Int. J. Advance Soft Compu. Appl, Vol. 18, No. 2, July 2026
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Detecting Plagiarized Text in Images using OCR and NLP-Based Deep Learning Approaches

Ahmad Alhami[1] and Belal Zaqaibeh[2]

[1]Department of Data Science and Artificial Intelligence, Jadara University, Jordan
e-mail: ahmadalhami1977@gmail.com
[2] Department of Data Science and Artificial Intelligence, Jadara University, Jordan
e-mail: zaqaibeh@jadara.edu.jo

Abstract

This paper evaluates plagiarism detection using deep learning and natural language processing techniques. A novel model, Detecting Embedded Plagiarized Text in Images (DEPTI), is introduced to identify plagiarized text embedded within images, demonstrating high accuracy and robust performance. DEPTI effectively recognizes paraphrased, translated, and artificial intelligence generated content, achieving strong detection capabilities across diverse scenarios. The model integrates PAN-PC-11, TF-IDF, Tesseract OCR, DistilBERT, and LSTM to extract and analyze text from images, enabling advanced plagiarism detection beyond conventional approaches. Experimental results confirm DEPTI's effectiveness, highlighting its potential as a reliable tool for safeguarding academic integrity in the digital era.

Keywords: *DEPTI, Text Image Plagiarism, Deep Learning, and Natural Language Processing.*

1 Introduction

Plagiarism is stringently conceptualized as the misappropriation of intellectual contributions including ideas, textual materials, or creative outputs—without due recognition of the original source, and is regarded as a profound infringement upon the principles of academic integrity [1]. The pervasive accessibility of unregulated online resources, in conjunction with the increasing reliance on algorithmic and software-assisted writing platforms, has exacerbated the incidence of plagiarism. Consequently, it has emerged as one of the most pervasive and critical forms of scholarly misconduct confronting contemporary educational and research environments [2].

Additionally, authors verify that plagiarism can be a direct copy, it can be an intelligent paraphrasing, a translation, and Artificial Inelegance (AI) generated text as well. They have also underscored the inadequacies of Turnitin in detecting those texts that are elaborately paraphrased or translated from other languages [3] As a result, it is necessary to investigate modern deep learning and Natural Language Processing (NLP)-based techniques that can serve to meet the growing demand. An objective of this research is to of new techniques, considering that it is of significance if the plagiarism detecting methods are accurate and valid.

Today, major technological advancements are recorded in various fields of science primarily due to the rapid and widespread availability of information and the over-reliance on digital sources of

Received 3 April 2026; Accepted 1 July 2026

information, which is a double-edged weapon if misused [4]. This has made it easier for individuals to steal information without citing it to its rightful owners [5]. Progress poses a significant challenge to the research and academic community and has led to an increase in plagiarism in scientific and academic research, as the issue of academic plagiarism has become one of the most important risks and challenges facing both academic and research communities [6] [7]. This makes it easier for individuals to steal information without attributing it to its rightful owners [5].

The proposed model which is called the Detecting Embedded Plagiarized Text in Images (DEPTI) is expected to remove the obstacles in documents that is not readable and scanned from the image to the scenario where the plagiarism detection will be a job that is entirely done, and even new image-file types that have text on them will not escape the view of the tool. The innovation emphasized would be of a certain value in the educational area of progression to organizations, being partners in the transparent and original distribution of knowledge and truth, and of course, instruments for recognizing the very plagiaristic places.

The importance of this research can be summarized in designing and developing an effective model for detecting plagiarized texts that are embedded in images via Optical Character Recognition (OCR), converting them into processable text. Once transformed, more sophisticated methods in NLP and Deep Learning (DL) with Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) are able to be applied for plagiarism detection with a higher level of accuracy.

This paper is organized as follows: a comprehensive introduction is represented in section 1, and the literature review is shown in section 2, section 3 shows the methodology of DEPTI, subsequently, section 4 shows the results and the conclusion and future work in section 5.

2 Related Work

The rapid advancement of contemporary technologies has markedly diminished the efficacy of conventional plagiarism detection systems. Most existing tools are optimized for textual analysis, yet they remain largely incapable of processing content embedded within non-textual formats such as images, scanned documents, or PDF files. This limitation enables plagiarized material concealed in visual or scanned formats to evade detection. Accordingly, the development of methodologies that enable the extraction of textual data from images and video sources [8][9], followed by its conversion into machine-readable digital formats for plagiarism screening, is imperative. Such approaches hold particular significance in scientific and academic contexts, where accuracy, transparency, and the safeguarding of authorship rights constitute foundational principles [10].

Artificial intelligence–based tools for plagiarism detection have been systematically investigated in [1] through the utilization of an academic dataset comprising diverse source types, with the objective of enhancing the authenticity of educational materials and mitigating plagiarism. The study employed a sample of 200 plagiarism cases, including 50 instances of direct copying, 75 instances of paraphrased or rewritten content, and 75 instances generated by machine algorithms. These cases were analyzed using three widely adopted plagiarism detection systems, Turnitin, Grammarly, and Copyleaks.

A novel methodology was proposed in [2] to address the increasing incidence of plagiarism arising from large language models. The approach employed advanced architectures such as GPT-3.5 and T5, integrating question generation with cosine similarity measures to quantify textual resemblance. This framework achieved a precision rate of 94%, thereby demonstrating its efficacy in distinguishing AI-generated text from authentic scholarly writing.

The MIT Plagiarism Detection Dataset was introduced in [11] to advance research on plagiarism identification in low-resource languages, with a particular focus on Marathi. The study explored

the application of Term Frequency–Inverse Document Frequency (TF-IDF) in conjunction with BERT embeddings and the MahaSBERT model. This approach embodies the principle of representing textual data numerically, enabling the model to learn not only from predefined similarity scores but also from intrinsic natural language patterns. Consequently, the system exhibits enhanced intelligence and adaptability in detecting plagiarism across linguistically underrepresented contexts.

The BERT framework incorporates the MahaSBERT-STS model, an adaptation of MahaSBERT specifically designed to generate embeddings for semantic textual similarity [12]. In addition to this approach, TF-IDF vectors were employed to provide a statistical representation of textual data. The integration of these methods yielded an F1 score of 88.5% for BERT, underscoring its effectiveness in semantic similarity tasks. Furthermore, the study presented in [13] investigates the application of machine learning algorithms to automate the evaluation of requirements articulated in natural language. The findings reveal that random forest outperformed all stemmers, achieving an accuracy of 95% with the ISRI stemmer, and 0.97 with the Arabic light stemmer. These results highlight the robustness and practicality of the random forest algorithm.

SNLI and MultiNLI datasets have employed deep learning in the form of BERT and RoBERTa supported by traditional techniques represented by TF-IDF [14]. The leading status of the plagiarism detection model is validated through the corresponding results, which dies on the part of their research have asked Pearson correlation to produce.

Analyzing texts, including uploaded reviews and user posts on e-commerce portals, forums, and social media platforms, regarding their opinions about the event or person, product, and service is proposed as Sentiment Analysis (SA) [15]. The SA task comprises analyzing text to determine whether the sentiment expressed is negative, positive, or neutral, aiming for precise subjective data analysis.

The LSTM network was used in [16] which turned out to be quite effective in the identification of academic plagiarism. The aforementioned choice performed by exclusively using Doc2Vec vectors and TF-IDF on the PAN-PC-11 dataset. The method happened to be success of 99% in detecting and handling academic plagiarism. Table 1, shows an overview of the significant points in the studies that have been done already in a simple manner.

Table 1: Algorithms and their accuracy

Algorithm	**Accuracy**
Copyleaks, Grammarly, Turnitin	92%
GPT-3.5, T5 Paraphrasing	94%
MahaSBERT, TF-IDF	88.5% (F1)
BERT, RoBERTa	87%
LSTM, Doc2Vec, TF-IDF	99%

The efficiency of deep learning methods shows a high level of accuracy in comparison to traditional ones. The GPT-3.5 and T5 models show detecting AI-generated texts and the outcomes demonstrate that it is essential to extend the use of these technologies in academic and research areas in order to come up with the drawbacks of traditional systems.

In [24], authors have utilized a genetic algorithm fine-tune OCR hyper parameters, ensuring maximal text extraction length, followed by NER processing to categorize the extracted information into required entities, adjusting parameters if entities were not correctly extracted based on manual annotations. Their study highlighted the approach of IE from images of documents by dynamically adjusting OCR parameters through an adaptive model, followed by the extraction of required entities using NLP techniques. The efficacy of these methods was

demonstrated through the comparison of the extracted entities like organization identification numbers, TINs, license numbers, payment amounts, and payment dates.

3 The Methodology of DEPTI

The proposed DEPTI relies on a dataset called the PAN Plagiarism Toolkit PAN-PC-11, where it is one of the most important datasets used to evaluate plagiarism detection algorithms. The dataset consists of original and suspected files, the DEPTI can extract text from images and then apply the data cleaning process, the text will be processed through features extraction using TF-IDF to extract statistical features and also using DistilBERT to extract text context as Fig. 1 shows.

The DEPTI model is an advanced approach that detecting plagiarism in images containing texts and evaluating automated plagiarism detection algorithms. In the initial step which is data preparation, the DEPTI reads the file concerned with detecting plagiarism from it (plagiarized and non-plagiarized) and also the PAN-PC-11 dataset, the dataset is split and divided into training and testing sets with 80/20 ratio using the stratified method to split the dataset into training and testing sets. thus, text in suspicious images is scanned and extracted using Tesseract OCR.

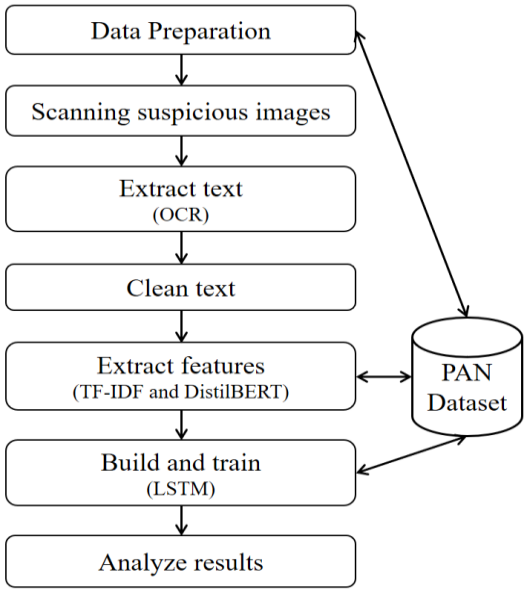

Fig. 1: The DEPTI model

Subsequently, text will be cleaned up using the following process:

- Removing special characters: Regular expressions are applied to kill the unwanted characters.
- Converting text to lowercase: Switching to lowercase in order to let the user easily distinguish one word from another hence ensuring clarity in the text. Besides, uniformity is ensured and in case of words distinctions and unnecessary are reduced among the words.
- Using NLP tools: to obtaining a much more coherent word description.
- Generating organized dataset for future used.

The DEPTI model uses combination of functions as follows:

- TF-IDF to detect duplicates and verbal features that shows the significance of words in text.
- DistilBERT to capture semantic, structural relationships, and recognize the context and the deeper meanings in the text. It helps to enhance the accuracy of plagiarism detection.

- LSTM to build a model as convolutional neural networks to overcome problems of fading, rapid explosion of values and are useful in generating long-term time series, and long texts while maintaining a high degree of accuracy in predictions.

This combination improves the accuracy of academic plagiarism detection and extends the range in regard to paraphrased, translated, or machine-generated texts. Enhancing DEPTI efficiency and accuracy of plagiarism detection systems is perfect harmony to the research objectives. These techniques are not only fast but also accurate. Employing Distil BERT enables quick processing, while LSTM and TF-IDF networks can effectively discover complex plagiarism patterns that adds to the DEPTI ability to cope with different types of academic plagiarism.

3.1. Confusion Matrix

Binary classification has four key values as Fig. 2 shows, in different rows and columns each row in the matrix represents actual class instances while each column represents predicted class instances.

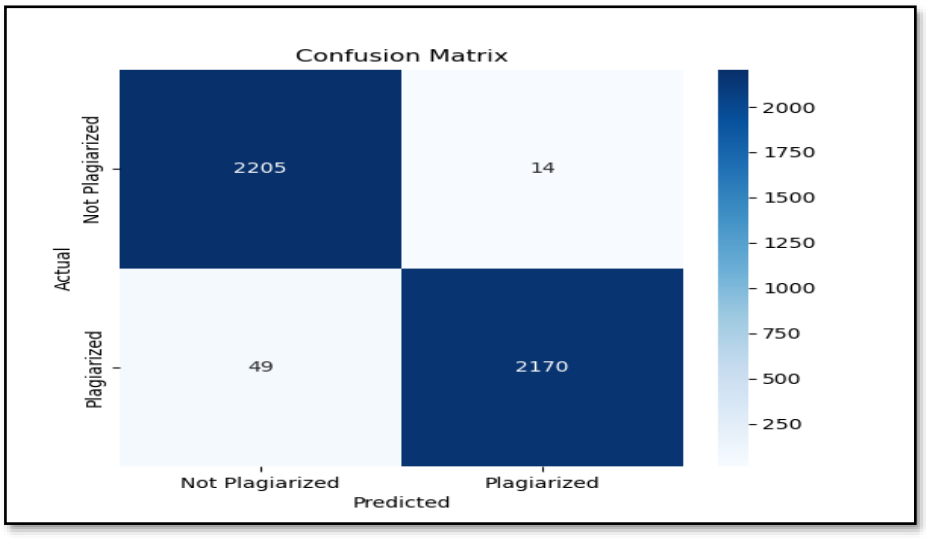

Fig. 2: Confusion matrix results

Since accuracy metrics are used to assess and test the model's performance and classification efficiency, the other performance metrics precision, recall, and f-measure that are associated with accuracy were also employed. The following definitions are used to determine these evaluation criteria, which are based on the confusion matrix:

1. True Positives (TP): Instances are identified as class 1 and indeed class 1.
2. True Negatives (TN): Instances from class 1 but are predicted to be class 2.
3. False Positives (FP): Instances are thought in class 1 but actually class 2.
4. False Negatives (FN): Situations are thought in class 2, but are actually class 1.

Table 2: Confusion matrix summary

Actual / Predicted	Not plagiarized	Plagiarized
Not Plagiarized	TN	FP
Not Plagiarized	FN	TP

The values in the confusion matrix are: TN = 2205, FP = 14, FN = 49, and TP = 2170.
In order to verify the system, an assortment of performance the model metrics is utilized as Fig. 3 shows:

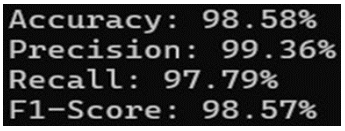

Fig. 3: Results of DEPTI

3.2 Building the LSTM

The extracted features are combined as vector representations of the extracted features. Each document is represented by a sequence vector that is composed of statistical word from TF-IDF and contextual embedding from DistilBERT. While LSTM networks are typically used with temporal sequences, therefore, each sequential feature vector will be treated as a one-step sequence. This gives the LSTM the ability to capture complex, non-linear interactions between input features.

The LSTM has a hidden dimension of 256, which is transformed into another dimension of 128 using a linear transformation. A non-linear function known as Rectified Linear Unit (ReLU) is also used, which assigns zero values to negative values and can represent more complex patterns [23]. A dropout layer is used to enhance the overfitting problem by randomly dropping some neurons. Finally, another feed-forward layer will be used to calculate the output which will be passed through a Sigmoid function, which will give the expected probability of plagiarism scale 0,1.

The training process, provides an accurate measure of the gap between probabilistic predictions and the actual values, helping the model optimize itself accurately and efficiently. Subsequently, Adam with Weight decay (AdamW) is used which is a modern version of the Adam optimizer where it handles weight decay more efficiently, solves structural problems in regular Adam regarding weight reduction, and delivers better performance, especially in modern deep learning models. Therefore, it is currently the preferred choice for advanced applications such as image classification, natural language processing, and plagiarism detection where the model is trained over five-time epochs.

To evaluate the effectiveness of the DEPTI, a series of systematic experiments were conducted. Apart from building and training an LSTM model, the system depended on merging optical character recognition Tesseract OCR with sophisticated text understanding models employing deep learning methods, particularly contextual processing models. The performance of DEPTI was evaluated using the following metrics: accuracy, precision, recall, and F1-score, in addition to the area under the receiver operating characteristic curve.

3.3 Accuracy

The expansive capabilities of DL and NLP in detecting academic plagiarism surpass those of conventional methodologies, thereby motivating their integration into higher education institutions as a means of safeguarding the authenticity of scholarly research and scientific writing. Accuracy is the correct identification of positive and negative instances (plagiarized and non-plagiarized) out of the entire sample set.

$$\text{Accuracy} = (TP + TN) / (TP+TN+FP+FN)$$
$$= (2170+2205) / (2170+2205+17+19) \approx 98.58\%$$

This indicates that the model correctly identifies approximately 98.58% of all texts (plagiarized + non-plagiarized).

3.4 Precision

Precision is the set of all texts marked as plagiarized.

$$\text{Precision} = TP/(TP+FP)$$
$$=2172 / (2170+14) \approx 99.36\%$$

This indicates that the correct classification accuracy for the 99.36% of the plagiarized samples was achieved. Basically, the DEPTI model detects an original text to be a plagiarism case with low False and Positive rate.

4 Results

The DEPTI model is developed and trained where the metric values are successfully achieved and Fig.d out the accuracy, confusion matrix, precision, and receiver operating characteristic curve. A summary of several well-known algorithms comparing with DEPTI is listed in Table 3, the results with varying accuracy depending on the methodology and datasets used in recent plagiarism detection studies.

Table 3: A compression between DEPTI and other Models

Methodology	Algorithm	Dataset	Accuracy
plagiarism detection framework with LESK (Word Sense Disambiguation), Semantic Analysis, and a Modified SVM classifier	LESK for word-sense detection, SVM Classifier combine multiple similarity features (lexical, syntactic, semantic)	PAN 2012,13,14	92.84%
Candidate Retrieval and Detailed Analysis in addition to Multilingual Word Clusters	Tokenization, lemmatization XLM-R fine-tuned for cross-lingual text alignment/translation detection	PAN-PC-11 and MRPC dataset	recall 79%, precision 85%
Document Segmentation, Feature Extraction, Computing Similarity Scores, Analysis of n-gram Length Impact	TF-IDF Weighting, n-grams Analysis, Cosine Similarity, Granularity and Plagdet Metrics	PAN-PC-11	Precision: 0.3381 Recall: 0.3117 F-measure: 0.3244 PlagDet: 0.3189
DEPTI	DistilBERT, LSTM, and TF-IDF	PAN-PC-11	Accuracy: 98.58% Precision: 99.36%

As well as algorithms listed in Table 3 that have achieved different accuracy scores. In [17] reaches 93% resolution using BERT-based attention LSTM on SNLI, MSRP and SemEval, while [18] lexical and semantic techniques (n-grams, TF-IDF, possibly BERT-based) and gets accuracy rates of 92%, and [19] determines the degree of PlagDet at 92.86%, unlike [20], while the highest accuracy was recorded in [18] who used OCR, Cosine, Jaccard, BERT, N-gram techniques where they achieved 85-95% success range depending on text quality and image complexity, and in [21] and [22] they got low performance comparing with others. A work is done in [15] with high accuracy but it works with clear text that is directly written. However, the DEPTI shows a great results of accuracy which is 98.58% and 99.36% precision comparing with other models.

5 Conclusion and Future Work

The DEPTI model, which is based on Tesseract and NLP, is capable of extracting contents from images, also it uses DiscilBERT and TF-IDF features on the one hand and on the other hand, it

classifies using a 256-unit LSTM with ReLU, AdamW, and the binary cross-entropy. Consequently, it delivers the accuracy of 98.58%, precision 99.36%, recall 97.79%, an F1-score of 98.57%, and 1.00 AUC of receiver operating characteristic, almost perfect discrimination ability between plagiarized and non-plagiarized text. These results are encouraging and indicate that the DEPTI is highly effective in detecting plagiarism in texts embedded within images.

It is recommended that the system be integrated with conventional text plagiarism detection tools like Turnitin or Grammarly as a future work. By highlighting the found similarities between the extracted texts and reference texts, explainable AI's explanation of detection results contributes to increased system confidence.

ACKNOWLEDGEMENTS

The author is grateful to the Deanship of Scientific Research at Jadara University for providing financial support for this publication.

References

[1] Leong, W. Y., & Zhang, J. B. (2025). AI on academic Integrity and Plagiarism Detection. ASM Science Journal, 20(1), 75.

[2] Quidwai, M. A., Li, C., & Dube, P. (2023). Beyond Black Box AI-generated Plagiarism Detection: From Sentence to Document Level. *arXiv preprint*.

[3] Moodley, K., & Nhavoto, R. (2023). An Evaluation of the use of Turnitin as a Tool for Electronic Submission, Marking, and Feedback in Higher Education in South Africa: Students' perspective. *Proceedings of The World Conference on Education and Teaching*, *1*(1), 31–43.

[4] Naik, R. R., Landge, M. B., & Mahender, C. N. (2015). A Review on Plagiarism Detection Tools. International Journal of Computer Applications, 125(11), 16-22.

[5] Alenezi, F. (2023). Artificial intelligence creates plagiarism or academic research? *European Journal of Arts, Humanities and Social Sciences*, *1*(3), 43–48.

[6] Maurer, H. A., Kappe, F., and Zaka, B. (2006). Plagiarism – A survey. Journal of Universal Computer Science, 12(8), 1050–1084.

[7] Rumanovská, Ľ., Lazíková, J., Takáč, I., & Stoličná, Z. (2024). Plagiarism in the academic environment. Societies, 14(7), 128.

[8] Alkhalifa, A.K., Zaqaibeh, B., Hassine, S.B.E.N.H.A.J., Yahya, A.E., Almukadi, W.S., Sorour, S., Alzahrani, Y., Majdoubi, J. (2024), "Motion Target Monitoring and Recognition in Video Surveillance using Cloud-Edge-IoT and Machine Learning Techniques", *Fractals*, Vol. 32, Issue 9-10, No. 2540013.

[9] Alomari, S., Alzboon, M., Zaqaibeh, B., Al-Batah, M.S. (2019), "An Effective Self-Adaptive Policy for Optimal Video Quality over Heterogeneous Mobile Devices and Network Discovery Services", *Applied Mathematics and Information Sciences*, Vol. 13, Issue 3, pp: 489-505.

[10] Srinivas, T. A. S., & Bharathi, M. (2023). Turnitin: The good, the bad, and the unseen dimensions. *Research and Applications of Web Development and Design*, *7*(1), 31–39.

[11] Mutsaddi, A., & Choudhary, A. P. (2025). Enhancing plagiarism detection in Marathi with a weighted ensemble of TF-IDF and BERT embeddings for low-resource language processing. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages* (pp. 89–100). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2024). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Posted October 11, 2018.

[13] Althunibat, Q. Alsawareah, B. Maidin, S, S. Hawashin, B. Jebril, I. Zaqaibeh, B. and Al-khawaja. H, A. (2024). Detecting ambiguities in requirement documents written in Arabic using machine learning algorithms, International Journal of Cloud Applications and Computing (IJCAC), Vol. 14, No. 1, pp: 1-19.

[13] Rosu, R., Stoica, A. S., Popescu, P. S., & Mihăescu, M. C. (2020). NLP based deep learning approach for plagiarism detection. International Journal of User-System Interaction, 13(1), pp: 48-60.

[15] Abdulkhaleq Q. A., Alshammari, A. Zaqaibeh, B. Swaileh, M. Allafi, R. Alazwari, R. Aljabri, Nouri, A, M. (2024). Enhancing Natural Language Processing Using Walrus Optimizer with Self-Attention Deep Learning Model in Applied Linguistics, World Scientific Publishing Co. Pte. Ltd., vol. 32, pages 1-19.

[16] Mansoor, M., & Al Tamimi, M. (2022). Plagiarism detection system in scientific publication using LSTM networks. *International Journal of Technical and Physical Problems of Engineering*, *4*(4), 17–24.

[17] Moravvej, S. V., Mousavirad, S. J., Oliva, D., & Mohammadi, F. (2023). A novel plagiarism detection approach combining BERT-based word embedding, attention-based LSTMs and an improved differential evolution algorithm. *arXiv preprint*.

[18] Kumar, P. S., and Prasad, K. (2024). Integrating OCR and NLP Techniques for Accurate Text Extraction and Plagiarism Detection in Image-Based Content. Library of Progress-Library Science, Information Technology and Computer, 44(3).

[19] Upadhyay, D. K., & Sinha, K. (2024). Enhancing academic integrity: An analysis of advanced techniques for plagiarism detection using LESK, Word Sense Disambiguation, and SVM. *International Journal of Experimental Research and Review*, *39*, 92–108.

[20] Kulkarni, S., Govilkar, S., & Amin, D. (2023). Text similarity from image contents using statistical and semantic analysis techniques. In D. C. Wyld et al. (Eds.), *Proceedings of AI&FL, SCM, NLPTT, DSCC 2023* (pp. 37–46). AIRCC Publishing Corporation.

[21] Avetisyan, K., Malajyan, A., Ghukasyan, T., & Avetisyan, A. (2023). A simple and effective method of cross-lingual plagiarism detection. *arXiv preprint*.

[22] Kadhim, N. J. (2024). Mining deviations in document writing style through vector dissimilarity. *Iraqi Journal of Science*, *65*(4), pp. 2322–2331.

[23] Nair, V., and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th International Conference on Machine Learning pp. 807-814.

[24] Malashin I., Masich I., Tynchenko V., Gantimurov A., Nelyub V., and Aleksei Borodulin, "Image Text Extraction and Natural Language Processing of Unstructured Data from Medical Reports", machine learning and knowledge extraction, Vol. 6(2), pp: 1361-1377, 2024.