Journal of Engineering Science and Technology
Vol. 11, No. 10 (2016) 1403 - 1420

PLAGIARISM DETECTION IN TEXT DOCUMENTS USING SENTENCE BOUNDED STOP WORD N-GRAMS

DEEPA GUPTA*, VANI K., LEEMA L. M.

Amrita School of Engineering, Amrita Vishwa Vidyapeetham,
Bangalore Campus, Bangalore -560035, India
*Corresponding Author: g_deepa@blr.amrita.edu

Abstract

With the evolution of technologies like internet search engines and improved text editors, plagiarism has become a critical issue. Many works are already available in verbatim plagiarism detection which is a type of simple copy and paste plagiarism but when it comes to intelligent plagiarism the scenario becomes more complex. Intelligent plagiarism includes plagiarism through idea adoption, translation and text manipulations which is more challenging to deal with. The paper makes an attempt to detect intelligent plagiarism using the structural information within the document. This is done by the extraction of stop words, in contrast to the other methods that usually rely upon content words. The proposed method enhances this existing idea by including the rough sentence boundaries along with stop word profiles. Further this method is extended using the part of speech tags and finally the system is evaluated using sample documents from PAN- 2010 data set. The results are compared with the baseline approach and performance is evaluated based on standard PAN measures.

Keywords: Plagiarism detection, Extrinsic plagiarism, Stop word, Sentence bounded, POS tagging.

1. Introduction

In today's world of information and modern technologies, copying text and manipulating them is not a difficult task. In this modern world where any information can be obtained with the help of digital devices, plagiarism can occur easily. Plagiarism is defined as a close intimation of an old idea and publishing it as one's own work without proper citation [1]. Intensity of plagiarism can be in

Nomenclatures

C	Most frequent stop words, a list
D_{rx}	Source documents retrieved for the suspicious document
D_s	Set of source documents
d_s	Subset of source documents
D_x	Set of suspicious documents
d_x	Subset of suspicious documents
$P\ (n, d_s)$	SWNG profile for source documents
$P\ (n, d_x)$	SWNG profile for suspicious documents
P_{SB}	Sentence bounded stop word profile
t_s	Corresponding passage in source document d_s.
t_x	Plagiarized passage in suspicious document d_x

Greek Symbols

θ	Threshold

Abbreviations

NLP	Natural Language Processing
POS	Part of Speech
SBSWNG	Sentence bounded stop word n gram
SBSWNG(N)	Sentence bounded stop word n gram+ Noun
SBSWNG(NV)	Sentence bounded stop word n gram + Noun + Verb
SBSWNG(V)	Sentence bounded stop word n gram + Verb
SWNG	Stop-Word N-gram

varying range, from making a true copy of text to hacking ideas and making their own texts. Plagiarism comes into picture when someone forgets the basic rule in research that if an idea of a third party is taken, then it has to be given credit.

Plagiarism detection is a necessity in today's world to ensure that users can access information freely without compromising on illegal copying and distribution of information. Plagiarisms are of different types and the approach of detection also varies accordingly. Copy & paste plagiarism, style plagiarism, idea plagiarism, word switch plagiarism, metaphor plagiarism and cross-lingual plagiarisms are some of the main categories [2]. These plagiarism cases mainly belong to the two general categories via, verbatim/literal plagiarism and intelligent plagiarism. In verbatim/literal plagiarism the plagiarized text is the exact copy of source or without any major modifications while in intelligent plagiarism the source content is manipulated and modified using different techniques to form the plagiarized text. Intelligent plagiarism is more complex in nature which includes idea adoption, translation and text manipulations. Sophisticated methods have to be employed for the detection of intelligent plagiarism [2, 3].

The two main generic detection methods used in text plagiarism are: Extrinsic and Intrinsic plagiarism detection methods. Extrinsic method compares a suspicious document with a reference collection, which is a set of documents assumed to be genuine [1]. Intrinsic method solely analyzes the text to be evaluated without performing comparisons to external documents. This approach

aims to recognize changes in the unique writing style of an author as an indicator for potential plagiarism [1]. The proposed method in this paper aligns to extrinsic plagiarism detection method. The presence of reference collections will usually help to find out the plagiarism cases more effectively. But blindly comparing two text documents (source and suspicious) will take a lot of time and effort. Hence algorithms for external plagiarism detection mainly focus on improving the detection efficiency by reducing the number of comparisons [3, 4].

In extrinsic detection there are different methods like character based, vector based, syntax based, semantic based, structural based and fuzzy based methods [1].The work proposed focuses on structural based method for detection of intelligent plagiarism and hence it has to deal with the replacements of words by metaphors, synonyms etc. These features are accomplished using the stop word structure of the documents that are compared to an extent. Enhancements are made by considering rough sentence boundaries along with these stop word structures. An attempt is also made by considering part of speech (POS) tags along with sentence bounded stop word profiles.

The following sections describe the details of the methods and algorithms used in the proposed work. Section 2 focuses on a brief discussion of previous works carried out in the areas of extrinsic plagiarism detection. Section 3 describes the baseline system and methods used. The enhancements applied to baseline method are explained in detail in Section 4 of the paper. Section 5 compares the result sets and its detailed analysis is performed. The work is concluded and future enhancements are proposed in Section 6.

2. Existing Works

Many impressing works are done by eminent researchers in the area of extrinsic plagiarism detection. Classical approaches to plagiarism detection can be broadly classified into following genres via. Grammar based, Semantic based and Grammar-Semantic hybrid based approach. The first method focuses on the grammatical structure of the document and uses a string based matching approach to measure similarity between the documents. HaCohen-Kerner et al. [5] proposes a method that uses this approach in which set of all possible sequential overlapping substrings are calculated. But this method can detect only simple plagiarism cases where plagiarized text is an exact copy of source text. Overlapping n-gram algorithm [6] also come under grammar based method. Here the document is split into sentences and then into word n-grams. Each word n-gram of suspicious document is searched over the reference document and similarity measure is taken. Experiments are conducted on METER corpus where performance of different n-gram comparisons via unigram, bigram and trigram are analyzed. But the problem with this approach is that it cannot detect synonym substitution, paraphrasing etc. Du Zou & Zhang Lin presents a cluster based plagiarism detection method [7] that has three main steps that include pre-selecting, locating and post-processing. Here a clustering approach is used to identify and find the position of plagiarized fragments using Winnowing's fingerprint extraction algorithm [8]. The system is tested on PAN-10 corpus which gives good results with verbatim plagiarism and slight obfuscation cases. The limitation with all these method is that it can detect only verbatim plagiarism

cases and is not suitable for the detection of intelligent plagiarism as it mainly follows a string-based matching approach.

The second approach is semantic based method which overcomes the above mentioned limitation to a good extent. It mainly utilizes the vector space concepts, i.e., the vector space model (VSM) and thereby the word frequency in a document to obtain its feature vector. Further it uses the dot product, cosine product etc to measure the document similarity. Ekbal et al. [9] proposed a method based on VSM mainly used for candidate selections from source documents. Here for the detection of similar passages graph-based approaches are used and finally the false detections are filtered out. Experimental results using PAN-11 corpus give promising results. The main problem is that when the number of common n-grams was not enough, the plagiarism detection performance reduces. VSM can be combined with natural language processing (NLP) approaches for more depth-wise intelligent plagiarism analysis. Graph based method by Osman et al. [10] is an approach which comes under the same category in which similarity check is done by comparing the concept of each sentence. Here a graph is constructed by grouping the terms in each sentence within one node. The resulting nodes are connected to each other based on the order of sentence within the document. All nodes are connected to the top level node called topic signature which is formed by extracting the concepts of each sentence terms and grouping them in this node. The topic signature node is then considered for the comparison between source documents and the suspected one. The method proposed by Kasprzak [11] splits the document into overlapping n-grams or tokens. Then hash values are calculated and mapped to an inverted index containing all attributes. The methods in this category are not always effective to detect cases of partial plagiarism. This approach cannot detect partial plagiarism as it is difficult to find the position of copied text using this method.

The third approach, grammar semantic hybrid method overcomes the problems of the first two approaches. This method gives a better result in the case of complex plagiarism cases like word reordering and rephrasing which cannot be detected by other two methods. It also takes care of the partial plagiarism cases since it can detect the location of the plagiarism passages within a document, which cannot be detected by semantic-based method. Indexing is also an effective method for external plagiarism detection. Muhr et al. [12] proposes a hybrid system that works with translated, non-translated and intrinsic plagiarism. Here the external plagiarism detection system is formulated as an information retrieval problem. The source documents are first split into overlapping blocks and then indexed by a lucene instance which is similar to clustering. The system is evaluated using PAN-10 data set and it works well with different intelligent plagiarism cases also. In the work proposed by Alzahrani et al. [13] structural information is utilized for candidate retrieval and further comparisons. Citation evidence is also considered in this by excluding those cases with proper citation and taking remaining cases as suspected. Experimental results show that the system gives efficient results when compared to the baseline methods via TF-IDF weighting with a Cosine coefficient, and shingling with a Jaccard coefficient. Chiang [14] proposed a plagiarism detection method that uses word-sentence based s-grams. S-gram is an n-gram unit that allows some terms to be skipped. The method gives better results when compared to classical VSM model.VSM with fuzzy-semantic similarity overcomes the limitation of system discussed in [9]. Nassem et al. [15] made such an attempt and the results got

improved when experimented with PAN-12 dataset. Thus grammar-semantic hybrid approach solves the limitation of both grammar-based and semantic-based method.

Considering the three approaches, grammar-semantic based approach is found to be more effective. Majority of the methods explained above focuses on content words of the document and stop words are completely removed in these methods. Stop words are those words which carry no specific semantic relevance in the document, while content words are the semantically important words in the document. One of the works under the grammar-semantic category which explores structural information of the document is reported by Stamatatos [16, 17]. Different from other methods, it extracts the structural information from the document by retaining the stop words and removing the content words. Stop words maintain the stability of a sentence and hence even though the sentence gets manipulated using synonym replacements or certain rearrangements the stop words usually remains the same. Thus stop words directly captures the syntax of a sentence and also indirectly captures its semantics. Using this idea the Stop Word N-grams (SWNG) are made for finding document similarity. The method gives good performance in terms of run time and it gives a clear cut idea about the plagiarized passage boundaries in source and suspicious document. In the proposed work the same idea is explored and possible enhancements are carried out. It is then compared with the baseline method [16, 17] and performance is analysed.

3. Methodology of Baseline System - Stop Word N-gram (SWNG)

In baseline method the structural information is captured using the stop words present in the document. The structure of a sentence usually resides on these stop words. In detection models, comparisons are usually done using content words. Here a method of structural analysis using stop words is considered by forming Stop Word N-grams (SWNG) of the document. By taking the overlapping SWNG of the source document and comparing this against the n-grams of the suspected text their similarities are calculated. Apart from similarity checking, boundary detection of each plagiarized passage in the source and suspected document is done. A similarity check score is used to differentiate the detected passages from true copies to the coincidentally related passages and provides the degree of plagiarism [16]. The architecture of the SWNG method in detail is given in Fig. 1. Here source and the suspicious documents are converted to stop word profiles. All proceedings in the algorithm are done using this stop word profile representation of the documents.

3.1. Pre-processing

In the initial stage certain pre-processing is applied to both source and suspicious documents which finally produces the n-gram profiles. Let D_s be the set of source documents and D_x be the set of suspicious documents. Initially all the text in the document is converted to lower case. Then these documents are compared against a standard list of stop words and all the characters other than stop words

are removed from the documents (50 most frequently occurring stop words as said in British National Corpus[1] is taken as the stop word reference here).

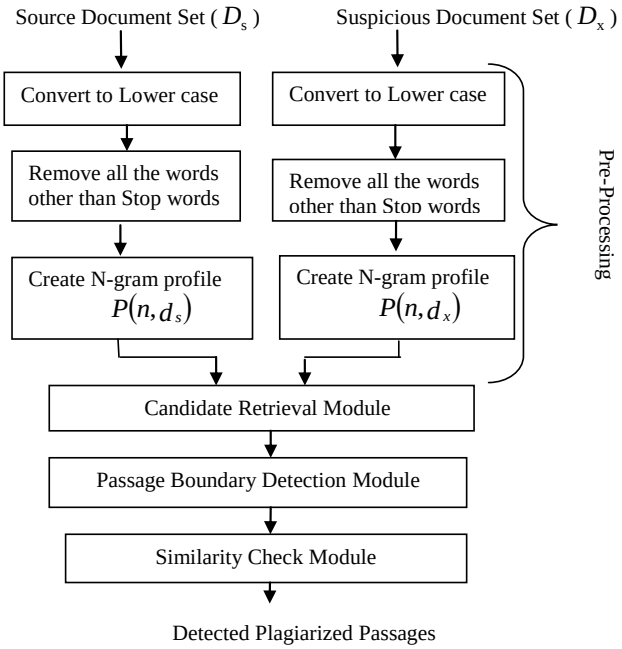

Fig. 1. Flow chart of SWNG method.

The stop word profiles thus obtained are then converted to n-gram, i.e., '*n*' number of stop words are there in each gram. This overlapping grams form the SWNG profiles of source and suspicious document. For a source document d_s, $P(n, d_s)$ is the SWNG profile and for a suspicious document d_x, it is $P(n, d_x)$. Due to the stability of stop words, plagiarism detection can be done efficiently. Figure 2 shows an original text and a plagiarized version of that text and its stop word profile. Despite the fact that the plagiarized version is highly modified, the structure of the sentence remains the same.

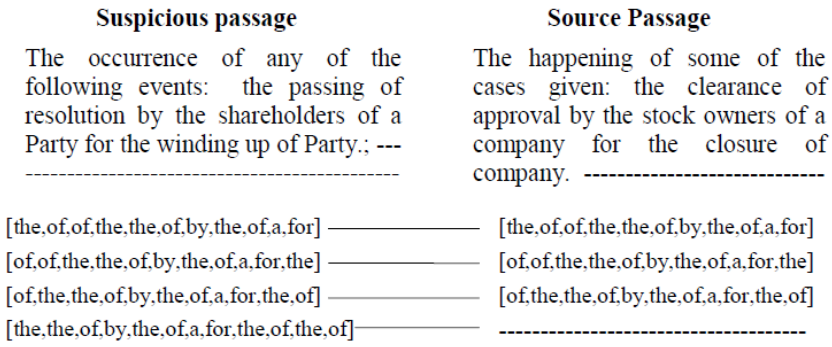

Fig. 2. SWNG representation of text and its similarity.

[1] http://www.natcorp.ox.ac.uk/

3.2. Candidate retrieval

This step includes the comparison of the suspicious document with all the source documents in the corpus in order to identify similarity. A subset of related source documents is then grouped and these retrieved sources are used for further proceedings. For candidate retrieval, the value of *n* is taken as 11(n_1). If the value of '*n*' is too high, then the elimination of gram which is highly modified is possible. Similarly if the value is too small, then there is a possibility of detecting coincidental matches as true cases. To avoid these coincidental similarities, a condition is included for detecting common n-grams. Let 'C*'* be the set of most frequent stop words used, C = {the, of, and, a, in, to,'s}. Let $d_x \in D_x$ and $d_s \in D_s$, then $P(n_1, d_x)$ and $P(n_1, d_s)$ are the corresponding profiles of the suspicious and source document. The passages are said to be plagiarized if and only if the following criteria given in Fig. 3 is satisfied:

for $P(n_1, d_s) \cap P(n_1, d_x)$
  if (*member*(g, C) < n_1 -1 and *maxseq* (g, C) < n_1-2) then
    "Coincidental match not detected"
  else "Coincidental match detected"

Fig. 3. Criteria for coincidental match in candidate retrieval step.

The function *member*(g, C) returns the number of stop words of the n-gram 'g' that belong to C and the function *maxseq*(g, C) will return the longest sequence of stop words of 'g' that belong to C.

3.3. Passage boundary detection

This module detects the boundaries of each passage within the source and suspicious documents which are likely to be plagiarized. Let $D_{rx} \subseteq D_s$ represents the set of source documents that have been retrieved for the suspicious document d_x. The maximal sequence of common SWNG in the profiles of d_x and each $d_s \in D_{rx}$ is found using this step. If the passages are exact copy of the source then it is not difficult to find a match but if the text is highly modified then it will be difficult to find the match. So the value of *n* should be taken in such a way to include all types of plagiarism cases. Here *n* is taken as n_2 and the value is predetermined as 8 ($n_2 < n_1$). A shorter n_2 will give a detailed match of the two profiles. Here the effect of coincidental matches also should be handled. Thus it is considered that C is the set of stop words $P(n_2, d_x)$ and $P(n_2, d_s)$ are the corresponding profiles of the suspicious and source documents respectively. An n_2-gram 'g' is a match between these documents if the following criteria given in Fig. 4 get satisfied. The function *member* (g, C) returns the number of stop words of the n_2- gram 'g' that belong to C.

Let $M(d_x, d_s)$ be the match profile of the source and suspicious document. If the gap between two plagiarized passage boundaries is not too significant, then they are merged. The threshold value taken for this is 100 here, i.e., θ_g = 100. If

the number of stop words in between two plagiarized passages is lesser than 100 then those two passages can be merged and taken as one. Another case that should be handled during detection and merging is the problem that may occur if two plagiarized passages in the source document are kept in a large distance and its corresponding passages in the suspicious document are kept within a threshold. To avoid the above said problem some procedures are assigned. Figure 5 gives the criteria to be followed for this process. Here M_1 represents suspicious passage and M_2 represents the corresponding source passage. Further details can be found in the base paper [11].

for $P(n_2, d_x) \cap P(n_2, d_s)$

if (*member*(g, C) < n_2) then "Coincidental match not detected"

else "Coincidental match detected"

Fig. 4. Criteria for coincidental match in passage boundary detection.

for $m_i \in M(d_x, d_s)$:
if $abs(m_i - m_{i+1}) > \theta_g$:
"Boundary Detected"
else:
"Boundary not detected"

for $m_i \in M(d_x, d_s)$:
if $abs(m_i - m_{i+1}) > \theta_g$: "Boundary Detected"
else:
"Boundary not detected"

Fig. 5. Criteria to find initial set of passage boundary and passage boundary for original passage set.

3.4. Similarity checking

This is the final step in SWNG algorithm. Here the merging errors that might have occurred during the process of passage boundary detection are eliminated. To detect the similarity between the actual texts (in the case of highly modified passages) the characters other than stop words are also taken into consideration. A similarity score is assigned to each passage to get an overview of how much similar are the passages in the real text document. This score is applied only to the passages detected, so that the computational over head is low. The similarity score is calculated using the formula in Eq. (1).

$$Sim(t_X, t_S) = \frac{|P_C(n_C, t_X) \cap P_C(n_C, t_S)|}{\max(|P_C(n_C, t_X)|, |P_C(n_C, t_S)|)} \quad (1)$$

The word n- grams of each passage are obtained and the match profile for this is created and then the similarity score is determined. Here the value of *n* for n-gram is taken as 3. Here $P_C(n_C, t_X)$ be the character n- gram of suspicious document and $P_C(n_C, t_s)$ be the character n-gram of source document. |P| shows the size of *P*. The detected plagiarism cases are considered true only if the

similarity score is above a threshold value of 0.3, as given in [16]. From this formula it is seen that that if the two passages are identical, and then the similarity score is 1. If one of the passages is much longer, then the denominator value will be very high and there by the score will be considerably reduced. This may help to eliminate the cases where adjacent passages are merged with errors.

4. Enhancements on Baseline Method

In baseline method the main limitation is that the similarities are checked using the structure of the text considering the position of stop words irrespective of the sentence boundary. Considering this, a new approach is proposed which forms sentence bounded stop word profiles. Figure 6 shows the high level design of proposed architecture. In the proposed method enhancement is done using two approaches.

- Firstly by considering sentence boundaries and within that the SWNG profiles are made (SBSWNG).
- Secondly by including positions of POS tags of the content words along with SBSWNG.

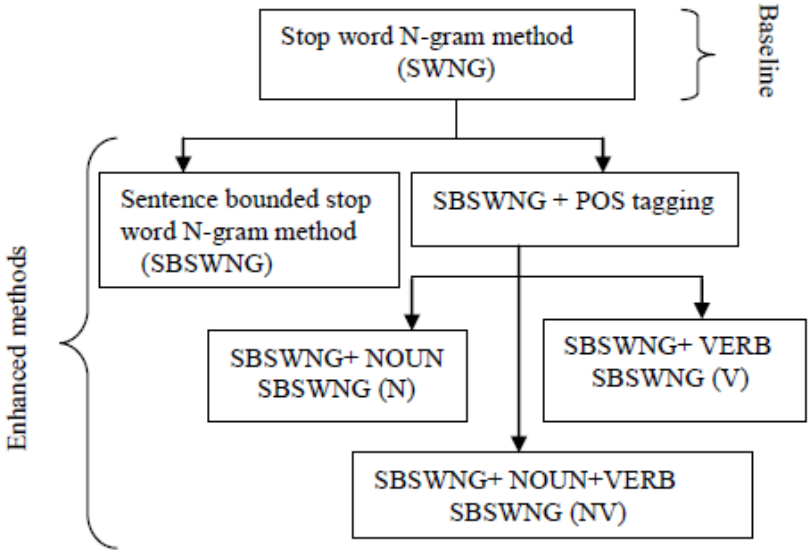

Fig. 6. High level design.

4.1. Sentence bounded stop word N-gram method(SBSWNG)

SBSWNG algorithm is an improved form of the baseline algorithm, i.e. SWNG algorithm [11]. Fig.7 depicts the general flow of SBSWNG algorithm. The SBSWNG algorithm also includes four steps that are discussed in baseline method. Other than pre-processing the remaining steps are same as that of the basic algorithm. But the threshold used in passage boundary detection is different from the baseline algorithm.

In the enhanced method the pre-processing part have two new modules, via sentence boundary identification and gram length calculation. These two modules are discussed below.

4.1.1. Sentence boundary detection

In English script sentence boundary usually get identified using full stops and question marks which are followed with a space. In a plain text where all the sentences are uniformly placed and contain only one full stop there is no problem for sentence boundary identification. But there are also cases where more than one full stop is there in a sentence and they are not sentence boundaries. These conditions occur in the following cases:

(1) If the text contain abbreviations (for e.g. - Jr., Dr., Br., etc.).
(2) If name of a person is mentioned with initials.
(3) If numbering is done using alphabets.

For eliminating the above said conditions full stops are considered as sentence boundaries only if more than one word is identified within a sentence. If only one word is there then full stops are removed and is attached as a part of next sentence and the search will go on to identify the next full stop. This helps in detection of sentence boundaries in a better way even though it cannot be claimed as a complete foolproof method. In the baseline method when a sentence is having less than 8 stop words it is difficult to find a match of the stop word profile. In proposed method since the stop word taken for comparison are sentence bounded grams less than 8 or 11 are also detected. Again when the stop word profiles are not sentence bounded, some of the n-grams become irrelevant due to its grammatical inaccuracy. Sentence bounded stop- word n-grams overcomes this limitation and hence increase the accuracy.

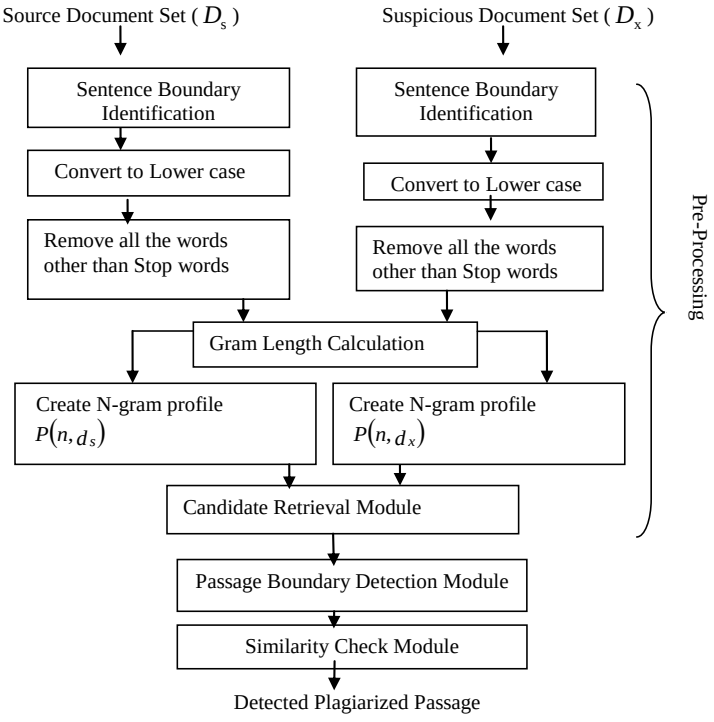

Fig. 7. Schematic diagram of SBSWNG method.

4.1.2. Gram length calculation

In this step the gram length for each sentence from source and suspicious documents are calculated. The input to gram length calculation module is the SWNG profile retaining the sentence boundaries.

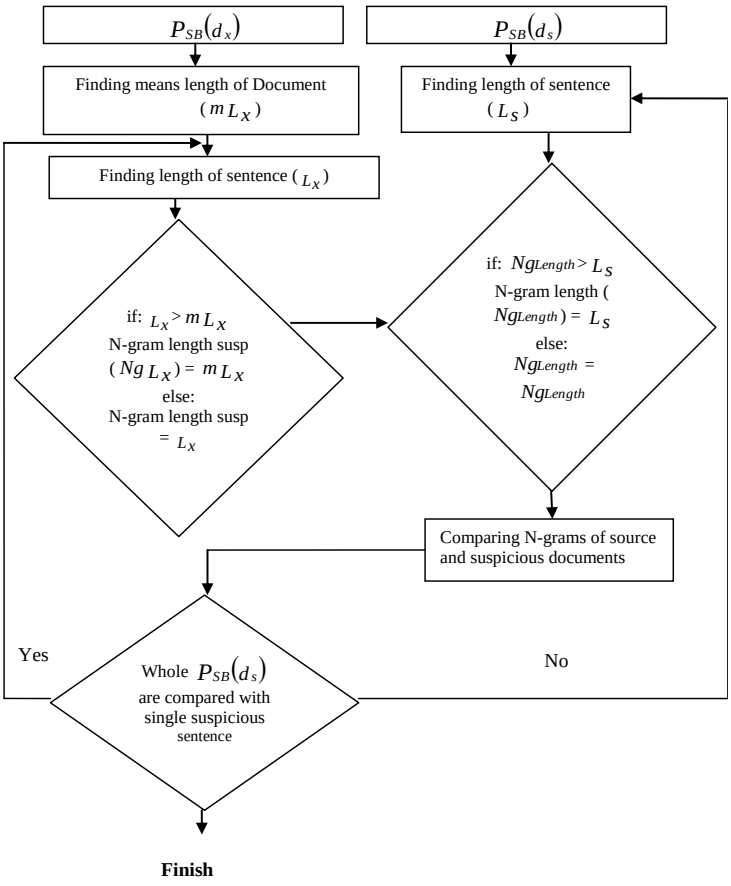

Fig. 8. Gram Length Calculation in SBSWNG Algorithm.

$P_{SB}(d_x)$ is the sentence bounded stop word profile for suspicious document and $P_{SB}(d_s)$ for source document. Stop word profiles for source and suspicious documents are taken and the gram length for each sentence is calculated. The detailed steps are explained using the schematic diagram given in Fig. 8. From the given schematic diagram it is observed that for gram length calculation first the sentence length of each sentence from source and suspicious document is to be calculated. After calculating sentence length the mean sentence length for suspicious document is calculated. If this mean gram length (mL_X) of suspicious document is less than the sentence length (L_X) then the gram length for suspicious ($Ng L_X$) will be L_X otherwise mL_X. After this step the gram length $Ng L_X$ is compared with the source sentence length L_S. If L_S is greater than $Ng L_X$ then

the gram length for comparison (Ng_{Length}) will be Ng_{L_X} otherwise L_s. After calculating gram length, Ng_{Length} for each sentence in suspicious and source, the N-grams are calculated and compared. The threshold used in the passage boundary detection, θ_g is 50 instead of 100. The candidate retrieval part and the passage boundary detection works as same as that of the basic SWNG algorithm.

4.2. SBSWNG with POS tagging

The next approach used is part of speech tagging to identify the structure of sentence. POS tagging is the process of naming out each word in a sentence with its syntactic label using a tag explaining its corresponding part-of-speech like noun, verb, adjective, adverbs, etc. The importance of part of speech in plagiarism is that verbs in a text usually determine the structure of a text. In plagiarism the plagiarist usually replaces the content words with its synonyms, but this will not change the word class. In this work NLTK[2] is used for tagging the text. This algorithm considers the part of speech tag of each content word with the stop words position. Here POS tags are also considered during n-gram extraction based on sentence length. Three versions of the same algorithm are implemented which are given below:

1. Nouns and stop words (SBSWNG (N))
2. Verbs and stop words (SBSWNG (V))
3. Nouns, Verbs and stop words (SBSWNG (NV)).

In the first version of algorithm, positions of stop words and nouns are considered to find the structure information of the text. All the words other than stop words and noun tags are removed from the text, still maintaining the sentence boundary. Gram length calculation for both source and suspicious document is done on this stop word and noun profile. Candidate retrieval and passage boundary detections are done in same way as in the baseline method. In the second version instead of nouns, verb tags are considered and in third both nouns and verbs are considered.

5. Result Evaluations

The baseline method and the proposed methods are evaluated using the documents selected from PAN-10 dataset and the performance is compared using standard PAN measures.

5.1. Data statistics

The evaluation is carried out in four sets of data which is taken from PAN corpus and is discussed in Table 1. Simulated plagiarism cases are those in which passages are obfuscated by humans. In verbatim plagiarism the content is merely exact copies of source documents and artificial plagiarism is produced algorithmically. In artificial high obfuscation the passages are obfuscated with large amount of text operations while in low obfuscation the passages are changed

[2] http://www.nltk.org/

by moderate shuffling. The sample of plagiarized suspicious-source pair texts for these obfuscations is given in Table 2. Due to hardware limitations only partial set of documents from each document type are considered for evaluation.

For evaluating number of n-gram comparisons made, one suspicious document and the respective source documents are selected randomly from each set. The statistics is given in Table 3.

Table 1. Corpus details.

Sets	Document Type	# of suspicious document	# of source document	# of plagiarized passages
1	**Simulated**	20	50	100
2	**Verbatim**	15	30	50
3	**Artificial: High Obfuscation**	30	60	130
4	**Artificial: Low Obfuscation**	50	75	160

Table 2. Sample of plagiarized documents with different obfuscations.

Verbatim	
Suspicious	**Source**
People don't typically discuss bathrooms in association with food but the restroom at Club 33 is special.	People don't typically discuss bathrooms in association with food but the restroom at Club 33 is special.
Artificial	
Suspicious	**Source**
Some attempts of farming were not irregular because of water and small cultivation. Some of them practised rice, but they still continued to depend of successful crops for food.	Some attempts at large scale rice farming were not successful because of an irregular supply of irrigation water and poor knowledge of cultivation methods.
Simulated	
Suspicious	**Source**
As a result of this belief in fairness, honesty, respect for people and the environment, there is a strong bond of mutual trust with the community.	As a solution of this faith in reasonable, reliable, honor for citizens and the surroundings, there is a firm adherence with the society.

Table 3. Sample statistics used for N-gram comparison.

Document Type	# of suspicious document	# of source document
Simulated	1	4
Verbatim	1	3
Artificial: High Obfuscation	1	2
Artificial: Low Obfuscation	1	6

5.2. Measures

The input is a set of source document and its corresponding suspicious document. Algorithm is evaluated on four different measures, Recall (rec), Precision (prec), Granularity (gran) and Plagdet_score (plag) discussed in PAN [18]. Recall is the ratio of matched characters of source and suspicious to the total length of expected plagiarized characters in source and suspicious document.

$$rec(S,R) = 1/|S| \sum_{s \in S} |\cup_{r \in R}(s \cap r)|/|s| \quad (2)$$

S denotes the set of plagiarism cases in the corpus, and R denotes the set of detections for the suspicious documents. 'r' corresponds to the detected characters from source and suspicious plagiarized passage, while 's' is the expected characters from source and suspicious plagiarized passage. Precision is fraction of retrieved document that are relevant to search, here detected plagiarized passage is treated as expected and vice versa.

$$prec(S,R) = 1/|R| \sum_{r \in R} |\cup_{s \in S}(s \cap r)|/|r| \quad (3)$$

Granularity is defined as the ratio of number of detected plagiarized source passage to given plagiarized source passage.

$$gran(S,R) = 1/|S_R| \sum_{s \in S_R} |R_s| \quad (4)$$

Plagdet_score combines recall, precision, and granularity to allow the ranking. The range of plagdet_score is between 0 and 1.

5.3. Evaluations and comparisons

The baseline method (SWNG) and the proposed methods are analysed and compared in this section. In Table 4, the SBSWNG method is compared with baseline method in terms of number of n-gram comparisons.

Table 4. N-gram comparison of SWNG and SBSWNG methods.

Set	Number of N-gram comparison				% of Reduction
	SWNG			SBSWNG	
	11 Gram	8 Gram	Total		
1	86,296,377	86,411,880	172,708,257	48,149,280	72.12
2	1,095,562	1,106,560	2,202,122	890,318	59.57
3	533,923,556	534,201,962	1,068,125,518	188,536,294	82.34
4	7,10,645	7,17,440	14,28,085	6,25,586	56.19

From Table 3, it can be observed that the proposed method takes less number of n-gram comparisons which speeds up the performance of SBSWNG method. It can be analyzed that a considerable reduction in number of comparisons is made using the proposed method with all document sets. Using data statistics in Table 1, the graphical plots of the baseline method and the proposed methods in terms of the standard PAN measures is given in Figs. 9 to 12.

Analysing Fig. 9, it can be observed that with simulated data the proposed SBSWNG method outperforms the other methods in terms of both precision and recall. It can be analyzed that the recall remains less in all methods. This reduced performance may be due to the intelligent obfuscations made by humans in this set which are difficult to detect. Among the methods, SBSWNG method shows the highest recall. Further, comparing the methods using other document sets as shown in Figs. 10 to 12 it can be noted that in all cases the proposed method surpasses the baseline method.

In these sets via verbatim, artificial high and artificial low, compared to baseline method the SBSWNG method shows a considerable increase of 60% to 85-98% in recall. Granularity remains one for all the methods with all the four document sets except for the SWNG method using artificial high obfuscation set. In this method the granularity is slightly higher than 1, which means that some of the unwanted passages are captured by the algorithm. This may be because of the large amount of shuffling and text operations made in this set. The methods integrated with POS tagging via SBSWNG (N), SBSWNG (V) and SBSWNG (NV) also outperform the baseline method but gives reduced precision and recall compared to SBSWNG method. In methods using POS tagging SBSWNG (NV) surpasses the other two methods in all the cases. Observing and analyzing each of these plots, it can be concluded that the proposed SBSWNG method outperforms the other methods in terms of the PAN measures.

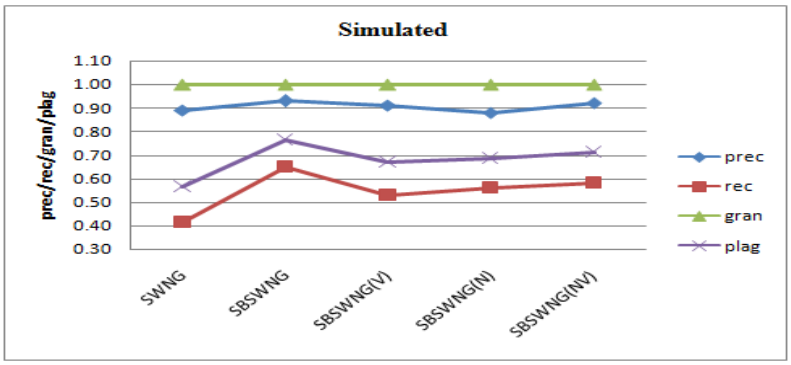

Fig. 9. PAN measures of simulated set.

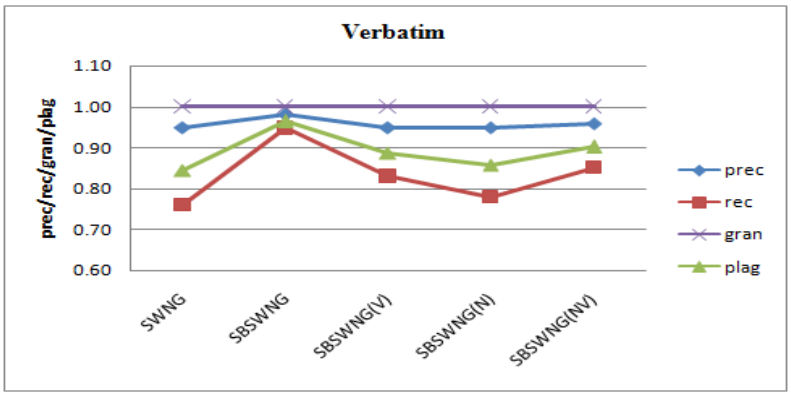

Fig. 10. PAN measures of verbatim set.

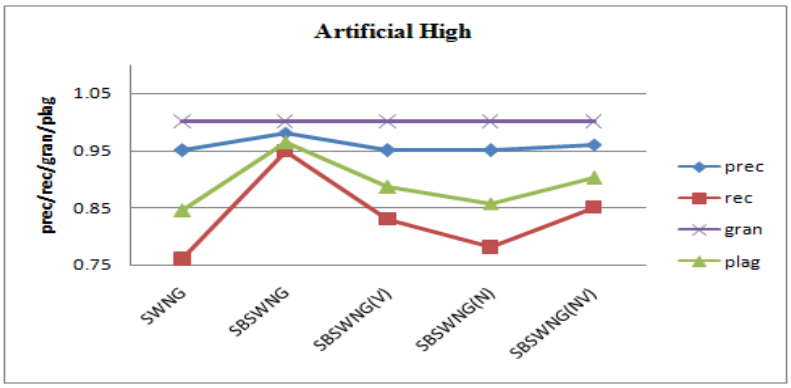

Fig. 11. PAN measures of artificial high set.

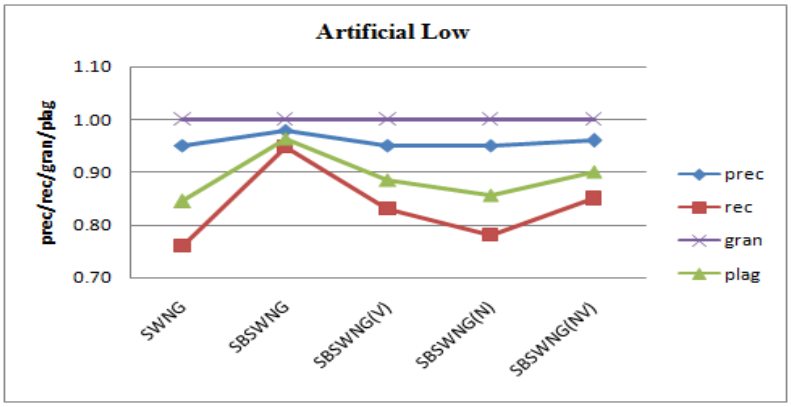

Fig. 12. PAN measures of artificial low set.

6. Conclusion and Future Work

The proposed work focuses on the detection of extrinsic plagiarism by utilizing the structural information from the given text. Initially the baseline/ SWNG method is evaluated, where the structure of a sentence is extracted merely using stop words. After several experimentations and analysis, the limitations of the baseline method are pinpointed. Here firstly, the sentence boundaries are not considered while formation of stop word n-grams. Secondly, the number of comparisons between the source and suspicious document is found to be quite high.

Considering the above drawbacks, the SBSWNG (Sentence Bounded Stop Word N-gram) method is proposed which considers the sentence boundaries. Hence the number of comparisons of the n-grams gets reduced considerably since only the sentence bounded grams are considered here. Further in the proposed method, merging errors are found to be less compared to SWNG method. An attempt is also made to improve SBSWNG method by incorporating POS tag (verb, noun, verb and noun) information along with the stop word positions. From the discussions and analysis, it can be observed that with verbatim plagiarism an overall 20% improvement is attained by the proposed method. With simulated plagiarism 21% gain is noted while with artificial high obfuscation set, a 7% gain is obtained and with artificial low set 8% increase in system performance is

achieved. Thus it can be observed that with all degrees of obfuscations the proposed method presents a significant improvement over the baseline method. With POS tag incorporation the system outperformed the baseline SWNG method, but did not surpass the performance obtained with pure SBSWNG method. Thus from the results and analysis done, it can be analyzed that the proposed SBSWNG outperforms the baseline method with notable improvement in terms of both accuracy and efficiency.

In future, the results can be improved by incorporating different natural language processing (NLP) techniques. NLP techniques like stemming, lemmatization, and chunking can be used so as to make complex intelligent plagiarism detections possible. Further the detection accuracy can be increased by using efficient similarity calculation methods.

References

1. Alzahrani, S.M; Salim, N.; and Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews*, 42(2), 133-149.
2. Ali, A.M.E.T.; Abdulla, H.M.D.; and Snasel, V. (2011). Survey of plagiarism detection methods. *Proceedings of 5t^h Asia Modelling Symposium*, *IEEE Computer Society*. Kuala Lumpur, 39-42.
3. Lukashenko,R.; Graudina,V.; and Grundspenkis, J.(2007). Computer-based plagiarism detection methods and tools: An overview. *Proceedings of International Conference on Computer Systems and Technologies-CompSysTech'07*. New York, USA, 1-6.
4. Osman, A.H.; Salim, N.; and Abuobieda, A. (2012) Survey of text Plagiarism detection. *Journal of Computer Engineering and Applications*, 1(1), 37-45.
5. HaCohen-Kerner, Y.; Tayeb, A.; and Ben-Dror, N.(2010). Detection of simple plagiarism in computer science papers. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING).* Stroudsburg, PA, USA, 421-429.
6. Barrón-Cedeño, A.; and Rosso, P. (2009). On automatic plagiarism detection based on N-grams comparison. *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*. Berlin, Heidelberg, 696-700.
7. Zou, D. Long, W.; and Ling, Z. (2010). A cluster-based plagiarism detection method - Lab report for PAN at CLEF 2010. *Proceedings of the 4th Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*. Padua, Italy, 32-37.
8. Schleimer, S.; Wilkerson, D.S.; and Aiken, A. (2003).Winnowing: local algorithms for document fingerprinting. *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*. New York, USA, 76-85.
9. Ekbal, A.; Saha, S.; and Choudhary, G. (2012) Plagiarism detection in text using vector space model. *Proceedings of 12th International Conference on Hybrid Intelligent Systems (HIS)*. Pune, 366-371.

10. Osman, A.H; Salim, N.; and Salem, M.B. (2010).Plagiarism detection using graph-based representation. *Journal of Computing*, 2(4), 36-41.
11. Kasprzak, J.; and Brandejs, M. (2010). Improving the reliability of the Plagiarism detection system - Lab report for PAN at CLEF 2010.*Proceedings of the 4th Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*. Padua, Italy, 51-62.
12. Muhr, M.; Kern, R.; Zechner, M.; Granitzer, M. (2010). External and intrinsic plagiarism detection using a cross lingual retrieval and segmentation system - Lab report for PAN at CLEF 2010. *Proceedings of the 4th Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*. Padua, Italy, 11-22.
13. Alzahrani, S.M; Palade, V.; Salim, N.; and Abraham, A. (2011).Using structural information and citation evidence to detect significant plagiarism cases in scientific publications. *Journal of the American Society for Information Science and Technology*, 2(64), 286-312.
14. Chiang, M. J. Sci. (2011). Automatic plagiarism detection using word-sentence based S-gram. *Journal of Science*, 3(8), 1-7.
15. Naseem, R.; and Kurian, S. (2013).Extrinsic plagiarism detection in text combining VSM and fuzzy semantic similarity scheme. *Journal of Advanced Computing, Engineering and application (IJACEA)*, 2(6), 112-116.
16. Stamatatos, E. (2011). Plagiarism detection based on structural information. *Proceedings of the 20th ACM International conference on Information and knowledge management*. Glasgow, United Kingdom, 1221-1230.
17. Stamatatos, E. (2011) Plagiarism detection using Stopword N-grams. *Journal of the American Society for Information Science and Technology*. Wiley, 62(12), 2512-2527.
18. Potthast, M.; Barrón-Cedeño, A.; Eiselt, A.; Stein, B.; and Rosso, P. (2010). Overview of the 2nd international competition on plagiarism detection. *Proceedings of the 4th Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*. Padua, Italy, 1-14.