4495 single words which have been extracted from 2268 sentences.

After forming bag of words for two pairs of suspicious and source documents, filtering at document level is done using Eq. (3). In fact, in this step, if the number of common words in two bags of words from suspicious and source documents is more than one third of number of words in bag of suspicious document, both documents are analyzed for more examinations; otherwise, the suspicious document is not compared to all source documents in the database and the problem's search space is limited to some extent.

$$\begin{cases} intD = Bag_{susp} \cap Bag_{sorc} \geq \frac{1}{3}\left(Bag_{susp}\right) sentenceProcess \\ intD = Bag_{susp} \cap Bag_{sorc} < \frac{1}{3}\left(Bag_{susp}\right) otherDocProcess \end{cases} \quad (3)$$

3.3. Filtering at sentence level

If the two examined documents pass the document level filtering, in this step, filtering is examined at sentence level. In this step like the filtering at document level, the common points of the sentences are examined using Eq. (4). To be precise, in filtering at sentence level, if two sentences from suspicious and source documents have more than 5

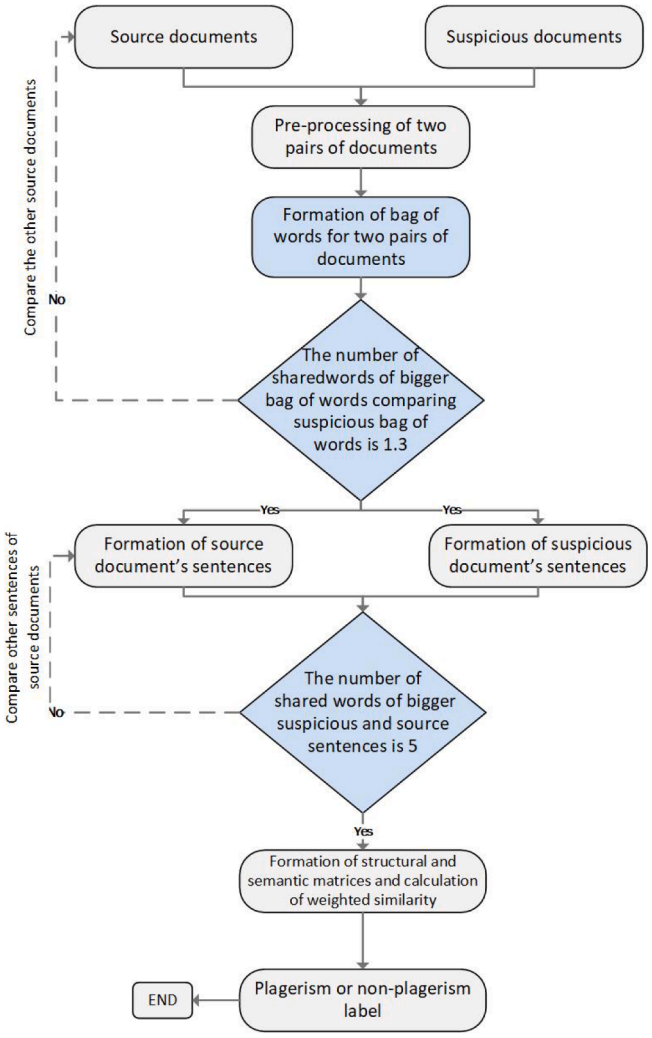

Fig. 9. The general diagram of two proposed models.

common words, they are considered for more examinations. Otherwise, the suspicious document sentence is compared to other sentences. This phase also causes all the sentences of suspicious document not to be compared to all the sentences of source documents and the problem's search space is limited to some extent and the model's time complexities reduce to some extent. In case two suspicious and source documents and their sentences pass both levels of filtering, in the next step, the formation process of semantic and structural process is done to examine the suspicious sentences precisely.

$$\begin{cases} ints = \boldsymbol{Sentence}_{susp} \cap \boldsymbol{Sentence}_{sorc} \geq 5 JointMatrixCreation \\ ints = \boldsymbol{Sentence}_{susp} \cap \boldsymbol{Sentence}_{sorc} < 5 otherSentenceProcess \end{cases} \quad (4)$$

3.4. Formation of semantic matrix

This part of two proposed models is different in formation of common dictionary and semantic matrix formation:

- **Formation of semantic matrix in the first model:**

1) In the first model, the unique word list is formed to form the semantic matrix of a pair of suspicious and source sentences. Then, for each word of source words, its first synonym word is extracted using WordNet ontology and in case the considered word does not exist in common word list, it will be inserted.
2) Regarding the formed common word list, to form semantic matrix, two kinds of distance, Wu_Palmer and Path_Similarity, are calculated using Eqs. (5) and (6) respectively in WordNet ontology between two groups of words related to source and suspicious sentences. Wu-Palmer equation calculates the similarity between two words based on their position comparing the position of the most common concept among them (Slimani, 2013).

$$Sim_{Wu\&Palmer}(C_1, C_2) = \frac{2 \times N}{N_1 + N_2 + 2 \times N} \quad (5)$$

In the equation above, N_1 and N_2 show the number of Is-a relation of two words of C_1 and C_2 comparing their common concept. N shows also the distance of common concept from the root. If two words of fever and diarrhea in Fig. 13 are considered, the hierarchical structure of WordNet ontology drawn in the figure shows that the common concept among them is signs_and_symptoms; the distance of both words comparing their common concept equals 2, the depth of common concept comparing hierarchical tree structure root also equals 3. Therefore, the Wu_Palmer distance is calculated as follows for them (Slimani, 2013):

$$Sim_{Wp}(fever, diarrhea) = \frac{2 \times 3}{2 + 2 + 2 \times 3} = 0.6$$

The criterion of Path_Similarity is also calculated based on the shortest distance between two concepts or words in Is-a classification. The output of this criterion is values between 0 and 1. If a word is compared to itself, the output will be 1. This criterion is calculated from Eq. (6) (Ahuja et al., 2020). After calculation of two groups of similarity and the distance above for common dictionary words, the final amounts of words of common matrix are calculated from Eq. (7). In this part, the amount of α was considered as 0.5.

$$Sim_{Path}(C_1, C_2) = \frac{1}{Distance + 1}, whereDistance(C_1, C_2) = ShortestPath \quad (6)$$

$$Semantic = \alpha^*(sim_{wu}) + (1 - \alpha)^*(Sim_{Path}) \quad (7)$$

- **Formation of semantic matrix in the second model:**

1) In the second model, words of common matrix are produced only using unique words in two suspicious and source sentences because word embedding network of FastText produces similar vectors for words with common meanings. Word embedding network of FastText produces a vector with dimensions of 300 for each word of common list. This network has 16 billion tokens. To reach the final amount for each word, the sum of amounts of vector of each word is simply inserted for that word. In Fig. 14, the output vector of FastText network has been shown for 15 words in the common matrix of second model. This matrix has 300 columns and n rows the number of which is different for each pair of suspicious and source sentences and depends on the number of single and common words in them. In Fig. 14, only 10 first columns of 300 columns of word embedding matrix have been shown because of space limitation. The final weight resulted for each word in common dictionary is inserted in the related element of the same word in semantic matrix using sum of amounts of its vector.

3.5. Formation of structural matrix

Formation of structural matrix is similar for both models. To form the structural matrix for suspicious and source sentences, TF-IDF weighting technique is used. TF-IDF weight of each of current word in suspicious and source sentences is calculated using bag of words of suspicious and source documents. To calculated TF-IDF weight, Eq. (1) is used. In Fig. 15, the amounts of TF-IDF weight of words from a suspicious and source sentence have been shown. The output matrix of TF-IDF is a sparse matrix of which just the numerical amounts related to considered words are extracted. After calculation of weight of each word, to form the structural matrix related to two suspicious and source sentences, the amounts of the given word's weight is inserted for each element. In this

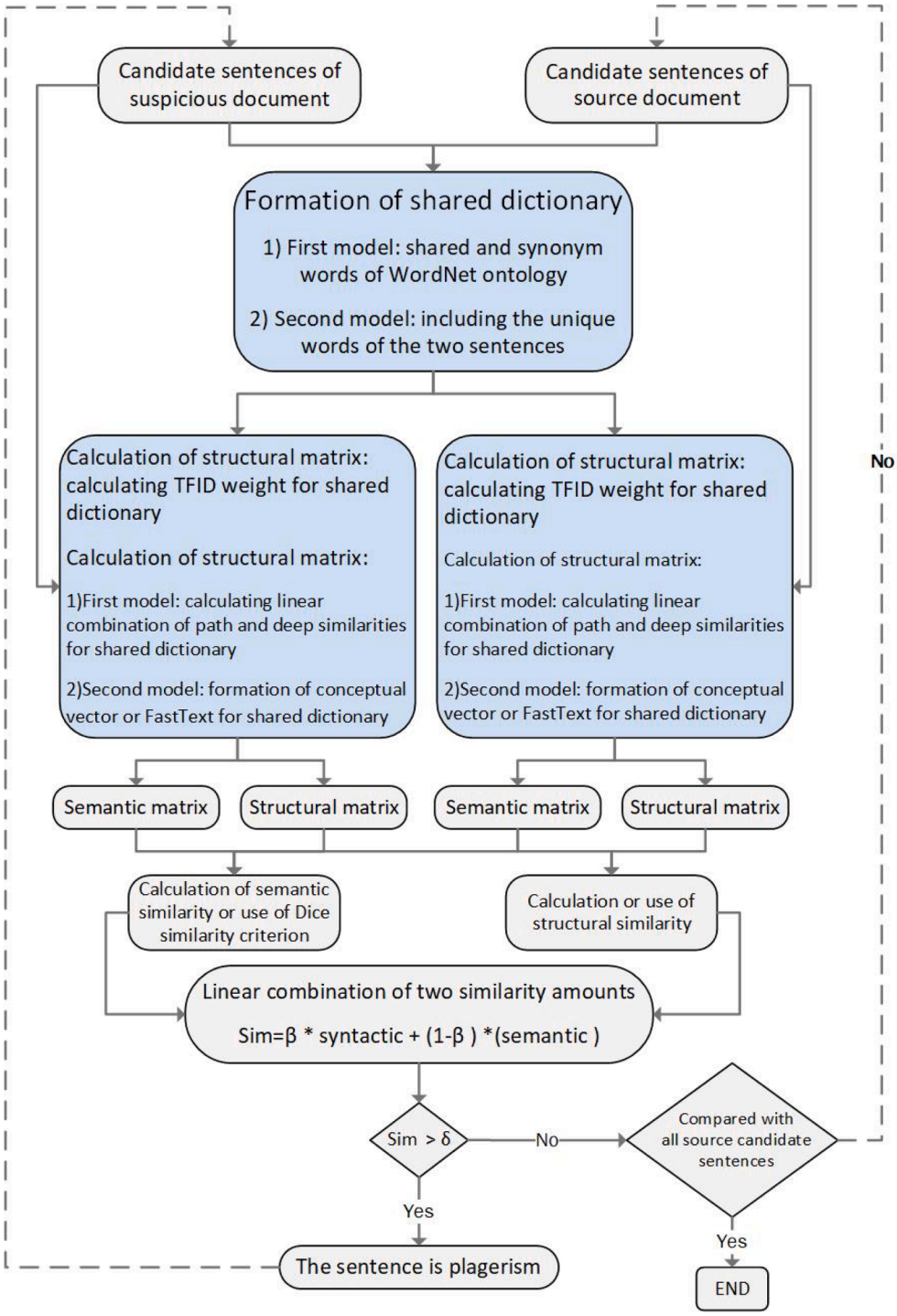

Fig. 10. The detailed diagram of two proposed models.

part, for the common words of two sentences, the resulting TF-IDF weights in both documents are added and inserted as the final weight for each common word. It should be mentioned that combination of TF-IDF weight of words is not the same as word order in common dictionary of suspicious and source sentences and to calculate the final weight amount of common words between suspicious and source sentences, their index has been used.

After calculation of semantic and structural matrix in the last step of the model, the final similarity of the sentences must be calculated.

3.6. Calculation of similarity between the sentences

After calculating two semantic and structural matrices for two pairs of suspicious and source sentences using two proposed models, semantic and structural similarity are calculated from Eqs. (8) and (9) respectively (Ahuja et al., 2020).

$$sim_{sem}(S_1, S_2) = \frac{2 \times \sum_{i=1}^{n} P_{1i}.P_{2i}}{\sum_{i=1}^{n} P_{1i}^2 + \sum_{i=1}^{n} P_{2i}^2} \tag{8}$$

$$sim_{syn}(S_1, S_2) = 1 - \frac{\|S_1 - S_2\|}{\|S_1 + S_2\|}, where$$

```
10 tokens:  Frontispiece : From it was evoked a monstrous shape
```
(1)

```
9 tokens:  Frontispiece From it was evoked a monstrous shape
```
(2)

```
5 tokens:  Frontispiece evoked monstrous shape
```
(3)

```
5 tokens:  frontispiece evoke monstrous shape
```
(4)

Fig. 11. Pre-processing stages of a sentence in a suspicious document.

```
bag_sus: 1x1 bagOfWords =

  bagOfWords with properties:

          Counts: [2268×4495 double]
      Vocabulary: [1×4495 string]
        NumWords: 4495
    NumDocuments: 2268
```

Fig. 12. Characteristics of the formed bag of words for a suspicious document.

$$\|S_1 - S_2\| = \sqrt[2]{(S_{11} - S_{21})^2 + \cdots + (S_{1n} - S_{2n})^2}$$

$$\|S_1 + S_2\| = \sqrt[2]{(S_{11} + S_{21})^2 + \cdots + (S_{1n} + S_{2n})^2} \quad (9)$$

In the equations above, S_1 and S_2 are the sentences of suspicious and source documents and P_{1i} and P_{2i} are semantic matrix elements of suspicious and source sentences. after calculation of two amounts of semantic and structural similarity for both sentences, the final amount of similarity of two sentences is calculated from Eq. (10):

$$sim_{sentences} = \beta^*(sim_{sem}) + (1-\beta)^*(sim_{syn}) \quad (10)$$

If the resulted amount from the equation above is bigger than defined threshold of δ, the given sentence in the suspicious document is labeled as plagiarism; otherwise, the next sentences are examined. In Fig. 16, the amount of semantic and structural similarity has been shown for two sentences of two models.

The amount of minimum threshold in the proposed model has been considered as 0.65. Regarding that the two resulted amounts for similarity between two sentences is smaller than minimum threshold, the sentence will not be plagiarism; therefore, the next sentences of the suspicious document are examined. The process above is iterated for other suspicious documents.

4. Experiments and results

In plagiarism detection methods, criteria of precision, recall, accuracy, f average and total rank are used to assess the models. These criteria are different according to consideration of plagiarism detection method in text alignment or retrieval groups. The criteria are calculated from confusion matrix similar to Fig. 17 and from Eqs. (11) to (14). In confusion matrix, True Positives (TP) variable shows the documents for which the plagiarism has been predicted and plagiarism exists in them. True Negatives (TN) variable shows the document for which non-plagiarism label has been predicted by the model and plagiarism does not exist in them. False Positives (FP) shows the documents for which the plagiarism has been predicted and plagiarism does not exist in them and False Negatives (FN) shows the document for which non-plagiarism label has been predicted, but plagiarism exists in them.

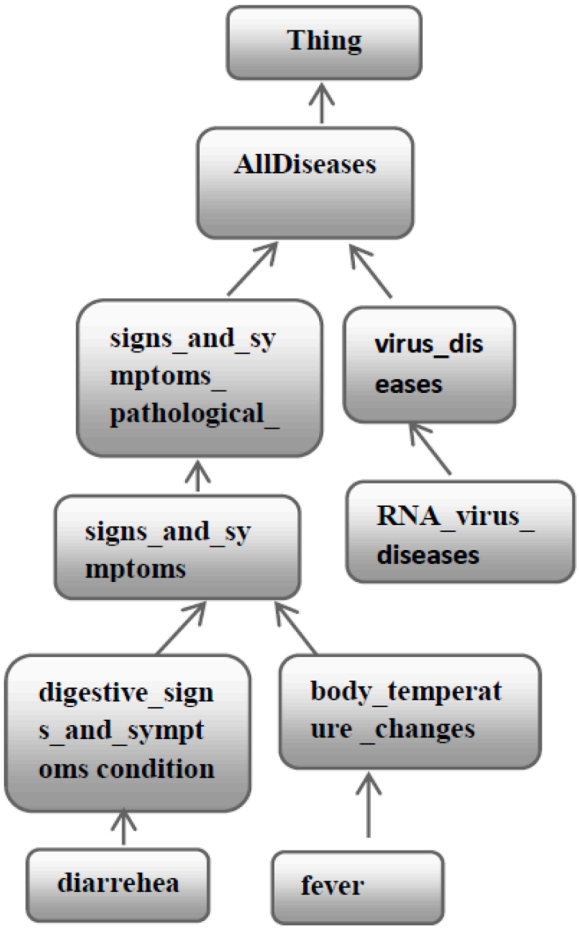

Fig. 13. A sample of tree structure in WordNet (Slimani, 2013).

$$Presicion = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$F_{measure} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (13)$$

$$Plagdet(S,R) = \frac{F_1}{\log_2(1 + gran(S,R))} \quad (14)$$

1	0.0573	0.0199	-0.0889	0.1225	0.0650	0.0679	-0.0235	-0.0029	0.0301	0.0392
2	-0.0019	0.0302	-0.0063	0.0662	0.0179	-0.0362	-0.0382	-0.0638	0.0354	-0.0175
3	-0.0657	0.0460	0.0251	-0.0332	0.0141	-0.2210	-0.0121	-0.0384	0.0709	-0.0542
4	-0.1106	-0.0377	0.1494	0.2807	0.0728	-0.0624	0.0867	0.0050	-0.0180	0.1056
5	0.0183	-0.1116	-0.2975	0.0242	-0.0045	-0.0678	0.1322	0.2278	-0.0656	-0.1138
6	0.0082	0.0446	0.1612	-0.0317	0.0049	-0.0621	-0.1180	-0.0024	0.1487	0.0818
7	-0.1476	-0.0070	-2.0000e-04	0.0270	0.0427	-0.0093	-0.0564	-0.1138	-0.1072	-0.2663
8	0.1506	-0.0998	0.0943	0.0520	-0.0106	-0.1593	0.0573	-0.1888	0.0922	-0.0845
9	0.1511	-0.0856	0.0087	0.0734	0.1880	0.0082	0.0504	-0.0357	-0.0820	0.1294
10	0.0183	-0.1829	0.0955	-0.0214	0.2068	0.1762	-0.0371	-0.0764	-0.1367	-0.0089
11	0.0083	-0.0348	0.0254	0.2156	0.0412	-0.0820	-0.1312	-0.2125	0.1735	-0.0571
12	-0.1775	-0.1043	-0.0019	0.0723	-0.0055	-0.0485	0.1293	0.1250	-0.2192	0.1778
13	-0.0350	0.0850	0.0474	-0.0285	-0.0350	-0.0649	-0.0635	-0.0193	0.0439	0.0828
14	0.0550	-0.0301	-0.1220	0.0445	-0.0056	-0.0503	0.0679	0.0728	-0.0012	0.1398
15	0.0916	-0.0278	0.0717	0.0192	-0.1441	0.0135	-0.0265	0.0325	-0.0644	0.0517
16	-0.1121	0.0279	-0.0237	0.0990	-0.0240	-0.1493	0.0326	-0.1620	0.1356	-0.0079
17	-0.0031	0.0908	0.0326	0.0724	0.0875	-0.0472	-5.0000e-04	-0.2428	-0.0428	0.0918
18	-0.0934	-0.0046	0.0345	-0.0632	-0.0948	-0.0144	-0.0656	0.2334	0.1027	0.1499

Fig. 14. The output matrix of FastText network for words in common dictionary of second model.

```
M: 1x1584 double =

   (1,146)      1.9957
   (1,216)      2.6198
   (1,479)      4.6347
   (1,480)      5.3279
   (1,481)      4.2293
   (1,482)      4.6347
   (1,483)      3.9416
   (1,484)      5.3279
   (1,485)      5.3279
   (1,486)      5.3279
```

```
M2: 1x2362 double =

   (1,52)       3.3322
   (1,70)       5.2781
   (1,125)      3.8918
   (1,234)      3.0809
   (1,407)      4.3618
   (1,618)      4.8726
   (1,619)      5.9713
   (1,620)      5.9713
   (1,621)      5.2781
   (1,622)      4.8726
   (1,623)      5.2781
   (1,624)      5.9713
   (1,625)      4.3618
   (1,626)      5.9713
   (1,627)      5.9713
   (1,628)      5.9713
   (1,629)      5.9713
```

Fig. 15. The TF-IDF weight amount for words of suspicious and source sentences.

The total rank criterion to compare the models results from the combination of accuracy, recall and granularity. Granularity here shows the number of time that a sentence was given plagiarism label (Ahuja et al., 2020).

In this article, PAN-PC11 received from the https://pan.webis.de/data.html address has been used for assessment. This database includes documents in which plagiarism has been placed automatically or manually. This set includes a number of suspicious and source text files and another text file that includes pairs of suspicious and source files in which plagiarism has occurred. Moreover, this set includes a file in XML format which determines the precise location of plagiarism. Source documents in this database have been extracted from Guttenberg project's books including 22,000 English books, 520 German books and 210 Spanish books. This database has separate documents to detect extrinsic and intrinsic plagiarism. In intrinsic part, it has 4753 documents and in extrinsic part, it has 11,093 suspicious and 11,093 source documents and together, it has 22,180 documents. In Table 2, the characteristics of this database have been shown briefly.

To implement and assess the proposed method, Matlab 2021 was used. All the tests were done in a 16G RAM system with 7 core processor and 3.4 GHZ and 64 bit Win 10 operating system.

4.1. Examination of number of suspicious documents

In this experiment, 2 to 200 documents as suspicious ones have been considered and the results of precision in both models for each group of documents have been reported. The results of this experiment have been shown in Fig. 18.

As observed in Fig. 18, in minimum number of suspicious documents i.e., 2, both models have reached 100% precision. Furthermore, the models could also show maximum precision in 10 suspicious documents. However, increasing the number of query documents, the model's precision has reduced. The results of this experiment show that in 200 suspicious documents, the models could achieve more than 90% precision that shows the proposed approaches can show acceptable performance in plagiarism detection.

Increasing the number of documents, the model based on WordNet ontology has reduced more than the model based on Fast Text words embedding. In fact, it is expected that by increasing the number of documents, this model has lower performance in comparison with FastText pre-trained word embedding network. The results of this experiment show that using word embedding can act better than ontologies in plagiarism to some extent. In this experiment, the precision of word embedding-based method in the minimum state i.e., in 200 query documents has been 1.3% better than ontology-based method in 50 query documents; this difference in precision has reached 2%. Generally, the results above show that using word embedding network in pre-trained state can yield promising results.

4.2. Examination of variable of $_\beta$ in linear combination of similarities

In the two proposed models, two kinds of similarities have been calculated and then, using the weighted variable of β in linear form, the amounts of two similarity kinds have been combined. This amount has a significant effect in the model results. Therefore, its amount was considered between 0 and 1 so that the models' results can be examined only one time in structural similarity state, one time in semantic one and one time in combined state. The precision results of the resulting two

```
sim_sem: 1x1 double =
    0.2919

sim_syn: 1x1 double =
    0.5185

sim_sentences: 1x1 double =
    0.4052

sim_sem: 1x1 double =
    0.5340

sim_syn: 1x1 double =
    0.6286

sim_sentences: 1x1 double =
    0.5530
```

Fig. 16. The amount of semantic and structural similarity for two sentences.

$$(11)\ Presicion = \frac{TP}{TP+FP}$$

$$(12)\ Recall = \frac{TP}{TP+FN}$$

$$(13)\ F_{measure} = \frac{2\times Precision\times Recall}{Precision+Recall}$$

$$(14)\ Plagdet(S,R) = \frac{F_1}{\log_2(1+gran(S,R))}$$

	Actual: plagiarized	Actual: non-plagiarized
Predicted: plagiarized	TP	FP
Predicted: non-plagiarized	FN	TN

Fig. 17. Confusion matrix.

Table 2
Database characteristics.

Plagiarism Type	Type Doc.	Num. Doc.	Total Doc.
Extrinsic -detection-corpus	Source-document	11,093	22,180
	Suspicious-document	11,093	
Intrinsic-detection-corpus	Suspicious-document	4753	4753

models have been reported in Fig. 19. This experiment can show that using which similarity can act more successfully in plagiarism detection and whether combination of two similarities can cause model success or not.

In Fig. 19, in case $\beta = 0$, only structural similarity has been used for plagiarism detection and when $\beta = 1$, only semantic similarity has been used.

In the figure above, it is observed clearly that in case only the structural similarity has been used for plagiarism detection, the results have reduced significantly comparing using only semantic similarity. In fact, these results show that using weighting technique of TF-IDF alone cannot have successful results for plagiarism detection. In this state, both models have achieved the approximate precision of 74%. In contrast, the results above show that when only semantic similarity was used for plagiarism detection, the results of both models are acceptable. In fact, the results show that using similarity based on FastText can

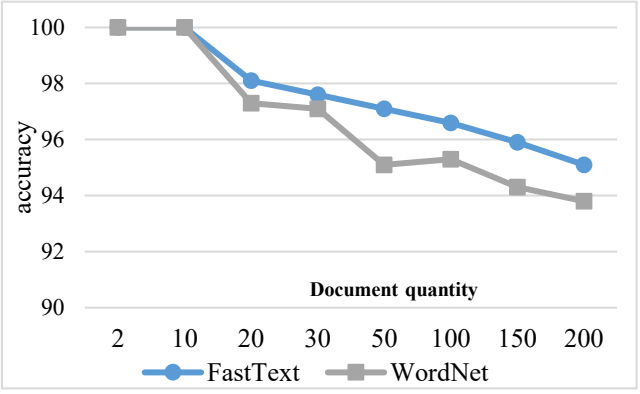

Fig. 18. Examination of the effect of number of suspicious documents on models' precision.

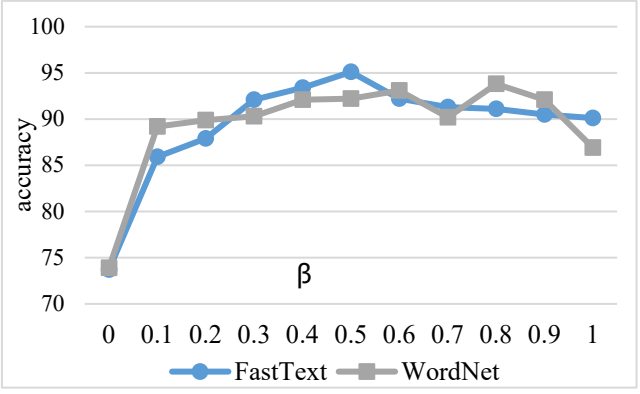

Fig. 19. Examination of the effect of similarity weight on models' precision.

provide a precision of 90%. In this experiment, the performance of the method based on FastText has been more successful than the method based on WordNet. When $\beta = 1$, the method based on ontology has achieved 86.9% that was 3.2% weaker comparing FastText. Finally, the results above show that combining two kinds of similarities with appropriate weight has a better performance comparing one kind of similarity. The results above show that combining structural and semantic similarities in the model based on FastText when $\beta = 0.5$ and in the method based on WordNet ontology when $\beta = 0.8$, can yield precision of 95.1% and 93.8% respectively.

4.3. Examination of minimum threshold in plagiarism detection

If the minimum threshold is considered as a small similarity, it causes the models to consider all sentences as plagiarism and if the minimum threshold is considered as a big similarity, it causes the models to consider all sentences as non-plagiarism. Therefore, to find an appropriate threshold minimum amount can have a high effect in the model's success. As a result, in this experiment, the minimum threshold was considered between 0.4 and 0.7. The precision results of this experiment have been shown in Fig. 20. As expected, reduction of threshold minimum to 0.4 has caused the models' results to reduce.

In this state, the precision of the model based on FastText word embedding network and the model based on WordNet ontology have been 65.9% and 62.2% respectively which are considered as very low amounts in plagiarism detection and the reason was that sentences with not a lot of similarities have also mistakenly received plagiarism label and therefore, the FP amount has increased significantly and the models' precision has decreased. In contrast, when the minimum threshold has been considered as 0.7, the models' results have again a decreasing trend. In this state, the precision of the model based on FastText word embedding network and the model based on WordNet ontology have been 93.4% and 92.1% respectively. These results show that a number of sentences which must receive plagiarism label has received non-plagiarism label and the models' precision has decreased. The results of the graph above show that both models have achieved their best results when minimum threshold equals 0.65. In this state, the precision of the model based on FastText word embedding network and the model based on WordNet ontology have been 95.1% and 93.8% respectively.

4.4. Comparison of results

After examination of two proposed models in different experiments, in the last experiment, a comparison was conducted between two proposed models and the methods of Ahuja et al. in 2020 (Ahuja et al., 2020) and Abdi et al. in 2015 (Abdi et al., 2015). In this experiment, the variables of precision, recall, harmonic mean f and Plagdet were compared with the previous models and the results have been reported in Fig. 21. In this test, the comparison of methods has been done in 200 suspicious documents. The minimum similarity threshold in all methods was considered as 0.65. The amount of variable of linear combination of β was also 0.5 in FastText model and 0.8 in WordNet model. The amount of α that is for linear combination of two criteria of deep similarity and the model based on WordNet was considered as 0.5.

The results of this experiment show clearly that the plagiarism model based on FastText and combining it with TF-IDF weighting technique has had a better performance comparing two previous models based on WordNet and the proposed model based on WordNet. One of the main

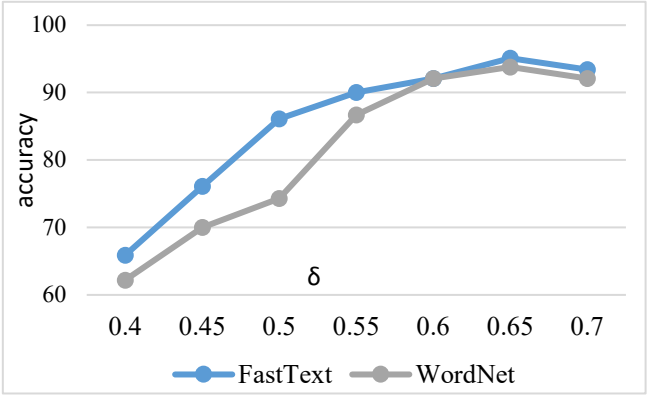

Fig. 20. The effect of results of minimum similarity threshold on models' precision.

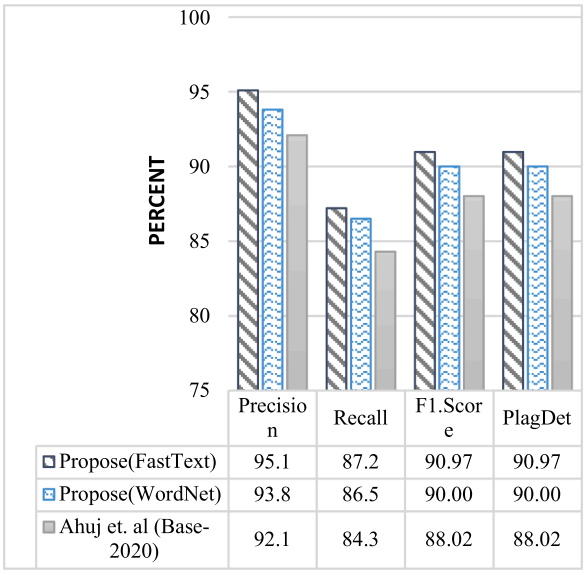

Fig. 21. Comparison of methods' results.

factors of having these results is using pre-trained word embedding. FastText network has 16 billion tokens and using it in plagiarism had successful performance. On the other hand, using words' weight instead of their placement index in the sentences caused the proposed models in this article to show better results in plagiarism. In the proposed model based on WordNet ontology, using structural matrix based on TF-IDF and also, insertion of synonyms of document words caused this method to make little changes comparing Ahuja's method which has achieved more desirable results. Generally, the results of the graph above show that the two proposed models in this article had better performance comparing Ahuja's method in extrinsic monolingual plagiarism detection.

The method based on FastText pre-trained word embedding network had a better performance comparing Ahuja's method in criteria of precision, recall, F1-score and plagdet by 3%, 2.9%, 2.95% and 2.95% respectively. The method based on FastText pre-trained word embedding network had a better performance comparing Abdi's method in criteria of precision, recall, F1-score and plagdet by 4.9%, 17%, 11.87% and 12.07% respectively.

The method based on WordNet ontology had a better performance comparing Ahuja's method in criteria of precision, recall, F1-score and plagdet by 1.7%, 2.2%, 1.98% and 1.98% respectively. The method based on WordNet ontology had a better performance comparing Abdi's method in criteria of precision, recall, F1-score and plagdet by 3.6%, 16.3%, 10.90% and 11.10% respectively. Considering TF-IDF weight of common words of suspicious and source documents to reach more similar vectors, using pre-trained word embedding network to reach a more detailed vector of each word and considering synonym words and insertion of some of them in the common matrix of suspicious and source sentences are some main factors of success of two proposed models comparing the previous method.

5. Conclusion and further works

In this article, two extrinsic monolingual plagiarism detection methods were introduced. These two methods have the capability of plagiarism detection in English. In these two methods, after preparation and pre-processing of text, the suspicious and source documents' bags of words were formed and in case the number of their common words is more than the minimum threshold, both documents will be sent to the next stages for more detailed processing and structural and semantic matrices of sentences using TF-IDF weighting technique and two

techniques of FastText pre-trained word embedding network and WordNet ontology were formed. Then, two similarity criteria of final similarity rate were determined using weighted combination. In case the similarity of sentences of suspicious and source documents was more than the minimum threshold, the given suspicious sentence would receive plagiarism label. The results of assessment of these models in PAN-PC11 showed that the model can have 100% result in low number of suspicious documents between 2 and 10; however, increasing the number of suspicious documents to 200 documents had also acceptable performance in plagiarism detection and can reach a precision of 95%. The results of examinations showed that combining word embedding network and TF-IDF weighting technique can result in plagiarism detection improvement. Moreover, combining TF-IDF weighting with WordNet ontology also caused the model to reach better precision comparing the previous method. In line with this research, the proposed model can be applied in Persian language and the rate of its success can be examined in another language. In this case, simple word embedding network can be used because FastText network cannot be used in Persian. Furthermore, document level filtering can be done using distance criteria such as cosine one and the problem search space will be more limited. Following this study, it can be sufficed to use only word embedding model to detect plagiarism at sentence level and the method can be examined by combining two models of pre-trained word embedding and the simple one.

CRediT authorship contribution statement

Hamed Arabi: Methodology, Software, Conceptualization, Writing – original draft. **Mehdi Akbari:** Validation, Writing – review & editing, Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdi, A., Idris, N., Alguliyev, R. M., & Aliguliyev, R. M. (2015). PDLK: Plagiarism detection using linguistic knowledge. *Expert Systems with Applications, 42*(22), 8936–8946.

Agarwal, B. (2019). Cross-lingual plagiarism detection techniques for English-Hindi language pairs. *Journal of Discrete Mathematical Sciences and Cryptography, 22*(4), 679–686.

Ahnaf, A., Saha, S., & Hossain, N. (2020). Closed Domain Bangla Extrinsic Monolingual Plagiarism Detection and Corpus Creation Approach. *Paper presented at the 2020 IEEE Region 10 Symposium (TENSYMP)*.

Ahuja, L., Gupta, V., & Kumar, R. (2020). A new hybrid technique for detection of plagiarism from text documents. *Arabian Journal for Science Engineering*, 1–14.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management, 39*(1), 45–65.

Al-Suhaiqi, M., Hazaa, M. A., & Albared, M. (2018). Arabic English cross-lingual plagiarism detection based on keyphrases extraction, monolingual and machine learning approach. *Asian Journal of Research in Computer Science*, 1–12.

Altheneyan, A. S., & Menai, M. E. B. (2020). Automatic plagiarism detection in obfuscated text. *Pattern Analysis Applications*, 1–24.

Alvi, F. (2020). *Monolingual plagiarism detection and paraphrase type identification.* University of Sheffield.

Alvi, F., Stevenson, M., & Clough, P. (2021). Paraphrase type identification for plagiarism detection using contexts and word embeddings. *International Journal of Educational Technology in Higher Education, 18*(1), 1–25.

Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(2), 133–149.

Aravind, K., Shanmukh, B., Charan, G. S., & Nellikoppad, C. S. (2020). A survey of cross-lingual plagiarism detection using natural language processing. *Journal of King Saud University - Computer and Information Sciences*, 1–14.

Barrón-Cedeno, A., Rosso, P., Pinto, D., & Juan, A. (2008). On Cross-lingual Plagiarism Analysis using a Statistical Model. *Paper presented at the PAN*.

Belguith, L. H. (2021). *Extrinsic Plagiarism Detection for French Language with Word Embeddings.* Paper presented at the Intelligent Systems Design and Applications: 19th International Conference on Intelligent Systems Design and Applications (ISDA 2019) Held December 3-5, 2019.

Boukhalfa, I., Mostefai, S., & Chekkai, N. (2018). A study of graph based stemmer in Arabic extrinsic plagiarism detection. *Paper presented at the Proceedings of the 2nd Mediterranean Conference on Pattern Recognition and Artificial Intelligence*.

Brin, S., Davis, J., & Garcia-Molina, H. (1995). Copy detection mechanisms for digital documents. *Paper presented at the ACM SIGMOD Record*.

Ceska, Z., Toman, M., & Jezek, K. (2008). Multilingual plagiarism detection. *Artificial intelligence: Methodology, systems, and applications*, 83-92.

Chen, K., Zhang, Z., Long, J., & Zhang, H. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Systems with Applications, 66*, 245–260.

De Prisco, R., Esposito, A., Lettieri, N., Malandrino, D., Pirozzi, D., Zaccagnino, G., & Zaccagnino, R. (2017). Music plagiarism at a glance: Metrics of similarity and visualizations. *Paper presented at the 2017 21st International Conference Information Visualisation (IV)*.

De Prisco, R., Malandrino, D., Zaccagnino, G., & Zaccagnino, R. (2017). Fuzzy vectorial-based similarity detection of music plagiarism. *Paper presented at the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*.

Dogan, T., & Uysal, A. K. (2019). Improved inverse gravity moment term weighting for text classification. *Expert Systems with Applications, 130*, 45–59.

Elamine, M., Bougares, F., Mechti, S., & Belguith, L. H. (2019). *Extrinsic plagiarism detection for French language with word embeddings.* Paper presented at the International Conference on Intelligent Systems Design and Applications.

Elhadi, M., & Al-Tobi, A. (2009). Duplicate detection in documents and webpages using improved longest common subsequence and documents syntactical structures. *Paper presented at the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*.

Ezzikouri, H., Madani, Y., Erritali, M., & Oukessou, M. (2019). A new approach for calculating semantic similarity between words using WordNet and set theory. *Procedia Computer Science, 151*, 1261–1265.

Ezzikouri, H., Oukessou, M., Erritali, M., & Madani, Y. (2019). Fuzzy cross language plagiarism detection approach based on semantic similarity and Hadoop MapReduce. In *Recent Advances in Intuitionistic Fuzzy Logic Systems* (pp. 181–190). Springer.

Fellbaum, C. (2010). WordNet. In *Theory and applications of ontology: computer applications* (pp. 231–243). Springer.

Fensel, D. (2001). Ontologies. In *Ontologies* (pp. 11–18). Springer.

Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys (CSUR), 52*(6), 1–42.

Gharavi, E., Veisi, H., & Rosso, P. (2020). Scalable and language-independent embedding-based approach for plagiarism detection considering obfuscation type: No training phase. *Neural Computing and Applications, 32*(14), 10593–10607. https://doi.org/10.1007/s00521-019-04594-y

Gillam, L., & Vartapetiance, A. (2016). From English to Persian: Conversion of Text Alignment for Plagiarism Detection. *PAN@ FIRE2016 Shared Task on Persian Plagiarism Detection and Text Alignment Corpus Construction. Notebook Papers of FIRE 2016*.

Giller, G. L. (2012). The statistical properties of random bitstreams and the sampling distribution of cosine similarity.

Gruner, S., & Naven, S. (2005). Tool support for plagiarism detection in text documents. *Paper presented at the Proceedings of the 2005 ACM symposium on Applied computing*.

Gunawan, S. P., Krisnawati, L. D., & Chrismanto, A. R. (2020). Analysis of stylometric features and segmentation strategies in intrinsic plagiarism detection system. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 4*(5), 988–997.

Huang, A. (2008). *Similarity measures for text document clustering.* Paper presented at the Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand.

Joseph, A., & Haroon, R. P. (2016). A survey on plagiarism detection in documents. *Imperial Journal of Interdisciplinary Research, 3*(1).

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., & Rudzicz, F. (2019). A survey of word embeddings for clinical text. *Journal of Biomedical Informatics, 100*, Article 100057.

Kim, S., Park, H., & Lee, J. (2020). Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications, 152*, Article 113401.

Lazemi, S., & Ebrahimpour-Komleh, H. (2020, 29-30 Oct. 2020). *ParsiPayesh: Persian Plagiarism Detection based on Semantic and Structural Analysis.* Paper presented at the 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE).

Ling, W., Dyer, C., Black, A. W., & Trancoso, I. (2015). Two/too simple adaptations of word2vec for syntax problems. *Paper presented at the Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*.

Lovepreet, V. G., & Kumar, R. (2019). *Survey on Plagiarism Detection Systems and Their Comparison.* Paper presented at the Computational Intelligence in Data Mining: Proceedings of the International Conference on ICCIDM 2018.

Lukashenko, R., Graudina, V., & Grundspenkis, J. (2007). Computer-based plagiarism detection methods and tools: An overview. *Paper presented at the Proceedings of the 2007 international conference on Computer systems and technologies*.

Mahdavi, P., Siadati, Z., & Yaghmaee, F. (2014). Automatic external Persian plagiarism detection using vector space model. *Paper presented at the Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on*.

Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology, 6*(1), 1–12.

Meuschke, N., Stange, V., Schubotz, M., & Gipp, B. (2018). HyPlag: A hybrid approach to academic plagiarism detection. *Paper presented at the The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

Meuschke, N., Stange, V., Schubotz, M., Kramer, M., & Gipp, B. (2019). Improving academic plagiarism detection for STEM documents by analyzing mathematical content and citations. *Paper presented at the 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Muhammad, S. (2020). *Mono-and cross-lingual paraphrased text reuse and extrinsic plagiarism detection*. Lancaster University.

Muneer, I., & Nawab, R. M. A. (2021). Cross-lingual text reuse detection using translation plus monolingual analysis for english-urdu language pair. *Transactions on Asian and Low-Resource Language Information Processing, 21*(2), 1–18.

Pinto, D., Civera, J., Barrón-Cedeno, A., Juan, A., & Rosso, P. (2009). A statistical approach to crosslingual natural language tasks. *Journal of Algorithms, 64*(1), 51–60.

Polydouri, A., Vathi, E., Siolas, G., & Stafylopatis, A. (2020). An efficient classification approach in imbalanced datasets for intrinsic plagiarism detection. *Evolving systems, 11*(3), 503–515.

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation, 45*(1), 45–62.

Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., . . . Stein, B. (2012). *Overview of the 4th International Competition on Plagiarism Detection.* Paper presented at the CLEF (Online Working Notes/Labs/Workshop).

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An evaluation framework for plagiarism detection. *Paper presented at the Proceedings of the 23rd international conference on computational linguistics: Posters*.

Raghavan, V. V., & Wong, S. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for information Science, 37*(5), 279.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. *Paper presented at the Proceedings of the first instructional conference on machine learning*.

Roostaee, M., Fakhrahmad, S. M., & Sadreddini, M. H. (2020). Cross-language text alignment: A proposed two-level matching scheme for plagiarism detection. *Expert Systems with Applications, 160*, Article 113718.

Roostaee, M., Sadreddini, M. H., & Fakhrahmad, S. M. (2020). An effective approach to candidate retrieval for cross-language plagiarism detection: A fusion of conceptual and keyword-based schemes. *Information Processing & Management, 57*(2), Article 102150. https://doi.org/10.1016/j.ipm.2019.102150

Saini, A., Sri, M. R., & Thakur, M. (2021). Intrinsic Plagiarism Detection System Using Stylometric Features and DBSCAN. *Paper presented at the 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Shivakumar, N., & Garcia-Molina, H. (1995). SCAM: A copy detection mechanism for digital documents.

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas, 18*(3), 491–504.

Slimani, T. (2013). Description and evaluation of semantic similarity measures approaches. *arXiv preprint arXiv:1310.8059*.

Staab, S., & Studer, R. (2010). *Handbook on ontologies*. Springer Science & Business Media.

Stein, B., Lipka, N., & Prettenhofer, P. (2011). Intrinsic plagiarism analysis. *Language Resources and Evaluation, 45*(1), 63–82.

Stein, B., & zu Eissen, S. M., & Potthast, M.. (2007). Strategies for retrieving plagiarized documents. *Paper presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141–188.

Vaz, D. d. A. (2021). Cross language plagiarism detection with contextualized word embeddings. 1-58.

Wakil, K., Ghafoor, M., Tariq, S., & Ahmad, M. (2017). Plagiarism detection system for the Kurdish language. *International Journal of Information Technology and Computer Science, 12*, 64–71. https://doi.org/10.5815/ijitcs.2017.12.08

Wang, S., Qi, H., Kong, L., & Nu, C. (2013). Combination of VSM and Jaccard coefficient for external plagiarism detection. *Paper presented at the Machine Learning and Cybernetics (ICMLC), 2013 International Conference on*.

Wong, S. M., & Raghavan, V. V. (1984). Vector space model of information retrieval: A reevaluation. *Paper presented at the Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval*.

Yilmaz, S., & Toklu, S. (2020). A deep learning analysis on question classification task using Word2vec representations. *Neural Computing and Applications, 32*(7), 2909–2928.

Zu Eissen, S. M., & Stein, B. (2006). Intrinsic Plagiarism Detection. *Paper presented at the ECIR*.

Zubarev, D., & Sochenkov, I. (2019). Cross-language text alignment for plagiarism detection based on contextual and context-free models. *Paper presented at the Proc. of the Annual International Conference "Dialogue*.