A CONTENT ANALYSIS BASED PLAGIARISM DETECTOR FOR UML DIAGRAMS IN ACADEMIC LITERATURE

Sumudu Tharaka Mohottige

A dissertation submitted in partial fulfilment of the requirement for
Bachelor of Engineering (Honors) degree in Software Engineering

Department of Computing

Informatics Institute of Technology, Sri Lanka
in collaboration with
University of Westminster, UK

2021

Abstract

Committing plagiarism is considered as a serious offence in the academic arena. Detection of plagiarism accurately has been vital due to the tremendous rise of plagiarism. Advanced plagiarism detection tools are used widely for the verification of the authenticity of academic documents. Even though the tools are capable of detecting numerous forms of text-based plagiarism, most of them fail to detect plagiarism on the image-based content. Discarding image-based content is inappropriate since modern academic documents are composed of a comparable amount of image-based content and it has also led to the creation of a loophole that advantage the authors. Image-based content in academic documents include diagrams, charts, and figures which are used to visually represent data related to the content of the document. Due to the flaw of the current plagiarism detection tools, the authors can plagiarize image-based content easily without being detected. Hence there is a need for a reliable approach for the detection of plagiarism in image-based content.

In this research project, the author focuses on developing a reliable and a scalable approach for the detection of plagiarism in Unified Modelling Language (UML) diagrams found in the form of image-based content. The proposed approach combines a series of image processing, computer vision, and feature extraction mechanisms to extract and analyze the vital features for the detection of similarity in UML diagrams. The implemented solution supports both intrinsic and extrinsic detection. The performance and the accuracy of the proposed solution were critically evaluated. Based on the output generated during the testing phase, the proposed solution was able to achieve a precision of value “0.94” and a recall of value “1”. The achieved precision and the recall values proves that the implemented solution validates the proposed hypothesis and is worthy to be deployed into the consumer level market.

Keywords: Image processing, Computer vision, Content analyzation, Plagiarism detection, Diagrammatic plagiarism, Unified Modelling Language