JCST Journal of Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766398

A STUDY OF TOOLS AND TECHNIQUES FOR DETECTION AND REMOVAL FOR SOME PLAGIARISM TYPES

Dr. Kiran Mayee Adavala
Professor in CSE, Vidya Jyothi Institute of Technology, Hyderabad, India
kiranmayee@research.iiit.ac.in

Abstract Plagiarism is nowadays all-pervasive. Spectrum 2.0 defines twelve types of plagiarism. Some of them can be detected and removed using software tools. Many such tools are commonplace on internet. One can identify a broad spectrum of features that make them popular. This paper describes some important types of plagiarism. It also focuses on some important features of Plagiarism detection and removal tools that make them popular. This paper also dwells briefly on the need for semantics based plagiarism checks for research papers.

Keywords Plagiarism, Spectrum 2.0, Plagiarism tools, Semantics

1. Introduction

Plagiarism is the use of pre-used text/speech/image/content in the present. The pre-used text could be in oral or written form. It could have been presented by self or by others [1]. It could be online or offline. It could be across regional and language boundaries. Plagiarism is flagged when similarities are above 15% as on the date of this writing. It is more prevalent in academic and research circles. It causes loss to the original, creative user of the text. It is therefore considered illegal with punishments prescribed for it. Every institution of repute has an anti-plagiarism policy and stringent punishment for lapses. Plagiarism can happen in many ways – intentionally or unintentionally. Spectrum 2.0 [3] from Turnitin[2] speaks about twelve different types of plagiarism. We study some of them in the next section.

2. Types of Plagiarism

2.1. Copy-Paste Plagiarism

Many students tend to copy-paste code or text from code geeks and other sites and pass it on as their own without citing the original authors. With the advent of project work as a mandatory part of coursework in Higher Education, this form of plagiarism is on the rise worldwide. It is difficult to detect such cases.

2.2. Contract Cheating

Many-a-times, even for offline assignments, students do not write the assignment themselves. The same is done by a classmate or someone else, which the student passes off as his/her own.

2.3. Data Plagiarism

Data acquisition is a very tough process. Data Sets available on various free sites need to be cited. Many a time, the user of this data does not cite the source from which it was downloaded.

2.4. Accidental or Inadvertent Plagiarism

Sometimes, plagiarism happens without user cognition. The user tries to present an idea in a similar manner as someone before had done, without knowing the same. It is this type of plagiarism that students look for in plagiarism tools as the final step before handing over their assignments.

2.5. Manual Text Modification

Plagiarism text is very meticulous. It looks for the right spellings as well as it looks for plagiarism. Several documents have wrong grammar and spellings or text modified in some other manner. This confuses the plagiarism tools.

2.6. Patchwork or Mosaic Plagiarism

Some authors mix up the sentences of others with their own, thereby creating a mosaic of plagiarized text that is difficult to detect.

2.7. Paraphrase Plagiarism

Many a times, when an author is faced with the dilemma of either citing the source or rewriting content, he/she chooses the later. Thus, we have the same semantics presented in many different ways, all contributing to the same understanding.

2.8. Self-Plagiarism

Many authors have no idea that the content published by them at an earlier date has to be cited before being used again in another context, whereas quoting oneself has no loss.

2.9. Automated text generation

It is possible to rewrite an entire paragraph of text to prevent plagiarism detection. This is a case of plagiarism as well. The author feels that focus on syntax of sentences and not semantics is the prime reason for the use of rewriting and paraphrasing.

2.10. Source-based Plagiarism

Some authors cite sources erroneously, such as, a missing character in the hyperlink, to ensure that the citation is not found. This actually adds some more percentage to their plagiarism report.

2.11. Student Collusion

Students usually work together for common assignments. When the assignment is copied across all students in a study group, it is a case of plagiarism.

3. Plagiarism detection and removal

The various essential and additional features of plagiarism tools [5][7] and comparative study are explained in this section. Saini and Prakash [13] present plagiarism in online learning. Cough [14] explains plagiarism occurrences in programs as well. Baba [23] discusses plagiarism detection through document similarity. Su et al. [15] show how levenshtein distance and smith-waterman algorithm can be used for plagiarism detection. Engels and Craig [16] show how plagiarism can be detected using Neural Networks. Suleiman et al. [22] also use deep learning for plagiarism detection, whereas AlSallal [24] uses a slightly different machine learning approach and Kong et al. uses logical regression modeling. Jain et Al. [21] use bitwise operations for plagiarism detection. Scherbinin and Butakov [18] show how Microsoft SQL server platform can be harnessed towards plagiarism detection. Barnbaum [19] presents a simple guide to recognizing plagiarism. A corpus for plagiarism is described by Motaj et al. [20]. Many papers present comparative studies of plagiarism tools [5][7][26].

4. Features of Some Existing Plagiarism tools

The various essential and additional features of plagiarism tools are explained in this section with examples.

4.1. Turnaround time

The time taken for a plagiarism tool to report its findings once the input is given is its turnaround time. Many tools have different turnaround times for free and paid subscriptions. The plagiarism checker by SmallSEO Tools [6] gives results in 0.83 seconds for 1,000 words per search.

4.2. Number of Words Allowed

The maximum count of words that are allowed for free and paid usage of the tools is limited. There is also a time-based restriction on the use of the tool. For example, SmallSEO allows 1000 words in its free version.

4.3. Checker or Remover

Plagiarism tools have some specializations. Some of these tools check for plagiarism (ithenticate) and others help in removing plagiarism in one or both of two ways – by presenting the list of matching sources to the user (Quillbot) and by replacing the plagiarized content with paraphrased ones (paraphraser).

4.4. Algorithm used

Plagiarism detection and removal is a subject belonging to the Artificial Intelligence - Natural Language Processing realm. There are several methods used by tools to determine if the content presented matches with content in its database or its crawled data. They are string tiling, Karp-Rabin algorithm, Haeckel's algorithm, k-grams, string matching algorithm. There are also others that refine existing algorithm with the use of use Bag-of-Words, Edit Distances, Levenshtein Distances and Smith-Waterman algorithm. Mostly, there is one or more of these Natural Language Processing algorithms at play. Most tools do not share their algorithm or code on the internet.

4.5. Number of documents compared with

Popular tools store millions of documents in their database, mostly from the academic community, and perform all comparisons with this ever-growing document-base. Ithenticate, turnitin’s tool, checks 89.4 million journals and counting at the time of this writing.

4.6. Multi-language Support

Plagiarism does not have language boundaries. It may happen in any language. Many Plagiarism tools consider material supplied only in English. There are others that provide multi-language support. Also, many tools provide multi-language user interface support as well. Copymatic [9] works with 25+ languages.

4.7. Operating System Support

Many diverse operating systems are installed on computers and software may or may not be compliant to every one of them. Some tools, however, have installation support for many operating systems. Ithenticate is compatible to most operating systems.

4.8. Support as a Plug-In/Extension

Plagiarism tools are more usable when they are connected to the source of the text. These could be word processing applications, applications for creation of material to be posted on internet and so on. Grammarly provides extension for Browser, MS Office, Desktop and Google Docs.

4.9. Safety of Submitted Data

All Plagiarism tools use internet for matching given text. There is every possibility of data being breached in transit. Similarly, many of these tools update themselves almost immediately with the text being checked, so that, if student does not make a submission, it becomes 100%

plagiarism the next time around. Thus, the policy of the tool owners storing user data is also under public scrutiny.

4.10. Cloud Usage

The use of cloud storage and implementation of plagiarism check as an Internet service under SaaS is nowadays popular with most plagiarism tools. SmallSEO supports Dropbox and Google Drive.

4.11. Document Format support

A document presented to a plagiarism tool may be a document in .docx, .pdf, .txt or some other format. Some tools permit all these file formats, whereas, other tools do not. Quillbot [28] permits txt, docx and html formats.

4.12. Document Input Type

Text is usually presented to plagiarism tools in one of three ways – as ordinary text that is pasted in an online text box, as a file upload or as a url. Paraphraser [12] permits entry of upload url.

4.13. Reporting option

Once plagiarism is found, the details of sentences or paragraphs containing them are to be presented to the user. Similarly, once plagiarism is corrected, the new plagiarism report has to be generated and presented in a similar fashion. Additionally, Easybib provides suggestions to improve grammar and style.

4.14. Sharing option

If a student wants to pre-op share his/her plagiarism report with a mentor, there has to be provision for this. The same also applies to a mentor sharing a report of a plagiarized paper with his/her student. This is a user-friendly option for ease of report and comment sharing. Most tools have this option.

4.15. User Privileges

Most plagiarism tools have user logins where users input their documents. It is important that they know beforehand if every document being submitted is being stored in database for future matching and also, to decide when this can be done. The timeline for doing so would vary from one user to another. For example, for a researcher, the timeline would be the publication of the document in a journal. For a High School teacher, the document would have to be stored and brought to use immediately to prevent peer plagiarism of the assignment.

4.16. Scope of Tool

Plagiarism tools can perform either or both of detection and removal tasks. The flow of control in a typical plagiarism checker-remover tool is illustrated in figure 1. Some tools perform removal by paraphrasing. Others show sources list and allow user to cite the sources. Some tools have both possibilities. There are a third category of tools that do not detect plagiarism, but paraphrase given text. For example, SEOMagnifier is an AI-based Paraphrasing tool. SmallSEO is another such tool used mostly by software developers.

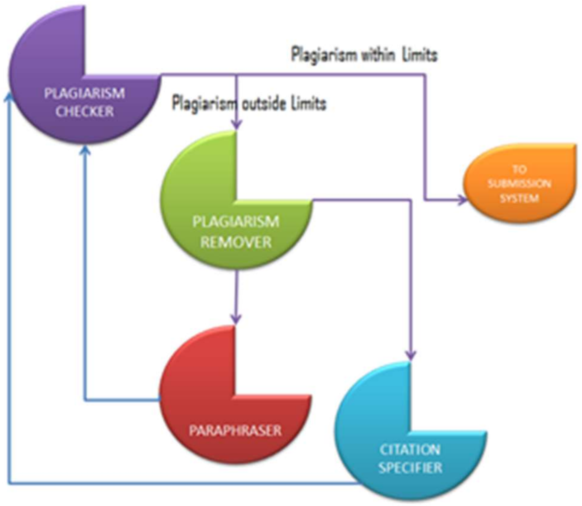

Figure1. Flow of Control in a Plagiarism Checker + Remover Tool

4.17. Depth of Search

Search performed by plagiarism can be superficial (sentence-level) or it can be an indepth phrasal search. Quetext uses DeepSearch technology. Also, tools like plagiarism detector.net have reverse image search option as well.

4.18. Option to exclude specific URLs

Search performed by plagiarism can be superficial (sentence-level) or it can be an indepth phrasal search. PrepostSEO [11] gives the option to exclude URLs.

4.19. API Support

Most tools have APIs with user-friendly GUIs for data/file/URL entry and display of results. Copymatic [9] provides API access. Another related topic is that of uptime – tools such as Unicheck have 99% uptime.

4.20. Exclusion of Quotes

Many plagiarism checkers give users the option of removing all quoted text for a more practical plagiarism check. PrepostSEO [11] provides quote exclusion option.

4.21. Method of Plagiarism Removal

Almost all tools today present the sources of plagiarism. There are advanced tools that also automatically cite all such sources and remove plagiarism. Quetext [27] has a citation assistant as well as a citation generator.

4.22. Provision for Human Expert Guidance for Plagiarism Removal

A few plagiarism tools also offer online human expert support for plagiarism reduction. Easybib allows submission of papers for a human expert check and feedback, with a turn-around time of 24 hours.

5. Bases for Plagiarism Detection

There are several methodologies employed for detecting Plagiarism. Some of them are purely statistical or syntactic. Very few of them focus on meaning or semantics of content.

5.1. Characters and Words

Here, a fixed number of characters and words are matched between a document A and a database of documents. The number of words used for matching is called n-gram, for example, if four words are matched, 4-gram. Simple string matching algorithm is used. Similarly, vector measures may also be used for similarity computation.

5.2. Syntax

Parts-of-Speech, simple grammar metrics are used in conjunction with words to detect plagiarism. This enables check of exact phrases – Noun Phrase, Verb Phrase etc. in existing documents.

5.3. Semantics

Semantic similarity between words used in two sentences is computed to find plagiarism. Sentences rewritten in active or passive voice and synonym replacements in sentences are detected. Word-similarity is looked up from existing Knowledgebases containing words, their synonyms and the similarity measure.

5.4. Semantics

Semantic similarity between words used in two sentences is computed to find plagiarism. Sentences rewritten in active or passive voice and synonym replacements in sentences are detected. Word-similarity is looked up from existing Knowledgebase containing words, their synonyms and the similarity measure. The same approach may also be applied to whole sentences as well as translated sentences.

6. Need for Deep-Semantics Based Plagiarism Check

The current algorithm used for plagiarism check focus mainly on the syntax of sentences. The better tools also look for sentences that are almost similar. For scientific papers, what is required is focus on the meaning of what is being said rather than the content itself. A researcher and a student cannot be evaluated for plagiarism with the same set of rules. Students doing assignments have smaller content that can probably be written in not more than, say, 10 different ways. A researcher cannot be permitted to write the same content in another way and get published. Most of the tools that remove plagiarism are focused on paraphrasing and assisting a researcher in doing exactly this. Thus, research is reduced to an exercise in English (or other) language usage. There has to be a more realistic approach to determining if a researcher has copied content and paraphrased it by examining the meaning of research content using intelligent algorithm and tools.

In the context of inter-language translation, the most stable solution is the creation of a Universal language that every other language can be translated to for ease of pattern matching towards plagiarism detection.

6. Conclusion

This paper presents some of the plagiarism types being actively discussed in academic circles today. It dwells on some of the features of plagiarism tools, methodologies and examples of tools that implement them. It also stresses on the need for semantics based plagiarism checks.

REFERENCES

[1] About Plagiarism, Online available from https://plagiarism.org/

[2] About Turnitin, Online available from https://en.wikipedia.org/wiki/Turnitin

[3] Plagiarism spectrum 2.0, Online available from https://www.turnitin.com/resources/plagiarism-spectrum-2-0

[4] Plagiarism Checker Free, Online available from https://www.studyinternational.com/news/plagiarism-checker-free/

[5] 5 Plagiarism remover tools for students, Online available from https://www.studyinternational.com/news/5-free-plagiarism-remover-students/

[6] SmallSEO Tools, Online available from https://smallseotools.com/plagiarism-checker/

[7] Top 10 Plagiarism tools for teachers, Online available from https://elearningindustry.com/top-10-free-plagiarism-detection-tools-for-teachers

[8] SEOMagnifier Tool, Online available from https://seomagnifier.com

[9] Copymatic Tool, Online available from https://copymatic.ai

[10] Rewriter Tools, Online available from https://rewritertools.com/article-rewriter

[11] PrepostSEO Tool, Online available from https://www.prepostseo.com/article-rewriter

[12] Paraphraser Tool, Online available from https://www.paraphraser.io/article-rewriter

[13] Saini D.K., Prakash L.S. Plagiarism Detection in Web based Learning Management Systems and Intellectual Property Rights in the Academic Environment, International Journal of Computer Applications, Vol.57, No.14, 6 – 11, 2012.

[14] P. Clough. Plagiarism in natural and programming languages: an overview of current tools and technologies. 2000.

[15] B. R. A. Z. Su, K. Y. Eom, M.-K. Kang, J.-P. Kim,and M.-K. Kim. Plagiarism detection using the levenshtein distance and smith-waterman algorithm. 3rd International Conference on Innovative Computing Information and Control, Washington DC, USA, 569, 2008.

[16] V. L. S. Engels, and M. Craig. Plagiarism detection using feature-based neural networks. Thirty-Eighth SIGCSE Technical Symposium on Computer Science Education, Covington, Kentucky, 34-38, 2007.

[17] Z. Su, B. R. Ahn, K.Y. Eom, M. K. Kang, J. P. Kim, M. K. Kim. Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm. 3rd IEEE International Conference on Innovative Computing Information And Control, 2008.

[18] V. Scherbinin and S. Butakov. Using Microsoft SQL server platform for plagiarism detection. PAN'09, 36-37, 2009.

[19] C. Barnbaum. Plagiarism: A Student’s Guide to Recognizing It and Avoiding It. 2002. Online available at http://www.cpalms.org/Public/PreviewResourceUrl/Preview/25915

[20] S. Mohtaj, H. Asghari, V. Zarrabi. Compiling a text re-use detection corpus from scientific papers with semi-real cases of plagiarism. IEEE International Conference In Asian Language Processing, 227-230, 201.

[21] S. Jain, P. Kaur, M. Goyal, G. Dhanalekshmi. CPLAG: Efficient plagiarism detection using bitwise operations, Tenth IEEE International Conference on Contemporary Computing (IC3), 1-5, 2017.

[22] D. Suleiman, A. Awajan, N. Al-Madi. Deep Learning Based Technique for Plagiarism Detection in Arabic Texts. IEEE International Conference on New Trends in Computing Sciences, 216-222, 2017.

[23] Baba K. Fast plagiarism detection based on simple document similarity. IEEE Twelfth International Conference on Digital Information Management, 54-58, 2017.

[24] AlSallal M, Iqbal R, Amin S, James A, Palade V. An Integrated Machine Learning Approach for Extrinsic Plagiarism Detection, 9th International IEEE Conference on Developments in eSystems Engineering (DeSE), 203-208, 2016.

[25] Kong L, Lu Z, Qi H, Han Z. High obfuscation plagiarism detection using multi-feature fusion based on Logical Regression model, 4th International IEEE Conference on Computer Science and Network Technology, 355-359, 2015.

[26] V. Sn´aˇsel, J. Pokorn´y, K. Richta, (Eds.): Dateso, Overview and Comparison of Plagiarism Detection Tools, 161–172, ISBN 978-80-248-2391-1, 2011.

[27] Chowdhury A. Hussain, Bhattacharyya K. Dhruba. Detection - Plagiarism: Taxonomy, Tools and Detection Techniques. Dept. of CSE, Tezpur University

[28] Quetext. Online available from https://www.quetext.com/

[29] Quillbot. Online available at https://www.quillbot.com/

[30] Easybib. Online available at https://www.easybib.com/

[31] Unicheck. Online available at https://www.unicheck.com/