

A proposed approach for plagiarism detection in Myanmar Unicode text

Sun Thurain Moe¹, Khin Mar Soe¹, Than Than Nwe²

¹Faculty of Computer Science, University of Computer Studies, Yangon, Myanmar

²Faculty of Information Science, University of Information Technology, Yangon, Myanmar

Article Info

Article history:

Received May 2, 2024

Revised Oct 27, 2024

Accepted Nov 14, 2024

Keywords:

Deep learning

Myanmar Unicode

Natural language processing

Plagiarism detection

Syllables segmentation

ABSTRACT

Around the world, with technology that improves over time, almost everyone can access the internet easily and quickly. With the increase in the use of the internet, the plagiarism of information that is easily available on the internet has also increased. Such plagiarism seriously undermines originality and ethical principles. In order to prevent these incidents, there is plagiarism detection software for many countries and languages, but there is no plagiarism detection software for the Myanmar language yet. In an attempt to fill that gap, this study proposed a deep learning model with Rabin-Karp hash code and Word2vec model and built a plagiarism detection system. Our deep learning model was trained by randomly obtaining information from Myanmar Wikipedia. According to the experiments, our proposed model can effectively detect plagiarism of educational content and information from Myanmar Wikipedia. Moreover, it is possible to distinguish plagiarized texts by rearranging words or substituting words with some synonyms. This study contributes to a broader understanding of the complexities of plagiarism in the Myanmar academic area and highlights the importance of measures to effectively prevent plagiarism. It maintains the credibility of education and promotes a culture that values originality and intellectual integrity.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Sun Thurain Moe

Faculty of Computer Science, University of Computer Studies

D2, Room (608), Mindama Pyin Nyar Yeik Thar, Yangon, Myanmar

Email: sunthurainmoe@ucsy.edu.mm

1. INTRODUCTION

In the fields of literature and journalism, including various academic areas, the submission or copying of intellectual property, which is someone else's efforts, without providing a reference or credit to the original owner is gradually increasing, and it is becoming a challenge for various fields. The rapid and easy access to vast amounts of information on the internet makes plagiarism attractive, and plagiarism detection methods struggle to keep up with the growth of technologies such as artificial intelligence used in plagiarism. The most advanced plagiarism detection systems available today use complex machine learning and natural language processing techniques to find syntactic and semantic patterns in text. However, there is a gap that needs to be filled, and that is the lack of proper application of these developments to Myanmar Unicode text.

Plagiarism detection is an ever-evolving field within natural language processing, driven by the increasing complexity of text and the sophisticated methods employed by those attempting to plagiarize. Researchers have continuously explored various algorithms and techniques to improve the accuracy and effectiveness of plagiarism detection systems. There are many different approaches in this sector, from rule-

based algorithms to advanced deep learning models, each contributing uniquely to improving detection accuracy and efficiency.

Moe and Nwe [1] developed a highly accurate rule-based Myanmar syllable segmentation (MSS) algorithm that achieves perfect segmentation accuracy on a large dataset of Myanmar Unicode text. This algorithm's success underscores the potential of rule-based systems for handling specific linguistic challenges. In the area of deep learning, El Mostafa and Benabbou [2] provided an extensive overview of various propositions for plagiarism detection, highlighting the limitations of word granularity and Word2vec methods in capturing the semantic nuances of sentences. Their study suggests the need for more sophisticated models to accurately detect semantic plagiarism. Ali and Taqa [3] reviewed both traditional and modern plagiarism detection techniques, concluding that intelligent and deep learning algorithms, which consider lexical, syntactic, and semantic properties, outperform traditional methods, especially for large corpora. This insight is pivotal for developing more effective plagiarism detection systems. Xiong *et al.* [4] introduced a novel approach that integrates bidirectional encoder representations from transformers (BERT), an enhanced artificial bee colony (ABC) optimization algorithm, and reinforcement learning (RL). This model addresses imbalanced classification and has shown superior performance compared to existing models. Focused on detecting plagiarism in social media content through a four-phase methodology involving data preprocessing, n-gram evaluation, similarity analysis, and detection [5]. Their ensemble support vector machine based African vulture optimization (ESVM-AVO) approach has demonstrated high accuracy and efficiency. Jambi *et al.* [6] evaluated academic plagiarism detection methods using fuzzy multi-criteria decision-making (MCDM), providing valuable recommendations for future systems. Eppa and Murali [7] proposed a multi-source plagiarism detection method for C programming assignments, utilizing an attention-based model and density-based spatial clustering of application (DBSCAN) clustering algorithm. Saeed and Taqa [8] developed an application combining term frequency-inverse document frequency (TF-IDF) text encoding, natural language processing, k-means clustering, and cosine similarity algorithms, while [9] enhanced plagiarism detection using natural language processing and machine learning techniques, achieving impressive results on benchmark datasets.

In cross-language plagiarism detection (CL-PD), Bouaine *et al.* [10] utilized Doc2vec embedding techniques and a Siamese long short-term memory model, achieving outstanding accuracy and performance metrics. Further advanced this field with transformer models and cross-lingual sentence alignment techniques [11], [12]. AlZahrani and Al-Yahya [13] explored Arabic pretrained transformer-based models for authorship attribution in Islamic law, fine-tuning models like ARBERT and AraELECTRA to achieve significant results. Arabi and Akbari [14] proposed methods for detecting extrinsic plagiarism using pretrained networks and WordNet ontologies, demonstrating high precision. Zouaoui and Rezeg [15] presented a multi-agent indexing system for Arabic plagiarism detection, while El-Rashidy *et al.* [16] developed a system using hyperplane equations for high accuracy, outperforming previous systems on standard datasets. Elali and Rachid [17] examined artificial intelligence-based chatbots for detecting fabricated research, and Anil *et al.* [18] compared the effectiveness of various plagiarism detection software on artificial intelligence generated articles. Elkhatat *et al.* [19] evaluated artificial intelligence content detection tools' ability to distinguish between human and artificial intelligence authored content, highlighting ongoing challenges in this area. Foltýnek *et al.* [20] tested web-based text-matching systems, revealing that some systems can detect certain plagiarized content but often misidentify non-plagiarized material. Muangprathub *et al.* [21] proposed a formal concept analysis-based algorithm for document plagiarism detection, particularly effective with Thai text collections. Tian *et al.* [22] introduced FPBirth for multi-threaded program plagiarism detection, demonstrating significant performance improvements. Tlitova *et al.* [23] reviewed methods for identifying cross-language borrowings in scientific articles, focusing on Russian-English pairs and the need for specialized tools in this area. Ansoerge *et al.* [24] presented a case study highlighting common errors in paraphrased plagiarized texts, while Pal *et al.* [25] demonstrated improved accuracy in plagiarism detection using natural language processing techniques, further advancing the field's capabilities.

These diverse studies collectively enhance our understanding and capability in plagiarism detection, addressing various languages, contexts, and methodologies to ensure the integrity of academic and creative works. The primary challenge addressed in this study is the lack of plagiarism detection tools for Myanmar Unicode text. Without reliable plagiarism detection mechanisms, academic institutions and content creators in Myanmar face difficulties in maintaining the integrity and originality of their work. Therefore, an innovative method that makes use of the most recent developments in natural language processing and machine learning is urgently needed in order to accurately detect plagiarism in Myanmar Unicode text.

Since the Myanmar language does not have a specific word cutoff, such as a space character, we have worked step by step through complex preprocessing such as syllable segmentation, word tokenization, stop word removal, and word embedding. Finally, we successfully built a very accurate and effective deep learning model that can automatically identify text plagiarism cases across various topics on Myanmar Wikipedia. To ensure the reliability and robustness of the model, we use a comprehensive evaluation process

comparing its performance with established plagiarism detection techniques and manually annotated datasets. In the sections on the following, we will explore the construction and preprocessing of the dataset, details of the deep learning model architecture, and evaluation measures of this approach.

2. METHOD

In order to detect Myanmar Unicode plagiarism, input text, or documents are first processed to separate syllables by the MSS algorithm. Then, using the pre-collected Myanmar word list and the longest matching algorithm, it is converted into words. After that, stop words are removed from these words, and sentences are segmented. Plagiarism detection will take a long time if the input sentences are compared with all the texts on Myanmar Wikipedia. Therefore, keywords are extracted from the input sentences with the yet another keyword extractor (YAKE) keyword extraction algorithm, and Wikipedia pages related to these sentences are selected using the Wikipedia search application programming interface (API). Only the text content is pulled from the selected Wikipedia pages, followed by syllable segmentation and word tokenization. Then stop words are removed and segmented into sentences. Our proposed system design is shown in Figure 1.

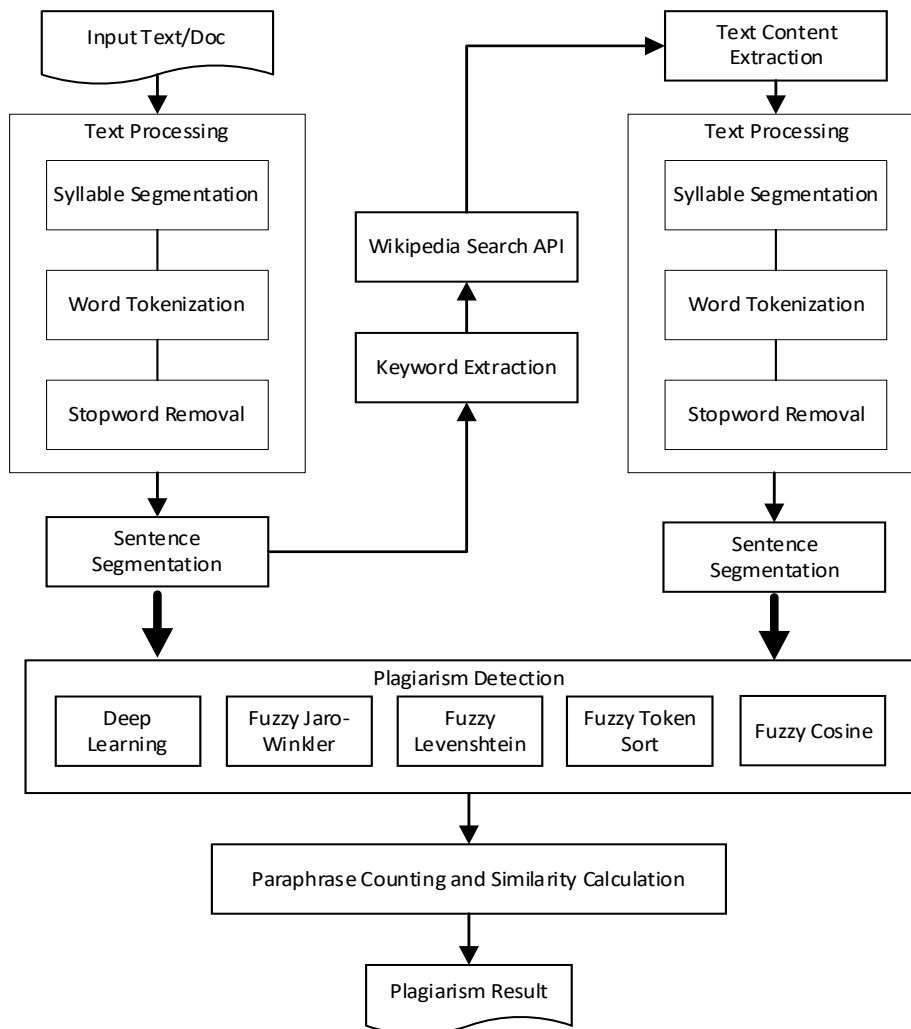


Figure 1. Myanmar Unicode plagiarism detection system

2.1. Myanmar syllable segmentation

The MSS algorithm extracts the Myanmar syllables from the input sentence based on the following four rules [1]: i) If the i^{th} syllabic element of input sentence is not a member of vowel_medial_group; ii) If

the two-dimensional space of t-distributed stochastic neighbor embedding (t-SNE) with principal components analysis (PCA), and each point represents a word.

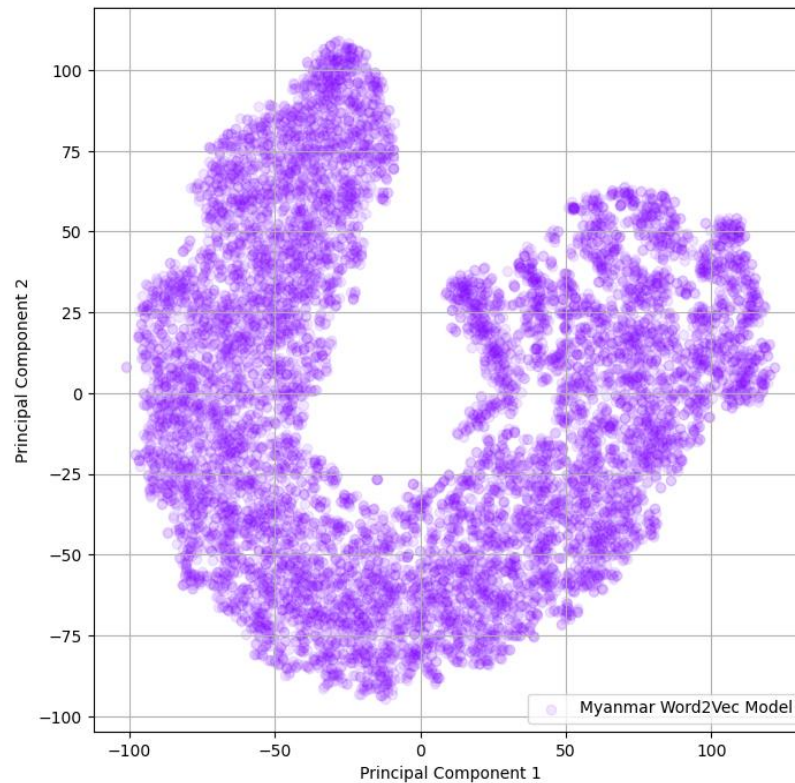


Figure 2. Word2vec model

2.6. Rabin-Karp hash function

The Rabin-Karp algorithm was originally a checker that determines whether two strings (or patterns) match. In this research, however, we employed the Rabin-Karp hashing approach to generate the hash codes for Myanmar words and convert them to vectors. Unlike the Rabin-Karp technique, we did not compare hashes according to linear time, but instead built a deep learning model and compared it. In this work, we used the polynomial rolling hash and modular arithmetic methods defined in (1) to produce hash codes for Myanmar words, ensuring that each Myanmar word received a unique hash code.

$$H = (c_1 * b^{m-1} + c_2 * b^{m-2} + \dots + c_m * b^0) \text{ mod } Q \quad (1)$$

Where H is the hash code, c is the integer American standard code for information interchange (ASCII) code of the character in the word, b is the number of all Myanmar characters, m is the number of characters in the word, and Q is a large prime number.

2.7. Deep learning model

After the preprocessing steps, we performed plagiarism detection on them using a deep learning model. The deep learning model was trained using two different sentence vectorization techniques, such as the Rabin-Karp rolling hash function and Word2vec. The training data includes 1,506 sentences randomly taken from Myanmar Wikipedia pages. The training vectors are obtained after all the words from the training sentences have been converted into hash code numbers, or Word2vec weights. The proposed model is illustrated in Figure 3.

However, the length of the training vectors varies depending on how long the sentence is. We then searched for the sentence with the most words and counted them, discovering that it contained almost 50 words. Each training vector was given a length of 50, and the blank spaces were filled with 1 s. Concatenating the resulting 50-length vectors into 100-length vectors also includes adding the class label, as

shown in Tables 1 and 2. When adding class labels, the class label is set to 1 if a vector repeats itself twice and to 0 if it is adjacent to another randomly chosen vector. In this manner, we obtain a 3011×101-dimensional training dataset. It is important to clarify why each training vector's empty spaces must be filled with 1s in this context.

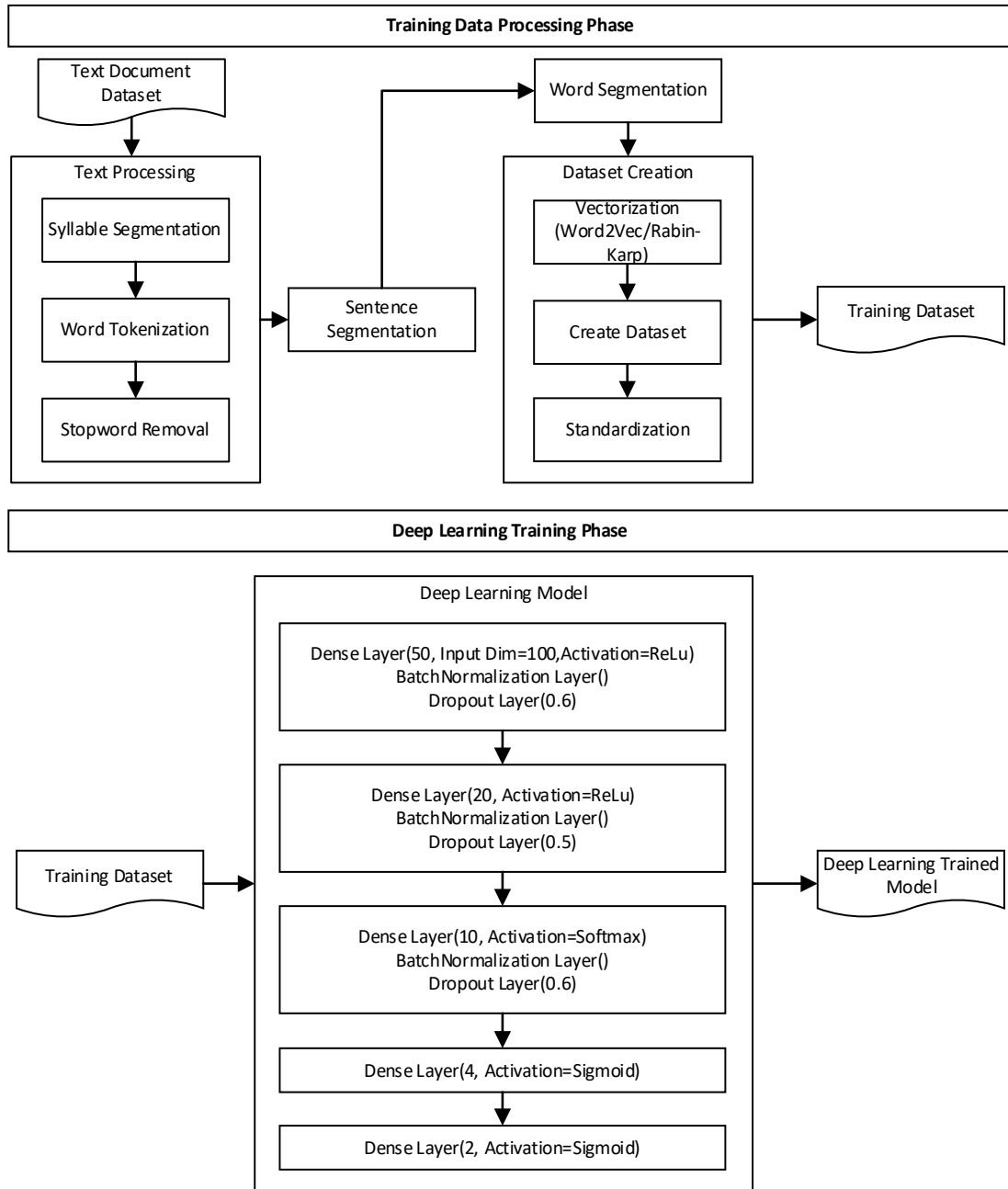


Figure 3. Proposed deep learning model

Table 1. Training data (Rabin-Karp rolling hash)

Src W01	Src W02	...	Src W49	Src W50	Tgt W01	Tgt W02	...	Tgt W49	Tgt W50	Class
1207	2223	...	1	1	1207	2223	...	1	1	1
830	1146	...	1	1	830	1146	...	1	1	1
⋮	⋮	...	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮
830	1146	...	1	1	236	467	...	1	1	0
455	920	...	1	1	515	1207	...	1	1	0

Table 2. Training data (Word2vec)

Src W01	Src W02	...	Src W49	Src W50	Tgt W01	Tgt W02	...	Tgt W49	Tgt W50	Class
-0.01706486	-0.01323607	...	1	1	-0.01706486	-0.01323607	...	1	1	1
-0.01706486	-0.01334877	...	1	1	-0.01706486	-0.01334877	...	1	1	1
⋮	⋮	...	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮
-0.01706486	-0.01334877	...	1	1	-0.03069331	-0.01706486	...	1	1	0
-0.01620913	-0.00342125	...	1	1	-0.01583702	-0.01572607	...	1	1	0

Checking for identity between the source string and the target string is the primary goal of plagiarism detection. We aim to detect any instances of plagiarism in Myanmar Wikipedia articles. The entire collection of articles on Myanmar Wikipedia will need to be used as training data if we choose to use a deep learning model, which is what we typically do for this kind of requirement. There are many challenges involved in doing this, including training data extraction, model training time, and hardware resources. Due to the existence of these challenges, our deep learning model was separated from traditional operation models and purposefully built as a probabilistic model based on weights similar to a logistics regression.

To maintain the weights of our trained model, we filled all empty spaces in the training vectors with 1 s. Our deep learning model does not use convolutional layers because our training dataset has a 1D structure. It has only 5 dense, fully connected layers, and rectified linear unit (ReLU), softmax, and sigmoid are used as activation functions. Using the holdout method, 20% of the 3,011 training datasets were divided, and 602 were used as testing data for the model evaluation. Our proposed model has a 98% accuracy rate for detecting plagiarism, as shown in Table 3.

Table 3. Results of proposed model

	Precision	Recall	F1-score	Support
Unmatched	1	0.96	0.98	302
Matched	0.96	1	0.98	300
Accuracy			0.98	602
Macro avg	0.98	0.98	0.98	602
Weighted avg	0.98	0.98	0.98	602

3. RESULTS AND DISCUSSION

As mentioned in subsection 2.7, we tested our proposed deep learning model using two different vectorization techniques. The results indicated that the deep learning model trained with sentence vectorization using the Rabin-Karp rolling hash function provided more accurate plagiarism detection results than the model trained with Word2vec sentence vectorization. Our experiment is the first in the field of Myanmar Unicode plagiarism detection, with no existing methods available for comparison. Plagiarism detection techniques used in other languages, including English, are not applicable to Myanmar Unicode.

To evaluate the performance of our proposed model, we compared the results with those obtained using well-known fuzzy string matching methods. The experiment involved 500 randomly selected sentences from Myanmar Wikipedia, containing nearly 3,000 paraphrases. These sentences were tested for direct copying and paraphrasing plagiarism, where word syntax was reversed. The results of this test are presented in Table 4.

The experiment revealed that our proposed method, which combines a deep learning model with Rabin-Karp sentence vectorization, produced the best results. However, our method still has some weaknesses. As the first research for Myanmar Unicode plagiarism detection, focusing on direct copying and pasting, the results shown in Table 4 are promising. Nevertheless, the method has limitations in detecting other types of plagiarism, such as outlining and summarizing, where only the concept or idea is taken. Further research is needed to develop adaptive methods for detecting these types of plagiarism.

Table 4. Experimental result

Method	Similarity score (%)	
	Complete plagiarism	Paraphrasing plagiarism
DL(Rabin-Karp)	95.6	95.6
DL(Word2Vec)	94.1	94.1
Fuzzy Jaro-Winkler	91.5	88.6
Fuzzy Levenshtein	92.4	59.4
Fuzzy Token Sort	91.9	91.5
Fuzzy Cosine	90.8	91.2

4. CONCLUSION

Our study marks a significant step forward in the field of Myanmar Unicode plagiarism detection. By testing a deep learning model trained with two different vectorization techniques, we found that sentence vectorization with the Rabin-Karp rolling hash function provided superior accuracy compared to the Word2vec-based approach. This experiment, the first of its kind for Myanmar Unicode, highlights the limitations of applying plagiarism detection methods developed for other languages. While our method showed promising results in detecting direct copying and paraphrasing, it still faces challenges in identifying more complex forms of plagiarism, such as outlining and summarizing. Future research should focus on developing adaptive methods to address these challenges, enhancing the robustness and accuracy of plagiarism detection in Myanmar Unicode. In the future, such progress can be extended and significantly contribute to maintaining academic and professional integrity while encouraging originality and creativity in written works.

ACKNOWLEDGEMENTS

We would like to express our heartfelt gratitude to all those who contributed to the successful completion of our publication paper. Additionally, we would like to express our gratitude for the support and resources provided by the University of Computer Studies, Yangon, Myanmar. Their contributions were essential to the completion of this work.




REFERENCES

- [1] S. T. Moe and T. T. Nwe, "An algorithm for Myanmar syllable segmentation based on the official standard Myanmar Unicode text," in *2023 IEEE Conference on Computer Applications (ICCA)*, Feb. 2023, pp. 6–10, doi: 10.1109/ICCA51723.2023.10181391.
- [2] H. El Mostafa and F. Benabbou, "A deep learning based technique for plagiarism detection: a comparative study," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 1, p. 81, Mar. 2020, doi: 10.11591/ijai.v9.i1.pp81-90.
- [3] A. Ali and A. Taqa, "Analytical study of traditional and intelligent textual plagiarism detection approaches," *Journal of Education and Science*, vol. 31, no. 1, pp. 8–25, Mar. 2022, doi: 10.33899/edusj.2021.131895.1192.
- [4] J. Xiong *et al.*, "Efficient reinforcement learning-based method for plagiarism detection boosted by a population-based algorithm for pretraining weights," *Expert Systems with Applications*, vol. 238, p. 122088, Mar. 2024, doi: 10.1016/j.eswa.2023.122088.
- [5] S. V. Vadivu, P. Nagaraj, and B. A. S. Murugan, "Ensemble machine learning technique-based plagiarism detection over opinions in social media," *Automatika*, vol. 65, no. 3, pp. 983–991, Jul. 2024, doi: 10.1080/00051144.2024.2326383.
- [6] K. M. Jambi, I. H. Khan, and M. A. Siddiqui, "Evaluation of different plagiarism detection methods: A fuzzy MCDM perspective," *Applied Sciences*, vol. 12, no. 9, p. 4580, Apr. 2022, doi: 10.3390/app12094580.
- [7] A. Eppa and A. H. Murali, "Machine learning techniques for multisource plagiarism detection," in *2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, Dec. 2021, pp. 1–5, doi: 10.1109/CSITSS54238.2021.9683752.
- [8] A. A. M. Saeed and A. Y. Taqa, "A proposed approach for plagiarism detection in article documents," *Sinkron*, vol. 7, no. 2, pp. 568–578, Apr. 2022, doi: 10.33395/sinkron.v7i2.11381.
- [9] F. K. AL-Jibory and others, "Hybrid system for plagiarism detection on a scientific paper," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 13, pp. 5707–5719, 2021.
- [10] C. Bouaine, F. Benabbou, and I. Sadgali, "Word embedding for high performance cross-language plagiarism detection techniques," *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 17, no. 10, pp. 69–91, May 2023, doi: 10.3991/ijim.v17i10.38891.
- [11] R. S. R. Raj and G. R. Ramya, "Detection of plagiarism in contextual meaning using transformer model and community detection algorithm," *Smart Trends in Computing and Communications*, 2023, pp. 777–795, doi: 10.1007/978-981-99-0838-7_67.
- [12] T. Ter-Hovhannisyan and K. Avetisyan, "Transformer-based multilingual language models in cross-lingual plagiarism detection," in *2022 Ivannikov Memorial Workshop (IVMEM)*, Sep. 2022, pp. 72–80, doi: 10.1109/IVMEM57067.2022.9983968.
- [13] F. M. AlZahrani and M. Al-Yahya, "A transformer-based approach to authorship attribution in classical Arabic texts," *Applied Sciences*, vol. 13, no. 12, Jun. 2023, doi: 10.3390/app13127255.
- [14] H. Arabi and M. Akbari, "Improving plagiarism detection in text document using hybrid weighted similarity," *Expert Systems with Applications*, vol. 207, Nov. 2022, doi: 10.1016/j.eswa.2022.118034.
- [15] S. Zouaoui and K. Rezeg, "Multi-agents indexing system (MAIS) for plagiarism detection," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 2131–2140, May 2022, doi: 10.1016/j.jksuci.2020.06.009.
- [16] M. A. El-Rashidy, R. G. Mohamed, N. A. El-Fishawy, and M. A. Shouman, "An effective text plagiarism detection system based on feature selection and SVM techniques," *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 2609–2646, Jan. 2024, doi: 10.1007/s11042-023-15703-4.
- [17] F. R. Elali and L. N. Rachid, "AI-generated research paper fabrication and plagiarism in the scientific community," *Patterns*, vol. 4, no. 3, Mar. 2023, doi: 10.1016/j.patter.2023.100706.
- [18] A. Anil *et al.*, "Are paid tools worth the cost? a prospective cross-over study to find the right tool for plagiarism detection," *Heliyon*, vol. 9, no. 9, Sep. 2023, doi: 10.1016/j.heliyon.2023.e19194.
- [19] A. M. Elkhatat, K. Elsaid, and S. Almeer, "Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text," *International Journal for Educational Integrity*, vol. 19, no. 1, Sep. 2023, doi: 10.1007/s40979-023-00140-5.
- [20] T. Foltýnek *et al.*, "Testing of support tools for plagiarism detection," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, Dec. 2020, doi: 10.1186/s41239-020-00192-4.
- [21] J. Muangprathub, S. Kajornkasirat, and A. Wanichsombat, "Document plagiarism detection using a new concept similarity in formal concept analysis," *Journal of Applied Mathematics*, vol. 2021, pp. 1–10, Mar. 2021, doi: 10.1155/2021/6662984.




- [22] Z. Tian, Q. Wang, C. Gao, L. Chen, and D. Wu, "Plagiarism detection of multi-threaded programs using frequent behavioral pattern mining," *International Journal of Software Engineering and Knowledge Engineering*, vol. 30, no. 11-12, pp. 1667–1688, Nov. 2020, doi: 10.1142/S0218194020400252.
- [23] A. Tlitova, A. Toshev, M. Talanov, and V. Kurmosov, "Meta-analysis of cross-language plagiarism and self-plagiarism detection methods for Russian-English language pair," *Frontiers in Computer Science*, vol. 2, Oct. 2020, doi: 10.3389/fcomp.2020.523053.
- [24] L. Ansoorge, K. Ansoorgeová, and M. Sixsmith, "Plagiarism through paraphrasing tools—the story of one plagiarized text," *Publications*, vol. 9, no. 4, Oct. 2021, doi: 10.3390/publications9040048.
- [25] S. K. Pal, O. J. Raffik, R. Roy, V. B. Lalman, S. Srivastava, and B. Sharma, "Automatic plagiarism detection using natural language processing," in *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2023, pp. 218–222.

BIOGRAPHIES OF AUTHORS






Sun Thurain Moe    is currently pursuing a Ph.D. degree at the University of Computer Studies, Yangon, Myanmar. He received a Master's degree (M.I.Sc.) in information science from the University of Computer Studies, Yangon. He has contributed significantly to the field of Myanmar syllable segmentation, Myanmar plagiarism and credit scoring, publishing three papers in IEEE. His research interests include Myanmar natural language processing, machine learning, and action recognition. He can be contacted at email: sunthurainmoe@ucsy.edu.mm.



Dr. Khin Mar Soe    received a Ph.D. (information technology) degree from the University of Computer Studies, Yangon, Myanmar. She is a professor at the University of Computer Studies, Yangon. Her research interests are in the areas of natural language processing, part-of-speech tagging, machine translation, and Myanmar name entity recognition. She can be contacted at email: khinmarsoe@ucsy.edu.mm.



Dr. Than Than Nwe    received a Ph.D. (information technology) degree from the University of Computer Studies, Yangon, Myanmar. She is a professor at the University of Information Technology, Yangon. Her research interests are in the areas of information retrieval, data mining, big data, and pattern recognition. She can be contacted at email: thanthannwe.cu@uit.edu.mm.