ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse201942020.pdf
https://doi.org/10.30534/ijatcse/2020/201942020

Towards the Enhancement of Text Plagiarism Detection Effectiveness: Experimental Study

Waseem Alromema[1], Essa Abdullah Hezzam[2]
[1]Dept. of Computer Science and Information,Taibah University, Medina, Saudi Arabia, wromema@taibahu.edu.sa
[2]Dept. of Information Systems,College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia, ehezzam@taibahu.edu.sa

ABSTRACT

Due to the need of discovering the originality of academic works. As a result of massive insufficient plagiarism information, many academic researchers need to check their documents for plagiarism. Most available Plagiarism Detection (PD) tools start the detection process with a preprocessing stage. However, some of the PD tools are fooled by some misleading punctuation marks, such as double quotation. This paper proposes a framework for the enhancement of PD based on document cleaning regardless of the preprocessing methods adopted and the PD methodology being used. An experiment conducted by checking the plagiarism on a dataset of research articles, collected from the Internet, using iThenticate and the proposed method, the proposed method showed an improvement percentage of 68% over the traditional method.

Key words: Plagiarism, detection, Natural Language, pre-processing, Limitations, text cleaning, iThenticate.

1. INTRODUCTION

According to Oxford Dictionary [1], plagiarism can be defined as "The practice of taking someone else's work or ideas and passing them off as one's own". The vast amount of material published on the Internet makes it easy to use and copy text without citing the reference, especially in the academic field. However, huge effort was paid for plagiarism detection and even plagiarism prevention systems.

Plagiarism Detection (PD) is one of the Natural Language Processing (NLP) applications that aims to recognize unethically reused text. Text plagiarism is classified into extrinsic (i.e. external) and intrinsic (i.e. internal) [2,3]. External plagiarism detection aims to compare a suspicious document with a stored database (called source documents or reference database), while intrinsic plagiarism detection aims to trace the writing style within the same document. Many methodologies have been proposed for PD in the literature, including machine learning algorithms, citation and mathematical content analysis, and similarity-based techniques. Most PD techniques, regardless of the method, start the detection process with a preprocessing stage. The preprocessing stage usually includes text tokenization, stop-words and numbers removal, stemming, Part of Speech Tagging (POST), etc. After preprocessing (suspicious and/or source documents), PD tools employ some methodology for detecting either extrinsic or intrinsic plagiarism cases. For detecting extrinsic plagiarism cases, a PD tool aims to find the similarity between the suspicious document and the reference database using many techniques, such supervised classification algorithms, Cosine similarity measure, Jaccard similarity measure, etc. The similarity detection phase can be semantic-based, considering synonyms replacement using external resources such as lexical thesaurus, or term co-occurrence-based [4] [5]. For detecting the intrinsic plagiarism, a PD tool aims to recognize the variation in writing style within the suspicious document, this is done by extracting some linguistic features that best reflect the author's writing style [6].

On the other hand, the over quoting is a straight replication of the exact text from the main source. The quotations types, namely, direct and indirect. Direct quotations use the exact language. Indirect quotations do not use the exact phrasing from the source, this gives limitations such as failure to process textual images for match checks [7] [8]. Therefore, there are limitation of over quoting PD using common PD tools such as iThenticate.

There are many tools [9-13] available on the internet that are used for plagiarism detection such as iThenticate, Turnitin, Dupli Checker, Plagiarism Checker, plagiarism detect, Plagiarisma.net, Eve Plagiarism Detection System, Plagiarism.org, Copy catch.com, heck for plagiarism, Essay Verification Engine. Therefore, in this study we used the iThenticate [9] as a case study, whereas iThenticate is the most trusted plagiarism detection and the first tool used by reviewers, researchers and expert writers to check their original works for possible plagiarism.

Although the preprocessing stage generally aims to eliminate unnecessary tokens that may be of no importance to the detection process, this stage may play a crucial role in the detection process and therefore, the results accuracy. Based on this idea, more attention must be paid for this elementary stage of PD.

In this paper, we propose a framework with an additional phase, namely, text cleaning, in which the suspicious document is passed before tokenization. The text cleaning phase aims to remove misleading punctuation marks, namely, double quotation. Most PD tools skip text within double quotation

from the comparison process, since quoting text (i.e. borrowing text within double quotation referring to its source) is acceptable and cannot be considered a plagiarism case.

The paper organized as follows: Section 2 explains the related work. Section 3 presents the proposed framework for enhancing PD. Section 4 discusses the experiments and discussion. Finally, section 5 concludes the paper.

2. RELATED WORKS

Many studies have been presented for PD and many methods have been proposed for PD improvement in terms of the detection methodology itself, but few were dedicated for the enhancement of the PD results in terms of the preprocessing stage, using traditional techniques such as POST and stemming.

In this section, we separately review the related literature regarding the plagiarism detection systems, the recent problem to detect the plagiarism based on the preprocessing/text cleaning document, and the solutions.

In [4], the author proposed a framework solution of the search engine to search for discretized and discretized-less Arabic text using query expansion techniques. The query expansion has been applied using Quran related limited thesaurus. This thesaurus contains 100 semantic groups, where each group consists of 3 to 6 synonyms, and used 40 diacritic-less queries obtained from Arabic native speakers. The authors concluded the query expansion for searching Arabic text is promising and it is likely that advanced NLP tools can further improve the efficiency. As far as we know, in the above context, the analysis of pre-processing and text cleaning for documents in terms of the plagiarism detection system is needed, as a preprocessing phase. Ghanem et al [14] have removed diacritics, non-Arabic letters, numbers, and words consisting of only one letter. Then named entities were extracted. As the proposed system, called HYPLAG, is said to detect both verbatim and rephrasing plagiarism cases, the detection process continues to apply stemming, part of speech tagging, and synonyms replacement. HYPLAG adopted a hybrid approach of corpus-based and knowledge-based PD approaches. The detection process starts by chunking both suspicious and original documents into n-terms sentences, then Arabic WordNet was used for extracting synonyms of the suspicious document, then the original sentences are ranked according to their similarity with those of the suspicious ones. Sentences with highest rank are classified as candidates for plagiarism and proceed to similarity comparison with the suspicious sentences using Vector Space Model (VSM) and TF*IDF weighting scheme. Finally, similarity results are classified into plagiarism cases or another phase of feature-based semantic similarity measurement according to a predefined threshold. HYPLAG was said to have 89% success rate. in addition, paper [15] presented a mechanism for using the information retrieval system in the process of plagiarism detection

Another PD system proposed for Arabic text by Khorsi et. al in [16], called A Two-Level Plagiarism Detection System (2L-APD), is said to detect both verbatim and rephrasing plagiarism cases. The system adopted two modules of detection, namely fingerprinting and word embedding, where Jaccard and cosine similarity measures were applied respectively. The preprocessing phase, which falls in the first module, includes tokenization, diacritics, non-letters, and stopwords removal, and finally, lemmatization using a tool called MADAMIRA. The experimental results showed an overall precision rate of (85%) and a recall rate of (87%) on ExAraDet-2015 corpus.

Cherroun et al in [17] have presented two approaches, word embedding and machine learning, for detecting different plagiarism cases in Arabic text. The proposed approaches are said to detect more disguised plagiarism cases than verbatim ones. The word embedding approach mainly employed the vector space model for measuring similarity among the suspicious and original sentences with the use of Term Frequency (TF) weighting scheme and POST. While the machine learning-based approached employed a set of supervised learning algorithms, namely, Support Vector Machine (SVM), Decision Trees (DT) and Random Forests (RF) for detecting the plagiarism cases. The preprocessing step is proceeded by sentence-level segmentation and consisting of removing diacritics and non-alphanumeric characters, normalizing. A set of experiments were conducted using a dataset called Tr-EXARA-2015 corpus, and the results are as follows: the best precision and recall by SVM classifier are about 0.89 and 0.92 respectively, while the best precision and recall values achieved by the RF classifier are about 0.86 and 0.82 respectively, and the best precision and recall values achieved by the DT classifier are about 0.91 and 0.85 respectively. The best precision value achieved by the word embedding approach is about 0.89, while the best recall value achieved is about 0.88 [17].

In [19], the authors have proposed a parallel cross-language PD system. Fuzzy semantic similarity among words was measured using two WordNet-based similarity measures. The proposed system is said to detect different types of plagiarism, including translation plagiarism, in Arabic and English texts. The studied documents were segmented into tri-grams and for preprocessing, tokenization, stopwords removal, lemmatization, and POST were applied. The parallelism came from the use of three Big Data technologies, namely, Apache Hadoop framework, Hadoop Distributed File System (HDFS), and MapReduce programming model. Experimental results showed best precision and recall for Fuzzy-Wup equal to 0.54 and 0.66 respectively.

Zaher et al in [6] have proposed an unsupervised model for PD in Arabic documents, called ASTAP. The proposed model is said to handle both handwritten and electronic Arabic documents, by using an Optical Character Recognizer (OCR) for converting the handwritten to the electronic form. ASTAP works as follows: (1) The detection process starts by preprocessing text, which includes tokenization, stopwords removal, stemming, and synonyms replacement (2) Then queries are generated and submitted to be searched for over the web (3) The retrieved documents are then represented in a document-tree-structure, where the tree root is the whole document, the next level stands for the paragraphs, and the leaves level stands for the sentences (4) finally, the suspicious document is compared for similarity with the retrieved documents on all trees' levels. The best precision value

reported by experimenting ASTAP on three datasets is 0.75. Also, for the Arabic document presented in the paper [21].

Machine learning was employed in [20] for detecting paraphrasing in Arabic documents. A vocabulary corpus of words' synonyms is said to be built with the use of POST and Word2Vec representation. The vocabulary corpus creation required a preprocessing phase that includes removing diacritics, extra white space, titles numeration, punctuation marks, special characters, duplicated letters and non-Arabic words. Global vector representation is used for the extracted features, and Convolutional Neural Network (CNN) is used for classifying documents are into plagiarized or non-plagiarized. experimental results showed 0.8 precision and 0.82 recall. Based on the reviewed studies, the existing enhancements of the PD results basically relies on the use of common techniques such as POST, stemming, synonym replacement, etc. Therefore, we come up with a framework for the enhancement of the whole PD process in terms of PD results.

3. THE PROPOSED MODEL FOR ENHANCING PD

This section introduces an architecture for enhancing text PD effectiveness and discovering the attempted misleading and decreasing the rate of plagiarism through the use of double quotation by researchers. Figure 1 shows the proposed methodology, which consists of three stages: Document Preprocessing, PD Process and Plagiarism Results. Document pre-processing starts with text cleaning, then proceeds to tokenizing, and indexing the suspicious document. While in PD Process, the core PD process takes place. Finally, the total percentage of plagiarism calculated in the stage of Plagiarism Results. These three stages are discussed below.

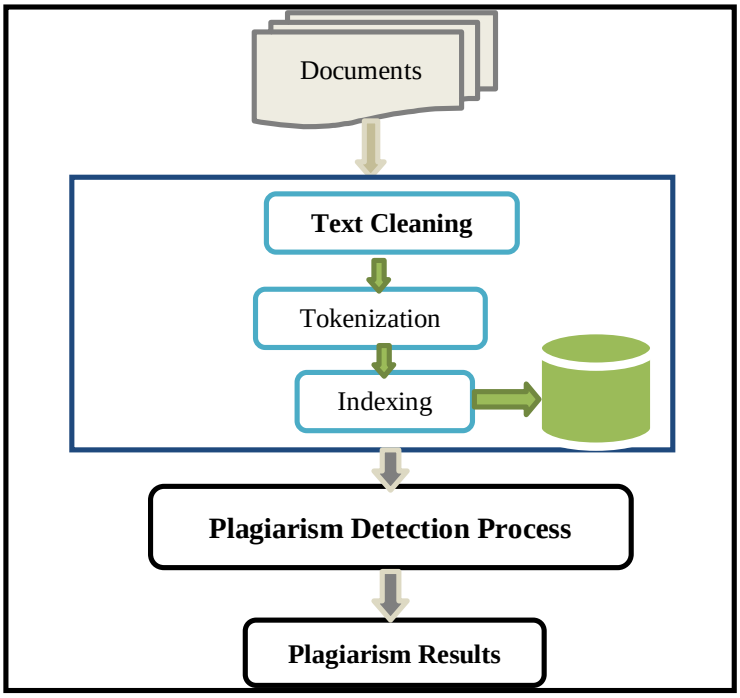

Figure. 1: Process of PD System

Stages of the Proposed Model

1. Document Preprocessing: In this stage, the incoming suspicious document is preprocessed before moving to the second stage, it consists of: (a) text cleaning, (b) tokenizing, and (c) indexing.

(a) Text cleaningis the main proposed idea in this study, which aims to scan the suspicious document, dedicated for removing double quotation (“”)if it has been found. The text cleaning step aims to improve the PD results regardless of the other preprocessing steps adopted and the methodology being used. In some NLP applications, such as Information Retrieval (IR) systems, double quotation plays an important role of retrieving required exact matches [2][4].Nevertheless, in PD systems, the text included within double quotation is usually skipped from the text comparison process. Ignoring such text may affect the performance of a PD system, since quoting is an acceptable behavior and does not indicate plagiarism. For example, iThenticate [9] successes to detect the following intentionally copied sentence within a suspicious document such as the paragraph below as an example to show the case:

The approach adopted here is to consider cognitive structure from a conceptual heuristic standpoint, which differentiates memory

While it failed to detect it as plagiarized when we added the double quotation to the same previous paragraph as follows:

“*The approach adopted here is to consider cognitive structure from a conceptual heuristic standpoint, which differentiates memory***”**

The previous example addresses how canbe fooling automated PD tools using double quotation by people committing plagiarism.

(b) Tokenization: is a mandatory step in almost all NLP applications that can be defined as the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens [14]**.** The resulting tokens include stop-words, defined by Lo, et al. as a varying list of meaninglessly frequent words, numbers, letters from a different language, punctuations and special characters, which are usually eliminated from the text [18].

(c) Indexing module aims to index the tokens by saving each token along with its weight based on some weighting scheme.

2. Inthe PD Process,the PD system deploys the detection methodology, such as string matching, classification algorithm, etc., for capturing copied text.
3. In Plagiarism Results stage the framework ends up by presenting a percentage of total copying in the suspicious text, usually with the source document from which the text was plagiarized.

4. EXPERIMENTS AND DISCUSSION

This section shows the results of the experiment conducted on two scenarios, the existing method of current PD tool (iThenticate) and proposed method for plagiarism detection. In both scenarios, the dataset used a set of research articles collected from the web written in English language. The well-known PD tool has been used in order to achieve and evaluate the results of the above mentioned two scenarios, with using ten research articles as dataset.

Scenario 1:

The dataset is passed to iThenticate. Table 1 shows the results of existing methods of current PD tool.

Table 1: Experimental Results of existing methods of current PD tool

Doc. #	Plagiarism Percentage in Doc.# with double quotations "unclean text" (existing method)
Doc.1	10%
Doc.2	12%
Doc.3	14%
Doc.4	6%
Doc.5	16%
Doc.6	12%
Doc.7	15%
Doc.8	16%
Doc.9	14%
Doc.10	12%

Table 1 demonstrates the results of unclean text in documents using double quotation. Therefore, the average results of plagiarism for ten articles is 13%. This way the research articles could be accepted when submitted for publication.

Scenario 2:

The proposed PD method is applied to the same dataset used for scenario 1. Table 2 shows the results of this experiment. In addition, based on the experimental results, the proposed approach succeeds to enhance the PD results, applied on iThenticate as a case study. For instance, testing Doc. # 1 without double quotation using iThenticate for plagiarism check shows a total percentage of plagiarism= 37%. However, when testing the same document (i.e. Doc.#1) with double quotation, the total percentage of plagiarism decreased to 10%. Table 2 shows the results of this experiment.

Table 2. Experimental Results of PDproposed method

Doc. #	Plagiarism Percentage in Doc.# without double quotations "clean text" (the proposed method)
Doc.1	37%
Doc.2	67%
Doc.3	45%
Doc.4	24%
Doc.5	44%
Doc.6	35%
Doc.7	42%
Doc.8	38%
Doc.9	45%
Doc.10	39%

Table 2 shows the average of plagiarism for ten documents is 42% without using the double quotation. With this higher similarity the papers will not be accepted for publication compared to the results in senario1. Additionally, the average improvement can be seen in Figure 2. Using the proposed method will enhance the quality of scientific publications. The average of improvement can be seen significantly by 68.86%.

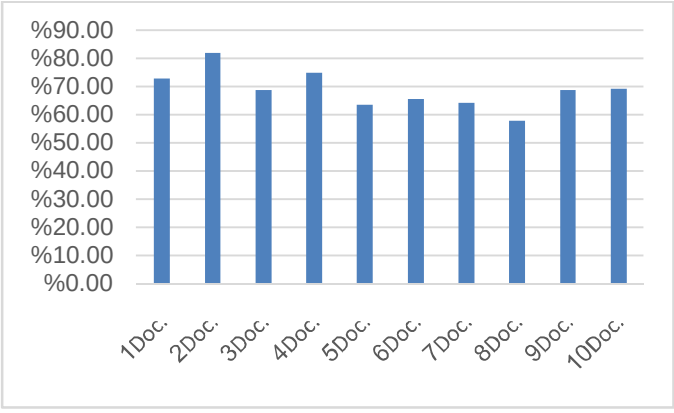

Figure 2: Average 68.86% improvement using the proposed method

5. CONCLUSION

This paper proposed a framework with a new approach of preprocessing text that aims to enhance the PD results. The proposed approach, namely, text cleaning is dedicated for the elimination of double quotation in the suspicious text. Many existing PD tools can be fooled by the double quotation as ethical quoting, but some plagiarizers use the double quotation, even writing in white color, to fool the automatic PD tools. The proposed approach aims to improve the PD results regardless of the adopted PD methodology. In the experiments, iThenticate has been used as an assessment tool for testing the existing and the proposed approach. The experimental results show the effectiveness of the proposed method in the enhancement of PD results.

REFERENCES

[1] https://en.oxforddictionaries.com/definition/plagiarism. Accessed Date [May-2020].

[2] AlSallal, M., Iqbal, R., Palade, V., Amin, S., & Chang, V. (2019). An integrated approach for intrinsic plagiarism detection. Future Generation Computer Systems, 96, 700-712.

[3] Gupta, D. (2016). Study on Extrinsic Text Plagiarism Detection Techniques and Tools. Journal of Engineering Science & Technology Review, 9(5). https://doi.org/10.25103/jestr.095.02

[4] Hammo, B.H. (2009). Towards enhancing retrieval effectiveness of search engines for diacritisized arabic documents. Information retrieval 12(3), 300–323

[5] Moawad, I., Alromima, W., & Elgohary, R. (2018). Bi-gram term collocations-based query expansion approach for improving Arabic information retrieval. Arabian Journal for Science and Engineering, 43(12), 7705-7718.

[6] Zaher, Mahmoud, Abdulaziz Shehab, Mohamed Elhoseny, and Farahat Farag Farahat. "Unsupervised Model for Detecting Plagiarism in Internet-based Handwritten Arabic Documents." *Journal of Organizational and End User Computing (JOEUC)* 32, no. 2 (2020): 42-66.

[7] Use of Quotations in Writing: Types of Quotations, https://beanerywriters.wordpress.com/2010/06/11/use-of-quotations-in-writing-types-of-quotations/. Accessed Date [jun-2020]

[8] Seadle, M. 2008. Copyright in the networked world: plagiarism and its ambiguities, Library Hi Tech, Vol. 26, Iss: 4, pp.691-695

[9] iThenticate plagiarism system, https://app.ithenticate.com/ Accessed Date [Jun-2020]

[10] turnitin detection system, www.turnitin.com

[11] Duplichecker, http:/www.duplichecker.com

[12] PlagiarismChecker.com", http://wwwplagiarismchecker.com/helpteach

[13] Plagiarism detector", http://www.plagiarism-detector.com/

[14] Ghanem, Bilal, Labib Arafeh, Paolo Rosso, and Fernando Sánchez-Vega. "HYPLAG: Hybrid Arabic Text Plagiarism Detection System." In International Conference on Applications of Natural Language to Information Systems, pp. 315-323. Springer, Cham, 2018 https://doi.org/10.1007/978-3-319-91947-8_33

[15] Hagen, M., Potthast,M., and Stein, B. Source Retrieval for Plagiarism Detection from Large Web Corpora: Recent Approaches. Proc. of CLEF 2015 Labs and Workshops, Notebook Papers, 8-11 September, Toulouse, France (2015).

[16] Khorsi, Ahmed, Hadda Cherroun, and Didier Schwab. "2L-APD: A two-level plagiarism detection system for Arabic documents." *Cybernetics and Information Technologies* 18, no. 1 (2018): 124-138.

[17] Cherroun, Hadda, and Ali Alshehri. "Disguised plagiarism detection in Arabic text documents." In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pp. 1-6. IEEE, 2018.

[18] Lo, Rachel Tsz-Wai, Ben He, and Iadh Ounis. "Automatically building a stopword list for an information retrieval system." In *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, vol. 5, pp. 17-24. 2005

[19] Ghanem, Bilal, Labib Arafeh, Paolo Rosso, and Fernando Sánchez-Vega. "HYPLAG: Hybrid Arabic Text Plagiarism Detection System." In International Conference on Applications of Natural Language to Information Systems, pp. 315-323. Springer, Cham, 2018 https://doi.org/10.1007/978-3-319-91947-8_33

[20] Mahmoud, Adnen, and Mounir Zrigui. "Distributional Semantic Model Based on Convolutional Neural Network for Arabic Textual Similarity." *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 14, no. 1 (2020): 35-50.

[21] Emad Al-Shawakfa et al., International Journal of Advanced Trends in Computer Science and Engineering, 9(1), January – February 2020, 98 – 109. https://doi.org/10.30534/ijatcse/2020/16912020