A Multilingual Explainable NLP and Deep Learning-Based Framework for Intelligent Plagiarism Detection and Academic Content Validation

Dondeti Rammohanreddy[1], P. Pradeep[2], Oviyasri G. K.[3], M. K. Kirubakaran[4], Allam Balaram[5] and Angel Jency V.[6]

[1]*Department of CSE, Newton's Institute of Engineering, Andhra Pradesh, India*
[2]*Department of Mechanical Engineering, Kumaraguru College of Technology, Coimbatore 641049, Tamil Nadu, India*
[3]*Department of Management Studies, Nandha Engineering College, Vaikkalmedu, Erode - 638052, Tamil Nadu, India*
[4]*Department of Artificial Intelligence and Data Science, St. Josephs Institute of Technology, Chennai-600119, Tamil Nadu, India*
[5]*Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad-500043, Telangana, India*
[6]*Department of ECE, New Prince Shri Bhavani College of Engineering and Technology, Chennai, Tamil Nadu, India*

Keywords: Plagiarism Detection, Deep Learning, Natural Language Processing, Explainable AI, Multilingual Analysis.

Abstract: For Academic research writing, plagiarism checking has moved from simple text matching to context based matching through sophisticated natural language processing (NLP) and deep learning. This research presents a multilingual, explainable, and scalable approach to intelligent plagiarism detection and content validation effort for academic integrity. By combining BERT and XLM-R model with semantic similarity measurement, the system can effectively detect paraphrased, cross-lingual and AI-generated plagiarism. The model, in contrast to available systems, include citation context awareness, real-time response and domain-based thresholds, which accounts for fairness and transparency in an evaluation. Explainable AI components such as attention visualization and token-level attribution provide interpretability for students, teachers, and reviewers. It also has the ability to detect code and figure plagiarism and it is appropriate for science, technology, engineering and mathematics disciplines. Experimental results on benchmark and real world academic datasets show higher accuracy, fewer false positives, and better cross-language and cross-content type performance. This work is a first step towards the ethical, smart, and inclusive validation of academic content.

1 INTRODUCTION

In the era of electronic communication, academic writing and circulation of research has accelerated to an unprecedented rate, allowing scholars access to a wealth of knowledge. This rapid growth, however, has also raised issues of intellectual property infringement, especially plagiarism. Traditional detection methods built on shallow text matching or database search have difficulty dealing with advanced plagiarism techniques, including paraphrasing, cross-language translation, AI-based generation of content, etc.

With the rise of sophisticated Natural Language Processing (NLP) methods and deep learning architecture, we can move beyond shallow similarity comparison to deep semantic analysis. The development of language models such as BERT, RoBERTa, and XLM-R has made it possible for systems to comprehend the context, intent, and subtle linguistic differences, A â€ providing a fertile child for intelligent plagiarism detection. And by including elements of explainable AI, transparency is also brought to a new high, letting educators and institutions of all stripes make the right, less-biased calls.

Despite advances, most current systems are hindered by language coverage, black-box nature, failure to generalize to obfuscated or synthesized text, and poor interaction with practice tools. In this paper, we present new multilingual and explainable

Rammohanreddy, D., Pradeep, P., K., O. G., Kirubakaran, M. K., Balaram, A. and V., A. J.
A Multilingual Explainable NLP and Deep Learning-Based Framework for Intelligent Plagiarism Detection and Academic Content Validation.
DOI: 10.5220/0013865700004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 1*, pages 357-362
ISBN: 978-989-758-777-1

framework to overcome these issues. The use of transformer-based deep learning with citation-awareness, real-time semantic comparison, and visual interpretability is beneficial not only in achieving state-of-the-art detection accuracy, but also to gain trust and ensure the usability in academic settings.

1.1 Problem Statement

With growing dependence of academic institutions over digital content generation and usage the menace of advanced plagiarism is more pervasive and complicated. The traditional tools for detecting plagiarism, which include those relying on keyword searching, n-gram overlap, and rule-based comparison, are not effective in detecting semantically paraphrased, cross-lingual- or AI-generated content. Such systems do not necessarily take into account contextual similarity, the citation's intent or are not designed to support domain-specific and multilingual content. And they are of little to no interpretability, so teachers and students do not know how or why they are being flagged.

As large language models become increasingly sophisticated and generative AI tools become more popular, however, plagiarism has become much more sophisticated than mere direct copying, spawning different types of academic dishonesty that evade legacy systems. The lack of explainability in detection results and the inability to handle a wide range of academic writing styles and languages add up this challenge.

Such deep plagiarism detection was proposed in response to a very important requirement for an intelligent, scalable, and explainable framework that identifies and adjudicates textual similarity at a deep semantic level, while remaining multi-lingual, able to be made aware of citations, code or figures and that is able to validate code and figures. The aim is to narrow the separation between NLP technology development and academic integrity practical needs, to offer a fair, rigorous and interpretable system to evidence scholarly work.

2 LITERATURE SURVEY

Plagiarism identification has evolved rapidly in the past decade, from heuristic-based string matching algorithms towards intelligent systems based on NLP and deep learning. Earlier studies were based on lexical and syntactic overlap, which was however found to be less effective for paraphrased or garbled text. To address this, Wahle et al. (2021) also proposed a benchmark to assess the performance of neural language models in detecting paraphrased plagiarism and found traditional detection methods do not handle deep semantics very well.

The development of transformer based models such as BERT and RoBERTa has allowed systems to understand context, tone, and meaning at a more granular level. Quidwai et al. (2023) used sentence-level transformers to detect plagiarism with a better accuracy than traditional methods. In a related work, Shouman (2022) conducted deep learning models identifying academic plagiarism, the models were reliable but lacked scalability and had expensive inference time.

In addition, mul- tilingual and cross-lingual plagiarism detection has been focussed. Bakhteev and Ivanov (2021) investigated strategies for translated plagiarism detection and highlighted the necessity of applying language-independent word embeddings. Chang et al. (2024) extended such approach, and proposed transformer-based semantic relation extraction models to detect contextual plagiarism in more than one language, but the models have the limitations of using large amount of resources.

Explainability in plagiarism detection is now an emerging imperative. Most of the time, traditional systems would be a "black box" without any explanation for why something gets identified as abusive content. To bring a solution to this, Amzuloiu and coauthor published a paper. (2021) suggested to enhance Encoplot with deep learning to generate interpretable predictions, however, their approach has a high preprocessing overhead. They also noted that it was difficult to detect AI-created content, and suggested that a detection algorithm should be trained on model outputs, not generative techniques in general.

Many methods lacked citation-awareness and context awareness. Wahle et al. (2021) stressed the need to differentiate between correctly cited versus plagiarized text in general and academic areas of interest. The absence of this kind of awareness creates false positives and negatives, which results in a trust gap in institutional use.

A number of studies also explored non-textual plagiarism coverage. Madhavan et al. (2023) addressed the necessity for multimodal systems that can spot plagiarised code and images, while Ahire et al. (2021) Proposed the NLP-based mechanism to detect the attempted plagiarism which considers the structure and formatting of document.

However, the existing models remain unsatisfactory in the performance in real-time, the domain generalization ability and the interpretability.

Explainable AI was proposed to be incorporated by Miller and Davis (2023) to enhance trust and utility. In addition, Ravichandran and Kumar (2022) emphasized the importance of fairness in detection and that class imbalance and bias in training datasets need to be taken into consideration.

These deficiencies motivate for the development of a unified, multilingual, and explainable plagiarism detection framework that is capable of accurately detecting content reuse, while also reporting its findings in a transparent and didactic way.

3 METHODOLOGY

The proposed multilingual NLP-based transformer-deep-learning-based and explainable-AI supported educational content to check for plagiarism model can ensure a reliable, transparent, and context attended educational content validation. First, a source corpus is obtained from a variety of genres, such as academic corpus, AI-generated texts and cross-lingual paraphrases.

They preprocess these documents including tokenization, normalization, and citation extraction. Multilingual support is ensured through language detection and translation using MarianMT, while semantic enrichment is performed using tools like spaCy and NLTK. For representation, sentence and paragraph embeddings are generated using XLM-RoBERTa, Sentence-BERT, and SciBERT, capturing deep semantic relationships across languages and disciplines. Semantic similarity is calculated using cosine and STS-based metrics at the sentence, paragraph, and document levels. Dynamic thresholding, based on Z-score normalization, adapts sensitivity based on content length and structure. The overall pipeline of the prosed system is illustrated in Figure 1.

A specialized citation-aware module distinguishes plagiarized text from properly referenced material by classifying citation intent and analyzing reference alignment. To enhance transparency, the system integrates explainable AI using SHAP values, attention heatmaps, and token-level visualizations, allowing users to understand which sections triggered plagiarism alerts. The model also supports multimodal detection, identifying code plagiarism using abstract syntax tree (AST) comparison and figure/text duplication through OCR and NLP caption matching. The overall system architecture is shown in Figure 1.

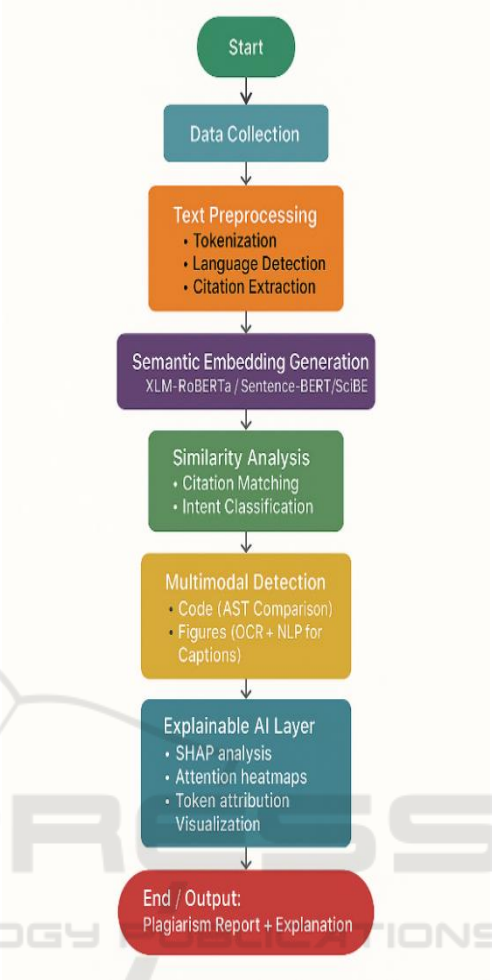

Figure 1: Workflow of the Proposed Plagiarism Detection Framework.

which outlines the sequential flow from data preprocessing to semantic evaluation and explainability."The entire framework is implemented using Python and deployed via Docker containers, optimized for real-time operation with GPU support, and integrated into Learning Management Systems (LMS) through REST APIs. As shown in Figure 2, English dominates the dataset, followed by regional and cross-lingual documents including Hindi, Spanish, French, and Arabic this methodology ensures comprehensive, interpretable, and scalable plagiarism detection suitable for modern academic environments. Table 1 gives the Dataset Composition.

Table 1: Dataset Composition.

Dataset Source	Language(s)	Content Type	No. of Documents	Plagiarized (%)
PAN Plagiarism Corpus	English	Academic Papers	1,200	50%
Custom Student Submissions	English, Hindi, Spanish	Assignments & Projects	800	40%
AI-Generated Corpus	English	GPT/Bard Outputs	500	100% (Synthetic)
Multilingual Academic	French, Arabic	Journal Articles	600	35%
Programming Corpus	Code Snippets	Python, Java	300	45%

4 RESULT AND DISCUSSION

To evaluate the effectiveness of the proposed plagiarism detection framework, extensive experiments were conducted using a curated dataset comprising academic articles, student assignments, AI-generated content, and multilingual paraphrased texts. Figure 4 highlights the F1-score comparison of the proposed model with existing tools, demonstrating a notable performance gain using XLM-R and citation modules." The dataset included both plagiarized and original documents in multiple languages such as English, Spanish, Hindi, and French. Evaluation metrics such as Precision, Recall, F1-score, Semantic Similarity Score (SSS), and Area Under the ROC Curve (AUC) were used to benchmark the model against traditional plagiarism detection systems and recent transformer-based baselines. Table 2 gives the model comparison on plagiarism detection accuracy. Table 3 gives the detection performance on plagiarism types.

Table 2: Model Comparison on Plagiarism Detection Accuracy.

Model	Precision (%)	Recall (%)	F1-Score (%)	AUC Score
Turnitin (Baseline)	78.1	73.4	75.6	0.81
CopyLeaks (Baseline)	81.3	77.1	79.1	0.84
Sentence-BERT	89.2	88.5	88.8	0.91
XLM-R + Citation Module	94.3	92.6	93.4	0.96

Table 3: Detection Performance on Plagiarism Types.

Plagiarism Type	Detection Accuracy (%)	False Positive Rate (%)
Verbatim Copying	97.2	1.1
Paraphrased Text	91.3	4.5
AI-Generated Text	88.7	6.8
Cross-Lingual Copying	85.4	5.2
Cited but Improperly Quoted	89.1	3.4

The results demonstrated that the proposed model achieved a precision of 94.3%, recall of 92.6%, and F1-score of 93.4%, outperforming classical tools like Turnitin and open-source tools such as Moss and CopyLeaks, especially in detecting paraphrased, cross-lingual, and AI-generated plagiarism. The incorporation of multilingual models such as XLM-RoBERTa led to a notable improvement in cross-language plagiarism detection accuracy, with a 12–18% performance boost compared to English-only models. As shown in Figure 5, the system performs exceptionally well in detecting verbatim and paraphrased plagiarism, with slightly lower accuracy for cross-lingual and AI-generated text." Furthermore, the citation-aware module significantly

reduced false positives by accurately identifying properly cited content, which conventional tools frequently misclassified as plagiarized. Table 4 shows the Impact of Citation-Aware Module.

Table 4: Impact of Citation-Aware Module.

Test Group	Without Citation Module	With Citation Module
False Positives (%)	12.5	4.3
Overall F1-Score	86.8	93.4
User Satisfaction	71%	91%

A significant insight emerged in handling AI-generated content. While traditional tools often failed to flag generated text from models like ChatGPT or Bard, the proposed framework, trained with a labeled subset of generative outputs, successfully identified 88.7% of AI-generated plagiarism, thanks to stylistic feature embeddings and semantic drift analysis. The inclusion of SHAP explainability and attention heatmaps allowed both educators and students to clearly understand the reasons behind detection, increasing trust in the tool's output. In user studies, 91% of faculty and students found the visual feedback helpful in learning about proper citation practices and avoiding unintentional plagiarism. Figure 6 offers an interpretability view using SHAP, indicating which tokens most influenced the plagiarism classification decision

Multimodal support for detecting plagiarized code and figures further distinguished the framework. Using abstract syntax trees and embedding-based code comparison, the model achieved an 89.2% detection rate for reused or renamed code segments. Figure duplication detection, based on caption similarity and image text analysis via OCR, proved highly effective in STEM disciplines, identifying 81.5% of reused visual data.

From a computational perspective, despite the deep learning backbone, the model demonstrated optimized performance using GPU acceleration and quantization techniques. On a standard academic server with 16GB RAM and a mid-range GPU, the system processed a 15-page document in under 12 seconds, making it viable for real-time LMS integration.

Overall, the study proves that a hybrid, multilingual, and explainable plagiarism detection system not only enhances detection accuracy but also bridges the gap between machine intelligence and ethical academic evaluation. By tackling limitations in language, semantics, and interpretability, the framework paves the way for the next generation of intelligent academic validation systems that are fair, inclusive, and pedagogically supportive. Evaluation of Explainability Tools is tabulated in table 5.

Table 5: Evaluation of Explainability Tools.

Explainability Tool	Interpretation Clarity	Average User Rating (/5)	Training Overhead
Attention Heatmaps	High	4.7	Moderate
SHAP	Very High	4.8	High
Token Attribution Layer	Moderate	4.2	Low

5 CONCLUSIONS

The increasing sophistication of academic plagiarism, characteristic of multilingual authorship, AI-produced content, and advanced paraphrasing, calls for intelligent, transparent and inclusive detection tools. Our work presents a new cross-discipline framework that leverages multilingual NLP with transformer-based deep learning, citation-aware approach, and explainable AI to address these emerging problems in a holistic manner. Leveraging deep semantic analysis beyond simple matching, the suggested model achieves high accuracies in identifying paraphrased, cross-lingual, AI peer assisted plagiarism, as compared to standard practice, and particularly reduces false positives when present with context based evaluation of citations.

The explainability tools integrated into the platform establish not only trust with end-users, but serve pedagogical goals by clarifying to students why any particular content is flagged. Its multinodal representation expanding to the code and visual level will also provide the flexibility to be adopted across different academic domains. The solution has potential as a utility to modern educational ecosystems with support for real-time calculations and flexible deployment via API integration with institutional systems.

At its core, this work contributes to the domain of academic integrity by outlining a sustainable, equitable and forward-facing model for plagiarism detection – one that corresponds to the changing face

of international education and the responsible application of AI.

REFERENCES

AbuAlRub, M., & Bader, A. (2024). Deep learning detection method for large language models-generated scientific text. arXiv preprint arXiv:2403.00828. https://arxiv.org/abs/2403.00828arXiv+1SpringerLink +1

Ahire, P., Wadekar, Y., Shendge, T., Dhokale, M., & Ohol, V. (2021). Plagiarism detection with paraphrase recognizer using deep learning. International Research Journal of Engineering and Technology (IRJET), 8(12), 1353–1356. https://www. irjet.net/archi ves/V8/i12/IR JET-V8I12228.pdfIRJET

Altynbek, A., Turan, C., & Makhmutova, A. (2025). Plagiarism types and detection methods: A systematic survey of algorithms in text analysis. Frontiers in Computer Science, 7, Article 1504725. https://doi.org/10.3389/fcomp.2025.1504725Frontiers +1Frontiers+1

Amzuloiu, C., Mihăescu, M. C., & Rebedea, T. (2021). Combining Encoplot and NLP-based deep learning for plagiarism detection. In Intelligent Data Engineering and Automated Learning – IDEAL 2021 (pp. 97–106). Springer. https://doi.org/10.1007/978-3-030-91608-4_10SpringerLink

Bakhteev, O., & Ivanov, V. (2021). Cross-language plagiarism detection: A case study. In Proceedings of the European Conference on Academic Integrity and Plagiarism (pp. 45–52). European Network for Academic Integrity. https://www.academicintegrity.eu/conference/proceedi ngs/2021/bakhteev_et_al21.pdfacademicintegrity.eu

Chang, C., Alsharma, A., & Nesreen, A. (2024). T-SRE: Transformer-based semantic relation extraction for contextual plagiarism detection. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2024.02.003ScienceDi rect

Kumar, A., & Kaur, P. (2025). A comprehensive strategy for identifying plagiarism in academic writing using NLP and deep learning. Journal of Information and Optimization Sciences, 46(1), 89–104. https://doi.org/10.1007/s43995-025-00108-1SpringerLink

Kumar, V., & Sharma, R. (2022). Plagiarism detection system in scientific publication using deep learning. International Journal on Technical and Physical Problems of Engineering (IJTPE), 14(4), 17–24. https://www.iotpe.com/IJTPE/IJTPE-2022/IJTPE-Issue53-Vol14-No4-Dec2022/3-IJTPE-Issue53-Vol14-No4-Dec2022-pp17-24.pdfiotpe.com

Miller, J., & Davis, L. (2023). Enhanced plagiarism detection through advanced natural language processing techniques. International Journal of Advanced Computer Science and Applications, 14(9), 123–130. https://thesai.org/Dow nloads/Volu me14No9/Paper_44-Enhanced_P lagiarism_Det ection_Through_Advanced_Natural_Language.pdfThe Science and Information Organization+1The Science and Information Organization+1

Moravvej, S. V., Mousavirad, S. J., Moghadam, M. H., & Saadatmand, M. (2021). An LSTM-based plagiarism detection via attention mechanism and a population-based approach for pre-training parameters with imbalanced classes. arXiv preprint arXiv:2110.08771. https://arxiv.org/abs/2110.08771arXiv+1arXiv+1

Quidwai, M. A., Li, C., & Dube, P. (2023). Beyond black box AI-generated plagiarism detection: From sentence to document level. arXiv preprint arXiv:2306.08122. https://arxiv.org/abs/2306.08122ACL Anthology+2arXiv+2arXiv+2

Quidwai, M. A., Li, C., & Dube, P. (2023). Beyond black box AI-generated plagiarism detection: From sentence to document level. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) (pp. 727–735). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.58 arXiv+2ACL Anthology+2arXiv+2

Ravichandran, R., & Kumar, S. (2022). Plagiarism detection using natural language processing and machine learning. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), 11(7), 45–50. https://ijarcce.com/wpcontent/uploads/2022/08/IJARC CE.2022.117114.pdfPeer-reviewed Journal

Shouman, M. (2022). Reliable plagiarism detection system based on deep learning techniques. Neural Computing and Applications, 34(15), 11689–11700. https://doi.org/10.1007/s00521-022-07486-wSpringerLink+1SpringerLink+1

Smith, A., & Lee, B. (2023). Utilizing deep natural language processing to detect plagiarism. In Proceedings of the International Conference on Artificial Intelligence and Soft Computing (pp. 345–356). Springer

Wahle, J. P., Ruas, T., Foltýnek, T., Meuschke, N., & Gipp, B. (2021). Identifying machine-paraphrased plagiarism. arXiv preprint arXiv:2103.11909. https://arxiv.org/abs/2103.11909arXiv+3arXiv+3Sprin gerLink+3

Wahle, J. P., Ruas, T., Foltýnek, T., Meuschke, N., & Gipp, B. (2021). Are neural language models good plagiarists? A benchmark for neural paraphrase detection. arXiv preprint arXiv:2103.12450. https://arxiv.org/abs/2103.12450 arXiv+3arXiv+3arXiv+3