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.org

Semantic Similarity Search Model for Obfuscated Plagiarism Detection in Marathi Language using Fuzzy and Naïve Bayes Approaches

Ms. Nilam Shenoy[1], Mrs. M. A. Potey[2]
[1](M.E student, Computer Department, DYPCOE, Akurdi, Savitribai Phule Pune University, India)
[2](HOD of Computer Department, DYPCOE, Akurdi, Savitribai Phule Pune University, India)

Abstract: *Plagiarism detection (PD) in natural language texts is an example of NLP applications that is linked with information retrieval (IR) and soft computing (SC) approaches. Obfuscated plagiarism cases contain invisible texts, which is difficult to find in existing plagiarism detection methods. In this paper fuzzy semantic-based similarity search model and Naïve Bayes model for uncovering obfuscated plagiarism for English and Marathi language are presented and compared with different state-of-the-art baselines (B1-W1G, B2- W3G, B3-W5G, B4-S2S). 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. Results are assessed using precision, recall, F-measure and granularity for Fuzzy and Naïve approaches and it is observed that Naïve Bayes model gives more appropriate result than fuzzy semantic based model.*

Keywords: *Fuzzy, Naïve Bayes, Obfuscated, Plagiarism detection, Semantic similarity.*

I. Introduction

The Plagiarism could be fuzzier than apparent, more difficult than trivial copy and paste [1]. The word plagiarize means kidnapping, which is discrete as to get ideas, articles, image, documents, audio, code, from other authors and pass them as one's own without giving credit to originator. Hence detecting plagiarism cases is a global problem [2], which can happen in several different areas of our life. The Plagiarism detection work for English is previously done using various techniques. Our proposed system is for Plagiarism detection in Marathi Language using Fuzzy and Naive Bayes Approaches. Our Literature survey says that, there are different types of Plagiarism detection, like in academic field Plagiarism can be a very de-motivating issue for teachers and students. This System will be more useful for Marathi Organizations and Researchers. If plagiarism is not addressed satisfactorily, plagiarists could increase unwarranted advantage, e.g. giving more marks for their assignments with less efforts. Plagiarism detection of document plays important role in other applications also, such as file management, plagiarism prevention and copyright protection.

The simplest and common way to execute plagiarism is to copy-paste texts from its original resources. This is called literal plagiarism and is easy to mark by current available Plagiarism detection tools. Another type of plagiarism called Obfuscated Plagiarism which includes different types of plagiarism like Cross-Language plagiarism, Idea plagiarism, Summarized plagiarism, Citation-based plagiarism. In Yerra and Ng paper [2], a copy detection approach for web documents was formulated using fuzzy based information retrieval (IR) model. The basic concept in fuzzy Information Retrieval shows that words in a document have definite degree with a fuzzy set that has words with associated meaning and two documents are considered similar although their semantic content may be different if they increase high similarity degree with the fuzzy set [1]. Thus, fuzzy Information Retrieval model has proved to work well for partially related semantic content in web retrieval. A recent literature review on the field of Plagiarism Detection (Alzahrani et al., 2012) has shown that there is a need for more effective algorithms to find deep plagiarism that are semantically, but not syntactically, the same with original texts [3]. Most of the current Detection methods fail to detect unseen deep (obfuscated) plagiarism cases because the similarity metrics of compared texts are calculated without any knowledge of the linguistic and semantic structure of the texts.

II. Related Work

A. Semantic Similarity Measure

In lexical categorization, such as the WordNet (Miller, 1995) [4], lexes are arranged into 'has-a' and 'is a' hierarchies wherein words with the same meaning are grouped together. They are called synsets. Synsets linked with more abstract words called hypernyms, and most particular words called hyponyms. The text features applied in Plagiarism Detection methods are Lexical features, Syntactic features, and Semantic features.

Table 1: Text Features

Sr. No	Text Features	Example
1	Lexical Features	n-grams Character, n-grams word
2	Syntactic Features	Word order Sentence, Chunks, Structure, Part-of-Speech, Phrase
3	Semantic Features	Synonyms, Hypernyms, Hyponyms

If two sentences are given then similarity between two texts T1 and T2 is defined as follows:

$$Sim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w \epsilon T_1} maxSim(w, T_2) \times idf(w)}{\sum_{w \epsilon T_1} idf(w)}\right) - \frac{1}{2}\left(\frac{\sum_{w \epsilon T_1} maxSim(w, T_2) \times idf(w)}{\sum_{w \epsilon T_1} idf(w)}\right) \quad (1)$$

B. Feature Extraction Method

Feature Extraction is the prepossessing Method which contains two types of textual structures. The first goal at describing the text as word k-grams where k is typically set before the experiments. The second goal at splitting the text into sentences using end-of statement delimiters (i.e., full-stops marks, question marks, and exclamation marks).

C. FEM framework

Feature Extraction Method is used to characterize text in the form of part-of-speech (POS) and Lexicons. Fig. 1 shows the major components of Feature Extraction Method.

1) **Tokenization**: The input text given to system is divided into tokens, and each token is marked as token [T], or end-of-sentence [E].

2) **Lemmatization**: A lemmatizer is applied on the extracted tokens from input text, WordNet (Miller, 1995) dictionary form is provided for each word. In Lemmatizer the tokens are replaced with lemmas [L]. This would help, to compare the semantic meaning of two sentences based on the semantic relatedness of their words derived from the WordNet.

3) **Stop words removal**: The most frequent English words like ‘a’, ‘an’, ‘the’, ‘is’, ‘are’, etc., are removed from the input text. In this step most of the conjunctions and interjections will be removed from input text. The stop words (127 words) list has been obtained.

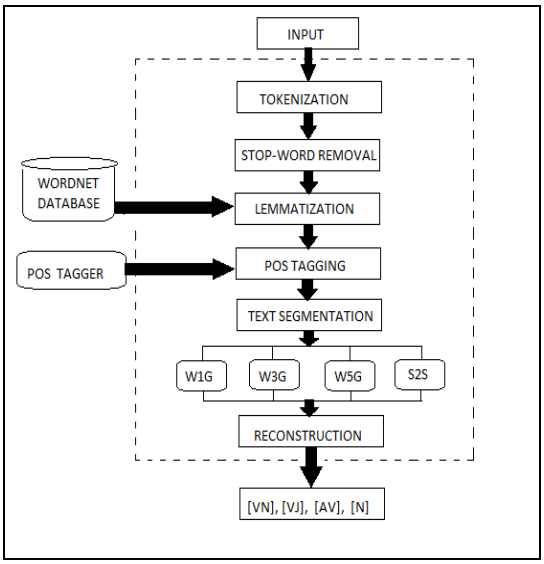


Fig.1. Feature Extraction Method

4) **Text segmentation**: The resulting text is segmented into word 1 grams (W1G), word 3-grams (W3G), word 5-grams (W5G) and sentence to sentence (S2S). These different segmentation schemes will be compared during the experimental work in terms of which approach can better handle obfuscated plagiarism cases along with the proposed fuzzy semantic-based similarity method and Naïve Bayes Model.

5) **Part-of-Speech (POS)**: The lemmas in each segment are categorized into the following tags: noun [N], verb [V], adjective [AJ] or adverb [AV]. In this regard, a transformation function is used to convert multiple Tags into our tags. For instance, [VB], [VBD], [VBN], [VBG] will be [V], and so on.

D. Fuzzy Semantic-Based Model

For Plagiarism Detection Deep word similarity detection analysis between two input texts utilizing their POS-related semantic spaces. Semantic relation between two words can be defined based on the is-a relationship from WordNet lexical taxonomies (Miller, 1995) [4]. According to Yerra and Ng (2005) [5], matching two sentences can be approximate, which can be modeled by considering that each word in a sentence is associated with a fuzzy set that contains the words with the same meaning, and calculates semantic similarity score (usually less than 1) between words (in a sentence) and the fuzzy set. If two words are exactly same then similarity score is 1 and if both word are totally different then similarity score is 0.

E. Naïve Bayes Model for Plagiarism Detection

Naïve Bayes classifier are suitable for pattern recognition can be used for source code author identification. 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 [6]:

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

Using conditional probability:

$$P(S|t_1, t_2..t_n) = P(S)P(t_1..t_n)|S \qquad (3)$$

Naïve Bayes classifier gives more appropriate result than Fuzzy-Semantic based Model.

III. Implementation Details

System Architecture for implementation of Semantic Similarity Search Model for Obfuscated Plagiarism Detection for Marathi Language using Fuzzy and Naïve Bayes techniques is

A. System Architecture

Figure 2 shows the general framework of this model. Two input texts (might be of document size) are used in the feature extraction method. The resulting features from the texts are used as inputs to the fuzzy inference system, whereby a semantic similarity measurement is modeled as a membership function. After the evaluation of the rules, the outputs are aggregated into a single value which can be interpreted as a similarity score between input texts.

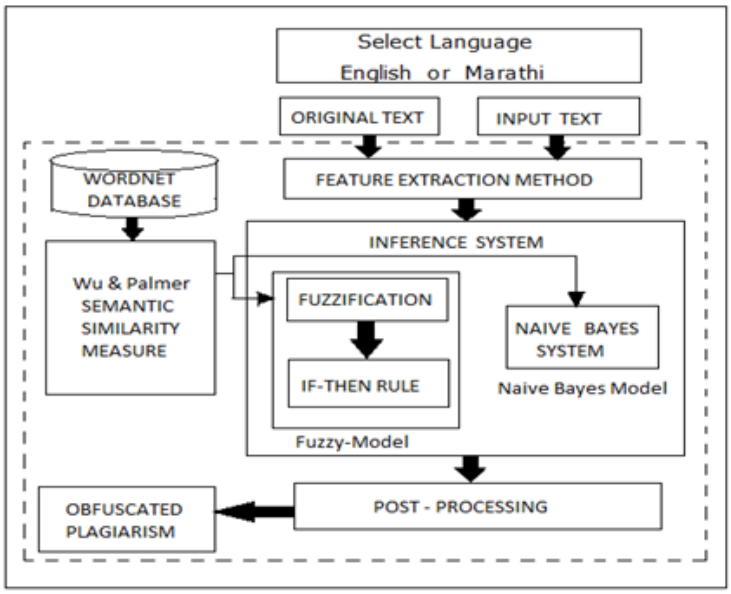


Fig.2. System Architecture

B. Mathematical Model

The Mathematical representation of Obfuscated Plagiarism Detection consist of system S = {I, F, O, D}, I = Set of Input, F = function, O = Set of output, D = Datasets used for training and testing purpose, I = {T}, T =

$\{t_1, t_2, t_3\ldots, t_n\}$ = Group of words, F = $\{f_1; f_2\}$, f1 = $\{T_1, S, P, L, T_2\}$ = Feature Extraction Method, T_1 = Tokenization, S = Stop-Word Removal, P = POS Tagging, L = Lemmatization, T_2 = Text Segmentation, f_2 = {Representation of I}, O = $\{O_1\}$, O_1 = {Literal or Obfuscated Plagiarism Detected cases}.

This System uses group of words $\{t_1, t_2, t_3\ldots, t_n\}$ as a input. Then Feature Extraction Method will apply on text and then similarity search will calculate using similarity search equation 1. Then Naïve Bayes classifier and fuzzification rules will apply on source text and check for literal or Obfuscated Plagiarism detection cases.

C. ***Algorithm***

A detailed checking should be carried out between source and suspicious texts in order to locate similar fragments. The final output of the algorithm is a list of segment pairs.

Algorithm 1 Detailed checking Algorithm

Input Text A
Input Text B \\(Input texts are either English or Marathi)
Choose segmentation method W1G, W3G, W5G, S2S
Apply FEM for Text A \\FEM-(Feature Extraction Method)
Apply FEM for Text B
For each segment $A_i \varepsilon$ A
For each Segment $B_j \varepsilon$ B
Input A_i and B_j to inference engine(either fuzzy model or Naïve Bayes model)
Compute SIM(A_i,B_j)
If PD(A_i,B_j) is true \\PD-(Plagiarism Detection) Output (A_i,B_j)

Algorithm 1 [1] provide a pseudo code for the detailed checking algorithm used in implementation. To reduce the algorithm complexity, first the source programs are divided to overlapped substrings with the length sequence of k which is regarded as the smallest unit hence we are using word-k-grams instead of sentences and Lemmatiser instead of the stemmer to get better results from WordNet. This algorithm gives more detection efficiency and less time complexity.

D. Experimental Setup

The fuzzy based and Naïve Bayes systems are built using Java framework (version jdk 8) on windows 7 platform. The NetBeans version (version 8.0.2) is used as development tool. The system doesn't require any specific hardware to run; any standard machine is capable to run the application.

E. ***Dataset***

In this system we have identified three datasets for English. The first two corpora, PAN-PC- 11 (Potthast et al., 2011) [7] and PAN-PC-10 (Potthast et al., 2010a, b) [8], include 7645 manual paraphrases and 34,310 automatic paraphrases [1]. PAN-PC-09 (Potthast et al., 2009b) involve 17, 127 artificial cases but no simulated plagiarism cases were found [1]. For Marathi language we have collected some manual and artificial paraphrases for testing.

IV. Result and Discussion

A. ***Result***

To evaluate the methods Precision (P), Recall (R), Granularity (G), Harmonic-mean (F), Plagiarism Score(S) is calculated.

$$P = \frac{TP}{TP + FP} \qquad (4)$$

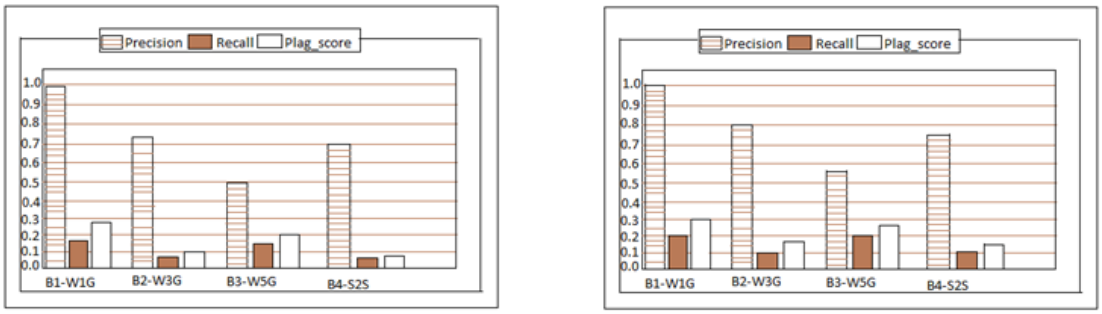
$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F = 2 * \frac{PR}{P + R} \tag{6}$$

Where TP refers to the number of correct plagiarism cases, FP refers to the number of false detections of cases annotated, FN refers to the number of plagiarism cases that are not detected as plagiarism. For calculating granularity equation 7 is used.

$$G = \frac{NP_{detected} : P_{detected} \subseteq P_{annotated}}{NP_{annotated}} \tag{7}$$

Where $NP_{detected}$ is the number of true detections, $N_{Pannotated}$ denotes the number of annotated cases. Highest results obtained by the state-of-the-art baselines and the proposed methods are shown in bold. The first four columns give the mean precision, recall, granularity and score of plagiarism. The last column shows the standard deviation. It is noticed that manual and artificial datasets behaved differently. The accuracy of the results on handmade paraphrases was overall exceeding that on artificial paraphrases given the same segmentation scheme.



(a) (b)

Fig 3: Recall, precision, and plagiarism score results from Fuzzy-semantic based (a) model and Naïve Bayes (b) model with four Baselines (B1-W1G, B2-W3G, B3-W5G).

Figure 3 shows the Recall, precision, and plagiarism score results from Fuzzy-semantic-based model and Naïve Bayes Model with four Baselines (B1-W1G, B2-W3G, B3-W5G). Results are showing that Naive Bayes gives more appropriate result than Fuzzy semantic based model.

V. Conclusion

A fuzzy semantic-based similarity model and Naïve Bayes model for uncovering obfuscated plagiarism in Marathi Language is presented and is compared with different state-of-art baseline (B1-W1G, B2-W3G, B3-W5G). Results are assessed using Precision, Recall, F-measure and Granularity for Fuzzy and Naïve Bayes approaches and it is observed that Naïve Bayes gives better detection performance than fuzzy semantic based model. The proposed system is more effective than the existing system, since it helps to detect more deep plagiarized text in Marathi Research work. Future work will include experiment on cross Language Plagiarism Detection between Marathi and English Language and integration on more semantic rules, which can be used by Marathi researchers and organizations.

References

Journal Papers:

[1] S. Alzahrani and N. Salim, "Fuzzy semantic-based string similarity for extrinsic plagiarism detection," Braschler and Harman, 2010.

[2] S. M. Alzahrani, N. Salim, and V. Palade, "Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model," Journal of King Saud University-Computer and Information Sciences, vol. 27, no. 3, pp. 248–268, 2015.

[3] S. M. Alzahrani, N. Salim, and A. Abraham, "Understanding plagiarism linguistic patterns, textual features, and detection methods," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 42, no. 2, pp. 133–149, 2012.

[4] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 133– 138, Association for Computational Linguistics, 1994.

[5] R. Yerra and Y.-K. Ng, "A sentence-based copy detection approach for web documents," in Fuzzy Systems and Knowledge Discovery, pp. 557–570, Springer, 2005.
[6] J. Z. Kolter and M. A. Maloof, "Learning to detect malicious executables in the wild," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 470–478, ACM, 2004.
[7] M. Potthast, A. Barrón-Cedeño, B. Stein, and P. Rosso, "Cross-language plagiarism detection," Language Resources and Evaluation, vol. 45, no. 1, pp. 45–62, 2011.
[8] L. Luo, J. Ming, D. Wu, P. Liu, and S. Zhu, "Semantics based obfuscation-resilient binary code similarity comparison with applications to software plagiarism detection," in Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering, pp. 389–400, ACM, 2014.
[9] S. M. Alzahrani and N. Salim, "On the use of fuzzy information retrieval for gauging similarity of arabic documents," in Applications of Digital Information and Web Technologies, 2009. ICADIWT'09. Second International Conference on the, pp. 539–544, IEEE, 2009.
[10] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 133–138, Association for Computational Linguistics, 1994.