AN IMPROVED TECHNIQUE FOR ANALYZING SIMILARITY INDEX USING TURNITIN

MUSA, M.A.[1]*, SOULEY, B.[1], ZAMBUK, F.U.[1] AND KABIR, R.K.[2]
[1]Department of Mathematical Sciences, Abubakar Tafawa Balewa University, Bauchi, Nigeria. [2]Computer Science Department, College of Education Zuba, Abuja, Nigeria

ABSTRACT

The detection of plagiarism by software in the checking of a student plagiarized document in higher institutions of learning (universities and colleges) plays a crucial role in preventing plagiarism amongst students. However, plagiarism detection services show a variation by violating the intellectual property rights of the students during the detection process. A control framework was developed, according to the legal requirement that protects intellectual property to protect the integrity of authors in the course of the plagiarism detection. Students' documents submitted to their lecturers were pre-processed. The framework filtered at least two letter words using python programming language and subjected and to see the extent of locating plagiarized sources in respect of the original thesis. An accuracy of about 70% was obtained from results obtained using Turnitin. The control framework prevented the violation of intellectual property by the use of the pre-processed thesis during the plagiarism detection process by a third party such as Turnitin. There is an encouragement in academic skilled writing and beyond in academia as every undertaking is the property of the owner. The control framework is, therefore, a preventive marker for the violation of intellectual property.

Keywords: Intellectual Property, Plagiarism, Thesis, Turnitin, Violation,
***Correspondence:** ammaryam@atbu.edu.ng, 08034984685

INTRODUCTION

In the new global economy, internet technologies have simplified sharing any information. Extremely notable is a thesis and research reports of students or academics. There is evidence that protecting intellectual property plays a crucial role in regulating plagiarism, which is one of the main challenges faced by students and academics. The violation of intellectual property (IP) is present in many educational institutions, from colleges to universities.

In today's, IP refers to creations of the mind: inventions; literary and artistic works; and symbols, names, and images used in commerce. IP is divided into two:

1. Industrial property includes patents for inventions, trademarks, industrial designs and geographical indications.
2. Copyright covers literary works (such as novels, poems and play), films, music, and artistic works (e.g. drawings, paintings, photographs, and sculptures).

IP rights are like any other property right. They allow the creators or the owners of patents, trademarks or copyrighted works benefit from their work or investments in creation. These rights are outlined in Article 27 of the universal declaration of human rights, which provides for the right to benefit from the protection of moral and material interest resulting from authorship of scientific, literary and artistic productions. The need to protect IP are that, the progress and well-being of humanity rest on its capacity to create and invent new works in the areas of technology and culture. Also, the legal protection of new creations encourages the commitment of additional resources to further innovation, and lastly the protection of IP spurs economic growth, creates new jobs and industries, and enhances the quality and enjoyment of life.

Interestingly, there is no copyright in ideas; the creativity protected by copyright law is creativity the choice of arrangement of words, musical notes, and shape. However, copying parts of another creator's work, or its arrangement or structure, might be considered plagiarism.

Plagiarism is a culpable activity that occurs at many different levels, from the student who incorporates pages from an article in his thesis/assignment without acknowledgement, to the scientist who eases to make of use his colleague's test results and publishes them under his name, or the author whose novel is a reworking of an obscure folktale.

Strictly speaking, plagiarism is when the perpetrator poses himself off as the originator of a work, whereas he is not. As for the real author, his copyright has been infringed and so has his moral right of paternity, the right to be identified as the author of the work.

It has been established that there are two types of plagiarism namely textual and source code that occur in educational and research areas [1] . Observations of plagiarism in practice reveal a number of commonly found methods for illegitimate text usage [2], the copy and paste plagiarism specifies the act of taking over parts or the entirety of a text verbatim from another author. Disguised plagiarism includes practices intended to mask copied segments. Undue paraphrasing defines the intentional rewriting of foreign thoughts in the vocabulary and style of the plagiarist without giving due credit to conceal the source [3]. Translated plagiarism is the manual or automated conversion of content from one language to another intended to cover its origin. Idea plagiarism encompasses the usage of a broader foreign concept without appropriate source acknowledgement. Existing text plagiarism detection system can be categorized into external and intrinsic [4].

Intrinsic plagiarism detection system statistically examines linguistic features of a suspicious text, a process known as stylometry, without performing comparisons to external documents while the external plagiarism detection system compares a suspicious document to a corpus of genuine works .Different comparison strategies have been proposed for external plagiarism detection system [5]. The most common ones are: Sub-string matching. This procedure is aimed at identifying long pairs of identical strings. Such strings are treated as indicators of potential. Plagiarism if they share concerning the entire text exceeds a chosen threshold. Most commonly suffix document models such as suffix trees or trays, have been used. In the last several years, plagiarism in colleges and universities particularly the internet or cut and paste plagiarism has increased in frequency [2].

In the United States, a 2003 Rutgers University study of 18000 students, 2600 faculty and 650 teaching assistants on 23 campuses found that “thirty eight percent of the undergraduate students completing the survey indicated they had engaged in one or more instances of cut and paste plagiarism using the internet in the past year, paraphrasing or copying a few sentences of material from the internet without citing the source [3]. This is a dramatic increase from the 10% who acknowledged ‘cut and paste’ plagiarism using the internet in a similar survey conducted only two years ago”. The new study, therefore, confirms internet plagiarism as prevalent.

In the past, programmes that promote education and the honor code as well as post plagiarism detection have been utilized to combat plagiarism. Detection, if done, can be a very labour intensive process and may be impossible to conduct within time and personnel resource constraints. The advent of plagiarism detection technology is a great benefit for institutions and lecturers that can effectively utilize the technologies.

There are some research studies that deal with detecting duplicated material available on the internet. This study has evolved from earlier studies, examining plagiarism detection in source code [6]. Accordingly, Al Jarrah *et al.* [1] based their study on the correlation between author, title, and content. They assumed plagiarism can be detected through the automation of passing text to search engines. Another study by Clough [7] laid emphasis on plagiarism detection in text documents by inspecting the suspicious documents using grammatical structures that authors use to build sentences and find inconsistences in syntax. Culwin and Lancaster [8] proposed a copy detection service that identifies partial or complete overlap of documents. A prototype of the service was implemented and experimental results suggested the service can indeed detect violation of interest. A considerable number of studies have been done by researchers on plagiarism detection local or across the globe. It is becoming a common practice to use software and services that automate the processes of plagiarism detection. The majority of these applications are based on the document source comparison [1]. The detection process starts with the submission of the suspected document to the system via a desktop application or web-based form, and a report is presented highlighting the matched sources and copy percentage. The ease with which such documents are accessed account for the authors violation of copyright [5]. The plagiarism detection services compromise the original documents, making a profit with them since they archive such documents.

In an attempt to fulfill this requirement in plagiarism detection systems, a preprocessing platform is included in the architecture for plagiarism detection. The proposed service available to University aims at recasting the suspected plagiarized document by removing at least two letter wordings prior submission to plagiarism detection services. The system makes use of off the shelf tools (web services) to extend plagiarism detection.

Plagiarism detection

First, Plagiarism detection has, of course, existed for as long as plagiarism itself, and many tutors would stress that they have long been adept at using their own low-tech, but highly intuitive methods to spot plagiarism in student essays. Automation in plagiarism detection has emerged more recently, with earlier research in the field focusing largely on detecting plagiarism in computer programs while recent years have seen considerable developments in the online detection of text-based plagiarism. Most automated plagiarism detection services’ aims are twofold to highlight possible plagiarism, and also to identify the potential source of the plagiarized paper. Hence, as El-Alfy *et al.* [9] writes, ‘the plagiarism detection task is different from authorship attribution but deeper than information retrieval.' The principles and practices behind the automated services vary considerably, but the overall strategy tends to remain the same [10].

Plagiarism detection services

Most textual plagiarism detection tools operate by using a variety of submission and search techniques, whereby the content of submitted essays is checked against various sources, such as web sites, paper mills, essay banks and other assignments uploaded to the service. Most use search engine technology to identify similarities between sections of the submitted text and web sites, usually looking for overlaps between strings of text. This is based on the premise that the two writers are unlikely to use the same sequence of words above and beyond a certain phrase length. The output of these services is usually in the form of a report, often using colour coding and hypertext links to enable the end-user to home in on both potentially plagiarized text in the submission, and also the possible Internet source.

Some of the well-known products in this field include the web-based service, Turnitin.com produced by iParadigms and used by various organizations, such as the Joint Information Systems Commission (JISC) Plagiarism Detection Service [6] as well as the downloadable Essay Verification Engine (EVE) [7]. As there is quite a degree of overlap, most products try to promote particular enhancements which differentiate their goods from the rest. For example, [8] proposed detecting slight linguistic modifications, such as a change of verb, as well as verbatim copying, and it also converts all submitted documents to PDF before running them through the plagiarism detection service to avoid problems of format incompatibility. My Dropbox also claims that it is not limited to uncovering verbatim copying, by utilizing innovative artificial intelligence module. Like Turnitin, it also broadens its coverage beyond the merely visible web by searching password-protected databases of journal articles, and other assignments submitted to the service.

Turnitin and the issues of IP

Since students frequently don’t allow for the copying of their work, Turnitin is the best known, and one of the longest running of today’s commercially available plagiarism detection services for archiving digital resources. The service had to search internet sites, and paper mills and the service could not maintain submitted papers in an internal database. Keeping such papers is a violation of students intellectual property rights because students neither agree that their course papers are accessible to anyone on the internet nor, even more disturbing, consent that their papers may be used for profit [12].

In a recent dispute by high students of Virginia against Turnitin, the service did not have the right to archive their intellectual property without permission. Turnitin legal representatives reportedly claim that its archival practices, saving all submitted papers to be screened for matches against future submissions fall under the “fair use” designation of legal reproduction of material for educational purposes. The students have disagreed, noting that the addition of their work to the database serves the purpose of a corporation’s monetary gain.

Moreover, in their study of plagiarism detection services, Mckeever [13] indicated that the University of Illinois at Urbana would not allow them to test services that kept copies of papers precisely because “students’ essays are their property. In this way, the university itself has taken a stand against services that retain student papers. Purdy [14] addressed how some college lawyers now advise institutions that plagiarism detection services that maintain copies of submitted student papers, specifically Turnitin, not only potentially violate students’ copyrights on their written work, but also violate privacy.

These studies conclude that the services sometimes used to ensure the integrity of students’ texts can be of questionable integrity, largely through the design of their archives. Therefore, measures are expected to be taken to tackle these problems.

METHODOLOGY

According to the world intellectual property organization, intellectual property rights are like any other property right. They allow creators or owners of patents, trademarks or copyrighted works to benefit from their work or investments in creation. These rights are outlined in Article 27 of the Universal Declaration of Human Rights, which provides for the right to benefit from the protection of moral and material interest resulting from authorship of scientific, literary and artistic productions. The importance of intellectual property was first recognized in Paris convention for the protection of industrial property (1883) and the Berne Convention for the protection of literary and artistic works (1886). Both treaties were administered by the world intellectual property organization.

The PDS architecture for papers submitted is very straight forward [15]. A school maintains a database of all students’ works and compares each new document with existing ones upon submission and the students’ submissions will remain as digital files in the school database. Performing search requires a web crawler which prohibits a university or college to maintain and therefore causes them to outsource to a company (e.g. Turnitin.com) that specializes in plagiarism detection so that a suspected plagiarized paper can be compared against all possible sources on the web. Outsourcing can be done in two ways: outsourcing the whole process or outsourcing the most data intensive part of it. The figures 1 and 2 shows the two processes. (e.g. Turnitin.com) That specializes in plagiarism detection so that a suspected plagiarized paper can be compared against all possible sources on the web. Outsourcing can be done in two ways: outsourcing the whole process or outsourcing the most data intensive part of it.

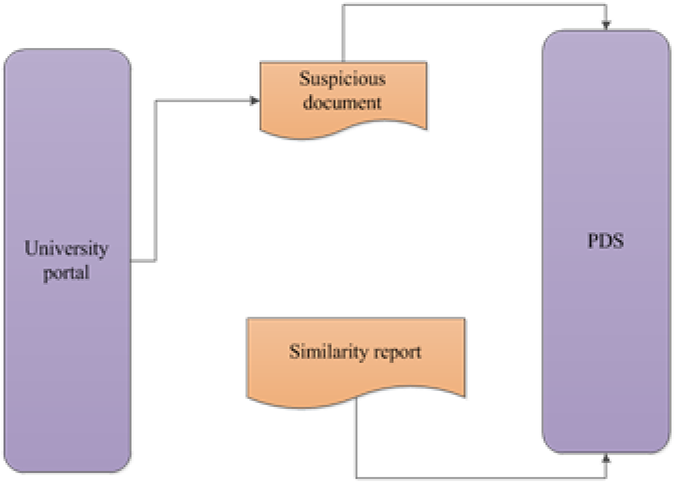

Figure 1: Documents presentation to Plagiarism Software

RESULTS

Table 1: Results Obtained from subjecting thesis/assignment to Turnitin

Thesis/ Assignment	Similarity Index (%)
Original Thesis	25
Modified Thesis	15

A similarity index of 25% and 15% was obtained from the Turnitin software, we deduce that the latter and former document have similarity content of 85%, signifying the modified document can equally serve in checking similarity reports. For more intuitive explanation this is presented in Figure 2.

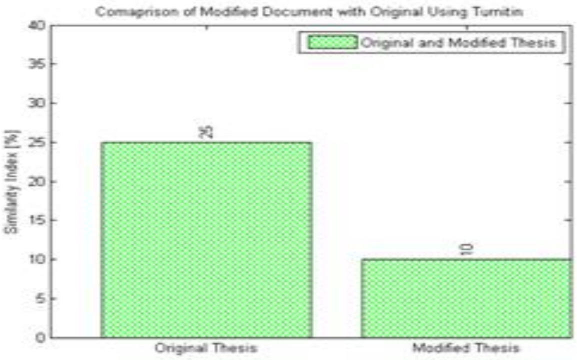

Figure 2: Original and Modified Thesis Plagiarism Index

DISCUSSION

A similarity index of 25% was generated when the original and modified thesis were subjected to the Turnitin.com respectively. A difference of 15% in their contents showed that they have 85% similarity in their contents as shown in Figure 2. To this end, study therefore, the modified thesis can be used as a substitute to the original thesis during the process of plagiarism detection.

The intellectual property management and control framework application preprocessed the students' thesis prior subjecting it to the plagiarism detection software (Turnitin.com). The difference in the plagiarism index obtained when both the original and modified thesis was submitted to the Turnitin.com was negligible. In the light of these our findings, we culminate that the preprocessed thesis can take care of the original thesis in the course of plagiarism detection. As Butakov and Barber [5] investigated, hiding some contents of a document is also protecting such a document from violation of intellectual property. This study therefore has protected the students' intellectual property during the process of plagiarism detection.

FURTHER RESEARCH

This research work filtered two at least two words and used the modified thesis for plagiarism detection using the Turnitin.com software. Our study however protected the intellectual property during the process of plagiarism detection. In future, improvement on this research may entail modification using larger words to find no or negligible difference in the plagiarism index or content similarity machine learning algorithms for more accurate results and to reduce computational cost. possible.

REFERENCES

1. AL JARRAH, A., I. ALSMADI, & Z. ZA'ATREH (2011). Plagiarism Detection Based on Studying Correlation between Author, Title, and Content. International Conference on Information Communication System (CICS).
2. BEASLEY, J.D. (2004). The Impact of Technology on Plagiarism Prevention and Detection: Research Process Automation, a New Approach for Prevention. Plagiarism: Prevention, Practice and Policies, p. 28-30.
3. BRAUMOELLER, B.F. & B.J. GAINES (2001). Actions Do Speak Louder Than Words: Deterring Plagiarism with the Use of Plagiarism-Detection Software. PS: *Political Science & Politics*, **34**(4): p. 835-839.
4. BRIN, S., J. DAVIS, & H. GARCIA-MOLINA. (1995). Copy Detection Mechanisms for Digital Documents. in Proceedings of the 1995 ACM SIGMOD international conference on Management of data.
5. BUTAKOV, S. & C. BARBER (2012), Protecting Student Intellectual Property in Plagiarism Detection Process. *British Journal of Educational Technology*, **43**(4).
6. CLOUGH, P. (2012). Plagiarism in Natural and Programming Languages: An Overview of Current Tools and Technologies. 2000.
7. CLOUGH, P. (2003). Old and New Challenges in Automatic Plagiarism Detection. *National Uk Plagiarism Advisory Service*.
8. CULWIN, F. & T. LANCASTER (2000). A Review of Electronic Services for Plagiarism Detection in Student Submissions. in LTSN-ICS 1st Annual Conference.
9. EL-ALFY, E.-S.M., et al. (2015). Boosting Paraphrase Detection through Textual Similarity Metrics with Abductive Networks. *Applied Soft Computing*, **26**: 444-453.
10. FOSTER, A.L. (2002). Plagiarism-Detection Tool Creates Legal Quandary. *Chronicle of Higher Education*, **48**(36): p. A37-38.
11. GIPP, B., N. MEUSCHKE, & J. BEEL (2011). *C*omparative Evaluation of Text-and Citation-Based Plagiarism Detection Approaches Using Guttenplag. in Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries.
12. LUKE, D., et al. (2014). Software Plagiarism Detection Techniques: A Comparative Study.
13. MCKEEVER, L. (2006). Online Plagiarism Detection Services—Saviour or Scourge? *Assessment & Evaluation in Higher Education*, **31**(2): p. 155-165.
14. PURDY, J.P. (2009). Anxiety and the Archive: Understanding Plagiarism Detection Services as Digital Archives. *Computers and Composition*, **26**(2): p. 65-77.
15. TSCHUGGNALL, M. & G. SPECHT (2013). Detecting Plagiarism in Text Documents through Grammar-Analysis of Authors. Datenbanksysteme für Business, Technologie und Web (BTW) 2028.