Paper ID No:224

Anulipi Porichhedok: A Bangla Plagiarism Detection Tool

Shejuti Binte Feroz, Zakia Tamanna, Muhammad Samee Sevas, Chowdhury Farjna-Tur-Santona

Department of Computer Science and Engineering Military Institute of Science and Technology, Mirpur Cantonment, Dhaka-1216, Bangladesh
shejutibinteferoz@gmail.com, zakiatamannanaora@gmail.com, samee.sevas@gmail.com, cftsantona@gmail.com

Extended Abstract:

In academics, plagiarism is a big problem, and finding it can be difficult. The need to create efficient methods for plagiarism detection has been increased due to the availability of more and more digital content. The goal of this project is to create a program that employs natural language processing to identify plagiarism in essays written in Bangla. The project intends to produce tools that would support academic integrity as well as promote NLP research in the Bangla language. It's crucial to find plagiarism in government job applications and Bangla posts to preserve the integrity and authority of the writing. We are proposing a system that will fill this gap by developing a Bangla plagiarism detection tool specifically designed for these domains.

This project's study field is Bangla natural language processing (NLP), with a particular emphasis on creating an essay plagiarism detection tool. By investigating new techniques for feature extraction and similarity assessment in the Bangla language, this research intends to further NLP in addition to creating a plagiarism detection tool for Bangla writings. Additionally, the research examines the effectiveness of various machine learning algorithms and determines the best strategy for Bangla plagiarism detection.

Our proposed model includes some basic steps: 1) Collecting dataset by scrapping digital platform such as NCTB books and converting it in the docs format; 2) Modifying the dataset; 3) Preprocessing the data ; 4) Compare it with the written data from which plagiarism needs to be detected; 5) Evaluate the accuracy of the model. At first, we will focus on detecting plagiarism in different Bangla essay writing events. Bangla 2nd paper essay writing books PDF need to be converted to docs in this context. It can be done by web scrapping different Bangla essay writing books and collecting the dataset. After that the dataset will be modified if needed. For preprocessing the data gained from our modified dataset, we propose NLP algorithms such as bag of words in order to extract the texts from books, removing the stop words and finally to generate the preprocessed format of data to feed the model. The model will be trained and evaluated after completion of previous steps.

In this project we will use Accuracy, F1_score, Precision, Recall etc in order to compare the outcomes to get a better result. The highest accuracy is expected to detect suspicious texts in our project. Moreover, the comparison clarity will be ensured in our project. Despite all of these we have limitations in our project. A developed dataset will be more appreciated in our project for the contribution. More efficient algorithms can be introduced to develop the accuracy of detecting plagiarism in different application area of our project. As a result we can say, if Bangla plagiarism detection can be applied successfully in many domains then it is possible to eradicate the issue from root and many valuable writings at various digital platform can be saved from it's grasp.

References

[1] Y. M. M. A. S. Mohammad Shamsul Arefin, "BAENPD: A Bilingual Plagiarism Detector," *JOURNAL OF COMPUTERS,* vol. 8, no. 5, pp. 1145-1156, 2013.

[2] S. S. N. H. Adil Ahnaf, "Closed Domain Bangla Extrinsic Monolingual Plagiarism Detection and Corpus Creation Approach," in *IEEE*, Dhaka, 2020.