JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 04 Issue 02
2023

Research Article
https://doi.org/10.56979/402/2023

An Efficient Machine Learning Approach for Plagiarism Detection in Text Documents

Muhammad Mubashir Zahid[1*], Kamran Abid[1], Abdul Rehman[2], M Fuzail[1], and Naeem Aslam[1]

[1]Department of Computer Science, NFC Institute of Engineering and Technology Multan, Pakistan.
[2]Faculty of Computer Science, Lahore Garrison University, Lahore, 54000, Pakistan.
[*]Corresponding Author: Muhammad Mubashir Zahid. Email: mubashirzahid128@gmail.com

Received: December 29, 2022 **Accepted:** February 27, 2023 **Published:** March 29, 2023.

Abstract: Plagiarism is when you use someone else's words or ideas as your own. On every subject, the internet is a reliable source of information. People can therefore simply copy data and use various techniques to cover up plagiarism. Extrinsic and intrinsic methods can both be used to detect plagiarism. Extrinsic plagiarism involves comparing the source and the allegedly plagiarised texts to one another in order to obtain precise similarity metrics like Jaccard and Cosine. Source materials are not necessary for intrinsic plagiarism recognition, though. Plagiarism can be detected by an author's writing style and other notable actions. Cross-Lingual Plagiarism (CLP) is a kind of plagiarism in which the author steals content by translating text from one language to another like Urdu-English. It is hard to identify CLP because the source and suspicious documents are in two different languages. In this regard, various approaches to tackling the problem of CPD in text documents were presented. We need to apply ML way to deal with the problems of CLPD. For PD task curpus is used to evaluate the performance of PD, so we use Urdu English language pair Corpus CLPD UE 19 [1]. The source text in the language-pair corpus (CLPD-UE-19) is written in Urdu, whereas the suspicious text is supplied in English. To construct a dataset that can be understood by machine learning tools and extract optimized features from a corpus using Python NLP techniques. Our created dataset is in the CSV format in which there are distinctive features of source, and suspected content is mentioned like Jaccard similarity and Cosine similarity. We have used one gram and tri-gram of the preprocessed text to get comparability measures. five ML classifiers, such as KNN, Naïve Bayes, SVM, Decision Tree, and Random Forest, are utilized to build models. Python language is used on PyCharm tool to build models from various classifiers. We use two methods to examine the models' accuracy (cross-validation and percentage split) in the python language. The trial shows that KNN, RF and have produce better results as compared to other models.

Keywords: Plagiarism detection, Cross-lingual plagiarism detection, Obfuscation level, Similarity matrix, Extrinsic plagiarism detection, Urdu English Plagiarism Detection.

1. Introduction

The Text reuse is the practice of using passages from previously published works by other authors without giving due credit to the original authors. If text is used without proper referencing, it becomes plagiarism which is considered as intellectual theft. The act of plagiarizing is the uncredited use of another person's ideas or words. Plagiarism comes in many forms and degrees, including paraphrase, direct copying, patchwork, self-plagiarism, and inaccurate citation. To paraphrase is to rewrite text in your voice. The act of copying and pasting the original content into a new document is known as verbatim plagiarism. Patchwork plagiarism is the practice of using passages from different sources to create one new work. Self-

plagiarism is the act of exploiting one's own prior work without citing it in a new paper. An improper citation is one in which the author fails to adequately cite the source text. Finding information on any subject is now simple because to the growth of online data. Finding the origins of a questionable document is the process of plagiarism detection. A suspect document is the one for which plagiarism detection is desired. Cross-lingual plagiarism (CLP) is plagiarism in which the suspicious material and the source text are written in separate languages. Both extrinsic and intrinsic strategies can catch plagiarism. Finding all of the sources of a questionable document from a certain document collection is the method of extrinsic PD. One method of identifying plagiarism within a single manuscript is intrinsic PD. In the intrinsic PD several features are used to perceive the plagiarism like author's writing style, vocabulary size and frequency of word usage. Through an information retrieval, the query is a suspicious document, and the system retrieves the potential sources of these worrisome documents. The suspect text is compared against unsure source texts using a similarity metric. Similarity measures include JS, CS, and longest common subsequence (LCS). Plagiarism in monolingual texts can be detected using a number of techniques, such as 1) Character 2) Vector 3) Syntax, and the 4) Sematic Based. N-grams of words are used to calculate how similar the source and the suspected content are in the character-based approach. "The source and suspicious texts are converted into vectors in the vector-based approach, and the correlation between two vectors can be found using the CS or JS metric" [1]. There are numerous ways to find CLP. The most used method among these is translation + monolingual analysis (T+MA). In T+MA method, a suspected document is translated into source documents language which makes the task a mono lingual task and then applying any plagiarism detection strategy, task could be performed easily.

2. Related Work

The Due to the rise in online papers, CLPD is currently one of the most difficult challenges. In numerous studies in the field of PD, a predictive model is built to calculate the similarities between the source and the suspect text documents. As a result, the section below highlights several aspects of connected research projects. Plagiarism is when you use someone else's words or ideas as your own. Literary theft is another name for it. Everyone can learn about a wide range of subjects on the internet. People can therefore simply copy data and use various techniques to cover up plagiarism. A serious problem in an academic setting is plagiarism. To decrease repetitive content and raise literacy, plagiarism detection is essential [1,5]. Plagiarism that reuses a text while maintaining the same language is known as mono-lingual plagiarism. However, CLP occurs when the recycled text is translated into a new language.

For CLPD, the cross-lingual repository is an essential part that is used to train statistical algorithms. There are two types of corpora used for CLPD i.e. 1) Parallel corpus: this corpus contains text in one language and their translations in another language whereas, 2) Comparable Corpus: it is the collection of similar texts in different languages. In 2019 Israr Haneef et al. [2] have created a CLPD corpus for the English-Urdu language combination. The Dataset contains of 2398 Text documents in XML format, with questionable content in English and Urdu as the source language. There are four classes defined in the corpus that are 1) heavy revision, 2) near copy, 3) no-copy, and 4) light revision. This corpus is openly accessible to researchers and other scholars [2]. Existing parallel corpus and their shortcomings are discussed in [3]. They have since established a new corpus for CLPD. Various methods are employed for CLPD, such as CL-Alignment-based Similarity Analysis (CL-ASA), CL-Character n-gram analysis (CL-CNG), and translation plus monolingual analysis (T+MA). The literature shows that CL-CNG has outperformed the CL-ASA if some vocabulary between the two languages is shared. A multilingual dictionary is yet another CLPD tactic. The source and suspect materials in the various languages are used in a parallel corpus to create a bilingual dictionary. The identification of near copy obfuscation levels can be done extremely effectively using this method [4]. For arbitrary languages, CL-ASA and CL-CNG show similar results however T+MA method produces better results comparatively.

In Oleg Bakhteev et al. [6] created a CLPD framework for Russian and English language pairings. They employed the T+MA approach to find plagiarism. Using a translation framework, the suspicious Russian document is translated into English. The candidate documents are then obtained, and a phrase embedding system is used to compare the source and the suspicious document [6]. Zubarev D. V. et al. have published another study to uncover cross-lingual plagiarism in the Russian and English language pair. They employed a cross-lingual corpus to detect plagiarism by using ML approaches [7].

The SVM-based classifier in [8] is used to determine the CLPD in Arabic-English documents. Another framework for the Arabic-English pair is proposed in [9], where the suspicious material is written in Arabic yet the source text is in English. To locate CLP, they applied the T+MA method. The word2vec model has been utilised by Raki Lachraf et al., 2019, to identify CLP in Arabic and English language pairings. They have used the skip-gram technique which shows better performance than CBOW [10]. Another ML-based work is done in [11] where Chinese-English document pairs are used. The corpus is produced using online document resources and for model training and classification, SVM is used. Ceska, Z., Toman, M., & Jezek, K, 2008, has put up a new foundation for CLPD in Chinese and English. To prepare the corpus, many online sources, including Wikipedia articles, were employed and train the models [12].

The use of WordNet is proposed by [13] to identify plagiarism for English-Chinese language pairings. For CLPD, Liu Gang employed a parallel corpus. Comparing the proposed method to cutting-edge methods like T+MA and CL-ASA, it performs better. Another similar method is presented by Victor Thompson et al, 2017. Using the Word2Vec model, they have computed the plagiarism. The results produced through the proposed method and T+MA are compared which show the better results of the developed method. They have improved the performance of the translation tool by using the word2vec model. Another work based on Word2Vec is presented by Parth Gupta et. al. [14].

A brand-new approach for detecting CLP in Urdu and English language pairs has been put out by F. Shahzad et al. "1000 documents total (500 source papers in Urdu and 500 suspicious documents in English) make up the corpus They have used Google Translate and Bing Translate for source document translation. In order to measure the performance of the system, T+MA of the method is used" [15]. In order to identify CLP detection Eriksson Haker and Martin Sch. 2014 looked at both automated and human translation tools. The outcomes demonstrate that machine translation performs better for cross-lingual plagiarism as compared to manual translation [16]. An automatic translation-based technique for plagiarism detection has been put out by Nava Ehsan et al. They have used Dict. cc, BabelNet, and Google Translate for translation. The trial revealed that Google Translator's translations result in superior outcomes [17]. 2017 saw the development of a system by Marcos Garcia and Pablo Gamallo to detect plagiarism in Romance languages. To detect plagiarism, the provided record's grammatical structure is labelled [18]. Laurent Besacier et al. 2017, have used a multilingual corpus. According to the experiment, this corpus responded best to translation combined with a monolingual evaluation of the text [19]. 2019 saw the introduction of a new framework by Le Thanh Nguyen and Dinh Dien [20] for the detection of plagiarism in the Vietnamese and English language pairs. Machine learning based methods, both supervised and unsupervised methods, for CLPD have shown better results as compared to traditional information retrieval methods. For CLP detection, a number of studies have used ML techniques. [7] [10] [11] [12] [14].

Table 1. Comparison of Different research on this topic

Ref.	Year	Technique	Algorithm	Language	Dataset	Type	Similarity metrics	Accuracy
Erfaneh Gharavi et al.	2016	Deep learning	Word2vector	Persian	PAN2016	Monolingual	Cosine similarity, Jaccard similarity	90.6 %
Basant Agarwal	2019	Machine learning	Semantic space	Hindi-English	PAN-CLEF	Cross-lingual	Semantical similarity, alignment-based similarity	92 %
Mokhtar Al-Suhaiqi et al.	2018	Machine learning	Linear logistic regression, naïve	Arabic-English	Real dataset	Cross-lingual	N-Grams Similarity, LCS, DC, Fingerprin	92 %

			Bayes, SVM,				t based JC and Fingerprint based	
Saeed Albukhitan et al.	2020	Deep learning	Word2Vec and Glove	Arabic	Arabic documents	Mono-lingual	Semantic annotation	80.8 %
Waqar Ali et al.	2018	Machine learning, NLP	Damerau Levenshtein Distance, , Vector Space Model	Urdu	OSCA corpus	Mono-lingual	intertextual and syntactic similarity	90 %
Salha Alzahrani and Hanan Aljuaid	2020	Deep learning, NLP	deep neural networks	Arabic-English	handmade data	Cross-lingual	Semantic text similarity	97 %
Habibollah Asghari et al.	2020	Artificial intelligence	Artificial intelligence	Persian	HAMTA	Mono-lingual	Semantic similarity	

2.1 Available Dataset for CLP detection

Table 3 contain the datasets that are available for CLP detection. The first column shows the citation and second column explain the languages that supported and number of documents in the in the corpus. Last column is the full name of the corpus that online available.

3. Proposed Methodology

We have suggested a method for identifying CLP for the Urdu-English languages in this chapter. There hasn't been enough work done on the English-Urdu language pair. Our suggested dataset was produced using the methodology depicted in Figure 3.1 and is built on the CLPD_UE_19 corpus. Some ML algorithms further utilises the dataset to determine the degree of obfuscation. Different parameters serve as features in our dataset. Each pair of documents also has a document id, domain, size, and level of obfuscation. For the train of the models, the variables as similar values of the source and suspected documents are used. Before using the similarity measurements, one and three grammes of the original and the suspected document are created.

3.1 The Dataset CLPD-UE-19

There is a number of parameters/features associated with CLPD-UE-19 dataset that include document index (article id), source in Urdu, source translated into English, presumed target, domain, size, and level of obfuscation of each pair of documents. There are 2379 document pairs in this dataset. In order to extract the features from CLPD-UE-19 dataset, the documents and associated features are fed to the designed models as shown in figure-2. Uni-gram and tri-grams of the source and the suspicious documents are created previously applying the classification procedure.

he CLPD UE 19 [2] corpus is organized in XML. To read this corpus and translate the Urdu source into the English source using Google Translation APIs, we used Python code. Figure 1 is presenting an example from CLPD-UE-19 dataset. The documents in dataset are divided into three groups based on their size which are small, medium and large documents. Four labels are defined for each pair of documents that include Non-Plagiarized (NP), Light Revision (LR), Heavy Revision (HR), and Near Copy (NC).

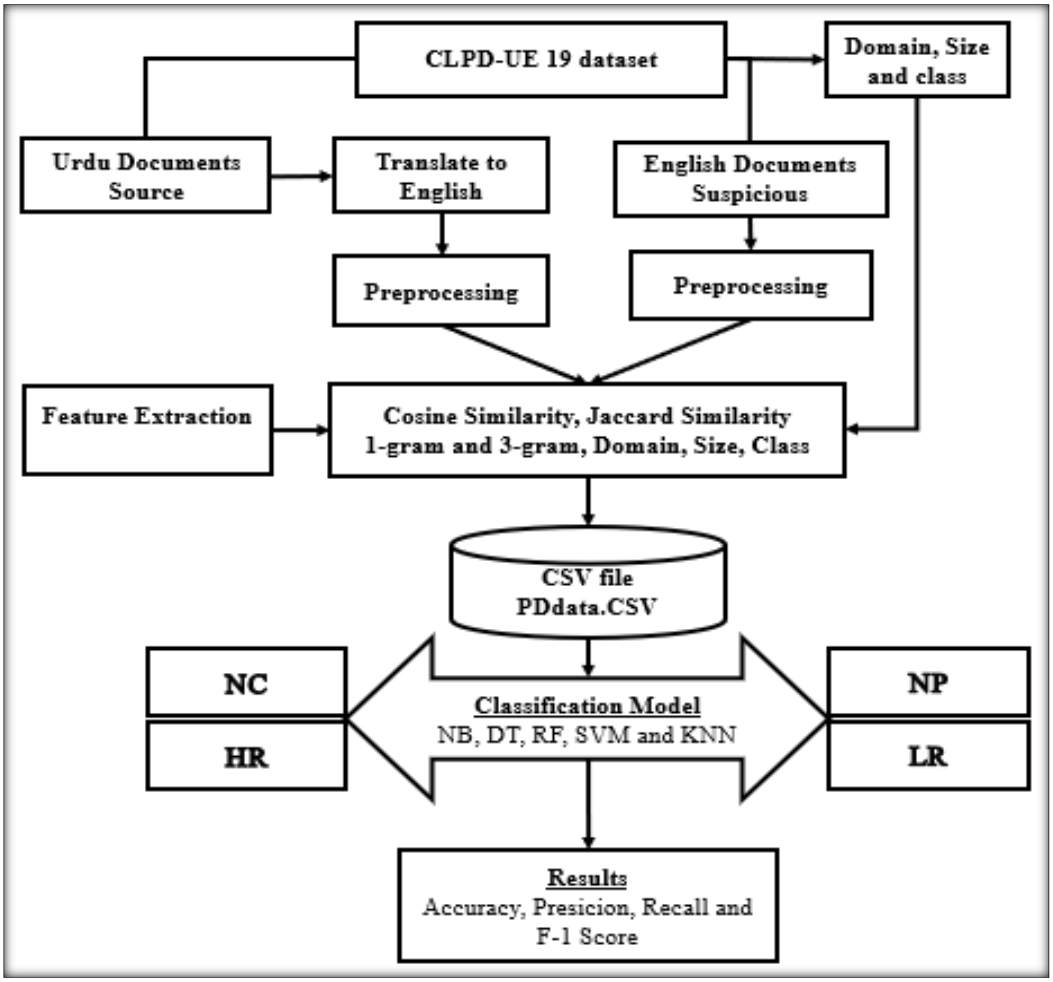

Figure 1. Proposed Methodology

3.2 Pre-processing

To make computation and classification process easier, a number of preprocessing steps are used. The preprocessing includes stop words removal, lemmatization using wordnet and conversion to unigram and trigrams. The correlation result of each pair of documents are added in the csv file containing associated features of dataset. An example of this is shown in the Table 4. In this table, multiple similarity measures are listed. The columns were also Jaccard similarity with uni-gram, cs is cosine similarity with uni-gram, LCS indicates the value of two papers, js fr is Jaccard similarity with 3-grams, cs fr is cosine similarity with 3-grams, size field refers the structure of the particles such as small, medium, and large, and the last attribute describes the plagiarized class of this pair.

Table 1 contains three rows first row show the Source text that is in Urdu language and the second row contain the English suspicious text the last row has the text of translated source text. In Table 2 two format of text are include in first row the text is show before the Preprocessing methods such as removal of punctuation and stop words. After applying the stop words and punctuation removal techniques the structure of text is changed and show in the second row.

We use CLDP-UE-19 [2] in our research work. This Curpus is the collection of English-Urdu pair documents.

4. Experimental Setup

In our processed dataset, there are 2379 records. The dataset is divided between training and test sets. Six different classifiers are used to classify data as given in table 3. The training data is compiled in the form of a csv file having source and suspicious documents along with 8 different features including the class label as shown in the Table-1. These models are evaluated using one-fold and 10-fold cross validation.

In uni-fold validation, we used 66% instances for training and 34% for testing purpose whereas in 10-fold cross validation, 90% instances out of 2379 are used for training and remaining 10% for testing and validation.

4.1 Accuracy

In this section we compare the training, Testing and Cross Validation accuracy of all the models.

Table 2. Accuracy comparison of models

Models	Training Accuracy	Testing Accuray	Cross_Validation Accuracy
DT	0.93	0.89	0.87
RF	0.99	0.89	0.90
NB	0.88	0.88	0.87
SVM	0.91	0.90	0.72
KNN	0.93	0.92	0.84

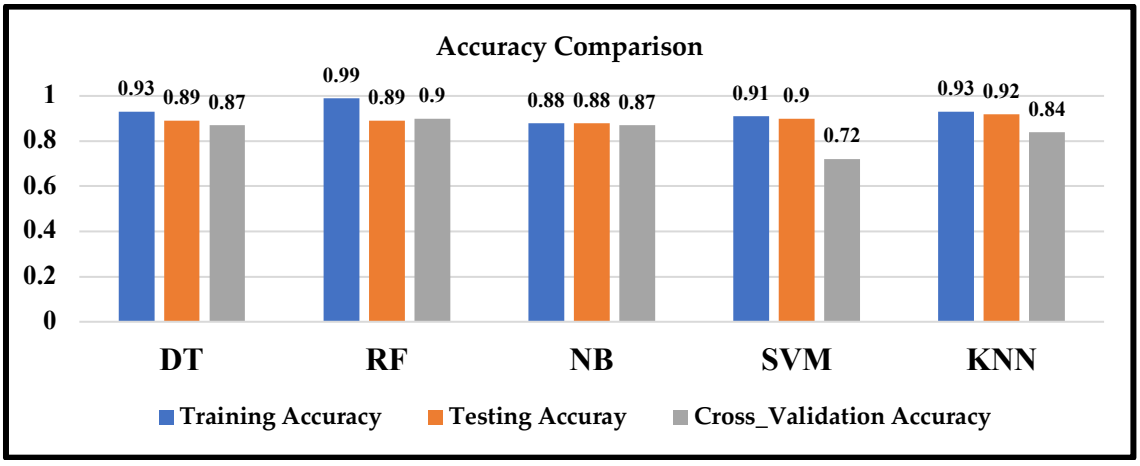

Figure 2. Classification results of cross validation

5. Conclusion

Plagiarism is when you use someone else's words or ideas as your own. Detecting plagiarism is the process of locating the suspect document's copied passage(s). Extrinsic plagiarism detection is one of two types of plagiarism detection methods that focuses on finding the source or sources for a particular suspect document. Typically, there are two parts to this process. In the first step, all papers containing even a little amount of the suspected information are gathered. It functions like an information retrieval system, retrieving several documents based on keywords provided by the customer. Second stage: using similarity measures like Longest Common Subsequence (LCS), n-gram overlap, cosine, etc., to identify the suspected document's plagiarised passage(s). Our study's goal is to create a technique for detecting CL plagiarism in the Urdu-English language combination. For ML algorithms to identify CLP, we have employed. To create the dataset that is used to train our model, we employed the CLPD_UE_19 corpus. In our situation, translation is combined with a monolingual analysis of the corpus. Five classifiers used to generate models are NB, NN, KNN, SVM, DT, and RF. In order to test our models, Python offers a variety of options, including cross-validation and percentage split. All of the dataset's features are used to train the models. The trial shows that KNN, RF and was produced superior results as compared to other models.

Future research on this subject can expand on it by employing numerous translators to translate from Urdu to English and then comparing the results of each pair using different similarity metrics. The CLPD_UE_19 corpus can be used to generate a multilingual lexicon, which can subsequently be used to translate documents. The dataset can be used for unsupervised learning to cluster the instances after it has been created.

References

1. Chowdhury, H. A., & Bhattacharyya, D. K. (2018). "Plagiarism: Taxonomy, tools and detection techniques". arXiv preprint arXiv:1801.06323.
2. Haneef, I., Adeel Nawab, R. M., Munir, E. U., & Bajwa, I. S. (2019). "Design and Development of a Large Cross-Lingual Plagiarism Corpus for Urdu-English Language Pair". Scientific Programming, 2019.
3. Ferrero, J., Agnes, F., Besacier, L., & Schwab, D. (2016, May). "A multilingual, multi-style and multi-granularity dataset for cross-language textual similarity detection". In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 4162-4169).
4. Barrón-Cedeno, A., Rosso, P., Pinto, D., & Juan, A. (2008). "On Cross-lingual Plagiarism Analysis using a Statistical Model". PAN, 212.
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
6. Bakhteev, O., Ogaltsov, A., Khazov, A., Safin, K., & Kuznetsova, R. (2019, September). "CrossLang: the system of cross-lingual plagiarism detection". In Workshop on Document Intelligence at NeurIPS 2019.
7. Zubarev, D. V., & Sochenkov, I. V. (2019). Cross-language text alignment for plagiarism detection based on contextual and context-free models. In Proc. of the Annual International Conference "Dialogue (Vol. 1, pp. 799-810).
8. Al-Suhaiqi, M., Hazaa, M. A., & Albared, M. (2018). Arabic English cross-lingual plagiarism detection based on keyphrases extraction, monolingual and machine learning approach. Asian Journal of Research in Computer Science, 1-12.
9. Alaa, Z., Tiun, S., & Abdulameer, M. (2016). Cross-language plagiarism of Arabic English documents using linear LOGISTIC REGRESSION. Journal of Theoretical & Applied Information Technology, 83(1).
10. Lachraf, R., Ayachi, Y., Abdelali, A., & Schwab, D. (2019, August). ArbEngVec: Arabic-English Cross-Lingual Word Embedding Model. In Proceedings of the Fourth Arabic Natural Language Processing Workshop (pp. 40-48).
11. Lee, C. H., Wu, C. H., & Yang, H. C. (2008, June). A platform framework for cross-lingual text relatedness evaluation and plagiarism detection. In 2008 3rd International Conference on Innovative Computing Information and Control (pp. 303-303). IEEE.
12. Ceska, Z., Toman, M., & Jezek, K. (2008, September). Multilingual plagiarism detection. In International Conference on Artificial Intelligence: Methodology, Systems, and Applications (pp. 83-92). Springer, Berlin, Heidelberg.
13. Gang, L., Quan, Z., & Guang, L. (2018, March). Cross-language plagiarism detection based on WordNet. In Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence (pp. 163-168).
14. Gupta, P., Singhal, K., Majumder, P., & Rosso, P. (2011). Detection of Paraphrastic Cases of Mono-lingual and Cross-lingual Plagiarism. ICON.
15. Shahzad, F., Jabeen, S., Pasha, M., Majeed, B., & Gao, X. (2018). DOMAIN-SPECIFIC CROSS-LINGUAL URDU TO ENGLISH (CLUE) PLAGIARISM DETECTION. Pakistan Journal of Science, 70(2), 195.
16. ERIKSSON, H., & SCHÖN, M. (2014). Using a machine translation tool to countercross-language plagiarism.
17. Ehsan, N., Tompa, F. W., & Shakery, A. (2016, September). Using a dictionary and n-gram alignment to improve fine-grained cross-language plagiarism detection. In Proceedings of the 2016 ACM Symposium on Document Engineering (pp. 59-68).
18. Garcia, M., & Gamallo, P. (2017, August). A rule-based system for cross-lingual parsing of Romance languages with Universal Dependencies. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (pp. 274-282).
19. Ferrero, J., Besacier, L., Schwab, D., & Agnes, F. (2017). Deep investigation of cross-language plagiarism detection methods. arXiv preprint arXiv:1705.08828.
20. Dien, D. (2019, September). Vietnamese-English Cross-Lingual Paraphrase Identification Using Siamese Recurrent Architectures. In 2019 19th International Symposium on Communications and Information Technologies (ISCIT) (pp. 70-75). IEEE.
21. E. M. Hambi and F. Benabbou, "A Multi-Level Plagiarism Detection System Based on Deep Learning Algorithms" IJCSNS International Journal of Computer Science and Network Security, VOL.19 No.10, October 2019.
22. E. M. Hambi and F. Benabbou, "A new online plagiarism detection system based on deep learning," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 9, pp. 470–478, 2020, doi: 10.14569/IJACSA.2020.0110956.
23. S. Alzahrani and H. Aljuaid, "Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases," J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2020,doi:10.1016/j.jksuci.2020.04.009.
24. M. Roostaee, S. M. Fakhrahmad, and M. H. Sadreddini, "Expert Systems with Applications Cross-language text alignment: A proposed two-level matching scheme for plagiarism detection," Expert Syst. Appl., vol. 160, p. 113718, 2020, doi: 10.1016/j.eswa.2020.113718.
25. F Jérémy, L. Besacier, Ferrero, Laurent Besacier, Didier Schwab, and Frédéric Agnes "Deep Investigation of Cross-Language Plagiarism Detection Methods," pp. 6–15, 2017.
26. M. Roostaee, M. H. Sadreddini, and S. M. Fakhrahmad, "An effective approach to candidate retrieval for cross-language plagiarism detection: A fusion of conceptual and keyword-based schemes," Inf. Process. Manag., vol. 57, no. 2, p.

27. M. Franco-Salvador, P. Gupta, and P. Rosso, "Cross-language plagiarism detection using a multilingual semantic network," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7814 LNCS, pp. 710–713, 2013, doi: 10.1007/978-3-642-36973-5_66. (2013).
28. M. Franco-Salvador, P. Gupta, P. Rosso, and R. E. Banchs, "Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language," Knowledge-Based Syst., vol. 111, pp. 87–99, 2016, doi: 10.1016/j.knosys.2016.08.004.
29. S. Zouaoui and K. Rezeg, "Multi-Agents Indexing System (MAIS) for Plagiarism Detection," J. King Saud Univ. - Comput. Inf. Sci, 2020, doi: 10.1016/j.jksuci.2020.06.009.
30. Safi, Faramarz & Rakian, Sh & Nadimi-Shahraki, Mohammad H.. (2017). English-Persian Plagiarism Detection based on a Semantic Approach. 5. 275-284.
31. Fazal, U., Khan, M., Maqbool, M. S., Bibi, H., & Nazeer, R. (2023). Sentiment Analysis of Omicron Tweets by using Machine Learning Models.
32. Maqbool, M. S., Hanif, I., Iqbal, S., Basit, A., & Shabbir, A. (2022). Optimized Feature Extraction and Cross-Lingual Text Reuse Detection using Ensemble Machine Learning Models.