Journal of Science-FAS-SEUSL (2021) 02(02) 47-64
ISSN:2738-2184

Plagiarism detection tools and techniques: A comprehensive survey

MAC Jiffriya[1]*, MAC Akmal Jahan[2], RG Ragel[3]

[1] Department of Information Technology, Sri Lanka Institute of Advanced Technological Education
[2]Department of Computer Science, South Eastern University of Sri Lanka
[3]Department Computer Engineering, University of Peradeniya

Abstract

Plagiarism is a rising issue in academics as increasing resources constantly, and it plays a high impact on student's performance and quality of education. Several studies have been carried out to mitigate the issue. This study provides a comprehensive analysis of the previous research, plagiarism detection approaches, existing plagiarism detection tools, type of the tools, features of popular plagiarism detection tools and the challenges in plagiarism detection. Plagiarism detection can be divided into source code and natural language plagiarism detection. Natural language plagiarism detection tools can be categorised based on the mode of detection, type of application, mode of services and languages. This study analyses existing plagiarism detection approaches and previous researchers' used approaches. In addition, it provides an overview of popular plagiarism detection tools, their features and the challenges when using them. Further, the study can support to develop an effective approach and tool to control the issue of efficiency in future. Although several tools are available for the plagiarism detection process, none of them is effective in accuracy and efficiency.

1. Definition of Plagiarism

The process of plagiarism is one of the prominent concerns in academia. The research on it has been carried out several decades to minimise the occurrence of plagiarism, maintain the quality of creating writing and protect copyright authorship. Plagiarism can be defined as violating the copyright of an author's or authors' literature work. It refers to copying someone's work or idea without proper acknowledgement, which diminishes the quality of the work. It is noted that digital plagiarism has many forms and definitions [1]. For instance, Alzahrani et al. quoted that plagiarism can range from copying part or all of the text without referring to the original composer, rephrasing the text by changing the words, expressing ideas of others work, translating creativity from a language to another language without the proper reference [2]. According to the Cambridge dictionary, plagiarism can be defined as "the process or practice of using another person's ideas or work and pretending that it is your own" [3]. In contrast, *Merriam-Webster*, which is one of the most reliable dictionaries in America, defines plagiarism as "to steal and pass off (the ideas or words of another) as one's own and: to commit literary theft" [4]. Further, plagiarism.org considers the following as plagiarism [5]

- Copying words, phrases or concepts from other's sources or converting one’s work as own work

Corresponding Author: jiffriya@sliate.ac.lk

- failing to mention the source of information and quotation marks or providing false information
- paraphrasing sentences or ideas without giving credit to the owner

Skandalakis mentioned plagiarism as "the theft of someone's words or thoughts" [6]. Based on the author's perception, plagiarism can be categorised as intention and un-intention. The former paraphrases one's ideas and thoughts without any intention due to huge volume sources, while the latter purposely steal others' ideas and thoughts without mentioning proper references [7]. Self-plagiarism is also another type of plagiarism that repeatedly publish the same work in a different form. Though plagiarism is taken place in several forms, such as plagiarism on literature, articles, book, poetry, song, documents, assignments, images, cinema, audio, web content, piece of art and so on, it becomes a critical challenge in academic writing. In universities, plagiarism is becoming a burning issue while submitting assignments, reports and creative publications. At this juncture, many scholars face the problem of plagiarised content while reviewing research articles and evaluating assignments and reports. Therefore, plagiarism is considered severe misconduct and intellectual dishonesty [8].

2. Occurrence of Plagiarism

Kauffman conducted an empirical study on digital plagiarism where 79.5% of writers were involved in it due to the easy accessibility of digital resources [9]. Students' attitudes were analysed towards written assignments and found that students understood plagiarism but were reluctant to practice this ethics [10]. The survey on plagiarism in the academic field carried out at the University of California in Berkley revealed the following facts: i) the percentage of plagiarism has increased by 74.4 % within four years period (1993 – 1997)[11]; ii) and more than 90% of high school students involve in plagiarism [1]. A trend of plagiarised assignment submission among students was analysed at the University of Bostwana, which depicts that, on averagely, 20.5% of students were involved in this immoral activity[12].

3. Reason for increasing plagiarism

The Internet widely opens the door for learning resources to learners and surfers. Electronic materials are widely used as resources in the academic community, which are simple to generate, modify, and delete. In addition, these are easy to store, maintain for an extended period and capable of exchanging or transmitting from one to others electronically through the network at high speed with the least cost. Therefore, most academics prefer electronic materials rather than printed materials, creating a paperless environment and leading a pathway for green IT world. Though the electronic resources positively impact the learning community, they generate adverse consequences through plagiarism among them. Because plenty of resources can be accessed from the Internet within a few seconds and reused easily by copying. As a result, plagiarism is reported widely in academic institutions. In addition, some educational institutions have a poor mechanism to prevent and control plagiarism. Some institutions try to hide the incidence of plagiarism due to maintaining the standard and hierarchy value of the institution. Also, at the state university of Sri Lanka, there is a lack of a proper mechanism to detect or eliminate plagiarism among students.

Further, students do not consider the importance of citation and reference details in their creative work. They provide poor attention on copyright authorship and offence for its violation. The punishment for violating the copyright authorship is primarily written in universities and other institutes, but the rules are rarely implemented. Moreover, academics are also reluctant to check the originality of students' creative work as it takes more effort and is time-consuming. Though we have various automated tools for it, the time consumption for plagiarism detection increases with the number of documents and the size of documents. Still, some automated tools' reliability is questionable, and there is lack of evidence to validate the results.

This study aims to provide an overview of plagiarism detection methods, techniques and tools. This paper analyses currently available detection tools and their features. This study categorises tools according to their features, which helps choose tools based on the writer's requirements. Another objective of this work is bringing out the gap in the field of plagiarism detection, which supports researchers to investigate that area and avoid repetition of the work and the study plan to analyse challenges in the domain of plagiarism detection.

4. The way for Controlling the Issue

Since plagiarism has become a growing trend among students in academic institutions, it needs to be controlled to maintain the academic program's quality. It can be controlled in two approaches such as prevention and detection. The former method focuses on making aware of plagiarism and its severe penalty. The latter approach involves detecting plagiarised content of a work. Some institutions provide grants for research or any part of the work in the field of literature where it is mandatory to examine the originality of the work. As the impact of rising plagiarism, scrutinising one's literary work is crucial for evaluating the works appropriately and equitably. To mitigate the issue and maintain the standard, detecting plagiarised content is effective. Hence, examiners would have to spend a massive amount of time reviewing the process to provide a high degree of judgment on students' creative work manually. Also, manual detection may be impossible when we increase sources and suspicious documents. Therefore, an automated tool for plagiarism detection is vital to detect plagiarism. It reduces the examiner's workload and helps evaluate students' creative writing properly without wasting valuable time.

5. Previous Surveys

A survey was carried out for plagiarism detection in 2006, which focused on text-based plagiarism detection and described some plagiarism detection tools with test cases and results [7]. In the following year, another survey was done by Lukashenko et al., which focused on a general way of minimising plagiarism and discussed metrics for calculating similarity scores. Although this study only showed a few attributes of seven plagiarism detection tools, this paper did not comprehensively analyse tools, detection algorithms and techniques [13]. In 2011, Garg conducted a study that briefly discussed plagiarism and described two source code plagiarism detection tools and six natural language plagiarism detection tools. Even though several plagiarism detection tools are available for source code and text, it explained only a few tools. This survey did not include the type of plagiarism detection, method and techniques for it and challenges [14]. Later a year, a study was conducted by Osman et al., who briefly explained the text-based plagiarism detection and existing plagiarism detection approaches [15]. Then Hiremath and Otari briefly described only five plagiarism detection tools and few detection techniques in 2014 [16]. Next year, Ahmed conducted an overview survey where he only discussed the existing 21 tools for textual and source code plagiarism. Still, it did not consider the type of plagiarism, detection methods, and challenges [17].

In contrast, Eisa et al. focused on the form of plagiarism and plagiarism detection techniques and their limitation, but it did not address the existing plagiarism detection tools [2]. Further, Naik et al. analysed the type of plagiarism and plagiarism detection methods and listed available detection tools for textual and source code, but it does not mention the challenges in these tools [18]. Following year, Vani and Gupta focused on extrinsic plagiarism detection techniques for textual plagiarism detection. Although many plagiarism detection tools are available, this study addressed only a few web-based plagiarism detection tools and their features [19]. In 2017, a comparison study on plagiarism detection approaches was carried out by Hourrane and Benlahmar [20]. Chowdhury and Bhattacharyya generated a taxonomy for the type of plagiarism and plagiarism detection tools for textual and source code detection methods. They briefly discussed 31 plagiarism detection tools for both sources. However, the study pointed out a few issues in plagiarism detection related to accuracy and reliability but did not mention them in detail [21]. A survey was carried out by Foltynek et al. in 2019. This survey focused on plagiarism detection between 2013 and 2018, which provides a comprehensive analysis of the type of plagiarism, plagiarism detection approach and methods. Though it mentions a lack of plagiarism detection and research gaps for future work, it does not describe plagiarism detection tools and challenges in this process [22]. Our study focuses on a comprehensive analysis of the previous research, existing plagiarism detection approaches, existing plagiarism detection tools, the type of the tools, features of popular plagiarism detection tools, and their challenges in plagiarism detection.

Categories of Plagiarism Detection Tools

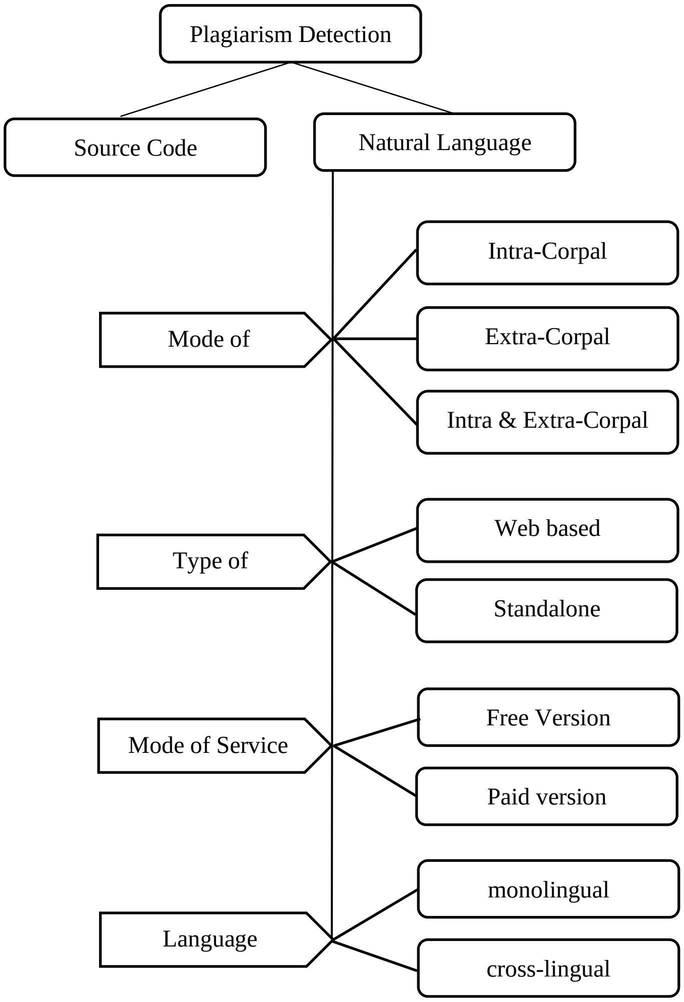

Fig 1: Categories of plagiarism detection tools

In the academic field, plagiarism detection is classified into two broad categories: a) source code; and b) natural language plagiarism detection. Also, Hussain A Chowdhury developed a taxonomy for plagiarism type where textual and source code plagiarism was the major type of plagiarism [23]. Source code plagiarism is reusing any pieces of programming code developed by other programmers without proper citation or even permission. In the early days, many researchers targeted to address plagiarism detection on automation of source code. The Table 1 shows source code plagiarism detection tools such as Plagio Guard [24][25], JPlag [26], Moss [27], Sherlock [28], Copy/Paste Detector (CPD) and Big Brother [29].

On the other hand, several natural language plagiarism detection tools are available, as shown in Table 1. The existing automated tools for natural language plagiarism detection can be categorised into intra-corpal and extra-corpal tools according to the area of the detection. Intra-corpal tools such as CopyFind [30] and CopyCatch [31] detect plagiarism within a set of grouped documents in a learning community, whereas EduTie [32], EVE2 [32], and Anti-P [33] are extra-corpal tools that check the similarity of available external sources of material on intranet and Internet [34]. Further, some tools such as iParadigms [32], WORDCheck [35] and Gossip [32] are capable of handling both intra and extra-corpals plagiarism detection.

Table 1: Plagiarism Detection tools and types

Type of Tools		**Plagiarism Detection Tools**
Source Code Tools		SIM, SID, MOSS, Jplag, Code Match, Plagio Gard, Detecta Copius, Copy/Paste Detector, Big Brother
Natural Language Tools	**Intra-Corpal**	CopyFind, CopyCatch, WORDCheck
	Extra-Corpal	EduTie, Thomas, EVE2, Anti-P
	Intra & Extra Corpal	iParadigms, Gossip, Turnitin

Based on the application, plagiarism detection tools can be categorised as web-based and standalone. We do not need to download and install the web-based tools, but it requires a high bandwidth internet connection. Tools such as PlagAware [36], PlagScan [37], iThenticate [38], CheckForPlagiarism.net [39] and plagiarismdetection.org [40] are web-based commercial tools which are commonly used by institutions and students. A comparison among these tools based on features and performance has shown that PlagAware and iThenticate were in first consecutive places [41]. On the other hand, standalone tools have to install on a user computer. WCopyfind [42], Plagiarism Detector and Desktop Plagiarism Checker [43] are a few examples for them.

Hussain A Chowdhury categorised plagiarism detection according to the number of languages such as monolingual and cross-lingual. In the former, source and suspicious documents are homogeneous, but later, the documents are heterogeneous. Monolingual plagiarism detection can be intrinsic detection, where it analyses the style of the author without checking the external source and extrinsic detection, where source document compares with all external on Internet and intranet [23]. Cross-language plagiarism detection mainly targets identifying plagiarised documents between documents in a different language.

Table 2: Web-based Plagiarism Detection Tools

S.No	**Web-Based Tools**	**URL**	**Type of Tool**
1	CheckForPlagiarism	https://www.checkforplagiarism.net/	Paid
2	Copycatch	https://copycatchlaw.com/	Paid
3	copyleaks	https://copyleaks.com/	Free
4	Copyscape	http://www.copyscape.com	Free
5	Doc Cop	http://www.doccop.com	Free
6	Docolc	http://www.docoloc.com	Paid
7	DupliChecker	http://www.duplichecker.com	Free
8	Dustball	http://www.dustball.com/cs/plagiarism.checker/	Free
9	EduBirdie	https://edubirdie.com/plagiarism-checker	Free

10	Ephorus	http://www.ephorus.com	Paid
11	Exactus Like	http://like.exactus.ru/index.php/en/	Free
12	GPSP (Glatt Plagiarism Screening Program)	http://www.plagiarism.com	Paid
13	Grammarly	https://www.grammarly.com/plagiarism-checker	Paid
14	iThenticate	http://www.ithenticate.com	Paid
15	PaperRater	https://www.paperrater.com/	Paid
16	Plagiarism Scanner	http://www.plagiarismscanner.com	Paid
17	Plagiarisma	http://www.plagiarisma.net	Free
18	PlagiarismChecker	http://www.plagiarismchecker.com	Free
19	PlagiarismDetect	http://www.plagiarismdetect.org	Free
20	Plagiarismhunt	https://plagiarismhunt.com/	Free
21	Plagium	http://www.plagium.com	Free
22	Plagpatrol	https://josemmo.github.io/plagpatrol/	Free
23	plagramme	https://www.plagramme.com/	Paid
24	PlagScan	http://www.plagscan.com	Paid
25	PlagTracker	http://www.plagtracker.com	Paid
26	Quetext	http://www.quetext.com	Free
27	SafeAssignment	https://www.blackboard.com/teaching-learning/learning-management/safe-assign	Free
28	Scribbr	https://www.scribbr.com/	Paid
29	Search Engine Report Plagiarism Checker	https://searchenginereports.net/plagiarism-checker	Free
30	SmallSeoTools	https://smallseotools.com/plagiarism-checker/	Free
31	Turnitin	http://www.turnitin.com	Paid
32	Unicheck	https://unicheck.com/	Paid
33	Urkund	http://www.urkund.com	Free
34	Viper	http://www.scanmyessay.com	Free
35	noplag	https://noplag.com/	Paid
36	plagly	https://plagly.com/	Free
37	PlagiarismSearch	https://plagiarismsearch.com/	Paid

6. Mode of Plagiarism Detection

Natural language plagiarism detection has been a challenging issue in recent years as the materials' source is increasing day by day on the Internet. The plagiarism detection process can be carried out mainly in two approaches: 1) intrinsic and 2) extrinsic. The intrinsic approach focuses on analysing stylometric features to identify how does writing style of the writer differ. Here, documents are investigated according to an author's style, assuming that each author has a unique style [44]. If the author's style differs from the unique writing style, then it is considered that the document contains the plagiarised text. In this approach, it does not retrieve related documents and compare them. This approach is ineffective in analysing students' creative writing. Because it would be impossible to recognise students' writing style as they submit an assignment for the first time. Popular detection tools such as Turnitin [45], Viper [46], Plagium [47], PlagTracker [48] and Paper Rater [49] do not have the facility to analyse the writing style of the author [50].

On the other hand, the extrinsic approach intends to compare suspicious documents with all other related documents retrieved from all-external sources. In this method, documents are analysed based on the logical structure, where it analyses how much similar content is available between the suspicious and the candidate documents. If the similarity score between these documents is greater than the threshold value, then the document is considered a plagiarised document. This approach has two major processes, such as candidate document retrieval and analysis of the similarity between the suspicious document and the candidate documents.

Candidate document retrieval: In this process, keywords are extracted from the suspicious document, then passed into the search engines and local and global databases to search relevant documents on the Internet and intranet. The retrieved documents are collected together, called candidate documents, to be compared with the suspicious document, which is highly influenced by keyword extraction.

Analysis of Plagiarism: In an analysis of plagiarism, there are several plagiarism detections approaches, and they are: i): lexical; ii) structural; iii) semantic; iv) stylometric; v) syntactic; vi) citation and vii) cross language-based [2]. The lexical approach focuses on a document's lexical structure, which can be at a character or a word level. Fingerprinting, N-gram, Standford Copy Analysis Mechanism (SCAM), Longest common subsequence and clustering method are more popular lexical methods. It analyses the grammatical structure of documents and is better to detect copy-paste, but it cannot identify plagiarism on paraphrased documents [20]. The structural approach focuses on how the words are distributed in a document and structural features such as keywords, header, paragraph and references [2]. Tree and graph are widely used structural approaches that support detecting the idea of plagiarism. The semantic approach analyses the meaning of a document by considering synonyms and antonyms. Latent Semantic Analysis, fuzzy and vector-based methods are common for the sematic based approach [20]. The stylometric approach is a statistical method that concerns an author's style using the assumption that each author has a unique style of writing. Still, this method is not appropriate for translated documents, and the reliability of detection is low when the number of authors documents increases. The syntactic approach focuses on parts of speech of the document by dividing the text into chunks with the concern of function word, punctuation and word root [51]. Citation based approach is concerned with comparing reference documents with source documents, while cross language-based approach focuses on comparing multiple language documents [15].

7. Plagiarism Detection Process

A typical plagiarism detection process consists of major four tasks: a) Pre-processing, b) Feature extraction, c) Feature representation and d) Similarity Score.

a) Pre-processing: Pre-processing means preparing the required document format by eliminating unwanted text from it.

b) Feature extraction

Feature extraction is selecting appropriate text or words to represent the document. According to the feature extraction, there are four types of features such as lexical, structural, syntactic and semantic.

i. *Lexical feature:* It focuses on the lexical structure of the text, which can be character level and word level. The former concerns the sequence of character and N-gram while the latter is bag of words that focuses on word frequency, word character length, word N-gram and vocabulary richness [52].

ii. *Structural feature**:*** It concerns keywords, header, reference and paragraph where the words are distributed throughout the documents. In a study, plagiarism detection was handled by extracting title, author and keywords from articles, which were then represented as a binary matrix and compared. However, extracting such keywords is critical for detection [33].

iii. *Syntactic feature:* It analyses the structural pattern of words and their position. It divides the document into chunks based on part of speech. Sentence based technique was used in a study where a sentence of each document was compared with a sentence of another document, and similarity was computed as a i) function of the word in common and ii) length of the sentence. However, this method did not detect all plagiarised documents in a large dataset [31].

iv. *Semantic feature:* It analyses the meaning of context by considering synonyms, antonyms and dependency meaning using WordNet or EuroVoc.

c) Feature representation

Feature representation is defined as representing the document in memory to be processed. Several approaches have been used in the existing research, such as a bag of words, vector, tree and graph [53] [54] [55]. Among the current document representation approaches, a bag of words is commonly used as it is simple, and documents are represented as N-gram in text-based applications [56].

PlagInn algorithm was developed with the assumption that every author has a unique style of writing. In this method, sentences in documents were represented as grammar trees, and similarity was measured by comparing distances between trees [44].

d) *Similarity Score*

Computing the similarity score is the eventual phase of plagiarism detection where the approaches of cosine similarity, Jaccard similarity, dice similarity, overlap similarity, Humming distance and Euclidean distance are very frequent in past research [57][58]. However, these similarity measurements do not apply to all types of documents because they depend on the document representation method.

8. Previous Plagiarism Detection Approaches

Table 3: Author who used to Plagiarism Detection Approaches

Reference	Main Approaches	Algorithm
[59]	This article focuses on cross-language plagiarism detection according to the Ontology Learning approach.	Ontology Learning System, with the help of machine learning techniques.
[60]	Text-based plagiarism detection for text messages based on the semantic feature using a deep learning technique	Convolutional neural network (CNN) and a recurrent neural network (RNN) for feature vector generation with Cosine similarity measure.
[61]	The paper focuses on integrating stylometric and sematic based features for plagiarism detection.	LSA, Multi-layer perception
[62][63]	Plagiarism is detected according to the character-based and cluster-based approaches	Tri-gram K-means algorithm for clustering
[62]	Plagiarism detection based on clustering approach	Machine learning language: k-Nearest Neighbour Algorithm
[51]	The study focuses on the integration of stylometry and a semantic-based approach	Latent Semantic Analysis (LSA) SVM for classification
[64]	Lexical based approach	Word level Tri-gram sequence matching with Jaccard similarity
[65]	Semantic and syntactic based approach	Cosine similarity measure
[66]	Cross-language plagiarism detection method	Knowledge graphs: word sense disambiguation, vocabulary expansion, and representation by similarities with a collection of concepts.
[67]	Lexical based approach	Vector Space model with cosine similarity
[68]	Semantic-based approach	Optimal abductive network models
[69]	Cluster-based approach	Bloom filters & Hashing function
[70]	Syntactic based approach	Pattern analyses based on the part of speech using natural language processing techniques with Jaccard coefficient for similarity.
[71]	Lexical based approach & cluster	Tri-gram sequence matching. K-Means algorithm for cluster
[54]	Character based-Finger printing approach with tree structure representation	Heuristic algorithm for searching Longest Common Substring metric generation

9. Plagiarism Detection Tools

Table 4: The popular plagiarism detection tools and their features

Detection Tools	**Features**
Plagiarisma [19][72]	It applies a simple string-matching algorithm. Multiple language support that more than 190. Support to multiple type file format as txt, html, rtf, doc, docx, xls, xlsx, pdf, odt, epub, fb2, and pdb. Free & paid version. The free version has a limited number of checks.
PlagScan [37]	Web-based commercial tool. It provides services to personal users and institutions. In a free trial, it can check up to 2000 words. It can handle many types of files, including doc, docx, odt, html, pdf and zip. User can compare side by side views. Maximum 30,000 words.
Viper [46]	Free online plagiarism detection tool. It provides unlimited resubmission and shows links to plagiarised work.
SmallSeoTools [19][73]	Free web service. Support up to 1000 words per search. It can check tex, txt, doc, docx, odt, pdf and rtf files.
Urkund:[17][74]	Web-based service. Document is submitted via email, and users can receive results through it.
Docol©c [23][75] www.docoloc.com	Commercial online tool. It searches on entirely Google API. User gets result of submission via email. It can support txt, pdf, doc, docx, odt and rtf. It lists link which has similar sentence of the submitted document.
SafeAssignment [76][18][77]	Free online plagiarism prevention service. It supports multi-language such as English, Arabic, Chinese, Dutch, French, German, Japanese, Spanish. User does not have control on the detection method. It is a part of Blackboard (virtual learning platform) product. It provides an overall match score. It is better to identify plagiarism on the web.
Copycatch [19]	Standalone tools use the local database while online versions use Google API It can support txt, rtf and doc file type.
Wcopyfind [76]	It is a desktop tool and uses a local repository extending to access Google API.
EVE2 [19]	"Essay Verification Engine" is a standalone tool was created by Canexus It searches on Web searching engines and does not have a local database.
Dupli Checker [78]	Free online tool. 50 checks per day. No paid version.
Copyleaks [79]	Used for education and business. Support multiple file formats. Search on e-learning content. Free for the first 10 pages. Allow to check freely 2500 words per month.

Plagiarism Checker	Totally online free tool. User friendly interface and easy handle.
Plagium [47]	Simple online tool. 5,000 characters per search freely check. It does a quick and deep search.
Ephorus [80]	Online commercial tool. It can support to popular twenty languages of the world.
iThenticate [19]	Paid online plagiarism detection service. It is more appropriate for universities and institutions. It is a well-known tool to publishers such as Elsevier, Springer, Wiley and IEEE.
Doc Cop [81]	Free web-based plagiarism detection tools. It generates a correlation report between documents and the web.
Turnitin [76][82][77]	Web-based commercial plagiarism tool. A product of iParadigms It is generated for students and academics, especially teachers It is popular in most institutions. It can support intra-corpal & extra-corpal detection. It is pretty expensive. It does not generate an instant response.

10. Challenges of plagiarism detection tools

We have analysed several studies related to plagiarism detection after 2000. In natural language plagiarism detection, there is a lack of study on detecting plagiarism in tables and figures, and existing tools are incapable of detecting plagiarised images, tables, figures, formulas and scanned documents [2]. There is another challenge in using the tools is security and privacy. Some tools save the submitted users' documents in their repository. For example, Turnitin is one of the famous commercial tools which saves students' assignments and writings in its database for future plagiarism detection. It is considered as illicit practice [83].

The detection tools generate false-positive results and fail to detect copied content. TurnitIn, SafeAssignment, Plagiarism-Finder and EVE were evaluated, but these showed poor performance in terms of accuracy [84]. It has been identified three critical reasons for it such as i) scope of the detection, ii) paraphrasing, and iii) cross-language [7]. The scope of the area is one of the significant factors to determine the effectiveness of the detection tool. Some tools compare documents with their repository only. It fails to detect plagiarism when the plagiarised document is not available in its repository. And also, some web content is invisible, and documents cannot be accessible since it is a password-protected database containing many journal articles [85]. Secondly, a paraphrased text which can be created in three ways: a) replacing synonym words, b) expressing the same context by altering syntactic form and grammatical structure, and c) rewording of the content. Plagiarism detection tools are struggling to detect paraphrased text. The third reason for failing to detect proper plagiarised documents is cross-language, which occurs during the process of translating an original document into another language, and people are unaware of the translated document and the ethical issues [66].

Citation based plagiarism detection is difficult to identify the plagiarism because there are several modes of citation in a variety of disciplines. Turnitin cannot differentiate appropriately cited sentences in quotation marks and illegally stolen text [85]. Also, Docol©c reported quoted sentences with citations as found sentences. User intervention is required to decide about the plagiarised content. Docol© produces a similarity score, but the user has to review the result and determine whether the document has been plagiarised or not [75].

Time consumption for processing and generating reports are demanding challenges. The time is increased with the document size, and it demands high bandwidth. Turnitin is considered one of the leading plagiarism detection tools due to its richness in functionality and a massive number of users. However, it consumes a larger amount of time to generate reports and criticism of slow reporting time [85]. In our previous study, 41 students' assignments were tested with *Plagiarism Checker X,* which consumed more than two hours [71]. Moreover, consumption time for the detection increased dramatically with respect to the number of documents as there was a higher number of comparisons and searching among them.

Technical restriction and limitation are other significant challenges in plagiarism detection. Web-based tools do not need to be downloaded and installed on a user computer but require a high bandwidth internet connection. Existing tools can handle a limited number of documents with a limited file size at a time. They utilise a significant amount of time when checking many documents. For instance, PlagAware cannot handle a document that is more than 15 MB in size.
Further, the execution time for the detection depends on the workload of servers and documents' size [36]. Turnitin is a web-based well-known plagiarism detection tool used by 35000 educational institutes worldwide. Though it can support 400 pages with a maximum size of 40 MB, the student can check only a maximum 0.5MB document size with a 150,000 characters limitation. An originality report may be generated within minutes to hours [38]. Also, iThenticate and Ferret do not allow checking a document of more than 25000 and 10000 words, respectively [86].

Moreover, plagiarismDetection.org cannot support multiple document comparisons and consumes more time to display the analysed results [41]. As a result, plagiarism detection tools become inefficient when handling a massive number of documents due to the consumption of a significant amount of time. Some tools are incapable of handling a large number of documents as well. Also, it is a significant bottleneck in the field of plagiarism detection. Although there are many tools available for plagiarism detection, none of them is more effective for it as these tools have some limitations in terms of reliability, accuracy and speed.

References

1. S. Butakov and V. Scherbinin, "The toolbox for local and global plagiarism detection," *Comput. Educ.*, vol. 52, no. 4, pp. 781–788, 2009, doi: 10.1016/j.compedu.2008.12.001.
2. T. A. E. Eisa, N. Salim, and S. Alzahrani, "Existing plagiarism detection techniques: A systematic mapping of the scholarly literature," *Online Inf. Rev.*, vol. 39, no. 3, pp. 383–400, 2015, doi: 10.1108/OIR-12-2014-0315.
3. Cambridge Dictionary (online), "PLAGIARISM | meaning in the Cambridge English Dictionary," *Cambridge Dictionary (online)*. 2020, [Online]. Available: https://dictionary.cambridge.org/dictionary/english/plagiarism.
4. Marriam-Webster, "Dictionary by Merriam-Webster: America's most-trusted online dictionary," *Merriam Webster*. 2002, [Online]. Available: https://www.merriam-webster.com/.
5. "What is Plagiarism_ - Plagiarism." [Online]. Available: https://www.plagiarism.org/article/what-is-plagiarism.
6. S. Article, "Arch Surg. 2004;139:1022-1024," vol. 139, pp. 1022–1024, 2004.
7. H. Maurer, F. Kappe, and B. Zaka, "Plagiarism - A survey," *J. Univers. Comput. Sci.*, vol. 12, no. 8, pp. 1050–1084, 2006.
8. W. M. Journal, "Plagiarism in Scientific Publishing Plagiarism in Scientifi c Publishing," *Acta Inf. med*, vol. 20, no. JUNE 2013, pp. 208–213, 2015, [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3558294&tool=pmcentrez&rendertype=abstract.
9. Y. Kauffman and M. F. Young, "Digital plagiarism: An experimental study of the effect of instructional goals and copy-and-paste affordance," *Comput. Educ.*, vol. 83, pp. 44–56, 2015, doi: 10.1016/j.compedu.2014.12.016.
10. P. Newton, "Academic integrity: a quantitative study of confidence and understanding in students at the start of their higher education," *Assess. Eval. High. Educ.*, vol. 41, no. 3, pp. 482–497, 2016, doi: 10.1080/02602938.2015.1024199.
11. R. Lukashenko, V. Graudina, and J. Grundspenkis, "Computer-based plagiarism detection methods and tools: An overview," *ACM Int. Conf. Proceeding Ser.*, vol. 285, no. January, 2007, doi: 10.1145/1330598.1330642.
12. T. Batane, "Turning to Turnitin to Fight Plagiarism among University Students," *Educ. Technol. Soc.*, vol. 13, no. 2, pp. 1–12, 2010.
13. R. Lukashenko, V. Graudina, and J. Grundspenkis, "Computer-based plagiarism detection methods and tools: An overview," *ACM Int. Conf. Proceeding Ser.*, vol. 285, pp. 1–6, 2007, doi: 10.1145/1330598.1330642.
14. U. Garg, "Plagiarism and Detection Tools: An Overview," *J. Eng. Sci.*, pp. 92–97, 2011.
15. A. Hamza Osman, N. Salim, and A. Abuobieda, "Survey of Text Plagiarism Detection," *Comput. Eng. Appl. J.*, vol. 1, no. 1, pp. 37–45, 2012, doi: 10.18495/comengapp. v1i1.5.
16. S. A. Hiremath and M. S. Otari, "Plagiarism Detection-Different Methods and Their Analysis: Review," *Int. J. Innov. Res. Adv. Eng.*, vol. 1, no. 7, pp. 2349–2163, 2014.
17. R. A. Ahmed, "Overview of Different Plagiarism Detection Tools," vol. 2, no. 10, pp. 2–4, 2015.
18. R. R., M. B., and C. Namrata, "A Review on Plagiarism Detection Tools," *Int. J. Comput. Appl.*, vol. 125, no. 11, pp. 16–22, 2015, doi: 10.5120/ijca2015906113.
19. D. G. Vani K, "Study on Extrinsic Text Plagiarism Detection Techniques and Tools," vol. 9, no. 4, pp. 150–164, 2016.

20. O. Hourrane and E. H. Benlahmar, "Survey of plagiarism detection approaches and big data techniques related to plagiarism candidate retrieval," *ACM Int. Conf. Proceeding Ser.*, vol. Part F1294, pp. 1–6, 2017, doi: 10.1145/3090354.3090369.
21. H. A. Chowdhury and D. K. Bhattacharyya, "Plagiarism: Taxonomy, tools and detection techniques," *arXiv*, no. February, 2018.
22. N. Meuschke and B. Gipp, "Academic Plagiarism Detection : A Systematic Literature," vol. 52, no. 6, 2019.
23. H. A. Chowdhury and D. K. Bhattacharyya, "Plagiarism: Taxonomy, tools and detection techniques," *arXiv*, no. 1, 2018.
24. S. Goel, D. Rao, and Others, "Plagiarism and its Detection in Programming Languages," 2005, [Online]. Available: http://pdf.aminer.org/000/591/297/yap_improved_detection_of_similarities_in_computer_program_and_other.pdf.
25. E. L. Jones, "Metrics based plagiarism monitoring," 2001.
26. L. Prechelt, G. Malpohl, and M. Philippsen, "Finding plagiarisms among a set of programs with JPlag," *J. Univers. Comput. Sci.*, vol. 8, no. 11, pp. 1016–1038, 2002.
27. S. Schleimer, D. S. Wilkerson, and A. Aiken, "Winnowing: Local Algorithms for Document Fingerprinting," *Proc. ACM SIGMOD Int. Conf. Manag. Data*, pp. 76–85, 2003.
28. M. Joy and M. Luck, "Plagiarism in programming assignments," *IEEE Trans. Educ.*, vol. 42, no. 2, pp. 129–133, 1999, doi: 10.1109/13.762946.
29. T. Lancaster and F. Culwin, "A comparison of plagiarism detection tools," *Comput. Sci. Educ.*, vol. 14, no. 2, pp. 101–112, 2004, [Online]. Available: http://www.cs.uu.nl/research/techreps/repo/CS-2010/2010-015.pdf.
30. "The Plagiarism Resource Site « Welcome." [Online]. Available: http://plagiarism.bloomfieldmedia.com/z-wordpress/.
31. D. R. White and M. S. Joy, "Sentence-Based Natural Language Plagiarism Detection," *ACM J. Educ. Resour. Comput.*, vol. 4, no. 4, p. 2, 2004, doi: 10.1145/1086339.1086341.
32. T. Lancaster and F. Culwin, "Classifications of plagiarism detection engines," *Innov. Teach. Learn. Inf. Comput. Sci.*, vol. 4, no. 2, pp. 1–16, 2005, doi: 10.11120/ital.2005.04020006.
33. A. Al Jarrah, I. Alsmadi, and Z. Za'atreh, "Plagiarism Detection based on studying correlation between Author, Title, and Content," *Int. Conf. Inf. Commun. Syst.*, pp. 22–24, 2011.
34. Y. Kats, Learning Management System Technologies and Software solution for online Teaching: Tools and Application. New York: Information Science Reference Hershey, 2010.
35. P. Clough, "Plagiarism in natural and programming languages: an overview of current tools and technologies," *Finance*, no. July, pp. 1–31, 2000, [Online]. Available: http://www.dcs.shef.ac.uk/nlp/meter/Documents/reports/plagiarism/Plagiarism.pdf.
36. "Plagiarism Check | For Universities, Students and SEO." [Online]. Available: https://www.plagaware.com/.
37. "Online Plagiarism Checking | PlagScan." 2019, [Online]. Available: https://www.plagscan.com/en/.
38. L. Turnitin, "Plagiarism Detection Software | iThenticate," *Web del sistema*. 2018.
39. "Check For Plagiarism - The Ultimate Plagiarism Checker." [Online]. Available: https://www.checkforplagiarism.net/.
40. "Plagiarism Checker for Teachers, Students, Bloggers." [Online]. Available: http://plagiarismdetection.org/.

41. A. M. El Tahir Ali, H. M. Dahwa Abdulla, and V. Snášel, "Overview and comparison of plagiarism detection tools," *CEUR Workshop Proc.*, vol. 706, pp. 161–172, 2011.
42. H. Dreher, "Automatic Conceptual Analysis for Plagiarism Detection," *Issues Informing Sci. Inf. Technol.*, vol. 4, pp. 601–614, 2007, doi: 10.28945/974.
43. S. I. Nugraha and K. Wachyudi, "THE 1 ST PROCEEDINGS OF NATIONAL SEMINAR ON TEACHING ENGLISH TO YOUNG LEARNERS UNIVERSITAS SINGAPERBANGSA KARAWANG 2019 Wednesday, August 7 th 2019."
44. M. Tschuggnall and G. Specht, "Detecting plagiarism in text documents through grammar-analysis of authors," *Lect. Notes Informatics (LNI), Proc. - Ser. Gesellschaft fur Inform.*, vol. P-214, pp. 241–259, 2013.
45. "Empower Students to Do Their Best, Original Work | Turnitin." [Online]. Available: https://www.turnitin.com/?svr=54&session-id=&lang=en_us&r=0.724938223983784.
46. "Plagiarism Checker _ Viper Online.".
47. "Free Plagiarism Checker: Plagium." [Online]. Available: https://www.plagium.com/.
48. PlagTracker, "Plagiarism Checker - the most accurate and absolutely FREE! Try now!" [Online]. Available: http://www.plagtracker.com/.
49. PaperRater, "Free Online Proofreader: Grammar Check, Plagiarism Detection, and more." 2019, [Online]. Available: https://www.paperrater.com/.
50. R. G. S. Berlinck, "The academic plagiarism and its punishments - a review," *Brazilian J. Pharmacogn.*, vol. 21, no. 3, pp. 365–372, 2011, doi: 10.1590/S0102-695X2011005000099.
51. M. Alsallal, R. Iqbal, S. Amin, A. James, and V. Palade, "An Integrated Machine Learning Approach for Extrinsic Plagiarism Detection," *Proc. - 2016 9th Int. Conf. Dev. eSystems Eng. DeSE 2016*, pp. 203–208, 2017, doi: 10.1109/DeSE.2016.1.
52. K. Lagutina *et al.*, "A Survey on Stylometric Text Features," *Conf. Open Innov. Assoc. Fruct*, pp. 184–195, 2019, doi: 10.23919/FRUCT48121.2019.8981504.
53. A. H. Osman, N. Salim, and M. S. Binwahlan, "Plagiarism Detection Using Graph-Based Representation," vol. 2, no. 4, pp. 36–41, 2010, [Online]. Available: http://arxiv.org/abs/1004.4449.
54. M. E. B. Menai, "Detection of Plagiarism in Arabic Documents," *Int. J. Inf. Technol. Comput. Sci.*, vol. 4, no. 10, pp. 80–89, 2012, doi: 10.5815/ijitcs.2012.10.10.
55. M. Zechner, M. Muhr, R. Kern, and K. Graz, "External and Intrinsic Plagiarism Detection."
56. T. Kučečka, "Plagiarism Detection in Obfuscated Documents Using an N-gram Technique," *ACM Slovakia, Spec. Sect. Student Res. Informatics Inf. Technol.*, vol. 3, no. 2, pp. 67–71, 2011, [Online]. Available: https://pdfs.semanticscholar.org/c311/ec10650512533cd11b44ed1153f975e21c7d.pdf.
57. C. Xiao, W. Wang, X. Lin, J. X. Yu, and G. Wang, "Efficient similarity joins for near-duplicate detection," *ACM Trans. Database Syst.*, vol. 36, no. 3, 2011, doi: 10.1145/2000824.2000825.
58. Q. Zhang, Y. Wu, Z. Ding, and X. Huang, "Learning hash codes for efficient content reuse detection," *SIGIR'12 - Proc. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 405–414, 2012, doi: 10.1145/2348283.2348339.
59. P. Pathak, S. Pillai, and P. R. Pal, "Ontology learning system for cross language plagiarism detection," *Int. J. Sci. Technol. Res.*, vol. 9, no. 3, pp. 6190–6196, 2020.
60. B. Agarwal, H. Ramampiaro, H. Langseth, and M. Ruocco, "A deep network model for paraphrase detection in short text messages," *Inf. Process. Manag.*, vol. 54, no. 6, pp. 922–937, 2018, doi: 10.1016/j.ipm.2018.06.005.

61. S. M. Alzahrani, N. Salim, and M. M. Alsofyani, “Work in progress: Developing arabic plagiarism detection tool for e-learning systems,” *2009 Int. Assoc. Comput. Sci. Inf. Technol. - Spring Conf. IACSIT-SC 2009*, pp. 105–109, 2009, doi: 10.1109/IACSIT-SC.2009.22.
62. M. Sahu, “Plagiarism Detection Using Artificial Intelligence Technique In Multiple Files,” *Int. J. Sci. Technol. Res.*, vol. 4, no. 8, pp. 111–114, 2015.
63. M. Abid, M. Usman, and M. W. Ashraf, “Plagiarism Detection Process using Data Mining Techniques,” *Int. J. Recent Contrib. from Eng. Sci. IT*, vol. 5, no. 4, p. 68, 2017, doi: 10.3991/ijes. v5i4.7869.
64. M. A. C. Jiffriya and M. A. C. A. Jahan, “Accelerating Text-based Plagiarism Detection Using GPUs,” pp. 395–400, 2015.
65. A. Abdi, N. Idris, R. M. Alguliyev, and R. M. Aliguliyev, “PDLK: Plagiarism detection using linguistic knowledge,” *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8936–8946, 2015, doi: 10.1016/j.eswa.2015.07.048.
66. M. Franco-Salvador, P. Rosso, and M. Montes-y-Gómez, “A systematic study of knowledge graph analysis for cross-language plagiarism detection,” *Inf. Process. Manag.*, vol. 52, no. 4, pp. 550–570, 2016, doi: 10.1016/j.ipm.2015.12.004.
67. M. A. C. Jiffriya, M. A. C. A. Jahan, and R. G. Ragel, “Plagiarism detection on electronic text-based assignments using vector space model,” *2014 7th Int. Conf. Inf. Autom. Sustain. "Sharpening Futur. with Sustain. Technol. ICIAfS 2014*, 2014, doi: 10.1109/ICIAFS.2014.7069593.
68. E. S. M. El-Alfy, R. E. Abdel-Aal, W. G. Al-Khatib, and F. Alvi, “Boosting paraphrase detection through textual similarity metrics with abductive networks,” *Appl. Soft Comput. J.*, vol. 26, pp. 444–453, 2015, doi: 10.1016/j.asoc.2014.10.021.
69. S. Geravand and M. Ahmadi, “An efficient and scalable plagiarism checking system using Bloom filters,” *Comput. Electr. Eng.*, vol. 40, no. 6, pp. 1789–1800, 2014, doi: 10.1016/j.compeleceng.2014.06.003.
70. A. R. Adam and Suharjito, “Plagiarism detection algorithm using natural language processing based on grammar analysing,” *J. Theor. Appl. Inf. Technol.*, vol. 63, no. 1, pp. 168–180, 2014.
71. M. A. C. Jiffriya, M. A. C. A. Jahan, R. G. Ragel, and S. Deegalla, “AntiPlag: Plagiarism detection on electronic submissions of text-based assignments,” *2013 IEEE 8th Int. Conf. Ind. Inf. Syst. ICIIS 2013 - Conf. Proc.*, pp. 376–380, 2013, doi: 10.1109/ICIInfS.2013.6732013.
72. “Plagiarisma _ Plagiarism Checker”.
73. “Plagiarism Checker - 100% Free Online Plagiarism Detector.” [Online]. Available: https://smallseotools.com/plagiarism-checker/.
74. “Urkund – Plagiarism prevention that simply works.” [Online]. Available: https://www.urkund.com/.
75. “Docol©c- Add Document”.
76. T. Kakkonen and M. Mozgovoy, “Hermetic and web plagiarism detection systems for student essays-an evaluation of the state-of-the-art,” *J. Educ. Comput. Res.*, vol. 42, no. 2, pp. 135–159, 2010, doi: 10.2190/EC.42.2. a.
77. N. C. Heckler, M. Rice, and C. H. Bryan, “Turnitin systems: A deterrent to Plagiarism in College Classrooms,” *J. Res. Technol. Educ.*, vol. 45, no. 3, pp. 229–248, 2013, doi: 10.1080/15391523.2013.10782604.
78. “Plagiarism Checker | 100% Free and Accurate,” *Duplichecker.com*. 2020, [Online]. Available: https://www.duplichecker.com/.

79. "Plagiarism Detector_ AI Based Anti-Plagiarism Online _ Copyleaks." .
80. "Plagiarism Checker - 100% Free to Detect Plagiarism Online." [Online]. Available: https://searchenginereports.net/plagiarism-checker.
81. "DOC Cop = Accurate + Fast + Free + Simple + Plagiarism Detection." [Online]. Available: http://www.doccop.com/index.html?nc=71423917.
82. E. R. Bensal, E. S. Miraflores, and N. C. C. Tan, "Plagiarism: Shall we turn to Turnitin," *Call-Ej*, vol. 14, no. 2, pp. 2–22, 2013.
83. S. P. and S. P. Hoshiladevi Ramnial and Abstract, "Authorship Attribution Using Stylometry and Machine Learning Techniques," 2016, doi: 10.1007/978-3-319-23036-8.
84. T. Kakkonen and M. Mozgovoy, "An evaluation of web plagiarism detection systems for student essays," *Proc. - ICCE 2008 16th Int. Conf. Comput. Educ.*, no. 1, pp. 99–103, 2008.
85. L. McKeever, "Online plagiarism detection services-saviour or scourge," *Assess. Eval. High. Educ.*, vol. 31, no. 2, pp. 155–165, 2006, doi: 10.1080/02602930500262460.
86. C. Lyon, R. Barrett, and J. Malcolm, "A theoretical basis to the automated detection of copying between texts, and its practical implementation in the Ferret plagiarism and collusion detector," *Proc. 1st Int. Plagiarism Conf.*, pp. 1–7, 2004.