A Survey on Plagiarism Detection Techniques for Indian Regional Languages

Harshall Lamba
Department of Computer Engineering,
Pillai College of Engineering,
Mumbai University, India

Sharvari Govilkar
Department of Computer Engineering,
Pillai College of Engineering,
Mumbai University, India

ABSTRACT
Plagiarism is an illicit act of using other's work wholly or partially as one's own in any field such as art, poetry literature, cinema, research and other creative forms of study. It has become a serious crime in academia and research fields and access to wide range of resources on the internet has made the situation even worse. Therefore, there is a need for automatic detection of plagiarism in text. This paper presents a survey of various plagiarism detection techniques used for different languages.

Keywords
Plagiarism Detection, Semantic Role Labeling, Latent Semantic Analysis, Fingerprint and Winnowing

1. INTRODUCTION
Plagiarism, which is the act of passing off somebody else's original words and ideas as one's own, is seen as a moral offence and often also as a legal offence. Plagiarism has an ancient root, as the word itself is derived from Latin words "plagiaries", which means abductor, and "plagiare", which means to steal [1].

Plagiarism may be broadly classified as literal and intelligent. In the former, the plagiarist simply copy and paste the text from the web. In intelligent plagiarism though they try to betray the users by using paraphrasing skills and making it look as their own.

This paper is organised into 6 sections. The section 1 presents the introduction, section 2 describes about plagiarism detection and its types. Related work is presented in section 3 which describes various plagiarism detection techniques applied on different languages. The techniques are discussed in section 4. Section 5 offers comparisons and observations of various plagiarism detection techniques. Conclusion is made in section 6.

2. PLAGIARISM DETECTION
There are mainly two methods of automatic plagiarism detection: Extrinsic/ External plagiarism detection and Intrinsic/ Internal plagiarism detection. Intrinsic plagiarism detection analyses the input document only to find some parts which are not written by the same author without performing comparisons to external corpus. External plagiarism detection needs a reference collection of documents which are assumed to be genuine. A suspicious document is compared to all the documents in this collection to find duplicates or near duplicates fragments in source documents [2].

The survey discussed in this paper deals with external plagiarism detection methods used in various languages.

3. RELATED WORK
Plagiarism detection approaches identify documents that are likely to be plagiarized from a source corpus. In this section, various external plagiarism detection techniques used for different languages have been cited.

Urvashi Garg and Vishal Goyal [3] present an automated plagiarism detection software tool Maulik, which divides the text into n-grams. Stop word removal and stemming has been used. Cosine similarity has been used for finding the similarity score. Findings: Similarity score of 96.3 has been achieved which is higher as compared to the existing Hindi plagiarism detection tools such as Plagiarism checker, Plagiarism finder, Plagiarisma, Dupli checker, Quetext.

Nilam Shenoy and M.A. Potey [4] present fuzzy semantic-based similarity search model and Naïve Bayes model for uncovering obfuscated plagiarism for English and Marathi language. The fuzzy model identification is based on 'If-then' fuzzy rules. Semantic relatedness between words is studied based on the part-of-speech (POS) tags and WordNet-based similarity measures. Naïve Bayes classifier is used to achieve better detection performance.

Vani K and Deepa Gupta [5] presented an investigation of different combined similarity metrics: Cosine similarity, Dice coefficient, Match coefficient and Fuzzy-Semantic measure with and without POS tag information for extrinsic plagiarism detection in English language. These systems are evaluated using PAN1 -2014 training and test data set.

Hui Ning, Cuixia Du, Leilei Kong, Haoliang Qi and Mingxing Wang [6] performed comparisons of keyphrase extraction methods like TF-IDF, weighted TF-IDF, TF-IDF based on passages and Weighted TF-IDF based on passages. All comparisons experiments are implemented by using vector space model. Experimental results show that TF-IDF based on passages is the best choice.

Ashraf S Hussein [7] proposed a content-based method for document similarity analysis devoted to Arabic language. The hidden associations between the unique n-gram phrases and their documents are investigated using Latent Semantic Analysis (LSA). Next, the pairwise document subset and similarity measures are derived from the Singular Value Decomposition (SVD) computations. The results of the proposed method were compared to that of Plagiarism-Checker-X, and the proposed method outperformed Plagiarism-Checker-X, especially for the intelligent similarity cases with syntactic changes.

Sindhu L and Sumam Mary Idicula [8] present a detection system based on fingerprinting for identifying copy in Malayalam text-based documents. In this paper, a procedure for plagiarism detection of Malayalam documents to identify

similarity between documents is presented. The winnowing algorithm is used to compute the fingerprints at sentence level. The method improves the search time with more accuracy in the detection process.

Vani K and Deepa Gupta [9] present comparison of different methods of document categorization in external plagiarism detection for English language. Their primary focus is to explore the unsupervised document categorization/clustering methods using different variations of K-means algorithm and compare it with the general N-gram based method and Vector Space Model based method.

Deepa Gupta, Vani K and Charan Kamal Singh [10] propose to detect intelligent plagiarism cases where semantics and linguistic variations play an important role. The paper explores the different pre-processing methods based on Natural Language Processing (NLP) techniques. It further explores fuzzy-semantic similarity measures for document comparisons.

MAC Jiffriya MAC Akmal Jahan and Roshan G. Ragel [11] propose an effective plagiarism detection tool on identifying suitable intra-corpal plagiarism detection for text based assignments by comparing unigram, bigram, trigram of vector space model with cosine similarity measure. In addition, the selected trigram vector space model with cosine similarity measure is compared with tri-gram sequence matching technique with Jaccard measure.

Peyman Mahdavi, Zahra Siadati and Farzin Yaghmaee [2] propose an external Persian plagiarism detection method based on the vector space model (VSM). To implement and examine this method, a Persian corpus has been developed. Several optimizations have been done during the study. These optimizations make the algorithm very fast and accurate. The test results of the proposed method shows an accuracy of 0.87 and a processing time cost of less than 10 minutes.

Sidik Soleman and Ayu Purwarianti [12] employed latent semantic analysis (LSA) as the term-document representation to handle the Indonesian intelligence plagiarism. The LSA was used in the Heuristic Retrieval (HR) component and Detailed Analysis (DA) component. Experimental results showed that the LSA outperformed the VSM (Vector Space Model), especially in test cases with intelligence plagiarism.

Sindhu L, Bindu Baby Thomas and Sumam Mary Idicula [13] present a plagiarism detection tool for plagiarism detection in Malayalam documents. The tool is based on a new comparison algorithm that uses some NLP techniques to compare suspect documents which may not be identified using existing methods for Malayalam document plagiarism detection.

Ahmed Hamza Osman and Naomie Salim [14] introduce an improved semantic text plagiarism detection technique based on Chi-squared Automatic Interaction Detection (CHAID). The proposed technique analyses and compares text based on semantic allocation for each term inside the sentence. It also captures the underlying semantic meaning in terms of the relationships between its concepts via Semantic Role Labelling (SRL).

Shuai Wang, Haoliang Qi, Leilei Kong and Cuixia Du [15] propose a hybrid similarity measure model on the basis of the fitting function of the optimal dividing line between plagiarism and non-plagiarism where they integrate VSM and Jaccard coefficient into a unified one.

Agung Toto Wibowo, Kadek W Sudarmadi and Ari M Barmawi [16] propose fingerprint and Winnowing algorithm for detecting plagiarism of scientific articles in Bahasa Indonesia. Plagiarism classification is determined from those two documents by a Dice Coefficient at a certain threshold value. The results showed that the best performance of fingerprint algorithm was 92.8% while Winnowing algorithm's best performance was 91.8%. Level-of-relevance to the topic analysis result showed that winnowing algorithm has got stronger term-correlation of 37.1% compared to the 33.6% fingerprint algorithm.

Sindhu L, Bindu Baby Thomas and Sumam Mary Idicula [17] propose a method of copy detection in short Malayalam text passages. An algorithm for plagiarism detection using the n-gram model for word retrieval is developed and found tri-grams as the best model for comparing the Malayalam text. The experiments show that trigram model gives the average acceptable performance with affordable cost in terms of complexity.

Ahmed Hamza Osman, Naomie Salim, Mohammed Salem Binwahlan, Ssennoga Twaha, Yogan Jaya Kumar and Albaraa Abuobieda [18] introduce a plagiarism detection technique based on the Semantic Role Labeling (SRL). The technique analyses and compares text based on the semantic allocation for each term inside the sentence. SRL is superior in generating arguments for each sentence semantically.

Asif Ekbal, Sriparna Saha and Gaurav Choudhary [19] propose a technique based on textual similarity for external plagiarism detection. The method proposed is based on POS classes, VSM and graph based approach.

4. PLAGIARISM DETECTION TECHNIQUES

Most of the approaches present today follow a common methodology, only a few deviate. The general approach includes the following steps: pre-processing, candidate document selection, document comparisons, passage boundary detection and evaluation [10].

4.1 Pre-Processing and NLP Techniques

In this phase the words of the text are transformed to their corresponding base form after eliminating the stop-words. To transform a text into a structured form for the plagiarism detection process, the following steps are performed:

4.1.1 Tokenization

Document is broken up into tokens or words where a token is a unit of a document that may be used [8].

4.1.2 Stop-word Removal

Stop-words are words that do not have any meaning on their own. They are used in languages to give a structure for a sentence. They can be removed without affecting the accuracy of similarity. It also reduces the number of false-positives [8].

4.1.3 Lemmatization

Words can have different forms, which are formed as a result of adding suffixes to the root forms of words. These suffixes can be removed by lemmatization. . Thus different forms of the same word are reduced to the same term [8].

4.1.4 Stemming

Stemming transforms words into their stems (root forms), which generalizes the texts for similarity analysis.

4.1.5 Synonym Replacement

A plagiarist never wants to be discovered. So they may either insert or remove some parts of a sentence or just paraphrase it. In this step one of a word and all its synonyms must be substituted for all of them. This way the algorithm can detect paraphrasing [2].

4.2 Candidate Document Retrieval

In this stage pair-wise comparison between each suspicious text against all source texts is done. One or more similarity metrics are applied to give each suspicious-source text pair a similarity score. The likelihood is determined by setting a threshold on the similarity scores. If a pair has reached a certain threshold, the pair is listed as a candidate pair; otherwise the pair is discarded as not plagiarized.

One of the most common similarity measures is Jaccard Similarity with overlapping word n-grams. In this approach texts are compared pair-wise and word n-grams are extracted from the documents and the Jaccard coefficient is used to find the similarity [8].

$$\text{Jaccard Similarity} = \frac{\text{amount of overlapping n-grams}}{\text{union of n-grams in both texts}}$$

4.3 Document Comparison Techniques

4.3.1 Term Frequency and Inverse Document Frequency (TF-IDF) with Vector Space Model (VSM)

The vector space model is a generative model, which is often applied to information retrieval or other text process tasks. For plagiarism detection, VSM could be seen as one global similarity measure method. Sentences extracted from the suspicious and source documents are seen as consisting term groups which are mutual independent. For each term, a weight which is computed by TF-IDF is given. The similarity of the two sentences could be measured by the cosine distance as the formula:

$$D(I_S, I_R) = \frac{\sum_{k=1}^{n} w_{S_k} * w_{R_k}}{\sqrt{(\sum_{k=1}^{n} w_{S_k}^2)(\sum_{k=1}^{n} w_{R_k}^2)}} \quad \text{..... (1)}$$

Where I_S and I_R are pair of sentences from the suspicious document S and source document R; W_{Sk} and W_{Rk} are the weights of terms in S and R respectively [14].

4.3.2 Multinomial Naïve Bayes

Naïve Bayes classifier [2] is suitable for pattern recognition can be used for plagiarism detection. This classifier is based on Bayes theorem. When S with small number of classes or outcomes conditional on several features denoted by t_1, $t_2 \ldots t_n$. using Bayes theorem:

$$P(S|t_1,t_2,\ldots,t_n) = \frac{P(S).P(t_1 \ldots t_n)|S}{P(t_1 \ldots t_n)} \quad \text{..... (2)}$$

Using conditional probability:

$$P(S|t_1,t_2,\ldots t_n) = P(S).P(t_1,t_2,\ldots,t_n)|S \quad \text{..... (3)}$$

4.3.3 Semantic Role Labeling

SRL is a process to identify and label arguments in a text. The basic idea is that the sentence level semantic analysis of text determines the object and subject of a text. It can be extended to the characterization of events such as determination of "who" did "what" to "whom", "where", "when", and "how". The predicate of a clause (usually a verb) establishes "what" took place, and other parts of the sentence express the other arguments of the sentence (such as "who" and "when"). The primary task of semantic roles labeling is to identify what semantic relation holds among a predicate and its associate participants or properties, with these relations drawn from a pre-defined list of possible semantic roles for that predicate or class of predicate. The typical labels used in SRL are Agent, Patient and Location for the entities participating in an event. Those labels can be extended to more specific arguments such as time and place in some text [18].

Plagiarism Detection using SRL

SRL aims to detect the arrangement similarity between concepts of the documents and possible semantic similarity between both documents. This step in the study used the role labels of the concepts for the documents and collected them as groups. The groups that were used provided a quick guide to capture the suspected part of the document [14].

4.3.4 Fingerprinting based Plagiarism Detection

For each document in the collection, fingerprints are generated. The fingerprints, which are hash codes (n-grams), represent the document. Hence comparison of the whole document is not done and only the fingerprints need to be compared. Extensive comparisons are thus reduced. Substrings are selected from the document using positional, frequency based, structural based selection and full fingerprinting methods. These substrings are converted as hashed index for further querying.

There are different schemes available for the selection of fingerprints [8]. They are:

4.3.4.1 N-gram

Every chunk of the document is selected as fingerprints. The advantage of this method is its easy implementation and the disadvantage is that it does not detect similarity when there is deletion, insertion, or reordering of text. As an example, all the fingerprints will be shifted by one position, even by the insertion of a single letter at the beginning of the document. So, no similarity detection will occur because, the altered document will have no common fingerprint with the original documents. For a document D, the number of fingerprints is computed as

$$M_{n\text{-gram}}(D) = L(D) - k + 1 \quad \text{..... (4)}$$

L(D) = the term count of document D.

4.3.4.2 0 mod p hash

For an inter value p this method selects fingerprints located at every 0 mod p. If the hashes of copied text belong to those selected by the 0 mod p, then copied content is detected. Here the number of the fingerprints is reduced by the value of p.

$$M_{0 \bmod p} = M_{n\text{-gram}}(D) / p \quad \text{..... (5)}$$

4.3.4.3 Winnowing

Schleimer et. al. proposed the selection strategy Winnowing. First the pieces of the documents are generated using the n-gram technique. Then hash values are produced using a hash function on these pieces of text. Hash values are numerical representations. Next another window of fixed size is used to iterate the previous step on the hash-values generated instead of the original text. At last, from each window the least hash-value is selected. The rightmost occurring hash-value is

selected, if multiple hash-values with the minimum value are present. For any input text document a set of fingerprints to represent it is obtained as a result of winnowing. The input in this work is a document that undergoes several steps of pre-processing, for extracting its text and normalizing it. Then, the normalized text will be segmented into sentences and each of these sentences is passed through the winnowing phase and its fingerprints are generated. Pairwise comparisons are done between all possible combinations of sentences fingerprints to detect plagiarism.

4.3.5 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is used to capture co-occurrence amongst words/features/attributes. It is a way of dimensionality reduction i.e. to identify which features play more important role in classification; and the idea is to take a combination of these features and work with a smaller number of features/attributes. This is achieved by Singular Value Decomposition (SVD).

The term-document matrix 'A' is decomposed into three independent matrices U, Σ and V. All matrices can be decomposed in a reduced latent space k to perform the best k-rank approximation of A so that singular values σ_{k+1}, σ_{k+2},…, σm are replaced by 0, where 1 <= k <= m. Then, matrix U is an n-by-k column orthonormal, whose columns are phrase singular vectors. Σ is a k-by-k diagonal matrix without negative and zero numbers that represents singular values.

$$\sigma_1 >= \sigma_2 >= .....\sigma_m > 0 \quad \text{……..…… (6)}$$

A characteristic feature of SVD is that the singular values on the diagonal of ∑ are placed in descending order and satisfy Eqn. 1. Matrix VT is a k-by-m row orthonormal, whose rows are document singular vectors. After the decomposition, matrix VT is an essential building element for further processing, since it contains independent profile vectors of the examined documents [7].

4.3.6 Fuzzy Semantic Similarity

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 [10].

Initially the basic pre-processing steps like tokenization, spelling corrections etc. are carried out. POS tagging is used for 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. 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.

In fuzzy-semantic similarity measure, different ranges of similarity scores are used. This reduces the pruning of matched fragments and inclusion of unmatched one's. The measure is as follows:

$$F_{q,k}(x) = \begin{Bmatrix} 1.0\ if\ s = 1.0 \\ 0.7\ if\ s \in [0.7, 1.0) \\ 0.5\ if\ s \in [0.5, 0.7) \\ 0.3\ if\ s \in [0.3, 0.5) \\ 0.2\ if\ s \in (0.0, 0.3) \\ 0.0\ if\ s = 0.0 \end{Bmatrix} \text{……..……..…… (7)}$$

$F_{q,k}(x)$ is the fuzzy semantic similarity between w_q (word in source sentence)and w_k (word in suspicious sentence).

The comparisons are done by formation of N-grams. In this 'N' consecutive words of suspicious and source document is compared. 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.

4.4 Passage Boundary Detection and Evaluation

Given a suspicious document and a source document, matches between the two documents are identified using some seed heuristic. The heuristic seeds are often extracted by the similarity measure between seeds. By coming up with as many reasonable heuristic seeds as possible, the subsequent step of growing them into aligned passages of text becomes a lot easier. The seeds could be sentence, word or character n-grams.

After seeding, the seed matches are merged into aligned text passages of maximal length between the two documents which are then reported as plagiarism detections. Rationale for merging seed matches is to determine whether a document contains plagiarized passages at all rather than just seeds matching by chance, and to identify a plagiarized passage as a whole rather than only its fragments.

Given a set of aligned passages, a passage filter removes all aligned passages that do not meet certain criteria. Rationale for this is mainly to deal with overlapping passages and to discard extremely short passages [15].

Finally, evaluation is done based on four standard measures-Recall (Rec), Precision (Prec), Granularity (gran) and F-measure.

5. COMPARISON OF PLAGIARISM DETECTION TECHNIQUES

In this section comparison of various plagiarism detection techniques for various languages is done.

Table 1. Comparative Analysis of Plagiarism Detection Techniques

Sr No	Name of Algo	Advantages	Disadvantages	Implemented on Languages	Observations
1	TF-IDF and VSM	Simple model based on linear algebra Term weights are not binary	Suffers from synonymy and polysemy Theoretically assumes that terms are statistically independent Cannot capture semantics	English	Performs well for no obfuscated data but not so well in case of paraphrased data. Proposed method made use of smaller feature set.
				Persian	Gives a good F-measure score and is also very fast. If S is the size of document collection and K be the number of features, then time complexity of the method is O(S*K)
2	Multinomial Naïve Bayes	Easy to implement Feature Independence and bag of words assumption.	Assumes feature independence and bag of words Data scarcity	English, Marathi	Gives better results than fuzzy semantic similarity
3	Semantic Role Labeling and CHAID	Captures the semantics of the sentences. Performs comparison based on labels only.	Giving a label to all the argument increases complexity and computational time	English	Gives better performance compared to fuzzy semantic based string similarity, LCS and semantic similarity Time complexity is O(n2) where n is the number of arguments
4	Fingerprinting and Winnowing	Document matching and document storage can be done in fingerprint feature only thus reducing the space required to store the data	Fingerprinting using n-gram does not detect similarity when there is deletion, insertion or reordering of text	Malayalam	k-gram method can be considered best when the collection is small For detecting near duplicates, winnowing gives better results
				Indonesian	Maximum accuracy value of fingerprinting is higher than winnowing but winnowing algorithm has better and more stable performance
5	Latent Semantic Analysis	The documents and words end up being mapped to the same concept space. The concept space has	LSA cannot handle polysemy effectively. LSA depends heavily on SVD which is	Arabic	Outperformed Plagiarism Checker X. Time complexity is O(n2) for pairwise matching.

		vastly fewer dimensions compared to the original matrix.	computationally intensive and hard to update as new documents appear.	Indonesian	LSA method gave higher performance compared to VSM. System detects all types of literal plagiarism and partial intelligence plagiarism.
6	Fuzzy Semantic Similarity	Best suitable for complex plagiarism cases. By incorporating NLP techniques, number of comparisons can be reduced, thus reducing the computational cost.	Hard to develop a model from a fuzzy system. Require more fine tuning and simulation before operational	English	POS method integrated with improved fuzzy semantic similarity measure surpasses the other method in term of accuracy and efficiency.
				Persian	To identify paraphrasing based on sentences, fuzzy method is effective
				Marathi	The proposed helps to detect more deep plagiarized text in Marathi research work

6. CONCLUSION

In this paper various techniques for plagiarism detection for different languages were discussed. A wide range of text pre-processing techniques have been explored in terms of the quality of plagiarism detection. Numbers of plagiarism detection tools have been developed for documents containing English text but very less work has been done for Indian regional languages like Hindi and Marathi. From the literature survey it is observed that Naïve Bayes, N gram, NLP Techniques might be suitable for languages like Hindi and Marathi and also Fingerprinting and Winnowing which has been implemented for Malayalam.

7. ACKNOWLEDGMENTS

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks to the Head of Computer Department and to the Principal of Pillai College of Engineering (PCE), New Panvel for extending their support.

8. REFERENCES

[1] Miranda Chong, "A Study on Plagiarism Detection and Plagiarism Direction Identification Using Natural Language Processing Techniques", PhD Thesis, University of Wolver Hampton, UK, 2013.

[2] Peyman Mahdavi, Zahra Siadati and Farzin Yaghmaee, "Automatic External Persian Plagiarism Detection Using Vector Space Model", Computer and Knowledge Engineering (ICCKE), 4th International eConference, Oct 2014.

[3] Urvashi Garg and Vishal Goyal, "Maulik: A Plagiarism Detection Tool for Hindi Documents", Indian Journal of Science and Technology, Vol 9(12), DOI: 10.17485/ijst/2016/v9i12/86631, March 2016.

[4] Nilam Shenoy and M.A. Potey, "Semantic Similarity Search Model for Obfuscated Plagiarism Detection in Marathi Language using Fuzzy and Naïve Bayes Approaches", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. V (May-Jun. 2016), PP 83-88 www.iosrjournals.or.

[5] Vani K, Deepa Gupta, "Investigating the Impact of Combined Similarity Metrics and POS tagging in Extrinsic Text Plagiarism Detection System", Advances in Computing, Communications and Informatics (ICACCI), International Conference, Aug 2015.

[6] Hui Ning, Cuixia Du, Leilei Kong, Haoliang Qi and Mingxing Wang, "Comparisons of Keyphrase Extraction Methods in Source Retrieval of Plagiarism Detection", Computer Science and Network Technology (ICCSNT), 4th International Conference, Dec 2015.

[7] Ashraf S Hussein, "Arabic Document Similarity Analysis using N-grams and Singular Value Decomposition", Research Challenges in Information Science (RCIS), IEEE 9th International Conference, May 2015.

[8] Sindhu L and Sumam Mary Idicula, "Fingerprinting based Detection System for Identifying Plagiarism in Malayalam Text Documents", Computing and Network Communications (CoCoNet), International Conference, Dec 2015.

[9] Vani K and Deepa Gupta, "Using K-means Cluster Based Techniques in External Plagiarism Detection", Contemporary Computing and Informatics (IC3I), International Conference, Nov 2014.

[10] Deepa Gupta, Vani K and Charan Kamal Singh, "Using Natural Language Processing Techniques and Fuzzy-Semantic Similarity for Automatic External Plagiarism Detection", Advances in Computing, Communications and Informatics (ICACCI, International Conference, Sep 2014.

[11] MAC Jiffriya MAC Akmal Jahan and Roshan G. Ragel, "Plagiarism Detection on Electronic Text based Assignments using Vector Space Model", Information and Automation for Sustainability (ICIAfS), 7th International Conference Dec 2014.

[12] Sidik Soleman and Ayu Purwarianti, "Experiments on the Indonesian Plagiarism Detection using Latent Semantic Analysis", Information and Communication Technology (ICoICT), 2nd International Conference, Oct 2014.

[13] Sindhu L, Bindu Baby Thomas and Sumam Mary Idicula, "Automated Plagiarism Detection System for Malayalam Text Documents", International Journal of Computer Applications (0975 – 8887) Volume 106 – No. 15, November 2014.

[14] Ahmed Hamza Osman and Naomie Salim, "An Improved Semantic Plagiarism Detection Scheme Based on Chi-squared Automatic Interaction Detection", Computing, Electrical and Electronics Engineering (ICCEEE), International Conference, Aug 2013.

[15] Shuai Wang, Haoliang Qi, Leilei Kong and Cuixia Du, "Combination of VSM and Jaccard Coefficient for external plagiarism detection", Machine Learning and Cybernetics (ICMLC), International Conference, July 2013.

[16] Agung Toto Wibowo, Kadek W Sudarmadi and Ari M Barmawi, "Comparison Between Fingerprint and Winnowing Algorithm to Detect Plagiarism Fraud on Bahasa Indonesia Documents", Information and Communication Technology (ICoICT), International Conference, Mar 2013.

[17] Sindhu L, Bindu Baby Thomas and Sumam Mary Idicula, "A Copy detection Method for Malayalam Text Documents using N-grams Model", 2013.

[18] Ahmed Hamza Osman, Naomie Salim, Mohammed Salem Binwahlan, Ssennoga Twaha, Yogan Jaya Kumar and Albaraa Abuobieda, "Plagiarism Detection Scheme Based on Semantic Role Labeling", Information Retrieval & Knowledge Management (CAMP), International Conference, Mar 2012.

[19] Asif Ekbal, Sriparna Saha and Gaurav Choudhary, "Plagiarism Detection in Text using Vector Space Model", Hybrid Intelligent Systems (HIS), 12th International Conference, Dec 2012.

[20] Salha Alzahrani, Naomie Salim, "Fuzzy Semantic based string similarity for extrinsic plagiarism detection", Lab Report for PAN CLEF, 2010.

[21] Shima Rakian, Faramarz Safi Esfahani, Hamid Rastegari, "A Persian Fuzzy Plagiarism Detection Approach", Journal of Information Systems and Telecommunications, Vol. 3, No. 3, July-September 2015.

[22] Stanford University video tutorial on SVD, https://www.youtube.com/watch?v=P5mlg91as1c (Last accessed on 11th Nov, 2016 at 4 pm.)