DESIGN OF HYBRID AUTOMATIC EXTRINSIC PLAGIARISM DETECTION FRAMEWORK USING MACHINE-LEARNING TECHNIQUES.

BY Gilbert Engelbert. (2017)

ABSTRACT

The evolution of internet access provided easy access to data and information. Hence it is easier to take credit for one's intellectual work without acknowledging the original author(s) – this is plagiarism. Academic institutions exercise policies and use software tools to detect them. However, plagiarism detection is affected by the use of different levels of obfuscation, large amounts of data, and a number of data sources. This makes some detection methodologies accurate for practical applications.

This research aimed at assessing and testing plagiarism detection methods using machine learning techniques including Genism, Fingerprinting, Support Vector Machines Bayes Naïve, and TensorFlow Neural Networks; and this led to designing a hybrid framework using the tested methods.

The designed hybrid framework was tested using source files containing originality and suspicious files with plagiarized passages. With Gensim which is designed to work efficiently with personal computers; was able to process 8650 files in less than 4 hours. Fingerprint produced 100% detections, and Gensim above 80%. The use of more than one method in the design to create a hybrid framework provided more detection options and accuracy in both literal and intelligent plagiarism.

Hence, the combination of fingerprint, Gensim and TensorFlow resulted in an automatic plagiarism detection hybrid framework design which was the main aim of this research; whose overall performance was more accurate than individual setup. With the fast growth of the amount of data, preprocessing requires more robust tools; new big data tools are recommended to mine the data in future studies.