ISSN: 0067-2904

Mining Deviations in Document Writing Style through Vector Dissimilarity

Nasreen J. Kadhim
Department of Computer Science/ College of Science, University of Baghdad, Baghdad, Iraq

Received: 10/2/2023 Accepted: 14/4/2023 Published: 30/4/2024

Abstract

Doubts arise about the originality of a document when noticing a change in its writing style. This evidence to plagiarism has made the intrinsic approach for detecting plagiarism uncover the plagiarized passages through the analysis of the writing style for the suspicious document where a reference corpus to compare with is absent.

The proposed work aims at discovering the deviations in document writing style through applying several steps: Firstly, the entire document is segmented into disjointed segments wherein each corresponds to a paragraph in the original document. For the entire document and for each segment, center vectors comprising average $tf - idf$ weight of their word $n - grams$ are constructed. Second, the degree of closeness is calculated through applying Cosine similarity to measure for each segment, the deviation of its center vector from the center vector of the entire document. Additionally, word n-gram length will be investigated to show its effect on the proposed system performance wherein, center vectors are computed considering word n-grams for different values of n (n= 1, 2, and 3).
Performance evaluation of the proposed method was accomplished through the use of Precision, Recall, F-measure, Granularity, and Plagdet as evaluation measures. Moreover, PAN-PC-09 and PAN-PC-11 were used for detecting intrinsic plagiarism as evaluation corpora. It is shown that the proposed approach has achieved results that are comparable to the state-of-the-art methods. Positive impact was observed through discovering deviations in document writing style by computing weight vectors dissimilarity rather than calculating the difference between the word n-grams that exist in segments and their corresponding word n-grams in the suspicious document. Furthermore, when considering the length of word n-gram, better results were recorded for system performance when word bi-grams was used compared to word uni-grams and word tri-grams.

Keywords: Intrinsic plagiarism detection, center vector dissimilarity, Cosine similarity, n-gram length.

التنقيب عن الانحرافات في أسلوب كتابة المستند عبر اختلاف المتجهات

نسرين جواد كاظم
قسم علوم الحاسوب,كلية العلوم,جامعة بغداد

*Email: nasreen.kadhim@sc.uobaghdad.edu.iq

الخلاصة

تظهر الشكوك حول أصالة المستند عند ملاحظة تغيير في أسلوب كتابته. هذا الدليل على الانتحال قد جعل النهج الجوهري للكشف عن الانتحال يكتشف المقاطع المسروقة من خلال تحليل أسلوب الكتابة للوثيقة المشبوهة في حالة غياب لأي مجموعة مرجعية للمقارنة معها.

يهدف العمل المقترح إلى اكتشاف الانحرافات في أسلوب كتابة المستند من خلال تطبيق عدة خطوات: أولاً ، يتم تقسيم المستند بأكمله إلى مقاطع منفصلة حيث يتوافق كل منها مع فقرة في المستند الأصلي. بالنسبة للمستند بأكمله ولكل مقطع ، يتم إنشاء متجهات المركز التي تشتمل على متوسط وزن tf–idf لل n–grams الموجودة فيهم. بعد ذلك ، يتم حساب درجة الأختلافات التقارب من خلال تطبيق تشابه جيب التمام لقياس انحراف متجه المركز لكل مقطع عن المتجه المركزي للمستند بأكمله. كما سيتم فحص طول الكلمة n–gram لإظهار تأثيرها على أداء النظام المقترح. يتم حساب متجه المركز مع الأخذ في الاعتبار كلمة n–grams لقيم مختلفة لـ n (n = 1, 2 and 3).

تم تقييم أداء العمل المقترح من خلال مقاييس Precision, Granularity, Recall, F–measure, Plagdet. بالإضافة الى ذلك ، تم تطبيق الطريقة المقترحة على PAN–PC–09 و PAN–PC–11 للكشف عن السرقة الأدبية الذاتية، من خلال النتائج المتحققة للنهج المقترح يتضح أنه من الممكن مقارنتها بأحدث الأساليب. حيث لوحظ تأثير إيجابي من خلال اكتشاف الانحرافات في أسلوب كتابة المستندات من خلال حساب درجة اختلاف متجهات الوزن عما تحقق عبر حساب الفرق بين n–grams الموجودة في المقاطع وال n–gramsالمقابلة لها في المستند المشبوه. علاوة على ذلك ، عند أخذ طول كلمة n–gram بنظر الاعتبار، قد تم تسجيل تفوق في أداء النظام عند استخدام bi–gramsعلى أستخدامuni–grams وtri–grams.

1. Introduction

In the modern world, a major problem that mainly affects education and research is *textual plagiarism*, which refers to the unacknowledged use of another author's work either as an exact copy or a version with a slight modification 1,2]. Easy access for anyone to Billions of web pages that came from the speedy development of the World Wide Web (WWW) has provided plenty of possible sources for plagiarism. In view of that, analysis and detection of automated plagiarism discovery have gained an increasing attention in both software industry and academia [3]. This phenomenon inspired several authors to attempt describing it [4, 5].
For detecting plagiarism, two main strategies have been considered by researchers [6]: The first is when no reference exists to compare with .The aim is to discover plagiarism through examining the input document only and giving a decision whether portions of it are not written by the same author. This strategy is called intrinsic plagiarism detection. Whereas the second strategy, which is referred to as detecting external plagiarism, comparing suspicious documents is achieved taking in consideration a collection of sources for the recognition of plagiarized segments.

For the traditional intrinsic plagiarism detection [7]: The task is to determine whether the suspicious document comprises plagiarized sections or a single author has written it. The suspicious documents is not compared against external sources in the detection process. There exists a crucial condition in the setting of traditional intrinsic plagiarism: At least 70% of the considered document was written by one main author. Thus, the common structure for the detection of internal plagiarism, which has been designated by the «one-main-author» condition [8, 9, 1, 10, 6] is: Firstly, dividing a document comprising text into a set of segments. Second, for each segment, a set of features is recognized and combined to refer to an author style. function to measure its correspondence to an author-style. Finally, values that are identified as critical that exist in author style function are used for discovering the plagiarized divisions.

For text-based data, outlier detection techniques have been used for developing internal plagiarism detection strategies by means of deviation parameters regarding the writing style of a given document. The authors in [8] proposed a *sliding window* approach; where a text document is separated into a set of intersecting divisions and as key component of an author style function, character 3-gram frequencies were used. The supplementary well-thought-out examples of style function are the n-gram classes [4], pronouns, punctuation, and part-of-speech tags count [9], and normalized word frequency class [1]. A style function counting an n-gram frequency relative deviation from its typical value was proposed.

For this proposed work, suspicious document are separated into disjointed segments considering the original paragraphs exist. Also, a different representation is put forward wherein the document has been represented as a weight vector demonstrating its main content. Next, a relative deviation is computed for an n-gram average weight from its representative value through building a style function.

In this paper, the proposed approach is illustrated through the following steps: Firstly, segmenting the entire document into disjointed segments, where each of them corresponds to a paragraph in the original document. For the entire document and for each segment, center vectors involving average $tf - idf$ weight of their word $n - grams$ are constructed. Second, the degree of closeness is calculated through applying Cosine similarity to measure. For each segment, the deviation of its center vector from the center vector of the entire document. Also, word n-gram length was investigated to show its effect on the proposed system performance.
This paper is organized as follows: Section 2 introduces the works related to the proposed work. Section 3 presents a detailed description for the proposed intrinsic plagiarism detection approach. Performance evaluation measures are discussed in Section 4. Section 5 presents performance evaluation results for the proposed system, as well as the comparison results with state-of-the-art methods. Finally, the conclusions and future work directions are provided in Section 6.

2. Related work

Once a text is compared to a reference set of possible sources, the difficulty to decide on the true set of documents for performing comparison will arise. Moreover, the opportunities brought to plagiarists through the Internet makes accomplishing this task more complex. Hence, writing style analysis can be achieved within the document, and the inconsistencies can be examined. The key aspect that provides an indication of plagiarism is describing a criterion to conclude if a significant change has occurred to the writing style. Achieving text style analysis and complexity can be constructed on certain parameters, such as text statistics, syntactic features, structural features, part-of-speech features, and closed-class word sets, as stated by the author of [11].

In [8], a method for detecting intrinsic plagiarism was presented. Through the use of a suitable dissimilarity measure, a style variation function was constructed. The function was firstly proposed for character n-gram profiles and author identification and attempted for quantifying the style change within a document. A sliding window was initially applied for constructing style profiles. Also, for constructing those profiles, character n-grams were proposed. The objective behind using n-grams was to acquire information about the writer's style, then analyzing profile deviations was performed to determine if a significant change happened to give an indication of another author's style [8].

As a technique for extracting structural information of text to detect intrinsic plagiarism, the authors of [12] introduced Kolmogorov Complexity measures. To detect style changes in a document, they performed an investigation with complexity features constructed on the Lempel-Ziv compression algorithm, hence revealing likely plagiarized segments [12].

Any natural language processing application considers text representation as one of its key building blocks. Using text character $n-grams$, its representation demands decomposing it into all the possible arrangements of n successive characters. For example, 3-grams of the word *computer* are: *com, omp, mpu, put, ute, ter*. So, the $n-gram$ profile of a given text is defined as the set containing all the $n-grams$ of a predetermined length, n, beside their frequency. The work in [13] summarizes the approaches for detecting intrinsic plagiarism where character $n-grams$ are used.

In [15], an approach was proposed wherein the representation of the suspicious document and its divisions was signified by the use of 3-grams profiles. Attaining the divisions was achieved through a sliding window of 1000 characters, and the movement was done by 200 characters in each step. Then, based on the variance between n-gram profile for each segment and the document n-gram profile, the style variation function was calculated. By making a comparison between a threshold parameter and the standard deviation of style variation function values, the suspicious document was whether classified as comprising plagiarized segments or written by one main author. Plagiarism is discovered and a segment is recognized as plagiarized if the value of its style variation function exceeds a determined threshold [15].

An issue that was discussed in [14] pertaining to dealing with long texts, as representation of documents is going to be computationally expensive when all their n-grams are taken in consideration. Accordingly, the fragments of the suspicious document were represented using a predefined set of 3-grams with high frequency. An authorship attribution research motivated this idea where the use of n-grams with high-frequency succeeded [15]. For detecting outliers, this method used the dissimilarity measure of [8]. However, the computation was achieved considering every pair of sections in a suspicious document.
For [6], the proposed work was fully built on representing the deviation word frequency as a key indicator of stylistic difference.

The authors in [16] used a set of features for representing each sentence encompassing: the rarest n-grams frequency, the most frequent n-grams frequency, and the mean of n-grams relational frequency. The last feature was a new feature, and the calculations were performed for every n-gram in a sentence. N-gram's relational frequency became higher if it was more specific to a sentence. When doing experiments with different lengths, the authors stated that n-grams of length (1, 3, and 4) were determined as the optimal lengths. Then, for generating a model for combining features and predicting for each sentence, a score representing its mismatch degree with the main author style, and the gradient boosting regression trees were used. Lastly, a plagiarized mark was given to the sentences having a score higher than a certain threshold.

Authors in [17] proposed a model that constructs profiles for the suspicious document and the spawned segments based on considering average weight of word uni-grams rather than their frequency.

3. Methodology

The proposed method considers the work introduced in [17] for building the profiles for the entire document and the produced segments. As a feature, it uses the significance of word

considering its average weight instead of its frequency. The approach proposed in this research paper for detecting intrinsic plagiarism states that the given suspicious document is divided into a sequence of separated divisions considering its paragraphs. Then, the center vectors are calculated for the entire document and for each of the segments produced from document's segmentation process. Each center vector contains words n-gram importance represented as their average weights over the formed segments and the original document. Afterwards, dissimilarity is measured between the center vector for every segment and the center vector for the complete document. Next, the document style function is constructed as an average of segments' dissimilarity to the entire document. Lastly, plagiarized sections are detected through calculating their deviation from the document's main style function. Moreover, word n-grams length has been investigated to show its effect on system performance.

Thus, given a suspicious document D containing k paragraphs such that $D = \{p_1, p_2, p_3, \dots, p_k\}$. For the proposed approach, to discover variations in the author's writing style, the applied steps are as follows: Initially, pre-processing D is performed comprising the tasks: excluding numbers, removing all non-alphabetic characters, lowercasing all characters, and then the word n-grams are considered without excluding stop words. the final result from pre-processing D will be the m word n-grams as $V = \{g_1^{ng}, g_2^{ng}, g_3^{ng}, \dots, g_m^{ng}\}$ where ng represents the length of the word n-gram. After pre-processing, the weighting process is achieved for the resulted m word n-grams with the use of $tf - idf$ weighting scheme [18, 19]. After that, the center of document D is calculated as a vector $\overline{V}$ reflecting its main content and containing the average $tf - idf$ weight for all m word n-grams where $\overline{V} = \{\overline{v_1}, \overline{v_2}, \overline{v_3}, \dots, \overline{v_m}\}$. The j^{th} coordinate, $\overline{v_j}$ of the center vector $\overline{V}$ is calculated as in $Eq.1$:

$$\overline{v_j} = \frac{\sum_{i=1}^{n} wt_{ij}}{n} \qquad j = 1, 2, 3, \dots, m \qquad (1)$$

Where wt_{ij} is the $tf - idf$ weight of word n-gram j at sentence i.

Now, for the segmentation process, the complete document is divided, taking into consideration paragraphs exist in it wherein sections seg are produced initially where$seg \in S$. Next, for each segment $seg \in S$, the weight vector $\overline{v_{seg}}$ is computed as in Eq. 1 representing its center that imitates average $tf - idf$ weight considering its word n-grams. After that, testing document self-similarity is achieved using algorithm 1 as follows:

Algorithm 1 Intrinsic plagiarism detector

Input: Document D contains k paragraphs; $D = \{p_1, p_2, p_3, \dots, p_k\}$

Threshold: τ

Step0: Start

Step1: Pre-process $D = \{p_1, p_2, p_3, \dots, p_k\} \Rightarrow V = \{w_1^{ng}, w_2^{ng}, w_3^{ng}, \dots, w_m^{ng}\}$ contains unique m word $n-grams$; such that m varies according to the fragment length ng.

Step2: Considering each segment as a document, weigh each word $n-gram$ w_j^{ng} in V using $tf - idf$ weighting scheme.

Step3: For the entire document D, Compute its center vector; $\overline{V} = \{\overline{v_1}, \overline{v_2}, \overline{v_3}, \dots, \overline{v_m}\}$ wherein each element $\overline{v_j}$ corresponds to a word $n-gram$ and is calculated as in Eq. 1

Step4: Segment D into k disjoint segments $seg \in S$ wherein each seg corresponds to a paragraph; $D = \{seg_1, seg_2, seg_3, \dots, seg_k\}$

Step5: for each $seg \in S$ do

Step6: Compute seg center vector; $\overline{v_{seg}}$ wherein each element in it is calculated as in Eq. 1 considering the word $n-gram$ exists in it.

Step7: $DisSim_{seg} \leftarrow 0$

Step8: Compute Dissimilarity between $\overline{v_{seg}}$ and document center vector $\overline{\mathrm{V}}$

$DisSim_{seg} = \left(1 - Sim_{Cos}\left(\overline{\mathrm{V}}, \overline{v_{seg}}\right)\right)$

end for

Step9: Calculate the writing style function of the entire document

$Style \leftarrow \frac{\sum_{seg \in S} DisSim_{seg}}{|S|}$

Step10: For each $seg \in S$, check their deviation from document writing style; $Style$

Step11: if $DisSim_{seg} <$ Style $- \tau$ then

Mark segment seg as an outlier segment.

end if

end for

Step12: Finish

The general document style is presented in Algorithm 1, which is characterized by the average of dissimilarities for all segments seg contained in the complete document. This algorithm takes into account the intuition; a low value will result from the comparison of the center vector of segment seg against the center vector of the entire document through measuring their dissimilarity if certain words are only used on a certain segment. Finally, the segment seg is classified as an outlier or not by considering its distance in relation to the document's style. Document's main style is represented by the average value resulting from comparing center vectors of all segments against document center vector. This value is roughly calculated by measuring Cosine dissimilarity [18] for all segments' center vectors $\forall seg \in S$ and the entire document center vector. If the similarity is significant, in this case, the value of the style function will be lower than the threshold, and then the segment is classified as suspicious.

Evaluation Metrics

Performance of the proposed approach was evaluated using the PAN corpora [20]. For Precision metric, which takes in consideration a pair of passages recognized as a case for plagiarism, it states the degree of copying between them. For Recall, it states the proportion of plagiarized passages if the classifier achieves their classification properly.

Interpretation of these measures can be performed in conjunction with the effectiveness of the classifier. True Positive means a detection that is correct. False Positive means a document that should have been recognized as plagiarized, was not. True Negative means incorrect classification for a document that was categorized as plagiarized. Lastly, False Negative describes an incorrect classification for a document that was not recognized as plagiarized, when the right classification is the opposite.

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (2)$$

$$Recall = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (3)$$

The two measures are sometimes used together in $F - measure$ metric, which represents the harmonic mean between them for providing a single measurement for a system, which is computed as in Eq. (4).

$$F - measure = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Granularity is a metric introduced in 2009 by the authors of [20] for evaluating algorithm usability. Every actual plagiarism case should be reported one time. If there are multiple

detections for only one case, the Granularity increases. The desired value for Granularity is 1. It is calculated using Eq. (5) :

$$Granularity(S,R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_s| \quad (5)$$

S_R is the cases that are correctly identified, whereas $|R_s|$ represents the number of times when case s is detected .

The measures are combined into a single score $Plagdet$ to make a unique rank among detection methods as in Eq. (6).

$$Plagdet = \frac{F-\text{measure}}{log_2(1+\text{Granularity}\ (S,R))} \quad (6)$$

5. Performance Evaluation and Discussion

The intrinsic part of two corpora used in the international competition of plagiarism detection in 2009 and 2011 (named PANPC-09 and PAN-PC-11, respectively) was used for the performance evaluation of the proposed approach as an evaluation corpora. These collections include XML clarifications that specify the positions of the plagiarized sections. For evaluating performance, evaluation metrics including: Macro-averaged F-measure, Recall, Precision, Granularity, and Plagdet were used [20]. The macro-averaged Precision, and Recall are not affected by the length of the plagiarism case. F-Measure is the weighted harmonic mean of Precision and Recall. Since nothing indicates that either Recall or Precision is more noteworthy, they are measured equally weighted. The Granularity metric captures the detection algorithm power, which means reaching the detection of a plagiarism case either done in one portion or in a number of portions. Precision, Recall and Granularity need to be combined for an overall score called Plagdet to achieve a total order for the reason that they permit for a partial ordering among algorithms that detects plagiarism.

Through the use of $PAN-PC-09$ and $PAN-PC-11$ as an evaluation dataset, Table 1 and Table 2 illustrate the results of comparing the approach proposed in this paper against the work of [6] and the approach proposed in [17].

Table1: Performance comparison of the proposed model against $Model_{[6]}$ and $Model_{[17]}$ in terms of $Precision, Recall, F-measure, Granularity,$ and $Plagdet$ evaluated using $PAN-PC-09$ corpora.

Evaluation metric	$Model_{[6]}$	$Model_{[17]}$	Proposed approach
Precision	0.3897	0.3308	0.3886
Recall	0.3109	0.4503	0.3498
F-measure	0.3458	0.3814	0.3682
Granularity	1.0006	1.1765	1.1002
Plagdet	0.3457	0.3399	0.3439

Table 2: Performance comparison of the proposed model against $Model_{[6]}$ and $Model_{[17]}$ in terms of $Precision, Recall, F-measure, Granularity,$ and $Plagdet$ evaluated using $PAN-PC-11$ corpora.

Evaluation metric	$Model_{[6]}$	$Model_{[17]}$	Proposed approach
Precision	0.3398	0.2806	0.3381
Recall	0.3123	0.4303	0.3117
F-measure	0.3255	0.3397	0.3244
Granularity	1	1.1111	1.0241
Plagdet	0.3255	0.3151	0.3189

For each document, only information that included is to be taken in consideration. As revealed in Table 1 and 2, the proposed approach has achieved a noticeable improvement over the work in [17] in terms of Granularity and also for Plagdet, which defines the complete performance of a plagiarism detection method. Moreover, stability in the proposed approach performance is recognized for both corpora. The attained results state that representing the suspicious document and its segments as vectors concerning average weight feature for word n-grams and computing dissimilarity between these vectors to distinguish deviation in writing style have affected performance positively, other than identifying deviation through computing difference between the corresponding word n-grams related to the pair under comparison. Also, when compared with work in [6], more stability was detected for the proposed work against work in [17].

Table3: Effect of n-gram length with (n=1, 2 and 3) on performance evaluation of the proposed model in terms of $Precision, Recall, F-measure, Granularity,$ and $Plagdet$ evaluated using $PAN-PC-09$ corpora.

Evaluation metric	N=1	N=2	N=3
Precision	0.3886	0.3899	0.3707
Recall	0.3498	0.3544	0.3409
F-measure	0.3682	0.3713	0.3552
Granularity	1.1002	1.0177	1.1041
Plagdet	0.3439	0.3666	0.3310

Table4: Effect of n-gram length with (n=1, 2 and 3) on performance evaluation of the proposed model in terms of $Precision, Recall, F-measure, Granularity,$ and $Plagdet$ evaluated using $PAN-PC-11$ corpora.

Evaluation metric	N=1	N=2	N=3
Precision	0.3381	0.3454	0.3329
Recall	0.3117	0.3206	0.3106
F-measure	0.3244	0.3325	0.3214
Granularity	1.0241	1.0107	1.0471
Plagdet	0.3189	0.3300	0.3110

Table 3 and Table 4 illustrate how n-grams length affects performance evaluation when used as a supplementary feature. It is shown that the short n-grams (for n=1) and the long n-grams (for n=3) have been outperformed by middle-sized n-grams (n=2). This reveals that the significant word bi-grams, when the fragments of the suspicious document use them for representation, help in detecting plagiarism more than fragment representation using word uni-grams and word tri-grams. It is shown that the tri-grams have lower performance when compared to unigrams and bigrams. This means that dividing the document into fragments with long length has a negative effect on the detection process.

6. Conclusions

A method for recognizing intrinsic plagiarism centres on making comparison of the writing style of a particular document has been introduced in this research paper. The aim was to determine if writing of a text was done by one or more authors. The experimental results showed that representing the suspicious document and its segments as vectors concerning average weight feature for word n-grams and computing dissimilarity between these vectors to detect

deviation in writing style affect positively on performance evaluation. More than distinguishing deviation through computing difference between the corresponding word n-grams related to the pair under comparison.
Furthermore, considering the effect of the word n-gram length, a positive impact in detecting intrinsic plagiarism was recorded when the suspicious document and its generated segments are fragmented into word bi-grams rather than using word uni-grams and word tri-grams. It is shown that the tri-grams have lower performance compared to unigrams and bigrams. This means that splitting the document into fragments with long-length affects negatively on the detection process.

In the detection of intrinsic plagiarism, the recorded results reveal the necessity of further research, and the need for developing new approaches to model the writing style. For future works, other features may be added to work as supplementary features beside the word n-gram length for improving performance. Work may also focus on improving Granularity through applying other segmentation schemes. Moreover, experiments may be extended to be performed on other languages other than the English language.

References

[1] Z. Eissen, and S. Benno, "Intrinsic plagiarism detection," In *European conference on information retrieval*, 2006, pp. 565-569. Springer, Berlin, Heidelberg.
[2] M. Mohammed, N. Kadhim, and A. Ibrahim, "Improved VSM Based Candidate Retrieval Model for Detecting External Textual Plagiarism," *Iraqi Journal of Science*, pp. 2257-2268, 2019.
[3] A. Maurer, K. Frank, and Z. Bilal, "Plagiarism-A survey," *Journal of Universal Computer Science*. vol. 12, no. 8, pp.1050-1084, 2006.
[4] R. Hunt, "Let's hear it for internet plagiarism." Teaching Learning Bridges vol. 2, no. 3, pp. 2-5, 2003.
[5] C. Park, "In Other (People's) Words: Plagiarism by university students--literature and lessons," *Assessment & Evaluation in Higher Education*, vol. 28, no. 5, pp. 471-488, 2003.
[6] G. Oberreuter, L. Gabriel, R. Gaston, A. Sebastián, and V. Juan, "Approaches for intrinsic and external plagiarism detection," *Proceedings of the PAN*, vol. 4, no. 5, 2011.
[7] M. Kuznetsov, M. Anastasia, K. Rita, and S. Vadim, "Methods for Intrinsic Plagiarism Detection and Author Diarization," In *CLEF*, pp. 912-919. 2016.
[8] E. Stamatatos, "Intrinsic plagiarism detection using character n-gram profiles," *threshold*, vol. 2, no. 1,pp.500, 2009.
[9] M. Zechner, M. Markus, K. Roman, and G. Michael, "External and intrinsic plagiarism detection using vector space models," In *Proc. SEPLN*, vol. 32, pp. 47-55. 2009.
[10] L. Bensalem, R. Paolo, and C. Salim, "Intrinsic plagiarism detection using n-gram classes," In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1459-1464. 2014.
[11] S. Eissen, S. Benno , and K. Marion, "Plagiarism detection without reference collections," In *Advances in data analysis*, pp. 359-366. Springer, Berlin, Heidelberg, 2007.
[12] L. Seaward, and M. Stan. "Intrinsic plagiarism detection using complexity analysis," In *Proc. SEPLN*, pp. 56-61, 2009.
[13] L. Bensalem, Imene, R. Paolo, and C. Salim, "On the use of character n-grams as the only intrinsic evidence of plagiarism," *Language Resources and Evaluation*, vol. 53, no. 3, pp. 363-396, 2019.
[14] CM. Kestemont, L. Kim, and D. Walter, "Intrinsic plagiarism detection using character trigram distance scores*," Proceedings of the PAN*, vol. 63, 2011.
[15] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods," *Journal of the American Society for Information*, vol. 60, no.3, ,pp. 538–556, 2009b. doi:10.1002/asi.
[16] M. Kuznetsov,M. Anastasia, K. Rita, and S Vadim, "Methods for Intrinsic Plagiarism Detection and Author Diarization," In *CLEF (Working Notes)*, pp. 912-919, 2016.
[17] N. J. Kadhim and M. I. A. Almulla khalaf, "An Improved Outlier Detection Model for Detecting Intrinsic Plagiarism", *Iraqi Journal of Science*, vol. 63, no. 12, pp. 5581–5588, Dec. 2022.

[18] G. Salton, and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 25,no.5, pp, 513–523, 1988.

[19] N. J. Kadhim, H. H. Saleh, and B. Attea, "Improving Extractive Multi-Document Text Summarization Through Multi-Objective Optimization", Iraqi Journal of Science, vol. 59, no. 4B, pp. 2135–2149, Nov. 2018.

[20] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An evaluation framework for plagiarism detection,",In *Coling,* 2010.