

## Detecting translation borrowings in huge text collections using various methods

Adel Al-Janabi<sup>1</sup>, Ehsan Ali Al-Zubaidi<sup>2</sup>, Baqer M. Merzah<sup>3</sup>

<sup>1</sup>University of Kufa, Najaf, Iraq

<sup>2</sup>Department of Environmental Planning, Faculty of Physical Planning, University of Kufa, Najaf, Iraq

<sup>3</sup>Department of Computer Science, Faculty of Education, University of Kufa, Najaf, Iraq

### Article Info

#### Article history:

Received Dec 19, 2022

Revised Jan 19, 2023

Accepted Jan 27, 2023

#### Keywords:

Deep learning

Distributional semantics

Machine translation

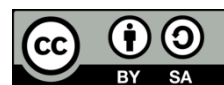
Natural language processing

Text borrowings detection

### ABSTRACT

The purpose of this work is to investigate the problem of detecting transportable borrowings and text reuse. The article proposes a monolingual solution to this problem: translating the suspicious material into language collections for additional monolingual analysis. One of the major requirements for the suggested technique is robustness against machine learning ambiguities. The next step in the document analysis is split into two parts. The authors begin by retrieving documents-candidates that are similarity to other types of text recurrence. The paper proposes retrieving texts utilizing word clusters formed using distributional semantic for robustness. In the second stage, the authors use deep learning neural networks to compare the suspected document to candidates utilizing phrase embedding. The experimentation is carried out for the language pair "English-Arabic" on both articles and synthetic data.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



### Corresponding Author:

Adel Al-Janabi

University of Kufa

Najaf, Iraq

Email: adelh.aljanabi@uokufa.edu.iq

## 1. INTRODUCTION

The problem of incorrect text borrowings is relevant for the field of education and scientific research [1]. According to the materials of the study by Al-Janabi *et al.* [2] conducted in 2019, more than 1,500 dissertations on historical sciences defended in Iraq after 2,000 contain significant borrowings from other dissertations. For the task of detecting borrowings within one language, industrial tools show high search completeness [1], whose work is based on the representation of documents in the form of a set of overlapping word-by-word n-grams (shingles) [3]. This approach allows you to effectively search for exact text borrowings but does not allow you to detect borrowings with a large proportion of paraphrased text or with inserts of text translated from another language. There are several approaches that describe the problem of finding translated borrowings for some pairs of languages [4], [5], for example, for the Spanish-English pair. This work is devoted to the detection of translated borrowings for Arabic-English pairs of languages. This pair is rarely found in the literature and is not related. The choice of a pair of languages Arabic-English is due to the predominance of English-language publications on the Internet and a better knowledge of this language compared to others. Similar to the works [6], [7], this article offers a description of the algorithm for the full cycle of borrowing search-first, the candidate documents are searched for in an external collection, then they are compared in detail with the document being checked. An algorithm is proposed based on monolingual analysis of documents, similar to the one carried out in [8], [9] the document being checked is translated into English using a machine translation system with the further comparison of text fragments inside the documents. Several works devoted

to the search for translated borrowings use additional resources, such as thesauri and ontologies. The authors suggest using knowledge bases to extract information about the proximity between texts [4], [5]. In work Kaliappan *et al.* [5], an algorithm based on a combination of neural networks and knowledge graphs is proposed. The main disadvantage of this approach is resource intensity: the use of multilingual ontologies and knowledge bases requires large computational capacities to build semantic graphs for each text fragment, as well as to compare the obtained semantic graphs. In this work, we propose a decomposition of the algorithm for detecting translated borrowings for searching through large text collections [10].

## 2. SEARCH FOR CANDIDATE DOCUMENTS

The most relevant candidate documents are found for the document being checked, for this purpose, a modification of the Shingle algorithm is used. The text is divided into fragments, each frame is mapped into a vector space. For each vector of the document being checked, the nearest vectors from the candidate documents are found, after which the pairs of these vectors are classified into similar and dissimilar pairs of text fragments. Since the proposed algorithm uses monolingual analysis of borrowings, the problem is close to the problem of detecting a paraphrased text. Several approaches [9], [11]–[14] to solving this problem use vector representations of phrases obtained using deep learning neural networks. The paper [13] proposes a neural bag of words (English Neural Bag-of-Words) and deep averaging networks (English deep averaging networks).

## 3. DOCUMENT COMPARISON

In this article, it is proposed to use the outputs of a neural network as vector representations of text fragments for a further approximate algorithm for finding the nearest neighbor [15]. The paper investigates the properties of the proposed method for detecting transferable borrowings. The analysis of deep learning models used at the stage of comparing documents, as well as a composite optimized function, is carried out. The quality of the proposed method is verified both on a synthetic sample and articles from journals. An error analysis is performed. The proposed method of finding borrowings is compared with the basic algorithm for finding borrowings based on the use of machine translation and the shingle algorithm. Problem statement Let the collections of documents in English be given  $D_e = \{d_e^j\}_{j=1}^N$  and Arabic  $D_a = \{d_a^i\}_{i=1}^N$ . Documents in Arabic and English are represented as a concatenation of text fragments:

$$d_e^j = [s_{e1}^j \cup \dots \cup s_{eh}^j]; d_a^i = [s_{a1}^i \cup \dots \cup s_{ah}^i] \text{ Let the sample be given:}$$

$$D = \{(d_e^l, d_a^l), AL^l\}_{l=1}^L$$

where each pair of documents  $(d_e^l, d_a^l)_{d_e^l \in D_e, d_a^l \in D_a}$

The list of fragment pairs is compared:

$$AL = \left[ (s_{e1}^l, s_{a1}^l), \dots, (s_{ek(l)}^l, s_{ak(l)}^l) \right]$$

for each pair  $(s_{ek}^l, s_{ak}^l)$ , it is known that the fragment  $s_{ak}^l$  is a translation of the fragment  $s_{ek}^l$ .

The model  $f$  is defined as the sequential execution of the filter and comparison functions, where filter:

$$(d_a^i, D_e)_{d_a^i \in D_a} \rightarrow D_e^{retrieved_i} \subset D_e$$

comparison:

$$(d_a^i, D_e^{retrieved_i})_{d_a^i \in D_a} \rightarrow AL^i$$

Here  $AL^i$  is a list of fragment pairs. The filter function is responsible for narrowing the number of documents in the collection compared with the document being checked and allows for furthermore detailed comparison of comparison using resource-intensive computational algorithms based on deep learning models. The quality of the model  $f$  is evaluated using the precision function and recall:

$$Precision = \frac{|(\cup_{l=1}^L AL^l) \cap (\cup_{i=1}^M AL^i)|}{|(\cup_{i=1}^M AL^i)|}, Recall = \frac{|(\cup_{l=1}^L AL^l) \cap (\cup_{i=1}^M AL^i)|}{|(\cup_{l=1}^L AL^l)|}$$

it is required to find the function  $f$  that maximizes  $F1$ , the harmonic mean of the precision and recall indicators:

$$\hat{f} = arg \max F1(f, D), F1 = \frac{2Precision \cdot Recall}{Precision + Recall} \text{ Where } f \text{ is a given family of models.}$$

#### 4. SEARCH FOR CANDIDATE DOCUMENTS

One of the algorithms for searching for candidate documents in the problems of detecting verbatim borrowings and searching for almost-duplicates of text is an algorithm based on the construction of an inverted index, in which each document in the collection is represented by a set of shingles [3], that is, a set of overlapping n-grams. The document being checked is also divided into shingles, after which the documents are searched for by the inverted index with the largest number of shingles. In this paper, we propose a generalization of the shingles algorithm, which allows us to improve the quality of the search for candidates in the case of detection of transferable borrowings. The filter function is proposed as follows:

$$(d_a^i, D_e) arg \max \sum_{d_e^j \in D_e} \sum_{h \in H(d_a^i)} I[h \in H(d_a^j)] / (|d_e^j \in D_e: h \in H(d_e^j)|^\alpha + const)$$

here  $H$  is a set of  $N$ -grams of the document, an ordered sequence of  $N$  cluster labels, where the procedure for forming clusters is described below;  $\alpha \in A$ ;  $K$  is an optimized hyperparameter. To reduce the impact of translation ambiguity on the search for candidate documents, it is proposed to replace words with the corresponding cluster labels:  $\{x_1, \dots, x_n\} \rightarrow \{class(x_1), \dots, class(x_n)\} = h$  where  $x_1, \dots, x_n$  words. Clusters are pre-selected from the text corpus and contain semantically similar words. To reduce the ambiguity of the translation, before splitting it into  $N$ -grams, it is proposed to remove stop words from the text and carry out lemmatization. To account for possible permutations of words that occur after the translation of the text, the words inside each of the  $n$ -grams are sorted in lexicographic order. In this paper, a vector representation model of words based on the distributive hypothesis is used to obtain clusters. Clustering is performed using the cosine distance function:

$$\cos(c_1, c_2) = \frac{\langle c_1, c_2 \rangle}{\|c_1\|_2 \|c_2\|_2} \tag{1}$$

where  $c_1$  and  $c_2$  are vectors from the same vector space. Below are examples of the resulting clusters:  
 i) [beers, beer, brewing, brew, brewery, brewed, pint, Guinness, stout, ipa, lager, ale, keg, pints]. and  
 ii) [excellent, brilliant, best, exceptional, super, outstanding, and amazing]. To compare the found candidate documents  $D_e^{retrieved_i}$  and the document being checked ( $d_a^i$ ), a vector representation model of the phrase is used-the texts are divided into fragments and the corresponding vectors are compared. Below are the details of the comparison algorithm, as well as an analysis of the proposed optimization problem.

##### 4.1. The model of the vector representation of the phrase

Let's take a closer look at the stage of building the mapping of a fragment into a vector. Let each word of the document in the language of the collection be associated with a vector  $v \in A^u$  of dimension  $u$ . For simplicity, we will assume that all fragments in the collection language have a bounded length  $n_{col}$ . Then the vectorization model of the fragment is the mapping  $h: W \in A^{u \times n_{col}} \rightarrow A^u$ , where  $W$  is the parameter space of the model. Objects from the set  $A^{u \times n_{col}}$  are the sequential concatenation of vectors of vector representations of words for sample fragments:

$$x \in [v_1, \dots, v_{n_{col}}]^T, x \in A^{u \times n_{col}}$$

To work with fragments less than  $n_{col}$  in length, we define some vector denoting an empty word. The model is optimized in the partially supervised learning mode. As an optimized function, a composite error function is used, which is the sum of the reconstruction error and the indentation error:

$$aE_{rec}(X_{rec}, w) + (1 - a)E_{me}(X_{me}, w) \rightarrow \min_{w \in W} \tag{2}$$

where  $E_{rec}$  is the reconstruction error;  $E_{me}$ -indentation error;  $X_{rec}$  and  $(X_{me}$  are training samples;  $w$  -model parameters;  $\alpha$  is a tunable hyperparameter. Let's consider in more detail each term of the error function. The

first term of the error function corresponds to the autoencoder model. Let a sample  $X_{rec} \subset A^{u \times n_{col}}$  be given. The  $h$  model acts as an encoding function for information about the  $X_{rec}$  sample. Let also be given an auxiliary decoding function  $g$  that restores the original vector representation of  $x$  from the outputs of the model.

$$h: r(x, w) = g(\cdot, w) \circ h(x, w) \approx x, x \in A^{u \times n_{col}}$$

The minimized reconstruction error looks like this:

$$E_{rec}(X_{rec}, w) = \frac{1}{|X_{rec}|} \sum_{x \in X_{rec}} \|x - r(x, w)\|_2^2. \quad (3)$$

the choice of the reconstruction error as an optimized function can be justified using the results of the article [16]. We will use the results proved in [16], where it was shown that an auto-encoder with a special type of regularization allows us to estimate the distribution of  $p(X)$  objects belonging to the general population. Theorem 1 [16]. Let  $p$  be a differentiable probability density and  $\forall x_i \in A^{u \times n_{col}} p(x_i) \neq 0$ . Let  $\mathcal{L}_{\sigma^2}$  be a loss function of the form:

$$\mathcal{L}_{\sigma^2} = \int_{A^{u \times n_{col}}} p(x) \left[ \|x - a(x, w)\|_2^2 + \sigma^2 \left\| \frac{\partial a(x, w)}{\partial x} \right\|_F^2 \right] dx$$

using the results of Theorem 1, we can make the following statement.

Theorem 2. The probability density is represented as:

$$\frac{\hat{a}_{\sigma^2}(x, w) - x}{\sigma^2} \approx -\frac{\partial}{\partial x} E(x), \text{ where } (1/Z) \exp(-E(x)), Z \text{ is the normalization constant.}$$

Proof:

$$a_{\sigma^2}(x, \hat{w}) = x + \sigma^2 \frac{\partial}{\partial x} \log p(x) + o(\sigma^2);$$

$$\frac{a_{\sigma^2}(x, \hat{w}) - x}{\sigma^2} = \frac{\partial}{\partial x} \log p(x) + o(1);$$

$$\frac{a_{\sigma^2}(x, \hat{w}) - x}{\sigma^2} \approx \frac{\partial}{\partial x} \log p(x).$$

Representing  $\log p(x)$  in the form  $-E(x) - \log Z$ , we get the desired expression. Thus, when the regularizer  $\sigma$  tends to zero, a language model is obtained, i.e., the probability distribution of a set of text sequences. The second term of the composite error function is the indentation error [17]. To optimize this error function, a sample of  $X_{me} = \{(x_i, x_j)\}$  consisting of pairs of objects is used:

$$X_{me} = [X_{me}^A, X_{me}^B] \subset A^{u \times n_{col}} \times A^{u \times n_{col}};$$

$$E_{me} = \frac{1}{|X_{me}|} \left( \sum_{(x_i, x_j) \in X_{me}} \max(0, \delta - c_-) + \max(0, \delta - c_+) \right) \quad (4)$$

$$c_- = \cos(h(x_i, w), h(x_j, w)) - \cos(h(x_i, w), h(x_{i'}, w));$$

$$c_+ = \cos(h(x_i, w), h(x_j, w)) - \cos(h(x_j, w), h(x_{j'}, w));$$

$\delta$ -indentation;  $\cos$ -distance (1),

$$x_{i'} = \arg \max_{x_{i'} \in X^B, x_{i'} \neq x_j} \cos(x_i, x_{i'})$$

$$x_{j'} = \arg \max_{x_{j'} \in X^A, x_{j'} \neq x_i} \cos(x_i, x_{j'})$$

The following theorem explains the behavior of this term during the optimization of the parameters  $w$  of the model  $h$ . Theorem 3. Let the following conditions be satisfied:

- i). The hyperparameter  $\delta \in (0, 2)$  is set.
- ii). The sampling power  $|X_{me}|$  is limited to the following value:

$$|X_{me}|(|X_{me}| - 1) \leq \sqrt{\pi} \frac{\Gamma(\frac{u-1}{2})}{\Gamma(\frac{u}{2})} (\int_0^{\cos^{-1}(1-\delta)} \sin^{u-2} x dx)^{-1} \tag{5}$$

iii). Subsamples  $X_{me}^A$  and  $X_{me}^B$  contain all elements in the singular, no element occurs in both samples. Then there is a continuous mapping  $\hat{h}$  from the set of vector representations of the words  $A^{u \times n_{col}}$  to the vector space  $A^u$ , delivering the global minimum of the function  $E_{me} = 0$ . Proof we construct the map  $\hat{h}$  explicitly. Let's say for each pair  $(x_1, x_2)$ :  $\hat{h}(x_1) = \hat{h}(x_2)$  then the time function looks like this up to a multiplier.

$$E_m = \sum_{(x_i, x_j) \in X_{me}} \max(0, \delta - 1 + \cos(\hat{h}(x_i), \hat{h}(x_j))) + \max(0, \delta - 1 + \cos(\hat{h}(x_j), \hat{h}(x_i)))$$

The range of values of the function is bounded from below by zero, which is achieved when the following conditions are met:  $1 - \delta \geq \cos(\hat{h}(x), \hat{h}(x'))$  For any pair  $x \in X_{me}^A, x' \in X_{me}^B, (x, x') \in X_{me}, (x', x) \notin X_{me}, (x, x') \notin X_{me}$ . The number of pairs described above in the set  $X_{me}$  when the third condition of the theorem is fulfilled is equal to  $|X_{me}|(|X_{me}| - 1)$ . We assign the value of the mapping  $\hat{h}$  for each such pair so that  $\cos(\hat{h}(x), \hat{h}(x')) \leq 1 - \delta$ . The existence of such a map follows from the problem of finding a spherical code of maximum size for a sphere in a space of dimension  $u$  and the  $\cos^{-1}(1 - \delta)$ . In, a lower estimate for the sample size satisfying the specified conditions is presented.

The estimate corresponds to the right side of the inequality (5). Since the  $X_{me}$  sample is finite, it is possible to use interpolation polynomials to construct a continuous function given by the conditions described above, which was required to be proved. Note that the mapping proposed in the theorem is continuous, so neural network models can be used to approximate this mapping. According to Tsybenko's n, mappings from the class of neural network models will approximate continuous models arbitrarily well [18]. Thus, the composite optimized function (2) allows us to obtain a model that, on the one hand, has generalizing properties that the language model (3) is responsible for, on the other hand, effectively separates similar and dissimilar phrases from the training sample (4). The hyperparameter  $\alpha$  is responsible for the contribution of each of the optimized terms to this function.

**4.2. The classifier**

For each vector of the phrase  $h(x_a^i)$  from the document being checked  $d_a^i$  is  $v$  nearest vectors by the cosine distance function (1) for fragments from candidate documents  $D_e^{retrieved}$ , using the method of approximate nearest neighbor search. The main purpose of this procedure is to reduce the number of pairs of fragments for classification to reduce the resource intensity of the document comparison stage. For a vector representation of a pair of  $(h(x_{eb}^j), h(x_a^i))$  the following decisive rule is considered:

$$fragments(h(x_{eb}^j), h(x_a^i)) = \begin{cases} 1, \cos(h(x_{eb}^j), h(x_a^i)) > t1, \\ p((h(x_{eb}^j), h(x_a^i))) > t2 \\ 0 \end{cases} \tag{6}$$

where  $p$  is the probability of the classifier;  $t1$  is the threshold of the cosine function of the distance (1);  $t2$  is the minimum threshold of the probability of the classifier. The features are the concatenation of the difference modulo and the component-by-component product of the components of the vector  $[[h(x_{eb}^j) - h(x_a^i)], h(x_{eb}^j) \odot h(x_a^i)]$ . The random forest model acts as a classifier.

**5. SYNTHETIC COLLECTION**

Documents from the English and Arabic versions of the Wikipedia website were used to generate translated borrowings. 100 thousand articles from the Arabic version of Wikipedia were used as a collection of documentaries [19]. A random subset of documents from the Arabic version of Wikipedia was used as a collection of  $D_a$  documents to be checked. To generate borrowings for each document  $d_a^i \in D_a$  the following

algorithm was used: i) Select candidate documents  $\{d_e^j\}$  from the  $D_e$  collection. To reduce the spread of vocabulary in the candidate documents and the document is checked, the selection of candidate documents was carried out from a subset of the 500 most relevant documents for the document being checked  $d_a^j$ . To determine the relevance measure was used. The number of candidate documents was randomly selected from 1 to 10; ii) Select proposals from the candidate documents  $\{d_e^j\}$  randomly and translate them into Arabic and iii) Replace random sentences from the document you are checking  $d_a^i$  with translated sentences from candidate documents. The share of replaced sentences from the document being checked was selected randomly from 20% to 80%.

### 5.1. Optimization of the parameters of the considered models

The fast text library [20] was used as a model for the vector representation of words, the optimization of the parameters of which was carried out on the English version of Wikipedia. The dimension of the vector space for the vector representation of words and fragments was set as 100. To optimize the model of the vector representation of text fragments, the AdaDelta algorithm was used with the parameters  $\varepsilon = 10^{-6}$ ,  $\mu = 0.94$  regularization  $\lambda_2 = 10^{-6}$ . For the final loss function (2), the following values of hyperparameters were set:  $\delta = 0.3$ ;  $\alpha = 0.1$ . The classifier thresholds (6) were selected based on the cross-validation procedure:  $t_1 = 0.6$ ;  $t_2 = 0.5$ . Agglomerative clustering on word vectors was used to build clusters. The cosine function of the distance (1) between the corresponding vector representations was considered as a measure of the proximity of words. The final model contained 29 thousand clusters for 776 thousand words. The recurrent model GRU (gated recurrent unit) was used as model for encoding h and decoding [21], [22]. He was used as a machine translation system, the model of which was trained on 18.4 million parallel sentences from Opus corpora [23], [24]. 10 million sentences from the English version of Wikipedia were used as a sample to minimize the  $E_{rec}$  (3) reconstruction error. The second term of the loss function (4) uses information about similar sentences  $X_{me} = \{(x_i, x_j)\}$ . Pairs of parallel sentences from the OpenSubtitles corpus were used as a sample of such sentences [25]–[27].

### 5.2. Details of the computational experiment

Three experiments were conducted on synthetic data. i) Search for candidates. In this experiment, the quality of the obtained model of word clusters was analyzed. As a basic experiment for comparison, an algorithm based on shingles without reducing words to cluster labels was considered; ii) Comparison of text fragments. In this experiment, we considered the case when the selection of candidates was carried out completely correctly:  $Recall@10 = 1.0$ . The shingle-based algorithm also served as the basic algorithm: the document being checked  $d_a^i$  was translated into English. After that, the resulting text was lemmatized and divided into a set of overlapping 4-grams. To account for possible permutations of words when translating, the words inside each 4-gram were sorted. The result of comparing two documents was a set of matching sorted 4-grams and iii) An experiment evaluating the quality of the entire algorithm (search for candidates and comparison of text fragments).

This experiment allowed us to evaluate the quality of the presented algorithm. The results of the candidate search experiment are presented in Table 1. The presented algorithm based on the construction of clusters gives better quality than the basic algorithm based on shingles. The results of experiments comparing text fragments are presented in Table 2. The presented algorithm shows an accuracy comparable to the accuracy of the basic algorithm and completeness significantly exceeding the completeness of the basic algorithm. The accuracy of the basic algorithm is explained by the fact that this algorithm takes into account the similarity of only almost duplicates of the text. In the third experiment, which took into account the quality of the presented algorithm as a whole, the following indicators were obtained: Precision=0.82; Recall=0.78; and F1=0.79s.

Table 1. Results of the candidate search experiment

Algorithm	Recall@10
Basic	0.88
Presented	0.90

Table 2. Results of experiments on searching for similar text fragment's

Algorithm	Precision	Recall	F1
Basic	0.93	0.14	0.25
Presented	0.87	0.79	0.84

## 6. RESULTS OF EXPERIMENTS ON A REAL COLLECTION OF SCIENTIFIC DOCUMENTS

To test the presented algorithm, an experiment was conducted to search for transferable borrowings on a collection of documents from an electronic library library.iugaza.edu.ps. This resource also contains additional metadata for each document: the title, the authors of the document, the language of the document and belonging to the subject corresponding to the state heading of scientific and technical information. 1.5 million documents in Arabic were prepared for testing the algorithm as verifiable Dr documents. As a collection of  $D_e$  documents, documents from the English version of Wikipedia were used, documents in English from the articles of the resource arXiv.org.

The total number of documents received was 2.4 million. Due to the large number of documents being checked, documents containing a significant number of found borrowings were considered for further analysis. 9 thousand documents with a significant number of borrowings were received. Of these, 5.3 thousand documents were analyzed, selected randomly. The main purpose of the experiment was to detect translated borrowings when the borrowing occurred from an English-language document to an Arabic-language document. At the same time, the analysis of the obtained results revealed several other positives of the presented algorithm, which was further divided into several types: i) conversion of borrowing-the document contains drawing translated from English issued for the original text; M other borrowings-borrowings from Arabic resources or borrowing, the direction of which cannot be determined by document date; ii) bilingual articles-works of the same author in two languages; iii) self-citation-citation by the author of his English-language work; iv) citing laws-using the wording regulations; and v) erroneous responses-false-positive responses of the presented algorithm.

## 7. CONCLUSION

The article discusses a technique for detecting transferable borrowings. The technique for finding transferable borrowings is decomposed, allowing for an efficient search of large text collections for borrowings. The proposed strategy for detecting borrowings is analyzed in detail, as is the composite error function used to optimize the deep learning model. To analyze the quality of the presented algorithm, experiments have been conducted on synthetic data for a pair of languages Arabic-English. The quality of the algorithm is also demonstrated in the collection of Arabic-language documents. In the future, it is planned to develop the proposed algorithm using a model of vector representation of sentences for the task of searching for candidates and improving the quality of the display that matches the vector phrase.




## REFERENCES

- [1] M. M and M. Veerachamy, "Quantitative analysis of plagiarism and academic integrity based on the gender category for the post graduate students to project the efficiency of higher education," *Ilkogretim Online*, vol. 20, no. 5, pp. 527–534, 2021, doi: 10.17051/ilkonline.2021.05.56.
- [2] A. Al-Janabi, E. A. Al-Zubaidi, A. Al-Sagheer, and R. Hussein, "Encapsulation of semantic description with syntactic components for the Arabic language," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, pp. 961–967, 2020, doi: 10.11591/ijeecs.v22.i2.pp961-967.
- [3] M. Muthukamatchi and M. Veerachamy, "Perspectives of students and faculty members towards plagiarism and academic integrity," *Ilkogretim Online*, vol. 20, no. 5, pp. 511–515, 2021, doi: 10.17051/ilkonline.2021.05.54.
- [4] S. Chawla, P. Aggarwal, and R. Kaur, "Comparative analysis of semantic similarity word embedding techniques for paraphrase detection," in *Lecture Notes in Electrical Engineering*, vol. 875, 2022, pp. 15–29.
- [5] D. J. Kaliappan, D. R. L. Kumar, D. N. Prasanth, S. Eshwar, and K. Sundararajan, "Plagiarism Detection using Attention based Text Similarity," *Solid State Technology*, vol. 63, no. 5, 2020.
- [6] V. Linardos, M. Drakaki, P. Tzionas, and Y. L. Karnavas, "Machine Learning in Disaster Management: Recent Developments in Methods and Applications," *Machine Learning and Knowledge Extraction*, vol. 4, no. 2, pp. 446–473, 2022, doi: 10.3390/make4020020.
- [7] C. Xiao, W. Wang, X. Lin, J. X. Yu, and G. Wang, "Efficient similarity joins for near-duplicate detection," *ACM Transactions on Database Systems*, vol. 36, no. 3, pp. 1–41, 2011, doi: 10.1145/2000824.2000825.
- [8] A. Kulmizev *et al.*, "The power of character n-grams in native language identification," in *EMNLP 2017-12th Workshop on Innovative Use of NLP for Building Educational Applications, BEA 2017-Proceedings of the Workshop*, 2017, pp. 382–389, doi: 10.18653/v1/w17-5043.
- [9] R. Xie, L. Zhu, and Y. Cheng, "Review of copy detection techniques for monolingual natural-language documents," in *Proceedings-2018 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2018*, 2019, pp. 631–634, doi: 10.1109/WI.2018.00-24.
- [10] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017, doi: 10.1109/MIS.2017.23.
- [11] J. M. List, "Automated methods for the investigation of language contact, with a focus on lexical borrowing," *Language and Linguistics Compass*, vol. 13, no. 10, 2019, doi: 10.1111/lnc3.12355.
- [12] J.-M. List and R. Forkel, "Automated identification of borrowings in multilingual wordlists," *Open Research Europe*, vol. 1, p. 79, 2021, doi: 10.12688/openreseurope.13843.1.
- [13] G. Sreenivasulu, N. T. Chitra, B. Sujatha, and K. V. Madhav, "Text summarization using natural language processing," in *Lecture Notes in Networks and Systems*, vol. 191, 2022, pp. 653–663.




- [14] T. Peng, I. Harris, and Y. Sawa, "Detecting phishing attacks using natural language processing and machine learning," in *Proceedings-12th IEEE International Conference on Semantic Computing, ICSC 2018*, 2018, vol. 2018-Janua, pp. 300–301, doi: 10.1109/ICSC.2018.00056.
- [15] P. Semberecki and H. Maciejewski, "Deep learning methods for subject text classification of articles," in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017*, 2017, pp. 357–360, doi: 10.15439/2017F414.
- [16] J. Cai, J. Li, W. Li, and J. Wang, "Deep learning model used in text classification," in *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2018*, 2019, pp. 123–126, doi: 10.1109/ICCWAMTIP.2018.8632592.
- [17] B. Alex, "Comparing corpus-based to web-based lookup techniques for automatic English inclusion detection," in *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, 2008, pp. 2693–2697.
- [18] J. A. Evans and P. Aceves, "Machine translation: Mining text for social theory," *Annual Review of Sociology*, vol. 42, no. 1, pp. 21–50, 2016, doi: 10.1146/annurev-soc-081715-074206.
- [19] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing statistical machine translation for text simplification," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 401–415, 2016, doi: 10.1162/tacl\_a\_00107.
- [20] A. Mevis, "Not so foreign," *Amsterdamer Beiträge zur älteren Germanistik*, vol. 82, no. 1, pp. 71–95, 2022, doi: 10.1163/18756719-12340242.
- [21] E. Geller, M. Gajek, A. Reibach, and Z. Lapa, "Applicability of wordnet architecture in lexical borrowing studies," *International Journal of Lexicography*, vol. 34, no. 1, pp. 92–111, 2021, doi: 10.1093/ijl/ecaa013.
- [22] R. Salah, M. Mukred, L. Q. B. Zakaria, R. Ahmed, and H. Sari, "A new rule-based approach for classical arabic in natural language processing," *Journal of Mathematics*, vol. 2022, pp. 1–20, 2022, doi: 10.1155/2022/7164254.
- [23] M. M. Mijwil and E. A. Al-Zubaidi, "Medical image classification for coronavirus disease (covid-19) using convolutional neural networks," *Iraqi Journal of Science*, vol. 62, no. 8, pp. 2740–2747, 2021, doi: 10.24996/ijs.2021.62.8.27.
- [24] A. Al-Janabi, E. A. Al-Zubaidi, and R. H. A. Al Sagheer, "Dependable estimations for education quality using fuzzy logic based strategy a case study (University of Kufa)," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, no. 1, pp. 472–480, 2019, doi: 10.11591/ijeecs.v17.i1.pp472-480.
- [25] M. F. Kadhim, A. Al-Janabi, A. H. Alhilali, and N. S. Ali, "Security approach for instant messaging applications: viber as a case study," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 2, pp. 1109–1115, 2022, doi: 10.11591/ijeecs.v26.i2.pp1109-1115.
- [26] M. M. Mijwil, "Implementation of machine learning techniques for the classification of Lung X-Ray images used to detect COVID-19 in humans," *Iraqi Journal of Science*, vol. 62, no. 6, pp. 2099–2109, 2021, doi: 10.24996/ijs.2021.62.6.35.
- [27] M. M. Mijwil, "Iraqi food image detection using convolutional neural network classification method," in *Lecture Notes in Networks and Systems*, vol. 170 LNNS, 2021, pp. 249–257.

## BIOGRAPHIES OF AUTHORS






**Adel Al-Janabi**    received the B.Sc. in computer science from the University of Babylon, Iraq, and the M.Sc. in computer science from Southern Russian state polytechnical university (NPI) of M. I. Platov, Russia. His research interests include the applications of artificial intelligence, image processing, security and natural language processing. He can be contacted at email: adelh.aljanabi@uokufa.edu.iq.



**Ehsan Ali Al-Zubaidi**    currently a Ph.D. student at the department of computer science and mathematics, University of Kufa. He also currently working as a faculty member at the faculty of Urban Planning. He graduated with a master's degree in computer science from Baghdad University in 2014. Research interest in data mining and machine learning. He can be contacted at email: ihsana.kareem@uokufa.edu.iq.



**Baqer M. Merzah**    He received a B.Sc. degree in computer science from the University of Karbala, Iraq, in 2010, and a master's degree in computer science from Amirkabir University of Technology, Iran, in 2018. Currently, He is a Ph.D. student at the department of computer science at the University of Tabriz, Iran. Also, he is a lecturer at the department of computer science, faculty of education, University of Kufa, Iraq. His research interests are in the areas of machine learning, natural language processing, and data mining. He can be contacted at email: baqirm.merzah@uokufa.edu.iq.