Role Term-Based Semantic Similarity Technique for Idea Plagiarism Detection

Ahmed Hamza Osman, Hani Moetque Aljahdali
Department of Information System
Faculty of Computing and Information Technology at Rabigh, King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—Most of the text mining systems are based on statistical analysis of term frequency. The statistical analysis of term (phrase or word) frequency captures the importance of the term within a document, but the techniques that had been proposed by now still need to be improved in terms of their ability to detect the plagiarized parts, especially for capturing the importance of the term within a sentence. Two terms can have a same frequency in their documents, but one term pays more to the meaning of its sentences than the other term. In this paper, we want to discriminate between the important term and unimportant term in the meaning of the sentences in order to adopt for idea plagiarism detection. This paper introduces an idea plagiarism detection based on semantic meaning frequency of important terms in the sentences. The suggested method analyses and compares text based on a semantic allocation for each term inside the sentence. SRL offers significant advantages when generating arguments for each sentence semantically. Promising experimental has been applied on the CS11 dataset and results revealed that the proposed technique's performance surpasses its recent peer methods of plagiarism detection in terms of Recall, Precision and F-measure.**

Keywords*—Plagiarism detection; semantic similarity; semantic role; term frequency; idea

I. Introduction

Given the bigness of the online, plagiarism, or the intended use of somebody else's original data while not acknowledge its supply, has been a heavy drawback in areas like Literature, Science, and Education. The convenience of access to proprietary contents has become an issue of concern additionally for scholars. The challenge is exacerbated when the suspected text generated semantically, which is known as idea plagiarism. It is not solely the extra problem of manually capturing the concept or idea performed, however additionally the people's lack of information concerning writing ethical issues and text paraphrasing. The different categories of plagiarism are cut-and-paste, ideas plagiarism, semantic plagiarism and paraphrasing, style plagiarism, authorship and citation plagiarism [1]. Several works had been done in text plagiarism detection based on the lexical and syntactic structure of the writing and failed to detect the semantic and idea plagiarism. However, most of these methods are created for verbatim duplicates, and similarity performance is decreased when dealing with plagiarism with heavy cases [2], due to paraphrasing and semantic similarity cases. Recently, different studies tried to develop and improve the accurate methods in semantic and idea text plagiarism detection domain such as [3-6] Alzahrani et al. [7]; Maurer et al. [4]; Gupta and Deep [6]; 2016; Vani and Gupta [5]; Juan D. Velásquez and et al. [8]; Weber-Wulff [9]).

The rest of the paper is ordered as follows: related work and Literature review in different type of plagiarism detection is considered in Section 2. In Section 3, a suggested proposed solution is presented. Section 4 discussed a full depiction of the idea plagiarism detection and role-based similarity that formulated in the proposed method. Corpus and experimental design that conducted in the suggested method is described in Section 5. In Section 6, output results and discussion is provided, whereas Section 7 is devoted to the conclusion.

II. Material and Methods

Several studies have discussed plagiarism in academia filed [4, 10, 11], and demonstrated different categories of available plagiarism detection techniques. For instance, Vani and Gupta[5] proposed an idea plagiarism detection method based on semantic syntax concept extraction. The extracting of the concepts was generated using a genetic algorithm. Their method detects the idea plagiarism based on two level Document level and word level. They tried to combine the similarity measure that employs the semantic concept extraction and then used for passage stage matching [5].

Palkovskii, Belov and Muzika [12] presented Exploring Fingerprinting as an outside plagiarism identification strategy to PAN-PC-2010. Their framework was initially created as a component of the proposal stockpiling framework utilized by the Zhytomyr State University. Palkovskii, Belov, and Muzika depended on fingerprinting and hash look techniques for finding likenesses between reports.

Osman and et al. [1] proposed a detection scheme based on SRL and concept extraction. The SRL used for extracting the roles and arguments for each sentence and the wordNet used to extract the sense of each term inside the sentence. The proposed method can use in different type of detection such as copy and paste, semantic and paraphrasing, structure plagiarism [1].

Palkovskii, Belov, and Muzyka, [13] likewise returned contenders. Like Encoplot, they proposed a WordNet-based semantic similitude estimation for the outside counterfeiting identification shown in PAN-PC-09 and was enhanced for PAN-PC-2010. For PAN-PC-2011, they exhibited a gauge for further review by demonstrating unmistakably characterized corpus measurements, for example, outside and inherent,

confusion techniques, point coordinate, case length and report length. They demonstrated that an immediate connection could be made between the complication technique and accomplished execution.

Sheffield University spoke to by Parth, Sameeer and Majumdar, [14] thought of a framework that was intended to identify extraneous counterfeiting. They utilized a three-organize technique for pre-preparing; record choice utilizing term n-grams and their last examination utilized a Running Karp-Rabin Greedy String Tiling string coordinating method. Their framework was granted a score of 0.20 for general execution with scores of 1.21 for Granularity, 0.16 for the Recall measure and 0.4 for Precision.

Chong Specia and Mitkov [15], proposed another system for unoriginality location given the string coordinating and Naïve-Bayes classifier. Guileless Naïve-Bayes algorithm is a straightforward classifier given the Bayes' hypothesis. The technique is accustomed to creating a probabilistic framework of the plagiarism short answer questions dataset. The point of a Naïve-Bayes algorithm is to take on their framework to order different cases given the dataset classes (cut-and-paste, light, heavy, and non-plagiarized). The strategy chose the arrangement of "best components" and utilized them to take in their framework to group a given potential plagiarism record as having a place with one of the plagiarism classes.

We show through the previous studies a various works has been done to detect and capture the text plagiarism. However, these approaches that have been introduced still need to be enhanced to finding the academic and scientific plagiarized idea, mainly in semantic structure part.

III. Proposed Method

It's very significant to note that exploiting relationship between terms roles and their verbs in the same text has promising possible for understanding the idea of the text. The ordinary method of text mining is to capture term frequency. In this paper, Semantic Role-based model for text similarity and idea plagiarism detection will be introduced. The main concept of the suggested technique is capturing the meaning construction of sentence terms within a text and documents; capturing and role terms frequency; and capturing the idea similarity based on semantic structure and role terms frequency together. The proposed method compromise of different phases starting from the preprocessing steps such as sentences chunking and stop term removal. Then, the semantic role labeling technique will be utilized to exploit the roles of all terms inside the sentences. The last steps are text similarity and idea extraction based on role frequency and role term similarity. The main framework of the suggested technique is shown in Fig. 1.

Every one of these means will be further talked about in the accompanying segments:

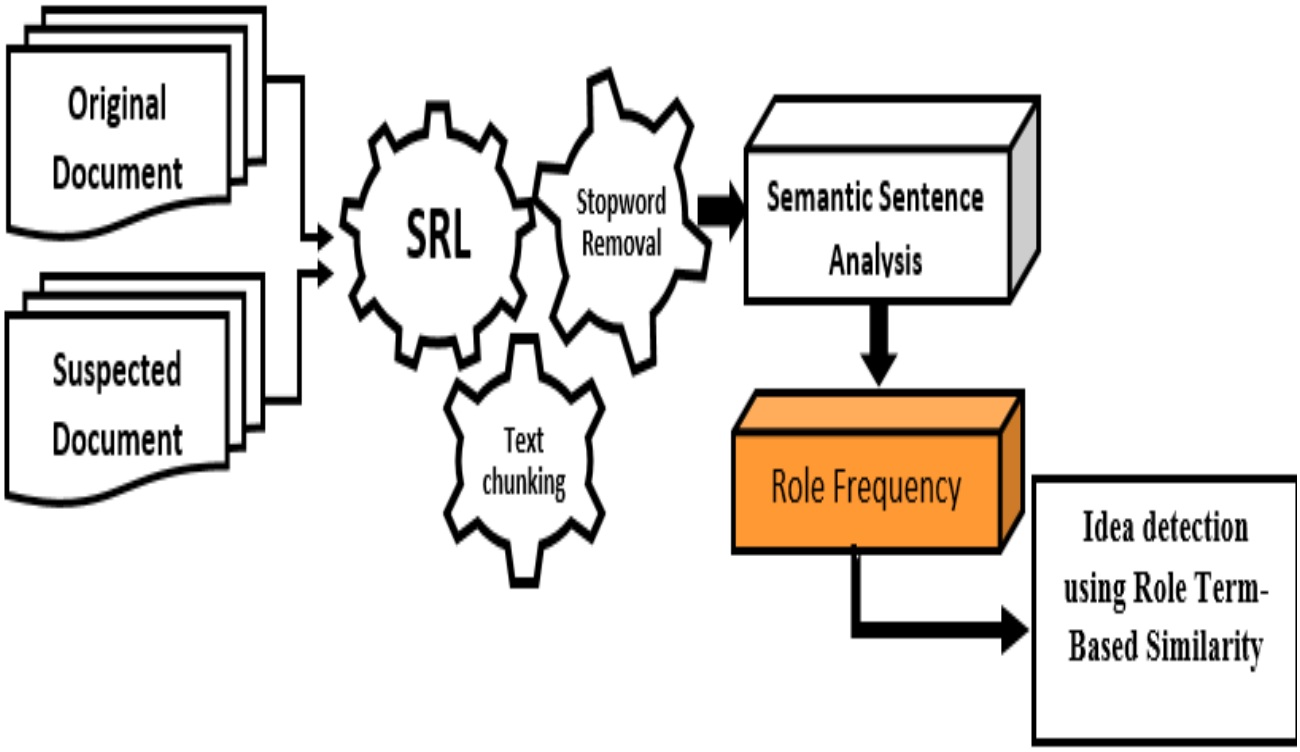

Fig. 1. Proposed Framework.

A. Semantic-Role Labeling (SRL)

Semantic constructions or linguistics frames were suggested by Fillmore [16] wherever general frames were utilized for general themes and roles such as and PropBank introduced by Palmer et al. [17]. FrameNet suggested by Baker et al. [18]. A statistical scheme is learned on the data from the FrameNet project to mechanically assign linguistics roles [17]. Pradhan et al.[19], Surdeanu et al. [20], and Xue and Palmer [21] followed this method by enhancing sets of data mining approaches. Barnickel et al. [22] proposed a large scale system based on SRL and neural network for extracting relation from biomedical data. This technique essentially used SENNA tool that is utilized in different text processing system. The SENNA tool can exploit the terms roles for each sentence based on the neural network technique whereas they modified this tool to exploit the relations among the biomedical terms semantically.

SRL is a procedure to the idea and terms roles in a text document. The main concept is that the subject and object roles in the text document are determined based on analysis of each sentence semantically. It can be produced to the categorization of actions such as determination of "who" did "what" to "whom", "where", "how", and "when". A verb is usually used as clause predicate begins "what" took place, and other fragments of the sentence definite the other roles of the sentence (such as "when" and "who"). The main function of SRL is to define the semantic relations between a predicate and its share properties or participants, with these relationships drawn from a predefined list of potential semantic roles for the class of predicate or the predicate itself. The dataset set that mainly used in the SRL and the SENNA tools is called Proposition Bank (Propbank) corpus.

In this study, role frequency of semantic-part marking in view of the sentence-based was suggested as a new technique for idea plagiarism identification. SRL intends to identify the game plan likeness among the ideas of the reports and conceivable semantic closeness among both records. This progression in the review utilized the part marks of the ideas for the text-documents and gathered them as clusters. The cluster that was utilized as a part of this technique gave a snappy manual for capturing the associated part with the text.

B. Text Preprocessing

In this phase, the text preprocessing stage contained three sub-stages which were text chunk, and stop words withdrawal. A text chunk partitioned a text archive into sub-sentences. Several studies concentrate on text preparing strategies in various fields, incorporate intrusion detection [23]. The step of stop terms removal for erasing meaningless terms was utilized. This progression separated the critical terms from the text and disregarded the rest of the terms. This may have unfavourably influenced the comparability between texts.

A basic prepreparing includes isolating the text into important parts and is defined text chunking. Text can be separated into words, themes, or sentences. The chunking is conducted by limit recognition and isolating a text into sub-sentences. By and large, an outcry stamp (!), a question mark (?), or a period (.) is the typical signs that show a sentences limit Mikheev [24]. This study utilized the sentence based text chunking as the initial phase in the suggested approach, where the first and suspected documents will be isolated into sentence pieces. This technique was picked on the grounds that our proposed strategy intends to contrast a speculated text and unique text in light of the sentence matching methodology.

Stop Terms are the Terms that every now and again happen in archives. They are Terms, for example, "a", "and" and "the". These terms don't provide any indication qualities or implications to the substance of the records, henceforth, they are dispensed with from the arrangement of file words [25]. The proposed strategy dispensed with all the stop terms in the documents to accelerate the system procedure. The introduced strategy utilized the list of the Buckley stop terms [26] that was utilized as a part of the SMART data recovery framework at Cornell University.

C. Term and Role Frequency

Term frequency is a one of the information retrieval process for highlighting the important terms within the text. Several page ranking techniques are working based on the term frequency to sort and rank the retrieved information based on the similarity and term frequency with the user query. Some retrieved documents can be similar and contribute more than other documents but will not be selected as relevant due to less term frequency between the query and the corpus. The idea of these documents can be identical semantically, but the similarity techniques ant not able to capture the similar pattern and meaning due to the lexical and character matching approached that was used in their techniques. This one of the main issues will be solved by the proposed method.

This research proposes a role frequency rather than term frequency to capture the idea of the documents. It was noticed that form the introduced technique the frequency of the term role can contribute more than the term frequency in some documents spatially when the people plagiarized the document idea by changing the structure of the sentences and reword the terms if text with their synonyms. The term can be either a verb or a role; either a word or phrase and the role can a labeled term.

A Verb argument structure is a useful example of semantic structure extraction and role term frequency capturing:

Martin eats the banana

eats*: the verb*

Martin *&* ***the banana****: roles of the verb "****eats****", Label: assigned to a role,* ***Martin****: subject, the* ***banana****: object.*

TABLE I. ARGUMENTS SORTS AND THEIR PORTRAYALS [1]

Argument Description	Argument Type	Argument Description	Argument Type
Agent	Arg(0)	Negation Marker	NEG
Object	Arg(1)	Location	LOC
Not-fixed	Arg(2 t0 5)	Purpose	PNC
Verb	V	Modal-verb	MOD
Manner	MNR	Direction	DIR
Time	TMP	Exit	EXIT
General purpose	ADV	Discourse connectives	DIS

The meaning of the arguments types illustrates in Table I.

Table I demonstrates the sorts of arguments that were utilized as a part of the analyses and their depiction or significance.

Analysis of the text document based on the term role of each sentence in the document is one of the main steps of the introduced method. Initially, the terms roles are extracting using the semantic role labeling NLP technique. The function of the SRL is to exploit roles and roles for all terms inside the sentence. The SRL used the analysis of the paragraph and sentence based on the sentence predicate (Subject, Verb, and Object) to define each the role of each term in the sentence and paragraph semantically using the PropBank dataset [17]. The verb argument plays an important role of the proposed method by extracting the semantic structure and the term frequency. The role term frequency (rtf) defined as the occurrence times of role term r in verb role structures of paragraphs. The role term (r) can be a word or phrase. The rtf of term (r) in document (d) can have various rtf parameters in a different paragraph in document (d), these parameters formulated as:

$$rtf = \frac{\sum_{n=1}^{sn} rtfn}{sn} \quad (1)$$

Where sn is the number of paragraphs or sentences that hold role term r in text document (d).

IV. Idea Plagiarism Detection

Theft and copy of ideas are considered one of the intelligent types of text plagiarism, especially in the scientific text and articles. The essential idea from the originals text is manipulated, exploited and represented in the suspicious text as a novel or new. Within a text, the idea can be presented in sentences, paragraphs, or phrases. The method developed extracts semantic ideas within a text based on term role frequency. The plagiarism detection is conducted based on two levels; Sentence level detection, and paragraph level detection. In the sentence detection level, the SRL explores sentences in the document based on the role of each term inside the sentences. Then, the frequency of extracted roles calculated and considered as important semantically. These roles frequencies are then utilized in capturing the similarity in both sentence level and paragraph level. On the other hand, the paragraph level detection is conducted based on the semantic meaning of the group of the roles terms frequencies that were extracted from the sentences level detection. Fig. 2 illustrates the idea extraction from the document.

The idea of the sentences extracts by focusing on the terms that have contributed more than other terms in the each paragraph. The term that has one frequency will be ignored and the terms that have a more frequency will be considered as significant terms. Finally, the ideas will be categorized into roles and arguments. The main idea of the paragraph is equal the group of sentences ideas in that paragraph. Typically, the main idea of the original and suspected documents is equal the group of paragraph ideas in that document and the title idea accordingly. An example of how the idea extracted based on term role frequency demonstrates in following original and suspected sentences:

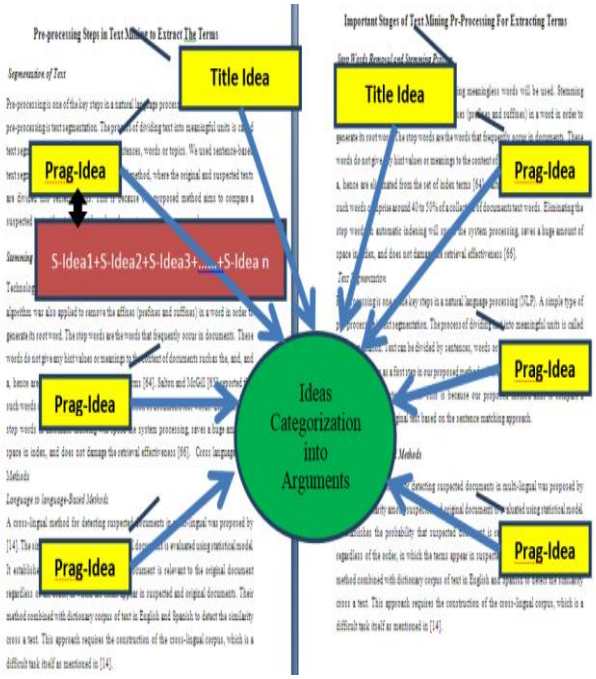

Fig. 2. Idea Extraction from the Document.

Example 1 supported by Shehataand et al. [27]:

Original sentence:

Texas and Australia researchers have created industry-ready sheets of materials made from nanotubes that could lead to the development of artificial muscles.

Suspected sentence:

The industry-ready sheets that were created by Malaysia and China researchers from resources made from nanotubes that could lead to the development of artificial muscles.

We noted from the example some important pints that should be the focus:

- Each Term has one frequency,
- What are the Important Terms?
- What is the term that has contributed more than the other terms?

The SRL analysis of calculating rtf of example 1 demonstrates as:

Original sentence:

First, the SRL will be employed to extract the verbs the sentence. In this example, three target words (verbs) is extracted as a verb argument structures:

- created
- made
- Lead

Fig. 3, 4, 5 and 6 illustrates the SRL analysis of the original sentence verbs [18] [33].

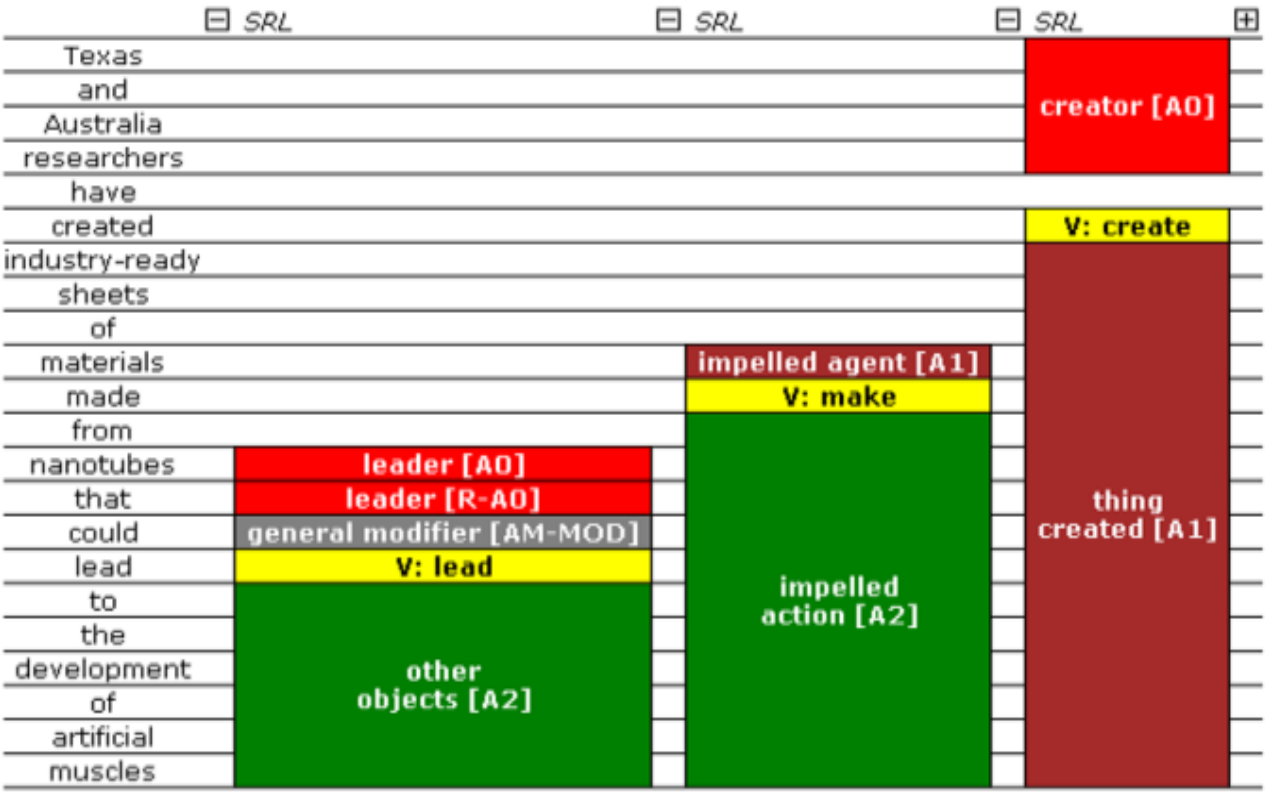

Fig. 3. Analysis the Original Sentence using SRL.

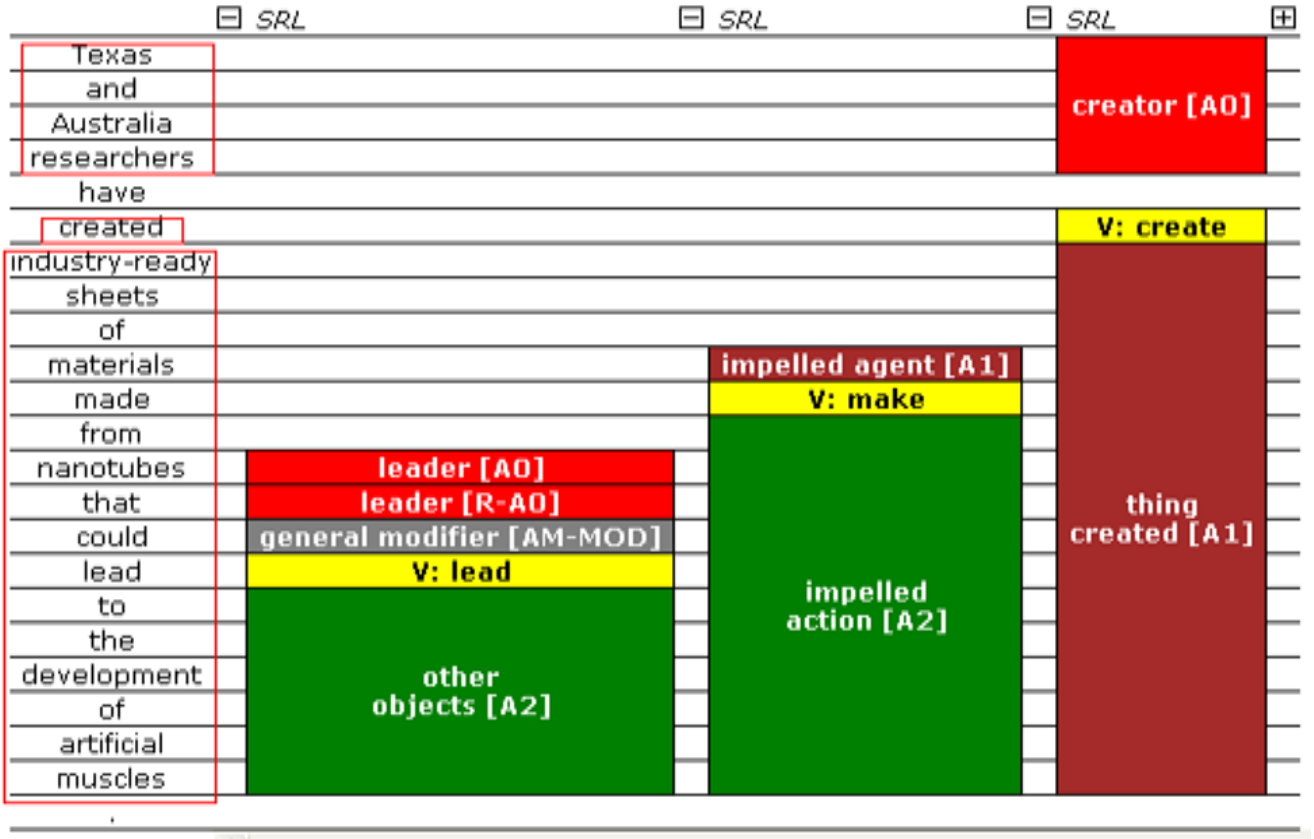

Fig. 4. Calculating rtf of the Created Verb Argument Structure.

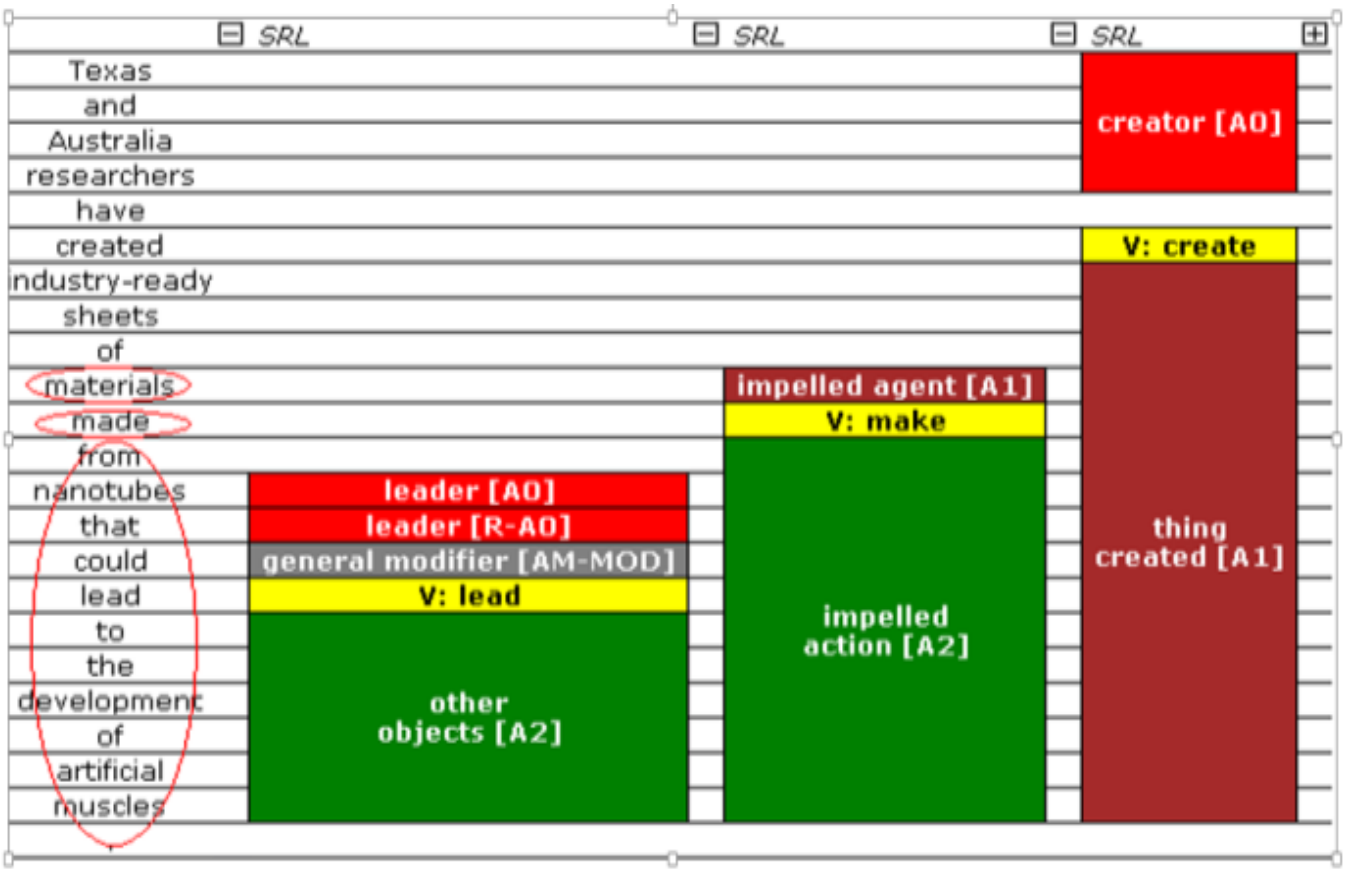

Fig. 5. Calculating rtf of the *Made* Verb Argument Structure.

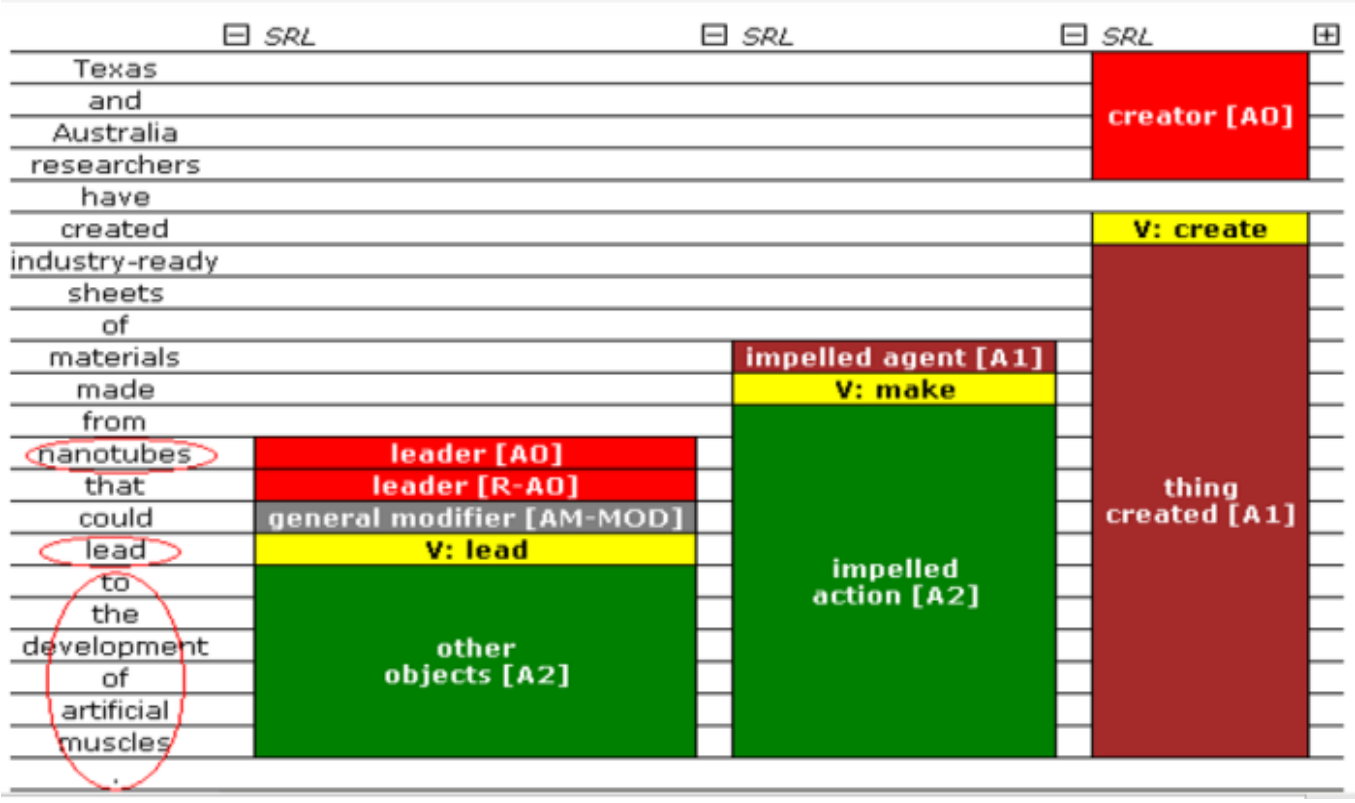

Fig. 6. Calculating rtf of the Lead Verb Argument Structure.

TABLE II. TERM ROLE FREQUENCY OF THE ORIGINAL SENTENCE

Role NO	Role-Term	Role Frequency	Role NO	Individual-Role-Term	Role term Frequency
1	-Texas-and-Australia-Researchers	1	1	- Texas	1
				-Australia	1
				Researchers-	1
2	Created-	1	2		1
3	-Industry-rely-sheets materials-made-nanotubes--lead development artificial-muscles	1	3	-Industry	1
				-rely	1
				-sheets	1
4	Materials-	1	4		1
5	nanotubes-lead-development artificial-muscles	1	5	Development-	3
				Artificial-	3
6	Nanotubes-	3	6	-	
7	-lead	3	7	-	
8	-development artificial-muscles	3	8	-muscles	3

Argument (1) is (Texas and Australia researchers), **the TARGET is** made**, and A**rgument (2) *is (*industry-ready sheets of materials made from nanotubes that could lead to the development of artificial muscles).

Argument (1) is (nanotubes), the TARGET is **lead**, and Argument (2) is (to the development of artificial muscles). The terms roles frequency at the first chunk of the original sentence in this example is calculating in both, phrase term and individual role term. Table II illustrates the term role frequency of the original sentence.

The Idea of the original sentence is "development artificial muscles" because the terms (development, artificial, and muscles) has a high role frequency and contribute more than other terms.

Suspected sentence:

The industry-ready sheets that were created by Malaysia and China researchers from resources made from nanotubes that could lead to the development of artificial muscles.

Three target words (verbs) for is also extracted from the suspected sentence. The extracted verb argument structures of this example are:

- created
- made
- Lead

We employed the same process that was used in the original sentence by extracting the argument structure of the verb ***mad,*** and ***lead*** is:

The verb (Made):

- Argument (1) is (materials), **the TARGET is made, and A**rgument (2) *is* (nanotubes that could lead to the development of artificial muscles).

The verb (Lead):

Argument (1) is (nanotubes), **the TARGET is lead, and A**rgument (2) *is* (to the development of artificial muscles). Table III illustrates the term role frequency of the suspected sentence

The Idea of the suspected sentence is "development artificial muscles" because the terms (development, artificial, and muscles) has a high role frequency and contribute more than other terms. We noted from calculating the rtf that it is possible two frequent terms have the same occurrence in their document, but one pays more to the meaning of its sentence than the counter one. This can occur when the people plagiarize the main idea of the documents with changing the structure of the documents or performing semantic plagiarism. This type called idea plagiarism. Additionally, we noted that the rtf could assist for capturing the similarly between the documents by extracting the main document idea.

TABLE III. TERM ROLE FREQUENCY OF THE SUSPECTED SENTENCE

Role -No	Role-Term	Role-Frequency	Role -No	Individual-Role-Term	Role-term-Frequency
1	Industry--rely-sheets	1	1	-Industry	1
				-rely	1
				-sheets	1
2	Created-	1	2		-
3	-Malaysia-China-researchers-resources-made-from-nanotubes-lead-development-artificial-muscles.-	1	3	-Malaysia	1
				-China	1
				-resources	1
4	-resources	2	4	-	-
5	nanotubes-lead-development-artificial-muscles	2	5	Development-	3
				Artificial-	3
6	Nanotubes-	3	-6	-	-
7	-lead	3	7	-	-
8	-development-artificial-muscles	3	8	-muscles	3

V. Experimental Design

A. Corpus and Dataset

The CS11 dataset comprises 100 human short answer questions samples of plagiarized text collected by Clough and Stevenson [2]. It offers cases of plagiarized short texts made in various plagiarism levels. The benefit of the CS11 dataset is that it is simulated and developed by a human; wherein the behaviour situation of plagiarized peoples is natural not artificial. The dataset involves of a 100 text 95 suspected short texts and 5 articles collected from Wikipedia website as the original documents. Non-native and native scholars to response five questions interrelated to the original articles wrote the suspicious texts. The responses were excepting for non-plagiarized samples based on the original documents with varied similarity, as well as the instructions are specified by the dataset inventers. The average terms in the short texts were among (200 - 300). Around 57 cases were noticeable: out of which 19 is heavy revision, 19 were nearby copy, and additional 19 marked as light revision samples; while the resting 38 cases were free plagiarized texts. The various kinds of suspected texts are defined as:

1) Near copy: this type focuses on a copy and paste from the original text;

2) Light revision: minor alteration of the original documents by substituting terms with their synonyms and giving a little linguistic modifications;

3) Heavy revision: rewriting and major alteration with rephrasing and restructuring in original documents;

4) Non-plagiarism: revised texts without any alteration in the original documents based on contributors' own terms.

B. Experimental Design

The experiments of the proposed method used the Clough09 Corpus for detecting the plagiarized idea. This is because of semantic characteristics imitation for plagiarism samples such as text rephrasing. To examine the proposed methods in, we utilized the CS11 dataset that was designed because of the PAN-PC dataset limitations. The limitations are that the common of the plagiarized samples were produced artificially. The experiments examined the amount of detecting plagiarized sentences from the original documents based on the Clough09 plagiarism Corpus. The suggested technique analysis the source and suspected documents based on the SRL. The analyzed sentences are then used to calculate the rtf from each sentence in the corpus. The idea extracted based on the rtf from the sources and suspected documents. The similarity between the ideas in the both documents is calculates based on the proposed role-based similarity metrics. The rft-based similarity metrics is adopted to detect the matching between the texts based on sentences and documents levels. To adopt the rtf-based similarity a role term-based analysis will be structured and formulated. This analysis considers two main issues; Similarity measure based on rtf, tf and df of matched role terms, and matched role terms (role terms that exist in two or more documents).

The Role Term-Based Similarity measure between two-text documents d1, d2, used factors:

- m: number of matching role terms between d1 and d2
- sn: the gross number of sentences hold similar role term ri in every text document d
- li: length of each role term in the verb role construction in very text document d
- Lvi: length of very verb role construction which holds the similar role term ri.
- N: gross number of text documents in the dataset

The similarity measure between two document d1, d2 is calculated as:

$$sim_c(d_1, d_2) = \sum_{i=1}^{n} max\left(\frac{l_{i1}}{l_{ui2}}, \frac{l_{i1}}{l_{ui2}}\right) * weight_{i1} * weight_{i1} \quad (2)$$

$$weight_i = \left(tfweight_i + cfweight_i\right) * \log\left(\frac{N}{df_i}\right) \quad (3)$$

$tfweight_i$=document level; $tfweight_i$ = Sentence level

Where $weight_i$ denotes the weight of the role term i in text document d.

The *tfweight* and *rtfweight* are normalized by length of document vector

$$tf\ weight_i = \frac{tf_{ij}}{\sqrt{\sum_{j=1}^{cn}(tf_{ij})^2}} \quad (4)$$

$$ctf\ weight_i = \frac{ctf_{ij}}{\sqrt{\sum_{j=1}^{cn}(ctf_{ij})^2}} \quad (5)$$

Where cn: the gross number of role terms which has a terms frequency values in text document d.

C. Performance Measures

This unit deliberates evaluation measures of plagiarism detection methods. The Precision and Recall factor is a common evaluation measure that normally used to assist the plagiarism detection. Potthast et al., [28-30] suggested a macro-averaged and a micro-averaged variant. The granularity or F-measure factor is an additional significant measure that was utilized in plagiarism detection performance [31][30][31][31][31](Potthast et al., 2010b), (Potthast et al., 2010b), (Potthast et al., 2010b), and (Potthast et al., 2010b). We use the micro-averaged Recall and Precision for evaluating our proposed method. The Precision and Recall of R under S are identified as follows:

$$\text{Precision}_{\text{micro}}(S, R) = \frac{\left|\cup_{(s,r)\in(S\times R)}(S \sqcap R)\right|}{\left|\cup_{r\in R} r\right|}, \quad (6)$$

Where, S and R present sets of plagiarized samples and detections, s denote plagiarized passage in a plagiarized documents, r denote associates a supposedly plagiarized passage in documents.

$$\text{Recall}_{\text{micro}}(S, R) = \frac{\left|\cup_{(s,r)\in(S\times R)}(S \sqcap R)\right|}{\left|\cup_{s\in S} S\right|}, \quad (7)$$

Where

$$S \sqcap R = \begin{cases} s \cap r & \text{if r detect s} \\ \emptyset & \text{Otherwise} \end{cases} \quad (8)$$

The granularity is the harmonic mean of recall and precision and computed based on the following formula:

$$\text{Granularity} = \frac{2\times Recall\times Precision}{Recall\times Precision} \quad (9)$$

VI. Results and Discussion

The results of the matching computation in the precision, recall, and granularity are specified in Fig. 7, 8, and 9 for different plagiarism classes Heavy, Light, and, Cut-and-paste respectively. The results of our proposed method are compared with other method reported by Chong [15]. It utilizing Naïve Bayes classifier with an arrangement of all elements, best components, and Ferret Baseline method [32]. These techniques were talked about before in Section 2. We select these techniques for correlation since it utilizes the CS11 human short answers question corpus. The aftereffects of the correlations show additionally in Fig. 7, 8, and 9.

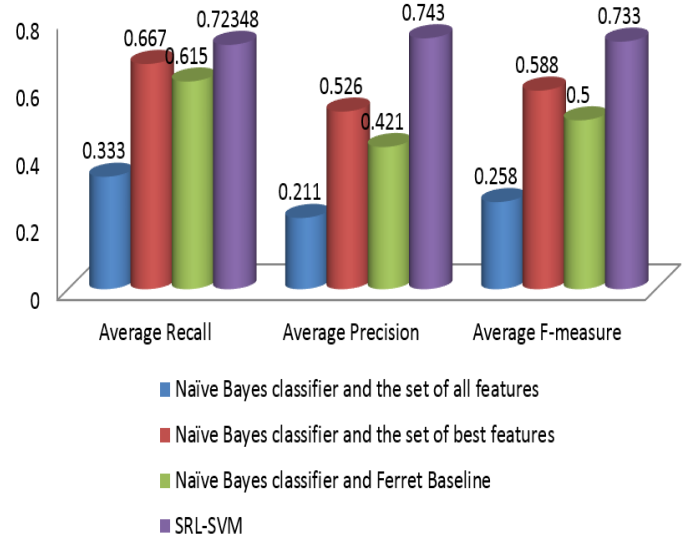

Fig. 7. CS11 Heavy Plagiarized Samples.

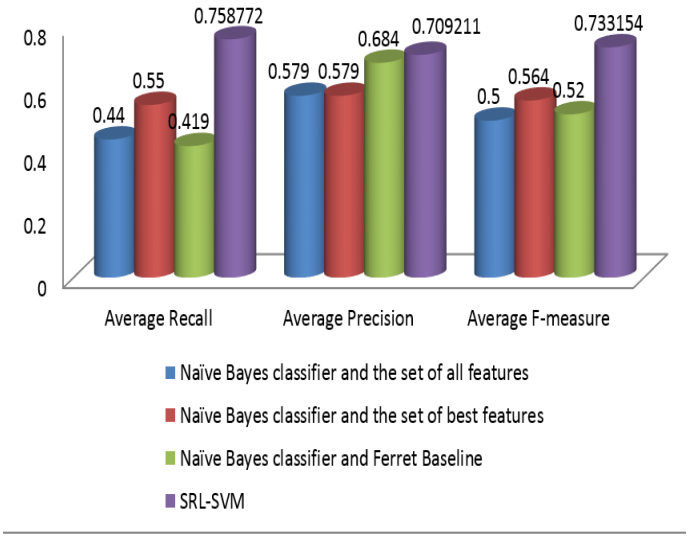

Fig. 8. CS11 Light Plagiarized Samples.

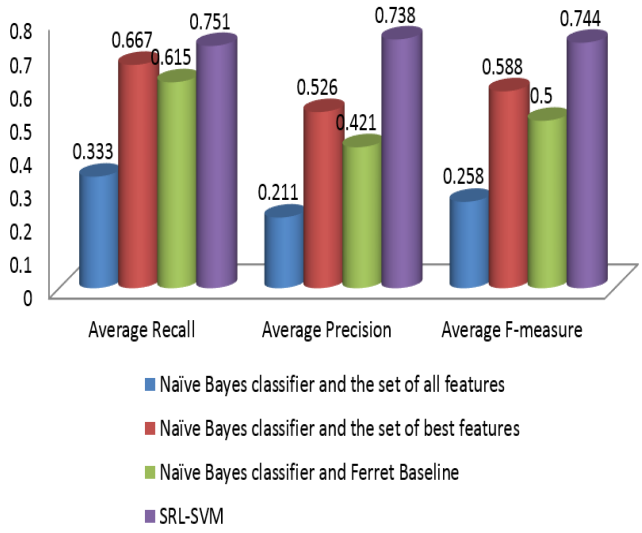

Fig. 9. CS11 Cut-and-Paste Plagiarized Samples.

Fig. 7 showed the correlation comes about between the suggested technique and different strategies based on Heavy copyright infringement class. We noticed that the suggested strategy accomplished high scores in term of Recall (0.723), Precision (0.743), and Granularity (0.733).

Fig. 8 showed the examination comes about between the suggested technique and different strategies based on light unoriginality class. We noticed that the suggested technique accomplished best scores in term of Recall (0.758), Precision (0.709), and Granularity (0.733).

Fig. 9 shown the cut-and-paste results in term of Recall, Precision, and Granularity with (0.751), (0.738), and (0.744) respectively.

The experimental output showed that the idea similarity crosses the CS11 corpus proved that the general performance in the precision, recall, and Granularity are achieved better results for capturing the main idea between the source texts and suspicion texts in the dataset. The proposed method tested with the different types of plagiarism in the CS11 corpus such as heavy, light, and copy-and-paste. Through the results, we observed that the suggested technique obtained good results compared with Naïve Bayes method with an arrangement of all elements, best components, and Ferret Baseline technique [32].

VII. Conclusion and Future Work

In this research, an idea plagiarism detection system using term role frequency is suggested and explained. The suggested method analyzed and compared the idea of the text based on role based-similarity and the frequency for each term in a text. It is possible that two frequent words have the identical occurrence in their document, but one pays more to the semantic of its sentence than the counter one. These documents can also have the same idea, but the main structure of the idea presentation totally differs spatially in the structure and semantic meaning. Semantic Role Labeling obtained significant benefits when it originated to produce meaning roles for every sentence individually. The utilization is to

detect the semantic matching among the passages. The main contributions and idea of the documents can be extracts by calculating the important roles using role term frequency. The results of the similarity performed and calculated cross the CS11 corpus and proved that the general performance of the suggested method is succeeded to capture the main idea between the source documents and suspicion documents in the dataset. The proposed method examined with samples of plagiarized text in diverse levels of plagiarism such as cut and paste, minor modification (light), and major rephrasing in source texts (Heavy). The results of the term role frequency discovered the benefits of the role-based term similarity for detecting the plagiarized idea between the source and suspected documents. In the future, the suggested method will be combining with optimization technique to improve the performance results. Another dataset using PAN-10 to PAN-12 will be tested and examined.

ACKNOWLEDGMENT

This work was supported by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under Grant No. (G: 466-830-1439). The author, therefore, gratefully acknowledge the technical and financial support from the DSR.

REFERENCES

[1] A. H. Osman, N. Salim, M. S. Binwahlan, R. Alteeb, and A. Abuobieda, "An improved plagiarism detection scheme based on semantic role labeling," Applied Soft Computing, vol. 12, pp. 1493-1502, 2012.

[2] P. Clough and M. Stevenson, "Developing a corpus of plagiarised short answers," Lang. Resour. Eval., vol. 45, pp. 5-24, 2011.

[3] S. Alzahrani, V. Palade, N. Salim, and A. Abraham, "Using structural information and citation evidence to detect significant plagiarism cases in scientific publications," Journal of the American Society for Information Science and Technology, pp. n/a-n/a, 2011.

[4] M. Hermann, K. Frank, and Z. Bilal, "Plagiarism - A Survey," Journal of Universal Computer Science, vol. 12, pp. 1050-1084., 2006.

[5] K. Vani and D. Gupta, "Detection of idea plagiarism using syntax–Semantic concept extractions with genetic algorithm," Expert systems with applications, vol. 73, pp. 11-26, 2017.

[6] D. Gupta, "Study on Extrinsic Text Plagiarism Detection Techniques and Tools," Journal of Engineering Science & Technology Review, vol. 9, 2016.

[7] S. M. Alzahrani, N. Salim, and A. Abraham, "Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. PP, pp. 1-1, 2011.

[8] J. D. Velásquez, Y. Covacevich, F. Molina, E. Marrese-Taylor, C. Rodríguez, and F. Bravo-Marquez, "DOCODE 3.0 (DOcument COpy DEtector): A system for plagiarism detection by applying an information fusion process from multiple documental data sources," Information Fusion, vol. 27, pp. 64-75, 2016.

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

[10] M. Potthast, B. Stein, A. Eiselt, A. Barrón-Cedeño, and P. Rosso, "Overview of the 1st International Competition on Plagiarism Detection," in PAN-09 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection, 2009, pp. 1-9.

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

[12] Y. Palkovskii, A. Belov, and I. Muzika, "Exploring Fingerprinting as External Plagiarism Detection Method," 2010.

[13] Y. Palkovskii, A. Belov, and I. Muzyka, "Using WordNet-based semantic similarity measurement in External Plagiarism Detection," in CLEF (Notebook Papers/LABs/Workshops), Amsterdam, The Netherlands, 2011.

[14] G. Parth, R. Sameeer, and P. Majumdar, "External Plagiarism Detection: N-Gram Approach using Named Entity Recognizer," presented at the CLEF (Notebook Papers/LABs/Workshops) 2010.

[15] M. Chong, L. Specia, and R. Mitkov, "Using Natural Language Processing for Automatic Detection of Plagiarism," Proceedings of the 4th International Plagiarism, 2010.

[16] C. J. Fillmore, "The case for case. In Emmon Bach and Robert T," Universals in Linguistic Theory. Holt, Rinehart, and Winston, New York, pp. 1–210, 1968.

[17] Martha Palmer, Daniel Gildea, and Paul Kingsbury, "The Proposition Bank: An Annotated Corpus of Semantic Roles," Comput. Linguist., vol. 31, pp. 71-106, 2005.

[18] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet Project," presented at the Proceedings of the 17th international conference on Computational linguistics - Volume 1, Montreal, Quebec, Canada, 1998.

[19] Pradhan, S. Sameer, H. Wayne, K. H. Ward, H. M. James, and Dan Jurafsky, "Shallow semantic parsing using support vector machines," In Proceedings of NAACL-HLT 2004, pp. 233–240, 2004.

[20] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth, "Using predicate-argument structures for information extraction," presented at the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, Sapporo, Japan, 2003.

[21] Xue, Nianwen, and Martha Palmer, "Calibrating features for semantic role labeling," n Proceedings of EMNLP 2004, pp. 88–94, 2004.

[22] T. Barnickel, J. Weston, R. Collobert, H. Mewes, and V. Stumpflen, "Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts," International Journal of Information Technology & Decision Making, vol. 4, p. e6393, 2009.

[23] A. Sharma, A. K. Pujari, and K. K. Paliwal, "Intrusion detection using text processing techniques with a kernel based similarity measure," Computers and Security, vol. 26, pp. 488-495, 2007.

[24] A. Mikheev, "Document centered approach to text normalization," In Proceedings of SIGIR, pp. 136-143 2000.

[25] C. Rijsbergen and J. Van, "A New Theoretical Framework for Information Retrieval.," 1979.

[26] C. Buckley, G. Salton, J. Allan, and A. Singhal, "Automatic query expansion using SMART: TREC 3," NIST SPECIAL PUBLICATION SP, pp. 69-69, 1995.

[27] S. Shehata, F. Karray, and M. Kamel, "An efficient concept-based mining model for enhancing text clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 22, pp. 1360-1371, 2010.

[28] M. Potthast, A. Barrón-Cedeño, A. Eiselt, B. Stein, and P. Rosso, "Overview of the 2nd international competition on plagiarism detection," Notebook Papers of CLEF, vol. 10, 2010.

[29] M. Potthast, A. Barrón-Cedeño, A. Eiselt, B. Stein, and P. Rosso, "Overview of the 3nd international competition on plagiarism detection," Notebook Papers of CLEF 11, vol. 10, 2011.

[30] B. Stein, M. Potthast, P. Rosso, A. Barrón-Cedeno, E. Stamatatos, and M. Koppel, "Fourth international workshop on uncovering plagiarism, authorship, and social software misuse," in ACM SIGIR Forum, 2011, pp. 45-48.

[31] M. Potthast, B. Stein, A. Barr, #243, n-Cede, #241, et al., "An evaluation framework for plagiarism detection," presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Beijing, China, 2010.

[32] C. Lyon, R. Barrett, and J. Malcolm, "A Theoretical Basis to the Automated Detection of Copying Between Texts, and its Practical Implementation in the Ferret Plagiarism and Collusion Detector," Plagiarism: Prevention, Practice and Policies Conference Newcastle, UK. 2004, 2004.