Using Natural Language Processing Techniques and Fuzzy-Semantic Similarity for Automatic External Plagiarism Detection

Deepa Gupta
Dept. of Mathematics
ASE, Bangalore
Amrita Vishwa Vidyapeetham
Bangalore, India
deepagupta.verma@gmail.com

Vani K
Dept. of Computer Science
ASE, Bangalore
Amrita Vishwa Vidyapeetham
Bangalore, India
vani5019@gmail.com

Charan Kamal Singh
Dept. of Computer Science
ASE, Bangalore
Amrita Vishwa Vidyapeetham
Bangalore, India
er.kamal26@gmail.com

Abstract— **Plagiarism is one of the most serious crimes in academia and research fields. In this modern era, where access to information has become much easier, the act of plagiarism is rapidly increasing. This paper aligns on external plagiarism detection method, where the source collection of documents is available against which the suspicious documents are compared. Primary focus is to detect intelligent plagiarism cases where semantics and linguistic variations play an important role. The paper explores the different preprocessing methods based on Natural Language Processing (NLP) techniques. It further explores fuzzy-semantic similarity measures for document comparisons. The system is finally evaluated using PAN 2012[1] data set and performances of different methods are compared.**

Keywords— External Plagiarism; Intelligent Plagiarism; Fuzzy-Semantic Similarity; POS Tagging; NLP techniques

1. INTRODUCTION

To plagiarize is to steal and pass off the ideas or words of another as one's own: use a created production without crediting the source [1]. Plagiarism is defined as appropriating someone else's words or ideas without acknowledgment. It is defined in dictionaries as the "wrongful appropriation," "close imitation," or "purloining and publication of another author's language, thoughts, ideas, or expressions and the representation of them as one's own original work" [2].

There are different types of text plagiarism, which are broadly categorized as literal and intelligent plagiarism. In the former, the plagiarists do not spend much time in hiding the academic crime they committed. For example, they simply copy and paste the text from the Internet [3]. In intelligent plagiarism, the plagiarists try to betray readers by changing the contributions of others in an intelligent way and make it appear as their own work. Intelligent plagiarists try to hide, obfuscate, and change the original work in various intelligent ways, including text manipulation, translation, and idea adoption. Text manipulation is related to the switching of the words in the original text and making a fake one. In idea adoption the whole idea from a text is copied. Thus intelligent plagiarism cases are more complex to detect and hence challenging to work with [3] [4].

In plagiarism detection there are mainly two methods, extrinsic/ external plagiarism detection and intrinsic/ internal plagiarism detection. Extrinsic compares a suspicious document with a reference collection, which is a set of documents assumed to be genuine [3]. Intrinsic 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 [3].

The work proposed in this paper deals with detection of intelligent plagiarism cases using an extrinsic plagiarism detection system. The system uses different pre-processing methods based on NLP techniques and considers fuzzy-semantic similarity measures for document comparisons.

2. RELATED WORKS

All the approaches available in external plagiarism detection mainly concentrate on improving efficiency of detection system. Most of the approaches present today follow a common methodology, only a few deviates. The general approach includes the following steps: pre-processing, candidate document selection, document comparisons, passage boundary detection and evaluation [3] [4].

Efstathios Stamatatos [5] proposes a N-gram based method where stop words are retained using a small 'list of stop words'. Using this stop word N-grams are formed for document comparisons based on the assumption that this helps to capture the syntactic similarities. Results are evaluated using a part of PAN 2010 corpus. Here all the content words

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

are removed and hence cannot be used for semantic based comparisons which will be more promising. Jiannan Wang, Guoliang Li & Jianhua Feng [6] presents an efficient method for Fuzzy-Token matching based on string similarity join. Fuzzy-Token similarity is a combination of token-based similarity and character based similarity. Different similarity measures like JACCARD, DICE, and COSINE are used in this method. YuriiPalkovskii, Alexei Belov, IrynaMuzyka [7], presents the use of semantic similarity measurement in external plagiarism detection. It uses WordNet[2] for extracting semantics, Porter stemmer for stemming, Brill Tagger for POS tagging and a Word sense disambiguation (WSD) process. This followed capturing the semantic similarity between sentences as a problem of computing a maximum total matching weight of a bipartite graph. Finally the system is evaluated using PAN-11 data set.

The approach in [8] uses semantic role labeling (SRL) which transforms the sentences into arguments. Extracted arguments are grouped into nodes based on argument type. For each argument group, all the concepts for each term are then extracted using WordNet. These are collected in one node known as topic signature node. For the comparison, the topic signature nodes of source and suspicious documents are used. The evaluation is done using PAN-09 corpus and the results are promising; but it is computationally expensive. In [9] pre-processing is performed for both texts followed by semantic labeling through Frame Net. Each training sentence is parsed into a syntactic tree. Features are marked with and then passed through the statistical classifiers. It compares the usefulness of different features and feature combinations in SRL task. It also explores the integration of SRL with syntactic parsing technique. Vector Space models (VSM) can be used in candidate retrieval stage, which retrieves a subset of source documents which are the possible sources of plagiarism. This method is followed in [10] where initially preprocessing is carried out. Then candidate document selection using VSM is done and in the third phase a graph based approach to find the plagiarized passages is adopted. The same approach with fuzzy-semantic similarity measurement in the third phase is used in [11].

Thus it is seen that different methods of external plagiarism detection were explored in the past by eminent research fellows. It varied from the most basic detection using word by word comparisons to the detections using different NLP techniques [12]. In most of the systems N-gram comparisons are used instead of sentence based approaches. Deep linguistic approaches are used for efficient semantic based detections as seen in [8].The problem with such approaches is that they are computationally expensive. In order to develop an effective system for intelligent plagiarism detection the trade-off between performance and processing speed has to be considered. This factor is the main limitation of many of the above systems.

A survey of different plagiarism detection methods are discussed in [3]. According to the findings, intelligent plagiarism detection is less explored due to its complex nature. From their studies it is also concluded that for the complex plagiarism cases, semantic-fuzzy based approaches are best suitable.

Such an approach is found in [13] which use fuzzy-semantic similarity measures for document comparison. Here in pre-processing stage, the text document is initially tokenized and stop-words are removed and then stemmed. This is followed by sentence based comparison of source and suspicious documents. Stop-Words are the most frequently used words having no semantic meaning. A list of the 50 most frequent words of the English language provided by the British National Corpus[3] which includes about 90 million tokens are usually used. Then the suitable matches are detected using similarity measures. Fuzzy similarity scores are adopted and semantic similarity of words is extracted using WordNet. Then post-processing is performed in which the boundary of detected plagiarized fragments is formed and finally evaluation is done using PAN-10 data set. To obtain the sentence based similarity, for each word the sentence correlation with other sentence is calculated using (1).

$$\mu_{q,x} = 1 - \prod w_k \in S_x(1 - F_{q,k}) \qquad (1)$$

$$F_{q,k}(x) = \begin{cases} 1.0 \text{ if } s = 1.0 \\ 0.5 \text{ if } s \in (0.0, 1.0) \\ 0.0 \text{ if } s = 0.0 \end{cases} \qquad (2)$$

$$Sim(S_q, S_x) = (\mu_{1,x} + \mu_{2,x} + \ldots. + \mu_{q,x} + \ldots. + \mu_{n,x}) / n \qquad (3)$$

Here S_x is the suspicious sentence and S_q is the source sentence. $\mu_{q,x}$ in (1) represents the word-to-sentence correlation factor for each word w_q in s_q and the sentence s_x. w_k is the word in S_x and $F_{q,k}(x)$ in (2) defines the fuzzy-semantic similarity between w_q and w_k. 'x' is a tuple (w_1, w_2), where w_1 and w_2 are the two words of different sentence and s represents the output of semantic similarity. The degree of similarity between S_q and S_x is calculated using (3) where n is the total number of words in S_q. Only those sentences whose similarity score is above a threshold value, $\alpha > 0.65$ is considered. Further deatils can be found in [13].

The above discussed approach in [13] is considered as the basic approach in proposed system, since the focus is on detection of intelligent plagiarism.The following shortcomings have been traced in the method discussed in [13] .

- It uses sentence based comparisons instead of N-gram comparison, which is the commonly used approach.

[2] http://wordnet.princeton.edu/

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

- The method is computationally expensive since comparison is done between all words in suspicious text against all words in source text after stop word removal.
- The fuzzy similarity scores are not so appropriate. The method assigns value 0.5 even if value of semantic similarity is 0.9 which may result in pruning of many matched fragments. Similarly if value of is 0.1 then $F_{q,k}(x)$ will assign 0.5 value, which means many unmatched fragment will be included.
- It uses stemming which is not so appropriate for semantic comparisons.

Considering the future enhancements mentioned in [13] and the above discussed limitations the proposed system is developed. NLP techniques are also incorporated as they can make the detection task efficient. The following section gives the detailed discussion of the methods used in the proposed system.

3. PROPOSED SYSTEM

The proposed system performs the following steps which are represented as the modules of the general system in Figure 1:

1. Pre-processing and NLP techniques
2. N-gram comparisons and Semantic-Similarity calculation
3. Passage Boundary Detection
4. Evaluation

As the main focus is to improve the document comparisons for an efficient detection, candidate retrieval is not considered as a separate stage in the system. The source documents are selected such that they are already candidate documents (i.e. they are already sources of plagiarism). Each step used in the proposed system is discussed in detail in the following sub sections.

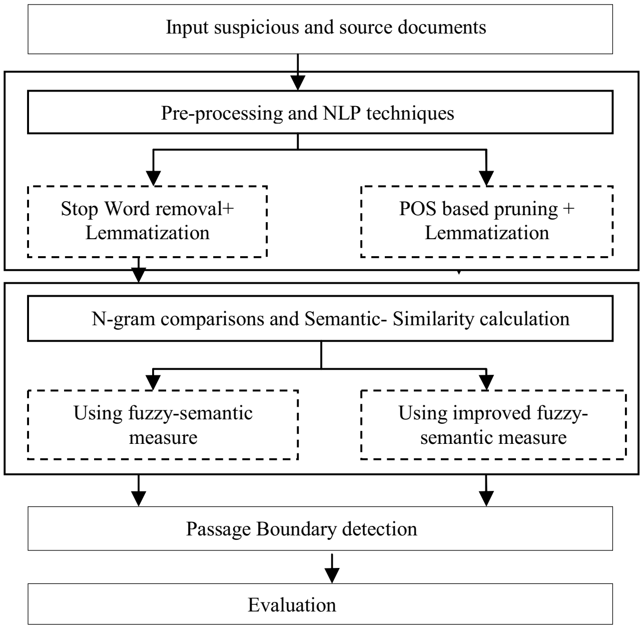
Input suspicious and source documents
Pre-processing and NLP techniques
Stop Word removal+ Lemmatization
POS based pruning + Lemmatization
N-gram comparisons and Semantic- Similarity calculation
Using fuzzy-semantic measure
Using improved fuzzy-semantic measure
Passage Boundary detection
Evaluation

Fig 1. Schematic diagram of the Proposed System

3.1. *Pre-*Processing *and NLP techniques*

In this stage, initially the basic pre-processing steps like tokenization, spelling corrections etc are carried out. In the base system [13] stop word removal and stemming are mainly used. As the approach is semantic based, matching stop words is meaningless and hence they are removed.

In the proposed system POS tagging with Stanford Tagger[4] is used for this pruning stage. In POS tagging each word in a sentence is tagged with their corresponding part of speech. In intelligent plagiarism, the plagiarist usually replaces the content words with its synonyms. The word replacement will change the words but this will not change the word class. This idea is explored in the current system. Here stop words are not removed. Instead after tagging, the words which belong to noun, verb, adjective and adverb class are retained and others are pruned. This is because other words like articles, conjunctions, prepositions etc will not contribute to the semantics of the sentence. Auxiliary verbs are also not considered since comparison of all these words are meaningless and increases computation time.

After this procedure, lemmatization of retained words is done instead of stemming which is used in base system. Lemmatization produces dictionary base forms which are appropriate for semantic comparisons. Stop word removal approach followed by lemmatization is also done and is compared with the POS approach. Consider the following English sentence given below:

E: They refuse to permit us to get the grant.

Tagged E: [('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('permit', ' VB'), ('us', 'PRP'), ('to', 'TO'), ('obtain', 'VB'), ('the','DT'), ('grant', 'NN')]

After SW removal: ['They', 'refuse', 'permit', 'us', 'obtain', 'grant']

After POS based pruning: [('refuse', 'VBP'), ('permit', ' VB'), ('obtain', 'VB'), ('grant', 'NN')]

It is clear that after POS based pruning more number of semantically irrelevant words are pruned out.

3.2. *N-gram Comparisons and Semantic-Similarity Calculation*

The fuzzy-semantic similarity measure in the base system is improved due to its limitation mentioned in Section 2. In improved fuzzy-semantic similarity measure, different ranges of similarity scores are used. This reduces the pruning of matched fragments and inclusion of unmatched one's, which

[4] http://nlp.stanford.edu/software/tagger

was the main problem of the other measure. The measure is as follows:

$$F_{q,k}(x) = \begin{cases} 1.0 \text{ if } s = 1.0 \\ 0.7 \text{if } s \in [0.7,1.0) \\ 0.5 \text{if } s \in [0.5,0.7) \\ 0.3 \text{ if } s \in [0.3,0.5) \\ 0.2 \text{ if } s \in (0.0,0.3) \\ 0.0 \text{ if } s = 0.0 \end{cases} \quad (4)$$

$F_{q,k}(x)$ is the updated Fuzzy-Semantic similarity between w_q and w_k . Other terminologies and equations (1) and (3) are similar to the basic fuzzy- semantic similarity measure discussed in Section 2.

In the method discussed in [13] sentence based comparisons are used, while in the current method comparisons are done by formation of N-grams. In this 'N' consecutive words of suspicious and source document is compared. Here N=3 ie, trigrams are considered. N-gram comparisons are done for both combinations, ie, stop words with lemmatization and POS with lemmatization. For each of these methods, similarity is calculated using basic fuzzy-semantic similarity measure discussed in Section 2 and the improved similarity measure as mentioned in (4). The results are then evaluated and compared, which is given in detail in Section 4. In POS based method, after pruning the semantically relevant tagged words are obtained. Then comparisons are done between words of same classes only. It is obvious that comparison of a noun with noun is meaningful rather than comparing it with verb and adjective. This helps in proper comparisons and is computationally efficient.

3.3. Passage Boundary Detection

After comparison of suspicious text with a source text, all the matched N-grams are stored. Passages are formed in both suspicious and source document based on a passage boundary condition which acts as a threshold value. This value depends on total word length of text. Let $\phi = G(w_1, w_2)$ in (5) represents the passage boundary threshold and w_1 denotes the total number of words in suspicious text and w_2 the total number in words in source text.

$$G(w1, w2) = \begin{cases} 50 \text{ if } w1 \leq 300 \text{ or } w2 \leq 300 \\ 100 \text{ if} w1 \in (300{,}10{,}000] \text{or} \\ w2 \in (300{,}10{,}000] \\ 150 \text{ if } w1 > 10{,}000 \text{ or } w2 > 10{,}000 \end{cases} \quad (5)$$

Here $G(w_1, w_2)$ represents a function which intakes the value of total length of source and suspicious text. Based on the function mentioned in (5) it gives the output i.e. the passage boundary threshold value.

Let F represents an array of matched fragments and $f_1, f_2, \ldots, f_n$ represents the matched fragment's position in the given source and suspicious text, i.e.

$F = [f1, f2, f3, \ldots. fm-1, fm \ldots\ldots fn]$

Two consecutive fragment values are checked, i.e. f_m and f_{m-1} .

```
if $fm - fm - 1 = \phi m > \phi$, then
        Split the passage at  $m^{th}$  position , i.e.
        $F = [[f1, f2, f3, \ldots. fm - 1] \; [fm \ldots\ldots\ldots fn]]$ .
else
        Merge the passages.
```

The above discussed procedure is used for passage boundary detection in the proposed system.

3.4. Evaluation

Algorithm is evaluated on four different standard measures, Recall (Rec), Precision (Prec), Granularity (gran) and Plagdet_score which are used in PAN competition [14]. Recall is the ratio of matched characters of source and suspicious to the total length of expected plagiarized characters in source and suspicious document.

$$rec\,(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\cup_{r \in R} (s \cap r)|}{|s|} \quad (6)$$

S denotes the set of plagiarism cases in the corpus, and R denote 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.

$$prec\,(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\cup_{s \in S} (s \cap r)|}{|r|} \quad (7)$$

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

$$gran\,(S, R) = \frac{1}{|SR|} \sum_{s \in SR} |Rs| \quad (8)$$

Plagdet_Score is a formula which combines recall, precision, and granularity to allow for ranking. The range of Plagdet_score is between 0 and 1.

4. RESULTS AND DISCUSSIONS

The input is a set of source document and its corresponding suspicious document. The evaluation is done by testing the algorithm in partial dataset that is taken from PAN-2012 corpus, which is the corpus used for PAN competition 2012 for textual plagiarism detection. The suspicious documents contain passages 'plagiarized' from the source documents obfuscated by some techniques. In PAN 2012 corpus the data set is divided into five categories according to the type of plagiarism they exhibit as follows:

- SET-1 represents Highly Plagiarized Passage
- SET-2 represents Low Plagiarized Passage
- SET-3 represents Simulated Paraphrase passage
- SET-4 represents No obfuscation passage
- SET-5 represents No Plagiarized Passage

Highly plagiarized passages are those document pairs where the plagiarized passages are obfuscated by the means of large amount of word shuffling. Low plagiarized passages are document pairs where the plagiarized passages are obfuscated by moderate word shuffling. Simulated paraphrase passages are document pairs where the plagiarized passages are obfuscated by humans. No obfuscation is the document pairs where the suspicious document contains exact copies of passages in the source document. No plagiarized passages mean document pairs without any plagiarism. Table 1 provides the data statistics of these sets.

Using the data statistics given in Table 1, results are evaluated and compared. The different pre-processing methods based on NLP techniques discussed in Section 3 are compared using fuzzy-semantic and improved fuzzy-semantic similarity measures. Hence it provides four possible combinations as given below. They are evaluated and efficiency is compared based on the measures given in Section 3.4.

- SWPFS: Stop-Word with Fuzzy-Semantic Similarity Measures
- SWPIFS: Stop-Word with Improved Fuzzy-Semantic Similarity Measures
- POSPFS:POS with Fuzzy-Semantic Similarity Measures
- POSPIFS:POS with Improved Fuzzy-Semantic Similarity Measures

In Table 2 the results obtained using each of the listed methods are shown. The graphical plots of performance analysis of these methods using the standard measures mentioned in Section 3.4 are shown in Fig. 2, Fig.3 and Fig. 4. Set 5 is not considered in the plot, since it is the data with no plagiarism cases and the result is always 100% for the proposed combinations of method using all the measures.

Table 1: Data Statistics

	Set 1	Set 2	Set 3	Set 4	Set 5
Suspicious # words	4,05,44	1,50,71	1 ,20,959	1,60,813	591
Source # words	4,92,113	1,99,635	1,89,703	3,22, 400	2,903
# expected plagiarized passages	11	9	4	4	0

Table 2: Result Statistics

Measures	SWPFS	SWPIFS	POSPFS	POSPIFS
Recall	80.2	81.6	83.8	85.2
Precision	82.6	84	84.2	85.8
Granularity	1	1	1	1
Plagdet_Score	0.818	0.832	0.84	0.854

From Table 2, it is seen that granularity is obtained as one for all the methods which is a significant improvement over the base system [13], where false detections are found.

Here SWPFS is considered as the reference method with which the other three methods are compared. This reference method is an improved form of the base method [13], as here N-gram comparisons and lemmatization are used. Analyzing the graphs given in Fig.1, Fig.2 and Fig.3 by comparing the SWPIFS with the reference method, it is seen that the improved fuzzy-semantic similarity measure gives promising results. In order to compare the two NLP techniques used via stop word removal and POS tagging, the results obtained using POSPFS is compared with SWPFS. From the analysis of the graphs and using Table 2, it is observed that POS based method produces an improvement over stop word method. There is an absolute three point improvement here. This will be more visible when large data sets are used. The refinement is mainly because of the use of N-gram comparisons with POS tags, i.e. the comparisons of words of same classes are done as discussed in Section 3.2. Further the analysis and comparison of all the proposed method combinations have projected out that the one which uses POS based technique with improved fuzzy-semantic similarity measure i.e. POSPIFS outperforms the other methods. It also reduces the number of comparisons and hence the computation time required.

5. CONCLUSION AND FUTURE SCOPE

Intelligent plagiarism using fuzzy-semantic based external plagiarism detection is explored in this paper. It uses the different pre-processing methods based on NLP techniques.

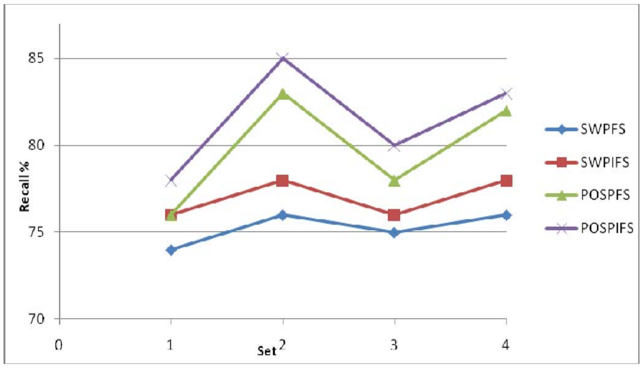

Fig 2. Performance analysis using Recall

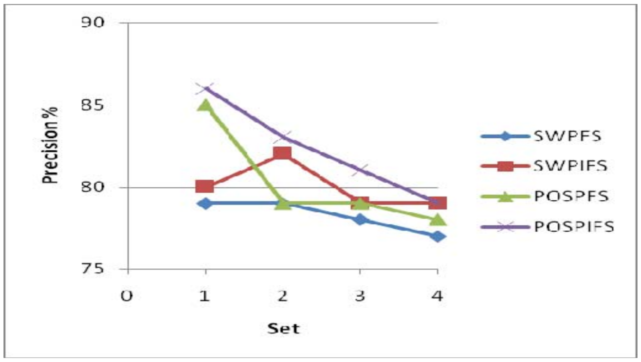

Fig 3. Performance analysis using Precision

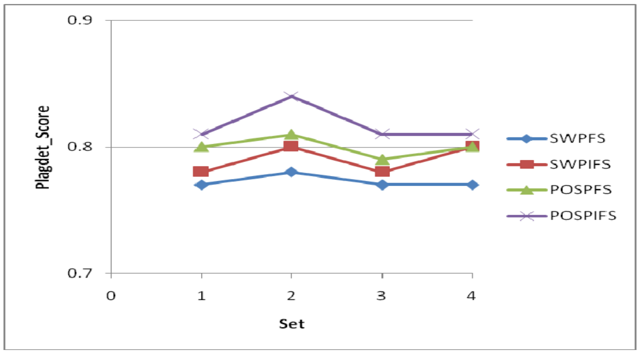

Fig 4. Performance analysis using Plagdet_Score

Here mainly lemmatization, stop word removal and POS tagging are explored. The paper provides an insight on how N-gram comparisons using POS tags can be done. It also throws light on how similarity calculation can be improved using fuzzy-semantic similarity measures. It introduces an improved fuzzy-semantic measure that can provide a significant improvement in the efficiency and accuracy of the system compared to the base method. The experimental results in terms of the PAN measures (discussed in Section 3.4) show that compared to the other methods, POSPIFS performs efficiently. According to the analysis and discussions done in Section 4, it can be observed that POS method integrated with improved fuzzy-semantic similarity measure surpass the other methods in terms of accuracy and efficiency.

In future, more efficient NLP techniques can be used to improve the performance of detection system. Results can be improved by using efficient passage boundary detection conditions. Evaluation can be performed using large data sets for proper analysis and comparisons. Advanced soft computing methods and optimization techniques can be used for enhancing the system performance.

6. REFERENCES

[1] Webster's New Collegiate Dictionary 9th ed, Springfield, Ma: Merriam 1981, pp. 870.

[2] Robert S. Nelson, Random House Compact Unabridged Dictionary: qtd. in Stepchyshyn, Vera; Library plagiarism policies.Assoc of College &Research Libraries, pp. 65.ISBN 0-8389-8416-9, 2007.

[3] Salha M. Alzahrani, NaomieSalim, and Ajith Abraham, " Understanding plagiarism linguistic patterns, textual features, and detection methods", IEEE transactions on systems, man, and cybernetics part c: application and reviews,42(2), march 2012.

[4] Ahmed Hamza Osman, Naomie Salim and Albaraa Abuobieda, "Survey of text plagiarism detection", Journal of Computer Engineering and Applications ,1(1), June 2012.

[5] Efstathios Stamatatos, "Plagiarism detection using stopword n-grams" Journal of the American Society for Information Science and Technology, 62(12), pp. 2512-2527, Wiley, 2011.

[6] Jiannan Wang, Guoliang Li, JianhuaFeng "Fast-Join: an efficient method for fuzzy token matching based string similarity join", Data Engineering (ICDE), 2011 IEEE 27th International Conference, March 2011.

[7] YuriiPalkovskii, Alexei Belov, IrynaMuzyka,"Using WordNet based semantic similarity measurement in external plagiarism detection", Notebook Papers of CLEF , 2011.

[8] Ahmed Hamza Osmana NaomieSalima, Mohammed Salem Binwahlanc,RihabAlteebd and AlbaraaAbuobieda, "An improved plagiarism detection scheme based on semantic role labelling", Applied Soft Computing 12 (2012) 1493–1502.

[9] Daniel Gildea, Daniel Jurafsky, "Automatic labelling of semantic roles", computational linguistics, 28(3), 2011.

[10] Asif Ekbal, Sriparna Saha and Gaurav Choudhary, " Plagiarism detection in text using Vector Space Model, In Proc. of 12th International Conference on Hybrid Intelligent Systems(HIS), pp.366-371, Pune, 2012.

[11] Rasia Naseem and Sheena Kurian, " Extrinsic plagiarism detection in text combining VSM and fuzzy semantic similarity scheme", Journal of Advanced Computing, Engineering and application(IJACEA),2(6), December 2013.

[12] Miranda Chong, Lucia Specia and Ruslan Mitkov, " Using natural language processing for automatic plagiarism detection", 4th International Plagiarism Conference, Northrumbia University, 2010.

[13] SalhaAlzahrani, NaomieSalim, "Fuzzy semantic-based string similarity for extrinsic plagiarism detection- lab report for PAN at CLEF 2010, In Proc. of 4th International Workshop PAN-10, Padua, Italy, 2010.

[14] Martin Potthast, Benno Stein, Alberto Barron Cedeno and Paolo Rosso, "An evaluation framework for plagiarism detection", In Proc. of 23rd International Conference on Computational Linguistics, COLING 2010,Beijing, China, 2010.