Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Computer Science

Designing and Implementing Intelligent Textual Plagiarism Detection Models

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Doctor of Philosophy in
Computer Science**

By

Ayoub Ali Mohammed Saeed

Supervised by

Asst. Prof. Dr. Alaa Yaseen Taha

2023 A.D. **1444 A.H.**

Abstract

Plagiarism is known as presenting other works as one's own without making proper citation or giving an explicit acknowledgment. The detection of plagiarism is interesting because it has become a significant topic in the ethics of scientific research especially in an academic environment.

This thesis has proposed two models for detection of textual plagiarism. The first model is called "Textual Plagiarism Detection Model" (TPDM) and it involves two key stages: Mini Batch Kmeans clustering algorithm and web scrape technique can be utilized to retrieve the candidate source documents in the first stage. The second stage starts with pre-processing and segmenting the texts of suspicious and source documents utilizing Natural Language Processing (NLP) techniques, then four proposed algorithms are used. The first is Exact Plagiarism Detection (EPD) algorithm which detects the exact copy-paste plagiarism using the Jaccard similarity measure; the second is the Lexical Plagiarism Detection (LPD) algorithm, which detects lexical changes in the source text using Term Frequency-Inverse Document Frequency (TF-IDF) and cosine similarity measure. The semantic changes are detected through Semantic Plagiarism Detection (SPD) algorithm utilizing pre-trained Deep Learning (DL) based Universal Sentence Encoder (USE) model and cosine similarity measure. Finally, the Merged Lexical-Semantic Plagiarism Detection (MLSPD) algorithm is utilized for detecting a lexical-semantic change utilizing both (and) and (or) operators.

The second proposed model is known as "Weight based Plagiarism Detection Model" (WTPDM) . It depends on the assigned weight values for each section of suspicious document. Like the first model, it involves four proposed algorithms. Same Weight Lexical Plagiarism Detection (SWLPD), Same Weight Semantic Plagiarism Detection

(SWSPD), Variant Weight Lexical Plagiarism Detection (VWLPD) and Variant Weight Semantic Plagiarism Detection (VWSPD). These algorithms are used to detect the plagiarism in two cases, if assigned weight values are the same for all sections of suspicious text or that assigned weights have different values.

After conducting several experiments on both models, the MLSPD with (or) operator algorithm gives the highest average of similarity ratios with 30.3% and 458.49 seconds as average of running time while EPD algorithm elapsed 8.212 seconds as the lowest running time average.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

تصميم وتنفيذ نماذج ذكية لاكتشاف الاستلال النصي

اطروحة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة دكتوراه فلسفة في
علوم الحاسوب

من قبل
أيوب علي محمد سعيد

بإشراف
أ.م. د. الاء ياسين طه

١٤٤٤هـ ٢٠٢٣م

الخلاصة

يُعرف الاســـتلال او الســـرقة العلمية بأنه تقديم أعمال باحثين اخرين على أنها أعمال خاصـــة بالباحث الحالي دون الإشـــارة الى المصـــدر الأصـــلي لهذه الاعمال. ان اهمية اكتشـــاف الاستلال النصـي ازدادت في الآونة الاخيرة لأن الاستلال أصـبح جزءًا مهمًا من أخلاقيات البحث العلمي خاصـــة في البيئة الأكاديمية. اقترحت هذه الأطروحة نموذجين للكشـــف عن الاســـتلال النصي.

النموذج الأول يسمى "نموذج الكشف عن الاستلال النصي" TPDM والذي يتكون من مرحلتين رئيسيتين: في المرحلة الأولى، يمكن استخدام خوارزمية العنقدة Mini Batch Kmeans أو تقنية قراءة البيانات من الويب web scrape لاسترداد مستندات المصدر ذات العلاقة بالمستند الذي يتم اختباره. تبدأ المرحلة الثانية بالمعالجة المسبقة وتجزئة نصوص المستندات قيد الاختبار ومستندات المصدر باستخدام معالجة اللغات الطبيعية (NLP)، تم استخدام أربعة خوارزميات مقترحة. الأولى هي خوارزمية كشف السرقة الأدبية (EPD) التي تكتشف الاستلال الدقيق (النسخ واللصق) باستخدام مقياس التشابه Jaccard ، الثانية هي خوارزمية الكشف عن الاستلال المعجمي (LPD) التي تكتشف التغييرات المعجمية في نص المصدر بأستخدام طريقة الترميز النصي تكرار المفردات – معكوس تكرار المستندات TF-IDF ومقياس الجيب تمام cosine للتشابه النصي. بينما يتم اكتشاف التغييرات الدلالية من خلال الكشف عن الاستلال الدلالي (SPD) باستخدام نموذج (USE) المدرب مسبقًا المبني على التعلم العميق (DL) لترميز النصوص وكذلك ومقياس الجيب تمام cosine للتشابه النصي. أخيرًا، تم استخدام خوارزمية اكتشاف الاستلال المعجمي–الدلالي المدمجة (MLSPD) لاكتشاف التغييرات المعجمية الدلالية في حالتين باستخدام علاقتي الربط (و) و (أو).

النموذج الثاني المقترح يسمى "نموذج الكشف عن الاستلال المبني على الوزن" WTPDM والذي يعتمد على الوزن المخصص لكل مقطع نصي في المستندات قيد الاختبار. كما في النموذج الأول يتضمن هذا النموذج أيضا أربع خوارزميات مقترحة. خوارزمية كشف الاستلال المعجمي بنفس الاوزان (SWLPD) ، خوارزمية كشف الاستلال الدلالي بنفس الاوزان (SWSPD) ، خوارزمية كشف الاستلال المعجمي بأوزان مختلفة (VWLPD) وخوارزمية الكشف عن الاستلال الدلالي بأوزان مختلفة (VWSPD). تستخدم هذه الخوارزميات لاكتشاف الاستلال في حالتين، إذا كانت قيم الاوزان هي نفسها لجميع أقسام النص قيد الاختبار أو أن هذه الأوزان لها قيم مختلفة.

بعد إجراء العديد من التجارب على كلا النموذجين، سجلت خوارزمية MLSPD التي تستخدم معامل (أو) أعلى معدل تشابه نصي بنسبة ٣٠,٣٪ وبمعدل وقت تنفيذ ٤٥٨,٤٩ ثانية بينما حصلت خوارزمية EPD على ٨،٢١٢ ثانية كاقل معدل وقت التنفيذ.