Analysis of Support Tools for Plagiarism Detection

Vrushali Bhuyar[1](✉) and S. N. Deshmukh[2]

[1] Maharashtra Institute of Technology, Aurangabad, India
vrushali.bhuyar@gmail.com
[2] Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India

Abstract. Plagiarism in academic research is on the rise. It is common practice to post stuff that was taken directly from the internet without giving the author credit. The researcher used a variety of tactics, including copy-paste, idea plagiarism, paraphrasing, artistic plagiarism, code plagiarism, forgotten or expired links to resources, improper use of quotation marks, misinformation of references, and translated plagiarism. Despite the availability of numerous commercial tools, these tools are unable to identify plagiarism. The analysis of various plagiarism detection systems, which are frequently used to find plagiarism in academic work, is the primary subject of this paper.

Keywords: Plagiarism Detection Tools · Dupli Checker · Plag Scan · Urkund · Turnitin

1 Introduction

According to Teddi Fishman, Plagiarism is defined as, "Plagiarism occurs when someone uses words, ideas, or work products, attributable to another identifiable person or source, without attributing work to the source from which it was obtained, in a situation where there is a legitimate expectation of original authorship, in order to obtain some benefit, credit, or gain which need not be monetary" [1].

Due to the widespread usage of internet technologies, data availability is expanding accordingly. There are many academicians, researchers and students utilizing data from internet and using it for their own purposes. To avoid being caught in the process of plagiarism, many people alter the text, replace terms, use synonyms instead of the original words, do paraphrasing, change active to passive or vice versa, and other techniques. Most often, it occurs in higher education, where students and instructors plagiarize by using previously published work.

The work of performing plagiarism detection at different stages is crucial for preventing plagiarism and maintaining the uniqueness of information source. For many years, research has been underway to achieve this goal. Numerous methods and technologies have been developed over time to identify plagiarism at different levels. Even though there are many techniques available for detecting plagiarism, it is still unclear how these tools will get developed in providing high degree of accuracy.

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 38–46, 2023.
https://doi.org/10.2991/978-94-6463-136-4_6

2 Literature Review

Similarity detection tools also known as "anti-plagiarism" software or plagiarism detection software is now widely available, both as commercial solutions and as open-source software. It is very difficult task to identify all sources of plagiarized document. Although software cannot detect plagiarism, it can assist in finding text similarities that may indicate plagiarism. Study shows that certain methods can assist in the detection of some plagiarized content, study showed that software were unsuccessful in detecting entire plagiarized contents and occasionally identified non-plagiarized content for plagiarism. There are plenty of papers available on plagiarism detection tools. Finding from some of papers are discussed below.

[2], tested plagiarism on three tools Turnitin and Mydropbox and Docol© with respect to cut-paste check, paraphrase check, tabular information processing, translation check, image/multi-media checks, reference validity check, exclusion/selection of sources. Study showed good result for verbatim but failure for paraphrasing, tabular information, translation, special character and cross lingual detection.

[3], discussed about plagiarism, plagiarism type and detection techniques. In this paper, author focused on different extrinsic plagiarism detection techniques with its pros and cons. 8 plagiarism checkers along with the features are mentioned in this study but comparison given between 3 tools small Seo, Plagiarisma and Turnitin on the basis of No obfuscation (copy-past), random obfuscation and translation obfuscation. Result showed that these tools worked effectively for copy paste but not for structural differences and paraphrasing.

[4], provided comparative analysis on 31 plagiarism detection tools based on its use (Extrinsic or Intrinsic), submission of single or multiple files, free or paid software and user friendliness. However, there is no guidance provided on the basis of tool performance or direction for tool selection.

[5], created large corpus of intentionally plagiarized document with the help of sources (Wikipedia, online articles, open access papers, student theses available online) and various plagiarism techniques (copy & paste, synonym replacement, paraphrase, translation). In this study researchers prepared the document in 8 different languages (Czech, English, German, Italian, Latvian, Slovak, Spanish, and Turkish). This plagiarized documents tested on 15 web-based text-matching tools (Akademia, Copyscape, Docol©, DPV, Dupli Checker, intihal.net, PlagAware, Plagiarism Software, PlagiarismCheck.org, PlagScan, StrikePlagiarism.com, turnitin, Unicheck, Urkund and Viper) using two main criteria (coverage and usability). They discovered that some tools are better suited to certain languages, and that system performance differs depending on the source of the plagiarised content. The performance in synonym substitution is only somewhat adequate, and it is completely terrible for paraphrased and translated texts.In multi-source documents, the systems appear to be better at detecting similarity than in single-source documents. They concluded that certain methods can assist in the detection of some plagiarized content but do not detect all plagiarism and occasionally identified non-plagiarized content for plagiarism.

[6], studied many plagiarism tools and provide information about URL and type of tool (Free/Paid). Study showed that none of the tool is effective in accuracy and efficiency.

3 Overview of Systems

In this section, we discussed about four web-based plagiarism detection tools that are widely used. Data collected here is based on the data provided on their websites.

Dupli Checker [7] is free web-based plagiarism detection tool. This tool is having 1000 words limit per search. It doesn't concentrate on any particular users or objectives. The website also provided paraphrasing tool and reverse image search. On the website, there is no information regarding who runs it.

PlagScan [8], is web based plagiarism detection tool. Only about 1000 words can be checked as a part of free trial. Results are easily understood as Citations, possible plagiarism, and duplicate text are all underlined in the text. The sources are easily visible and available. It shows the result in three colors. Green color is for <1% matching in other document. Yellow is for 1–5% similarity. And red for >5% similarity in other document. It was introduced in 2009 and is run by the German business PlagScan GmbH. They claim to serve more than 1,500 businesses as clients. PlagScan is accessible to single users as well, despite their focus on corporations, high schools, and institutions of higher learning.

Urkund is a web-based application for detecting plagiarism that works at the server side. This is a paid service that uses email credentials. This is an automatic and integrated plagiarism detection method [9]. It was founded in 1999. Regardless of language, Urkund is an automatic text-recognition system designed for identifying, avoiding, and managing plagiarism.

Turnitin [10], is a commercial plagiarism detection tool and used for document analysis. Plagiarism is detected by comparing the document to several web sources and its own database using various methods. The final report links to the probable sources and highlights or colors comparable sentences. Turnitin, which is utilized by 15,000 institutions across 150 countries, was created in 1998 by four students and focuses solely on institutional users.

4 Results and Discussion

For this research, we have used four web-based plagiarism detection tools. Dupli Checker [7] and Plag scan [8] are free to use, whereas Urkund [9] and Turnitin [10] are paid services.

We used an abstract from a study published in 2014 [11], to test the efficiency of plagiarism detection tools. Figure 1 depicts the paper's original abstract, which has not been altered. Original abstract of article supplied to QuillBot AI paraphrasing tool [12], for paraphrased plagiarism abstract. Figure 2 presents a paraphrased version of the paper's abstract.

The original abstract and the paraphrased abstract were initially given to the Dupli checker. Result shows that for copy paste or verbatim, 33% of the text was plagiarised and 67% of the text was unique, from Fig. 3(a), whereas for paraphrase, 0% of the text was plagiarised and 100% of the text was original, from Fig. 3(b). For copy paste and paraphrased, detection efficiency were not satisfactory using this tool.

Abstract: Indian economy is highly depends on agriculture. Agriculture is the main source of income for most of the population. So farmers are always curious about yield prediction. To increase yield production many factors are responsible like soil, weather, rain, fertilizers and pesticides. Now a days Data mining plays an important role in agriculture. The large amounts of data that is available with agriculture universities are mainly restricted to labs and research centers. There is a need to transform this huge data into technologies and make them available to the farmers. It can be possible with data mining. This huge amount of data can be utilized to mine nuggets of knowledge that can be useful for farmers and decision makers to take efficient, effective and prompt decision. In this paper one of the parameter which is used to increase yield production is considered; that is soil. Different classification algorithms are applied to soil data set to predict its fertility. This paper focuses on classification of soil fertility rate using J48, Naïve Bayes, and Random forest algorithm. J48 algorithm gives better result than other algorithms. Decision tree form by J48 algorithm helps the farmer and decision makers to identify the soil fertility rate and on the basis of nutrients found in the soil sample different fertilizers can be recommended

Fig. 1. Original abstract of paper published in 2014 [11]

The report also indicates the percentage of similarity between matching sources. However, there is no information about the plagiarism detection process, such as whether it is based on words, sentences, or other techniques.

Secondly, original abstract and the paraphrased abstract were given to the Plag Scan.

From Fig. 4(a), the result shows that 97.9% of the text was plagiarised for copy paste or verbatim, whereas 11.7% of the material was plagiarised for paraphrase Fig. 4(b). As a result, the detection was effective when the text was a simple copy-paste or literal plagiarism. With paraphrased plagiarism, the detection performance was shown to be lower.

The percentage of similarity between internet sources is also shown in the report. However, no information about plagiarism detection techniques is available.

The results of Ouriginal (Urkund) tool with original abstract and the paraphrased abstract were verified.

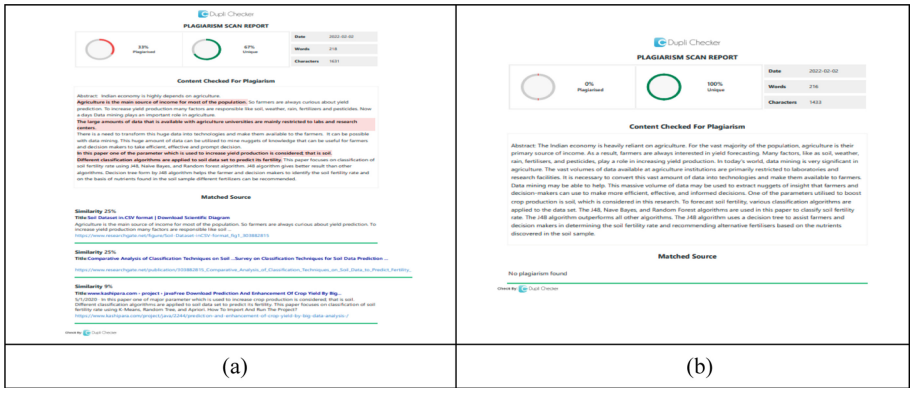
Result shows that for copy paste or verbatim, 100% of the text was similar from Fig. 5(a), whereas for paraphrase, 0% of the text was similar, from Fig. 5(b). This tool is highly good at detecting copy-paste plagiarism, however it isn't good at finding paraphrased content.

Report also shows the side by side comparison of submitted text and matched text. However, no information provided about plagiarism detection techniques.

The results of Turnitin tool with original abstract and the paraphrased abstract were given below.

Abstract: The Indian economy is heavily reliant on agriculture. For the vast majority of the population, agriculture is their primary source of income. As a result, farmers are always interested in yield forecasting. Many factors, like as soil, weather, rain, fertilisers, and pesticides, play a role in increasing yield production. In today's world, data mining is very significant in agriculture. The vast volumes of data available at agriculture institutions are primarily restricted to laboratories and research facilities. It is necessary to convert this vast amount of data into technologies and make them available to farmers. Data mining may be able to help. This massive volume of data may be used to extract nuggets of insight that farmers and decision-makers can use to make more efficient, effective, and informed decisions. One of the parameters utilised to boost crop production is soil, which is considered in this research. To forecast soil fertility, various classification algorithms are applied to the data set. The J48, Nave Bayes, and Random Forest algorithms are used in this paper to classify soil fertility rate. The J48 algorithm outperforms all other algorithms. The J48 algorithm uses a decision tree to assist farmers and decision makers in determining the soil fertility rate and recommending alternative fertilisers based on the nutrients discovered in the soil sample.

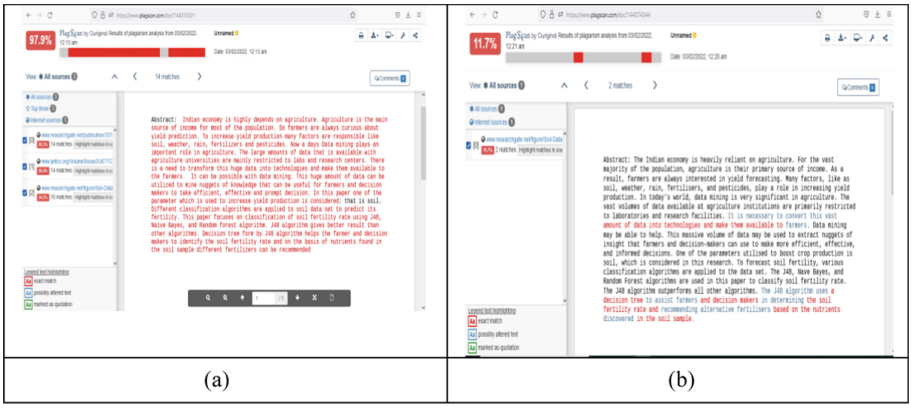
Fig. 2. Paraphrased abstract of paper



(a) (b)

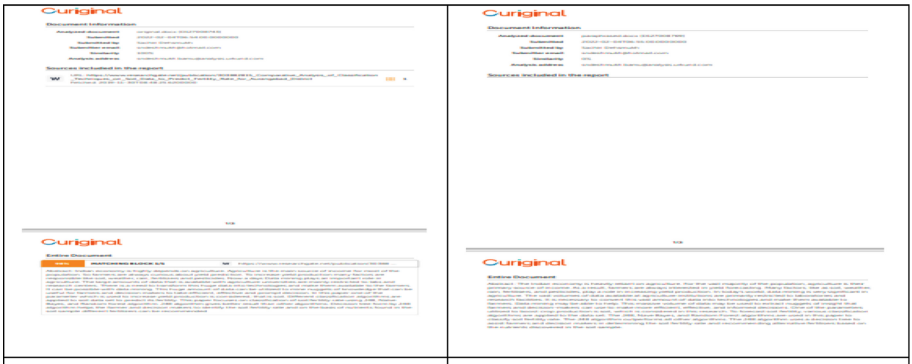
Fig. 3. (a) Dupli Checker result for copy paste abstract. (b) Dupli Checker result for paraphrased abstract

The result reveals that Fig. 6(a) shows a 12% similarity index for copy paste or verbatim, whereas Fig. 6(b) shows a 0% similarity index for paraphrase. If the original document isn't in the database, this tool struggles to detect copy paste plagiarism and is completely incapable of detecting paraphrased plagiarism.



(a) (b)

Fig. 4. (a) Plag Scan result for copy paste abstract. (b) Plag Scan result for paraphrased abstract



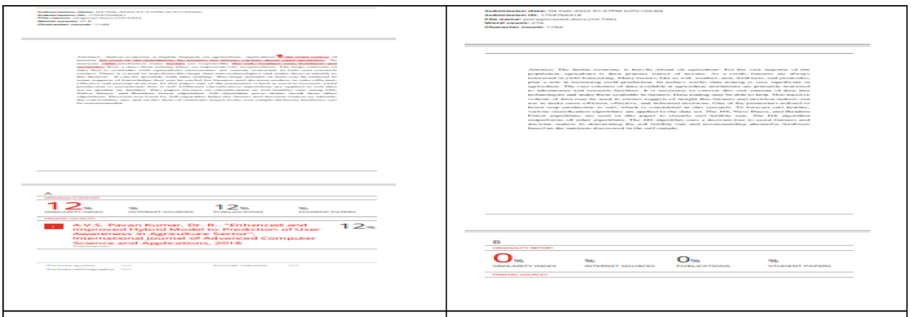
(a) (b)

Fig. 5. (a) Ouriginal (Urkund) result for copy paste abstract. (b) Ouriginal (Urkund) result for paraphrased abstract

Report also shows the side by side comparison of submitted text and matched text. However, no information provided about plagiarism detecting systems.

The Table 1 compares four plagiarism detection tools and their percentages of plagiarism. It has been discovered that Ouriginal (Urkund) and Plag Scan can accurately detect copy paste abstracts. Dupli Checker and Turnitin were both unsuccessful at detecting copy paste abstracts, however none of the four tools were able to detect a paraphrased abstract. As a result, more effective strategies for detecting all types of plagiarism are required.

Also we have performed experiment using urkund and turnitin. Results of 25 papers using Urkund and Turnitin are given in Table 2.



(a) (b)

Fig. 6. A (a) Turnitin result for copy paste abstract. Turnitin result for paraphrased abstract

Table 1. Comparison of different plagiarism detection tools based on similarity

Similarity Detection tools	Plagiarism % for copy paste	Plagiarism % for paraphrased
Dupli Checker	33%	0%
Plag Scan	97.9%	11.7%
Ouriginal (Urkund)	100%	0%
Turnitin	12%	0%

Table 2. Comparison of Urkund and Turnitin based on similarity

Paper no	Plagiarism % using Urkund	Plagiarism % using Turnitin
1	5%	18%
2	9%	22%
3	1%	16%
4	1%	7%
5	0%	11%
6	2%	26%
7	0%	2%
8	1%	7%
9	3%	25%
10	10%	27%
11	1%	16%
12	10%	31%

(*continued*)

Table 2. (*continued*)

Paper no	Plagiarism % using Urkund	Plagiarism % using Turnitin
13	29%	58%
14	17%	50%
15	6%	18%
16	1%	20%
17	6%	91%
18	35%	62%
19	2%	97%
20	6%	23%
21	3%	60%
22	7%	40%
23	38%	26%
24	3%	34%
25	46%	79%

5 Conclusion

In this paper, we have covered a variety of plagiarism-detection tools, including Turnitin, Urkund, Dupli Checker, and Plag Scan. Abstracts that have been copied and pasted, were recognized easily by Plag Scan and Ouriginal (Urkund). While Dupli Checker and Turnitin both failed to detect copy-pasted abstractions. None of the four plagiarism detection tools were capable of detecting a paraphrased passage efficiently. Results depicted the detection discrepancies between Urkund and Turnitin tools. Development of an efficient system is required, which will address each and every problem aimed at improving their accuracy and correctness.

References

1. Fishman, T.: We know it when we see it" is not good enough: toward a standard definition of plagiarism that transcends theft, fraud, and copyright. Proceedings of the Fourth Asia Pacific Conference on Educational Integrity (4APCEI), University. (2009)
2. Hermann Maurer, F. K.: Plagiarism - A Survey. Journal of Universal Computer Science, vol. 12, no. 8, 1050-1084. (2006)
3. Vani K, D. G.: Study on Extrinsic Text Plagiarism Detection Techniques and Tools. Journal of Engineering Science and Technology Review, 9–23. (2016)
4. Bhattacharyya, H. A.: Plagiarism: Taxonomy, Tools and Detection Techniques. arXiv preprint arXiv:1801.06323. (2018)
5. Foltýnek, T. D.-N.-D.-W.: Testing of support tools for plagiarism. International Journal of Educational Technology in Higher Education, 17–46. (2020)

6. M Jiffriya, M. J.: Plagiarism detection tools and techniques: A comprehensive survey. Journal of Science-FAS-SEUSL, 47–64. (2021)
7. DupliChecker Homepage, https://www.duplichecker.com, last accessed 2022/8/10
8. PlagScan Homepage, https://www.plagscan.com, last accessed 2022/8/10
9. Ouriginal Homepage, https://www.ouriginal.com, last accessed 2022/8/10
10. Turnitin Homepage, https://www.turnitin.com, last accessed 2022/8/10
11. Bhuyar, V.: Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District. International Journal of E merging Trends & Technology in Computer Science (I JE TTCS), 200–203. (2014)
12. Quillbot Homepage, https://quillbot.com, last accessed 2022/8/10