Authors' Writing Styles Based Authorship Identification System Using the Text Representation Vector

Nacer Eddine Benzebouchi*
Computer Science Department, Labged Laboratory
Badji Mokhtar Annaba University, PO BOX 12, 23000
Annaba, Algeria
nasrobenz@hotmail.fr

Nabiha Azizi
Computer Science Department, Labged Laboratory
Badji Mokhtar Annaba University, PO BOX 12, 23000
Annaba, Algeria
azizi@labged.net

Nacer Eddine Hammami
Faculty of Computer and Information Sciences
Jouf University
Aljouf, KSA
nacereddine.hammami@gmail.com

Didier Schwab
LIG-GETALP
Univ. Grenoble Alpes
Grenoble, France
didier.schwab@imag.fr

Mohammed Chiheb Eddine Khelaifia
Annaba, Algeria
chihebkhelaifia@outlook.fr

Monther Aldwairi
College of Technological Innovation
Zayed University, P.O. Box 144534
Abu Dhabi, United Arab Emirates
Monther.Aldwairi@zu.ac.ae

***Abstract*—*Text mining is one of the main and typical tasks of machine learning (ML). Authorship identification (AI) is a standard research subject in text mining and natural language processing (NLP) that has undergone a remarkable evolution these last years. We need to identify/determine the actual author of anonymous texts given on the basis of a set of writing samples. Standard text classification often focuses on many handcrafted features such as dictionaries, knowledge bases, and different stylometric characteristics, which often leads to remarkable dimensionality. Unlike traditional approaches, this paper suggests an authorship identification approach based on automatic feature engineering using word2vec word embeddings, taking into account each author's writing style. This system includes two learning phases, the first stage aims to generate the semantic representation of each author by using word2vec to learn and extract the most relevant characteristics of the raw document. The second stage is to apply the multilayer perceptron (MLP) classifier to fix the classification rules using the backpropagation learning algorithm. Experiments show that MLP classifier with word2vec model earns an accuracy of 95.83% for an English corpus, suggesting that the word2vec word embedding model can evidently enhance the identification accuracy compared to other classical models such as n-gram frequencies and bag of words.*

Keywords*—*Authorship Identification, Text Mining, Natural Language Processing, Word2Vec, MLP classifier

I. Introduction

An enormous quantity of new information being created is stored in diverse unstructured formats. Text mining is a thrilling multidisciplinary domain of research that excerpts helpful information from unstructured or semi-structured text data sources in document sets by exploring and identifying interesting patterns using machine learning (ML) techniques, information retrieval, computational linguistics and natural language processing (NLP).

Authorship Identification (AI) is an important topic in the field of text mining that has undergone a remarkable evolution in recent years. In the typical problem of AI, an unknown author text is attributed to a candidate author, given a set of candidate authors for whom samples of disputed paternity text are available. Figure 1 shows an example of authorship identification process.

In other words, the AI is intended to identify the most probable author of a disputed or anonymous document from a collection of candidate authors [1].

The AI is generally considered a text classification task, starting with data preprocessing, and then with feature engineering (extraction), representing document/text as a characteristic vector. Feature engineering (FE) is a crucial topic in the ML process. This is the use of data domain knowledge for vector representation of raw data and the creation of characteristics that enable ML algorithms to function.

Lately, the practical applications of AI have developed in various fields such as criminal law [2], opinion detection [3], civil law [4] and intelligence agencies work [5]. AI has also become an important component of other identification technologies, such as cryptography, signature and intrusion detection systems [6].

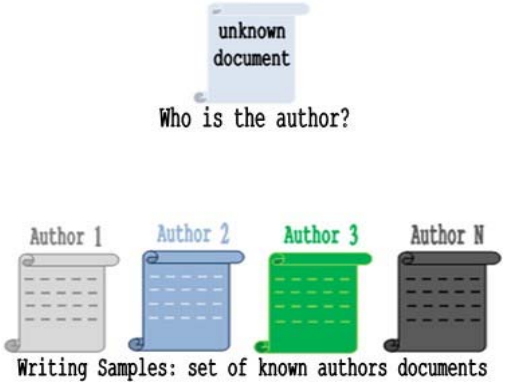

Fig. 1 Authorship Identification Problem

One of the main tasks of AI problem is the extraction of the most relevant characteristics to represent the author's writing style. Indeed, numerous researches have been done in this domain and many techniques have been suggested [7,8]. Most standard studies are based on human-designed characteristics, such as lexical [9], syntactic [10], content-specific [11] and stylometric features [12].

Mohsen et al. [11] proposed an author identification system using variable size characters n-grams as the

document representation to extract characteristics using deep learning. A Stacked Denoising AutoEncoder is used for extracting document characteristics. For the classification phase, authors applied the Support Vector Machines (SVM) classifier. Sarwar et al. [12] presented a cross-lingual authorship identification approach of several authors based on stylometric features such as Vocabulary, Structural and Punctuations characteristics. The kNN classifier is used for the classification stage. Zhang et al. [13] presented an authorship identification method based on the stylistic characteristics of texts using principal components analysis (PCA) and linear discriminant analysis (LDA). The authors use a semantic association model on voice, dependency relations between words and non-subject stylistic words to describe each author's writing style. Zhang et al. [14] suggested a source code authorship identification approach based on a logic model of continuous word-level n-gram and discrete word-level n-gram, and a multi-level context model for building author profiles of programmers. A sequential minimal optimization algorithm for SVM formation is applied for identification. Table 1 summarizes some state-of-the-art techniques suggested for AI.

TABLE I. Related studies for authorship identification

Ref	Classification Method	Nbr of Authors	Nbr of Docs	Nbr Lang	Features/ Method
[11]	SVM	50	5000	1	variable size character n-grams
[12]	KNN	400	825	6	Vocabulary Richness, Structural, Punctuations
[15]	KNN	136	2386	1	Lexical, Syntactic, Structural
[16]	SVM	6	34	2	Part of Speech
	Logistic Regression	6	34	2	Word 1,2,3-gram
[17]	Random Forest	8	120	4	Vocabulary Richness
[18]	Conv. Neural Net	6	-	1	Characters
Proposed	MLP	8	72	1	Word2vec

FE is the foundation of any ML application [19] and is both difficult and expensive. The need for manual FE can be obviated by automated feature learning [20, 21].

Recently, Word Embeddings (WE) have grown rapidly and have created a new search domain in many tasks of text mining and natural language processing (NLP) [22, 23]. WE automatically extracts the most relevant information from raw text and provides distributed word vectors that take into account relatively small size syntactic and semantic aspects to enable better processing of machine learning over a simple word coding [24].

This study suggests an authorship identification system using feature engineering with network embedding for the distributed vector representation of words, taking into account each author's writing style. This method consists of two learning parts. On the one hand, we generate the semantic representation of several authors by using *word2vec* word embeddings model to learn and extract the most appropriate characteristics to describe the author's writing style from the raw document. On the other hand, we apply the multilayer perceptron (MLP) classifier to fix the classification rules using the *backpropagation* learning algorithm.

This paper is organized as follows. Section 2 illustrates the general concept of *word2vec* WE models and Section 3 explains the principal steps of the proposed method. In section 4, we present the obtained results of our work. Finally, section 5 describes the conclusion of this study.

II. LEARNING WORD REPRESENTATIONS: *WORD2VEC*

Choosing a good word representation is often necessary for many text mining tasks and NLP. The *word2vec* (W2V) is a widely used tool in recent times because of its better results and fast training. W2V is a set of patterns designed to generate word embeddings (WE). WE is a recent and very popular method in the domain of text mining and NLP that allows learning vector representations of words from raw documents. WE also captures syntactic and semantic relationships between the words of relatively small size, contrary to other classic methods such as one-hot representation, knowledge-bases and n-grams primitives.

Word2vec models are shallow two-layer neural networks that are formed to rebuild the linguistic contexts of words. Its entry is a large corpus of text that will then generate a vector space, with for each word of the corpus a corresponding vector in space. W2V consists of two models, namely a continuous bag of words (CBOW) and continuous skip-gram in order to generate a distributed vector representation of words. The CBOW model uses the context to predict a target word and the skip-gram model does the opposite of the latter, uses a word to predict a target context (see Fig.2).

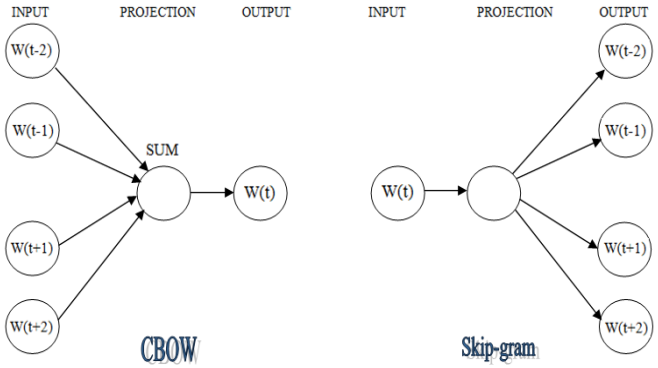

Fig. 2 Word2vec Architectures

III. Proposed Approach

This study is part of the general framework of authorship identification, and more precisely the analysis of the vector representation for the extraction of robust and adaptive primitives. We opt for the Word2vec (W2V) method in this critical phase to ensure the semantic aspect and to maintain the characteristics obtained very close to the real data in order to obtain an effective authorship identification system.

The architecture that we propose for author identification task by using W2V can be divided into two phases: the offline phase for learning the general model. It allows the creation of W2V models for each of the authors from the text documents associated with each of them. The classification phase (online), it consists of assigning an author number to an input document. During this phase, we proposed two methods for generating the vector representation of an unknown entry document.

A. Creation of datasets:

In this work, we treat a problem with N authors, each author P documents. Initially, these N × P documents are in textual form. By applying Word2vec to a text file, we will have its vector model; each word will have its own vector.

The following steps introduce the creation of the database from N × P text files:

- Group the texts of each author into a single text file.
- The application of Word2vec on these latter to create the vector model of each author.
- After creating the vector models of each author, we can now create the vector models of each document from the authors' models. that is, author A, who has his own documents, has his own vector model; we take for example document 1, in this step we will not apply the word2vec, since we have the vector model of the author A, so to generate the vectors of the words of document 1, take each word of the latter and go look for its vector in the author's model A.
- We obtain as results N × P vector models that correspond to the N × P textual documents.

At this point, we will reduce the size of these N × P vector models, each model of the latter containing a very large number of words with their vectors. After several empirical tests, we will choose from this large number of words only six random words that will represent the documents. This choice gave better results.

So, to summarize this step, each document model will contain only six representative words with their vectors of size T for each word. That is, the vector size that will represent the document will be 6 × T, plus 1 which represents the class (author). Figure 3 illustrates the principle of the offline phase.

B. Supervised learning algorithm: Multilayer Perceptron (MLP) classifier

After the process of representing each document with a feature vector, the Multilayer Perceptron (MLP) classifier with a single hidden layer is adopted for the formation of our system (to fix the classification rules). The MLP classifier has proven its ability to successfully deal with non-linear phenomena. Its operation is based on the retro-propagation of the error gradient in multilayer systems. We use MLP classifier as a supervised learning algorithm: As an input of MLP, N × P vector models correspond to the N × P textual documents, taking into account the semantic and syntactic aspect between the words of the documents. The output layer consists of N neurons that represent the N authorships.

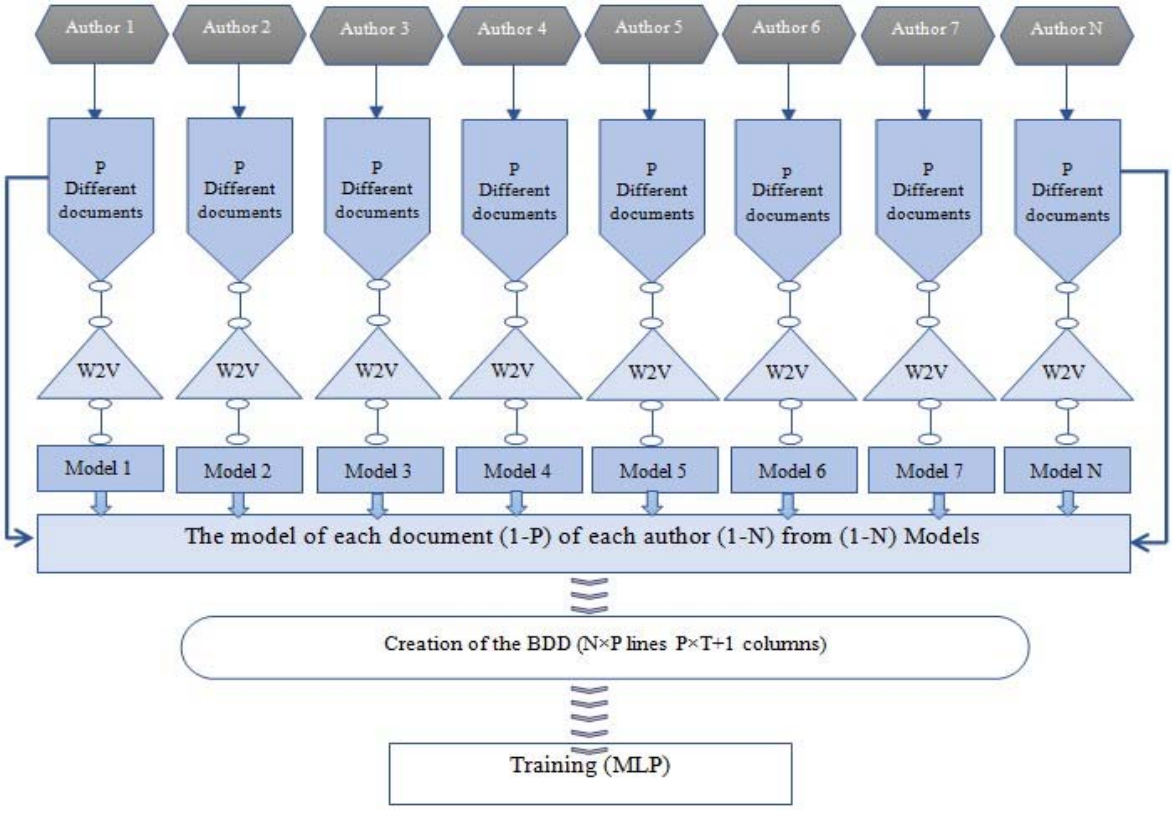

Fig. 3 Creation of Datasets- Word2vec and Model Learning

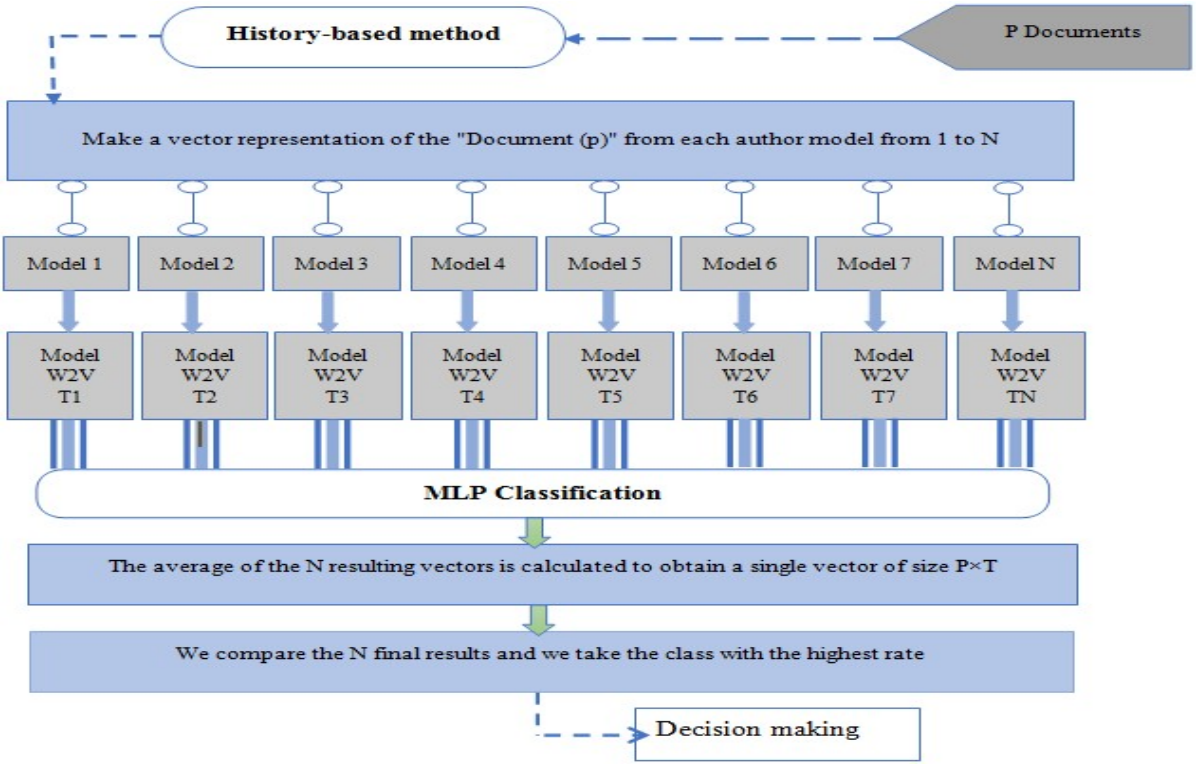

Fig. 4 Validation Process of Proposed Approach

During the classification phase (Online), each unknown document having M input words must absolutely be represented by a feature vector using the W2V vector representation based on the authors' history, i.e. the author's writing style. In this case, a document may have several vector representations of words (W2V) depending on the author that it has written. This is ensured by browsing the vector models of each of the N authors. The Figure above (Fig. 4) shows the process of the online phase.

The MLP classifier will return a probability of belonging to each of the predetermined classes (authors), the final decision will be based on these probabilities, and the class that will have the highest probability will be the output class.

IV. Experimental Results

In order to evaluate the use of the text vector representation for document classification and validate the performance of our system according to the adopted Word2vec model (Skip-Gram or CBOW), this section tests the ability to classify unknown documents, which are not previously learned from our approach, and present the details of the experiments and their results.

A. Used dataset

We use PAN 2012[1] datasets written in English. This database has been developed with 8 authors, including 9 documents for each of them. To start the execution of our authorship identification model, we have set the parameters by referring to the database: N=8 represent the number of authors and P=6 describe the number of documents. To validate our approach, we divided the database into two parts: 48 for learning and 24 for testing (6 documents for each author to learn and 3 documents for each author to be tested).

B. Feature Engineering using Word2Vec Model

We construct various models based on different architectures and the dimensionality of words vector using *word2vec (w2v)*. Table 2 illustrates the training parameters used for several models based on the *Skip-Gram* and *CBOW* models.

TABLE II. Model Configurations

Model / parameters	CBOW	Skip-gram
LayerSize	50	50
WindowSize	5	5
MinWordFrequency	1	1
Iterations	3	3
LearningRate	1.0E-4	1.0E-4
Sampling	1.0E-5	1.0E-5

The *word2vec* configuration accepts a number of hyper-parameters. Some explanations of these parameters are presented below:

- **Layerize:** defines the features number in the word vector. This is equal to the number of dimensions in the feature space.
- **WindowSize:** when processing a word in the corpus, number of contextual words to be taken into account for sampling purposes.

[1] http://pan.webis.de/data.html

- **MinWordFrequency:** is the minimum number of occurrences of a word that the corpus must contain for this word to appear in a generated ontolection. Here, if it appears less than 5 times.
- **Iterations:** this is the number of times you allow the network to update its coefficients for a batch of data.
- **LearningRate:** is the step size for each update of the coefficients, when the words are repositioned in the feature space.
- **Sampling:** is the word sampling frequency.

In this study, we opt for the *CBOW* model, which offers good performance and produces more accurate results.

C. Results and Discussion

In order to evaluate the performance of our system, we use the accuracy evaluation criterion defined in Eq.1:

$$Accuracy = \frac{Number\ of\ success}{Number\ of\ document} \times 100\% \quad (1)$$

The proposed method is based on the vector models of the authors in the creation of the models of the test files; but instead of applying the *w2v* on these files to create the models, we will create them from the models of the authors already created before depending on the history -written style-, and then the resulting models are passed to the MLP classifier

The MLP classifier will return a probability of belonging to each of the predetermined classes (authors), the final decision will be based on these probabilities, and the class that will have the highest probability will be the output class. The results show an accuracy of 95.83%, 23 documents were well ranked among the 24 incoming texts.

In order to better analyze the *w2v* method, we performed several empirical tests on some parameters, namely: *WindowSize* (ws) and *MinWordFrequency* (*mwf*), the table below summarizes some results.

TABLE III. Obtained Results from Different Experiments About *ws* and *mwf* Parameters

WindowSize	MinWordFrequency	Accuracy (%)
5	1	95.83
1	1	79.16
5	5	83.3
1	5	87.5

Note that when *mwf* = 1 parameter allows ignoring all the words with a lower total frequency 1, in this case it will take all the words, and a value of *ws* = 5 implies that the maximum distance between the current word and the word in the sentence is 5.

From these results, we can conclude that the frequency of the words as well as the size of the window treating the neighboring words can affect the performance of the global classification system, hence the need to perform these experiments for each new database.

According to our case of experiments, we find that *mwf = 1* and *ws = 5* give better results.

V. Conclusion:

Text mining is currently evolving as technology and parallel architectures and new approaches to data mining are advanced. This system that we have designed and realized is intended for the authorship identification from documents written by them, based on a popular approach and technique, namely the *Word2Vec (W2V)* vector representation.

Word2Vec has attracted a lot of attention in data representation thanks to its efficiency and the consideration of syntactic and semantic aspects as well as the relations that can exist between the words of the texts. According to our knowledge, this is the first study that uses Word2vec Word Embedding for the authorship identification task.

This approach offers new techniques for extracting adaptive features from any textual database, making our system flexible and accurate. The experimental results of this method prove that the *Word2Vec* method and more particularly the *CBOW* technique is very effective for authorship identification.

As an extension of our work, we intend to expand our system by offering:

- It can be used in various fields except for authorship identification, such as plagiarism detection, the fight against spam, machine translation, etc.
- The enrichment of our system by the integration of other techniques such as *GloVe* and *ELMo* [25].
- Analyze the system on a larger scale: by increasing the number of authors or the number of documents per author.

Acknowledgment

This work has been supported in part by Zayed University Research Incentives Grant R18054

References

[1] K. Burns, "Bayesian inference in disputed authorship: A case study of cognitive errors and a new system for decision support," Information Sciences, vol.76 , pp.1570–1589, 2006

[2] F. Iqbal, H. Binsalleeh, B.C.M. Fung, M. Debbabi, "A unified data mining solution for authorship analysis in anonymous textual communications," Information Sciences, vol.231, pp.98–112, 2013

[3] A. Ziani, N. Azizi, Y.T. Guiyassa, "Combining random sub space algorithm and support vector machines classifier for Arabic Opinions Analysis," Advances in Intelligent Systems and Computing 358, Springer 2015, ISBN 978-3-319-17995-7, pp.175-184

[4] T. Grant, "Quantifying evidence in forensic authorship analysis," International Journal of Speech, Language & the Law, vol.14, 2007

[5] A. Abbasi, H. Chen, "Applying authorship analysis to extremist group web forum messages," IEEE Intelligent Systems, vol.20, pp.67–75, 2005

[6] A. Orebaugh, "An instant messaging intrusion detection system framework: Using character frequency analysis for authorship identification and validation," In Carnahan Conferences Security Technology, Proceedings 40th Annual IEEE International, pp. 160–172, 2006

[7] M. Al-Ayyoub, Y. Jararweh, A. Rabab'ah, M. Aldwairi, "Feature extraction and selection for Arabic tweets authorship authentication," Journal of Ambient Intelligence and Humanized Computing, Vol. 8, No. 3, pp.383–393, 2017. https://doi.org/10.1007/s12652-017-0452-1

[8] A. Rabab'ah, M. Al-Ayyoub, Y. Jararweh, M. Aldwairi, "Authorship Attribution of Arabic Tweets," In the Proc. of the13th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2016), Agadir, Morocco, Nov 29 – Dec 02, 2016. pp.1-6. https://doi.org/10.1109/AICCSA.2016.7945818

[9] E. Stamatatos, "On the robustness of authorship attribution based on character n-gram features," Journal of Law & Policy, vol.21, pp. 421–439, 2013

[10] R. Zheng, J. Li, H. Chen, Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," Journal of the American Society for Information Science and Technology, vol.57(3), pp. 378–393, 2006

[11] A. M. Mohsen, N. M. El-Makky, N. Ghanem, "Author Identification Using Deep Learning," 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016

[12] R. Sarwar, Q. Li, T. Rakthanmanon, S. Nutanong, "A scalable framework for cross-lingual authorship identification," Information Sciences, vol.465, pp. 323–339, 2018

[13] C. Zhang, X. Wu, Z. Niu, W. Ding, "Authorship identification from unstructured texts," Knowledge-Based Systems, vol.66, pp. 99–111, 2014

[14] C. Zhang, S. Wang, J. Wu, Z. Niu, "Authorship Identification of Source Codes," In: Chen L., Jensen C., Shahabi C., Yang X., Lian X. (eds) Web and Big Data, APWeb-WAIM , Springer, pp.282-296, 2017

[15] S. Nutanong, C. Yu, R. Sarwar, P. Xu, D. how, "A scalable framework for stylometric analysis query processing". In ICDM, 2016

[16] D. Bogdanova, A. Lazaridou, "Cross-language authorship attribution" .In LREC, pp. 2015–2020, 2014

[17] M. Llorens-Salvador, S. J. Delany, "Deep level lexical features for cross-lingual authorship attribution," In ECIR, pp. 16–25, 2016

[18] Y. Kim, "Convolutional neural networks for sentence classification," CoRR, abs/1408.5882, 2014

[19] N. Azizi, N. Farah, M. Sellami, "Off-line handwritten word recognition using ensemble of classifier selection and features fusion," Journal of Theoretical & Applied Information Technology , vol.10, pp. 141-150

[20] N.E. Benzebouchi, N. Azizi, K. Ayadi, "A Computer-Aided Diagnosis System for Breast Cancer Using Deep Convolutional Neural Networks." In: H. Behera, J. Nayak, B. Naik, A. Abraham (eds) Computational Intelligence in Data Mining. Advances in Intelligent Systems and Computing, Springer, Singapore, vol 711, pp.583-593, 2019. https://doi.org/10.1007/978-981-10-8055-5_52

[21] N.E. Benzebouchi, N. Azizi, S.E. Bouziane, "Glaucoma Diagnosis Using Cooperative Convolutional Neural Networks," International Journal of Advances in Electronics and Computer Science, Vol.5, No.1, pp. 31-36, 2018, ISSN 2393-2835.

[22] N. E. Benzebouchi, N. Azizi, M. Aldwairi, N. Farah, "Multi-classifier system for authorship verification task using word embeddings," 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), Algiers, Algeria, 25-26 April 2018, IEEE, pp.1-6. DOI: 10.1109/ICNLSP.2018.8374391

[23] R.A. Stein, P.A. Jaques, J.F. Valiati, "An Analysis of Hierarchical Text Classification Using Word Embeddings," Information Sciences, vol.471, pp. 216-232, 2019

[24] T. Mikolov, W.t. Yih, G. Zweig, "Linguistic regularities in continuous space word representations," Proceedings of NAACL-HLT, pp.746–751, 2013

[25] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, "Deep contextualized word representations," Proceedings of NAACL-HLT, pp. 2227–2237, 2018