CHAPTER 1: INTRODUCTION

1.1. INTRODUCTION TO PLAGIARISM

The oxford dictionary defines plagiarism as “The practice of taking someone else’s work or ideas and passing them off as one’s own”. The word “plagiarism” originated in early 17th century from the Latin word “plagiarius” which means kidnapper [1]. The problem of plagiarism has increased to a great extent due to increased use of digital media for storage, retrieval and communication of the information. Earlier, the plagiarism was restricted to close geographical locations, but with the advent of Internet and Communication Technologies, it has spread across the length and breadth of the world. Plagiarism may be intentional when the author makes a copy of someone else’s work intentionally, or it may be un-intentional when the author does not know the rules and regulations for research publications and fails to give credits/citations to the original author. Copying the work published by one self at some other place without giving a reference to it in the current work is known as self-plagiarism. Plagiarism is one of the most serious problem that researchers are facing these days. While the plagiarism detection methods keep on improving day by day, the young programmers, authors and researchers are continuously finding ways to escape from the plagiarism detection software.

1.2. TYPES OF PLAGIARISM

The researchers have identified a number of parameters to classify the types of plagiarisms [2]. Apart from being Intentional or unintentional, the extent of plagiarism may vary from merely copying a few words to copying full sentences and paragraphs or even copying the entire ideas. If not all, the Plagiarism detection software has to address most of these type of plagiarisms. Based on the defined parameters, the plagiarism can be categorized into the following broad types –
1.2.1 Copy & Paste Plagiarism

Picking sentences or phrases or paragraphs from the original work done by oneself or by someone else and not giving any reference/due credit to the original author is known as copy and paste plagiarism. It is the simplest form of plagiarism and is very easy to detect using the inspection method or by using the available automated tools. [3] The Copy and Paste plagiarism is most widely spread form of plagiarism and is usually found to be practiced by the youngsters and fresh entrants in the research domain. The highlighted text in the diagram below depicts an example of Copy & Paste Plagiarism in a news article –

1.2.2. Word Switch Plagiarism

The copy paste plagiarism is easy to detect. Any simple string matching algorithm can be used to predict this type of plagiarism. However the smarter researchers have found alternative ways to escape from these simple plagiarism detection tools. The technique called word switching can easily fool these simple tools. Picking the sentences or phrases or paragraphs from the original work done by oneself or by someone else, making a few changes in the sentences by changing the order of the words to hide the originality and bypass string matching tools for plagiarism
detection and not giving due credits to the author is called word switch plagiarism. The highlighted text below shows an example of word switch plagiarism –

**Figure 2. Word Switch Plagiarism**

### 1.2.3. Idea Theft

Picking the idea from someone else’s work without giving a due credit to the original source of idea is known as idea theft. This type of plagiarism is hard to detect as the ideas can be presented using different words, keeping the meaning of the text same. The example below shows the idea theft –

**Figure 3. Idea Theft by rewriting the passage**

In this given example the author is talking about some eye disease common in the children and discusses how to handle the patients and gives a remedial solution for it. The second author copies the idea of the first author and rewrites the Punjabi text as shown in the box written on the right. This is an example of the idea theft plagiarism.
1.2.4. Missing Citations

Most of the research work done happens to be extension of some previous work done by other researchers. As per the research ethics, whenever the authors need to use the research work done by someone else, they should give due credit to the original authors by mentioning their publications in the references section. This technique is known as citation. Sometimes, the authors genuinely mention the previous work so as to co-relate it to their own work, but fail to give citations to the original author. This makes their text liable to be treated as plagiarized text. Due to this negligence, the original contributors to the topic suffer. Sometimes this happens unintentionally due to lack of knowledge, however, this can also happen intentionally as the researcher may be reluctant to mention this information in his research work. The lack of knowledge however should not be an excuse to escape from plagiarism, proper guidance should be given to the authors so that they must mention each and every author who has originally contributed to the work.

1.2.5. Invalid citations

Citations play a major role in the evaluation of any research paper. The citations mention the sources of statements written in any research paper which are derived from other research papers. Many evaluators do take notice of the number of citations before starting the evaluation of the research work. This gives an idea of how much literature survey has been done in writing the paper under review. Sometimes, the authors fail to perform the requisite literature survey before starting with their own research work and in order to support their concepts, they provide fake or invalid citations to the papers that actually do not exist or have no connection with the topic under consideration. This unlawful activity falls under the purview of invalid citations. The invalid citations on one end waste the time of evaluators and misguide the readers of the research paper, on the other side, many research papers get undue citations to their research that inadvertently increases their citation index value.
1.2.6. Self-Plagiarism

The authors can use their original research work in more than one research papers to extend the previously done work. Although the work is the original workmanship of the author, but copying one’s own work and present it as something new, without giving any reference to the previous work is known as self-plagiarism. If in case, the authors need to refer to any work done by them in some other research paper, they should treat that research paper just like any other research paper and provide valid citations to it. This will also increase the citation index of the previously done and duly cited research work.

1.2.7. Dual Submission

Sometimes, the authors submit the same research work to more than one journals for evaluation and publication, and fail to retract the paper from one of the journals, thereby getting their same research work published twice under different index numbers. Submitting the same work to two or more journals is also a type of plagiarism. Due tracking should be done by the researchers about the research work submitted to multiple journals.

1.2.8. Translation

Translation is one of the most popular and dangerous forms of plagiarism, where the author translates the original work done by someone else into some other language and presents as his own work, without giving a reference to the original author. This type of plagiarism is very hard to detect. The example below lists such a case, where the same paragraph written in Punjabi Language is translated into English language. All the words of first passage have been changed to words in other language. The simple string matching algorithms will simply tell these two texts apart. Thus, for plagiarism detection software to detect such cases, good language translation tools need to be developed, which in itself is a major task.
1.2.9. Plagiarism of secondary sources

Martin [4] has mentioned an indirect category of plagiarism in which the author reads some quoted text from a research paper and cites the original author of the quoted text, without mentioning the secondary source from where he actually read the quoted text. This type of plagiarism is debatable and very hard to judge.

A number of other categories of plagiarism have also been identified by [5] and [6] in their research publications.

1.3. REASONS BEHIND THE PREVAILING PLAGIARISM

Plagiarism exists and flourishes due to many reasons. The researchers investigated various reasons as to why plagiarism is prevalent amongst the scholars and have found the following reasons for plagiarism –

1.3.1. Lack of Knowledge

Most of the time, lack of knowledge is the main reason behind the plagiarism. The scholars are not well versant with the topic in which they are carrying their research. So they try to copy the text from related articles to complete their research writing. This tendency of copying text and not developing innovative ideas leads to the deterioration of the quality of the research work.
1.3.2. Lack of Interest

The scholars might not feel interest in the topics assigned to them or the topic appears to be boring to them, thereby provoking them to copy the text from other sources and present as their own assignment.

1.3.3. Lack of time

The strict timelines sometimes provoke the scholars to copy the text from other sources. The scholars fail to manage their time and when the deadlines are approaching, they try to complete their assignments by copying the ideas or text from the established researches, without giving due credit to the original author or without citing the referred paper.

1.3.4. Poor writing skills

The scholars might have novel ideas with them but they might be having poor technical writing skills. Sometimes, the scholars are not well versed in English language. So they may tend to copy the text from other sources in a zeal to get their work published.

Besides this, there are a number of reasons like lack of awareness about plagiarism, lack of competent plagiarism prevention policies, lack of plagiarism detection tools etc. which exaggerate the situation as well.

1.4. DEALING WITH PLAGIARISM

Plagiarism now a days is treated as a social evil in researchers. With the advent of internet, it has spread worldwide. This social evil has to be dealt with very strictly, otherwise the society will have to bear its bad or even worse consequences. The Researchers have identified two prominent methods to deal with plagiarism. These methods are Plagiarism prevention and plagiarism detection.
1.4.1. Plagiarism prevention

The plagiarism can be dealt with providing information to the students and researchers. The teachers can adopt strict policy regarding plagiarism such as giving negative marks to the plagiarized text, lowering the grades of students or even suspension from the class for a specific term. The institutes can frame strict policies that may impose heavy penalties on students, expulsion from the university or even debarring the student from taking any course in any other university for a certain term.

The researchers need to be guided about the plagiarism at the time of their enrollments. The universities should issue strict guidelines against the plagiarism in the form of print material and publish anti-plagiarism policies on their websites as well. The stringent actions should be taken against the defaulters. Also the universities should conduct awareness workshops from time to time and encourage the students to do quality research work. University of Bradford in 2008 [7] made a change in their strategies and regulations in handling plagiarism in scholarly articles. The researchers found that there were more number of cases due to ignorance of plagiarism and referencing and not because of intentionally copying the text. The university started with “Plagiarism Awareness Programme (PAP)” for students getting enrolled in various courses and a significant reduction in number of plagiarism cases was observed over a period of four years. The plagiarism prevention methods have slow convergence, but have deeper impact in the long run [3].

1.4.2. Plagiarism Detection

Due to the slow convergence of plagiarism prevention methods, the immediate relief can be sought of by using the plagiarism detection. The researchers have proposed different techniques for detecting different types of plagiarisms. The most commonly used plagiarism detection techniques can be summarized as follows:

1.4.2.1. Manual Inspection Method. In this method, the evaluator picks up some keywords and key phrases from the suspect document and searches for it on Internet using some
popular search engine. The results obtained can be manually inspected for finding any copied text. This method is time consuming and can only be used to detect the copy-paste type of plagiarism. The results of this method are also dependent on the choice of search engine as well as on the ability of the evaluator to do the word by word comparison. This method is suitable for preliminary inspections only and the detailed comparisons should be done only using the automated tools.

1.4.2.2. Automated Tools. The plagiarism can occur in written text, in source code of computer programs or even in images. A number of commercial as well as free plagiarism detection tools have been developed till date. These tools are capable of detecting various forms of plagiarism ranging from simple copy-paste plagiarism to word switch, sentence and paragraph rephrasing etc. There are a number of desktop based as well as web based plagiarism detection tools now a days. These tools are available in offline as well as online modes. Majority of these tools do the source comparison based on ASCII data only, however the most modern tools now use UNICODE data for document comparison. Some of the existing tools have their local repositories built up and perform local comparisons only, whereas others use internet sources for finding the matching documents. Some tools perform the comparisons on the fly, whereas others accept the user documents as jobs submitted to the engines via online forms or via emails and do the processing in the background, thereby intimating the results via email to the users. More mature systems like Turnitin and Urkund also allow the management of instructor and student accounts for the job submission and analysis respectively. Lancaster & Culwin [8] have suggested various metrics based on the number of documents processed at one time. These metrics are –

- **Singular Metrics** – The Singular metrics process single document at a time, measuring certain attributes within the document, such as mean number of words per sentence, proportion of words what are synonyms to each other etc.
- **Paired Metrics** – The paired metrics work on two documents simultaneously, measuring the usage of Capital letter words in two documents, measuring the longest occurrence of common words in two sentences etc.
- **Multi-dimensional Metrics** – The Multi-dimensional metrics work on chosen group of documents, finding the occurrence of group of words in the chosen documents.
• **Corpal Metrics** – The Corpal Metrics work on entire corpus at a time, for example for finding some general property of the text.

Based on these metrics, various plagiarism detection tools and techniques have been designed as per the classifications below –

• **Superficial Plagiarism Detection** – In this technique, no knowledge of linguistics is necessary. The document structure is also not significant. The software such as CopyCatch, DetectaCopius, EVE2, WordCheck use this method for plagiarism detection.

• **Structural Plagiarism Detection** – This method requires the knowledge of the underlying language. The document is first reduced to a smaller size by replacing the words into their root forms and then similarity calculation is done. The software such as CopyFind, EduTie, Turnitin use this method for plagiarism detection.

• **Source Code plagiarism Detection** – This technique is used for detecting plagiarism in source code of various programming languages. The software needs to know the structure of the program and identify its tokens and functions, based on which the similarity detection is done. The method also counts the number of variables, functions, function calls etc. to predict the similarity. jPlag uses this technique for plagiarism detection in source code.

• **String Matching Based Plagiarism Detection** – String matching is the simplest technique of all the plagiarism detection techniques. In this technique, a word by word comparison is made between the query document and the suspected sources. The different algorithms have been proposed by researchers for optimizing the string matching. These algorithms have different complexities associated with them. The algorithms are language independent, however, the implementations which use Unicode based characters for comparison are successful, whereas the others fail to do a string comparison on Unicode characters. For plagiarism detection in multiple languages, the Unicode support is a must. Even in case of multi-lingual documents, if the string matching algorithm is not Unicode aware, then only non Unicod text will be compared, thereby skipping all non Unicode suupoted words. The table below shows
various string matching algorithms along with their complexities in terms of processing time and string matching time.

### Table 1. String Matching Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Preprocessing Time</th>
<th>Matching Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve string search algorithm</td>
<td>0 (no preprocessing)</td>
<td>$\Theta(nm)$</td>
</tr>
<tr>
<td>Rabin–Karp string search algorithm</td>
<td>$\Theta(m)$</td>
<td>average $\Theta(n + m)$, worst $\Theta((n−m)m)$</td>
</tr>
<tr>
<td>Finite-state automaton based search</td>
<td>$\Theta(mk)$</td>
<td>$\Theta(n)$</td>
</tr>
<tr>
<td>Knuth–Morris–Pratt algorithm</td>
<td>$\Theta(m)$</td>
<td>$\Theta(n)$</td>
</tr>
<tr>
<td>Boyer–Moore string search algorithm</td>
<td>$\Theta(m + k)$</td>
<td>best $\Omega(n/m)$, worst $O(mn)$</td>
</tr>
<tr>
<td>Bitap algorithm (shift-or, shift-and, Baeza–Yates–Gonnet)</td>
<td>$\Theta(m + k)$</td>
<td>$O(mn)$</td>
</tr>
<tr>
<td>Two-way string-matching algorithm</td>
<td>$\Theta(m)$</td>
<td>$O(n+m)$</td>
</tr>
<tr>
<td>BNDM (Backward Non-Deterministic Dawg Matching)</td>
<td>$O(m)$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>BOM (Backward Oracle Matching)</td>
<td>$O(m)$</td>
<td>$O(n)$</td>
</tr>
</tbody>
</table>

- **Naïve string search algorithm** – It is the simplest string matching algorithm. In this algorithm, the source string (called needle) is to be matched within the destination string (called haystack). The first character of the needle is checked against the first character of haystack. If a match is not found, then the second character of haystack is checked for a possible match and so on. On an average, the algorithm takes $O(m+n)$ steps for the comparison, where $n$ is the length of needle and $m$ represents the length of haystack. In worst case, it takes $O(mn)$ steps for comparison.

source as well as query documents is represented by a unique hash value. The hash values are calculated using some mathematical functions. These hash values are then used to compare the two strings. If the query text has length \( n \) and there are \( p \) patterns of combined length \( m \) from which the match is to be found, then its average and best case running time are \( O(n+m) \) in space \( O(p) \), and the worst-case time is \( O(nm) \).

- **Finite State Automata Based Search** – Finite state automata is best suitable for searching regular expression based patterns. In this approach, Deterministic finite automaton (DFA) is used for recognizing the search string, avoiding backtracking hassles.

- **Knuth–Morris–Pratt String Matching algorithm** – Knuth–Morris–Pratt [10] algorithm is faster as compared to the simple string matching algorithm. In simple string matching algorithm in worst case, the performance is \( O(k-n) \), whereas the KMP spends a little time precomputing a table (on the order of the size of search word, \( O(n) \)), and then it uses that table to do an efficient search of the string in \( O(k) \).

- **Boyer–Moore string search algorithm** - Boyer–Moore [11] string search algorithm has been very efficient in cases where the string being searched is much shorter than the text in which it is to be searched for. That is why this has been considered as a benchmark while evaluating the pattern matching algorithms. It is also suitable for finding multiple occurrences of strings in two or more documents. In this algorithm, the pattern being searched is pre-processed instead of the text in which the pattern is to be found. Generally, the more is the length of the pattern, the better the algorithm performs.

- **Bitap Algorithm** – This algorithm is also known as shift-or, shift-and, Baeza-Yates–Gonnet algorithm [12]. It is an approximate string matching algorithm based on Levenshtein distance. The search pattern and the source text are considered equivalent if they are calculated to be \( k \) distance apart. In preprocessing task, the algorithm calculates a set of bitmasks containing one bit for each element of the pattern. The remaining matching operations are carried by using bitwise operations, which are extremely fast.
- **Intra-Corpal Plagirism Detection** – Detects the plagiarism with the source as well as copy being present inside the corpus. Most of the plagiarism detection tools support this feature.

- **Extra-Corpal Plagiarism Detection** – Detects the plagiarism with source present outside the corpus and the copy present inside the corpus. EduTie, EVE2 and Turnitin support this feature.

- **Local Similarity Detection** – This technique reduces the document to smaller parts called segments and a unique integer (minuta) is generated for every distinct segment by using some mathematical functions. Each minuta is compared against the local fingerprints stored in the repository. The matching minutae are retrieved and counted to predict the similarity score between the two documents.

- **n-gram based Plagiarism Detection** – An n-gram is a substring of n consecutive words from a single string. The popular forms of n-gram include Uni-gram (1 word), Bi-gram (2 words), Tri-gram (3 words), Four-gram (4 words) and Five-gram (5 words). The sentence is broken into the chosen size of ngrams. A fingerprinting can be done for the generated ngrams or even the direct ngrams can be used in string matching algorithms to detect the similarity.

- **Global Similarity Detection** – This technique deals with the larger chunks of the document and considers the whole document for the calculation of the similarity score. It also performs citation pattern analysis and stylometry for the calculation of the final score.

- **Citation Pattern Analysis** – This technique is suitable for detecting plagiarism in scholarly articles. The Citation order from the source document is collected and a fingerprint is generated, which can be matched with the Citation order of other documents for detecting the similarity. [13]

- **Stylometry for Plagiarism Detection** – The stylometry deals with the analysis of writing styles of the authors. Writing styles of different authors vary from small to greater extents and however, the style of single author should remain consistent within the document. The stylometry analyses the document paragraph by paragraph and generates various statistics for the different paragraphs e.g. percentage of nouns, pronouns, verbs, adjectives etc. For single author, the statistics should remain
consistent throughout the document, allowing variations up to a certain degree only. If there are large variations within the document, then there are more chances of the document being plagiarized.

1.5. PLAGIARISM CASE STUDIES

The plagiarism has been around for many years and especially with the increasing internet connectivity across the various parts of the world, the situation is going even worse. A number of plagiarism cases are being reported each day. In some cases, the necessary actions are being taken, whereas many other cases go un-reported. The news extracts below reveal a few such plagiarism cases reported across the different parts of the world –

- “Three scientists caught in plagiarism row, top publisher retracts article after 14 years” [14]

“A top international scientific publication house has retracted an article by three Indian scientists on plagiarism charges 14 years after its publication. The article was withdrawn after an inquiry, with the company issuing a retraction note on its website on July 7.

The article ‘The Nitrile — degrading enzymes: Current Status and future prospects’ by Prof A Banerjee, Prof Rohit Sharma and Prof U C Banerjee was published in Springer’s Applied Microbiology and Biotechnology journal in October 2002.”

The three authors are scientists working in the field of biotechnology. At the time of the article’s publication, all three were affiliated with the Mohali-based National Institute of Pharmaceutical Education and Research (NIPER).

The note on Springer’s website said the article had been retracted “at the request of the Editor-in-Chief, as it contains portions of other authors’ writings on the same topic in other publications, without sufficient attribution to these earlier works being given. The principal authors of the paper acknowledged that text from background sources was mistakenly used in this article without proper reference to the original source”.
• “Journal retracts Pondy varsity VC's article, to send findings to President” [15]. The editor of International Journal of Legal Information (IJLI) confirmed the plagiarism case in an article written by Chandra Krishnamurthy – “Legal education and legal profession in India” published in their journal in 2008. The article was retracted from the journal after the decision of board of directors on the case and no communications /clarifications were received from the author on the said article.

• The Indian Journal of Dermatology banned a group of Tunisian researchers from publishing in the journal after the authors were found to have plagiarized [16]. The article titled “Dermatology Life Quality Index scores in vitiligo: Reliability and validity of the Tunisian version”, written by Jalel A, Soumaya GS, Hamdaoui MH, published in Indian Journal of Dermatology 2009; 54 (4): 330–333, was found to be plagiarized. The article was retracted and the authors were barred from submitting their manuscript(s) to IJD for the next 5 years. The retraction information was also posted on the journal website –

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2807707/

• In a recent article published on Retraction watch [17], a total of 58 papers were retracted for plagiarism and fake reviews issues by Springer, BMC. The papers were published in seven journals. The springer and BioMed Central were investigating several other journals for the possible cases of plagiarism. The retraction notices were published on their respective journal websites.

The cases above revealed well that the plagiarism cases existed since long and their number is increasing day by day. Many plagiarized articles remained un-detected for several years and had got significant number of citations as well till someone pointed out for the suspected plagiarism and the plagiarism was confirmed by using some well known tools and techniques (mostly Turnitin software) as cited in the “Retraction Watch” website. The defaulters faced several consequences ranging from meager retraction of the paper to a ban on further publications in the same journal or even denial of fellowships and expulsion from the university or their present jobs. The harsh punishments given to the authors who did plagiarism did set examples for the other fellow researchers to refrain them from this ill-practice.
1.6. EXISTING PLAGIARISM DETECTION TOOLS

A number of plagiarism detection tools are available till date. A brief introduction to these tools is given in the section below -

- **Turnitin** [18] provided by iParadigms, LLC is a web based plagiarism detection tool. The Turnitin website accepts the user documents via web uploads and plagiarism detection is done remotely. A complete fingerprint of the document is created and stored in the remote repository. The system offers the document comparisons using locally archived and indexed web pages, journals and books available from ProQuest™ database and millions of documents already submitted their database. The customization of reports is also possible, where the instructors can omit certain sources from comparison while calculating the similarity scores.

- **Urkund** [19], is an online tool that accepts the document uploads to the remote server and does the background job for document comparisons. This tool is also instructor-student based that allows the self-signup of the students and links the students accounts to their instructor accounts via instructor’s email ID. The tools is Unicode based and thus can detect the Punjabi text.

- **JPlag** [20] is an online tool for detecting plagiarism in software programs. Developed by Guido Malpohl as a student research project, the software gradually evolved into a full fledge plagiarism detection engine for software programs. The software works only on C,C++,Java and natural language text. The tools does not support internet search and is primarily used for checking the student assignments.

- **WCopyfind** [21] is a windows based tool used to find plagiarism based on words and sentences. It performs string matching only on local documents. The tool provides several customizations such as ignoring numbers & punctuation marks, selecting minimum match length, case sensitive search etc. The output in the side by side comparison format is stored in local storage in the form of HTML pages.

- **Moss** [22] is an online tool for detecting software plagiarism in various programming languages viz C/C++/Java/Python and many other popular programming languages. It is specially used for evaluating the student assignments in a class. The registration to
Moss is free for non-commercial use and is available via email. The Moss accepts the source program and suspected copies via a script that can be uploaded to the Moss server. It uses some proprietary algorithms that counts the number of tokens, lines, function calls etc and checks their positions in the program. The copied programs have a tendency of replacing variable names, however, the program constructs and their line numbers remain the same. The Moss uses this information to predict the level of plagiarism in the query document.

- **Glatt** [23] is an online plagiarism assessment program. Glatt is based on the principal that every person has his own writing style (unique linguistics pattern). The approach here is based on Wilson Taylor's cloze procedure [24]. The Glatt screens the writing style of the authors by taking every fifth word and replacing it with a blank space in the query document. The author of the script is then asked to fill in the blank spaces. The count of correct and wrong responses along with the time taken to fill the blanks is used for plagiarism detection.

- **EVE2** [25] is plagiarism assessment tool specially designed for students and teachers. The tool accepts the query documents online in various formats such as MS-Word, Corel Word-perfect and Plain Text etc. and compares it with various sources over the internet. It uses complex algorithms to search for the matching text over the internet and records the URLs where a suspected match is found. At the end of the search, a comprehensive report is given to the teacher indicating the various sources of plagiarism along with the percentage match with each source. The copied text in query document is also highlighted and annotated indicating the sources of copy.

- **DocCop** [26] is an online plagiarism assessment tool that accepts the paragraphs in plain text as well as MS-Word and Acrobat PDF format. The Doccop uses Google search engine to find the pages having text similar to that provided in the query document. The search ignores the Google Books and Research databases. Once the search is complete, a comprehensive email is sent along with the links of the websites from where the similarity has been detected. It provides a side by side comparison of the documents by highlighting the sentences similar in both the documents.

- **Plagiarisma** [27] is an online duplicate finder program that receives the documents in paragraph format. Plagiarisma checks for the similarity using various search engines
viz. Google, Yahoo, Babylon, Google Scholar and Google Books. Plagiarisma does not archive the user documents on their portal. The software claims to be compatible with various desktop and mobile platforms.

- **Dupli Checker** [28] is an online software for plagiarism detection. It requires the users to copy–paste the text on their web interface. The search text is limited to 1000 words per submission. The registration is free for commercial as well as non-commercial use. The unregistered version however limits only one submission in a day. The software is suitable for checking the duplicate versions of the web pages on the internet.

- **Viper** [29] is a free plagiarism assessment tool that can be download on the local system. It claims to scan the documents over two billion sources. Besides internet based comparison, It also provides comparisons on local sources (User folders). It also provides side by side comparison of the similar documents.

### 1.7. PLAGIARISM DETECTION IN PUNJABI TEXT

India is a country that has a heritage of 415 living languages, with 23 constitutionally recognized official languages, having more than 57 Million people who speak Punjabi language [30]. The Punjabi language is morphologically rich and is entirely different from English language. Although, prior to the prevalence of Unicode encodings, the Punjabi text if any in electronic form was written using ASCII encodings, but most of this electronic text was used for printing purposes and is not available in online form. The online plagiarism detection tools that accept ASCII encoding of data absolutely fail to report the plagiarism in Punjabi text documents, however, the Unicode supporting tools provide only basic functionality for non-English text, thereby performing “String Matching only” for plagiarism detection. The plagiarism detection methods proposed by researchers involve a variety of sub-tasks to be performed beginning with repository building for local comparisons, document procurement, document text preprocessing, document indexing, keywords identification and storage, similarity detection to final calculation of the degree of plagiarism in the query document. These sub-tasks are dependent on the language
specific resources and techniques, and that is the main reason why the prevalent tools do not perform well when confronted with Punjabi Text.

1.8. OBJECTIVES

1. To develop a plagiarism detection software for documents in Punjabi Language that is capable of detecting copy-paste, word switch, paraphrasing and synonym replacement in the query documents.
2. To build a tool for converting Non-Unicode text to Unicode format.
3. To build a database of reference documents for testing purpose.
4. To device a mechanism for automatic indexing of documents based on important keywords.
5. Develop a stemmer / lemmatizer for Punjabi language that can be used in pre-processing phase of the plagiarism detection.
6. Develop a tool for building thesaurus database for Punjabi language that can be used for detecting synonym replacements in the query document.
7. To find the suitable technique for detecting the similarity score amongst the Punjabi documents.
8. To develop the tool that is user friendly and easy to operate.
9. To generate PDF as well as interactive HTML based reports for the convenience of users.

1.9. CHALLENGES IDENTIFIED

Plagiarism detection in itself is a big challenge for the programmers. As earlier stated, the Punjabi language is morphologically rich and there is a very limited Punjabi corpus available in digitized form. Due to this, very less or negligible work has been done in the Information Retrieval (IR) field on Punjabi corpus. Existence of more than one keyboard
layouts and the lack of Unicode text further exaggerates the problem. Some of the important challenges identified in plagiarism detection task are mentioned as below:

1. **Diversity in Punjabi fonts in input documents**
   There are a number of non-standard fonts available for Punjabi language. These fonts use different keyboard layouts (Phonetic and Remington) with different character mappings. Since most of these fonts use ASCII / ISCII encoding schemes, so these Non-Unicode fonts need to be converted to Unicode format before starting with the plagiarism detection process.

2. **Lack of Punjabi literature / source documents/Online literature**
   A limited amount of Punjabi literature is available in online/digital form. Particularly, the availability of scholarly articles in Punjabi language in digital form is almost nil. Even if some digitized books are available, but most of those books used non-unicode fonts which makes them unavailable for the direct use in plagiarism detection process. Although the testing of the tool can be done using the articles from newspaper domains, but the production version will certainly require the genuine research articles as its corpus.

3. **Efficient Punjabi Stemmer and Lemmatizers**
   Stemming and lemmatization are important tasks in the Information Retrieval process. The stemming and lemmatization reduces the inflated words to their root forms. These pre-processing tasks significantly reduce the size of corpus to be compared thereby improving over the overall speed of the system. Also the stemming and lemmatization helps in improving the similarity index.

4. **Keyword identification**
   Internet is comprised of a collection of huge volumes of documents. Keywords identification is required to reduce number of overall comparisons by excluding the irrelevant documents from the corpus. The relevant documents can be identified using the keywords that are most relevant to the document in question.

5. **Choice of method for similarity calculation**
   A number of methods have been proposed by researchers for similarity calculations. The suitable method for plagiarism detection task needs to be identified keeping in mind the complex morphology of the Punjabi language.
1.10. APPLICATIONS OF PLAGIARISM DETECTION TOOL

The plagiarism detection tool can be used in several applications as mentioned below:

- Detecting Plagiarism in
  - Student Reports/assignments
  - Research Papers submitted to Conferences & Journals
  - Doctoral Thesis
  - Magazine articles

- Similarity Detection in
  - Writing Styles of two authors
  - News Contents
  - Ancient Literatures
  - Medical diagnosis