RESEARCH ARTICLE OPEN ACCESS

Neural Networks in Arabic Text processing

Asmaa Salem*
**(Department of Computer Science, Pavol Jozef Šafárik University in Košice,*

ABSTRACT
In this paper we aim to compare two developed methods to find text parts of long text which were not written by the same author or somebody manipulated with the text. Our work consists of two parts : in the first we developed a combined system of Replicator Neural Networks and ART2 to find outliers in some texts, in the second part we used Convolutional Neural Networks, the method was done by clustering of text parts and then classification of them. The analyzed texts were chosen from benchmark "King Saud University Corpus of Classical Arabic" of Arabic texts as long texts.

Keywords: Replicator Neural Networks, ART2NN, Convolutional Neural Networks, Clusters.

Date of Submission: 20-12-2018 Date of Acceptance: 04-01-2019

I. INTRODUCTION

Arabic language is a Semitic language spoken by more than 330 million people as a native language, in an area extending from the Arabian\ Persian Gulf in the East to the Atlantic Ocean in the west. Arabic is highly structured and derivational language where morphology plays a very important role. Arabic NLP applications must deal with several complex problems pertinent to the nature and structure of the Arabic language for example, Arabic is written from right to left. There is no capitalization in Arabic. Arabic letters change shape according to their position in the word [1].

Arabic text processing has become important in the area of natural language processing. Many problems that are connected with Arabic texts can be solved using different methods.

Many authors [1, 2, 3, 4] solve text-processing problem but solving of our special problem [5, 6, 7, 8] is not so known and it brings many possibilities to solve it using neural networks [9, 10, 11,12, 13, 14, 15]. Information on common languages is in [16].

We can describe our problem as to find outliers from the same texts. Our work is focused on long texts; we would like to do more analysis to get reasonable results.

We developed two methods, the first is combination of Replicator Neural Networks and ART2, the second method is Convolution Neural Networks to compare our results.

Arabic texts are from the corpus of classic Arabic texts built at King Saud University [17]. The text files are arranged into many folders representing the main types of the corpus. They were chosen from Religion Part 1, Religion Part 2. Information about types and groups of chosen texts is in Table 1.

The paper contains the following sections: In the second section, we present the results obtained using combined system of (ReNN and ART2 Neural Networks). The third section contains a description of our developed method which uses Convolutional Neural Networks (CNN). New results are written in the fourth section. The fifth section is conclusion.

II. COMBINED SYSTEM OF (REPLICATOR AND ART2)

We analyzed long texts and tried to find some parts of texts they have some anomalies and probably they were written by different author.

2.1 Replicator Neural Networks

Replicator neural networks (ReNN) are based on the feed-forward neural network (FFNN) models, and they are known in the literature as autoencoders [9]. They have a special numbers of neurons in layers. The number of neurons in the input layer is the same as in the output layer as in Fig.1.

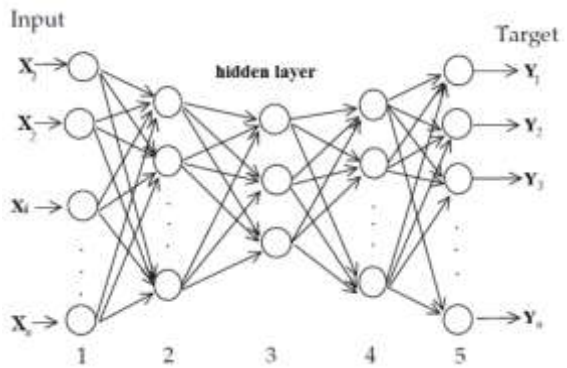

Fig 1. Replicator Neural Networks Architecture [18].

Let the number of layers in ReNN be k, k>2. The numbers of neurons in layers i+1 and

$$k-i,\ for 1 \le i \le (k/2)$$

are equal. If these numbers are decreasing as in Fig. 1, the network is called compression autoencoder, if the numbers are increasing the network is de-noising autoencoder. All neurons work in the classical way: some activation function is applied to the potential of neuron (the sum of its weighted inputs).

Training the ReNN: ReNN is a supervised network. The expected values on the output layer are the same as the values on input layer, it means the output of the network is compared to input values and modifications of network weights start according to the result. For the modifications of weights, methods of training FFNN can be used, for example Back-Propagation algorithm (BP) [10].

2.2 ART2 Neural Networks

The ART2 neural networks belong to the class of unsupervised and competitive learning algorithms. Adaptive resonance theory (ART) is a theory developed by Stephen Grossberg and Gail Carpenter on aspects of how the brain processes information [10, 19]. The structure of ART2 network is in Fig 2. The network is mainly used for building of clusters [18].

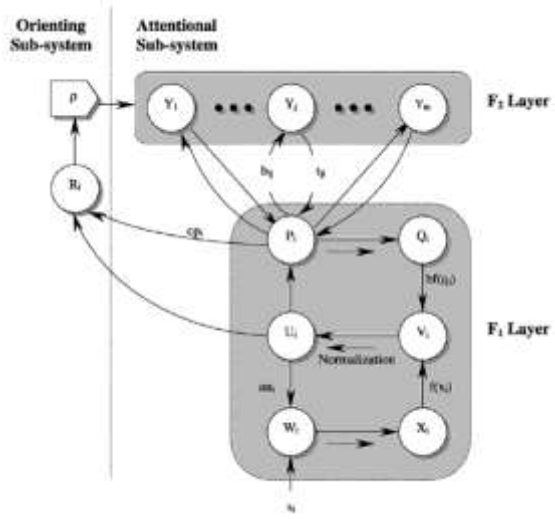

Fig2. ART2 Architecture [18].

The learning algorithm:

1. initialize the network and set all parameters
2. while input exists, do the steps 2.1. - 2.4.
2.1. read input
2.2. update activations in the input F1 layer
2.3. while not founded winner clustering node, do the following steps:
2.3.1. find a node in the output F2 layer with the highest activation
2.3.2. update activations of input layer for a comparison of conformity
2.3.3. if the vigilance is OK, the neuron is the winner
2.3.4. if the vigilance is not OK, set a reset signal to the neuron
2.3.5. if all nodes in the output layer have the reset signal set, create in the layer F2 a new neuron,which will be the winner neuron
2.4. modify the weights of the winner neuron

The update processes of weights in the steps (2.2.) and (2.3.2) are quite complex and they can be found in [11]. We used classical model with the length of input prepared according to the length of the average length of sentences in texts [18].

Table 1: Statistics of 3 analyzed original Arabic texts AR3, AR5, AR6 and three 3 modified Arabic texts AR1, AR2, AR4. The number of words by length for 1 – 8 columns, L-mx-sen means the maximal length of sentences in words.

3.2 Data Preprocessing

The texts are processed as the following steps:

1. remove titles of sections; the titles of sections, references and captions to figures are not normal sentences;
2. transform capital letters to small letters;
3. remove undesirable symbols' - ( ) `[ ] : ; ,
4. remove undesirable words - the, of, and, to, in, a, by, that, this, an, these; these words are not important for our analysis.
5. convert uppercase to lowercase; the meaning of the words will not change.
6. encode of the words; [18].

	Name of Texts					
	AR1	AR2	AR3	AR4	AR5	AR6
# words	1532	1385	2278	4307	4307	1468
#	21524	18065	33252	80819	8003	1913
Length words						
1	192	182	295	360	354	201
2	519	479	891	2070	2050	570
3	863	722	1311	2881	2858	783
4	861	718	1398	3322	3297	656
5	474	580	1107	2889	2857	470
6	212	517	867	1970	1952	211
7	129	184	361	938	930	121
8	32	94	167	642	458	15
L-mx-sen						
	195	186	296	830	824	200

4.2 Encoding of texts

In our experiments, we used the following method of encoding. By using frequencies of words divided by the total number of the words in the text; in this case different words have different code, but

words with the same frequencies have the same code. If the word S has its frequency Fs in the text T and $\max_{FS}$ is the maximal frequency of words in T then the code of the word S is

$$C_S = \frac{F_S}{\max Fs}, or \;\; C_S = \frac{F_S}{|V_t|}$$

The method gives codes with higher numbers but still from the interval (0,1) [18].

III. EVALUATION OF THE DEVELOPED METHOD

3.1 System of the processing

The process of outlier detection was done in the following steps:

1. ART2 neural network was used to classify all sentences in a text into classes. Similarly encoded sentences belong to the same class. It is supposed that they were written by the same author. The classes with the number of sentences higher than some threshold br (the number is an experimental parameter in our experiments) are supposed not outlier sentences. These all sentences belong to the set of good sentences Sg. The rest of the sentences are potential outliers and belong to the set of bad sentences S_{po}.
2. ReNN is used for analysis of sentences that belong to S_{po}. The training set S_{tr} is prepared from the set S_g. The validation set S_v is prepared from both set S_{po} and S_g,

$S_v \cap S_{tr} = \phi$ As a testing set we used the set S_{po} and some sentences from S_g [18].

3.2 The evaluation method

The main criteria for the evaluation of ReNN work is the error of computation given by

$$err_l = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - o_{ij}^{l})^2 ,$$

where m is the number of sentences in the training set (or validation, or testing set), n is the length of sentences * the average length of sentences in texts, 1 2), x_{ij} is the encoded input sentence, o_{ij}^{l} is output of ReNN in l-th iteration.

The threshold for the evaluation of outlier sentences is given by percentage of the validation error that is computed for the validation set after the training ReNN. We illustrate the results of the system processing for two different texts in both languages, two of analyzed texts (one from each language) are combined from two different texts, and a position of combinations or positions of inserted sentences are known too.

Let |T| be the number of sentences in the text T. the set of all sentences is input to the ART2 network. ART2 creates two subsets of sentences: (1) a set of A of good sentences in strong category and (2) in a set B with all remaining bad sentences. The training set for ReNN network is created from sentences in A, it is used 60% sentences. 30% belongs to a validation set and the rest 10% belongs to the testing set together with all sentences from B [18].

3.3 Analyzed texts.

Information on six types of used texts from [17] is in Table1. The Arabic texts AR3, AR5, AR6 are original texts from, and other texts AR1, AR2, AR4 are combined texts from three different texts. Arabic text AR1 is contains some inserted sentences from the other text in the positions 6, 7, 8, 10, 11, 179, 182, 185, 187.

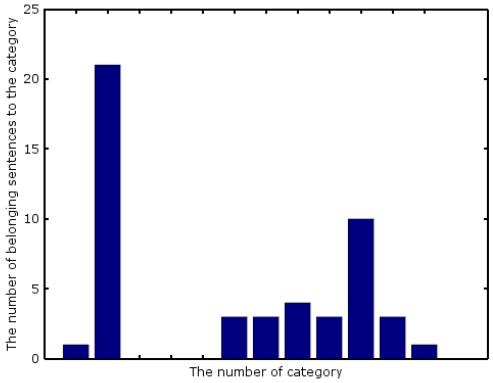

Fig 1. The number of categories in the text AR1 with their capacity. The category with one or two sentences is not visible

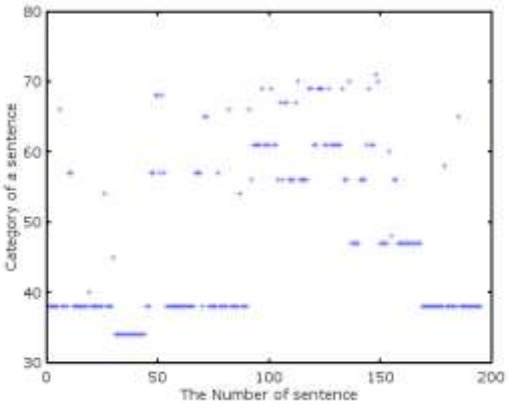

Fig 2. The analysis of the text AR1 according to ART2 network. In the graph, there are plotted the numbers of sentences in categories.

Table 2: Evaluation of the combined text AR1 for the testing set with 63 testing sentences. We show only the results for inserted sentences and some other types of results for the characterization. The value of the used validation error is 0.0126. in

the testing set, the thresholds for classification were 20%, 40% and 60%.

ART2	ReNN			% of Results AR1 text
	20%	40%	60%	
0	0	0	0	**53.96**
0	0	0	1	**12.69**
0	0	1	1	**0**
0	1	1	1	7.93
1	0	0	0	1.58
1	0	0	1	11.11
1	0	1	1	0
1	1	1	1	7.93

Table 3: Evaluation of the combined text AR1 for the testing set with 63 testing sentences. In the table we show only the results for inserted sentences and some other types of results for inserted sentences and some other types of results for the characterization. The bold numbers are the numbers of inserted sentences-real outliers in the text.

# Sent.	ART2 error	ReNN error	Evaluation to val. error 20%	40%	60%
6	0	0.030889	0	0	0
7	1	0.002801	1	1	1
8	1	0.032288	1	1	1
10	0	0.039091	0	0	1
11	0	0.060470	1	1	1
26	0	0.022980	0	0	0
30	0	0.019725	0	0	0
47	0	0.064649	0	0	0
48	0	0.041096	0	0	0
49	0	0.091132	0	0	0
50	0	0.043856	0	0	1
51	0	0.002784	1	1	1
52	0	0.014783	0	0	0
53	0	0.031588	0	0	1
54	1	0.013632	1	1	1
62	1	0.019190	0	0	1
97	0	0.001912	0	0	0
118	0	0.009656	0	0	0
120	1	0.018787	0	0	0
122	0	0.011924	0	0	1
129	0	0.019418	0	0	1
145	0	0.005493	0	0	0
154	0	0.027166	0	0	0
156	1	0.011867	0	0	0
159	1	0.019288	1	1	1
161	1	0.001493	0	0	0
167	1	0.020125	0	0	1
179	0	0.029480	0	0	0
182	1	0.009446	0	0	1
185	0	0.014312	0	0	0
187	1	0.032151	0	0	1

Results for the text AR1 described in Table 3 and characterized by graphs in Fig. 3 and Fig. 4. The graphs illustrate the clusters (categories) computed by ART2 neural network. In the clusters, there are concentrated the sentences with the similar codes.

According to results partially given in the Table 3, we have the analysis of the combined text AR1. The size of the testing set was 63 sentences. The set was prepared in random sequence of 10 % of sentences from the set A and the rest sentences from the set B. in Table 2, we can see a percentage evaluation of all sentences using both networks. The recognition of a sentence as a good sentence is written by 1 and the recognition of sentence as a bad sentence is written by 0. In the recognition of ReNN were used 3 thresholds for the recognition (20%, 40%, 60% of the validation error).

The bad sentences are potential outliers; it means that their analysis is more important. 53.96 % of sentences were classified as **bad** by ART2 and bad by ReNN as well. 7.93% sentences, ReNN recognized as **good** sentences. The rest 20.62% sentences ART2 classified as **good** but 1.58% of them ReNN classified as **bad** sentences.

3. Clustering Combined with CNN method

Clustering is one of the most popular data mining algorithms and it is extensively studied in the text context. It has many applications for example in a classification of short texts as advertisement. The clustering is the task to find groups of similar texts in a collection of texts. The similarity is measured by using a similarity function. Text clustering can be in different levels of granularity where clusters can be documents, paragraphs, sentences or terms.

Our approach is oriented to long texts. Long text can be split into paragraphs, we call them segments. We analyze this similarity and clustering of segments.

Convolutional neural networks are used in practice and realize good results specifically in the area of image processing as it is presented in [20]. But for word processing there are models that have explored their use and achieve great results. We have used some modification of a convolutional neural network for the sentence processing and the advertisement 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 [21], and we tried to find parameters such that the network would well evaluate our data using the knowledge found by [12]. Our developed method works in two steps.

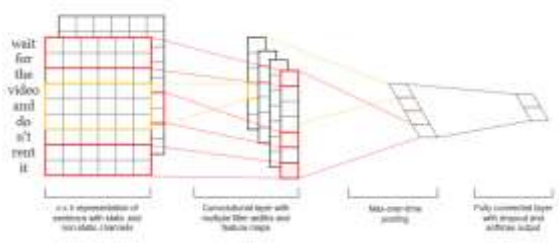

Fig 3. Convolutional Neural Network Architecture [9].

Algorithm CL-CNN:
Let l(s)…, l(s) be lengths of segments in the words. ns 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(s)} we worked with 4 and 5 clusters. The algorithm maps each segment into one class.

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 [11].

IV. RESULTS AND EVALUATION

We will illustrate our method on 3original Arabic (AR2, AR3, AR5) texts and 3 combined Arabic texts (AR1, AR4, and AR6). The segments of texts were mapped into classes according to clusters. We worked with 4 and 5 clusters. In the second step of our method we used results for 4 and 5 clusters. We used the following parameters in the evaluation:
(a) Accuracy: Calculates how often predictions match 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.

The results of the segment classification for 3 original Arabic texts AR3, AR5, AR6 and 3 combined Arabic texts AR1, AR2, AR4 are in Table 4, 5 and 6.

Table 4: Clusters of segments for 3 original Arabic texts AR3, AR5, AR6 and 3 combined Arabic texts AR1, AR2, AR4. The numbers in the columns 2 and 3 inform about numbers of sentences in clusters.
"Bad" in the columns 4 and 5 means that in the previous analysis [10, 22] the text was classified as text with outliers and it is important to analyze it again.

Text	5clusters	4clusters	Previous results ART2\ ReNN	
AR1 488 seg	1:251, 2:68, 3:45, 4:26, 5:24	1:249, 2:71, 3:48, 4:46	Bad 74.58%	Bad 79.34%
AR2 340 seg	1:177, 2:45, 3:30, 4:27, 5:10	1:175, 2:52, 3:49, 4:13	Bad 92.16%	Bad 53.03%
AR3 560 seg	1:227, 2:88, 3:63, 4:59, 5:39	1:235, 2:145, 3:57, 4:39	Bad 67.51%	Bad 84.38%
AR4 850 seg	1:447, 2:130, 3:119, 4:93, 5:61	1:448, 2:149, 3:140, 4:113	Bad 50.82%	Bad 47.88%
AR5 850 seg	1:517, 2:106, 3:103, 4:88, 5:36	1:526, 2:117, 3:114, 4:39	Bad 60.08%	Bad 71.99%
AR6 360 seg	1:161, 2:45, 3:39, 4:31, 5:30	1:159, 2:73, 3:41, 4:33	Bad 52.62%	Bad 47.36%

Information about 6 analyzed texts in the Table 4. We analyzed 3 original texts and 3 combined texts using 4 and 5 clusters in the first step of the algorithm, results of texts in the accuracy are in interval (0.6071; 0.7894) using 4 clusters and (0.6078 ; 0.7471) using 5 clusters.

Table 5: Statistics of classification results for three analyzed original Arabic texts AR3, AR5, AR6 and three combined Arabic texts AR1, AR2, AR4, the number of clusters is 4, the number of training iterations 1000.

Text	set	accu-racy	false Neg.	false Pos.	prec-sion
AR1	TR:85% TE:15%	0.6891 0.6884	14.0 82.0	0.0 0.0	1.0 1.0
AR2	TR:85% TE:15%	0.6862 0.7059	12.0 55.0	0.0 0.0	1.0 1.0
AR3	TR:85% TE:15%	0.6071 0.6492	24.0 118.0	0.0 0.0	1.0 1.0
AR4	TR:85% TE:15%	0.6133 0.6212	3.0 10.0	22.0 140.0	0.8503 0.8333
AR5	TR:85% TE:15%	0.7067 0.7894	5.0 19.0	23.0 80.0	0.8175 0.8919
AR6	TR:85% TE:15%	0.6667 0.6339	13.0 74.0	0.0 0.0	1.0 1.0

Table 6: Statistics of classification results for 3 original Arabic texts AR3, AR5, AR6 and 3 combined Arabic texts AR1, AR2, AR4, the number of clusters is 5, the number of training iterations

10000.

Text	set	accu-racy	false Neg.	false Pos.	prec-sion
AR1	TR:85%	0.6486	0.0	6.0	0.9189
	TE:15%	0.6932	0.0	26.0	0.9371
AR2	TR:85%	0.6078	13.0	0.0	1.0
	TE:15%	0.6989	54.0	0.0	1.0
AR3	TR:85%	0.6905	16.0	0.0	1.0
	TE:15%	0.6345	126.0	0.0	1.0
AR4	TR:85%	0.74	27.0	0.0	1.0
	TE:15%	0.7471	168.0	1.0	0.9939
AR5	TR:85%	0.64	0.0	19.0	0.8707
	TE:15%	0.6094	8.0	103.0	0.8736
AR6	TR:85%	0.5741	19.0	0.0	1.0
	TE:15%	0.6536	68.0	0.0	1.0

According to the results on the both sets (training and testing), we can say they are closed to each other. That means for Bad texts it is not possible to evaluate them as Bad texts.

V. CONCLUSION

In the paper we developed two methods: the first method is the system with two types of neural networks for detection of outliers sentences in some text, the second method for classification of them. According to the second method, we have got different results on good texts, if the texts were tested as Bad on the first method. The second method according to good results (97%) in the accuracy belongs to the set of good texts. We need to do more analysis if we comparable results especially if the texts are tested as Bad texts. For original texts AR3, AR5, AR6 we have got the results that the texts are not Bad texts. As future work, we will continue using different encoding of words and put more information to codes, and we will do more analyze on long texts.

ACKNOWLEDGEMENTS

The Slovak Scientific Grant Agency VEGA, Grant No. 1/0056/18, supports the research. I thank to my supervisor Assoc. prof. G. Andrejkova and Bc. P. Kozak for help in writing of the paper.

REFERENCES

[1]. A. Farghaly, K. Shaalan, Arabic Natural Language Processing: Challenges and Solutions, ACM Transactions on Asian Language Information Processing, Vol. 8, No. 4, Article 14, Pub. December 2009.

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

[3]. D. Jurafsky, and J. H. Martin, Speech and Language Processing. Prentice-Hall, 1. Vyd, 2000.

[4]. E. Stamatatos, A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol, 2010, 538–556.

[5]. A. Salem, Convolutional Neural Networks in Text Processing, Spring School of PhD students, Local Conference in Liptovský Ján, 11 – 14. 6. 2018, P. J. Safarik university in Kosice, Faculty of Science.

[6]. A. Salem, A. Almarimi and G. Andrejkova, Text Dissimilarities Predictions using Convolutional Neural Networks and Clustering, 1st World Symposium on Digital Intelligence for Systems and Machines August 23-25, 2018, International Conference in Technical University of Košice, Slovakia, ISBN: 978-1- 5386-5101-8, p. 343-348.

[7]. A. Salem and G. Andrejková, Analysis of text advertisements using convolutional neural networks, In proceedings of Cognition and artificial life, Brno, 30. 5. - 1. 6. 2018, ISBN 978-80-88123-24- 8, 59-61.

[8]. A. Almarimi, and G. Andrejkova, Discrepancies detection in Arabic and English documents. ACSIJ Advances in Computer Science: an International Journal, 4 (2322-5157), 2015, 69–75.

[9]. H. A. Dau, V. Cieselski, and A. Song, Anomaly detection using replicator neural networks trained on examples of one class. In Proceedings of the SEAL, LNCS 8886, 2014, p. 311–322.

[10]. J. Hertz, A. Krogh, and R. G. Palmer, Introduction to the Theory of Neural Computation. Addison-Wesley, 1991.

[11]. G. A. Carpenter, and S. Grossberg, Adaptive resonance theory. Encyclopedia of Machine Learning and Data Mining, 2014, p. 22–35.

[12]. Y. Zhang, and Byron Wallace, A sensitivity analysis of convolutional neural networks for sentence classification, arXiv:1510.03820. http://arxiv.org/abs/1510.03820, 2015.

[13]. G. Barreto, and L. Aguayo, Time series clus-tering for anomaly detection: Using competitive neural networks. Proceedings WSOM, LNCS (5629), 2009, 28–36.

[14]. ReNN (2016). Replicator NN, https:// blog.acolyer.org/2016/06/23/ai2-training-a-big-data-machine-todefend/.

[15]. R. E. R. Christ, T. Basani, J. C. Nievola, and C. N. Silla, The use of ART2 to create summaries from texts of different areas. https://www.researchgate.net/publication/242514338, 2005, 1–6.

[16]. N. Chomsky, Three factors in language design. (Lingistic Inquiry, 2005, 36:1–22).

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

[18]. A. Salem, G. Andrejkova and P. Kozak, Replicator and ART2 neural networks in text outliers identifications, In Proceedings of cognition and artificial life 2017, International conference in Trenčianske Teplice, ISBN: 978-80-223-4346-6, p. 130-135.
[19]. ART2(2015).https://www.researchgate.net/figure/222517262 fig3 fig-3-typical-architecture-of art2- neural-network.
[20]. Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, and L. Jackel, Handwritten digit recognition with a back-propagation network. Advances in Neural Information Processing Systems, 2:396404, 2009.
[21]. Y. Kim, Convolutional neural networks for sentence classification, arXiv: 1408.5882. web-page:https://arxiv.org/abs/1408.5882, 2014.
[22]. D. M. Hawkins, Identification of Outliers. Chapmann & Hall, 1980.

Asmaa Salem" Neural Networks in Arabic Text processing"International Journal of Engineering Research and Applications (IJERA) , vol. 8, no.12, 2018, pp 05-11