# CROSS-LINGUAL PLAGIARISM DETECTION USING NLP AND DATA MINING

**Sanyukta Kamble[1], Prof. Madhuri Thorat[2]**

[1]*Student, Information Technology, AISSMS Institute of Information Technology, Pune, Maharashtra, India*
[2]*Professor, Information Technology, AISSMS Institute of Information Technology, Pune, Maharashtra, India*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** *As technology is growing in no time and usage of computer systems is increased as compared to the past, plagiarism could be a phenomenon that's increasing day by day. Wrongful appropriation of somebody else's work is understood as plagiarism. Manually detection of plagiarism is difficult so this process should be automated. There are various tools that may be used for plagiarism detection. Some works on intrinsic plagiarism while other work on extrinsic plagiarism. data processing is that the field which will help in detecting plagiarism yet as can help to enhance the efficiency of the method. Different data processing techniques may be wont to detect plagiarism. Text mining, clustering, bi-gram, trigrams, n-grams are the techniques which will help during this process.*

***Key Words*: Plagiarism, Paraphrasing, Data mining, Text mining, MDR, trigram, n-gram, Clustering, Similarity, Intrinsic plagiarism, Extrinsic plagiarism.**

## 1.INTRODUCTION

In this contemporary time, with the advancement of internet, easy availability of the computers over the world has made it easy to access other's work which ends up in plagiarism. Plagiarism is understood because the act of using some other persons work without the information of author or without giving acknowledge thereto corresponding person with the advancement of technology, use of computers is growing very vastly and it is seen that they're used everywhere in schools, institutes, and industries. More often, assignments of scholars are submitted in electronic forms. As e-form is straightforward and suitable for teachers and students still, but it leads towards the simple opportunity of plagiarism.[8]

With the widespread of data over the world, it's very easy to copy the info from different sources which has internet, papers, books over the internet, newspapers etc. and paste it during a single work without giving any acknowledge to the sources. These actions lead towards lack of learning in students. So, there's a requirement of detecting the plagiarism to extend and improve the education of students.[10] Therefore, plagiarism will be classified into various forms. Some are easily detectable, and a few are complex. several the forms are:

1. Coping pasting: The sort during which one sentence, a full paragraph or a complete page of written communication is copied with none reference. [15]

2. Re-using existed work: Using again the present work or already written e-data

3. Manipulating the text: The kind of plagiarism where text is modified and its appearance it changed

4. Translating the text: When data is translated from one language to a different without giving any reference of the source.

5. Plagiarizing the idea: One the foremost form during which someone else's idea is employed without acknowledging the owner.[6]

6. Incorrect citation: Citation of unread sources and without giving acknowledge to the other sources from where the information has been read.[15]

7. Self-plagiarism: The kind within which author uses his own previously done work and presenting as new one with any reference of prior work.[15]

The plagiarism is difficult to detect manually so it must be automated in order that it can be done efficiently. For this purpose, there are different techniques and ways to implement this for example:

- Algorithms to check documents.
- Crawler to go looking data from the websites

## 1.1. OBJECTIVE

1. To check plagiarism in multiple languages
2. To check with syntactical and semantic approach.
3. To make exceptional changes like diagram and tables for checking plagiarism
4. To detect plagiarism and generate report.
5. To add missing citations or rewrite your text

## 2. LITERATURE SURVEY

Plagiarism is defined as the unauthorized use or stealing of someone's ideas and presenting them as one's own. Data theft and copying have become quite common in recent times. Plagiarism detection of copied data dates to the 1970s, when common literary communication processing (NLP) methods for detecting copied data were introduced in three separate techniques: [14]

- Grammar-based method
- Semantic-based method
- Grammar semantic hybrid method

**Online Assignment Plagiarism Checking Using Data Mining and NLP:** In this paper they proposed a system to detect plagiarism in academic assignments, which will help to prevent students from copying other students' assignments, improve the quality of education, and help students improve their personal skills. Students can also check the grammar from the assignment. The plagiarism detector in this system compares comparable texts and detects plagiarism. In addition, semantical checking will be performed in relation to the assignment.

**The Longest Common Consecutive Word algorithm** The Longest Common Consecutive Word algorithm, proposed in paper [11], analyses the entire paragraph as a single unit and tracks the word positions. Then a word by-word comparison is done to find common terms, which results in the plagiarized version and document similarity.

**MDR (Match Detect Reveal)** MDR (Match Detect Reveal) is a method for checking plagiarism in which the document being tested is first broken into fixed length strings using a suffix tree. For comparison, a string-matching technique is utilized, and the longest common strings can be discovered in the suffix tree. The similarity index and document location can be acquired this way. This strategy is ineffective since it employs exact match terms, resulting in an ambiguous plagiarized text version.[7]

**Methods for Cross-Lingual Plagiarism Detection:** Plagiarism detection in many languages is a difficult issue. It necessitates a thorough understanding of various languages. Finding the right similarity metric for such a strategy is also crucial. These solutions rely on cross-lingual text features to function. (1) cross lingual syntax-based methods, (2) cross-lingual dictionary-based methods, and (3) cross-lingual dictionary-based methods are examples of these methods. In, a detailed survey on cross-linguistic approaches is carried out. The resemblance of two papers is assessed using a statistical model, regardless of the sequence in which the terms appear in the suspect and original documents.[2]

**Grammar Semantics Hybrid Plagiarism Detection Methods:** Because of their use of natural language processing, these methods are successful in detecting plagiarism. They are capable of accurately identifying plagiarism and paraphrase copy/paste. These solutions overcome the drawbacks of semantic based strategies. A semantic-based method cannot usually detect and pinpoint the position of plagiarized content in a document, but a grammar-based method can effectively handle this problem.[2]

## 3. METHDOLOGY

1. **Trigram and clustering method:** The tri-gram values are used to construct a plagiarism detection method that compares the sequences. The electronic assignments are pre-processed and run via a clustering algorithm in this manner. Then a tri-gram analysis is carried out, and similarity results are generated and given as a percentage.[12]

2. **Collecting the data and converting files:** Electronically submitted assignments are divided into three data sets. Because each assignment has a different format, they are all transformed to the same format.[3]

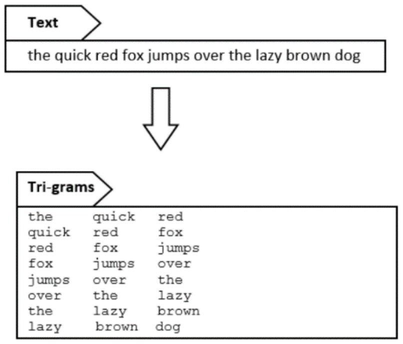
3. **Pre-processing:** It's a crucial step in detecting plagiarism. In this step, data is transformed into a format that may be used in the detection process. The documents submitted are in a variety of forms, including lower- and upper-case letters.[3]

4. **Constructing the trigrams:** Trigrams are three-word sequences that follow each other in a line. They're made once the assignments have been processed.[3]

5. **Measuring the similarity:** The trigram comparing method is used to compare tri-gram structures and the similarity is calculated. The calculated similarity is represented as a percentage. The higher the percentage, the greater the similarity.[3]

6. **Clustering:** The clustering strategy can improve the detecting process' efficiency. The K-means method can be used to do this. The "K means" technique has a lot of advantages for clustering documents (Sharma, Bajpai, Mr., 2012) [3]

7. **Stemming:** This method is used to convert a collection of terms to their root words to see how much this method affects plagiarism efficiency.[3]



Text

the quick red fox jumps over the lazy brown dog

Tri-grams

| | | |
|---|---|---|
| the | quick | red |
| quick | red | fox |
| red | fox | jumps |
| fox | jumps | over |
| jumps | over | the |
| over | the | lazy |
| the | lazy | brown |
| lazy | brown | dog |

**Fig. 1.** tri-gram formation [15]

## 4. PROPOSED METHODOLOGY

Plagiarism may be readily and effectively identified utilizing data mining tools. It is a platform to construct adaptive data driven approach that supports the algorithms for finding patterns, as diverse data mining activities are tradition following, data analyzing technique according to hypothesis. In terms of constructing models or spotting patterns, there are basically two types of data mining techniques. The following methodology is proposed for this purpose:[5]

**1. Collection of assignments:** All assignments and papers will be gathered electronically. So that plagiarism can be easily recognized

**2. Pre-processing:** Pre-processing is a crucial stage in the process of converting all the assignments into a usable format. All the assignments must follow the same format. Numbers, figure values, photographs, and anything else not in the a-z group should be left out of the documents.

**3. Classification:** To extract and split the elements of a sentence into alternative words, text classification should be used. This tool can be used to find key words in a statement.

**4. Text analysis:** The data will next be subjected to a text analysis procedure. This method might be repeated depending on the situation. Furthermore, depending on the nature of the content and the institutes' goals, several text analyzing approaches might be applied.

**5. Processing and analyzing the trigrams:** Every line will have three trigrams, which are three consecutive words. They are made up of a cluster of trigrams derived from a collection of assignments.

**6. Similarity measures:** The sequence of trigrams created from the processed documents is then compared using sequences comparing algorithms later in the process.

**7. Clustering the plagiarized data:** To determine the similarity score, clusters of comparable trigrams are produced. Clusters will assist with calculations and speed up the process.

**8. Similarity score:** The grouping of comparable trigrams will be used to determine the similarity score. The degree of similarity will be computed as a percentage. A high percentage value indicates a high similarity score.

## 5. ALGORITHMS

**1. Rabin-Karp algorithm:** It's a hashing-based search technique that looks for a sub-string pattern in a text. It's great for matching words with several patterns. This function allows you to modify the amount of accuracy. The hash function calculates the feature value of a given syllable fraction. It turns each string into a hash value, which is a number. The Rabin-Karp algorithm uses the same phrase to calculate the hash value.[4]

**2. KMP:** The technique is scanning the text from left to right for a pattern. The shifting in KMP is like that of the brute force technique, but it is done more wisely

**3. SCAM:** A Stanford Copy Analysis Mechanism (SCAM) based on word occurrence frequency is presented in this paper. It primarily keeps track of registered documents that are utilized for copy detection. To compare vectors in the database, a vector of words with their frequency is employed.[9]

**Step 1: Prepossessing:** In this step, the text of input document is isolated from the references mentioned therein. Separating the references from the text can be manually or programmatically.

**Tokenization**

**Step 1:** Declare String array text[], text2[], Declare String line. Initialize Integer C1text, C2text2, sumtext, sum2text to 0

**Step 2:** Set line = in.readLine(); // to fetch line from file

**Step 3:** Do WHILE line is not equal to null set text []=line.split(" "); // split the line based on space increment sumtext; // sum = length of array text[]. ENDWHILE; Set next line by: line = in.readLine(); // fetch the next line Until (end of file). // Now, all the lines are in array text[].

**Step 4:** WHILE C1text < sumtext

Set text2[C2text2] = text[C1text].

replaceAll("[\\W]", ""); //delete delimiters

Increment C1text

ENDWHILE;

C2text2 = sum2text2;

**Step 5:** Print text2[c2text2]

**Stop Words Removing**: Stop words are common in the English language, yet they don't convey any information. These words could be pronouns, conjunctions, or prepositions, among other things. This stage's output is a text that is clear of stop words and initially lowercase's all letters. Words that have been removed from the text should not be used again. The pre-processing stage produces a text that is ready for semantic plagiarism detection.

**Step 2: Document Disciplinary:** Before detecting semantic plagiarism, a procedure of identifying the document's specialist is carried out. Plagiarism will be detected only for documents that fall under the specialty of computer science, while documents from other disciplines will not be detected.

**Word Frequency:** The occurrence of each word in the input document will be computed after the pre-processing stage, based on how many times it appears in the document.

i. **Descending Order:** Frequencies that found in the previous step will arranged in descending order.

ii. **N specification:** N was determined within the program at this point, and it will be taken from the total number of it to represent the highest frequencies.

iii. **Word (N):** The words with the highest (N) frequencies will be shifted to one side to reveal the relationship between the original document and the computer science domains.

iv. **Decision Making:** Finally, the fields that are related to this document will be displayed and continue working.

**Step 3: Semantic Plagiarism Detection:** Then, to help detecting semantic plagiarism, we propose to use semantic similarity between documents based on information extracting techniques. Semantic plagiarism will be detect based on WordNet.

i. **Text:** The text will be taken again to complement this work if the document passed the threshold of a computer science specialisation test.

ii. **WordNet:** To determine the amount of the semantic plagiarism, use WordNet to find synonyms for each word after taking the text supplied in the previous step. Every word in the specified text will have its synonyms retrieved. As a result, while detecting plagiarism, synonyms will be treated as the appearance of the word itself.

iii. **WordNet Expansion:** At this stage, WordNet expansion has been proposed by specific words doesn't exist in its dictionary

iv. **Documents of Database:** At this point, the documents stored in the database will be withdrawn one by one, and the text for these documents will be taken in its entirety rather than just a specific text in it, to eliminate the possibility of plagiarising a text that exists in different places of the database document and putting it in another place of the source document, these places could be abstract, results.

### 4. Cross-Lingual Plagiarism Detection

The monolingual approach is the basic concept of the CrossLang1 system. We have a collection of questionable Russian documents and English references. Because the reference collection is in English, we simplify the work by translating the suspicious document into English. We proceed to the next step, which is document analysis. As a result, the fundamental problem with the Cross Lang design is that the algorithms must be stable in the face of ambiguous translations. Figure 1 depicts the major stages of the Cross Lang service. When a user sends a document for originality checking, Cross Lang receives it from the Antiplagiat system. The data is then sent between the following phases via the Entry point — primary service:[16]

i. **Machine Translation System (MTS) -**

is a microservice that converts questionable documents to English. Transformer Vaswani et al., an open-source neural machine translation framework, is used for these objectives.[16]

ii. **Source retrieval –**

Two microservices are combined at this stage: Shingle index and Document storage. Shingle index returns the document ids from the reference English collection to the entry point, which receives the translated suspicious document's shingles (n -grams). We employ a modified shingle-based technique to cope with the translation ambiguity. By using these ids, document storage returns the Source texts from the collection.[16]

iii. **Document comparison –**

This microservice compares a suspicious document's translation to its source documents. The vectors corresponding to the sentences in these texts are compared, not the texts themselves. As a result, we address the issue of ambiguous translation.[16]

## 6. ARCHITECTURE

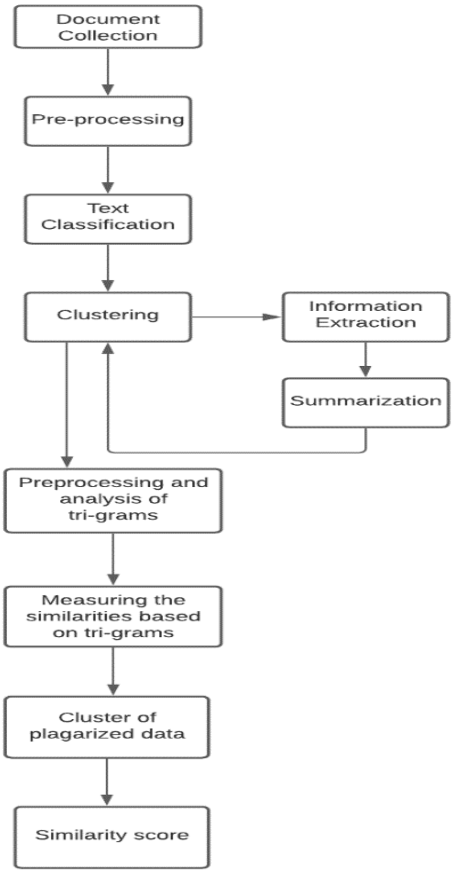


**Fig. 2.** Proposed Methodology [5]

## 7. CONCLUSIONS

To be effective, the plagiarism detection method should be automated. Data mining techniques can be utilized to improve the plagiarism detection process.

In this study, a methodology based on data mining techniques is proposed, with the goal of increasing the efficiency of the process. To reduce the overhead of the operation, pre-processing and clustering techniques might be applied. Furthermore, a similarity score can be derived using clusters of plagiarized data to increase efficiency.

Plagiarism detection is critical for ensuring the integrity of written output. All institutions and teachers, it is determined, should be aware of plagiarism and antiplagiarism software. We've devised a simple way for detecting instances of plagiarism in school and college assignments. Our method is simple to adapt to the wide range of programming languages in use, and it is robust enough to be very useful in an educational setting

## REFERENCES

[1] I Widaningrum et al. "Evaluation of the accuracy of winnowing, Rabin karp and Knuth Morris pratt algorithms in plagiarism detection applications". In: Journal of Physics: Conference Series. Vol. 1517. 1. IOP Publishing. 2020, p. 012093.

[2] Hussain A Chowdhury and Dhruba K Bhattacharyya. "Plagiarism: Taxonomy, tools and detection techniques". In: arXiv preprint arXiv:1801.06323 (2018).

[3] Mahwish Abid, Muhammad Usman, and Muhammad Waleed Ashraf. "Plagiarism detection process using data mining techniques". In: International Journal of Recent Contributions from Engineering, Science & IT (iJES) 5.4 (2017), pp. 68–75.

[4] Andysah Putera Utama Siahaan et al. "K-Gram as a determinant of plagiarism level in Rabin-Karp algorithm". In: (2017).

[5] Mohamed Abdul Cader Jiffriya. "Plagiarism Detection On Electronic Submissions Of Text Based Assignments". PhD thesis. 2013.

[6] Ali El-Matarawy, Mohammad El-Ramly, and Reem Bahgat. "Plagiarism detection using sequential pattern mining". In: International Journal of Applied Information Systems 5.2 (2013), pp. 24–29.

[7] Narendra Sharma, Aman Bajpai, and Mr Ratnesh Litoriya. "Comparison the various clustering algorithms of weka tools". In: facilities 4.7 (2012), pp. 78–80.

[8] Salha M Alzahrani, Naomie Salim, and Ajith Abraham. "Understanding plagiarism linguistic patterns, textual features, and detection methods". In: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42.2 (2011), pp. 133–149.

[9] Daniele Anzelmi et al. "Plagiarism detection based on SCAM algorithm". In: Proceedings of the International MultiConference on Engineers and Computer Scientists. Vol. 1. Citeseer. 2011, pp. 272–277.

[10] Alberto Barr´on-Cede˜no and Paolo Rosso. "On automatic plagiarism detection based on n-grams comparison". In: European conference on information retrieval. Springer. 2009, pp. 696–700.

[11] Agung Sediyono and Ku Ruhana Ku-Mahmud. "Algorithm of the longest commonly consecutive word for plagiarism detection in text-based, document". In: 2008 Third International Conference on Digital Information Management. IEEE. 2008, pp. 253–259.

[12] Tao Wang, Xiao-Zhong Fan, and Jie Liu. "Plagiarism detection in Chinese based on chunk and paragraph weight". In: 2008 International Conference on Machine Learning and Cybernetics. Vol. 5. IEEE. 2008, pp. 2574–2579.

[13] Miguel Roig. Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing. 2006.

[14] Jun-Peng Bao et al. "A survey on natural language text copy detection". In: Journal of software 14.10 (2003), pp. 1753–1760.

[15] Paul Clough. "Plagiarism in natural and programming languages: an overview of current tools and technologies, Department of Computer Science, University of Sheffield". In: URL http://ir. shef. ac. uk/cloughie/papers/plagiarism2000.pdf(2000)

[16] Einstein, A., B. Podolsky, and N. Rosen, 1935, "Can quantum-mechanical description of physical reality be considered complete?", Phys. Rev. 47, 777-780.

## BIOGRAPHIES

Sanyukta Kamble
Third Year IT
AISSMS Institute of Information Technology

Prof. Madhuri Thorat
She is a professor at Department of Information Technology Department at AISSMS Institute of Information Technology, Pune