CCS Combine Approach for detect para pharsing using Machine Learning Techniques

Dr. Durga Bhavani Dasari

Assistant Professor, Dept of CSE, Konerulakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India.

Dr. Venu gopala Rao. K

Professor, G. Narayanamma Institute of Technology and Science, Hyderabad, AP, India

Abstract: Due to heavy digital content generation it is very needful to protect data from copying of others content. Plagiarism is very useful for document ownership protection in the field of academic , research, journal and patents. Most of the existing algorithms are based on simple text similarity modals. Most popular plagiarism detector Turnitin is, because it is maintaining large web documents links, journals list and local database of submitted assignments. Reason tools to get popularity is speed and data size as local. Paraphrasing is article rewriting system where idea thief's can steel /copy with out giving credits to original author. No tool/system is available focusing on fully paraphrased content.

First time by doing combination of context, concept and semantic similarity derivations we are proposing a new super fast plagiarism detection system which over come copying and paraphrasing problems.

Keywords: Text Mining, Classification, Semitic Mining, Context & Concept Analysis

I. INTRODUCTION

Copyright infringement by understudies, educators, industrialist or analyst is viewed as scholarly misrepresentation. Literary theft is characterized in various manners like replicating others unique work without recognizing the creator or source. Unique work is code, recipes, thoughts, research, methodologies, composing or other structure. Discipline for literary theft comprises of suspension to end alongside loss of validity. Accordingly, distinguishing counterfeiting is fundamental. Exploration paper determination is repeating action for any meeting or diary in the scholarly community. It is a multi-measure task that starts with a call for papers.

Importance of plagiarism

1. **Stealing from Another**

- It is unethical
- Wouldn’t want to steal *your* work
- Not citing the source

II. LITERATURE REVIEW

2.1 Existing Plagiarism Detection Systems

The systems, designed to find similarities in the natural language texts, mainly search the Internet for the possible matches.

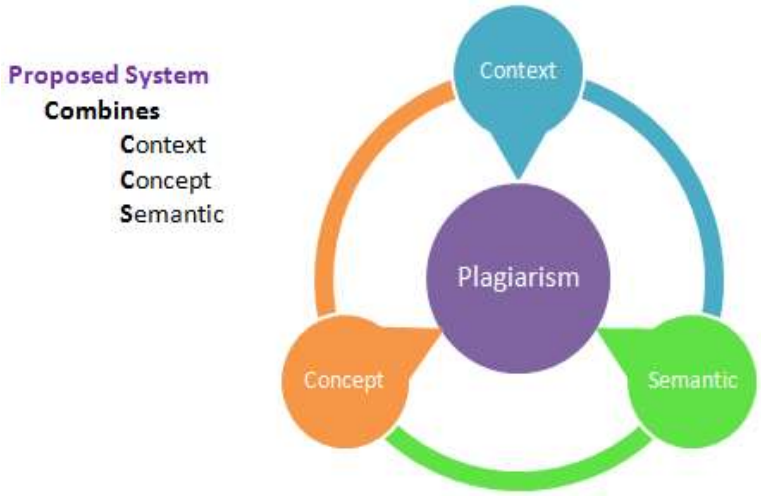
Generally, they do not use sophisticated comparison methods, aiming mostly at processing speed and wide coverage (e.g. the developers of Turnitin system claim they maintain "a huge database of books and journals, and a database of the millions of papers already submitted").

"Hermetic" systems for plagiarism detection in the natural language texts exist as well, though they are little-known. We can mention, e.g. CopyCatch Gold , YAP3 , and WCopyfind .

As a rule, the detection software can find only partial exact matches: rephrasing and rewording can conceal the evidence of plagiarism.

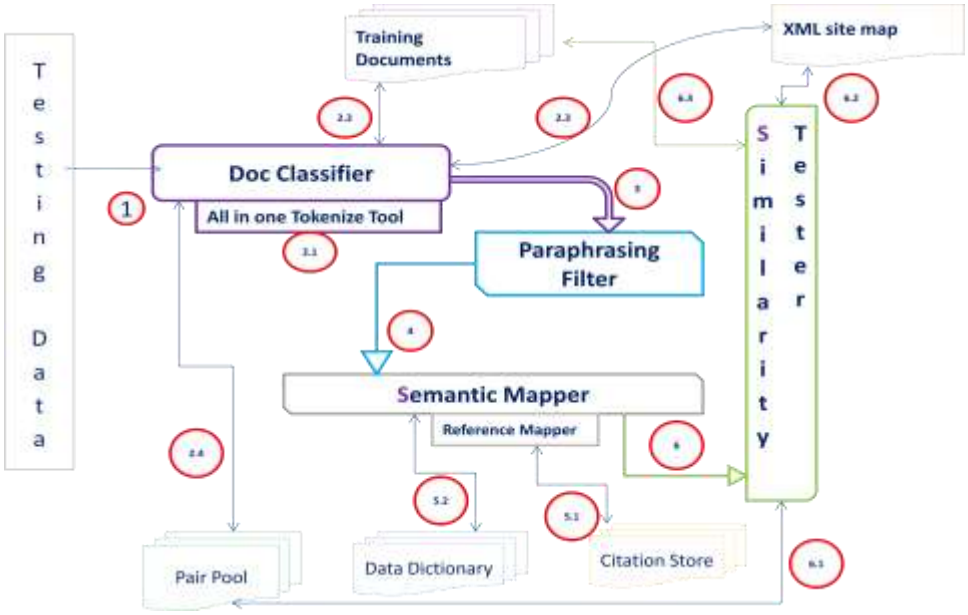
2.2 . Proposed System

As observations in existing systems/tools it is clearly stated that no approach is focusing on plagiarism by considering **Context, Concept and Semantic as combined.**



III. ASSOCIATE WORK

Figure 1: **Different phases of Architecture**



Proposed system architecture include following blocks/components

Training Documents: These are local copy of text documents in readable format like text and hyper text file format.

XML site map: Tag structured web pages list, to check for text copy from web.

Pair Pool: This is a CSV file which is maintain list of related documents with their identity.

Data Dictionary : Bin of terms synonyms and antonyms for implementing semantic

Citation Store : Maintain list of references defined of plagiarism checker

Doc Classifier: This is a functional block which will identify class of incoming testing data

Paraphrasing Filter: Main component of the architecture is this to check modifications in data like adding, replacing or removing characters.

Semantic Mapper : This functional component is for managing relation among testing words and data dictionary words.

Similarity Tester : Similarity among full document or blocks of document can be calculated between testing data and training data.

Reference Mapper : This functional tool is for managing citations.

All in one Tokenize Tool : Extracting sentences for separating full string into tokens. Replace stop words.

2.3. Different Phases

In phase-1: suspicious documents are sending to Doc Classifier

Suspicious documents are in text format or Hyper text format.

In Phase-2 : Doc Classifier and All in one tokenization toll is performing following functionalities

- Classifying input document into specified category
- Generating Pair details

In Phase-3: With term weight age and pattern matching , it can be detected whether sentence is paraphrased or not .

In Phase-4: Paraphrasing Filter is key component which takes summarized document to identify NP (Noun Phrase), VP (Verb Phrase) , CW (Connective Words)

In Phase-5:

- Word Relation Matrix is calculated
- Data Dictionary is used for calculating WRM
- Citation Store is for managing references.
- In Phase- semantic comparability between two sentences can be calculated by using WRM. Plagiarism status in percentage can be displayed

In Phase-6: Semantic comparability between two sentences can be calculated by using WRM. Plagiarism status in percentage can be displayed

IV.PROPOSED ALGORITHMS

4.1. CLASSIFICATION ALGORITHM

N is assigned number of class,

M is assigned number of associated sets

Step I: For each and every class i is assigned to 1 to N do

Step II . Set pval is assigned to 0, nval is assigned to 0, p is set to 0, n is set to 0

Step III. For each an every set s is assigned to 1 up to M do

Step IV. If the P(class (i)) for the set s is maxm then

++ pval else -- nval

Step V. If 50% of the set s is similarity with the keywords

set do step VI else do stepVII

Step VI. If maxm prob(p) matches the class of i then

++ p

Step VII. If maxm prob(p) does not similarity the class of i

++ n

Step VIII. If (s<=M)

go to step 3

Step IX. Find the % of matching in +ve sets for the class of i

Step X. Find the % of not matching in –ve sets for the class of i

Step XI. Find the total prob(p) as the sum of the results gained from

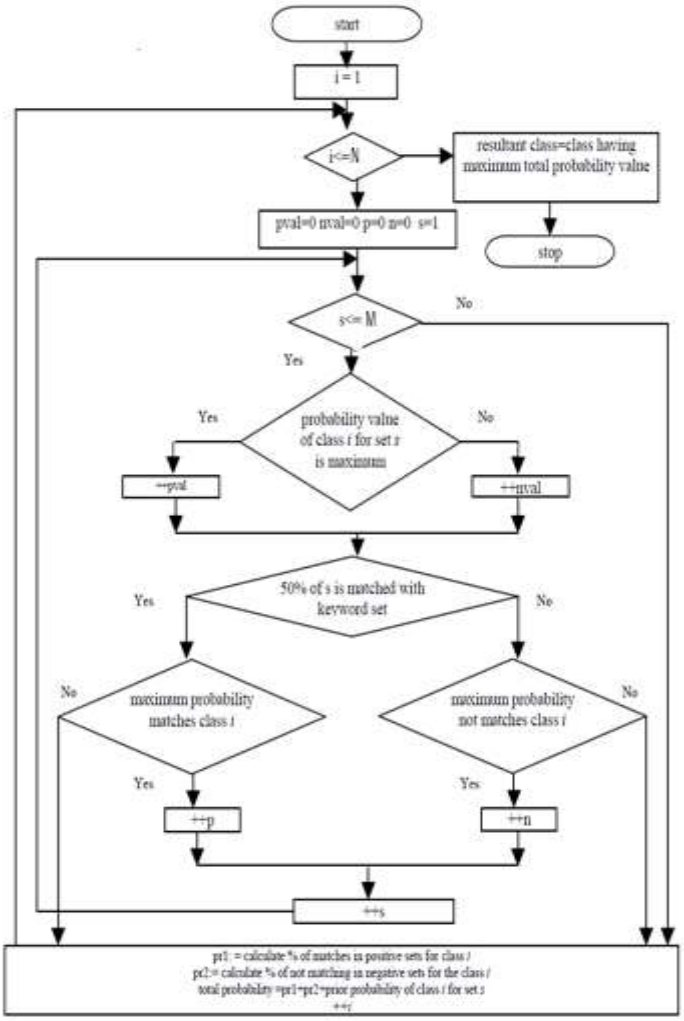
step IX and X and also the prior prob(p) of the class of i in set s

Step XII. If (i<=N)

go to step I

Step XIII. Set the class having the maxm prob(p) value as the result

4. 2 . CLASSIFICATION ALGORITHM FLOW CHART



V.SEMANTIC SIMILARITY

We will now describe the overall strategy to capture semantic similarity between two sentences. Given two sentences X and Y, we denote m to be length of X, n to be

length of Y. The major steps can be described as follows:

Step I:.Tokenization.

Step II: Perform word stemming.

Step III: Perform part of speech tagging.

Step IV: Word sense mapping nature.

Step V: Building a semantic similarity Relative matrix R[i, j] of each pair of word

senses, where R[i, j] is the semantic similarity between the most appropriate

sense of word at position i of X and the most appropriate sense of word at

position j of Y.

5.1. Pseudo code for computing similarity of two sentences X and Y

Step I : X is equal to zero
Step II : Y is equal to zero
Step III : Initialize to I is zero
I less than x
Max of I zero
Step Iv: J initialize to zero
J less than y
If I,j is garter than max of i
Step V: Sum of x =max of i
Step VI: Then max of j zero
Step VII: If I,j is grater than max of j
Step VIII: Sum of y = max of j

5.1. Semantic Similarity with Result

Given two sentences X and Y,

X and Y have lengths of 3 and 2, respectively.

The bipartite matcher returns that X[1] has matched Y[1] with a score of 0.8, X[2] has matched Y[2] with a score of 0.7:

Using Matching average, the overall score is : 2*(0.8 + 0.7) / (3 + 2) = 0.6.

Using Dice with a threshold is 0.5, since both the matching pairs have scores greater than

the threshold, so we have total of 2 matching pairs

VI. CONCLUSION

It is imperative that the plagiarism detection tools offer excellent service in terms of detecting the text that has similarities between the document sets.

Many of the plagiarism detection solutions have limitations in terms of identifying rightly the text that is cited and the ones that are plagiarized.

Despite of many developments that has emerged, still in terms of identifying the in-depth analysis of plagiarism, there is significant scope for development.

It certain contemporary range of plagiarism detection models are discussed with finite objectives of evaluating the concept, semantic, and context relevance.

REFERENCES

[1] Radev, D. R., Hovy, E., and McKeown, K. (2002). Introduction to the special issue on summarization. Computational Linguistics., 28(4):399{408.

[2] Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM, 16(2).

[3] H. Dalianis, "SweSum-A Text Summarizer for Swedish", Technical Report, TRITA-NA-P0015, IPLab-174, KTH NADA, Sweden, 2000.

[4] T. Chang and W. Hsiao, "A hybrid approach to automatic text summarization", 8th IEEE International Conference on Computer and Information Technology (CIT 2008), Sydney, Australia, 2008.

[5] Ghadeer Natshah, Yasmeen Ta‟amra, Bara Amar and Manal Tamini, "Text Summarization: Using combinational Statistical and Linguistic Methods"

[6] Vishal Gupta and Gurpreet Singh Lehal "A survey of Text summarization techniques "Journal of Emerging Technologies in Web Intelligence VOL 2 NO 3 August 2010.

[7] Suneetha Manne and S.Sammen Fatima‟s"A feature Terms based Method for Improving Text summarization with supervised POS Tagging".

[8] R.Shams, A.Elsayed and Q.M Akter, "A corpus-based evaluation of a domain-specific text to knowledge mapping prototype", A special issue of Journal of Computers, Academy Publisher, 2010(In Press)

[9] Chengcheng L.(2010).Automatic Text Summarization Based On Rhetorical Structure Theory. IEEE. 2010 International Conference on Computer Application and System Modeling (ICCASM 2010).

[10] http://www.summarization.com/mead/

[11] Joachim T. (2002) 'Learning to Classify Text Using Support Vector Machines', Methods Theory and Algorithms, Kluwer/Springer.

[12]. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, CRC Press, 1984.

[13]. Shariff M.N., Saisambasivarao B., Vishvak T., Rajesh Kumar T. (2017),'Biometric user identity verification using **speech recognition** based on ANN/HMM',Journal of Advanced Research in Dynamical and Control Systems,9(12 Special issue),PP.1739-1748.

[14]. Bhimanpallewar R., Narasinagrao M.R. (2017),'A **machine learning approach** to assess crop specific suitability for small/marginal scale croplands',International Journal of Applied Engineering Research,12(23),PP.13966-13973.

[15]..Ayushree, Balaji G.N. (2018),'Comparative analysis of coherent routing using machine learning approach in MANET',Smart Innovation, Systems and Technologies,77 (),PP. 731-741

[16]..Jahnavi P., Vamsidhar E., Karthikeyan C. (2019), 'Facial expression detection of all emotions and face recognition system', International Journal of Emerging Trends in Engineering Research, 7(12), PP.778-783.

[17]..Vijaya Chandra J., Challa N., Pasupuletti S.K. (2019), 'Machine learning framework to analyze against spear phishing', International Journal of Innovative Technology and Exploring Engineering, 8(12), PP.3605-3611.

[18]. Lalithabhavani B., Krishnaveni G., Malathi J. (2019), 'A comparative performance analysis of different machine learning techniques', Journal of Physics: Conference Series, 1228(1), PP.-.

[19]. **Sheela Rani**, Vuyyuru Tejaswi, Bonthu Rohitha, himavarapu Akhil, "Pre filtering techniques for face recognition based on edge detection algorithm", International Journal of Engineering & Technology, 7 (1.1) (2018) 213-218.

[20]..Gaikwad Kiran P, **Dr C M Sheela Rani**, "Comparative Analysis of Emotion states Based on Facial Expression Modality", Jour of Adv Research in Dynamical & Control Systems, Vol. 11, 01-Regular Issue, 2019, pp. 462 to 466.

[21]. K.Snigdha, P. Vishal, Shahana Bano, "**An Efficient face recognition system using Local Binary Pattern**", International Journal of Recent Technology and Engineering (IJRTE), 2019.

[22]. DurgaBhavaniDasari,Dr. VenugopalRao K (2018). Similarity check by Concept Relevance (SCCR): Plagiarism Detection in Text Documents. International Journal of Pure and Applied Mathematics. Volume 119 No.15, 1953-1967.

[23]. DurgaBhavaniDasari, Dr. VenugopalRao K (2018). Semantic Relevance Scale for Text Data Plagiarism Detection. Journal of Research in Dynamical & Control systems, Volume 10 01-Secial Issue – 2018

[24]. DurgaBhavaniDasari, Dr. VenugopalRao K (2018). Context Similarity for Text Data Plagiarism Detection. International Journal of Engineering and Technology. 7(2.32((2018)14-17