RESEARCH ARTICLE OPEN ACCESS

Anomalies in texts using Clustering and Convolutional Neural Networks*

Asmaa Salem*, Gabriela Andrejková**
**(Department of Computer Science, Pavol Jozef Šafárik University in Košice,*
*** (Department of Computer Science, Pavol Jozef Šafárik University in Košice,*
Corresponding Author: Asmaa Salem

ABSTRACT
Anomaly detection is an identification of an unknown behavior of the system. Finding anomalies in a text document may be useful, for example, in detecting internal or external plagiarism. Our goal is to analyze text segments of some long text and find segments which have some anomalies, a different stylometry in comparison to the other segments or to the full text. We present a new two-steps method: (Step 1) clustering of segments, and (Step 2) classification of segments using convolutional neural networks. The method was tested on ten Arabic and ten English long texts. Our new method contributes to the previous results about texts. The algorithm classify texts into two classes: a reliable or a suspicious text and in many cases it confirms the previous evaluation of analyzed texts. The results are compared to results computed by three other in short described methods.

Keywords – . Text, Segment, Anomaly Detection, Convolutional Neural Network, Clustering.

Date of Submission: 30-03-2019 Date of acceptance: 12-03-2019

I. INTRODUCTION

Text documents are analyzed on content, significance, and similarity to other documents, in order to better understand, to process, to efficiently investigate or to detect external and external plagiarism. The results can be used for example in artificial intelligence, when an unknown author is attributed to a text from a large number of candidates and in many other areas such as psychiatry (symptoms of some mental illusions of the author may be reflected in his texts) or in crime.
The authorship verification task solves the problem of determining an author (one of more authors) in an anonymous text. In solving of this problem it is possible to try to use linguistic features of the author texts, features that will allow to create a model of the writing style of this author and to measure how similar may be some unknown text to texts written by that author [7], [8]. The texts can be analyzed from the syntax or semantics point of view. The paper deals with the analysis of a text document from its author the point of view. Every author writes texts in his/her own personal style, he uses similar means of expression in different texts, uses similar words, words are composed to sentences similar to the frequency of occurrence in various texts of the same author. Attributes describing the author's expression can be imagined by many ways. Such attributes describe how the author expresses the text document and describes the author's stylistic space. The problem is more difficult if an author is not known or if we need to cover if the full text was written by the same author. We follow some text and try to find many characteristics of it. Our problem can be described: To cover stylistic anomalies among text segments of the same text. We are concentrated to long texts. Since we obtained the results by dealing with the long texts in the previous analysis [1], [2], [5], [6], we can say that we need to conduct further analysis in order to obtain satisfactory results. We developed in [1] a new method (CL-CNN algorithm) based on clustering and convolutional neural networks to support our results. New results about texts are presented in the paper using the method with ART2 clustering [2]. The paper contains the following sections: In the second section, we present the results obtained using two systems which do not use clustering methods. The systems were applied to long texts and their results are important in a comparison with new results. The third section contains a description of two methods which use clustering, our previous method (using Self Organizing Maps) and our new method which uses convolutional neural networks (CNN). The results are written in the fourth section. The fifth section is conclusion and it contains an evaluation of results and our plan for the next research.

1. Previous Analyzing Methods, Without Clustering

We analyze long texts and we find some parts of texts they have different stylometric properties than the other parts or the full text.

A. Information about used texts

We work with long texts, it means the length of one text is more than 150 000 letters.

Arabic texts:

The Arabic texts are chosen from [4]. They are named A1-A10. The used texts: Almaghnii book (speaks about the rules of purity in Islamic religion); Seaba'weah book (speaks about Arabic grammar); History book (speak about the liberation of Egypt and Morocco); Alkhasaes book (speaks about philology); The text speaks about the prophet Mohammed's friends; The book of Lineage (speaks about Nazar bin Maad); The text speaks about find out the best country in the world; The text speaks about Manuscripts; Islamic book speaks about Adam prophet; The text speaks about History of Arabic Literature. The basic statistics of 10 Arabic texts is described in Table 1.

Table1. Statistics of 10 Arabic texts, the number of words by length from 1-4.

Name of texts	# words	# letters	#words by Lengths 1-4			
A1	94847	489223	5	13218	23297	22532
A2	48430	246278	48	7358	12130	11636
A3	95302	469098	90	15398	23522	25065
A4	32205	167485	84	4810	7794	6328
A5	58990	318847	12	8079	12381	12102
A6	81501	408295	14	21107	17911	19991
A7	68181	376480	79	8139	14251	15431
A8	78984	400665	750	11837	18457	19385
A9	92240	492066	172	12503	21464	21573
A10	106579	555981	932	17312	23435	24782

B. Element n-gram profiles

The method is based on a similarity/dissimilarity of the text parts and their occurrences of n-grams in a comparison to the full text. The idea of the method was developed by Stamatatos [8]. n-grams are built from symbols or words. In our previous analyzes [5] we showed that the number of words with 4 letters in Arabic texts is higher than the number of words with 3 letters. It is illustrated in Tables I. We decided to work with 4-grams of symbols for Arabic language [5]. Normalized dissimilarity measure of the text T and its coherent part A is defined as nd by (1):

$$nd(A,T) = \frac{1}{|{}^nA|} * \sum_{{}^ng \in p(A)} \left[\frac{|{}^nT| - |{}^nA| * k^n_{A,T}({}^ng)}{|{}^nT| + |{}^nA| * k^n_{A,T}({}^ng)} \right]^2$$

where M = |T| is the length of T in symbols. ${}^nT$, ${}^nA$ are the sets of all different n-grams in the text T and in its part A, ${}^nA \subseteq {}^n$ T ${}^nT$, ${}^nA$ are the numbers of n-grams in T and A. $k^n_{A,T}(g) = |o^n_T(g)|/o^n_A(g)|, |o^n_T(g)|.|o^n_A(g)|$ are the numbers of the n-gram g in T and in A respectively. $|o^n_T(g)| \geq |o^n_A(g)| \geq 1$. In the formula (1) only n-grams from ${}^nA$ are analyzed, it means the dominator in $k^n_{A,T}(g)$ is $|o^n_T(g)| \geq 1$. The denominator $|{}^nA|$ ensures that the values of the dissimilarity function belong to the interval $\langle 0,1 \rangle$ (highest similarity is 1). It is possible to define the stylometric function of a text T using n-gram profiles of the moving windows as follows:

$$sf(i,T) = nd(W_i,T), i = 1..\lceil |M|/s \rceil,$$

where W_i is a window of the length l ≥ 1 , s ≥ 1 is the moving distance of the window, $\lceil M/S \rceil$ is the total number of windows in a text.

We developed the algorithm [5] which works with stylometric functions on moving text windows across the full texts. In the algorithm there are described conditions for removing of some windows from the text. The rest windows are evaluated in percentages of dissimilarities and it is given border for an evaluation of text as suspicious or reliable. The percentage of dissimilarities of Arabic texts is shown in Table 2. We use border 40% for classification. The percentage which is more than 40% means that the text is a suspicious text.

C. Histogram of words and symbols

The motivation to the method we found in [9]. The full text T is split into r coherent text parts $T_1, T_2, ..., T_r, r \geq 1$, N_i is the length of $T_p, 1 \leq p \leq r$. Each text part T_k is mapped into interval $\langle 0,1 \rangle$. The histograms are prepared for text parts using some kernel smoothing function and evaluated using three chosen distances. Dissimilarities of the text parts T_{i1} and T_{i2}, $i1 \neq i2, 1 \leq il, i2 \leq r$ were analyzed using distances of histograms on words.

If histograms will be normalized into interval $\langle 0,1 \rangle$ then they are comparable using similarity functions of two sequences. A comparison of the normalized histograms of the full document to the normalized histograms of the document parts shows anomalies between the text and its parts. These three functions can show the values of an intersection for two compared text parts. The used distance functions:

- Euclidean distance function.
- Intersection distance function - its ability is to handle partial matches when the areas of two histograms are different.
- x^2 statistics distance function - it measures how unlikely it is that one distribution was drawn from the population represented by the other.

For example, Arabic text A4 was split into 4 coherent parts T1, T2,T3, T4 as it is shown in Table 2 . Five histogram center positions for μ were analyzed for each histogram of text parts. Distances between each pairs of text part histograms were computed. The results of the text A4 show that all distances of the text part T3 to the other text parts have higher values than distances between the residual text parts, it is shown in Table 2.

We recommend to use the mean value of all distances as border parameter values, but the text part T3 has the higher distances to all residual parts for all 5 histogram vectors. It means the part T3 of the Arabic text A4 is critical (suspicious)[5].

D. Results about used texts

Both methods were applied to an evaluation of above described texts A1-A10.

Table2. Results of 10 Arabic texts using N-GRAM and Histogram methods

Text	4-Gram %	Text Results	Histogram of words			Text Results
A1	31.074	Reliable	T2-T5x	T2-T5x	T2-T5x	Suspicious
A2	26.256	Reliable	T1-T4x	T1-T4x	T1-T3x	Reliable
A3	35.233	Reliable	T4-T5x	T4-T5x	T4-T5x	Suspicious
A4	34.077	Reliable	T3-T5x	T3-T5x	T3-T5x	Suspicious
A5	38.302	Reliable	T4-T5x	T4-T5x	T4-T4x	Suspicious
A6	34.668	Reliable	T4-T5x	T4-T5x	T4-T5x	Suspicious
A7	33.655	Reliable	T4-T5x	T4-T5x	T4-T5x	Suspicious
A8	33.046	Reliable	T2-T4x	T2-T5x	T2-T5x	Suspicious
A9	35.805	Reliable	T2-T3x	T2-T4x	T2-T4x	Reliable
A10	33.774	Reliable	T4-T5x	T4-T5x	T4-T5x	Suspicious

2. Clustering Combined with neural networks methods

Clustering is one of the most popular data mining algorithms and have extensively used in a text context. It has mostly of applications for example in a classification of short texts as advertisement. The clustering of texts is a problem of uncontrolled learning, whose task is to divide a set of textual documents (input data) based on criteria (certain common features and mutual similarity between them) into the clusters we are writing about. Concerning documents, we can write about categorization and clustering. When categorizing texts where we are talking about the problem of controlled learning, we have information about each category before we start the learning. We do not have predetermined clusters when clustering texts. Algorithms learn by themselves during the clustering process. We do not have information about clusters, they are not given in advance and according to the chosen method they can create new ones as needed. The similarity is measured by using a similarity function. Text clustering can be in different levels of granularity where clusters can be texts, paragraphs, sentences or terms. Algorithms for clustering texts that use neural networks include Self-Organizing Maps and ART neural networks.

Our approach is oriented to long texts. Long text can be split into paragraphs, we call them segments. We analyze then similarity and clustering of segments. In [15], it is done very good survey of clusters method used in text processing. But we concentrate to the clustering of segments and sentences in some text. The paper [11] proposes three approaches: Unsupervised, Semi Supervised techniques and Semi Supervised with a reduction of dimension to construct a clustering based classifier for Arabic.

A. System of Self Organizing Maps (SOM)

Self-organizing maps are a model of neural networks based on the principle of uncontrolled learning. According to [10] they map the high-dimensional flag space into the lowdimensional space we are talking about. The neurons are arranged in a regular pattern, a grid that represents the exit space. The view in the self-organizing map preserves the topology. Each neuron has its vector weight whose size is equal to the size of the input vector, and the distance between neurons is counted as the Euclidean distance. Neurons have their potential or activation. This counts when the input vector arrives at the input. The winning neuron is the neuron with the greatest activation. SOM models are interesting models [10] usable in the text processing. We developed a system of modified SOM networks working on probabilistic sequences built from a text. The system has two different shapes but we describe in details one of them.

The system 4SOM ^{A} [5] is drawn in Fig. 1. We use input sequences such as words, 2-gram of words, 3-gram of words and 3-gram of symbols: SOMx ; x $\in$ {words, 2-gram of words, 3-gram of words, 3-gram of symbols} The main evaluation is based on a cumulative error and text part evaluation.

In the first layer, it has four SOM networks, SOMx ; x $\in$ {words, 2-gram of words, 3-gram of words, 3-gram of symbols}, they are trained to different sequences built from the text T.

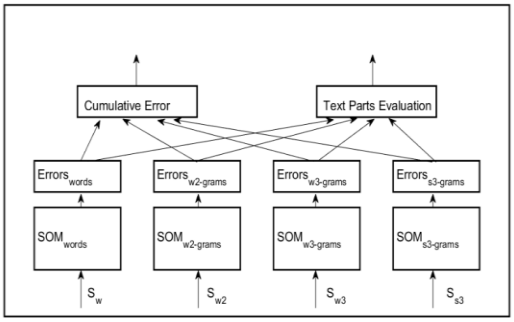

Fig.1. System $4SOM^A$ for anomaly detections. A cumulative error is based on the results of the quality and text part evaluation. The training of four SOM networks is done using sequences Sw, Sw2, Sw3, Ss3 - words, 2-gram of words, 3-gram of words and 3-gram of symbols, respectively.

The computation of a cumulative error: After the SOMx was trained, we evaluate how good was the training done by an evaluation of errors for all input vectors (all text parts in the text). We will use a quantization error Er_x defined by:

$$Er_x(x^+, w_{i^*}) = \left\| x^+ - w_{i^*} \right\|, \qquad (2)$$

as a measure of proximity input vector x+ to the learned winner vector w_{i^*} of $i^* - th$ neuron (winner for input vector x+) in the SOM_x. Using formula (2) it is possible to compute the vectors of quantization errors

$$\{Er_x(x^+(t), w_{i^*}(t))\}_{t=1}^{R} \qquad (3)$$

where R is the number of training vectors, t is the order of the member in input sequence. Let α be a significance level (α = 0:01 or α = 0:05). We suppose the percentage of normal values of the quantization error will be 100 * (1 - α). Let Nα be the real number such that a percentage 100 *(1 - α) of the error values is less than or equal to Nα.

Then

Lower limit: $\lambda^- = N_{1-}\alpha_{/2;}$ Upper limit: $\lambda^+ = N\alpha_{/2}$

The important interval is $\langle \lambda^-, \lambda^+ \rangle$ the values out of it could be detected as anomalies.

The text parts evaluation: The text T is split into r; r > 1 disjunctive coherent parts, $T = T_1T_2 \ldots T_r$. For each text part T_k, $1 \le k \le r$ the following evaluation will be done: $T_{-k} = T_1 \ldots T_{k-1}T_{k+1} \ldots T_r$ will be used as a training text and T_k will be a testing text. After the training using T_{-k} in the system, the testing text T_k will be evaluated using the quantization vector (3) and intervals $\langle \lambda^-, \lambda^+ \rangle$, the interval is built on training data. The percentage of values x , $x \in \langle \lambda^-, \lambda^+ \rangle$ expresses how the text T_k is similar to the training text.

Table3. The results of the evaluation using all developed methods- text A4 belongs to the set of suspicious texts[5].

Method	A4, the no of symbol =135573, W=31656			
System $4SOM_A$	Words	2-gram of words	3-gram of words	3-gram of symbols
$A4_{T1}$	90.584 %	89.749%	80.512 %	89.392 %
$A4_{T2}$	85.731 %	90.931 %	89.359 %	88.633 %
$A4_{T3}$	**72.430 %**	84.764 %	**29.262 %**	92.744 %
$A4_{T4}$	91.497 %	92.455 %	89.101 %	91.377 %

The results for Arabic text A4 in the system $4SOM^A$ are illustrated by the similarity percentage of input sequences: words, 2-gram of words, 3-gram of words, 3-gram of symbols. we divided the text into 4 parts. According to the similarity percentage we can see in the columns "words" and " 3-gram of words" the third part is not similar to the another parts. the values is 72:430% and 29:262%.

The system $4SOM^B$

The shape of the system is very similar to the previous system. The differences are the following: Three input sequences are prepared from symbols, only one SOM network is trained to 3-grams of words. We illustrate the similarity percentage of input sequences such as: symbols, 2-gram of symbols, 3-gram of symbols, 3- gram of words.

We can see in the column " 3-gram of words" the third part is not similar to the another parts. the value is 29:262%. It means the values is less than b = 75:000%, that means the text part T_3 of A4 is suspicious [5].

Table4. The results of the evaluation using all developed methods- textA4 belongs to the set of suspicious texts[5]

Method	A4,the noof symbol =135573, W=31656			
System $4SOM_B$	Symbols	2-gram of symbols	3-gram of symbols	3-gram of words
$A4_{T1}$	88.319%	91.418 %	89.392 %	80.512 %
$A4_{T2}$	88.291%	89.127 %	88.633 %	89.359 %
$A4_{T3}$	86.940%	90.931 %	92.744 %	**29.262 %**
$A4_{T4}$	87.694%	92.095 %	91.377 %	89.101 %

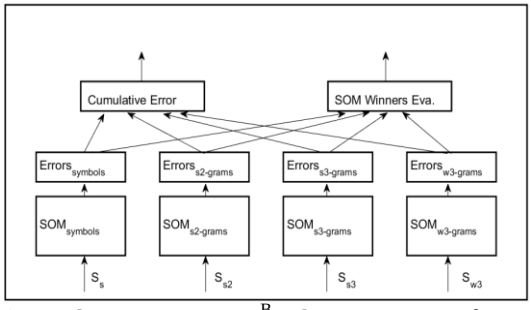

Fig.2. The system 4SOM^B. The percentage of a text parts similarity in Arabic text A4. Input sequences :symbols, 2-gram of symbols, 3-gram of symbols, 3-gram of words.

In Fig.3 we show the clusters of all four trained SOM networks. The training of symbols and 2-grams of symbols has special shape of winners. The figures show that declared clusters of SOM networks trained for symbols and 2 grams of symbols have any information. The better situation is in the case of 3-grams of symbols and 3-gram of words. We can say that 3-grams of words are very short sentences in many cases and they can give more information about clusters in some text.

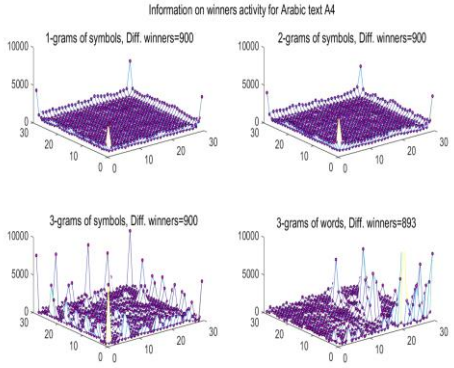

Fig.3 The system 4SOM$_B$. Evaluation of SOM winners for Arabic text A4.

B. **Clustering Combined with CNN method**

Convolutional neural networks are used in practice and realize good results specifically in the area of image processing as it is presented in [14]. But for a word processing there are models that have explored their use and achieve great results too. We have used some modification of a convolutional neural network for the sentence processing and advertisements classification in a condition that full advertisement text is one sentence created by words. We developed a similar network structure which was used for the processing of sentences in some texts as suggested by [12], and we tried to find parameters such that the network would well evaluate our data using the knowledge found by [13].

Our developed method works in two steps:

Algorithm CL-CNN:

Let $l(T_1)\ldots$, $l(T_r)$ be lengths of segments in the words. r be the number of segments in the text.

1. Clustering of segments using k-means Algorithm. Segments were mapped to vectors of dimension mean $\{l(T_i)\}$ Let k be the number of used clusters. The algorithm maps each segment into one class exactly. To choose some recommended clustering algorithm, for example k-means – algorithm, ART2 network have many clusters [2], ART2 offer some number of clusters.

2. CNN for classification of segments.

The training set is prepared from text segments together with the numbers of classes. CNN is trained on a set of segment and evaluated on the other set of segments. In CNN each segment looks like matrix of encoded words. We used back-propagation algorithm in the training procedure of

CNN, an error redistribution algorithm. It is the algorithm in which network errors are scrolled back across the layers so that the respective weights can be appropriately modified, and the network outputs are progressively improving. The second steps will give us results about a quality of previous clustering. We suppose that the accuracy of CNN for suspicious texts will be lower than for reliable texts.

3. Results and Evaluation

We will illustrate our developed method on 10 Arabic texts. The segments of texts were mapped into classes according to clusters. We worked with 4 and 5 clusters [1] using k-means - algorithm for clustering. Results of clustering to 4 and 5 clusters are described in [1]. The analysis using ART2 method to clustering gave us an recommendation to use 9 clusters [2]. Previous results about texts [5] are in the first column in Table 5, (Rel - Reliable text, Susp –Suspicious text). In each training, there were TR=85% of data as training data and TE=15% as testing data. In the second step of our method we used results of 9 clusters. We used the following parameters in the evaluation: (a) accuracy: Calculates how often predictions matches labels; (b) false negatives: Computes the total number of false negatives; (c) false positives: Sum the weights of false positives; (d) precision: Computes the precision of the predictions with respect to the labels.

Results on Arabic texts

In the Table 5, there are written results of 10 Arabic texts (3 of them are suspicious and 7 reliable according to previous methods - column 1;

the results are given by some criteria applied to 4 previous results on the texts). We used 9 clusters in both steps of the algorithm CL-CNN. The results of the accuracy and the precision values are written in Table 5.

Table5. Stastics of classification results for 10 Arabic texts, the number of clusters is 9, the number of training iterations 10000

Text of seg.	set	accu-racy	false Neg.	false Pos.	prec-sion
A1,Rel 69	TR:85% TE:15%	0.2906 0.1905	61.0 15.0	0.0 0.0	1.0 1.0
A2,Susp 248	TR:85% TE:15%	0.3243 0.1429	0.0 0.0	2.0 1.0	0.9429 0.8333
A3, Rel 137	TR:85% TE:15%	0.4354 0.5385	11.0 0.0	51.0 9.0	0.8198 0.8269
A4, Susp 136	TR:85% TE:15%	0.625 0.6364	58.0 9.0	0.0 0.0	1.0 1.0
A5, Rel 40	TR:85% TE:15%	0.2647 0.1667	0.0 0.0	2.0 2.0	0.9418 0.8333
A6, Rel 339	TR:85% TE:15%	0.2059 0.1667	0.0 0.0	4.0 0.0	0.8824 1.0
A7,Susp 325	TR:85% TE:15%	0.4429 0.4545	127.0 23.0	0.0 0.0	1.0 1.0
A8**,** Rel 338	TR:85% TE:15%	0.2857 0.3333	0.0 0.0	1.0 1.0	0.9642 0.8333
A9**,** Rel 305	TR:85% TE:15%	0.2619 0.125	31.0 7.0	0.0 0.0	0.0 0.0
A10,Rel 317	TR:85% TE:15%	0.2222 0.5481	0.0 0.4571	3.0 1.0	0.9333 0.8888

According to the results in Table 5, it is possible to evaluate the suspicious text A4 as reliable. The text A7 is still on the border, but the text , A2 is still suspicious Results of all Arabic texts in the accuracy parameter are in interval $\langle 0.125;0.6364 \rangle$ but the precision is above 0.8.

II. CONCLUSION

Arabic texts were analyzed from many statistical characteristics point of view. There were discovered some statistical differences between both languages. In the previous research, there were analyzed results of 40 Arabic texts from the database [4]. All results except results of the new developed method are described in [5]. The results from all methods trying to discover anomalies of text parts in each text show anomalies and they call for an attention to the text (or not) if the text parts were written by the same author (or not).

According to the results of all methods we could say that 38% of Arabic texts belong to the texts which are critical (suspicious), 61.5% belong to the texts which are reliable, and 5% belong to the texts which are very reliable. [5], [6].

According our new method we can give the definitive result on some reliable text. If the text was evaluated as suspicious according to previous methods, the new method according to good results (above 90%) in the precision parameter should move the text to the set of reliable texts. But if we have similar results (mainly, the text is suspicious) using more methods then we can recommend to do a new analysis of the text.

For Arabic text A4 we have got the good result, it means the text is not suspicious. The texts A2, A7 we will analyze again.

In our approach we will continue using different encoding of words, encoding n-gram of words. The idea of phrase clustering and new methods for clustering is interesting for our following research. But very important is the number of used clusters and the length of segments. We will continue in the research and we will analyze above named parameters.

ACKNOWLEDGEMENTS

We thank to Bc. Š . Horváth from P. J. Šafárik University in Košice for his help in a programming in Python.

REFERENCES

[1]. A. Salem, A. Almarimi and G. Andrejkov´a, Text Dissimilarities Predictions using Convolutional Neural Networks and Clustering, Proceedings of DISA 2018 - IEEE World Symposium on Digital Intelligence for Systems and Machines, Košice, August 23-25, 2018, p. 343-348, ISBN: 978-1-5386-5101-8, IEEE Catalog number: CFP18P13-US.

[2]. A. Salem, "Neural Networks in Arabic Text processing", Journal of Engineering Research and Application, Vol. 8, Issue 12 (Part -II) Dec 2018, pp. 05-11.

[3]. I. Bensalem, P. Rosso and S. Chikhi, " A new corpus for the evaluation of Arabic intrinsic plagiarism detection," CLEF. LNCS 8138 pp. 53–58, 2013.

[4]. King Saud University Corpus of Classical Arabic. http://ksucorpus.ksu.edu.sa.

[5]. A. Almarimi, "Dissimilarities Detections in Arabic and English Texts Using n-grams, Histograms and Self Organizing Maps". Pavol Jozef Šafárik University in Košice, PhD Thesis, 2016, 10, pp. 1–111.

[6]. A. Almarimi, G. Andrejkov´a and A. Salem, " Anomaly Searching in Text Sequences," CEUR, ISSN 1613-0073, Vol-2046 urn:nbn:de:0074-2046-8, Proceedings of the 11th Joint Conference on Mathematics and Computer Science Eger, Hungary, May 20-22, 2016.

[7]. E. Stamatatos, "Authorship attribution based on feature set subspacing ensembles," International Journal on Artificial Intelligence Tools, 15.5, pp. 823-838.

[8]. E. Stamatatos, " Ensemble-based author identification using character *n*-grams, "In Proceedings of the 3rd International Workshop on Text-based Information Retrieval, pp. 41-46, Riva del Garda, Italy.
[9]. H. J. Escalante, T. Solorio, M. Montes-y-Gomez, "Local Histograms of Character N -grams for Authorship Attribution." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp.288-298, Portland, Oregon, June 19-24, 2011. Association for Computational Linguistics.
[10]. T. Kohonen: Self Organizing Maps. Prentice-Hall, 2 ed, 2007.
[11]. A. K. Sangaiah, A. E. Fakhry, M. A. Basset, I. El-henawy, "Arabic text clustering using improved clustering algorithms with dimensionality reduction." Springer Science-Business Media, LLC, part of Springer Nature 2018, Cluster Computing, https://doi.org/10.1007/s10586-018-2084-4.
[12]. Y. Kim, " Convolutional neural networks for sentence classification,"arXiv:1408.5882. web-page: https://arxiv.org/abs/1408.5882, 2014.
[13]. Y. Zhang, ByronWallace, " A sensitivity analysis of convolutional neural networks for sentence classification," arXiv:1510.03820. 2015.
[14]. Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, L. Jackel, "Handwritten digit recognition with a back-propagation network." Advances in Neural Information Processing Systems, 2:396404, 2009.
[15]. Ch. C. Aggarwal, ChengXiang Zhai, "A survey of text clustering algorithms," Springer Science, Business Media, LLC 2012, DOI 10.1007/9781461432234 4.

Asmaa Salem" Anomalies in texts using Clustering and Convolutional Neural Networks" International Journal of Engineering Research and Applications (IJERA), Vol. 09, No.03, 2019, pp. 56-62