Plagiarism Detection in the Bengali Language: A Text Similarity-Based Approach

Satyajit Ghosh
Department of Computer Science & Engineering
Adamas University
Kolkata , India
satyajit.ghosh@stu.adamasuniversity.ac .in
Abhishek Roy
Department of Computer Science & Engineering
Adamas University
Kolkata , India
abhishek.roy@adamasuniversity.ac.in
Aniruddha Ghosh
Department of Computer Science & Engineering
SRM University
Chennai , India
ag7434@srmist.edu.in
Bittaswer Ghosh
Department of Computer Science & Engineering
Adamas University
Kolkata , India
bittaswer.ghosh@stu.adamasuniversity.ac.in

Plagiarism means taking another person's work and not giving any credit to them for it. Plagiarism is one of the most serious problems in academia and among researchers.Even though there are multiple tools available to detect plagiarism in a document but most of them are domain-specific and designed to work in English texts, but plagiarism is not limited to a single language only. Bengali is the most widely spoken language of Bangladesh and the second most spoken language in India with 300 million native speakers and 37 million second-language speakers. Plagiarism detection requires a large corpus for comparison. Bengali Literature has a history of 1300 years. Hence most Bengali Literature books are not yet digitalized properly. As there was no such corpus present for our purpose so we have collected Bengali Literature books from the National Digital Library of India and with a comprehensive methodology extracted texts from it and constructed our corpus. Our experimental results find out average accuracy between 72.10 % - 79.89 % in text extraction using OCR. Levenshtein Distance algorithm is used for determining Plagiarism. We have built a web application for end-user and successfully tested it for Plagiarism detection in Bengali texts. In future, we aim to construct a corpus with more books for more accurate detection.

Keywords— Plagiarism Detection, Bengali text, OCR, Levenshtein algorithm, Bengali Corpus

INTERNATIONAL JOURNAL OF SCIENCE AND INNOVATIVE ENGINEERING & TECHNOLOGY