Detection of computer-generated papers using One-Class SVM and cluster approaches

Conference Paper · July 2018

CITATIONS
2

READS
133

2 authors:

Zeev Volkovich
ORT Braude College
153 PUBLICATIONS **522** CITATIONS
SEE PROFILE

Renata Avros
ORT Braude College
23 PUBLICATIONS **58** CITATIONS
SEE PROFILE

Detection of computer-generated papers using One-Class SVM and cluster approaches

Renata Avros[1] and Zeev Volkovich[1]

[1] Department of Software Engineering
ORT Braude College
Karmiel, 21982, Israel
ravros@braude.ac.il, vlvolkov@braude.ac.il

Abstract. The paper presents a novel methodology intended to distinguish between real and artificially generated manuscripts. The approach employs inherent differences between the human and artificially generated wring styles. Taking into account the nature of the generation process, we suggest that the human style is essentially more “diverse” and "rich" in comparison with an artificial one. In order to assess dissimilarities between fake and real papers, a distance between writing styles is evaluated via the dynamic dissimilarity methodology. From this standpoint, the generated papers are much similar in their own style and significantly differ from the human written documents. A set of fake documents is captured as the training data so that a real document is expected to appear as an outlier in relation to this collection. Thus, we analyze the proposed task in the context of the one-class classification using a one-class SVM approach compared with a clustering base procedure. The provided numerical experiments demonstrate very high ability of the proposed methodology to recognize artificially generated papers.

Keywords: Artificial Papers Detection, Text Mining, SVM.

1 Introduction

Latterly, various fake senseless scientific papers written by computer programs were over and over again accepted for publication in various scientific conferences and journals. At first glance, such documents being completely meaningless look like human composed manuscripts, since they effectively imitate standard scientific papers by their structure, grammar, and other attributes.

Producing of fake documents has, evidently, begun after the establishment of an open artificial papers generator “SCIGen” innovated in 2005 by three graduate students of the Massachusetts Institute of Technology - Jeremy Stribling, Maxwell Krohn, and Dan Aguayo. The “SCIGen” software is intended to generate synthetic manuscripts in the computer science discipline. Later, other scientific document generators such as “SCIgen-Physic” focusing on physics, “Mathgen” concentrating on math and the “Automatic SBIR” (Small Business Innovation Research) Proposal Generator dealing with grant proposal fabrication were suggested.

It is indeed impressive that computer programs are now good enough to create a “tolerable gibberish”, but the computerized recognition of these works by some authoritative journals has taken on a wide scale and has become a serious problem. In order to detect fake scientific papers, many approaches have been proposed. The fact that counterfeit articles created by “SCIgen” or by other mentioned generators are always structured in a similar way suggests that a detection would be possible.

For example, in order to detect fake generated papers, the authors in [1] suggested measuring the keywords occurrences in the title, abstract and paper body, which is expected to be "uniform" in real papers. Moreover, it is natural to assume that the mentioned keywords are expected to appear frequently in real documents; yet rarely in fake articles.

An inter-textual similarity was used in [2] by counting the differences in word frequencies of two texts aiming to differ real and fake documents. This idea was extended in [3] by counting occurrences of word combinations in phrases. The list of references and keywords of the cited articles can serve as additional criteria to recognize artificial papers. As shown in [4], the various reference in a forged paper cannot be found on the Internet.

Another approach [5] is based on analyzing the texts' compression profiles suggests that there is a meaningful disagreement in the compression ratio between human-written and computer-generated texts.

As was found in [6], there is a significant dissimilarity in the topological properties of the natural and the generated texts. Various methods to identify synthetic scientific papers were discussed in [7] and [8].

In this paper, we present a novel methodology intended to distinguish between real and artificially generated manuscripts. This approach relies on inherent differences between the human and artificially generated wring styles. Namely, taking into account the nature of the generation process, we can suggest that the human style is essentially more “diverse” and "rich" in comparison to an artificial one. In fact, a counterfeit paper is formed by randomizing existing prechosen components. In consequence, the separate parts of a fake document are almost unrelated one to another, although they are obviously associated within a real text. Thus, the generated papers are much similar in their own style and significantly differ from the human written documents, which can be varied among themselves in their own style, but more significantly differ from fake texts.

In order to assess dissimilarities between fake and real papers, a distance between writing styles is evaluated in this paper using the methodology discussed in [9]-[13]. A set of documents generated by SCIGen is captured as the training data so that a real document is expected to appear as an outlier in relation to this collection. Therefore, we analyze the proposed task in the context of the one-class classification.

One class classification can be considered as a special case of two-class classification problem in the situations where only data of one class is accessible and well described, while the data from other class is hard or impossible to obtain. The usage of this methodology is reasonable in our approach since we suppose that the artificially created papers are grouped into a target class sharing the common characteristics,

while the real manuscripts "fall out" of it. A real document, seemingly not belonging to the target class, is expected to be labeled as "outlier" representing the second class.

In our methodology, all texts under consideration are divided into chunks and the distance values between them are calculated according to the replica introduced in [9]-[13] intending to quantify the difference in the documents' writing style. Afterwards, a One-class Support Vector Machine aiming to allocate where the chunks of the tested manuscript are placed, is constructed. If the majority of them are located outside of the training class then it is concluded that the document is real. Otherwise, the text is recognized as a fake one.

An additional approach presented in this paper is based on as a partition of all attained chunks into clusters provided by means of the mentioned earlier distance using the Partition Around Medoids (PAM) algorithm. A manuscript is recognized as a real or a fake one according to the majority voting of its own chunks. As we will see, applying this distance makes it possible to detect almost surely the fake documents by means of the two proposed methods.

This paper is organized as follows. In Sec. 1, we present the introduction with related approaches aiming at the identification of fake scientific manuscripts. The background is presented in Sec. 2. The methods employed for the characterization and classification of texts are presented in Sec. 3. The numerical experiments and obtained results are presented in Sec. 4, and the conclusion is presented in Sec. 5.

2 Preliminary

2.1 Dynamic dissimilarity between texts' styles

To pattern the human writing process the following model was proposed in [9]. Let us take a group of texts Δ, where each document $D \in \Delta$ is considered as a sequence of its own chunks

$$D = \{D_1, ..., D_m\}$$

of the same size L. In order to evaluate the development of a text within the writing process, the Mean Dependency characterizing the mean relationship between a chunk D_i, $i = T+1, ..., m$ and the set of its T "precursors" is outlined:

$$ZV_T(D_i, \Delta_{i,T}) = \frac{1}{T} \sum_{D_j \in \Delta_{i,T}} s(D_i, D_j),\ i=T+1,...,m, \tag{1}$$

where $\Delta_i = \{D_{i-j},\ j = 1, ..., T\}$ is the set of T "precursors" of D_i, and s is a similarity measure. In this model, s is constructed using the common Vector Space Model. Every chunk is expressed as a terms' frequency vector of terms taken from a given dictionary $\mathcal{D}$, and s is calculated as the Spearman's ρ (see, e.g. [14]):

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i}{n(n^2 - 1)},$$

where d_i, $i=1,..n$ are the differences between the corresponding ranks in the scales, and n is the dictionary size. ρ handles the frequency tables as a kind of ordinal data related to the frequencies rank placing. Rank ties take a rank equal to the mean of their position in the ascending order of the values. The Spearman's ρ is actually the famous Pearson's correlation amid the ranks.

Resting upon the ZV_T measure, a distance (dynamic dissimilarity) between documents' styles is produced as:

$$DZV_T(D_i, D_j) = abs\left(ZV_T\left(D_i, \Delta_{i,T}\right) + ZV\left(D_j, \Delta_{j,T}\right) - ZV\left(D_i, \Delta_{j,T}\right) - ZV\left(D_j, \Delta_{i,T}\right)\right).$$

Subsequently, the association of a text with the "precursors" of another text is derived from the association with its own neighbors. It is easy to see that it is merely a semi-metric. Even though, if the documents are almost consistently associated with their own "precursors" and the "precursors" of another document, then they are practically indistinguishable from the writing style standpoint.

An example presented in Fig. 1 is a 3D scatter plot of three main components of DZV_T computed for a collection of two books "2010: Odyssey Two" by A. C. Clarke and "Harry Potter and the Philosopher's Stone" by J. K. Rowling.

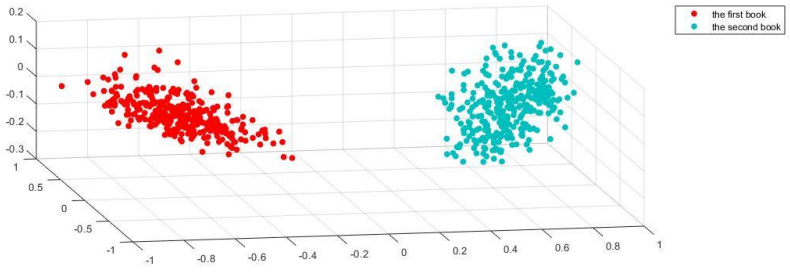

Fig. 1. Scatterplot plot of three main components of DZV_T of a two books' collection.

Two well separated bunches composed of the corresponding chunks unambiguously reflect a significant difference in the styles. This illustration demonstrates the ability of the proposed distance to distinguish the styles of distinct authors.

2.2 One-Class SVM

The One Class SVM (One Class Support Vector Machine) methodology was actually proposed in [15] as adjustment of the classical SVM (Support Vector Machine) methodology to the one-class classification problem. Here, after converting the data using a kernel transform just the origin is treated as the member of the second class,

and the usual two-class SVM methodology is applied. Formally, the problem is described as follows.

Let us consider a training sample $\{x_1, x_2,..., x_k\}$ from a group X , where X is a compact subset of a Euclidean space R^n , together with a kernel map $\Phi : R^n \to H$. Feature space H is implicit (and usually unknown) in all kernel methods. Aiming to isolate the data from the origin, the subsequent quadratic programming problem is solved:

$$\min_{w\in H,\xi_i\in R^n,\gamma\in R} \frac{1}{2}\|w\|^2 + \frac{1}{k\eta}\sum_{i=1}^{k}\xi_i - \gamma$$

subject to

$$(w\cdot\Phi(x_i)) \geq \gamma - \xi_i, \quad i = 1,2,...,k, \quad \xi i \geq 0 .$$

Here

- ξ_i - slack variables (one for each point in the sample);
- γ - the distance to the origin in feature space;
- w - the parametrization of the hyperplane separating the origin from the data in H;
- η - the expected fraction of data points outside the estimated one class support.

If w and ρ are a solution, then the decision rule

$$f(x) = sign((w\cdot\Phi(x)) - \gamma)$$

takes positive values for most points located in the training set. The dual form of the problem is

$$\min_{\alpha_i\in R}\frac{1}{2}\sum_{i,j=1}^{k}\alpha_i\alpha_j K(x_i, x_j)$$

subject to

- $0 \leq \alpha_i \leq \frac{1}{k\eta}$;
- $\sum_{i=1}^{k}\alpha_i = 1.$

where $K(x_i, x_j)$ is a kernel function defined as $\Phi(x_i)\bullet\Phi(x_j)$. One obtains the decision function

$$f(x) = \sum_{i=1}^{k}\alpha_i K(x_i, x) - \gamma.$$

Analogously to the discussed earlier primal form, a negative value of this function reveals outliers. The data points satisfying

$$0 < \alpha_i < \frac{1}{k\eta}$$

are the support vectors located exactly on the separating hyperplane.

3 Approach

We base our consideration on a common view on the human writing process (see, e.g. [16]) proposing that this development consists of four main elements: planning, composing, editing and composing the final draft. Such an assumption leads to the evident conjecture about the natural connection of the sequentially written parts of a human created manuscript. On the other hand, SCIgen based on context-free grammar methodology randomly produces nonsense texts in the form of computer science research papers including graphs, diagrams, and citations. Thus, these artificial texts do not pass the mentioned steps of the human writing process. It is very natural to expect that this contrived style is essentially different from one inherent to people. However, all generated papers are supposed to be very close in their writing style. The proposed model is created resting upon these matters in the framework of the one-class classification methodology.

3.1 One-Class SVM classifier

Applying a One-Class SVM approach, the target class is constructed using the artificially generated papers, whereas outliers are associated with the humanly written papers. In this fashion, a collection D_0 of artificially papers is generated, and a dictionary $\mathcal{D}$, the delay parameter T, and size of the chunks L are selected planning to apply the Dynamic Similarity model presented in Section 2.1.
Afterwards, a tested document is divided into chunks, and each chunk is transformed using the Vector Space Model in an occurrences vector. These vectors together with the target class vectors form a collection under consideration. A quadratic matrix of the DZV_T pairwise distances between all chunks in this expanded group having at least T "precursors" is calculated:

$$Dis = \left\{ DZV_T\left(D_{i_1}^{(m_1)}, D_{i_2}^{(m_2)}\right) \right\},$$

where $m_1, m_2 = 1,.., M+1$. i_1 and i_2 are the sequential indexes of the documents' chunks taking values from 1 to n_1-T and n_2 –T, correspondently. n_1 and n_2 are numbers of chunks in the documents. We create the target class from the rows of this matrix using apparently an embedding of the chunks set into a Euclidian space with the dimensionality:

$$Dim = \sum_{i=1}^{M+1} n_i - (M+1)T.$$

This embedding procedure increases the resolution of the method so that each piece is associated with a vector possessing coordinates corresponding to its distances to all other chunks.
A one-class SVM classifier is constructed on the matrix *Dis*. An example of such a matrix is given in Fig. 2.

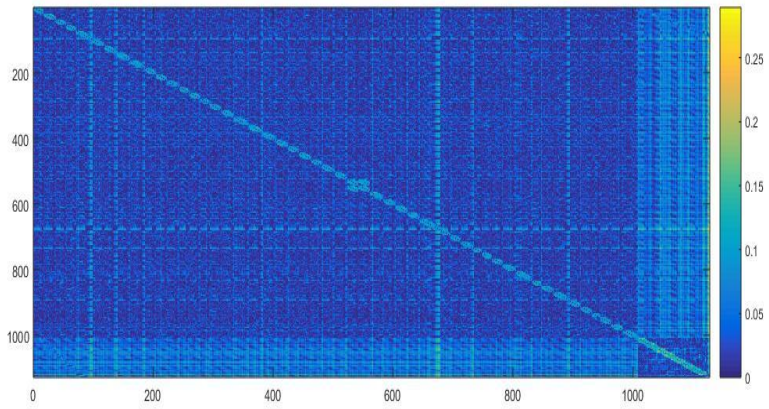

Fig. 2. An example *Dis* matrix.

The brighter areas indicate the distances between the tested texts and the elements of the target class. They are obviously larger than the corresponding ones inside distances. Therefore, it is natural to expect that the document in question is not fabricated.
A conclusion according to the total allocation of the tested document is made in our methodology using the following rules:

- **Majority voting of the chunks (SVM Rule 1)**. The document is assigned to the target class (a fake paper case) or recognized as an outlier (a real paper case) if more than half of the chunks belong to the consistent class.
- ***P*-value rule (SVM Rule 2)**. The average score of the decision function is calculated for each paper from the target class and for the tested paper. After, the sample *p*-value of the mean corresponding to the text in question is compared with the fraction of its chunks in the total their number. The article is accepted as a real one if the found *p*-value is greater of the fraction.

3.2 Clustering based classifier

Clustering is as a dependable unsupervised process for discovering prototypes in unlabeled data. The key benefit of a general clustering procedure is the proficiency to uncover intrusions in the checked data without any prior information. In our ap-

proach, a type of clustering is employed to reveal differences between an artificial and the human writing styles.
Our perspective suggests that the chunks of a tested real paper could produce a group located sufficiently far from the target class, i.e. to be a separate cluster. In this manner, a meaningful partition of the analyzed collection into two clusters indicates a difference in the styles and approve the authenticity of the verified paper.
The procedure is implemented in the following way. The matrix $\boldsymbol{Dis}$ of the pairwise DZV distances is constructed by the described in the previous subsection technique for chunks of the target class papers and chunks of a paper being examined.
In the next step, we partition the rows of this matrix into two clusters aiming to turn out a cluster corresponding to the artificial articles and a cluster corresponding to the tested manuscript. Each document is apportioned to a cluster (i.e. style) with the highest winning rate of the matching chunks, and the conclusion is made consistent with this cluster assignment. Therefore, if the majority of the target class texts are located in a group different from one containing the majority of the tested text's chunks then one concludes that this article is not false. Otherwise, the document is recognized as an artificial one.

4 Experiments

In order to evaluate the ability of our proposed methodology, we provide several numerical experiments performed in the MATLAB environment through the chunk sizes L = 200 and 400 with the delay parameter T value equal to 10.

4.1 Experiments setup

4.1.1 Material

One hundred artificial papers are constructed by means of the SCIGen procedure and fifty genuine manuscripts are drawn from the "arXiv" repository [17]. Fifty artificial papers form the target class, and the residual ones are tested together with the drawn real manuscripts. The results are represented via the True Positive Rate (TPR, otherwise known as Sensitivity or Recall), which evaluates the part of papers (false or real) that are rightly recognized as such.

4.1.2 Dictionary construction

Two different types of the dictionary of terms $\mathcal{D}$ are used.

- **The *N*-grams model**. Here, within a preprocessing procedure, all uppercase letters are transformed to the corresponding lowercase letters, and all other characters are omitted. The vocabulary contains all *N*-grams appearing in the trading class documents. Recall that an *N*-gram is an adjoining *N*-character piece formed by the characters occurring in a sliding window of length *N*. We use in our experiments no more 10000 of the most common *N*-grams with lengths of three, four and five.
- **The content-free word model**. Content-free words do not express any semantic meaning on their own. This kind of terms can be associated with a stylistic "glue" of the language appearing to set up the connection between all terms that. Joint occurrences of the content-free words offer a stylistic indication for authorship verification [18], [19]. This approach was essentially applied in a study of quantitative patterns of stylistic influence [20] and modeling of the writing style evolution [13].

4.1.3 SVM parameters

Discovering an appropriate data representation from their primary attributes is an essential part of any classification problem. As a rule, not each characteristic is constructive for classification and can even harm the result. With the intention, just the necessary information has to be retained for further study. The following figure represents an example of a scatterplot of the two-dimensional principal subspace found in the matrix ***DIS***.

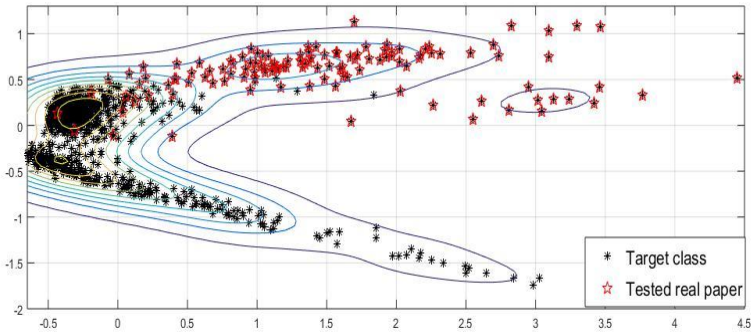

Fig. 33. Scatter plot of the two leading components of ***DIS***.

The two leading components explain 64.5% of the total variation in the data. The plot shows that a linear separation between the target class and its outliers is hardly ex-

pected. Thus, in our experiments the Gaussian Radial Basis Function kernel (the RBF kernel)

$$K(x_i, x_j) = exp\left(-\frac{\left(x_i - x_j\right)^2}{2\sigma^2}\right)$$

is employed. The standard MATLAB 2017b SVM toolbox is applied. All columns of the data are standardized. The width parameter σ is estimated using the inbuilt heuristic procedure.

The threshold η corresponds to the expected fraction outliers among the artificially generated papers. Picking a suitable level for η is very important for constructing an effective one-class SVM model. Small values can enlarge the false positive rate (the ratio between the number of artificial papers mistakenly documented as real ones and the total number of actual fabricated texts). On the other hand, a large η can obviously increase the false negative rate, calculated as the fraction of real papers recognized as fabricated.

To estimate its value we chose 20 synthetic paper and 20 real papers and run the parameter η starting at 0.01 and gradually increasing the value with an increment of 0.01 until 0.3. Each selection of η provides a pair of the false positive rate and the false negative rate. An appropriate value of η is chosen a balance point between these characteristics likewise to the ROC methodology. In our case, it is 0.15.

4.2 Results

4.2.1 N-grams model

Dealing with an *N*-gram based Vector Space Model we can naturally to expect that for sufficiently small lengths the chunks the obtained vector representation tends to be more and more independent of the text, de facto random. To illustrate it, let us consider histograms of the ZV_T values constructed for *N=3*, *4* and *5* and *L=200* within the training class. The mean values are 0.15, 0.06 and 0.04 correspondently.

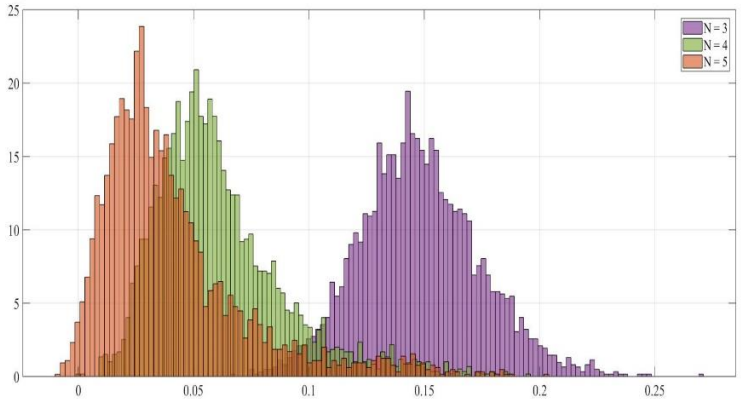

Fig. 4. Histograms of the ZV_T values within the training class calculated for *N=3*, *4* and *5* and *L=200*.

The histograms lean to concentrate around the origin. Therefore, the association between the sequential parts text vanishes. A similar situation appears once real papers are considered.

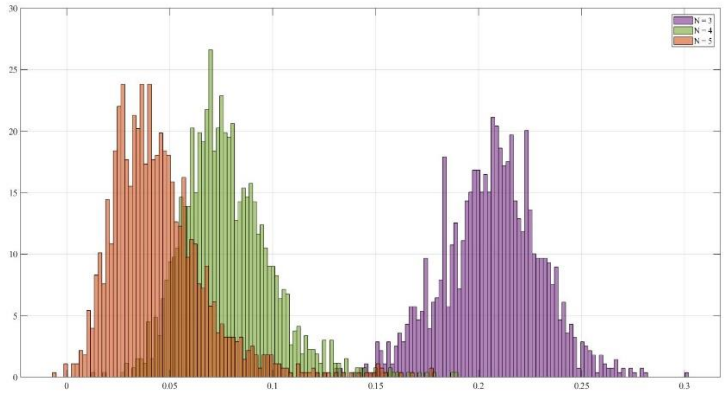

Fig. 5. Histograms of the ZV_T values within 50 real papers calculated for *N=3*, *4* and *5* and *L=200*.

However, the histograms of real papers seek much slower. These observations make possible to suggest that the methods are expected sufficiently well to recognize the fake papers even for small values of *L*, but can fail with the real articles. The following tables approve this suggestion.

Table 1. TRP calculated for fake papers for the chunk size *L=200.*

	N=3	*N=4*	*N=5*
Clustering	1	1	1
SVM Rule 1	1	1	1
SVM Rule 2	1	1	1

Table 2. TRP calculated for real papers the chunk size *L=200.*

	N=3	*N=4*	*N=5*
Clustering	0.86	0.82	0.64
SVM Rule 1	0.96	0.88	0.80
SVM Rule 2	0.92	0.90	0.90

This can be corrected by increasing the length of the chunk. The subsequent tables provide outcomes obtained for the chunk size *L=400.*

Table 3. TRP calculated for fake papers the chunk size *L=400.*

	N=3	*N=4*	*N=5*
Clustering	1	1	1
SVM Rule 1	1	1	1
SVM Rule 2	1	1	1

Table 4. TRP calculated for real papers the chunk size *L=400.*

	N=3	*N=4*	*N=5*
Clustering	1	1	1
SVM Rule 1	1	0.96	0.98
SVM Rule 2	0.92	0.92	0.92

4.2.2 Content-free word model

As was mentioned earlier, the content-free word model is being a kind of statistical glue. While it does not have any sense, it helps to join the words into meaningful texts.

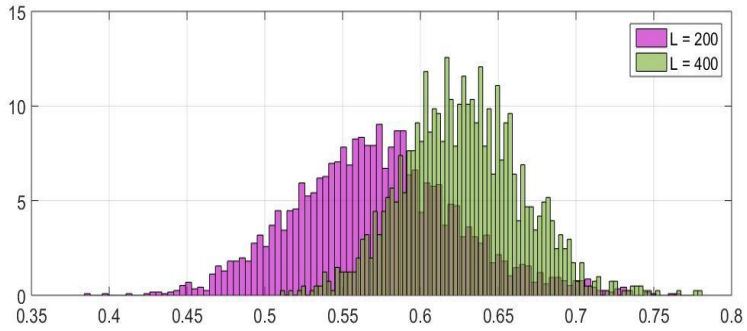

Fig. 6. Histograms of the ZV_T values within the training class calculated for *L=200* and *L=400* using the content-free word.

This figure represents histograms of the Mean Dependency found for two different values of the chunks size *L*. The corresponding means are 0.57 and 0.63. The fake papers are surely recognized again:

Table 5. TRP calculated for fake papers for the chunk size *L=200* and *L=400.*

	L=200	*L=400*
Clustering	1	1
SVM Rule 1	1	1
SVM Rule 2	1	1

However, detection of real papers is successful just for sufficiently large values of *L*.

Table 6. TRP calculated for real papers the chunk size *L=200, L=400,* and *L=500.*

	L=200	*L=400*	*L=500*
Clustering	0.62	0.96	0.96
SVM Rule 1	0.58	1	0.98
SVM Rule 2	0.78	0.78	0.78

5 Conclusion

This paper proposes a novel method constructed to make a distinction between computers generated scientific and human written papers. The problem, treated in the framework of the one class classification methodology, is considered using the one-class SVM methodology compared with a clustering approach. N-grams based and the content-free word based dictionaries are applied for the building of vector representations of texts. The key issue providing the highly reliable results is the dynamic distance measuring dissimilarity between particular implementations of the writing process. The fake papers are surely identified for all configurations of the system

parameters. The approach also classifies sufficiently well the human written papers for suitably chosen parameters values. We are planning to extend the proposed research aiming to test other outliers detection techniques.

References

1. Lavoie, A., Krishnamoorthy, M.: Algorithmic Detection of Computer Generated Text. arXiv:1008.0706, 2010arXiv1008.0706L, AUG 2010, aRXIV (2010).
2. Labbe, C., Labbe, D.: Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? Scientometrics, 94(1), 379–396 (2013).
3. Fahrenberg, U., Biondi, F., Kongshøj, K.C.J., Axel, L.: Measuring global similarity between texts, In Statistical Language and Speech Processing Second International Conference, pp. 220-232, SLSP 2014, Grenoble, France (2014).
4. Xiong, J., Huang, T.: An effective method to identify machine automatically generated paper. In Knowledge Engineering and Software Engineering. KESE'2009. Pacific-Asia Conference on. IEEE, pp. 101–102 (2009).
5. Dalkilic, M. M., Clark, W. T., Costello, J. C., Radivojac, P.: Using compression to identify classes of inauthentic texts. In Proceedings of the 2006 SIAM Conference on Data Mining (2006).
6. Amancio, D.R.: Comparing the topological properties of real and artificially generated scientific manuscripts. Scientometrics 105(3), 1763–1779 (December 2015).
7. Williams, K., Giles, C.L.: On the use of similarity search to detect fake scientific papers. In Similarity Search and Applications - 8th International Conference, SISAP 2015, 332–338 (2015).
8. Nguyen, M.T., Labbe, C.: Engineering a tool to detect automatically generated papers, in BIR@ECIR, ser. CEUR Workshop Proceedings, P. Mayr, I. Frommholz, and G. Cabanac, Eds., vol. 1567. CEURWS. org, 2016, pp. 54–62 (2016).
9. Volkovich, Z., Granichin, O., Redkin, O., Bernikova, O.: Modeling and visualization of media in Arabic, Journal of Informetrics, 10(2), 439–453 (May 2016).
10. Volkovich, Z.: A Time Series Model of the Writing Process, Machine Learning and Data Mining in Pattern Recognition, In 12th International Conference, MLDM 2016, Proceedings, New York, NY, USA, July 16-21, pp. 128-142 (2016).
11. Volkovich, Z., Avros, R.: Text Classification Using a Novel Time Series Based Methodology, In 20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES 2016, 5-7 September 2016, York, United Kingdom, Procedia Computer Science 96, pp. 53 – 62 (2016).
12. Korenblat, K., Volkovich, Z.: Approach for Identification of Artificially Generated Texts, HUSO 2107: In the Third International Conference on Human and Social Analytics.
13. Amelin, K., Granichin, O., Kizhaeva, N., Volkovich, Z.: Patterning of writing style evolution by means of dynamic similarity, Pattern Recognition, vol. 77, pp.45-64 (2018).
14. Kendall, M.G., Gibbons, J.D.: Rank correlation methods. London: Edward Arnold (1990).
15. Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. In Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS'99), pp. 582-588, S. A. Solla, T. K. Leen, and K. Müller (Eds.). MIT Press, Cambridge, MA, USA (1999).
16. Harmer, J.: How to teach writing. Delhi, India: Pearson Education (2006).
17. Some Webpage, URL: www.arXiv.org/archive/cs [accessed: 2017-07-02].

18. Juola, P.: Authorship attribution, Foundations and Trends in Information Retrieval, 1 (3), pp. 33-334, (2006).
19. Binongo, J.: Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution, Chance 6 (2) , pp 9-17, (2003).
20. Hughes, J. M. , Foti, N. J. , Krakauer D. C., Rockmore D. N.: Quantitative patterns of stylistic influence in the evolution of literature, in: Proceedings of the National Academy of Sciences, Vol. 109, pp. 7682–7686, (2012).