Internet of Things and Cloud Computing
2019; 7(1): 19-24
http://www.sciencepublishinggroup.com/j/iotcc
doi: 10.11648/j.iotcc.20190701.13
ISSN: 2376-7715 (Print); ISSN: 2376-7731 (Online)

Effectiveness Level of Online Plagiarism Detection Tools in Arabic

Ghadah Mohammed Abdullah Adel, Yuping Wang

Computer Science and Technology, Xidian University, Xi'an, China

Email address:
Ghada.adel23@gmail.com (G. M. A. Adel), ywang@xidian.edu.cn (Yuping Wang)

To cite this article:
Ghadah Mohammed Abdullah Adel, Yuping Wang. Effectiveness Level of Online Plagiarism Detection Tools in Arabic. *Internet of Things and Cloud Computing*. Vol. 7, No. 1, 2019, pp. 19-24. doi: 10.11648/j.iotcc.20190701.13

Received: April 9, 2019; **Accepted**: May 20, 2019; **Published**: May 23, 2019

Abstract: Plagiarism affects education quality, academic research results and publishers reputation. Consequently, many online plagiarism tools have been developed to detect and reduce such affects. However, most of these tools were evaluated according to their abilities to reveal different rates of plagiarism in English text. While evaluating their capability in detecting different plagiarism rates from different patterns in Arabic text is still vague. This paper aims to evaluate the efficiency level of online academic plagiarism detection tools (PlagScan, iThenticate and CheckForPlagiarism.net) in detecting different plagiarism patterns' amounts in Arabic language. A comparison was made between, PlagScan, iThenticate and CheckForPlagiarism.net, detection capabilities by merging university theses and dissertations with eight plagiarism patterns (whole document, some parts, insertion, sentence split or join, phrase reordering, syntax, lexical and morpho-syntactic) with the ratio between 90% , 30% and 10% respectively. Experiment's results showed that iThenticate is the most efficient online plagiarism detection tool in Arabic for eight plagiarism patterns between 90% and 80% ratio Arabic language. While none of the three online plagiarism detection tools are efficient for less than 80% plagiarized text from any of the eight plagiarism patterns. Hence, mechanism enhancements and consideration to the Arabic anguage structure are recommended for online plagiarism detection tool in Arabic.

Keywords: Academic Plagiarism, Plagiarism Levels and Patterns, Online Plagiarism Detection Tools, Arabic Plagiarism Detection, Effectiveness of Plagiarism Detection Tools

1. Introduction

Plagiarism is extracting a whole or a part of someone else work's either as a free text or as a program source code in order to use it or imputes it to the person's own [1]. Thus, free text plagiarism is the most common style in research papers [2]. The occurrence of plagiarism is either done by purpose or because of the luck of people awareness in importance of citation and the ways to avoid it [3]. There are two types of plagiarism, literal and intelligent [4]. Literal plagiarism refer to either copy text exactly as whole or some parts, or nearly copy text by either insertion or deletion, split or join sentences and substitution, or to modify the sentence syntax or order. While intelligent plagiarism points to using idea adaption, translation, summarizing or paraphrasing the text. Regardless which plagiarism patterned was used, the harmless of such an action leads to lower academic research quality and publishing standards.

Due to plagiarism affects, universities and publishers have specify some plagiarism levels and ratios to react upon. For example, plagiarizing more than 50%, less than 20% or between are considered as the top three levels of plagiarism violation defined by Institute of Electrical and Electronics Engineers (IEEE) [5]. As a result, many universities, which are interested in the research field and higher studies, have developed effective plagiarism detection tools to prevent such behavior and to publish researches with a high quality. However, most of the current plagiarism detection systems were developed for identifying plagiarism in English language [4, 6], their effectiveness with Arabic language needs to be evaluated in order to implement them in Arab universities to produce better Arabic academic writing. As a consequences, researchers' and the publishers' reputation are affected. Furthermore, many researches and publishers are frightened of their intellectual property [7]. For these reasons, it is so critical to detecting precisely the plagiarism percentage in whatever the plagiarism pattern is. Detection should be done using an effective plagiarism detection

software even if the human interference is needed [8].

The main goal of this paper is to determine the most effective online plagiarism tool for the Arabic language among the best ones mentioned in [9]. The efficiency of the selected plagiarism tool will be assessed according to the amount of theses and plagiarism patterned detected from each of the top three plagiarism levels defined by IEEE in [5].

The rest of this paper is organized as follows. Section 2 coves the study background about the plagiarism. Section 3 illustrates our experiment and implementation process. Section 4 presents the results and discussion. Finally, the conclusion and future work is presented in section 5.

2. Related Work

2.1. Plagiarism Detection Approaches and Tools

There are two approaches to detect plagiarism which are intrinsic and external [10]. Each approach can be implemented using different techniques. Intrinsic approach's techniques are based on natural language processing (NLP) to detect similarity [11]. Depending on text features such as grammar, structure and syntax or part of speech (POS) plagiarism can be located [4, 12]. While the external approach based on string matching, vector model, statistical and probability, fingerprinting and winnowing techniques [3, 11, 13, 14].

On the other hand, desktop and online tools have been developed for the seek of detecting plagiarism. These tools varied according to their features and language support. An overview of some of the popular paid plagiarism detection tools such as Ephorus, Plagiarism Scanner, Turnitin..etc. and free plagiarism detection tools such as Plagium, SID, Plaggie was elaborated in [11].

2.2. Textual Features of Arabic Language

Arabic language is one of the Semitic languages that is spoken by over than 300 million persons. Arabic language consist of 28 (ي ... ث ,ت ,ب ,أ) letters with 3 vowels (ي ,و ,ا) [15]. Some letters are written with dots and according to specific patterns such as (فعل , مفعول) and sometimes using diacritics Dhamma (ُ) , FatHa (َ) and Kasra (ِ). Moreover, the writing orientation is from the right to the left though the main sentence structure is verb, subject and object [16].

2.3. Arabic Plagiarism Detection Methods

Some recent studies have focused on developing systems or tools to detect plagiarism in Arabic text. From these studies, a study developed an E-learning system based on statement-based fingerprints matching and fuzzy-set information retrieval methodology [4]. This methodology helps in detecting some plagiarism patterns such as copy paste, restructure, paraphrased and reorder statements.

Another study demonstrated a new plagiarism detection tool which was called "APlag for Arabic text" [17]. This tool depends on the content based method and the logical representation of the document in order to identify similarities issues for Arabic text. The mechanism of detection is based on text preprocessing, fingerprinting, document representation, and similarity metrics.

One more study launched RDI_RED plagiarism detection system for Arabic documents [18]. A search engine was used to select the source documents. Then the plagiarized texts were aligned with the original documents. After that some filtering rules was applied for plagiarism detection. The system can be applied online easily but it cannot recognize paraphrasing or substitution plagiarism patterns.

Other study [19] proposed a system based on character based technique in detecting plagiarism. Considering the text features and n-grams of the Arabic documents are used for filtering the plagiarized text. The Latent Semantic Analysis (LSA) and Singular Value Decomposition (SVD) are used for analysis [20].

On other hand a quantitative and qualitative comparison were made between; PlagScan, iThenticate, CheckForPlagiarism.net, online plagiarism detection tools in Arabic documents [9]. The comparison was aimed to identify the top plagiarism tool in detecting as much as literal and paraphrasing plagiarism patterns and in which percentage. The study concluded that Plagscan can detect most of selected plagiarism patterns. Although Plagscan and iThenticate results' was nearly closed to the plagiarized text percentage of almost all literal plagiarism patterns but unable to recognize paraphrasing and syntax patterns. However, measuring the efficacy of each plagiarism detection tool in detecting different pattern from different plagiarism amount or levels was recommended.

3. Experiment's Framework Design and Implementation

Implementing a systematic framework and methods usually lead to accurate and efficient results. According to that, our experiment was established and divided into two stages: dataset preparation and detection of plagiarism level percentage. In the dataset preparation stage, the experiment's dataset were molded, the instruments and plagiarism patterns were chosen. While in the detection of plagiarism level percentage stage, a measurement was contacted based on the three plagiarism levels defined by IEEE in [5]. As a result, a percentage amount was recorded for each plagiarism tool's ability to detect the same amount of plagiarism in each of the experiments' selected patterns from [4].

3.1. Dataset Preparation

The dataset formulation process was the most important step in this paper. For that some factors were needed to be considered and several steps were followed for selection and preparation. These factors were determining the original and plagiarized texts' source and selecting plagiarism patterns. For the original text, it was extracted from 200 universities' theses and dissertation because the higher impacts are usually

occurred in the academic communities [21]. However, the plagiarized text was extracted from one of the Internet websites as it is the most common source of plagiarism after books [22]. While Eight plagiarism patterns (whole document, some parts, insertion, sentence split or join, phrase reordering and morpho-syntactic) were selected from [4]. Also as another study stated that these patterned are the most common forms of plagiarism among the academic and research community [23].

Therefore Three steps were needed to form the suitable dataset for the experiments. These steps are illustrated in Figure 1. Firstly, we extracted Arabic text from one of the Internet websites to use it as the plagiarized text. Next, we proceeded with formulation plagiarized text into the selected eight plagiarism patterns (whole document, some parts, insertion, sentence split or join, phrase reordering, syntax, lexical and morpho-syntactic). Then IEEE plagiarism levels were used as a standardized percentage rate to determine the plagiarized text amount in each plagiarism pattern [5]. The percentage rate distribution on each plagiarism pattern was selected as between 90% and 80% from level 1, 30% from level 2 and 10% from level 3 ass represented in Table 1. Lastly, the plagiarized and the original texts are merged into one theses or dissertation and distributed according to each plagiarized pattern and level to generate the final sample test.

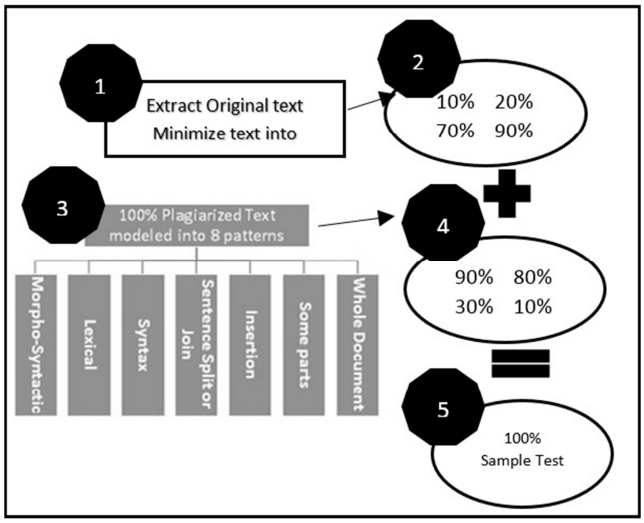


Figure 1. *Dataset Preparation Stages.*

Table 1. *Distribution of Plagiarism levels % among Plagiarism pattern.*

Pattern	Level							
	> 50%		> 0% & < 20%		> 20% & < 50%		> 50%	
	% Original Text	% Plagiarized Text	% Original Text	% Plagiarized Text	% Original Text	% Plagiarized Text	% Original Text	% Plagiarized Text
Whole Document	10%	90%	80%	20%	70%	30%	×	×
Some Parts	×	×						
Sentence Split or Join	×	×						
Insertion	×	×						
Phrase reordering	×	×	80%	20%	70%	30%	20%	80%
Syntax	×	×						
Lexical	×	×						
Morpho-Syntactic	×	×						

3.2. Instruments Selection

Identifying the appropriate online plagiarism detection tools were defined by Three features. The first feature was the available submission credit while the second was the maximum number of words count per account and finally, the support of Arabic language [1, 9]. By examining these features and compare them with ours, three checkers matched the conditions which are: PlagScan, iThenticate and CheckForPlagiarism.net.

4. Results and Discussion

Two main criteria are applied in the measurement process. The first one was used to determine how many plagiarized theses had been detected by each online plagiarism checker. While the second one was used to indicate the ability of each online plagiarism checker to extract the exact plagiarized amount in the Arabic theses from each plagiarism levels. Both of the criteria's are measuring the effectiveness of each tool.

4.1. Number of These Detected in Each Plagiarism Level

Several tests were made for each plagiarism checkers in order to get reliable results. From these tests the capability of each instrument which is highlighted in Table 2. These figures are showing the number of the plagiarized Arabic theses that each checker had detected from 200 theses. The detection process was according to the three levels illustrated in Table 2.

Table 2. *No. of Theses detected by each Checker in each level of the eight plagiarism patterns.*

Plagiarism Patterns	No. of Theses Detected by each Checker in each Level								
	iThenticate			CheckForPlagiarism.net			PlagScan		
	L1	L2	L3	L1	L2	L3	L1	L2	L3
Whole Document	200	×	×	160	×	×	200	×	×

Plagiarism Patterns	No. of Theses Detected by each Checker in each Level								
	iThenticate			CheckForPlagiarism.net			PlagScan		
	L1	L2	L3	L1	L2	L3	L1	L2	L3
Some Parts	200	200	20	190	180	90	200	110	30
Insertion	200	200	20	80	40	0	200	200	30
Sentence Split or Join	200	200	20	60	40	20	200	170	30
Phrase Reordering	200	200	20	60	40	50	200	150	30
Syntax	10	20	30	30	0	0	20	20	40
Lexical	20	20	20	40	20	20	30	10	20
Morpho-Syntactic	200	200	20	130	60	10	110	40	50

So as Table 2 elaborates that iThenticate checker discovered the plagiarized text from the whole 200 theses in level 1 and 2 from six patterns out of eight. Therefore the highest number of theses detected of syntax and lexical patterns was 30 theses from level 3. While the highest number of theses detected by CheckForPlagiarism.net checker was 190 from level 1, 180 from level 2 and 90 from level 3 all from some parts pattern. Furthermore, even though Plagscan shares the same highest number of theses detected in level 1 and level 2 (200 theses) with iThenticate, but it detected 20 theses more in level 3 than iThenticate.

These experiments led to three conclusions regarding to the checkers efficiency. The first one is about iThenticate. This checker was able to notice the whole 200 plagiarized theses from level 1 and 2. These plagiarized theses were formed from five plagiarism patterns which are whole document, some parts, insertion, sentence split or join, morpho-syntactic and phrase reordering. While it could not notice any 200 plagiarized theses from any pattern in level 3. It means that iThenticate checker is more effective on level 1 and 2 on whole document, some parts, insertion, sentence split or join, morpho-syntactic and phrase reordering patterns' then the other patterns of Arabic text.

The second conclusion is about CheckForPlagiarism.net checker. This checker was able to get the highest detection number from level 1, 2, 3 in some parts pattern. However, it could not get any full 200 plagiarized Arabic theses in any level. It means that CheckForPlagiarism.net checker is somehow effective on some parts only then the other patterns for Arabic text.

The third conclusion is about PlagScan checker. This checker was able to notice level 1 from five plagiarism patterns which are whole document, some parts, insertion, sentence split or join and phrase reordering. While it noticed only full plagiarized theses in some parts pattern. Whereas it could not detect any 200 theses in level 3 from any pattern. This means that PlagScan checker is more effective on level 1 of whole document, some parts, insertion, sentence split or join and phrase reordering patterns' then the other patterns for Arabic text. Moreover, level 2 from some parts pattern also has gain an effective results.

In contrast with previous studies which implement a quantitative comparisons between these online checkers in term of features [1, 9], this study compare the quality of each checker in Arabic language. Moreover, instead of building a new tool to detect plagiarism [4, 16, 17, 18], our study used the most common online plagiarism checkers however, our results confirmed with the previous studies in term of checkers ability to detect literal plagiarism patterns and fails on detecting intelligent types. In addition, our results confirm that iThenticate plagiarism checker was able to detect most of the plagiarized documents which contrary with top-3 comparison which mentioned that Plagscan is the top checker [9].

4.2. Amount Detected in Each Level

Measuring the effectiveness of each online plagiarism checker is not depending only on the amount of plagiarized these discovered. However, determining the plagiarized parentage discovered for each level in each pattern is more critical scale for effectiveness measurement. In this section, tests were made on each online checker to determine each checker effectiveness level.

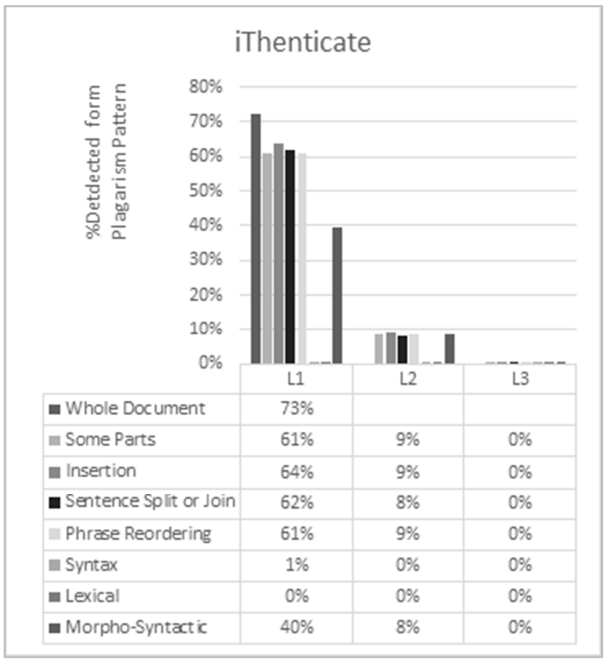
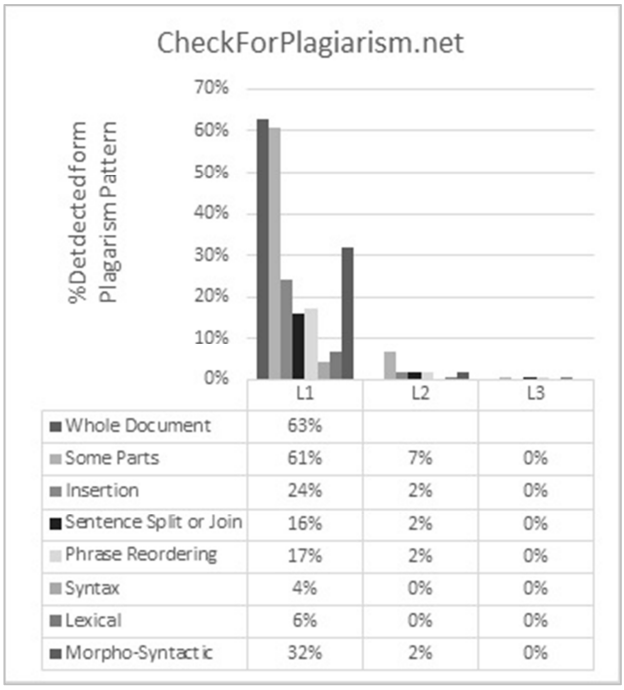


Figure 2. *Plagiarized detected percentage by iThenticate in each plagiarism patterns' levels.*

Figure 2 highlights the iThenticate checker detection percentage for each pattern's levels. The highest percentage from level 1 was 73% for whole document pattern. While the lowest percentage was for syntax pattern with 1% only. Although for level 2 the highest plagiarized percentage was 9% for some parts, insertion and phrase reordering patterns and the lowest was 0%. Thought iThanticate checker was

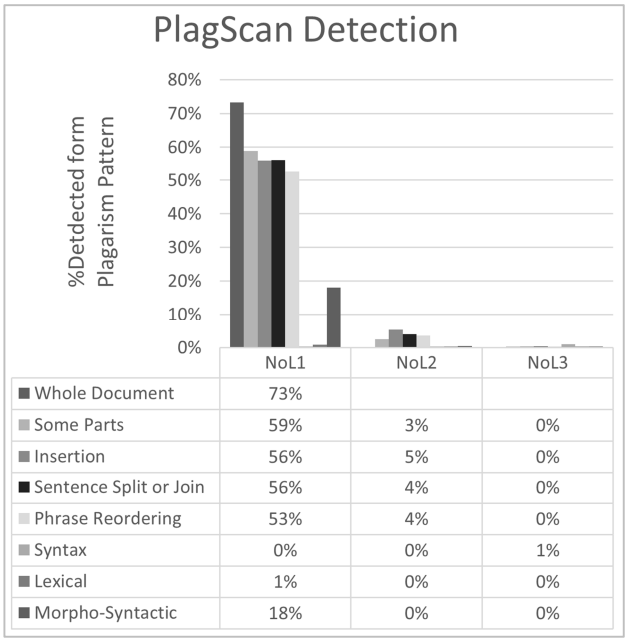
not able to detect any plagiarized text from level 3.



	L1	L2	L3
Whole Document	63%		
Some Parts	61%	7%	0%
Insertion	24%	2%	0%
Sentence Split or Join	16%	2%	0%
Phrase Reordering	17%	2%	0%
Syntax	4%	0%	0%
Lexical	6%	0%	0%
Morpho-Syntactic	32%	2%	0%

Figure 3. *Plagiarized detected percentage by CheckForPlagiarism.net in each plagiarism patterns' levels.*

Figure 3 shows the CheckForPlagiarism.net checker detection percentage for each pattern's levels. The highest percentage from level 1 was 63% for whole document pattern. While the lowest percentage was for syntax pattern with 4% only. Although for level 2 the highest plagiarized percentage was 7% for some parts pattern and the lowest was 0% for syntax and lexical patterns. Thought CheckForPlagiarism.net checker was not able to detect any plagiarized text from level 3.



	NoL1	NoL2	NoL3
Whole Document	73%		
Some Parts	59%	3%	0%
Insertion	56%	5%	0%
Sentence Split or Join	56%	4%	0%
Phrase Reordering	53%	4%	0%
Syntax	0%	0%	1%
Lexical	1%	0%	0%
Morpho-Syntactic	18%	0%	0%

Figure 4. *Plagiarized detected percentage by PlagScan in each plagiarism patterns' levels.*

Figure 4 shows the PlagScan checker detection percentage for each pattern's levels. The highest percentage from level 1 was 73% for whole document pattern. While the lowest percentage was for syntax pattern with 0% only. Although for level 2 the highest plagiarized percentage was 5% for insertion pattern and the lowest was 0% for syntax, lexical and moropho-syntactic patterns. Thought PlagScan checker was able to detect only 1% plagiarized text from level 3 for the syntax pattern.

From the result analysis above, assumptions on the effectiveness on each checker can be summarized. iThenticate are the most efficient checker for detecting plagiarism in Arabic text in almost all the patterns from level 1 except for syntax and lexical patterns. While none of the three checkers were able to detect the same plagiarized amount in level 2 or level 3 from all the patterns.

Our results were matched with other study that iThenticate is the most suitable online plagiarism checker for Arabic language [9]. However, the checker could not detect any plagiarized document with less than 10% from any plagiarism pattern. Though the highest plagiarism percentage detected between 30% and 10% was 9%. This outcome confirm that only literal patterns can be detected and only with a percentage more than 30%.

5. Conclusion and Future Work

In this paper performance comparisons were conducted between three plagiarism software checkers for Arabic language, iThenticate, CheckForPlagiarism.net and PlagScan. The goal was to measure the effectiveness of each checker in detecting the polarized text percentage impeded in each Arabic plagiarism patterns. Two comparisons were made; one to determine each checker ability to recognize the plagiarized document while the other was to identify how much percentage can be detected from each plagiarism pattern in each level.

Experiments have shown that iThenticate is the most effective online plagiarism checker in determining 100% to 50% from literal plagiarism patterns except for syntax pattern. Though CheckForPlagiarism.net is efficient in detecting intelligent plagiarism patterns which are only greater than 50%. However, for PlagScan even though was able to identify most of the plagiarism patterns but failed to determine the same plagiarism amount in each level.

The present findings confirm that even though the online plagiarism detection tools were able to detect plagiarism but it still could not detect the exact plagiarized amount in each level. Moreover, Syntax, lexical and morpho-Syntactic plagiarism patterns are difficult to recognize by the online plagiarism detection tools even if it is from level 1 plagiarism.

Therefore further consideration on expending the paper dataset and applying comparison on more of plagiarism intelligent patterns. Additionally, conducting comparison on standalone checkers and document to document plagiarism detection are also one of the issues that could be worked on in the future.

References

[1] Ali, A. M. E. T., Abdulla, H. M. D., & Snasel, V. (2011, May). Survey of plagiarism detection methods. In 2011 Fifth Asia Modelling Symposium (pp. 39-42). IEEE.

[2] Honig, B., & Bedi, A. (2012). The fox in the hen house: A critical examination of plagiarism among members of the Academy of Management. Academy of Management Learning & Education, 11 (1), 101-123.

[3] Plagiarism. (n. d.). turnitin. Retrieved from http:// www.turnitin.com/.

[4] Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42 (2), 133-149.

[5] IEEE (n. d). Plagiarism Levels and Corrective Actions, as taken from Section 8.2.4.D of the PSPB Operations Manual. Retrieved November 17, 2018 from https://www.ieee.org/content/dam/ieee.org/ieee/web/org/pubs/ Level_description.pdf.

[6] Abakush, I. (2016). Methods and tools for plagiarism detection in Arabic documents. In Sinteza 2016-International Scientific Conference on ICT and E-Business Related Research (pp. 173-178). Singidunum University.

[7] Beth Calvano (2012). Plagiarism in Higher Education Research. Retrieved from http://www.ithenticate.com /plagiarism-detection-blog/bid/87315/Plagiarism-in-Higher-E ducation-Research.

[8] Lukashenko, R., Graudina, V., & Grundspenkis, J. (2007, June). Computer-based plagiarism detection methods and tools: an overview. In Proceedings of the 2007 international conference on Computer systems and technologies (p. 40). ACM.

[9] Ghadah M. Adel, Abdullatif Ghallab (2014). Performance Comparisons on Online Plagiarism Detection Software in Arabic Theses. International Conference on e-Commerce, e-Administration, e-Society, e-Education and e-Technology (e-CASE & e-Tech 2014), Nagoya University, Japan.

[10] Naik, R. R., Landge, M. B., & Mahender, C. N. (2015). A review on plagiarism detection tools. International Journal of Computer Applications, 125 (11).

[11] Ali, A. M. E. T., Abdulla, H. M. D., & Snasel, V. (2011). Overview and Comparison of Plagiarism Detection Tools. In DATESO (pp. 161-172).

[12] Chow, T. W., & Rahman, M. K. M. (2009). Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection. IEEE Transactions on Neural Networks, 20 (9), 1385-1402.

[13] Osman, A. H., Salim, N., & Abuobieda, A. (2012). Survey of text plagiarism detection. Computer Engineering and Applications Journal (ComEngApp), 1 (1), 37-45.

[14] Ali, Y. A. A. (2018). A Model and Framework for Plagiarism Detection in Arabic Documents in Arabic Language (Doctoral dissertation, Sudan University of Science & Technology).

[15] Y. A. Abdelrahman, et al., "A Method For Arabic Documents Plagiarism Detection," International Journal of Computer Science and Information Security, vol. 15, p. 79, 2017.

[16] Alzahrani, Salha., Salim, Naomie (2008). Plagiarism detection in arabic scripts using fuzzy information retrieval, In Student Conference on Research and Development Student, 281, p. 1– 4, 2008.

[17] Menai, M. E. B., & Bagais, M. (2011, August). APlag: A plagiarism checker for Arabic texts. In 2011 6th International Conference on Computer Science & Education (ICCSE) (pp. 1379-1383). IEEE.

[18] Khorsi, A., Cherroun, H., & Schwab, D. (2018). A Two-Level Plagiarism Detection System for Arabic Documents. Cybernetics and Information Technologies, 20.

[19] Ashraf S Hussein. A plagiarism detection system for arabic documents. In Intelligent Systems' 2014, Springer International Publishing, 2015. p. 541-552.

[20] Ceska, Z. (2008, August). Plagiarism detection based on singular value decomposition. In International Conference on Natural Language Processing (pp. 108-119). Springer, Berlin, Heidelberg.

[21] Bull, J., Colins, C., Coughlin, E., & Sharp, D. (2007). Technical review of plagiarism detection software report.

[22] Weber-Wulff, D. (2014). False feathers: A perspective on academic plagiarism. Springer Science & Business.

[23] McCabe, D. L. (2005). Cheating among college and university students: A North American perspective. International Journal for Educational Integrity, 1 (1).