The Potential of Deep Learning and Natural Language Processing Methods in Revealing Academic Plagiarism

https://doi.org/10.63962/OSED8159

Ahmad Alhami
Department of Robotics and AI
Jadara University
Irbid, Jordan
ahmadalhami1977@gmail.com

Belal Zaqaibeh
Department of Software Engineering
Jadara University
Irbid, Jordan
zaqaibeh@jadara.edu.jo

***Abstract*—Plagiarized text in images will be detected, assessed, and evaluated through deep learning and natural language processing techniques. A new Detecting Embedded Plagiarized Text in Images (DEPTI) model is proposed to detect embedded plagiarized text in images where it is expected to show a high performance and accurate results. The DEPTI will include functions that use algorithms like DistilBERT and Long Short-Term Memory (LSTM) and Term Frequency–Inverse Document Frequency (TF-IDF) dataset. The expected outcomes of DEPTI may prove that it has the power of recognizing paraphrased, translated, and AI-generated content and will be done with a quite large of accuracy.**

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

I. INTRODUCTION

The aim of this study is to develop an end-to-end model that can detect text-based plagiarism in images. Therefore, the text in images will be digitalized using OCR, then extract text features by using TF-IDF and Destilbert, building LSTM approaches to get accuracy and high performance. Subsequently, the objectives of this research study are listed as follows:

- Design a deep learning model that can detect plagiarized text in images.
- Develop plagiarism detection model that can ease plagiarism detection with high performance.

Plagiarism can be strictly defined as the act of using the work, ideas, or texts of others without proper acknowledgement of the authors, and this is taken as a serious violation of academic integrity [1]. Uncontrolled internet access and various software applications for writing have led to increase plagiarism, which is now described as the most common problem in the educational and research era [2].

The proposed method will be explained in the following steps which are: A dedicated dataset will be utilized for the purpose of plagiarism detection, and each of the files that are identified as being plagiarized will be represented as a single image to create the research scenario. Subsequently, optical character recognition will be utilized for the extraction of text from images.

II. LITERATURE REVIEW

The continuous development of current technology has significantly reduced the efficiency of plagiarism detection tools, where most plagiarism detection tools focus on detecting plagiarism in texts, while most of these tools are unable to handle text stored in images or embedded in images, or scanned files such as PDF files, which allows these systems to bypass these texts inside images without any plagiarism scan. Therefore, it is essential to develop an approach that facilitates the extraction of the text from images and converts these texts into digital files, which can be checked to identify a copied material that is concealed into image files or scanned documents. This method is of great importance in scientific and research fields where correctness, transparency, and respect for the rights of the author are one of the foremost [4].

A study in [7] explores the use of machine learning algorithms to automate the assessment of requirements expressed in natural language. The study aims to compare various machine learning algorithms according to their abilities in classifying requirements

A study has investigated AI tools for plagiarism detection such as Copyleaks, Grammarly, and Turnitin [1]. The research was done at an academic dataset collection that included many types of sources and aimed at assisting in the issue of authenticity of educational materials and combating the problem of plagiarism. The author had at his disposal 200 cases of plagiarism, of which 50 were straightforward copies, 75 were rewritten versions and the remaining were machine generated. These samples were employed on three different plagiarism checkers, namely, Turnitin, Grammarly, and Copyleaks, and a conclusion that the Copyleaks software can produce 92% of successful detection of the summarized passages was reached.

A study introduced a novel approach to identify AI-generated text such as an academic paper in response to the rise in plagiarism from large language models (LLMs) [2]. Highly innovative methods are used like GPT-3.5 and T5 Paraphrasing [5] that is apart from creating a number of questions, also getting the cosine similarity to calculate the similarity. Their work achieved a 94% precision rate.

MIT Plagiarism Detection Dataset was represented in [3] to discuss the problem of plagiarism identification in low-resource languages such as Marathi, where they experimented on using TF-IDF in combination with BERT embedding as well as the MahaSBERT model. It is the embodiment of the principle of representing the text with the number directly. Thus, the model becomes increasingly intelligent because it is not only learning from the similarity scores which are pre-determined but also from the natural language patterns.

BERT has the MahaSBERT-STS model, which is an implementation of the MahaSBERT model specifically designed for generating embedding for semantic textual similarity [6], as one of their approaches, they also run TF-IDF Vectors to get the statistical representation. This led to a BERT F1 score of 88.5%.

Table 1 provides an overview of the significant points in the studies that have been done already in a simple manner in the first instance.

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%

Table 1 illustrates the accuracy of deep learning methods, especially BERT and LSTM models, when they want to get a very high level of accuracy in comparison to traditional ones. Besides, it has been mentioned that the GPT-3.5 and T5 models did a great job in detecting AI-generated texts. 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.

III. THE METHODOLOGY

Detecting Embedded Plagiarized Text in Images (DEPTI) is the proposed model that will rely on a dataset called PAN-PC-11, which is one of the most important free datasets used to evaluate plagiarism detection algorithms. The dataset consists of original and suspected files, the proposed model DEPTI can extract text from images and then apply the data cleaning process such as deleting punctuation and converting text into small letters, the text will be processed through features extracted using TF-IDF to extract statistical features and use DistilBERT to extract text context [8] as Fig. 1 shows.

- Most plagiarism detection systems focus on digital texts, without paying attention to examining non-textual data.
- Plagiarism rates have been observed to be high in academic and research communities, both in texts and even in images that have been embedded with text such as screenshots and scanned documents, and detecting them is costly and time consuming.

The DEPTI model is an advanced approach that detecting plagiarism in images containing texts to evaluate automated plagiarism detection algorithms. The initial step was converted suspicious text files into images, each file being converted into a single image. The files were read with Tesseract OCR, and then texts were gotten, purged, stop words, and punctuations were removed. Text features after extracting TF-IDF in DEPTI were used to show the significance of words in the text. Furthermore, the DEPTI will use DistilBERT model as a function to recognize the context and the deeper meanings in the text and also the LSTM to build a model as convolutional neural networks. The combination of functions is listed down as follows:

- TF-IDF to detect duplicates and verbal features.
- DistilBERT to capture semantic and structural relationships in the text. It helps to enhances the accuracy of plagiarism detection or similarity recognition.
- LSTM to build convolutional neural networks to overcome the 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.

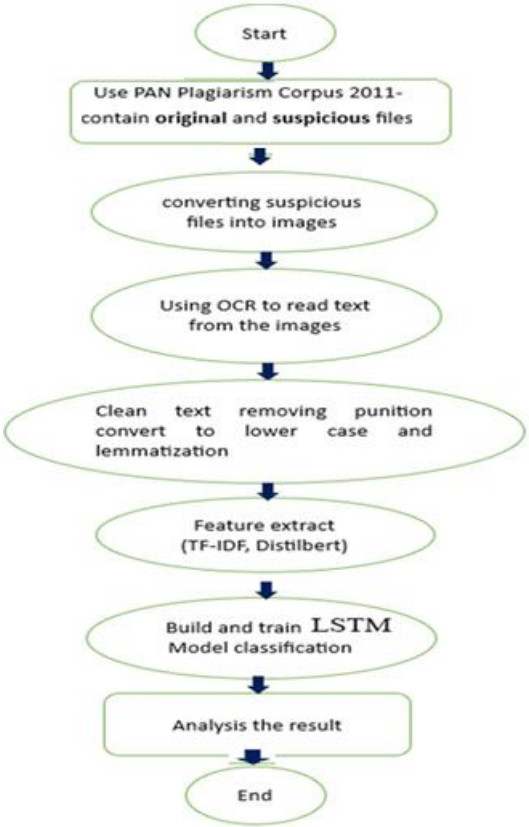

Fig. 1: the DEPTI model

IV. Expected Results

The findings suggest that the far-reaching abilities of deep learning and NLP in detecting academic plagiarism exceed those of traditional methods, thus prompting the use of these techniques in universities and colleges as a means to secure the genuineness of research and scientific writing. Subsequanlty, there is a need to:

- Developing a model that has the ability to detect plagiarism in texts within images with high accuracy and efficiency
- Bridging the gap in plagiarism detection systems in texts within images, as many of these systems cannot do this
- Improving the accuracy achieved by training the model

V. Conclusion

The DEPTI model, which is based on Tesseract and NLP and it will be capable to extract contents from images, also it will use DiscilBERT and TF-IDF features. it will deliver a high accuracy and precision. Generally, the ability between plagiarized and non-plagiarized text will be discovered. These expectations are encouraging that the DEPTI will have a highly effective in detecting plagiarism in texts embedded within images. It is recommended that the DEPTI be integrated with conventional text plagiarism detection tools like Turnitin or Grammarly.

References

[1] Leong, W. Y., & Zhang, J. B. 2025, AI on Academic Integrity and Plagiarism Detection. Learning, 92(12), 75.

[2] Quidwai, M.A., Li, C., & Dube, P. (2023). Beyond Black Box AI-Generated Plagiarism Detection. arXiv preprint.

[3] Mutsaddi, A., & Choudhary, A. (2025). Enhancing Plagiarism Detection in Marathi with a Weighted Ensemble of TF-IDF and BERT Embeddings for Low-Resource Language Processing. arXiv preprint arXiv:2501.05260.

[4] Aditya, T., Srinivas, T. A. S., and Munnuru, B. (2023). Turnitin: The Good, the Bad, and the Unseen Dimensions. Zenodo.

[5] Hassan ipour, S., Nayak, S. S., Ali Bozorgi, M. H., Keivanlou, T. D., Alotaibi, A., Joukar, F., ... & Amini-Salehi, E. The ability of Chat-GPT in paraphrasing texts and reducing plagiarism.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2024). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 1810.04805 [preprint] https://doi. org/10.48550/arXiv. 1810.04805. Posted October 11, 2018. Accessed May.

[7] Ahmad Althunibat, Bayan Alsawareah, Siti Sarah Maidin, Belal Hawashin, Iqbal Jebril, Belal Zaqaibeh, and Haneen A. Al-khawaja. 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.

[8] Avetisyan, K., Malajyan, A., Ghukasyan, T., and Avetisyan, A. 2023. A simple and effective method of cross-lingual plagiarism detection. arXiv preprint arXiv:2304.01352.