Plagiarism Detection with Word Embedding Model in Myanmar Unicode Text Documents

Sun Thurain Moe
University of Computer Studies
Yangon, Myanmar
sunthurainmoe@ucsy.edu.mm

Khin Mar Soe
University of Computer Studies
Yangon, Myanmar
khinmarsoe@ucsy.edu.mm

Than Than Nwe
University of Information Technology
Yangon, Myanmar
thanthannwe.cu@gmail.com

***Abstract*— Plagiarism of intellectual property has increased in numerous industries as the internet grows quicker and simpler to use. Plagiarism has grown to be a significant issue, particularly in academic sector where copying and pasting is common. Manually checking for such plagiarism is a highly challenging task, thus finding new means to do so has become a crucial procedure. There are numerous researches out there today that suggest strategies to detect plagiarism. However, most of them are solely for English, and no study has been done on plagiarism in Myanmar. The goal of this study is to fill that gap. The word embedding model is employed in this study to identify plagiarism in texts written in Myanmar Unicode. The model was trained by Myanmar Unicode text from 1,000 Myanmar Wikipedia pages. We used word mover's distance and fuzzy string-matching methods to measure plagiarism. According to experimental results and visual examination, the word embedding model can yield reliable results in identifying plagiarism in Myanmar.**

Keywords*— *Plagiarism Detection, Natural Language Processing, NLP, Fuzzy String Matching, Myanmar Syllables Segmentation

I. INTRODUCTION

Information and knowledge sharing in Myanmar have undergone significant changes as a result of widespread access to digital communication and the internet. A significant step for Myanmar, which has a rich linguistic and cultural legacy, is the adoption of Unicode. However, this development has also highlighted the problem of plagiarism in Myanmar Unicode text. The appropriation of digital content, frequently written in Myanmar, without the proper acknowledgement or agreement of the original creators is considered plagiarism in Myanmar Unicode text. In the academic and artistic fields, plagiarism has long been a problem in its different manifestations.

The development of online platforms has given this problem in Myanmar a new angle. This is disrespectful to the authors who have spent time and effort creating original works as well as to the literary history of Myanmar. Additionally, scholars and content creators face a serious threat from it. The goal of this study is to characterize the complex issue of Myanmar Unicode piracy, effects, and potential solutions.

While advances in plagiarism detection technologies have been developed in many languages, this is not the case for languages like Myanmar Unicode, which have complex scripts and few linguistic resources. Myanmar presents particular difficulties in creating effective plagiarism detection systems due to its unusual script and linguistic nuances.

This study will also look into the technological advancements and solutions available to stop the piracy of Myanmar Unicode, give people and organizations the power to protect their digital content, and encourage a culture of originality and intellectual property. It is critical to address the problems caused by piracy of Myanmar Unicode as we make our way through the complicated world of digital communications. By doing this, we can protect legacy of Myanmar culture, encouraging innovation and creativity, and make sure that the internet continues to be a place of respect and dignity for all of its users.

By applying the Word2Vec embedding technique and tailoring it to the complexities of Myanmar Unicode text, this research aims to bridge the gap in advanced plagiarism detection tools for the Myanmar Unicode script. The outcomes of this study have the potential benefit to Myanmar and other languages facing similar challenges in preserving the integrity of textual content.

II. RELATED WORK

The author proposed a new Myanmar syllable segmentation (MSS) algorithm based on official Myanmar Unicode text [1]. The algorithm introduces the four rules to segment the Myanmar syllables and proves 100% segmentation accuracy in 33,500 syllables from randomly selected 500 Myanmar Unicode sentences. [2] studied and compared many different methods used in plagiarism detection and pointed out the weakness of cosine similarity in keeping the semantic aspect. The author proposed the RNN, which will be more reliable in terms of plagiarism in the semantic aspect. [3] proposed multi-level plagiarism detection by using LSTM and CNN algorithms. The proposed method focuses on semantic plagiarism detection using the approaches of doc2vec and LSTM for duplicate query detection. CNN will take as an input the LSTM representation provided by the first stage of learning to be able to initiate data classification or accurately label this representation as a type of plagiarism. [4] presented a method, DeezyMatch, a free open-source library, to rank and match strings that can be used in plagiarism detection. It used vector representation and transfer learning in fuzzy string matching to find the similarity between large knowledge base and query sets. [5] studied traditional and intelligent methods for detecting textual plagiarism. The author

discussed various approaches to representing textual features, including contextual techniques, discrete representation, and distributed representation. According to the review, traditional approaches only consider the text's lexical features, whereas intelligent and deep learning approaches can identify the text's lexical, syntactic, and semantic characteristics. That makes it better than the traditional methods, particularly when the corpus is very large.

III. Methodology

Myanmar script, with its unique feature, poses distinctive challenges in developing robust plagiarism detection systems. Especially, Myanmar script has no boundary between Myanmar words and the author usually makes a cut with a space as he sees fit, to make it convenient to read.

Moreover, the appropriate cut-offs between each Myanmar word are different from one author to another, so there are many difficulties in preprocessing Myanmar script with currently used NLP tools. Therefore, Myanmar syllable segmentation algorithm (MSS) and longest matching word tokenization algorithm were used to separate Myanmar words. Our proposed system design is as shown in Fig. 1.

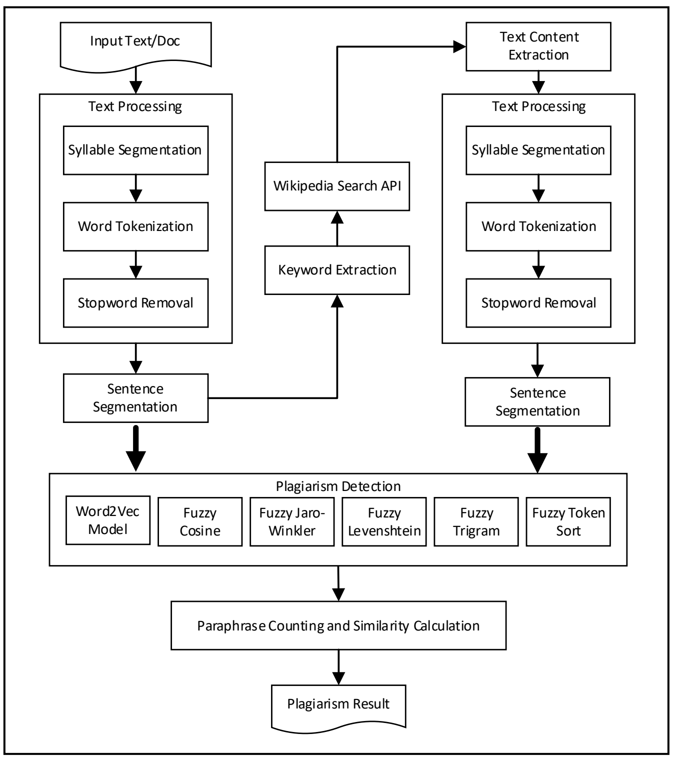

Fig. 1. Proposed system design

A. *Myanmar Syllable Segmentation(MSS)*

MSS is a kind of rule-based syllable segmentation algorithm. It is effective, easy to understand and straightforward. It divides the input sentence into syllables and compares it to the following rules.

- If the ith syllabic element of input sentence is not a member of vowel_medial_group.
- If the (i-1)th or ith or (i+1)th syllabic element of input sentence is not a consonant pairs symbol ‘◌္’.
- If the (i+1)th syllabic element of input sentence is not a consonant pairs symbol ‘◌့’.
- If the (i+1)th syllabic element of input sentence is not a ‘Asat’ symbol ‘◌်’.

In the set of rules that defines the border of a Myanmar Unicode syllable that we found, the ‘ေ◌’ character pronounced as "tha wai htoe", consonant, vowels and medial are grouped together [1].

vowel_medial_group = [‘ေ◌’, ‘◌်’, ‘◌့’, ‘◌ွ’, ‘◌ု’, ‘◌ူ’, ‘◌ှ’, ‘◌ျ’, ‘◌း’, ‘◌ါ’ , ‘◌ံ’, ‘◌ဲ’, ‘◌ိ’, ‘◌ီ’, ‘◌ြ’, ‘◌ာ’]

B. *Myanmar Word Tokenization*

Proposed system tokenized Myanmar terms against almost 50,000 pre-collected Myanmar words using a greedy longest-match-first method. The longest n-prefix of the remaining syllables that matches a word in the pre-compiled Myanmar word list is chosen by the algorithm. Unknown syllables are treated as stop words and excluded from the pre-compiled list of Myanmar words.

C. *Word Embedding*

Word embedding is a popular framework for representing each word as a vector space with multiple dimensions based on the semantic context in which it is found. Many machine learning and natural language processing projects employ this technique[6]. The phrase "word embed" refers to both vector within a word and numerous word vectors in a data collection. One of the models in Google's embedding projector, a trained model that is made available to the public, is Word2Vec[7]. Words with similar meanings are closer to each other in this vector space, making it a powerful tool for understanding textual content.

We have collected 45,399 sentences of Myanmar Unicode text from 1,000 Myanmar Wikipedia pages and which were first preprocessed and tokenized into words or phrases in order to perform plagiarism detection. Then the vector is used to represent each word. In this phase, the entire text content of Myanmar Wikipedia page is transformed into a matrix of vectors, with each row representing a word. Fig. 2 shows the visualization of our Word2Vec model in two-dimensional space of t-SNE with PCA and each point represents a word.

The similarity between the text can be assessed using a variety of methods, including cosine similarity or Euclidean distance, once the documents are represented as Word2Vec vectors. In this study we used word mover's distance method, due to its simplicity and hyper-parameter-free nature, was utilized to assess how similar the sentences were.

Plagiarism detection systems typically require a threshold value to determine when a given document with a reference page are considered similar enough to be flagged for potential plagiarism. This threshold is often set based on experimentation

and domain-specific requirements. In this study, the similarity threshold was set at 0.75.

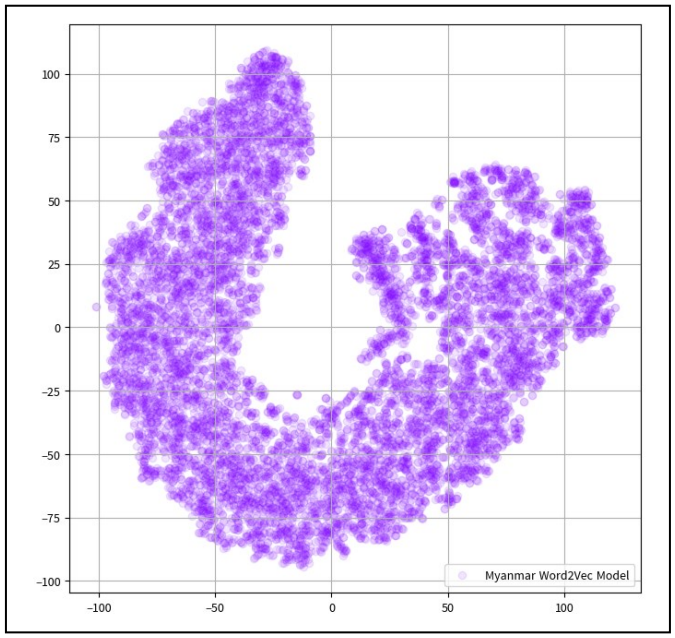

Fig. 2. Word2Vec model

IV. Experimental Setup

We gathered a random selection of sentences from Myanmar Wikipedia pages, separated it into three groups and evaluated our proposed model. Direct copying and pasting was examined for the first group, and plagiarism in paraphrasing was evaluated for the second group. The third group was evaluated for incorporates phrases from different sources. Each group in our observation included almost 3,000 paraphrases.

Syllable segmentation and word tokenization are applied to the input text in the experiment. The key words are then chosen from the remaining text after eliminating the stop words. Using the Wikipedia search API, we searched for the pages that might be potentially plagiarized. Then, checked the plagiarized text in the input text after extracting the text content from all potential pages. In this evaluation, we compared the results of our proposed model with 5 fuzzy similarity methods.

TABLE I. Evaluation Result

Method	Similarity Score (%)		
	Complete plagiarism	***Paraphrasing plagiarism***	***Patchwork plagiarism***
Proposed Method	93.00	92.88	92.88
Fuzzy Jaro-Winkler	87.00	84.13	84.13
Fuzzy Levenshtein	89.75	54.63	54.63
Fuzzy Trigram	92.13	92.00	92.00
Fuzzy Token Sort	91.75	91.50	91.50
Fuzzy Cosine	91.00	90.88	90.88

The analysis of the test results and visual inspections revealed that our proposed model can accurately identify the plagiarism of Myanmar Unicode text.

The evaluation results from the experiment are shown in Table I. Both the effects of paraphrasing and the results of patchwork plagiarism are the same. This is because the proposed methodology only checks for plagiarism at the sentence level. This means that when plagiarism is checked at the sentence level, paraphrasing plagiarism produces the same results as patchwork plagiarism composed of sentences from different sources.

Acknowledgment

I would like to thank all of the professors and friends who provided valuable advices and feedbacks to help make this research a success. I'd also like to thank my colleagues and students who helped me with my requirements.

References

[1] S. T. Moe and T. T. Nwe, "An Algorithm for Myanmar Syllable Segmentation based on the Official Standard Myanmar Unicode Text," 2023 IEEE Conference on Computer Applications (ICCA), Yangon, Myanmar, 2023, pp. 6-10, doi: 10.1109/ICCA51723.2023.10181391.

[2] el Mostafa, Hambi & Benabbou, Faouzia. (2020). A deep learning based technique for plagiarism detection: a comparative study. IAES International Journal of Artificial Intelligence (IJ-AI). 9. 81. 10.11591/ijai.v9.i1.pp81-90.

[3] el Mostafa, Hambi & Benabbou, Faouzia. (2019). A Multi-Level Plagiarism Detection System Based on Deep Learning Algorithms.

[4] Kasra Hosseini, Federico Nanni, and Mariona Coll Ardanuy. 2020. DeezyMatch: A Flexible Deep Learning Approach to Fuzzy String Matching. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 62–69, Online. Association for Computational Linguistics.

[5] Ali, Ayoub & Taqa, Alaa. (2022). Analytical Study of Traditional and Intelligent Textual Plagiarism Detection Approaches. JOURNAL OF EDUCATION AND SCIENCE. 31. 8-25. 10.33899/edusj.2021.131895.1192.

[6] Mewa, T. (2020). 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings' by Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, & Adam Kalai (2016). In Identifying Gender and Sexuality of Data Subjects. Retrieved from https://cis.pubpub.org/pub/debiasing-word-embeddings-2016.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space." Online, Available: http://arxiv.org/abs/1301.3781