# INTEGRADTED TOOL DEVELOPMENT FOR TEXT AND CODE PLAGIARISM DETECTION USING TF-IDF AND LEVENSHTEIN DISTANCE ALGORITHM .

[1] **MrunaliDeshmukh** , [2] **KshitijaGhogare**, [3] **BhaktiGunjal**, [4] **ShitalPatil** , [5] **Prof KetakiNaik**

**Abstract :**

Today Plagiarism is most common approach taken to complete research, completing college assignments and proving innovation. Sometimes we copy others work as our innovation, for that we copy contents of their work. Real researchers and fake researchers are not identified. This problem occurs in both the cases code as well as text. To identify this problem there are many systems available on web forum but that system having some limitations like they work on either text or code? In the proposed system we are going to develop integrated tool for text and code plagiarism detection system. To solve this problem we have Use Levenshtein Distance and TF-IDF technique (Term frequency and inverse document frequency). Levenshtein Distance for Code level and TFIDF for document level. In TF-IDF, it finds importance of words in the form of weight after removing stop words. In this system real time dataset is used which contains documents having text and code.

***Index Terms*** **- Component,formatting,style,styling,insert.**

## I. INTRODUCTION

Plagiarism and its automatic retrieval have attracted large attention from study to industry: various papers have been published on the topic, and many commercial software systems are being developed [1].They may also get their code assignment from external public resources, especially the Internet. In some places, local companies may offer helping students partially or completely in those code projects. The Internet also includes several websites in which students can submit their code assignments but this may get duplicates. There are two main areas of possible plagiarism in the academia. Those are plagiarism in research papers, projects and publications. It also includes plagiarism that is especially applicable for students in the information technology majors [2].Now a day's huge amount of digital information is both advantageous as well as disadvantageous too [4]. Advantageous means that we can get each and every information on the net easily for reference and hence searching time for required information has reduced a lot [5]. Plagiarism has been termed as stealing, theft of concept, idea writings of other research scholar and presenting as inventor of it. In context to research today scholars are aware that plagiarism should be avoided but lack to understand what, why and how plagiarism occurs. Many successful plagiarism detection tools and software products have been developed. However, the detection of paraphrasing or confusing plagiarism remains a challenge because most of the existing tools are only able to detect copy-paste cases of plagiarism. Change in pattern of writing or language is most common technique employed for restructuring sentence to misguide scientific community and detection of search work is also major challenge [15]. Many Current techniques having capacity of exactly matched substring or some kinds of texture fingerprinting but that may not be sufficient as cases of rephrasing and rewording the content treated as different. Therefore, this work considers the problem of finding the suspected fragments that have the same semantics with the same/different syntax [4]. This research work focuses on effort estimation with plagiarism analysis for research articles and code assignments. This is a research towards idea based plagiarism detection. Existing techniques focuses on keyword matching and fail to detect hidden patterns of plagiarism. Proposed research focuses on diverse patterns of plagiarism with innovative framework design [12].

II. Proposed research focuses on diverse patterns of plagiarism with innovative framework design [12].

**Keyword:**

Plagiarism, Levenshtein Distance, TF-IDF, Core NLP

**Related Work**

Plagiarism seeds identification for the high-obfuscation proposed by Leilei Kong in which presents a multi-features fusion method. From suspicious document and source document, integrated lexicon features, syntax features, semantics features and structure features are extracted using this method. A multi-feature fusion classifier based on Logical Regression model is proposed to decide whether a text fragment pair can be regarded as plagiarism seeds or not [1]. Haolianget. al. proposed an effective method in which high-obfuscation plagiarism seeds presents a significant research problem in the field of plagiarism detection. To capture plagiarism seeds the conventional methods of plagiarism detection are used based on single type of features [2]. Rada Mihalceaet. al. uses a methods like measuring the semantic similarity of texts and using corpus-based knowledge-based similarity measures[3]. S. Santhosinidevi Proposed a system in which presents a measure of semantic similarity in an is-a taxonomy based on the notion of shared information content. Experimental evaluation against a benchmark set of human similarity judgments demonstrates that the measure performs better than the traditional edge-counting approach [4]. Alexander Maedche proposed a system in which Ontology serve as a means for communication at a similarity and semantic level of the text contents [5]. Samuel Fernando proposed a system in which presents a novel technique to the problem of paragraph identification. Although paraphrases often make use of identified or near words, many previous approaches have either ignored or made limited use of information about similarities between word meanings [6]. James O'Shea proposed a system in which describes a comparative study of STASIS and LSA. These measures of semantic similarity can be applied to short texts for use in Conversational Agents (CAs). CAs are computer programs that interact with humans through natural language dialogue [7]. TarasFinikov proposed a system in which influence of transformation processes in higher education to lower academic standards, changes and deformation in ethical field of global and national higher education. We considered the genesis and modern standards of academic integrity [8].

**Mathematical Model**

## Mathematical Model:

Mathematical model set theory S= {s, e, X, Y, Φ}
s= Start of the program
1. Register/Login into the system
2. Provide Dataset (Text Document And Code Document).
e= End of the program
Identify the Plagiarism Detection
X = input of the program= {P, R, Q,D}
P = data
R = Dataset
Q = Total Number of User Data Line
D = Total Number of Line In Dataset
Y = Output of program = Plagiarism Detected Line With Percentage

Let D be the set of User Data

D = {D1, D2, D3... ..... Dn}
Let A be the set of categories Dataset (Code & Text)
therefore,
T = {T1, T2, T3...,..... Tn}
C = {C1, C2, C3,..., Cn}
Text data is evaluated with TFIDF algorithm and code data is evaluated with Levenshtein distance algorithm.

**System Architecture:**

Figure Shows detailed flow of Plagiarism Detecting System. In this user can upload text and code document as an input. Given text file or code file will be processed. We are going to perform operation like stemming, stop words removal and parsing technique. We are going to use Levenshtein Distance for code level and TFIDF for text level. Based on the similarity check TF-IDF values will be calculated of words present in already uploaded document. Based on TF-IDF values plagiarism report will be returned to the user in the form of duplicate content and graphical representation.

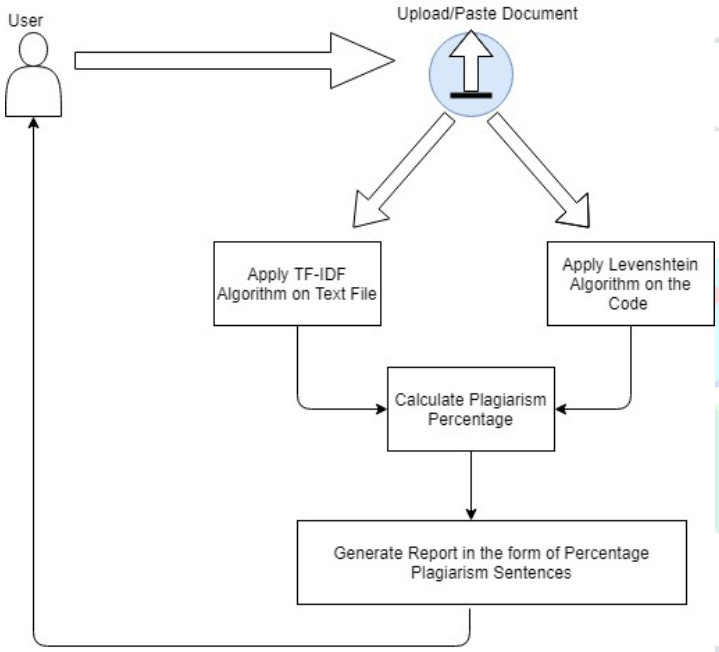


Fig 2. System Architecure

**TFIDF**

1.Term Frequency is used to calculate how frequently word occurs in document:

$$TF(term,document) = \frac{\text{Number of times term appears in doc}}{\text{Total number of words in doc}}$$ .....eq1

2.Inverse Document Frequency is used to calculate how frequently word occurs in all documents:

$$\text{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$ .....eq2

where,
t:terms
D:number of documents
$|\{d \in D: t \in d\}|$ : number of documents where term t appears

3.On equation 1 and 2 calculate overall TFIDF of word,

TF-IDF= TF * IDF …..eq3

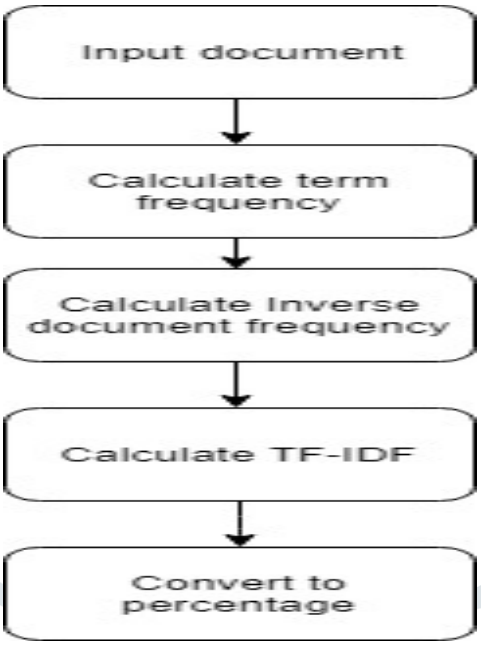


Fig 1. TFIDF Algorithm Working

**LEVENSHTEIN DISTANCE ALGORITHM**

If we select two directories of Java files for checking plagiarism of source directory.Our tool take one Java file of source directory at a time and compare it with all Java files of target directory.This process of picking a Java file from source directory and comparing it with all files of target directory will continue till the end of all files of source directory.

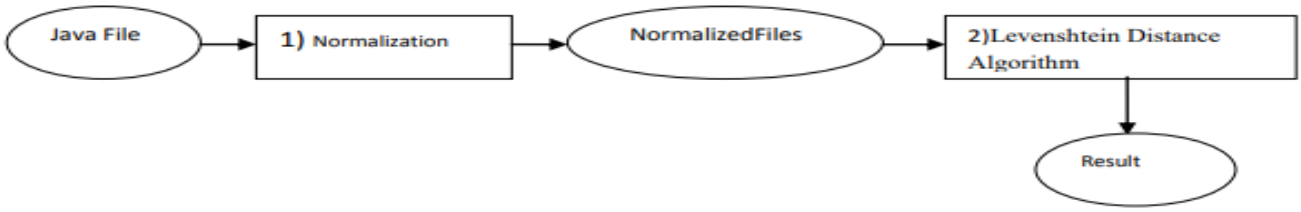


Fig 3.Levenshtein Distance Algorithm Main Idea

**Conclusion:**

In this paper we are describing our preliminary work on Levenshtein Distance and TFIDF and their possible usage for content detection in the task of plagiarism identification. This identification develop by TFIDF technique for text documents and Levenshtein distance algorithm for code document, TFIDF technique finds weight of words after removing stop words present in documents and generate percentage report in the form of file and graph.

## ACKNOWLEDGMENT

It gives us great pleasure in presenting the preliminary project report on '**INTEGRADTED TOOL DEVELOPMENT FOR TEXT AND CODE PLAGIARISM DETECTION USING TF-IDF LEVENSHTEIN DISTANCE ALGORITHM** '

We would like to take this opportunity to thank my internal guide for giving me all the help and guidance we needed. We are really grateful to them for their kind support. Their valuable suggestions were very helpful. We are also grateful to our Head of Information Technology Department, for her indispensable support and suggestions.

Name of Students

[1] **MrunaliDeshmukh** , [2] **KshitijaGhogare**, [3] **BhaktiGunjal**, [4] **ShitalPatil**

.

## REFERENCES

[I] L. Kong, Z. Lu, H. Qi, and Z. Han, "Detecting High Obfuscation Plagiarism: Exploring Multi-Features Fusion via Machine Learning," 1nl. 1. u-and e-Service. Sci. Technol., vo!.7, no. 4, pp. 385-396, 2014.

[2] Izzat Alsmadi1, Ikdam AlHami2 and Saif Kazakzeh3 " **Issues Related to the Detection of Source Code Plagiarism in Students Assignments**" International Journal of Software Engineering and Its Applications Vol.8, No.4 (2014), pp.23-34 http://dx.doi.org/10.14257/ijseia.2014.8.4.03

[3]M. K. Shenoy, K. C. Shet, and U. D. Acharya, "Semantic Plagiarism Detection System Using Ontology Mapping," Adv. Compul.An Inl. 1., vol. 3, no. 3, pp. 59-62, May 2012.

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

[5] S. Harispe, D. Simchez, S. Ranwez, S. Janaqi, and J. Montmain, "A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain," 1. Biomed.In{iirm., vo!. 48, pp. 38-53, Apr 2014.

[6] J. O. Shea, Z. Bandar, K. Crockett, and D. Mclean, "A Comparative Study of Two Short Text Semantic Similarity Measures," Arlij: Inlell.,vo!. 4953, pp. 172-181,2008.

[7] K. Bazdaric, V. Pupovac, L. Bili6-Zulle, and M. Petrovecki, "Plagiarism as a violation of scientific and academic integrity," 2009.

[8] E. S. AI-Shamery and H. Q. Gheni, "Plagiarism Detection using Semantic Analysis," Indian 1.Sci. Technol., vol. 9, no. I, Feb. 2016.

[9] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," 2006, vo!. 6, pp. 775-780.

[10]Agrawal, Mayank, and Dilip Kumar Sharma."A state of art on source code plagiarism detection."*Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on*. IEEE, 2016

[11]Ragkhitwetsagul, Chaiyong, and Jens Krinke."Using compilation/decompilation to enhance clone detection."*Software Clones (IWSC), 2017 IEEE 11th International Workshop on*. IEEE, 2017

[12]Chaddah, Praveen. "Not all plagiarism requires a retraction: papers that plagiarize only text can still contribute to the literature, but any errors or omissions should be prominently corrected, says Praveen Chaddah." Nature 511.7508 (2014): 127-128.

[13]Oberreuter, Gabriel, and Juan D. VeláSquez. "Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style." *Expert Systems with Applications* 40.9 (2013): 3756-3763.

[14] Nilsson, Lars-Erik, Anders Eklöf, and Tina Kullenberg. "Categorizing students, categorizing texts: will plagiarism detection leave blood on the tracks?." *Earli 2017 Biennal conference*. 2017.

[15] Vrbanec, Tedo, and Ana Meštrović. "The struggle with academic plagiarism: Approaches based on semantic similarity." *The 40th Jubilee International ICT Convention–MIPRO 2017*. 2017.

[16]https://en.wikipedia.org/wiki/Plagiarism

[17]https://www.uow.edu.au/~bmartin/pubs/08plagiary.html

[18]http://www.ithenticate.com/resources/6-consequences-of-plagiarism