Journal of Engineering Science and Technology
Vol. 12, No. 12 (2017) 3226 - 3250
© School of Engineering, Taylor's University

IDENTIFYING DOCUMENT-LEVEL TEXT PLAGIARISM: A TWO-PHASE APPROACH

VANI K.[1], DEEPA GUPTA[2,*]

[1]Department of Computer Science & Engineering,
[2]Department of Mathematics, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amrita University, Bangalore, India
*Corresponding Author: g_deepa@blr.amrita.edu

Abstract

The rapid evolution of information content and its ease of access have made the field of research and academia so vulnerable to plagiarism. Plagiarism is an act of intellectual theft and information breach which must be restricted to ensure educational integrity. Usually in plagiarism checking, exhaustive document comparisons with large repositories and databases have to be done. The paper presents a two phase document retrieval approach which can effectively reduce the search space for plagiarism detection task. An initial heuristic retrieval process is carried out before the actual exhaustive analysis to retrieve the globally similar documents or the near duplicates corresponding to the suspected document. The work proposes a two phase candidate retrieval approach for an offline plagiarism detection system that can identify the plagiarized sources having different complexity levels. This means that we already have the source document data base offline and hence the work is not focusing on online source retrieval. It explores and integrates the prospective aspects of document ranking approaches with vector space model in first phase and N-gram models in second phase for candidate refinement stage. The proposed approach is evaluated on the standard plagiarism corpus provided by PAN-14 text alignment data set and the efficiency is analyzed using the standard IR measures, viz., precision, recall and F1-score. Comparison is done with the vector space model and N-gram models to analyse the performance efficiency. Further statistical analysis is done using paired t-test with means of F1-scores of these techniques over the samples extracted from the PAN-14 set. Experimental results show that the proposed two phase candidate selection approach outperforms the compared models specifically when it comes to comparison and retrieval of complex and manipulated text.

Keywords: Information retrieval, Plagiarism detection, Candidate retrieval, Document ranking, Two phase approach.

Nomenclatures

$avgdiff$	Average or mean difference
C	Number of N-grams computed
$Cand$	Candidate set
D	Generic document
D_{susp}	Suspicious document set
D_{src}	Source document set
$diff$	Difference between two observations
H_0	Null hypothesis
H_a	Alternate hypothesis
k	No. of top ranked documents
n	Total number of documents
N	N-gram value
$prec$	Precision
rec	Recall
S	Sample size or no. of observations
sd	Standard deviation
se	Standard error
t	Term
$t\ Stat$	t-Statistics
$t\ Critical$	t-Critical value
V	Vocabulary size
X_{src}	Source document
$\overrightarrow{X}_{src}$	Source vector
X_{srng}	Source N-gram profile
X_{spng}	Suspicious N-gram profile
X_{susp}	Suspicious document
$\overrightarrow{X}_{susp}$	Suspicious vector

Greek Symbols

α	Overlap coefficient threshold
β	VSM threshold
γ	N-gram threshold

Abbreviations

CR	Candidate Retrieval
IR	Information Retrieval
NO	No Obfuscation
PDS	Plagiarism Detection System
RO	Random Obfuscation
TO	Translation Obfuscation
TF-IDF	Term Frequency-Inverse Document Frequency
VSM	Vector Space Model

1. Introduction

The breakneck evolution in information technology and communication has increased the availability of bulks of information content in different forms. Further, in the modern world with the developments in information handling and data mining, the access to this knowledge and information has become so easier through various mediums, viz., search engines, digital libraries etc. This in turn has opened the doors to plagiarism which is now a prime concern in different domains, especially in educational fields. In layman's terms plagiarism leads to the use of writings, ideas, innovations, etc. of others and reuse of them partially or completely without proper citation or reference to the source. According to Maurer et al. [1], plagiarism is a serious scholastic misconduct and it can be accidental, unintentional, and intentional or even a self-plagiarism. Liu et al. [2] discussed about the problems especially of non-native English speakers and developed a tutorial system to limit the citation problems. It is indicated in [3] that plagiarism can arise in any field not only in academia and research but even in journalism, literature, politics, business etc. The impact of plagiarism in teaching, learning and research are studied through surveys in [4]. According to Alzahrani et al. [5], the act of plagiarism can range from the simple copy paste of contents to the more complex scenarios where the research works and ideas of a third party is manipulated and published. Here plagiarism emanates as a serious intellectual theft. The former is termed as literal plagiarism and the latter as intelligent plagiarism/ paraphrasing.

With information being easily available through search engines and online digital libraries, the restriction of plagiarism has become a challenging task. To ensure the quality of information content varying from simple student assignments to important projects and research publications, it is quite vital that there should be some mechanism to curb these unprofessional practices. Thus there is an urgent need of intelligent automated plagiarism detection systems (PDS) to curtail this information breach. In any PDS or a plagiarism tool, the main task is to search the large databases to identify the plagiarized sources with respect to the submitted suspicious document. The general text based plagiarism detection tools or systems identify the plagiarized fragments in the given suspected document and their corresponding source counterparts. To do this various algorithms are employed, which perform the exhaustive comparisons of the suspicious document with the sources at different levels. This comparison search space can be reduced effectively if an initial document level retrieval is made which can identify and retrieve the near documents corresponding to the suspicious document at hand. This process of retrieving the near duplicates is termed as candidate retrieval and it is an information retrieval process which can substantially reduce the search space of a PDS [5, 6]. There are mainly two types of PDS; External/Extrinsic PDS and Internal/Intrinsic PDS [7]. In the former, the suspected document is compared with an available set of source documents for detecting the possible plagiarism cases while in the latter case, reference source collection is unavailable. In Intrinsic PDS, the suspected document is analyzed solely, based on authors writing style, structural arrangements of document etc. Mostly the available tools and software follows extrinsic detection as they have the sources available in the form of some repositories specific to the domain or the World Wide Web (WWW) itself.

Candidate document selection is an IR process; where in the documents relevant to a particular query is retrieved similar to IR systems, viz., search engines, question answering (QA) systems etc. [8]. Here the source documents related to the given suspicious query or document has to be retrieved which forms the candidate set for the suspected document at hand. With the initial identification of candidate set, the suspected document has to be compared only with these globally similar source counterparts which reduce the overall computational effort. Hence the results of this stage contribute to the entire system efficiency and accuracy of the PDS. Any document missed in this task will not be considered in further stages. Chong [9] performed binary and multi-class classification of plagiarized and clean documents using PAN-10 corpus for evaluations. Thus author used the corpus information for document-level experiments. Clough [10] conducted a detailed survey on plagiarism techniques and tools along with experiments to evaluate the effect of paraphrasing on these tools. According to the survey, authors claim that to build an effective PDS, a two-step candidate retrieval process followed by exhaustive analysis is required. In the two-step candidate retrieval, authors suggest a fast algorithm to determine near duplicates in the first phase and a better algorithm to prune out further dissimilar documents in second phase for refining the search.

Using this idea, the work proposes a two-phase approach for this candidate retrieval process. The two phases are: (1) Candidate document selection module and (2) Candidate document refinement module. In the initial phase a document ranking approach using VSM is proposed. In the second phase an N-gram model based approach is used to filter the false detections. The method is applied for offline plagiarism detection task, wherein the suspected source document database is already available. PAN-14 text alignment data is used for evaluation of proposed approach and its comparisons. Even though this data set is meant for text alignment task, we have used it for evaluating the proposed CR approach in offline PDS. Since in the PAN text alignment task, the candidate pair information is already available, the performance evaluation can be done more accurately. The performance is evaluated using IR measures, viz, recall, precision and F1-score. The proposed approach is compared with the N-gram and VSM based candidate selection methods which are mostly used in this process. Further, statistical significance tests are carried out and search space complexity analysis is done. In Sect. 2, the work done by different prominent researchers in the related areas and their findings are reported.

2. Literature Survey

N-gram models and VSM are mainly employed for the candidate retrieval task. It is found in the literature that the candidate retrieval task is carried out mainly in two ways (1) Comparing the entire suspicious document with sources available and employing different document comparison techniques (offline) and (2) Formulating queries from suspicious document by extracting relevant key words and then employing the search process (online) [6].

In the initial category, mainly N-gram based works and a few VSM based approaches are reported. A word-level N-gram model is used in [10] where the word-length encoding of texts is done by substituting each word by its length. Jaccard coefficient is used as the similarity metric and the model is evaluated on

PAN[1]-09 corpus. An N-gram based model is used in [12, 13] for both candidate retrieval and detailed comparison stages of plagiarism detection task. A heuristic retrieval process using three approaches is discussed in [14] for cross-lingual plagiarism detection system. Here a key-word based retrieval is used in first two approaches, while fingerprinting with hash-based model is used in the third approach. A search space reduction technique using Kullback-Leibler symmetric distance is proposed in [15]. The evaluation on METER corpus shows that the method reduced the search space of the PDS efficiently. Alzahrani and Salim [16] utilize N-gram models with Jaccard coefficient for detection of the candidate sets. Stop-word N-grams (SWNG) are used instead of content words in [17] for candidate selection and exhaustive analysis. Here overlapping SWNG in suspicious and source documents are analyzed to detect the document similarity. An extension of this model using sentence bounded stop word N-grams is proposed in [18]. AN N-gram frequency approach is utilized for mono-lingual text based detections in [19].

A VSM based PDS for extrinsic as well as intrinsic plagiarism detection is proposed by Zechner et al. [20]. Here the documents are represented using TF-IDF weighting scheme and then the vector similarities are calculated using cosine measure. Further a threshold rule is imposed to select the candidate sources. A similar approach is used in [21] for candidate retrieval task with different parameter settings. A cluster based-approach using K-means algorithm with VSM representation is presented in [22]. Here the source documents that are globally similar to the suspicious document are clustered using cosine similarity metrics. In [23], the clustering approach using Fuzzy C-Means approach is proposed, which avails soft clustering instead of the hard clustering approach in [22]. Thompson and Panchev [24] used inverted indexing and document ranking approaches for candidate document selections. The indexed terms from each suspicious document is selected to retrieve the globally similar source documents from the inverted index table and then ranking is applied to select those sources that contains a certain amount of the indexed terms of the suspicious document.

In the second category, most of the works reported is based on VSM approaches for online source retrieval. Kong et al. [25] proposed a method based on VSM with TF-IDF weighting to extract the key words from the suspicious documents to formulate queries and process them. Then a Lucene ranking method is employed for candidate document reduction. A ranking model for extracting the best key word phrases to formulate suspicious query was presented in [26]. Suspicious queries are constructed by combining each non-overlapping *K* keywords selected by ranking model. Prakash and Saha [27] presented a method based on term-frequency and co-occurrence to extract query terms from non-overlapping chunk of topically related sentences. Suchomel et al. [28] formulated diverse queries from the suspicious document using TF-IDF scores to retrieve the plagiarized sources. Expansion of queries and filtering of duplicate downloads is also employed. A method for extraction of key phrases to form suspicious queries after segmenting the document into topics is proposed in [29] while Elizalde [30] presented a method that uses both independent term and phrasal queries using either TF-IDF frequency or noun phrases. Paragraph chunking followed by the extraction of nouns, adjectives and verbs is used in [31, 32]. Further ranking with

[3]http://pan.webis.de/

TF-IDF values to select top key words for suspicious query was done. Suspicious queries are formulated from top marked sentences and then the noun phrases and key word phrases are extracted by Rafiei et al. [33]. A cross-lingual candidate retrieval algorithm using topic based segmentation of suspicious document is proposed in [34].

The techniques employed in the first category analyse the document as a whole using the text representation models. Then the similarity computation of suspicious and source documents are done at document level to retrieve the near duplicates. In the second category, the main focus is on query formulation techniques to download the most appropriate sources for plagiarism. Many of these discussed works (in second category) are those reported in PAN plagiarism competition, which is an international competition held in plagiarism domain yearly since 2009. The two categories reported are held as separate tasks, viz., text alignment and source retrieval, for the first and second category respectively. Text alignment focuses on more exhaustive comparisons of a PDS while source retrieval mainly focuses on the retrieval of sources based on the formulated suspicious query [7].

The techniques employed in the first category analyse the document as a whole using these text representation models. Then the similarity computation of suspicious and source documents are done at document level to retrieve the near duplicates. In the second category, the main focus is on query formulation techniques to download the most appropriate sources for plagiarism. Most of the discussed works are those reported in PAN plagiarism competition. The two categories reported are held as separate tasks in PAN, viz., text alignment and source retrieval, for the first and second category respectively [7].

From the literature, it is found that either N-gram or VSM approaches or their extensions and variations are mainly employed for candidate retrieval task. Exploring the ideas contributed by the eminent researchers, it is found that both N-gram and VSM models individually can contribute to the detection of near duplicates or candidate retrieval task. But the cogency of these models in combination is not explored for the detection of candidate documents in existing works. Further, most of the existing techniques utilized a single step process for this stage. As discussed in Section 1, candidate retrieval will be more efficient if it is implemented as a two-stage process with appropriate techniques. This helps in providing faster and efficient comparisons in the subsequent stages of plagiarism detection task. The proposed work aims to explore the idea of combining the potencies of the N-gram and VSM based approaches for availing a two phase candidate retrieval task that can prune out maximum false detections and retrieve the near duplicates effectively. The current work is implemented as an offline candidate retrieval task (first category). It utilizes a two-phase approach where in phase 1 (candidate document selection), a document ranking approach using TF-IDF weighting schemes is used. In phase 2 (candidate document refinement), for pruning out further false detections, N-gram based models are utilized. As the proposed model focuses on the first category of PAN task, evaluation is carried out using PAN- 14 text alignment data set.

In Section 3, the comparative baselines models are discussed and in Section 4 the proposed method is explained. Section 5 describes the data statistics and measures used for the model evaluation and in Section 6, results are analyzed, discussed and compared. Finally in Section 7 the work is concluded and future aspects of the work are reported.

3. Comparative Baseline Models

The comparative baseline models used in the proposed work are briefly discussed in this section.

3.1. N-gram model

N-gram is a traditional model widely used in candidate retrieval task of extrinsic plagiarism detection process. Let D_{susp} and D_{src} be the set of suspicious and source documents represented as $\{X_{susp1}, X_{susp2},.., X_{suspn}\}$ and $\{X_{src1}, X_{src2},.., X_{srcm}\}$ respectively. Now let $X_{suspj} \in D_{susp}$ be the j^{th} suspicious document and $X_{srci} \in D_{src}$ be the i^{th} source document. The general naming conventions remain the same throughout the paper. Each of these documents is converted into N-gram profiles after required pre-processing. In N-gram representation, the documents are converted to N consecutive words where N a user specified value. Let X_{spngj} and X_{srngi}represents the N-gram profiles of X_{suspj} and X_{srci} respectively. Then the suspicious and source N-grams are compared and similarity is calculated using some metrics, viz, Dice coefficient, Jaccard coefficient, etc. Here Dice Coefficient is employed as the metric for experimentations and is computed using Eq. (1). Numerator here denotes twice the number of shared N-grams and denominator the sum of N-gram lengths.

$$Dice(X_{spngj}, X_{srngi}) = \frac{2\left|X_{spngj} \cap X_{srngi}\right|}{\left|X_{spngj}\right| + \left|X_{srngi}\right|} \tag{1}$$

After calculation of similarity using Eq. (1), a threshold (γ) is defined and the source documents with a similarity greater than the threshold is selected for candidate set of particular suspicious document.

3.2. Vector space model (VSM)

In this approach, the source and suspicious documents are converted to vector representations with some feature reduction techniques. Here TF-IDF weighting scheme is used and computed using the Eqs. (2), (3) and (4).

$$tf - idf(t,d,D) = tf(t,d) * idf(t,D) \tag{2}$$

$$tf(t,d) = f(t,d) \tag{3}$$

$$idf(t,D) = \log \frac{n}{\left|\{d \in D; t \in d\}\right|} \tag{4}$$

In Eq. (2), $tf - idf$ represents the TF-IDF weight computed as the product of $tf(*)$ and $idf(*)$ using Eqs. (3) and (4) respectively. $tf(t,d)$ represents the term frequency, i.e., the frequency of the term 't' in document 'd'. In Eq. (4), 'n' represents the total number of documents in the given data set for source and suspicious documents respectively. Here the denominator represents the number of documents where the term 't' appears. Once the documents are represented using TF-IDF scheme, the similarity computation is done. Let $\overrightarrow{X}_{suspj}$ and $\overrightarrow{X}_{srci}$ be the vector space representations of the suspicious document X_{suspj} and source document X_{srci} respectively. Then the generally used cosine similarity metric is employed, to compute the similarity between the document vectors using Eq. (5).

$$Cos(X_{suspj}, X_{srci}) = \frac{\vec{X}_{suspj} \bullet \vec{X}_{srci}}{\left\|\vec{X}_{suspj}\right\| \left\|\vec{X}_{srci}\right\|} \tag{5}$$

The numerator in Eq. (5) denotes the dot product of these document vectors and the denominator denotes the product of their Euclidean norms. Then all the source documents with a similarity greater than a threshold β are considered as candidates of the suspicious document at hand.

4. Proposed Model- Two-Phase Candidate Document Retrieval Approach (Two Phase-CR)

The proposed two-phase approach attempts to exploit document ranking using VSM with TF-IDF scheme and candidate refinement using N-gram approach. Suspicious and source documents are initially subjected to some of the common pre-processing steps, viz., tokenization, stop word removal, and lemmatization. Initially the documents are converted to word tokens and stop words are pruned out. Then the remaining tokens are converted to their dictionary base forms, i.e., lemma by lemmatization.

The proposed two phase-CR algorithm is discussed in Algorithm 1. It is implemented in two phases, viz, Phase I and Phase II. Phase I uses a VSM representation with document ranking approach and Phase II utilizes an N-gram model. As discussed, let D_{susp} be the suspicious document set and D_{src} the source set. $\vec{X}_{suspj}$ and $\vec{X}_{srci}$ are the vector representations of documents $X_{suspj} \in D_{susp}$ and $X_{srci} \in D_{src}$ respectively. In Phase I, after the TF-IDF representation of documents, the cosine similarity for a suspicious document to all the source documents is computed. Then the IR ranking method is adopted where the 'k' top ranked source documents which are most related to the suspicious document at hand is selected. This can be done in terms of maximum similarity or minimum distance computation. Thus each suspicious candidate set will have 'k' top similar source documents in it. This forms the initial candidate sets for each suspicious document and completes the Phase I of proposed algorithm.

Algorithm 1: Two-Phase Candidate Document Selection Approach

```
Input: $D_{susp}$ & $D_{src}$; Suspicious & Source document collections
Output: Candidate Sets, $cnd$ for each suspicious document
Phase I- Candidate Retrieval Process
    1. Convert each $X_{suspj} \in D_{susp}$ & $X_{srci} \in D_{src}$ into vectors, $\vec{X}_{suspj}$ & $\vec{X}_{srci}$ ; j=1 to m & i=1 to n
    2. For each suspicious vector $\vec{X}_{suspj}$ do steps 3 to 5.
    3. Compute the cosine similarity $Sim = Cos(*)$ with all $\vec{X}_{srci}$ using Eq. (5)
```

```
4. Retrieve $k$ top ranked source documents with maximum $Sim$
5. Store output of Step 4 in the candidate set;
   $Cand(X_{suspj}) = \{X_{src1}, X_{src2,..,} X_{srck}\}$
Phase II- Candidate Refinement Process
1. For each $X_{suspj}$ do step 2 ; j=1 to m
2. Compute N-gram profiles for $X_{suspj}$ and all $X_{srci} \in Cand(X_{suspj})$
3. For each suspicious-source N-gram pair $(X_{spngj}, X_{srngi})$ do steps 4 and 5;
   j=1 to m & i=1 to k
4. Compute $Overlap(*)$ using Eq. (6).
5. If $(1 - Overlap(*)) \geq \alpha$ : Remove $X_{srci}$ from $Cand(X_{suspj})$
```

Once the initial candidate sets are formed, the Phase II is carried out. Here a filtering stage is imposed which utilizes the N-gram model and helps in pruning out the unrelated documents in each candidate set. Thus the process facilitates to improve the accuracy of the candidate task by reducing false detections. Each of the candidate sets formed using the VSM ranking approach in Phase I is considered here and then the source and suspicious documents are converted to N-gram profiles. Further the document similarity is computed using the overlap coefficient measure as discussed in Eq. (6).

$$Overlap(X_{spngj}, X_{srngi}) = \frac{\left|X_{spngj} \cap X_{srngi}\right|}{min\left(\left|X_{spngj}\right|, \left|X_{srngi}\right|\right)} \tag{6}$$

The terminologies are similar to Eq. (1). The overlap coefficient computes the ratio of the number of matched tokens between source and suspicious N-grams to the minimum of the length from the two compared N-grams. Then the source documents with an overlap distance (1- overlap similarity) greater than the threshold α are pruned out from the particular suspicious document's candidate set. After completion of Phase II, the final candidate sets are obtained which completes the candidate retrieval stage.

It is suggested that in candidate retrieval stage, recall must be high while maintaining a good precision/ F1-score. This is because if recall is less it means some documents are missing and these documents will not be considered in the further comparisons which degrades the overall system performance. But at the same time a good precision should also be maintained. This is because as precision decreases, false detection increases and a very less precision indicates that detailed comparison has to be done with many unwanted documents which actually nullify the incorporation of this search space reduction task. Thus a high recall and an average F1-score is considered usually. The threshold α facilitates in filtering of unrelated documents in Phase II. In the proposed work, the main focus is given in recall improvement while maintaining a good F1-score.

5. Data Statistics and Standard Measures

The algorithms discussed are evaluated on the entire PAN-14 training and test data sets for text alignment task. PAN is an international competition to find out the best systems for plagiarism detection. In PAN, the dataset is categorized based on the level of complexity. Complexity in turn is defined on the basis of obfuscations imposed on the data [35, 36]. The data sets used for the evaluation in this work are, viz, NO: No Obfuscation, RO: Random Obfuscation and TO: Translation Obfuscation. In No Obfuscation set, data is not obfuscated, which means it contains document pairs where the suspicious documents are exact copies of the source document. This is also termed as literal/ copy-paste plagiarism. Random Obfuscation is a simple approach to obfuscation in which intelligent text operations are done at random. The operations such as shuffling, adding, deleting and replacing words are performed here. In Translation Obfuscation, the given text is run through a series of translations, so that the output of one translation becomes the input to the next one and the final language in the sequence is the original text language. The data statistics of the complete training and test data used are given in Table 1.

For computing the performance efficiency, the standard IR measures, viz, recall, precision and F1-score is used as given in Eqs. (7), (8) and (9) respectively. Instead of general PAN measures, IR measures are used here, since the evaluation of candidate retrieval stage alone is done which is an IR task.

$$rec = \frac{\#\,of\ relevant\ documents\ retrieved}{Actual\,\#\,of\ relevant\ documents} \tag{7}$$

$$prec = \frac{\#\,of\ relevant\ documents\ retrieved}{Total\,\#\,of\ \ documents\ retrieved\ by\ system} \tag{8}$$

$$F1 - Score = 2 * \frac{(rec * prec)}{(rec + prec)} \tag{9}$$

Recall is defined as the number of relevant documents retrieved to the actual number of relevant documents to be retrieved. Precision is defined as the number of relevant documents retrieved to the total number of documents retrieved by the system. F1-score defines the harmonic mean of precision and recall or a weighted average of these two measures.

Statistical analysis tests are also done to evaluate the significance of proposed results over comparative baselines. This is performed using t-test. For statistical analysis using t-test, initially the null hypothesis (H_o) and the alternate hypothesis (H_a) is stated. Next t- statistic, *t Stat* is computed. For this the mean/average difference $avgdiff$ is computed using Eq. (8) which is the average of differences in F-score ($F1 - Score$) over all the samples. Here $diff = a_i - b_i$, where a_i and b_i refers to the plagdet obtained by the i^{th} samples of the compared techniques. S is the number of observations.

$$avgdiff = \frac{\sum_{i=1}^{s} diff_i}{s}. \tag{10}$$

Then the standard deviation (*sd*) is computed using $avgdiff$ as given in Eq. (9). Further standard error (*se*) is computed using Eq. (10) and then the *t Stat* is calculated using Eq. (11).

$$sd = \frac{\sum_{i=1}^{s} \left(diff_{i} - avgdiff \right)^{2}}{s-1} \tag{11}$$

$$se(avgdiff) = \frac{sd}{\sqrt{s}}. \tag{12}$$

$$t\ Stat = \frac{avgdiff}{se(avgdiff)} \tag{13}$$

Using the Students t distribution table, the t Critical value, *t Critical*, is also computed. Now if the *t Stat* > *t Critical*, reject the null hypothesis and accept the alternate hypothesis, otherwise do the reverse.

Table 1.Data statistics.

Training Sets	**# of Suspicious Documents**	**# of Source Documents**
NO	170	923
RO	166	942
TO	162	938
Test Set1		
NO	70	108
RO	67	93
TO	69	103
Test Set2		
NO	166	906
RO	170	931
TO	161	922
Test Set3		
NO	179	1427
RO	176	1395

Training Sets	**# of Suspicious Documents**	**# of Source Documents**
NO	170	923
RO	166	942
TO	162	938
Test Set1		
NO	70	108
RO	67	93
TO	69	103
Test Set2		
NO	166	906
RO	170	931
TO	161	922
Test Set3		
NO	179	1427
RO	176	1395

6. Results and Discussions

The proposed algorithm is evaluated on the datasets described in Table 1. The algorithm is compared with the comparative baselines discussed in Sect. 3 and their performance efficiency is evaluated.

6.1. Parameter selection

For proper selection of the parameter values, evaluation of the algorithms on training set is done. Initially the parameters for comparative baseline are selected.

6.1.1. Baseline parameter selection

Both the comparative baselines approaches (N-gram and VSM) are evaluated on the training dataset. The evaluation is conducted using different values of the parameters that regulate these algorithms to analyse and decide the best parameter values. Initially, N-gram model is evaluated on the training sets, using different N values and threshold (γ) values. Figures 1, 2 and 3 plot the IR measures, viz recall, precision and F1-score with N = 3, 5, 8 and 10 and multiple γ values on No Obfuscation (NO), Random Obfuscation (RO) and Translation Obfuscation (TO) sets respectively. With NO set in Fig.1, the variations are less prominent due to the low plagiarism levels. In Figs. 2 and 3 the performance of the sets RO and TO are plotted where the obfuscation complexity is high. With respect to both these sets, it can be clearly observed that the recall is maximum at N = 3 and drops as the N value increases. Now analysing the threshold values, in all cases beyond γ = 0.1, the F1-score drops substantially.

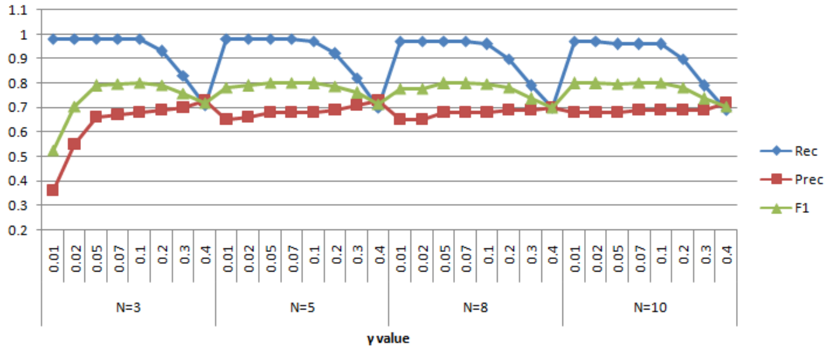


Fig.1. Evaluation of N-gram model using different N and γ values on training set NO.

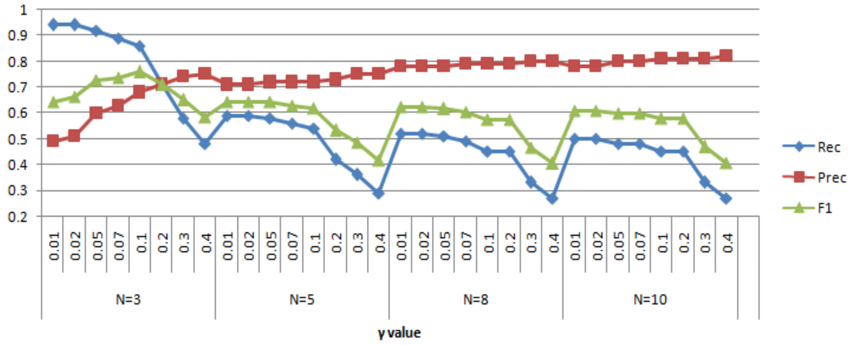


Fig.2. Evaluation of N-gram model using different N and γ values on training set RO.

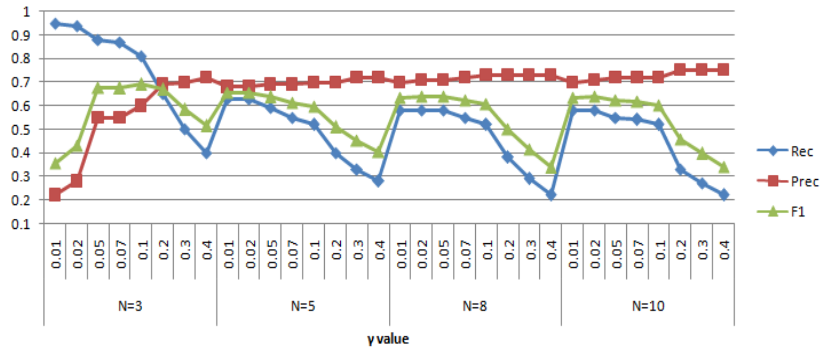


Fig.3. Evaluation of N-gram model using different N and γ values on training set TO.

With γ = 0.02, a high recall and good F1-score is exhibited by set NO and RO but with set TO, it is found that with this γ value the precision is very low, about 28% which reduces the F1-score. Considering a γ value that presents a high recall but at the same time exhibits a good F1-score, γ = 0.05 is selected as the optimal threshold value for all the sets. The selection should be such that there is a good performance with all sets while preventing the over fitting with training data. Considering these factors, for the further testing and comparisons, N=3 and γ = 0.05 is fixed.

Secondly, the VSM candidate retrieval algorithm is evaluated on the training set using multiple threshold (β) values. Figure.4 plots the performance of the VSM with different β values on the training set.

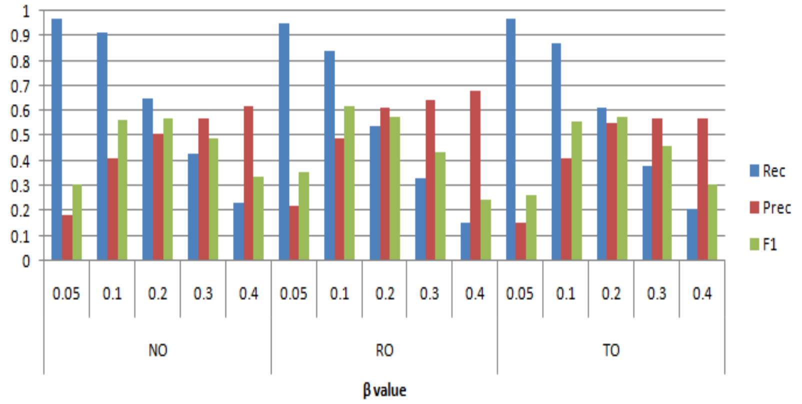


Fig.4. Evaluation of VSM algorithm using different β values on training sets.

From the plot, it is clearly evident that with all the sets, as the value of β decreases, recall increases. It can be noted that with β = 0.05, the recall is above 95% for all the sets while precision is only 18% with set NO, 22% with set RO and only 15% precision with set TO which reduces the F1-score highly. With β =

0.1, a substantial improvement in F1-score is noted while still maintaining a high recall above 85% with all the sets. Now with $\beta \geq 0.2$, it is clearly visible that the recall drops considerably in all sets. As discussed, considering high recall and a good F1-score, $\beta = 0.1$ is selected for further evaluations.

6.1.2. Proposed two phase-CR parameter selection

The proposed Two Phase-CR algorithm is evaluated for parameter settings, viz, k and α. Initially evaluation with different k values is done to determine the number of top ranked source documents to be selected in each candidate set in Phase I. This value thus regulates the IR ranking method of the proposed algorithm (Algorithm 3). Figure 5 depicts the performance exhibited by the proposed algorithm when evaluated using different k values for each training data set. It can be observed from Fig. 5, that with all sets recall increases as the k value increases up to k =25, and beyond that it drops. In candidate retrieval task, as discussed the main focus is given on improving recall while maintaining a good F1-score which is in turn contributed by a reasonable precision value. Considering the behaviour of all the three sets with respect to the plotted recall and F1 curves, k =25 is decided as the best value for the proposed algorithm.

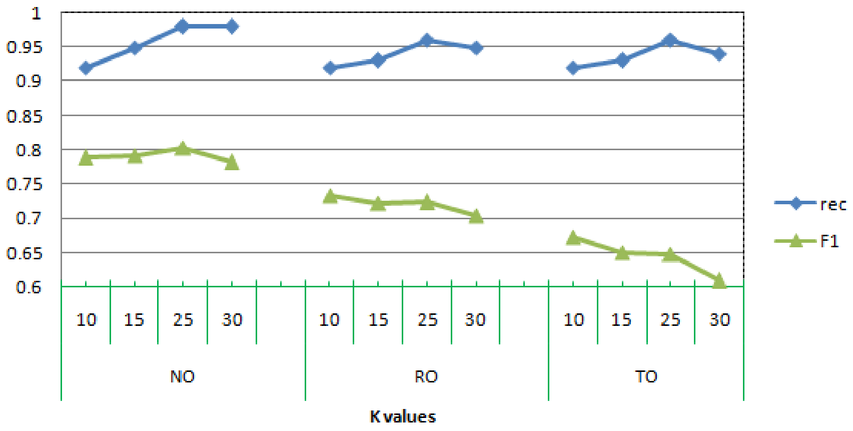


Fig.5. Evaluation of two phase-CR algorithm using different k values with training sets.

Table 2.F1-score for each set based on different α values.

	Recall, F1-Score		
α	NO	RO	TO
0.95	0.96, 0.775155	0.9, 0.690411	0.92, 0.622158
0.96	0.97, 0.785521	0.91, 0.693333	0.93, 0.624429
0.97	0.97, 0.792561	0.91, 0.700946	0.93, 0.633191
0.98	0.98, 0.795879	0.92, 0.703893	0.95, 0.628873
0.99	0.98, 0.802892	0.92, 0.711467	0.97, 0.633194
1.0	0.98, 0.78878	0.92, 0.696216	0.96, 0.621972

Further as discussed in Algorithm 3, the threshold α regulates the filtering in Phase II of proposed algorithm. To find out the optimal value of α, evaluation is done with different αvalues. The (Recall, F1-score) obtained by the two phase-CR algorithm with different αvalues is reported in Table 2. From Table 2, it can be observed that with all sets, α= 0.99 exhibits the maximum F1-score compared to other values even though recall remains almost the same. The parameter values selected for each algorithm based on experimentations are summarized in Table 3.

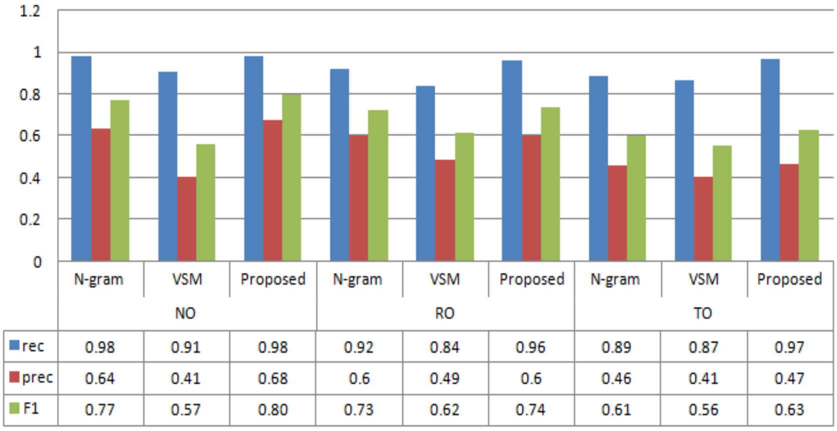
Table 3.Parameters fixed for the comparative baselines and the proposed approach.

Algorithm	Parameters Selected
N-gram	N =3, γ = 0.05
VSM	β = 0.1
Two phase-CR (Proposed)	K=25, α = 0.99

With these optimal parameter values, the two phase-CR algorithm and the compared N-gram model and VSM are evaluated on the PAN-14 training and test sets. Further, their efficiency is computed using the IR measures and the performance is compared.

6.2. Comparison of proposed approach with comparative baselines

Figure 6 depict the performance comparison of the proposed and compared approaches when evaluated on the training sets using the tuned parameters discussed in Table 2.



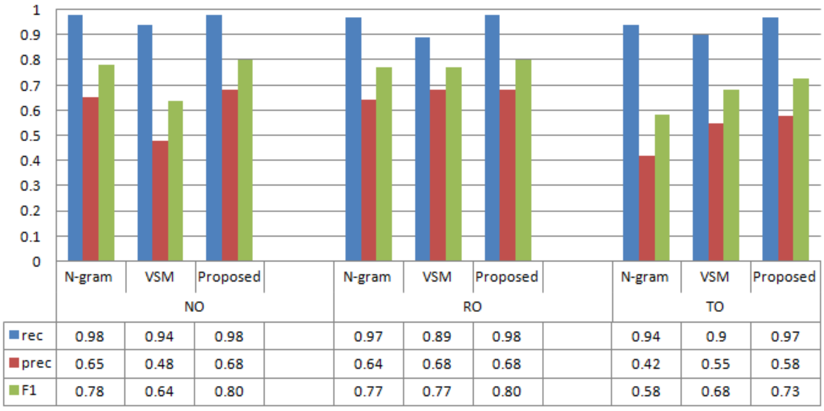
	N-gram	VSM	Proposed	N-gram	VSM	Proposed	N-gram	VSM	Proposed
	NO			RO			TO		
🟦 rec	0.98	0.91	0.98	0.92	0.84	0.96	0.89	0.87	0.97
🟥 prec	0.64	0.41	0.68	0.6	0.49	0.6	0.46	0.41	0.47
🟩 F1	0.77	0.57	0.80	0.73	0.62	0.74	0.61	0.56	0.63

Fig. 6. Comparison of proposed approach with comparative baselines over training sets.

It is observed that using the set with no obfuscations (NO), compared to VSM, the proposed approach presents 7.69% improvement in recall and 65.85% increased precision. In all the cases, the relative improvement is measured and expressed in terms of % gain. The improvement can be clearly noted with F1-

score where 40.35% gain is visible with proposed model. With TO and RO set, a recall improvement of 14.28% and 11.49% respectively is noted. Similarly with these sets, F1-score improves by 19.35% and 12.5% respectively using the two phase model. Thus it is quite obvious that with all types of obfuscations, the two phase algorithm outperforms the VSM approach. Considering the N-gram approach, it is noted that with set NO, the recall remains the same while 3.89% gain in F1-score is exhibited. While with sets RO and TO, a recall gain of 4.34% and 8.99% respectively is obtained. It is observed that F1-score also increases in all sets, even though the gain is only a few percentages.

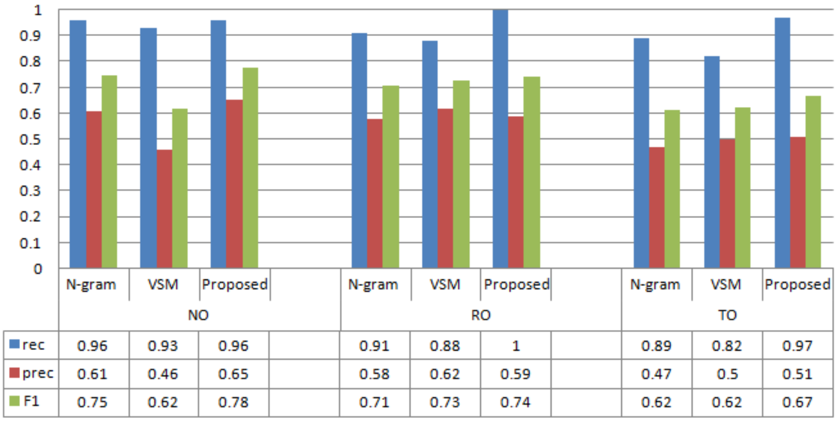
To assess the consistency of the proposed approach, the algorithms are evaluated on the three test sets provided by PAN-14 plagiarism corpus. Figures 7, 8 and 9 plot the recall and precision of each algorithm using Test Set1, Test Set2 and Test Set3 respectively. In Test Set3, PAN-14 provides only two of the data sets, viz, NO and RO sets. From Fig. 7, with respect to VSM approach, the proposed two phase algorithm presents a recall gain of 4.26% in set NO, 10.11% in set RO and 7.78% in set TO. Observing the F1-score, it is visible that with all sets two phase model outperforms VSM approach. Now, compared to the N-gram approach, the proposed algorithm exhibits a recall gain of 3.19 % with set TO and in terms of F1-score the improvement is noted with all sets. With complex obfuscations, the gain is more evident, as with TO set 25.86% increase is presented compared to N-gram model. Considering the Test Set2 results, it is clearly noticeable from Fig. 8 that the proposed two phase-CR algorithm outperforms the VSM approach considerably with all the three sets having various degrees of obfuscations.



	NO N-gram	NO VSM	NO Proposed	RO N-gram	RO VSM	RO Proposed	TO N-gram	TO VSM	TO Proposed
🟦 rec	0.98	0.94	0.98	0.97	0.89	0.98	0.94	0.9	0.97
🟥 prec	0.65	0.48	0.68	0.64	0.68	0.68	0.42	0.55	0.58
🟩 F1	0.78	0.64	0.80	0.77	0.77	0.80	0.58	0.68	0.73

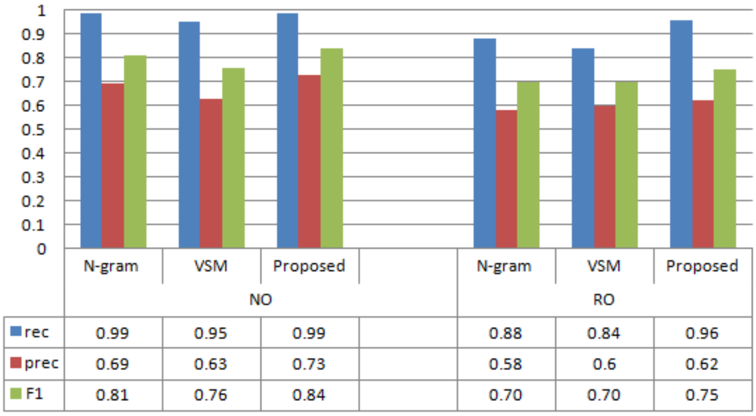
Fig.7. Comparison of proposed approach with comparative baselines using Test Set1.

A recall gain of 3.25% in NO set, 13.63% in RO set and 18.29% in TO set is observed. With N-gram model, the gain in recall is clearly visible with RO and TO sets. A 9.89% gain in set RO and 8.9% in set TO is observed and with all sets the F1-score improves. Similar observation can be made from Fig.9 which plots the performance with Test Set3. Here compared to VSM approach, 4.21% recall gain with set NO and 14.2% gain with set RO is noticed. Hence it is obvious that the two phase approach surpasses the VSM method in all these sets in terms of overall performance. Compared to N-gram model also the proposed two phase-CR algorithm presents an improvement in overall performance.



	N-gram	VSM	Proposed		N-gram	VSM	Proposed		N-gram	VSM	Proposed
		NO				RO				TO	
🟦rec	0.96	0.93	0.96		0.91	0.88	1		0.89	0.82	0.97
🟥prec	0.61	0.46	0.65		0.58	0.62	0.59		0.47	0.5	0.51
🟩F1	0.75	0.62	0.78		0.71	0.73	0.74		0.62	0.62	0.67

Fig. 8.Comparison of proposed approach with comparative baselines using Test Set2.



	N-gram	VSM	Proposed		N-gram	VSM	Proposed
		NO				RO	
🟦rec	0.99	0.95	0.99		0.88	0.84	0.96
🟥prec	0.69	0.63	0.73		0.58	0.6	0.62
🟩F1	0.81	0.76	0.84		0.70	0.70	0.75

Fig.9. Comparison of proposed approach with comparative baselines using Test Set3.

From the analysis and discussions, it is obvious that the proposed two phase algorithm exhibits an improvement in overall performance compared to both N-gram and VSM approaches. This is visible throughout all the data sets, in both training and test sets with all types of obfuscations. Among the compared approaches, N-gram approach exhibits a high accuracy in performance compared to VSM approach. In terms of overall performance, the two phase-CR algorithm surpasses the N-gram method but in some cases the gain in terms of recall is less. But still comparing the complexity of algorithm in terms of number of comparisons required, the proposed model requires lesser number of comparisons compared to N-gram model.

6.3. Statistical test results

To understand the statistical significance of proposed approach over the baseline approaches, initially the suspicious documents from the different sets are randomly sampled without replacement. The source data base remains the same which is provided by PAN-PC-14 text alignment corpus for each set. A total of 11 samples are extracted upon which paired t-test on the mean differences of F1-score is done.

Table 4. Statistical analysis of proposed approach & VSM using sample paired t-test.

Proposed (F1)	VSM(F1)	diff	Statistics	Two tailed test
0.8	0.57	0.23	H_0	Two phase-CR = VSM
0.74	0.62	0.12	H_a	Two phase-CR ≠ VSM
0.63	0.56	0.07	α level	0.05
0.8	0.64	0.16	Sample size, S	11
0.8	0.77	0.03	$avgdiff$	0.001123
0.73	0.68	0.05	sd	0.067501
0.78	0.62	0.16	se	0.020352
0.74	0.73	0.01	df	10
0.67	0.62	0.05	t Stat	± 4.511447
0.84	0.76	0.08	P(T<=t) two-tail	0.001123
0.75	0.7	0.05	t Critical	2.228139

Decision: Accept alternate hypothesis, H_a
Confidence Level (95%): 0.045348
Confidence interval: 0.04647< $avgdiff$ < 0.137166

Table 5. Statistical analysis of proposed approach & N-gram using sample paired t-test.

Proposed (F1)	N-gram (F1)	diff	Statistics	Two tailed test
0.8	0.77	0.03	H_0	Two phase-CR = N-gram
0.74	0.73	0.01	H_a	Two phase-CR ≠ N-gram
0.63	0.61	0.02	α level	0.05
0.8	0.78	0.02	Sample size, S	11
0.8	0.77	0.03	$avgdiff$	0.040909
0.73	0.58	0.15	sd	0.038067
0.78	0.75	0.03	se	0.011478
0.74	0.71	0.03	df	10
0.67	0.62	0.05	t Stat	± 3.564252
0.84	0.81	0.03	P(T<=t) two-tail	0.005145
0.75	0.7	0.05	t Critical	2.228139

Decision: Accept alternate hypothesis, H_a
Confidence Level (95%): 0.025574
Confidence interval: 0.015335< $avgdiff$ <0.066483

The statistical significance test is then carried out over the given combinations: (1) Proposed & N-gram (2) Proposed & VSM. For a particular sample, the difference in F1-score is computed. The Null hypothesis, H_0, and alternate hypothesis, Ha, is stated.

H_0*: Proposed = Comparative Baseline (N-gram/VSM) i.e., proposed approach results are statistically insignificant*
Ha*: Proposed ≠ Comparative Baseline (N-gram/VSM); i.e., proposed approach results are statistically significant.*

The test results are reported in Table 4 and 5. In Table 4, the statistical results based on Proposed & N-gram approaches are reported. With *t Stat*=± 3.564252, *t Critical*=2.228139 at α= 0.05, it is observed that *t Stat*> *t Critical*. Thus with 95% confidence, null hypothesis is rejected here.

Table 5 reports the two tailed t-test statistics of proposed & N-gram approaches. Using the tests, it is observed that here proposed results are statistically significant compared to VSM results with *t Stat*=± 4.511447, *t Critical*=2.228139 at α= 0.05. A 95% confidence interval about the mean difference is given 0.04647< *avgdiff* < 0.137166.It is noted that *t Stat*> *t Critical* and hence the null hypothesis is rejected with 95% confidence. This show that results of proposed approach is statistically significant compared to both baseline approaches.

6.4. Space complexity analysis

In N-gram model complexity, n= number of documents, C is the number of N-grams computed which is $|D| - N + 1$, where D is any generic document and N is the N-gram value. In VSM complexity let V represents vocabulary size.

For proposed approach the VSM complexity holds for phase I, while in phase II only k top documents has to be compared. For computing complexity of proposed approach, let us consider, $d1$ is the number of suspected documents and $d2$ is the number of source documents, $c1$ is the number of suspicious N-grams and $c2$ is the number of source N-grams computed. Further $v1$ is the suspicious vocabulary size and $v2$ is the source vocabulary size. Using the following values are computed:

$n = d1\, d2$, $V = v1v2$, $C = c1c2$

Next the *Reduced* n is computed. This is document reduction in phase II after candidate refinement stage. Next we compute the reduction obtained using the following steps:

$$\frac{n}{n_{red}} = \frac{d2}{k} \Rightarrow n = n_{red}\, \frac{d2}{k}$$

Finally the complexity of proposed approach is be formulated which is $\theta\left(nV + n_{red} C^2\right) \Rightarrow \theta\left(d1d2V + d1kC^2\right)$. This means phase II of proposed approach requires $\left(\frac{k}{d2}\right)$ times lesser number of comparisons when compared to comparative N-gram model. Compared to actual N-gram model the search space gets reduced substantially.

This is illustrated by using Translation Obfuscation (TO) set described in our paper. For computing and comparing the total comparisons required for the N-gram approach and proposed algorithm, PAN-14 training translation obfuscation (TO) set is considered. From Table1, it can be noted that the TO training set have 162 suspicious documents and 938 source documents. For the complexity evaluation of both the algorithms, an assumption is made that every document within the source and suspicious set has a fixed length, $|X_{susp}|$ and $|X_{src}|$ respectively. These lengths are taken as the average of all document lengths within the source and suspicious set respectively. Evaluating the TO set, it is found that the average suspicious document length, $|X_{susp}|$ = 3022 and average source document length $|X_{src}|$ = 845. Using these values, N-gram algorithm is evaluated taking N = 3 as discussed in Sect 6.1. With document length as $|X|$ and N as the N-gram value, the number of N-grams computed is given by $C = |X| - N + 1$. Further as N value decreases, C value increases and hence computation increases. As each suspicious document has to be compared with all the source documents, in the example considered with TO set, a total of 162*938 document comparisons will be required. Here with N = 3, $|X_{susp}|$ = 3022 and $|X_{src}|$ = 845, using C, 3020 suspicious N-grams and 843 source N-grams have to be formed and thus the total N-gram comparisons for each suspicious-source pair will be 3020*843. The total comparison required is the product of the total document comparisons and N-gram comparisons which is given as:

Total #.of comparisons of N-gram algorithm = (162*938)*(3020*843) =151,956*2,545,860 ≈3.87*1011

Thus it can be noted that the number of comparisons required for N-gram method is quite large. Here the worst case is considered, i.e., when the entire documents have a size equal to the average length. Now considering the complexity of the proposed two phase approach, the complexity of two phases has to be evaluated. As discussed in Sect 4.1, the proposed algorithm is implemented in two phases where phase I uses a VSM approach with simple document ranking with the clustering concept. The complexity of general VSM algorithm is measured based on the number of documents and the vocabulary size of each document. Hence to analyse the complexity of phase I, the vocabulary sizes of documents are required. Here the assumption is made that all documents in the source and suspicious set has a fixed vocabulary size, $|voc_{susp}|$ and $|voc_{src}|$ respectively. These sizes are computed as the average of vocabulary sizes of all documents in the source and suspicious set respectively. On evaluation with TO set, the average vocabulary sizes for source and suspicious set are found as $|voc_{susp}|$ = 812 and $|voc_{src}|$ = 207 respectively. Hence the total number of comparisons required in Phase I is given as:

Total #.of comparisons of Phase I = (162*938)*(*812**207) =151,956*168,084 ≈2.55*1010

In Phase II, filtering is carried out to improve the accuracy in terms of precision using the N-gram approach is used. But unlike the compared method, here the comparison of a suspicious document has to be done only with those source documents within the initial candidate sets formed by phase I of the proposed algorithm. This reduces the N-gram comparisons considerably and hence the overall complexity. As the K value selected for proposed evaluation is 25, after Phase I each suspicious document has to be compared with 25 source candidates. Hence 162 * 25 document comparisons are required here. The N-gram comparison part remains the same as 3020 * 843. Thus the total number of comparisons in Phase II is given as:

Total #.of comparisons of Phase II = (162*25)*(3020*843) = 4050*2,545,860 ≈1.03*1010

Now the total comparisons in proposed two phase-CR algorithm are computed as:

Total # of comparisons of Proposed Two phase-CR algorithm = Total # of comparisons in Phase I + Total # of comparisons in Phase II ≈ (2.55*1010) + (1.03*1010) ≈ 3.58*1010

Thus it is found that the proposed algorithm achieves ≈90.77% reduction in terms of relative % difference thus reducing the search space substantially. Relative % difference is computed using Eq. (14).

$$Relative\,\%\,Difference = \left(\frac{Proposed\ Result\,Value\text{ - }Baseline\,Result\,Value}{Baseline\,Result\,Value}\right)*100 \tag{14}$$

This reduction is also found with respect to other sets, viz, NO and RO. The general representation of complexities for baselines and the proposed approach is given in Table 6.

Table 6. Complexity analysis in terms of search space.

Methods	Complexity
N-gram	$\theta\left(nC^2\right)$
VSM	$\theta\left(nV\right)$
Proposed Approach	$\theta\left(nV + n_{red}C^2\right)$; $n_{red} << n$

Thus the proposed approach exhibits an overall complexity which is quite less than the compared N-gram method and a good performance improvement with respect to VSM. Hence the proposed algorithm integrates the relevant aspects of the compared approaches to avail an effective document level plagiarism detection in terms of both accuracy and performance efficiency. In educational domains effective detection of plagiarism is highly needed. Further, as a futuristic thought, it is important that, for maintaining the integrity of the country’s academic sector, proper identification of these unethical intellectual acts must be done. This will in turn restrain individuals from doing plagiarism and enable to increase the quality and individuality of the work**.**

7. Conclusions and Future Work

The paper proposes a two-phase candidate retrieval approach for identifying plagiarism at document level. It utilizes the information retrieval concepts to reduce the search space and complexity of plagiarism detection task. It attempts to explore the potential of combining the VSM and N-gram concepts in candidate retrieval task of extrinsic plagiarism detection. The VSM model with document ranking approach is adopted in Phase-1, which is followed by a filtering stage with N-gram approach for pruning out more false detections. From the analysis and discussions, it is quite evident that the proposed two phase approach surpasses the comparative baselines, viz., basic N-gram and VSM approaches. Compared to VSM approach, the performance improvement of the proposed approach is quite obvious from the result analysis. Compared to N-gram approach also, proposed approach exhibits an improvement especially when the manipulation complexity increases. On test sets, an average recall gain of 6.33% in set RO and 4.03% in Set TO is obtained. An average precision gain of 5.66% in set NO, 2.87% in set RO, and 9.98% in TO, is achieved. In terms of F1-score, the proposed algorithm exhibits an improved performance compared to both the baselines with data sets of different obfuscation levels. Further, complexity of the two phase approach in terms of ‘number of document comparisons’ is found to be considerably less compared to the baseline, viz., N-gram approach. About 90-92% relative reduction is obtained, which drastically reduces the computational complexity. Thus, considering the overall improvement, it can be found that the proposed two phase candidate retrieval algorithm reflects the potency of blending the potentials of individual N-gram and VSM approaches for availing an effective candidate retrieval module in extrinsic text plagiarism detections.

The parameter tuning here is done specifically for PAN sets. For real applications, there can be variations, specifically in *k* value used in IR ranking approach. In future, the algorithm can be evaluated on real world data sets to assess the performance stability and can be incorporated for availing effective online plagiarism detection systems.

Acknowledgements

The authors gratefully acknowledge Department of Science and Technology, Govt. of India (www.dst.gov.in) for sponsoring this research project, Sanction No.SB/FTP/ETA-0212/2013.

References

1. Maurer, H.; Kappe, F.; and Zaka, B. (2006). Plagiarism - A survey. *Journal of Universal Computer Science*, 12 (8), 1050-1084.
2. Liu, Gi-Zen.; Lo, Hsiang-Yee.; and Wang, Hei-Chia (2013).Design and usability testing of a learning and plagiarism avoidance tutorial system for paraphrasing and citing in English: A case study. *Journal of Computers & Education*, 69, 1-14.
3. Barrón-Cedeño, A (2012). On the mono- and cross-language detection of text re-use and plagiarism. *Ph.D. dissertation*, Universitat Politènica de València, Spain.
4. Dey, S.K.; and Sobhan, M.A. (2006). Impact of unethical practices of plagiarism on learning, teaching and research in higher education: Some

combating strategies. *In Proceedings of* the *7th International Conference on Information Technology Based Higher Education and Training ITHET '06*. 388-393.

5. 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.
6. Vani,K.; and Gupta, D. (2016). Study on extrinsic text plagiarism detection techniques and tools. *Journal of Engineering Science and Technology Review*, 9(4), 150-164.
7. Potthast,M.; Stein, B.; Barrón-Cedeño, A.; and Rosso, P. (2010). An evaluation framework for plagiarism detection. *In Proceedings of 23rd International Conference on Computational Linguistics*, COLING 2010. Beijing, China.
8. Manning,C.D.; Prabhakar, R.; and Schiitze, H. (2008) Introduction to information retrieval, Cambridge University Press, ISBN: 0521865719.
9. Chong, M. (2013). A study on plagiarism detection and plagiarism direction identification using natural language processing techniques. University of Wolverhampton, *Ph.D. dissertation.* U.K.
10. Clough, P. (2000). Plagiarism in natural and programming languages: An overview of current tools and technologies. *Research Memoranda: CS-00-05.*Department of Computer Science, University of Sheffield, U.K.
11. Barrón-Cedeño, A.; Basile, C.; Esposti, M.D.; and Rosso, P. (2010).Word length n-Grams for text re-use detection*. In CICLing 2010, LNCS6008*, 687-699.
12. Shrestha, P.; and Soloria, T. (2014). Machine translation evaluation metric for text alignment- lab report for PAN at CLEF 2014. *In Proceedings of 6th International Workshop PAN-14,* Sheffield, UK.
13. Shrestha, P.; and Soloria, T. (2013). Using a variety of n-grams for the detection of different kinds of plagiarism -lab report for PAN at CLEF 2013. *In Proceedings of 5th International Workshop PAN-13,* Valencia, Spain.
14. Potthast, M.; Barrón-Cedeño, A.; Stein, B.; and Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources & Evaluation Journal*, 45(1), 45-62.
15. Barrón-Cedeño, A.; Rosso, P.; and Benedí (2009).Reducing the plagiarism detection search space on the basis of the Kullback-Leibler distance. *In Gelbukh A. (ed.) CICLing 2009, LNCS 5449*, Springer-Verlag (2009), 523-534.
16. Alzahrani,S.M.; and Salim, N. (2010). Fuzzy semantic-based string similarity for extrinsic plagiarism detection- lab report for PAN at CLEF 2010. *In Proceedings of 4th International Workshop PAN-10*. Padua, Italy.
17. Stamatatos, E. (2011). Plagiarism detection using stop word n-grams, *Journal of the American Society for Information Science & Technology*, 62(12), 2512-2527.
18. Gupta, D.; Vani, K.; and Leema, L.M. (2016). Plagiarism detection in text documents using sentence bounded stop word n-grams. *Journal of Engineering Science and Technology*, 11(10), 1403-1420.

19. Korpal, R.; and Bose, S. (2016). A framework for detecting external plagiarism from monolingual documents: use of shallow NLP and N-gram frequency comparison. *In Proceedings of the 2nd International Conference on Information and Communication Technology for Competitive Strategy*. Pune, India.
20. Zechner,M.; Muhr, M.; Kern, R.; and Granitzer, M. (2009). External & intrinsic plagiarism detection using vector space models. *In Proceedings of SEPLN*, Spain, 47-55.
21. Ekbal, A.; Saha, S.; and Choudhary, G. (2012) Plagiarism detection in text using vector space model. *In Proceedings of 12th International Conference on Hybrid Intelligent Systems (HIS)*. Pune, 366-371.
22. Vani, K.; and Gupta, D. (2014). Using K-means cluster based techniques in external plagiarism detection. *In Proceedings of International Conference on Contemporary Computing & Informatics (IC3I).* Mysore, India, 27-29.
23. Ravi, N.R.; Vani, K.; and Gupta, D. (2016). Exploration of fuzzy C means clustering algorithm in external plagiarism detection system. International *Symposium on Intelligent Systems Technologies and Applications (ISTA)*, 384,127-138.
24. Thompson, U.T.; and Panchev, C. (2015). A hybrid Algorithm for Identifying & Categorizing Plagiarised Text Documents. In Proceedings of the World Congress on Engineering, WCE 2015, July 1-3, London, U.K., 297-302.
25. Kong, L.; Haoliang, Q.; Shuai, W; Cuixia, D.; Suhong, W.; and Yong, H. (2012). Approaches for Candidate Document Retrieval and Detailed Comparison of Plagiarism Detection- Notebook for PAN at CLEF 2012. *In Proceedings of the 4th International Workshop PAN-12,* Rome, Italy.
26. Kong, L.; Haoliang, Q.; Shuai, W; Cuixia, D.; Suhong, W; and Yong, H. (2014). Source retrieval based on learning to rank & text alignment based on plagiarism type recognition for plagiarism detection- Lab report for PAN at CLEF 2014. *In Proceedings of 6th International Workshop PAN-14*, Sheffield, UK.
27. Prakash, A.; and Saha, S.K. (2014). Experiments on document chunking & query formation for plagiarism source retrieval- lab report for PAN at CLEF 2014. *In Proceedings of 6th International Workshop PAN-14*, Sheffield, UK.
28. Suchomel, S.; Kasprzak, J.; and Brandejs, M. (2013). Diverse queries and feature type selection for plagiarism discovery- lab report for PAN at CLEF 2013. *In Proceedings of 5th International Workshop PAN-13*, Valencia, Spain.
29. Haggag, O.; and El-Beltagy, S. (2013). Plagiarism candidate retrieval using selective query formulation and discriminative query scoring -notebook for PAN at CLEF 2013. *In Proceedings of 5th International Workshop PAN-13*, Valencia, Spain.
30. Elizalde,V. (2014). Using noun phrases and tf-idf for plagiarized document retrieval- notebook for PAN at CLEF 2014. *In Proceedings of 6th International Workshop PAN-14*, Sheffield, UK.
31. Ravi N, R.; and Gupta, D. (2015). Efficient paragraph based chunking and download filtering for plagiarism source retrieval-notebook for PAN at CLEF 2015. *In Proceedings of 7th International Workshop PAN-15*, Toulouse, France.
32. Ravi N, R.; and Gupta, D. (2016). A plagiarized source retrieval system developed using efficient download filtering and pos tagged query formulation

with effective paragraph based chunking. International Journal of Artificial Intelligence, 14(1).

33. Rafiei,J.; Mohtaj, S.; Zarrabi, V.; and Asghari, H. (2015). Source retrieval plagiarism detection based on noun phrase and keyword phrase extraction-notebook for PAN at CLEF 2015. *In Proceedings of 7th International Workshop PAN-15*, Toulouse, France.
34. Ehsan, N.; and Shakery, A. Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information. *Journal of Information Processing and Management*, 52(6), 1004-1017.
35. Potthast, M.; Hagen, M., Gollub, T.; Tippmann, M.; Kiesel, J.; Rosso, P.; Stamatatos, E.; and Stein, B. (2013). Overview of 5th International Competition on plagiarism detection. *In Proceedings of CLEF 2013 Evaluation Labs & Workshop- Working Notes Papers*, September, Valencia, Spain.
36. Potthast, M.; Hagen, M.; Beyer, A.; Busse, M.; Tippmann, M.; Rosso, P.; and Stein, B. (2014). Overview of 6th International Competition on Plagiarism Detection .*In Proceedings of CLEF Evaluation Labs & Workshop- Working Notes Papers*, September, Sheffield, UK.