INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Classifying Personally Identifiable Information and Payment Card Industry Data in Text Contents Using Machine Learning

A dissertation by

Mr. Milinda Arambawela

Supervised by

Mr. Achala Aponso

Submitted in partial fulfilment of the requirements for the MSc in Advanced Software Engineering degree at the University of Westminster.

July 2023

ABSTRACT

With the achievement of the Internet, lives become easier with online services. Daily tasks such as purchasing goods and placing an appointment with a doctor using the internet are quicker and easier than they were. Improvement of many online services attracts many people to do their activities online. This makes larger amount of personal and payment transactions data recorded in the many forms of storage such as, databases, logfiles. This sensitive data should be protected and regulated according to the guidelines provided in GDPR and PCI-DSS compliances.

Identifying the exposed personal data is not an easy task. In this research, a novel approach has been introduced to identify personally identifiable information (PII) and payment card industry data (PCI). A machine learning based text classification model that uses Support Vector Machine model to identify PII and PCI data in the given text has been proposed in this research project.

The CNN model has been built and benchmarked against SVM, Naive Bayes, Random Forest, and gradient boost models. Among all the models, the CNN model achieved the highest accuracy of 0.96 (96%). The F1 scores for each class were also impressive, with PII scoring 0.96, PCI scoring 0.96, and Normal scoring 0.96. After building and training the model, it was utilized with the saved tokenizer's word indexes and label encoder classes in the classification tool, which was developed to identify exposed PII and PCI data. As promised, the classification tool successfully displayed the results of exposed PII and PCI data.

Keywords: Machine Learning, CNN, PII, PCI, Text Classification