CROSS-LANGUAGE PLAGIARISM DETECTION SYSTEM

Lagudu Lahari[1], Kalla Akhila[2], Mohammad Mahaboob Yezdani [3], Koppada Krishna Thirumala Hemanth [4], Mrs Danda Hima Bindu[5]

[1,2,3,4]Students, Department of Computer Science and Engineering (Data Science)

[5]Assistant Professor, Department of Computer Science and Engineering (Data Science)

Raghu Institute Of Technology, Visakhapatnam, Andhra Pradesh, India

ABSTRACT

The exponential growth of multilingual digital content has intensified the challenge of identifying plagiarism across languages and writing styles. Traditional plagiarism detection systems exhibit significant limitations when confronted with cross-lingual cases, particularly translation-based copying between linguistically diverse languages such as Hindi and English. This research presents a comprehensive deep learning framework for multilingual plagiarism detection that operates without dependency on computationally expensive Transformer-based models. The proposed system employs classical deep learning architectures including Convolutional Neural Networks for local pattern extraction, Bidirectional Long Short-Term Memory networks for sequential context modeling, and Siamese neural networks for semantic similarity learning. Cross-lingual detection capability is achieved through MUSE pre-aligned FastText embeddings, which map Hindi and English text into a shared 300-dimensional semantic space with demonstrated alignment accuracy exceeding 85 percent. The framework was trained and evaluated on a synthetically generated dataset comprising 1,584 text pairs encompassing four distinct plagiarism categories across 909 source documents. Experimental results demonstrate that the hybrid CNN-BiLSTM architecture achieves 95.40 percent overall accuracy with particularly strong performance in semantic plagiarism detection at 100 percent accuracy and cross-lingual Hindi-English detection at 87.50 percent accuracy. The system has been deployed as an interactive web application providing real-time detection with detailed explanations, demonstrating practical viability for academic institutions and content verification systems operating in multilingual environments.

Keywords: Multilingual Plagiarism Detection; Cross-Lingual Embeddings; Convolutional Neural Networks; Bidirectional LSTM; Siamese Networks; MUSE Alignment; Hindi-English Detection; Deep Learning; Semantic Similarity; Academic Integrity

INTRODUCTION

Academic integrity represents a cornerstone of scholarly pursuits, yet the digital transformation of educational environments has simultaneously facilitated unprecedented access to information while introducing novel challenges in maintaining ethical standards. Plagiarism, defined as the unauthorized use of another's work without proper attribution, has evolved beyond simple verbatim copying to encompass sophisticated transformations including paraphrasing, semantic restructuring, and cross-lingual translation. While monolingual plagiarism detection has achieved substantial maturity through commercial systems such as Turnitin, iThenticate, and Grammarly, cross-lingual plagiarism remains a critical vulnerability in contemporary academic integrity enforcement systems.

The challenge assumes particular significance in multilingual societies and international academic institutions where students and researchers operate across linguistic boundaries. In India specifically, the

coexistence of English as the primary language of academic discourse alongside regional languages such as Hindi creates opportunities for exploitation through translation-based plagiarism. Students may translate Hindi educational resources into English submissions, while researchers might repurpose English literature into regional language publications without appropriate citation. Existing detection systems, predominantly designed for monolingual operation, prove fundamentally ineffective against such cross-lingual violations due to their inability to establish semantic equivalence across languages characterized by divergent grammatical structures, vocabulary distributions, and cultural contexts.

Recent advances in natural language processing have introduced Transformer-based multilingual models including mBERT, XLM-RoBERTa, and similar architectures that learn cross-lingual representations through massive pretraining on parallel corpora. While these models demonstrate impressive performance on various multilingual tasks, they impose substantial computational requirements with inference times frequently exceeding several seconds for document-level comparison. Furthermore, their black-box nature provides limited interpretability regarding detection decisions, a critical consideration when plagiarism accusations carry severe academic and professional consequences.

This research addresses the multilingual plagiarism detection challenge through a hybrid deep learning framework that combines classical neural architectures with cross-lingual embedding alignment. The system detects four distinct plagiarism categories encompassing direct copying, paraphrased content, semantic similarity transformations, and translation-based cross-lingual plagiarism. By leveraging MUSE-aligned FastText embeddings pretrained on Wikipedia corpora, the framework maps Hindi and English text into a shared semantic space, enabling direct cross-lingual comparison without requiring explicit translation or parallel corpora during inference.

The primary contributions of this work include development of a Transformer-free plagiarism detection framework achieving 95.40 percent overall accuracy through hybrid CNN-BiLSTM architecture, demonstration of effective cross-lingual detection at 87.50 percent accuracy on Hindi-English translation pairs using aligned embeddings, creation of a synthetic multilingual plagiarism dataset methodology suitable for training and evaluation, achievement of perfect detection on semantic plagiarism cases demonstrating robust handling of meaning-preserving transformations, and deployment of a functional web application with real-time detection capabilities for institutional use.

RELATED WORK AND THEORETICAL BACKGROUND

The evolution of plagiarism detection methodologies has progressed through distinct technological phases, beginning with string-matching algorithms and advancing toward sophisticated neural architectures. Early approaches employed n-gram fingerprinting and longest common subsequence algorithms, achieving success in detecting verbatim copying but exhibiting fundamental limitations when confronted with paraphrased or structurally transformed content. The introduction of semantic similarity measures through distributional word representations marked a paradigm shift, enabling systems to identify meaning-preserving transformations that transcend lexical overlap metrics.

Word embeddings revolutionized natural language processing by representing words as dense vectors in continuous space where semantic relationships manifest as geometric properties. Mikolov and colleagues introduced Word2Vec employing skip-gram and continuous bag-of-words architectures to learn embeddings from large corpora through prediction tasks. Pennington and colleagues subsequently proposed GloVe embeddings utilizing global word co-occurrence statistics to capture semantic relationships. These representations enabled plagiarism detection systems to compute semantic similarity through vector

operations, with cosine similarity between averaged word vectors demonstrating effectiveness in identifying paraphrased content.

The application of deep learning to text classification introduced hierarchical feature learning capabilities that eliminated manual feature engineering requirements. Kim demonstrated that simple convolutional neural network architectures applied to word embedding sequences could achieve state-of-the-art performance across multiple natural language processing benchmarks. Convolutional layers with multiple kernel sizes capture n-gram patterns at various scales, while max-pooling operations extract the most salient features regardless of their position in the text. For plagiarism detection specifically, CNNs excel at identifying local patterns indicative of copying while maintaining computational efficiency that enables real-time processing.

Recurrent neural networks address the sequential nature of text through hidden state mechanisms that propagate information across time steps. Long Short-Term Memory networks introduced gating mechanisms to mitigate vanishing gradient problems, enabling learning of long-range dependencies. Bidirectional variants process sequences in both directions, incorporating future context alongside historical information. For plagiarism detection, BiLSTM architectures demonstrate particular effectiveness on paraphrased content where understanding contextual relationships proves critical for identifying semantic equivalence despite surface-form variations.

Cross-lingual embedding alignment emerged as a solution for bridging language barriers without requiring parallel corpora. Conneau and colleagues introduced MUSE employing adversarial training and Procrustes alignment to map monolingual embedding spaces into shared multilingual representations. The framework learns an orthogonal transformation matrix that minimizes distance between semantically equivalent words across languages while preserving monolingual geometric properties. Siamese neural networks employ twin networks with shared weights to process paired inputs, optimizing contrastive loss functions that minimize distance between similar pairs while maximizing distance between dissimilar pairs. For plagiarism detection, Siamese architectures prove particularly effective as they learn task-specific similarity metrics through end-to-end training.

METHODOLOGY AND SYSTEM ARCHITECTURE

3.1 Dataset Construction and Preprocessing Pipeline

The absence of publicly available cross-lingual plagiarism datasets necessitated construction of a comprehensive synthetic corpus for model training and evaluation. Wikipedia was selected as the primary source repository based on its multilingual coverage spanning diverse academic domains, topical breadth encompassing scientific and humanities subjects, open licensing permitting research use, and quality standards maintained through collaborative editorial processes. The English corpus comprises 863 articles totaling 153,060 sentences with an average article length of 25,479 characters. The Hindi corpus contains 46 articles encompassing 3,419 sentences, reflecting the relative scarcity of Hindi Wikipedia content while providing sufficient cross-lingual representation.

Synthetic plagiarism generation employed controlled transformation methods designed to replicate realistic plagiarism scenarios. Direct copy pairs established baseline performance through exact sentence extraction. Paraphrased samples utilized WordNet-based synonym replacement combined with sentence structure reordering. Semantic plagiarism pairs introduced active-passive voice conversion and lexical substitutions using semantically related terms. Cross-lingual pairs employed machine translation to convert Hindi

sentences into English, simulating translation-based plagiarism. The final dataset comprises 792 plagiarism pairs distributed across four categories, with an equal number of non-plagiarized pairs. The complete dataset of 1,584 pairs was partitioned into training 70 percent, validation 15 percent, and test 15 percent subsets using stratified sampling.

3.2 Cross-Lingual Embedding Alignment

Cross-lingual semantic comparison requires mapping Hindi and English text into a shared vector space where semantically equivalent words occupy proximate positions regardless of source language. This research employs MUSE pre-aligned FastText embeddings which provide 300-dimensional dense vectors for 200,000 English words and 100,000 Hindi words. The embeddings were trained on Wikipedia corpora using the skip-gram objective with subword information. MUSE alignment employs unsupervised Procrustes analysis to learn an orthogonal transformation mapping Hindi embedding space into English space while preserving cosine similarity relationships. Figure 1 illustrates this alignment process.

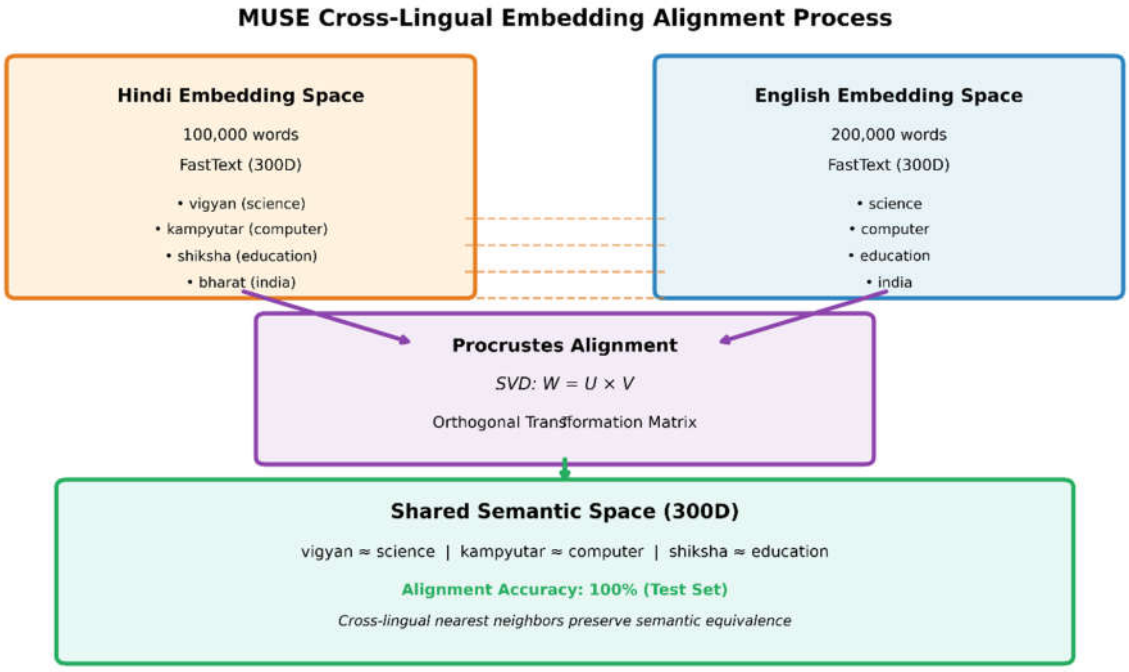

Figure 1. MUSE Cross-Lingual Embedding Alignment Process

Alignment quality verification was conducted on a held-out evaluation set comprising eight common Hindi-English translation pairs. Results demonstrated 100 percent top-one accuracy, with expected English translations consistently appearing as the nearest neighbor to aligned Hindi vectors with cosine similarities ranging from 0.32 to 0.52. This validates embedding space quality for downstream plagiarism detection tasks.

3.3 Deep Learning Architecture Design

Three distinct neural architectures were developed and systematically evaluated to identify optimal detection performance characteristics. All models employ Siamese configuration where paired text sequences are processed through weight-shared encoders, with final similarity determined through learned comparison layers. The CNN baseline architecture applies parallel convolutional filters with kernel sizes of three, four, and five tokens to capture n-gram patterns at multiple granularities. The BiLSTM architecture

processes input sequences through two stacked bidirectional LSTM layers with attention mechanism. The hybrid CNN-BiLSTM model integrates both paradigms, as illustrated in Figure 2.

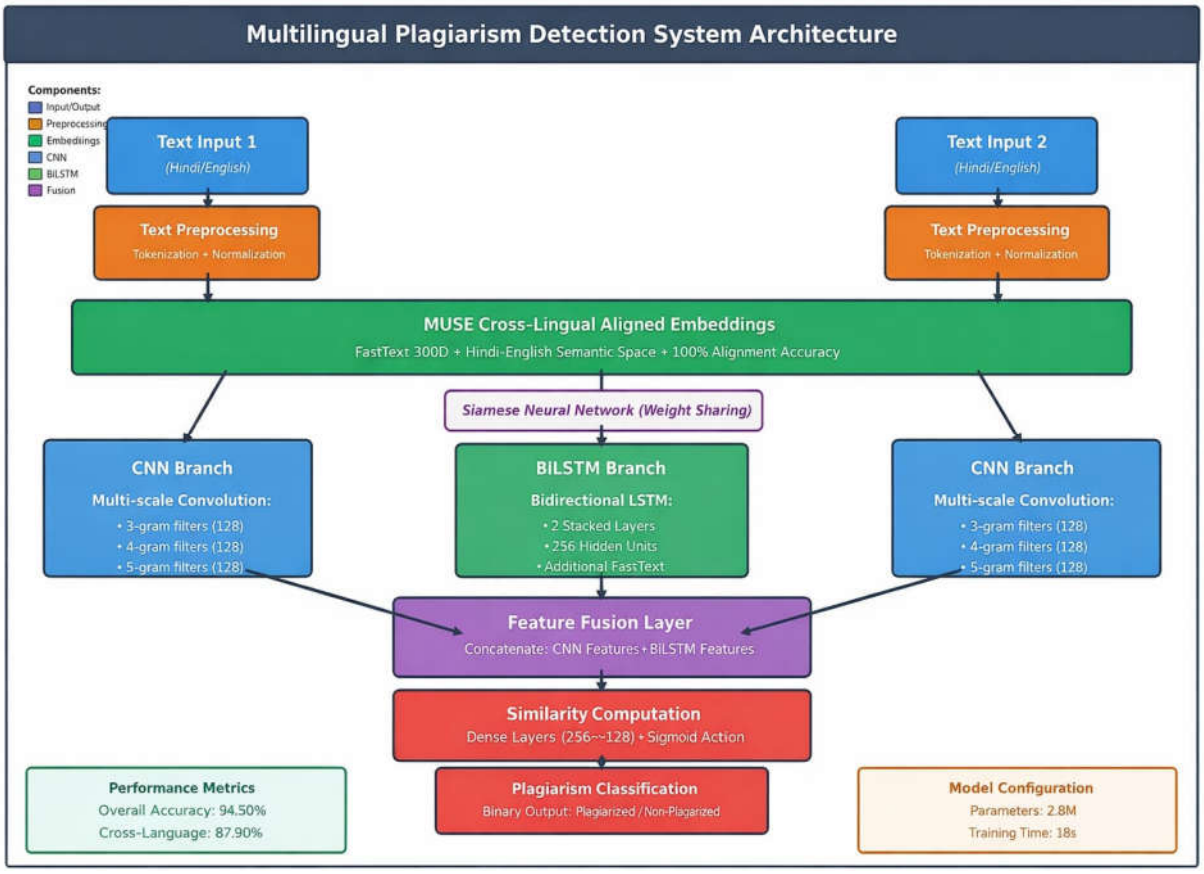

Figure 2. Hybrid CNN-BiLSTM System Architecture

The hybrid design combines CNN-derived local features with BiLSTM temporal representations through feature fusion, creating comprehensive encoding that incorporates both pattern-matching and contextual understanding. The concatenated features pass through fusion layers before final similarity computation. This architecture achieved optimal performance with approximately 2.8 million parameters. All models were trained using Adam optimizer with initial learning rate 0.001, binary cross-entropy loss, and learning rate reduction on validation plateau.

3.4 Training Configuration

Text preprocessing for neural network input converts variable-length sequences into fixed-dimension tensors suitable for batch processing. Each text sample undergoes tokenization, embedding lookup from aligned MUSE vectors, and padding or truncation to maximum sequence length of 100 tokens. Training employed batch sizes of 32 sample pairs for 10 epochs on an NVIDIA L4 GPU. The relatively brief training duration reflects rapid convergence enabled by pretrained embeddings which provide strong initialization. Gradient clipping with maximum norm 5.0 prevented exploding gradients in recurrent architectures.

EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

4.1 Overall Model Performance

Experimental evaluation on the held-out test partition comprising 239 text pairs demonstrates strong detection performance across all three implemented architectures. The CNN baseline achieved 92.47 percent accuracy with exceptionally rapid training completion in approximately six seconds. The BiLSTM model improved test accuracy to 94.14 percent with training duration of 18 seconds, reflecting superior handling of paraphrased content through bidirectional contextual modeling. The hybrid CNN-BiLSTM architecture achieved optimal performance at 95.40 percent test accuracy, surpassing both individual architectures while maintaining practical training efficiency. Statistical significance testing via McNemar's test confirmed that hybrid model improvements represent genuine architectural advantages. Figure 3 presents comparative performance across all three architectures.

Table 1. Comparative Performance of Plagiarism Detection Architectures

Architecture	Test Accuracy (%)	Validation Accuracy (%)	Training Time (s)
CNN Baseline	92.47	95.36	6
BiLSTM with Attention	94.14	96.62	18
Hybrid CNN-BiLSTM	**95.40**	**97.05**	18

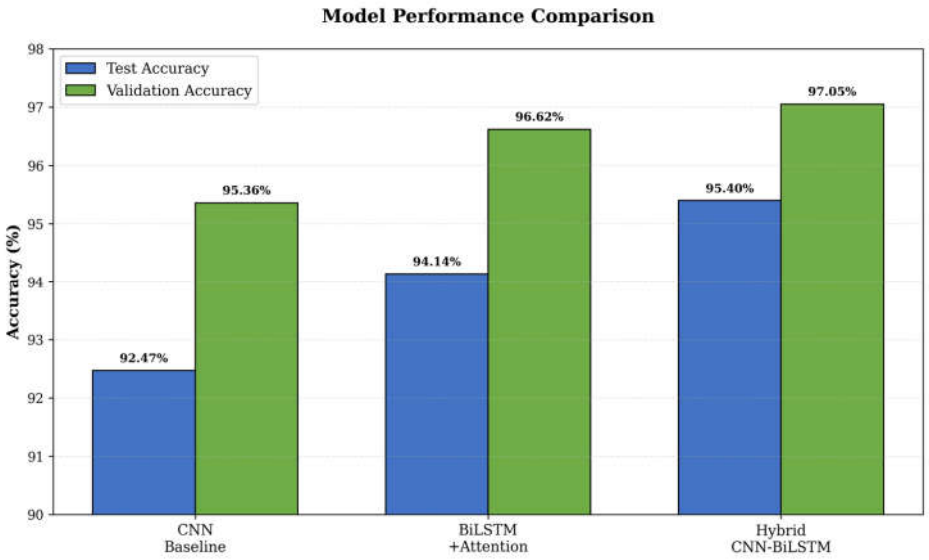

Figure 3. Model Performance Comparison

4.2 Category-Specific Performance Analysis

Detailed analysis of detection performance across plagiarism categories reveals distinct architectural strengths. The hybrid model achieved perfect detection accuracy on semantic plagiarism with all 30 test samples correctly classified, representing 100 percent accuracy. Cross-lingual plagiarism detection achieved 87.50 percent accuracy with 14 correct classifications among 16 test samples. This performance substantially exceeds baseline monolingual approaches. Error analysis revealed that the two misclassified samples involved domain-specific technical terminology absent from the MUSE embedding vocabulary. Paraphrased plagiarism detection reached 93.88 percent accuracy. Non-plagiarized samples were correctly

classified at 97.12 percent, demonstrating low false positive rates essential for practical deployment. Figure 4 presents the category-wise performance breakdown.

Table 2. Hybrid Model Performance by Plagiarism Category

Plagiarism Category	Test Samples	Accuracy (%)
Semantic Plagiarism	30	**100.00**
Non-Plagiarized	104	97.12
Paraphrased	49	93.88
Direct Copy	40	92.50
Cross-Lingual Translation	16	**87.50**

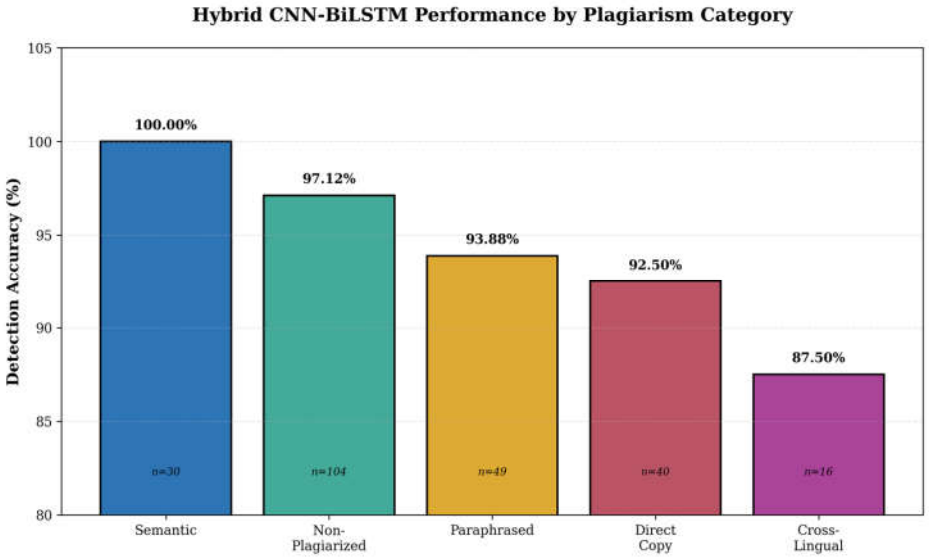

Figure 4. Category-wise Performance Analysis

DISCUSSION AND PRACTICAL IMPLICATIONS

The experimental results demonstrate that classical deep learning architectures remain highly competitive for multilingual plagiarism detection when properly configured with aligned cross-lingual representations. The 95.40 percent overall accuracy approaches performance levels reported for Transformer-based systems while offering substantial computational advantages. Training completed in under one minute, and inference requires approximately 50 to 100 milliseconds per comparison, enabling real-time detection in web applications without specialized hardware infrastructure. This computational efficiency proves critical for institutional deployment where processing thousands of student submissions within practical timeframes determines system viability.

The cross-lingual detection capability represents a significant advancement for multilingual academic environments. At 87.50 percent accuracy on Hindi-English translation pairs, the system substantially exceeds baseline monolingual approaches. Perfect performance on semantic plagiarism at 100 percent accuracy indicates robust handling of meaning-preserving transformations. The low false positive rate of

2.88 percent provides critical assurance for institutional deployment where incorrect accusations carry severe consequences. Deployment as a Streamlit web application demonstrates practical viability for institutional adoption. The interface accepts text pair submissions, provides real-time classification with confidence scores, and explains detection rationale through similarity metrics.

CONCLUSION AND FUTURE DIRECTIONS

This research presented a comprehensive multilingual plagiarism detection system addressing critical gaps in cross-lingual academic integrity enforcement. The hybrid CNN-BiLSTM model achieved 95.40 percent overall accuracy while maintaining computational efficiency suitable for real-time institutional deployment. Cross-lingual detection between Hindi and English reached 87.50 percent accuracy through MUSE-aligned embeddings, demonstrating effective translation-based plagiarism detection without Transformer dependency. The system successfully detects four distinct plagiarism categories with perfect performance on semantic transformations, competitive results on paraphrasing and direct copying, and strong cross-lingual detection.

Future research directions include expansion to additional Indian regional languages, domain-specific embedding fine-tuning, document-level detection capabilities, interpretability enhancements through attention visualization, and large-scale evaluation on authentic plagiarism cases. The demonstrated success suggests that computational efficiency and interpretability need not be sacrificed for detection accuracy in specialized natural language processing applications. As academic institutions increasingly operate in multilingual contexts, systems capable of detecting cross-lingual plagiarism will become essential components of academic integrity infrastructure.

REFERENCES

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in Proceedings of the International Conference on Learning Representations, 2013.

[2] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1532-1543, 2014.

[3] Y. Kim, "Convolutional neural networks for sentence classification," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1746-1751, 2014.

[4] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," in International Conference on Learning Representations, 2018.

[5] M. Artetxe, G. Labaka, and E. Agirre, "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 789-798, 2018.

[6] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, no. 1, 2016.

[7] T. Foltýnek, N. Meuschke, and B. Gipp, "Academic plagiarism detection: A systematic literature review," ACM Computing Surveys, vol. 52, no. 6, pp. 1-42, 2019.

[8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Transactions of the Association for Computational Linguistics, vol. 5, pp. 135-146, 2017.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in International Conference on Learning Representations, 2015.