Approach for Identification of Artificially Generated Texts

Katerina Korenblat and Zeev Volkovich

Department of Software Engineering
ORT Braude College
Karmiel, 21982, Israel
Email: katerina@braude.ac.il, vlvolkov@braude.ac.il

***Abstract*—The paper is devoted to a new method for the identification of the artificially composed scientific papers. We consider this problem from the general point of view of the writing style. It is natural to suppose that the style of artificial generated manuscripts has to be substantially different from this one of the human generated articles because the human writing process is established in inherently another manner. The Mean Dependency Distance introduced in previous authors' works is used to quantify the writing process developing. A set of artificially generated manuscripts is taken and the distance values are calculated to sequential chunks of all papers. A suspected document is also divided into chunks, and a version of the known KNN method is applied together with a distance-based outlier detection method to classify it as a real or a fake document. The provided numerical experiments demonstrate high ability of the method to distinguish between two types of documents.**

Keywords–Scientific Frauds; SCIgen; Classification.

I. INTRODUCTION

In 2005, three computer science Ph.D. students at the Massachusetts Institute of Technology-Jeremy Stribling, Max Krohn, and Dan Aguayo proposed a program, named SCIgen, intended to produce senseless manuscripts in the computer science field. Afterward, a group of scientific document generators was invented, including SCIgen-Physic concentrating on physics, Mathgen focusing on math and the Automatic SBIR (Small Business Innovation Research) Proposal Generator dealing with grant proposal fabrication. Initially, the generators were created as hoaxes with the aim to unmask scientific conferences that really rip off researchers with publication and fees. At the first glance, generated papers appear to be sensible, because they are structured to have all needed components of a paper, such as an abstract, an introduction, graphs, diagrams, citations and so on. The papers are reasonably organized employing context-free grammar and could confuse inexperienced persons. Not formally speaking, the named generators have learned the overall rules commonly used during during writing scientific papers and successively imitate this process. Each person can compose a fake paper using the site [1].

Some articles studied automatic identifying of SCIgen papers. For example, the problem was respected in [2] by proving of the external references. A paper is considered as artificial if a portion of its unrevealed references is sufficiently large. The paper [3] deals with distribution of a papers keyword inside the document, which is natural expected to be appropriately uniform. In [4], a compression profiles of texts are analyzed. The conclusion is based on difference in the compression rate between the authentic and computer generated texts. The ROUGE metrics [5] was used in [6]. Paper [7] along the lines of [8] suggested to use the structural distance between texts. In [9] topological properties of the natural and the generated texts were compared. Different measures to uncover artificial scientific papers were evaluated in [10] and [11].

In this paper, we consider artificial generated manuscripts from the point of view of their own writing style. It is natural to suppose that the style of artificial generated manuscripts has to be substantially different from this one of the human generated articles because the human writing process is completely established in inherently another manner. One of the common viewpoints on the human writing process (see, for example [12]) considers this process as composed of four key elements: planning, drafting, editing, and writing the final draft. Thus, it is natural to presume that dependency between sequential written text parts has remained at the almost uniform level if the text is composed by the same author. On the other site, SCIgen operates with a context-free grammar to produce a text. Essentially, the generator does not compose a paper, but goes along the predefined pattern by randomizing out prior components. An approach quantifying writing style development was introduced in [13]. In this paper, we use a distance between writing styles in order to evaluate dissimilarities between fake and real documents. A set of artificially generated manuscripts is taken, and the distance values are calculated to sequential chunks of all papers. A suspected document is also divided into chunks, and a variant of the known KNN method is applied together with a distance-based outlier detection method to check if it is a real or a fake document.

The remainder of the paper is organized as follows. In Section II, we provide the background on the theory proposed in [13]. The suggested methodology is explained in Section III. Section IV is devoted to numerical experiments. We conclude our paper in Section V.

II. MEAN DEPENDENCY

Let us consider $\mathbf{D}$ as a collection of texts and get a semi-distance function Dis defined on $\mathbf{D}\times\mathbf{D}$:

- $Dis(\mathcal{D}_1, \mathcal{D}_2) \geq 0$ for all $\mathcal{D}_1, \mathcal{D}_2 \in \mathbf{D}$.
- $Dis(\mathcal{D}, \mathcal{D}) \geq 0$ for all $\mathcal{D} \in \mathbf{D}$.

It is not suggested that $Dis(\mathcal{D}_1, \mathcal{D}_2) = 0$ implies that $\mathcal{D}_1 = \mathcal{D}_2$. In the framework of our model, we set a chunk size L and consider a document $\mathcal{D} \in \mathbf{D}$ as a series of sequential sub-documents: $\mathcal{D} = \langle \widehat{\mathcal{D}}_1, ..., \widehat{\mathcal{D}}_m \rangle$ of the length L. In the formal language theory terminology, $\mathcal{D}$ is the concatenation of $\widehat{\mathcal{D}}_1, ..., \widehat{\mathcal{D}}_m$.

Our perception suggests that a document $\mathcal{D}$ is considered as an outcome provided by "a random number generator" reflecting the writing style of the authors. Aiming to quantify the evolution of a text within the writing process, we introduce the Mean Dependency characterizing the mean relationship between a chunk $\widehat{\mathcal{D}}_i$, $i = T+1, ...m$ and the set of its T "precursors":

$$ZV^{(L)}_{T,Dis}(\widehat{\mathcal{D}}_i, \Delta_i) = \frac{1}{T} \sum_{\widehat{\mathcal{D}} \in \Delta_i} Dis(\widehat{\mathcal{D}}_i, \widehat{\mathcal{D}}), \qquad (1)$$

where $\Delta_i = \left\{ \widehat{\mathcal{D}}_{i-j},\ j = 1, ..., T \right\}$ is the set of T "precursors" of $\widehat{\mathcal{D}}_i$. To distinguish styles a function measuring dissimilarity among texts pieces is proposed by the following way:

$$DZV^{(T)}_L(\widehat{\mathcal{D}}_i, \widehat{\mathcal{D}}_j) = \qquad (2)$$

$$= \left| \begin{array}{c} ZV^{(L)}_{T,Dis}(\widehat{\mathcal{D}}_i, \Delta_i) + ZV^{(L)}_{T,Dis}(\widehat{\mathcal{D}}_j, \Delta_j) - \\ -ZV^{(L)}_{T,Dis}(\widehat{\mathcal{D}}_i, \Delta_j) - ZV^{(L)}_{T,Dis}(\widehat{\mathcal{D}}_j, \Delta_i) \end{array} \right| .$$

It is easy to see that $DZV^{(T)}_L$ is also a semi-metric. Once $DZV^{(L)}_T(\widehat{\mathcal{D}}_i, \widehat{\mathcal{D}}_j) = 0$ the sub-documents $\widehat{\mathcal{D}}_i$ and $\widehat{\mathcal{D}}_j$ exhibit close relationships with the own previous neighbors and the previous neighbors of another one. From the writing style standpoint the sub-documents appear to be very similar.

Distance function choice is essential in the proposed approach. A relevant distance function may be extracted to reflect writing style attributes. In the text mining domain, it is more acceptable to convert texts into a probability distribution and afterwards to use a distance between them. We suggest that there is a transformation $\mathcal{F}$, which maps the documents belonging to $\mathbf{D}$ into the set $\mathbf{P}$ of the probability distributions on $[0, 1, 2, ...]$, and

$$Dis(\mathcal{D}_1, \mathcal{D}_2) = dis\left(\mathcal{F}\left(\mathcal{D}_1\right), \mathcal{F}\left(\mathcal{D}_2\right)\right),$$

where dis is a distance function (a simple probability distance) defined on $\mathbf{P}$. In the current paper we use the following Spearman's correlation distance function:

$$Dis(\mathcal{D}_1, \mathcal{D}_2) = S(\mathcal{D}_1, \mathcal{D}_2) = 1 - \rho(\mathcal{F}(\mathcal{D}_1), \mathcal{F}(\mathcal{D}_2)),$$

where ρ is the Spearman's ρ (see, [14]), which is calculated for distributions of $\mathcal{F}(\mathcal{D}_1)$ and $\mathcal{F}(\mathcal{D}_2)$ treated as a kind of ordinal data such that the frequency values are regarded as the rank positioning.

As usual, a transformation $\mathcal{F}$ is constructed by means of the common Vector Space Model. This model disregards grammar and the order of terms, but keeps the collection of terms. Each document is described via a terms frequency table in contradiction of the vocabulary containing all the words (or "terms") in all documents in the corpus. The tables are considered as vectors in a linear space having a dimensionality equal to the vocabulary size.

In the Bag of Words Model a document is represented as the distribution of its words. To reduce the space dimensionality, the stop-words are commonly removed. The Keywords Model is an offshoot of the previously discussed model, where a document is represented not as a bag of all terms in the corpus but as a bag of selected words. In the N-grams Model the vocabulary consists of all N-grams in the corpus. An N-gram is a contiguous N-character slice of a longer text constructed frequently by means of the symbols occurring in a slide window of length N.

III. Methodology

We handle the considered task in the framework of the one-class classification methodology. One-class classification is based on the presumption that merely data of one of the groups, named the target class, are accessible, although there are no information of the other class (also called the outer class).

In our model the target class is composed from artificially generated papers, while the outliers class is suggested to contain the human written papers. By this way we take a collection of artificially generated papers $\mathbf{D}_0$ and chose the N-grams order N, the delay parameter T and size of the chunks L. All documents from $\mathbf{D}_0$ are divided into chunks having size L and a "cloud" of all chunks $CH\left(\mathbf{D}_0\right)$ is constructed as $\left\{ DZV^{(T)}_L(\widehat{\mathcal{D}}_i, \widehat{\mathcal{D}}_j) \right\}$ calculated for all possible, having at least T "precursors", chunks of the documents from $\mathbf{D}_0$. An example of the principal-component analysis plots of such a distance matrix is given in Figure 1, where a cloud is marked in red. A tested document's chunks are marked in blue.

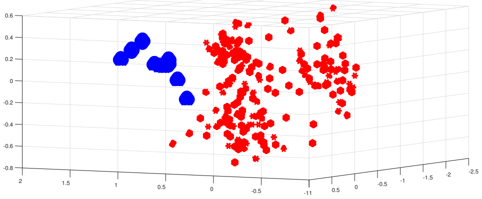

Figure 1. Principal-component analysis plots of a distance matrix

As can we see, the points corresponding to a test document are actually outliers in the red marked documents of the artificially generated papers, because they are located outside of their cloud. There are a lot of methods of the one-class classification (see, for example [15]). We use in this paper two of the most common approaches:

1) The KNN classification, where a text segment is assigned or not to the class by a majority vote of its k nearest neighbors. This algorithm is very intuitive one natural suggesting that each segment is similar to its nearest neighbors.
2) Distance-based outliers (DBO) detection:
    a) Determine a central point of $\mathbf{D}_0$:

$$\widehat{\mathcal{D}}_0 = \underset{\widehat{\mathcal{D}} \in CH(\mathbf{D}_0)}{\arg\min}\ mean\left(DZV^{(T)}_L(\widehat{\mathcal{D}}_0, \widehat{\mathcal{D}}) \right).$$

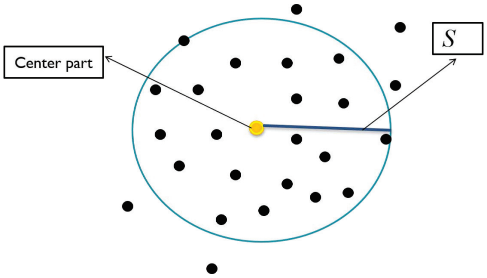

Figure 2. Distance-based outliers detection

b) Calculate $M = mean\left(DZV_L^{(T)}(\widehat{\mathcal{D}}_0, \widehat{\mathcal{D}})\right)$ and $S = std\left(DZV_L^{(T)}(\widehat{\mathcal{D}}_0, \widehat{\mathcal{D}})\right)$, where $\widehat{\mathcal{D}} \in CH(\mathbf{D}_0)$ and $std()$ is a standard deviation function.

c) A chunk $\widehat{\mathcal{D}}$ is recognized as an outlier if $DZV_L^{(T)}(\widehat{\mathcal{D}}_0, \widehat{\mathcal{D}}) > M + S$.

d) A paper is assigned to outliers (a human written paper) if a majority of its own chunks are recognized as outliers.

An illustration of the (DBO) method is given in the scheme presented in Figure 2. Here, the cloud is modelled as a circle having a radius equal to the cloud standard deviation S and the center located at the suggested center of gravity of the cloud. The outliers are associated with this case with point found outside of the circle. Aiming to evaluate significance of a voting in the applied procedures, a p-value is calculated within the null hypothesis according to the theoretical fraction pr of the majority voting chunks

$$H_0 : pr = \frac{1}{2}$$

against an alternative hypotheses

$$H_0 : pr > \frac{1}{2}.$$

For the KNN classification p-value is found as

$$p = CBIN\left(\widehat{pr} * k, k, \frac{1}{2}\right),$$

where k is the number of the nearest neighbors, $\widehat{pr}$ is a fraction of the majority voting ones, and $CBIN$ is the Cumulative Binomial Distribution.

In case of DBO, $\widehat{pr}$ is the observed proportion of the the majority voting chunks within the total number of a document's chunks m. Here

$$p = \Phi\left(\frac{\widehat{pr} - \frac{1}{2}}{\frac{1}{2}}\sqrt{m}\right),$$

where Φ is the cumulative function of the Standard Normal Distribution.

A value of p greater than some predefined threshold, typically 0.95, indicates a significant voting.

IV. Numerical experiments

As fake documents, one hundred papers are generated by the SCIGen procedure. One hundred real manuscripts are recovered from the ”arXiv” repository [16]. This number of the artificial and human written papers appear to be very reasonable and provides fair results. However, we are going in the future to study a possibility to reduce the size of the training set.

In the texts involved in the experiments any uppercase characters are converted to the corresponding lowercase characters, and all other characters are unchanged. The experiments are provided through the chunk size $L = 100, 200$ and 400 with $T = 10$. The number k is 10 in the KNN-classification.

Fifty artificial papers are used as a training corpus: each document is divided into chunks of size L and the distance values are calculated according to (2). In the first series of the experiments, the real papers were compared with training set using two mentioned approaches. The results obtained from considering real papers are presented in Table I by means of the positive predictive values (precision) calculated within the experiments. Recall that precision is an attained proportion of the true positive results. Almost all $p-$values are properly close to one. It is easy to see that the KNN method exhibits very stable behavior, which is practically independent of the choice of L. The achieved outcomes are very precise in all cases. So, almost in all experiments the considered real papers are recognized as human written.

TABLE I. POSITIVE PREDICTIVE VALUES CALCULATED FOR 100 REAL PAPERS.

L	100	200	400
KNN	0.98	0.99	0.98
DBO	0.63	0.81	0.88

In the second experiments series, fifty additional artificial papers were checked against the fifty training ones.

TABLE II. POSITIVE PREDICTIVE VALUES CALCULATED FOR 50 ADDITIONAL ARTIFICIAL PAPERS.

L	100	200	400
KNN	1	1	1
DBO	1	1	1

So, each fake document is acknowledged with probability one.

V. Conclusion

The article proposes a new technique for recognition artificially generated scientific papers resting upon written style characteristics. Texts are split into chunks and shown by means of a histogram shape defined via the 3-gram frequency ranks. Then, the Mean Dependency describing the mean rank correlation of a sub-document with its numerous precursors provides a distance amid chunks styles. Two classifiers by means of this distance within the KNN and the distance-based outlier detection methodologies are constructed and tested. It turns out that a KNN based approach trained on a sufficient number of fake papers is capable almost surely to distinguish between fake and real papers. The second classifier demonstrates less robust results. We are planning to extend our method aiming to construct new classification rules using one-class and two-class classification approaches.

REFERENCES

[1] “Some Webpage,” URL: https://pdos.csail.mit.edu/archive/scigen/ [accessed: 2017-07-02].

[2] J. Xiong and T. Huang, “An effective method to identify machine automatically generated paper,” in Knowledge Engineering and Software Engineering, 2009. KESE’09. Pacific-Asia Conference on. IEEE, 2009, pp. 101–102.

[3] A. Lavoie and M. Krishnamoorthy, “Algorithmic Detection of Computer Generated Text,” eprint arXiv:1008.0706,2010arXiv1008.0706L, AUG 2010, aRXIV.

[4] M. M. Dalkilic, W. T. Clark, J. C. Costello, and P. Radivojac, “Using compression to identify classes of inauthentic texts,” in Proceedings of the 2006 SIAM Conference on Data Mining, 2006.

[5] C.-Y. Lin and E. Hovy, “Automatic evaluation of summaries using n-gram co-occurrence statistics,” in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, ser. NAACL ’03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 71–78. [Online]. Available: http://dx.doi.org/10.3115/1073445.1073465

[6] D. P. F. Labbé, Cyriland Labbé, Detection of Computer-Generated Papers in Scientific Literature. Cham: Springer International Publishing, 2016, pp. 123–141.

[7] U. Fahrenberg, F. Biondi, K. C. J. Kongshøj, and L. Axel, “Measuring global similarity between texts,” in Statistical Language and Speech Processing Second International Conference, SLSP 2014, Grenoble, France, October 14-16, 2014, Proceedings, 2014, p. 220232.

[8] C. Labbé and D. Labbé, “Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science?” Scientometrics, vol. 94, no. 1, 2013, pp. 379–396.

[9] D. R. Amancio, “Comparing the topological properties of real and artificially generated scientific manuscripts,” Scientometrics, vol. 105, no. 3, 2015, pp. 1763–1779. [Online]. Available: http://dx.doi.org/10.1007/s11192-015-1637-z

[10] K. Williams and C. L. Giles, ”On the Use of Similarity Search to Detect Fake Scientific Papers”. Similarity Search and Applications, OCT 2015, pp. 332–338.

[11] M.-T. Nguyen and C. Labb, “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. CEUR-WS.org, 2016, pp. 54–62.

[12] J. Harmer, Ed., How to teach writing. Delhi, India: Pearson Education, 2006.

[13] Z. Volkovich, O. Granichin, O. Redkin, and O. Bernikova, “Modeling and visualization of media in arabic,” Informetrics, vol. 10, no. 2, 2016, pp. 439–453.

[14] M. G. Kendall and J. D. Gibbons, Rank correlation methods. London: Edward Arnold, 1990.

[15] S. S. Khan and M. G. Madden, “One-class classification: Taxonomy of study and review of techniques,” CoRR, vol. abs/1312.0049, 2013. [Online]. Available: http://arxiv.org/abs/1312.0049

[16] “Some Webpage,” URL: www.arXiv.org/archive/cs [accessed: 2017-07-02].