https://dx.doi.org/10.4314/dujopas.v9i3a.31
ISSN (Print): 2476-8316
ISSN (Online): 2635-3490
Dutse Journal of Pure and Applied Sciences (DUJOPAS), *Vol. 9 No. 3a September 2023*

Automatic Plagiarism Detection Using Fuzzy-Logic

Adedayo Sobowale, Adebimpe Esan, Adebiyi Tomilayo,
Bunmi Jooda and Adekemi Bolajoko

Department of Computer Engineering,
Federal University Oye-Ekiti,
Ekiti state,
Nigeria

Email: adebimpe.esan@fuoye.edu.ng

Abstract
Plagiarism occurs when a researcher copies fellow researcher's work verbatim without acknowledging the author. This work developed an automatic plagiarism detector using fuzzy logic. The system developed was tested with 4 different text documents and evaluated using portability, efficiency, functionality, ease of use and accuracy metrics. Results show that the developed plagiarism detector is very easy to use with high functionality and accuracy as well as moderate efficiency and portability based on user's assessment. However, future work can increase the data size for model training and consider machine learning techniques to improve accuracy.

INTRODUCTION
Plagiarism entails imitating the thoughts of other researchers without acknowledging them (Wadsworth, 2004; Knight et al., 2004). It is done by paraphrased works and the similarities between keywords and verbatim overlaps and change of sentences from one form to another (Ozlem *et al*., 2005; Abdelmalek *et al*., 2010). Plagiarism is common in an academic setting and it can threaten the foundation of knowledge as well as infringe on author's future rights. In addition, plagiarism harms educational establishments, hence the need for plagiarism detectors to detect plagiarized works (Benno, 2011; Antonio 1997). Manual plagiarism detection was formally used in plagiarism detection but it is time consuming and expensive, hence the introduction of automatic plagiarism detection. Automatic text plagiarism detection is the process of checking and detecting text document(s) being copied from original works using computer program or plagiarism detection software (Potthast 2009**)**. It is faster, less time consuming, more efficient and accurate than the traditional method.

Automatic Plagiarism detection can be done using two major techniques, which are: the Intrinsic Technique and the Extrinsic Technique. The intrinsic technique uses stylometric features from the text to detect passages that are different from the rest of the passages while the Extrinsic technique finds those that are similar to the passage in an external corpus (Snijders et al., 2012). Previous approaches for automatic plagiarism detection include: TF/IDF (Yuuki Mori *et al.,* 2015), Cosine Similarity (Umezawa *et al.,* 2011; Huang, 2008), N-gram based methods (Barr´on-Cede˜no *et al.,* 2009), (Kuta *et al.,* 2014) and (Stamatatos, 2011). TF/IDF and cosine similarity does not consider word order because it is based on set theory. There are also methods of detecting similarity using editing cost (Ueta *et al.,* 2010) but these also need a large amount of calculation. In this research, a plagiarism detector was developed using fuzzy logic-based approach.

**Author for Correspondence*

Ďuračíka *et al.,* (2017) and Agrawal *et al.,* (2016) developed source code plagiarism detection using Running-Karp-Rabin and Greedy-String-Telling (RKR-GST) algorithms. Nathaniel *et al*., (2008) develop a sentence-based plagiarism detector and research show that the system recorded good accuracy. Osman et al., (2012) and Gupta et al (2014) proposed an approach based on a Fuzzy Inference System and Semantic Role Labeling (FIS-SRL) for plagiarism detection. Hermann *et al*., (2006) developed plagiarism detection system using statistical approach. Francisco *et al*., (2008) used four similarity criteria to measure the similarities between two documents and the accuracies were recorded. Alzahrani and Salim (2010) proposed a semantic plagiarism detection technique using fuzzy semantic-based string similarity. The research include: pre-processing of the dataset, retrieval of candidate documents using Jaccard coefficient and shingling algorithm, comparison of suspicious documents to candidate documents and post-processing to join consecutive sentences.

Chow and Salim (2010), proposed a semantic-based plagiarism detection where the similarity between the suspected and original documents were calculated according to the predicates of the sentences. The shortcoming of this approach is that it is limited to specific parts of the sentence. Kundu and Kartik (2017) employed Latent Semantic Analysis (LSA) for word-based plagiarism detection. Alireza *et al.,* (2017) proposed a novel technique for plagiarism detection using PersianPlagDet dataset. Autade *et.al.,* (2017) employed a Multi-agent approach to detect plagiarism. The dataset used in training the model was pre-processed to remove unwanted data and the correlation similarity was calculated thereafter. Zhao *et al.,* (2015) proposed abstract syntax tree (AST) for plagiarism detection while Sediyono *et al.,* (2008) proposed longest common consecutive series (LCCW) algorithm for plagiarism detection. When compared to suffix tree algorithm, LCCW algorithm was found to perform better.

All previous related works are mainly used to detect plagiarism in online available sources, this research detected plagiarism in student's assignment. This means the system developed is capable of identifying students who copied each other.

METHODOLOGY

The system consists of the following Stages: data collection, data pre-processing, and similarity analysis. The corpus for building the fuzzy logic model was collected from a higher institution in Nigeria. Each student is assigned to write an assignment on various topics. A set of 30 students from computer engineering and civil engineering department 500 level were splitted into 6 groups (each group has 5 members). The data collected was pre-processed by the removal of stop words and unwanted words. The plagiarism detector was designed using fuzzy-logic. The resulting system was evaluated using accuracy, recall, precision and F1-measure. The validation time was also recorded. The architecture of the developed system is shown in Figure 1.

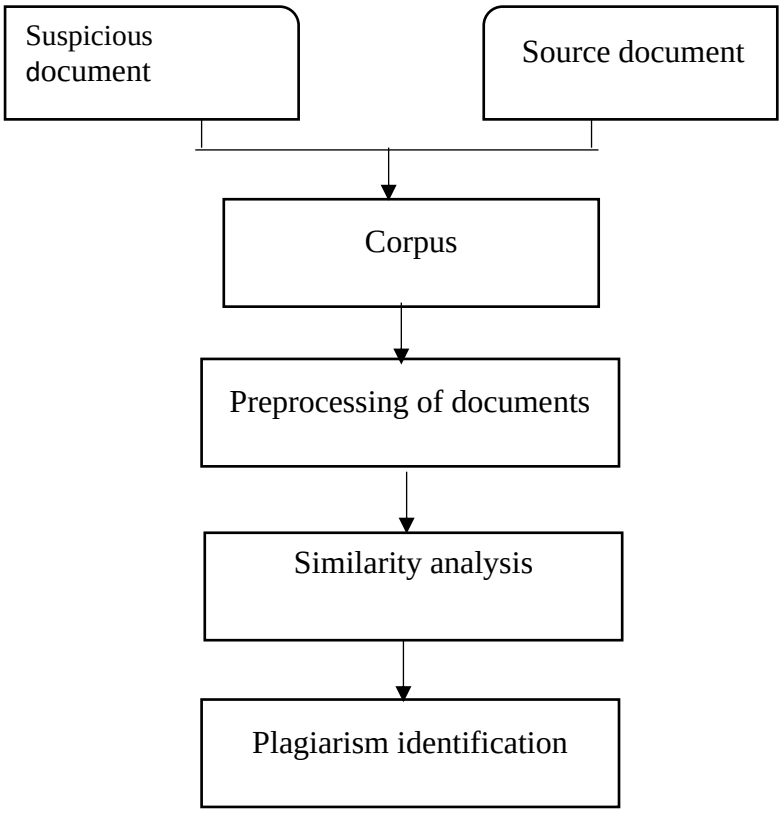

Figure 1: Block diagram of the developed system

RESULT AND DISCUSSION

Four different documents which were obtained from two departments in a higher institution in Nigeria were used to test the developed plagiarism detector. The result obtained after testing the developed system is shown in Table 1. Figure 2 shows the snapshot of the results obtained while testing the system.

Table 1: Results of the tested documents

S/N	Dataset	Number of words	Similarity percentage
1	Document 1	2222	20%
2	Document 2	694	60%
3	Document 3	6291	94%
4	Document 4	2391	98%

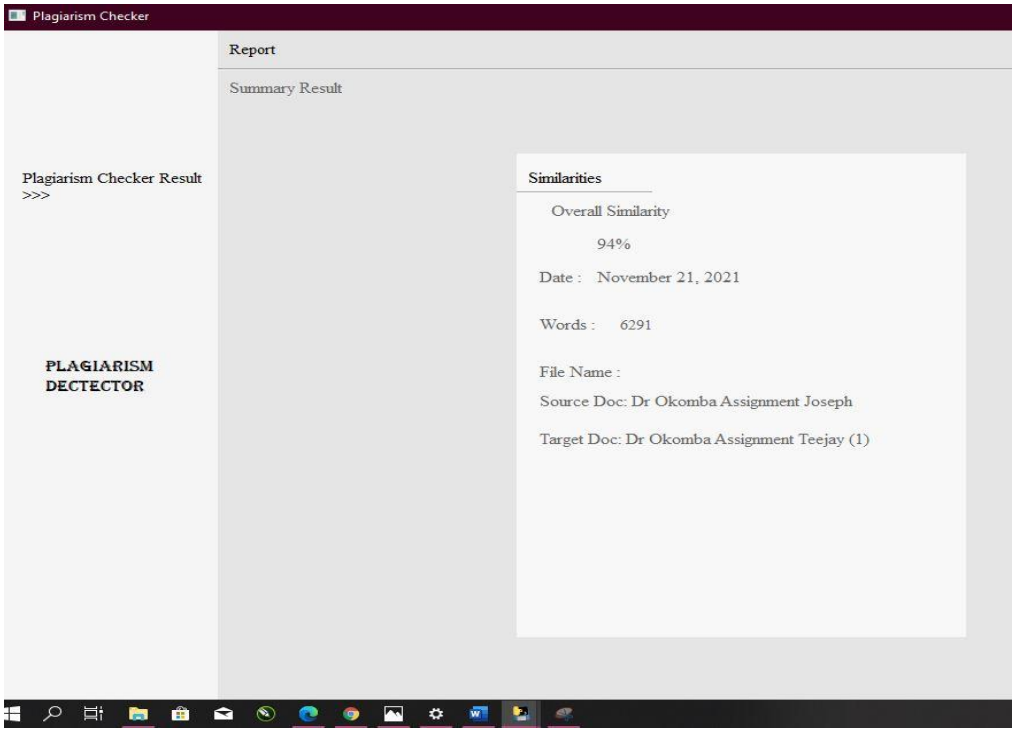

Figure 2: Snapshot showing the GUI displaying results from testing the developed system

Result from Evaluation of the developed plagiarism detector

The developed system was tested by ten lecturers of a higher institution in Nigeria. Questionnaires were administered to them to enable them assess the performance of the system developed based on the following metrics: functionality, accuracy, reliability, ease of use, efficiency, and portability. The results obtained from user's assessment is shown in Table 2 and Figure 3.

Table 2: Analyzed data from users' responses

S/N	Metrics	Excellent	Very Good	Good	Fair	Poor
1	**Functionality**	**7(70%)**	**3(30%)**	**0(0%)**	0(0%)	0(0%)
2	**Reliability**	**4(40%)**	**5(50%)**	**1(10%)**	0(0%)	0(0%)
3	**Ease of use**	**9(90%)**	1(10%)	0(0%)	0(0%)	0(0%)
4	**Efficiency**	5(50%)	3(30%)	**2(20%)**	0(0%)	0(0%)
5	**portability**	**6(60%)**	2(20%)	**1(10%)**	**1(10%)**	0(0%)
6	**accuracy**	**7(70%)**	2(20%)	1(10%)	0(0%)	0(0%)

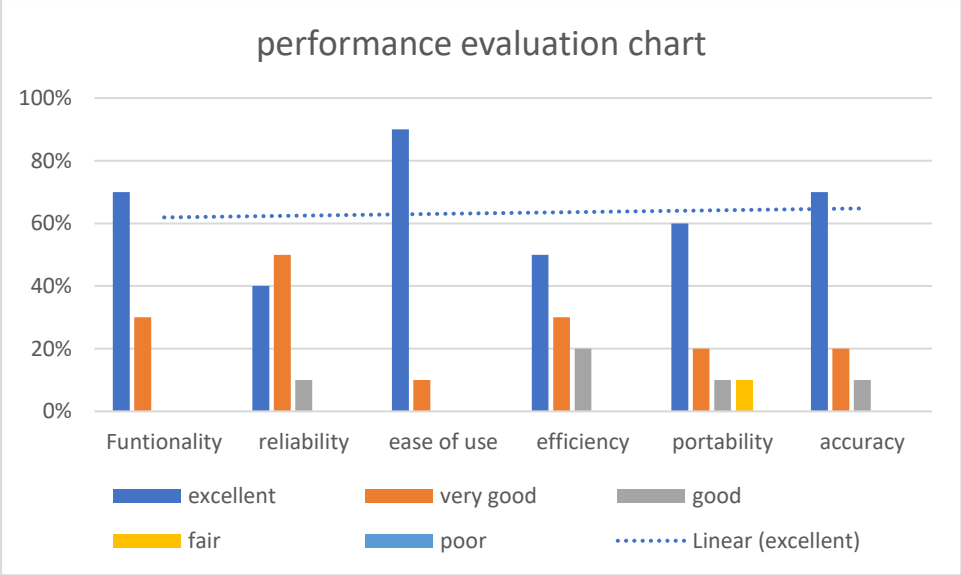

Figure 3: Performance evaluation chart

DISCUSSION

Results show that *ease of use* metric was rated excellent, which shows that users found the system very easy to use while *functionality* and *accuracy* were rated very good by the users, this indicated that the system functioned very well and accurately. Also, *portability and efficiency* metrics were rated good and this show that users found the system efficient and portable. The results obtained also show that the system can be used to detect plagiarism in student's assignments.

CONCLUSION

This research developed an automatic plagiarism detector using fuzzy-logic. The developed system was tested and evaluated based on: functionality, accuracy, reliability, ease of use, efficiency, and portability metrics. Results show that the developed plagiarism detector is very easy to use with high functionality and accuracy as well as moderate efficiency and portability based on user's assessment. However, future work should consider machine learning and deep learning techniques for plagiarism detection to improve system's accuracy.

REFERENCES

Abdelmalek A., Zakaria E., and Michel S., (2010) Evaluation of Text Clustering Methods Using WordNet, *The International Arab Journal of Information Technology*, (vol. 7, no. 4, pp. 349-357)

Sediyono A.(2008*), Algorithm of the Longest Commonly Consecutive Word for Plagiarism Detection in Text Based Document.* (pp 253-259).

Knight, A., Almeroth, K. and Bimber B. (2004). An Automated System for Plagiarism Detection Using the Internet, *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, Chesapeake*, (pp. 3619-3625)

Barr´on-Cede˜no A, Rosso P, M Potthast (2009), On Automatic Plagiarism Detection Based on n-Grams Comparison, *Proceeding of European Conference on Information Retrieval, Advances in Information Retrieval* (pp. 696-700).

Tebpour, A., Laskoukelayeh, M., Aminolroaya, Z. (2017) *Plagiarism Detection Based on a Novel Trie-tree based data structure*. (pp. 214-217).

Huang A. (2008). Similarity measures for text document clustering. *Proceedings of the sixth New Zealand Computer Science Research Student Conference* (pp. 49-56).

Si, A., Leong A. and Rynson W. (1997), *CHECK: a document plagiarism detection system, In Proceedings of the 1997 ACM Symposium on Applied Computing*, (pp. 70–77.)

Autade S.N., Prof.S.Z.Gawali and Prof. Dr. D. M. Thakore (2017), EMAS Framework for Text Plagiarism Detection, *International Journal of Applied Engineering Research ISSN 0973-4562* (Volume 12, Number 8 pp. 1584-1590)

Benno S., Nedim Lipka, P. Prettenhofer. S. Efstathios and K. Moshe, (2011) Intrinsic plagiarism analysis, *Language Resources and Evaluation* (Vol.45 Issue 1: pp. 63–82.)

Stamatatos. E. (2011). Plagiarism detection using stop word n-grams", *Journal of the American Society for Information Science and Technology* (Volume 62 Issue 12, pp. 2512-2527).

Francisco R., Antonio G., Santiago R., Jose L., Pedraza M., and Manuel Nieto. (2008*), Detection of Plagiarism in Programming Assignments,* IEEE Transactions on Education, (vol. 51, no. 2, pp. 174-183).

Gupta, D., Vani, K., and Singh, C. K. (2014*). Using Natural Language Processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection, In Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on (pp. 2694-2699). IEEE.*

Hermann M., Frank K., and Bilal Z. (2006) "Plagiarism -A Survey. *Universal Computer Science (vol 12, no. 8, pp. 1050-1084).*

Zhao, J., Xia, K., Fu Y. and Cui B. (2015), An AST-Based Code Plagiarism Detection Algorithm, *10th International Conference on Broadband and Wireless Computing, Communication and Applications*.

Umezawa K. (2011). A Study on Identifying Similar Documents based on the Dimension Reduction of a Document Vector. *The 3rd Forum on Data Engineering and Information Management, A6-1 (in Japanese).*

Ueta K. and Tominaga H. (2010). Plagiarism Detection methods based on Similarity by Distance for programing Reports. *Information Processing Society of Japan, SIG Technical Report (in Japanese)* (Vol.2010-CE-107 No.9)

Marcin K. and J Kitowski (2014), *Optimisation of Character n-gram Profiles*, *International Conference on Artificial Intelligence and Soft Computing* (pp 500-511)

Potthast M, B. Stein, Andreas Eiselt and Alberto Barr´on-Cedeno Paolo *Rosso. (2009). Overview of the 1st international competition on plagiarism detection. In 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse,* (page1.)

Agrawal M. and Sharma D. (2016), *2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16*

Ďuračíka M. (2017), Current trends in source code analysis, plagiarism detection and issues of analysis big datasets in TRANSCOM 2017*, International scientific conference on sustainable, modern and safe transport Procedia Engineering 192*

Nathaniel G., Maria P., and Yiu N. (2008). Nowhere to Hide, Finding Plagiarized Documents Based on Sentence Similarity. *in Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, NSW.* (pp. 690-696)

Osman, A. H., Salim, N., Kumar, Y. J., and Abuobieda, A. (2012). Fuzzy Semantic Plagiarism Detection. In AMLTA (pp. 543553).

Ozlem U., Boris K., and Thade N. (2005.). Using Syntactic Information to Identify Plagiarism. *In Proceedings of Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory Cambridge,* USA. (pp. 37-44).

Kundu R. and Karthik. K (2017), *Contextual plagiarism detection using latent semantic analysis*, *International Research Journal of Advanced Engineering and Science*, (Volume 2, Issue 1, pp. 214-217)

Snijders, C.; Matzat, U.; Reips, U.-D. (2012). Big Data: Big gaps of knowledge in the field of Internet. *International Journal of Internet Science.* (7: pp. 1–5).

Wadsworth, (2004) available at: http:// www.wadsworth. Com/English/special-features/ plagiarism/,

Yuuki Mori. (2015). Developing a Plagiarism Detection System based on Citations for Academic Reports. *Information Processing Society of Japan, Kansai-Branch Convention, C-05 (in Japanese).*