The College of Graduate Studies and the College of Information Technology Cordially Invite You to a

Master Thesis Defense

Entitled

UNVEILING THE ORIGINS OF SOURCE CODE THROUGH AUTHORSHIP ATTRIBUTION: A COMPARATIVE STUDY OF AI AND HUMAN CODING PATTERNS

by

Shamma Humaid Alalawi

Faculty Advisor

Dr. Saed Alrabaee

College of Information Technology

Date & Venue

Thursday, 4 April 2024

10:30 AM – 12:30 PM

Room 1005, H4 Building

Abstract

In recent years, artificial intelligence (AI) techniques have been used for source code authorship attribution, which is the process of identifying the original author of a given piece of code. With the advancement of AI technologies like ChatGPT, which can generate code, there is a need to accurately identify whether a piece of code is written by a human or generated by a machine. This is crucial for intellectual property protection, cybersecurity, and software forensics. The main objective of this thesis is to review existing research on source code authorship attribution and conduct several experiments to determine the best AI models for identifying the authorship of source code. This includes distinguishing between human-written and ChatGPT-4-generated codes and providing insights into the gender and region. A dataset of 600 source codes was utilized, focusing on extracting lexical and layout features. The study applied several information retrieval and ranking techniques such as TF-IDF, MI, and IG to extract features and understand author characteristics. It also employed various machine learning models, such as SVM, logistic regression, MLP, XGBoost, and random forest. It also employed deep learning models like LSTM, RNN, and CNN to analyze the data. The research achieved up to 94.7% accuracy with the random forest model using TF-IDF in machine learning, and a 95% accuracy rate with the CNN model in deep learning. These results demonstrate the effectiveness of these models in authorship attribution of source code. This work contributes to the field by identifying effective AI models for source code authorship attribution. It provides a comprehensive analysis of how machine learning and deep learning can be used to attribute authorship, differentiating between human and AI-generated code, and identifying gender and region. By achieving high accuracy in identifying the authorship of source codes, this thesis fills a gap in the current understanding and methodologies in source code authorship attribution. It offers new insights and methods in distinguishing between human- and AI-generated code to address the ethical concerns.

Keywords: Authorship Attribution, Machine Learning, Deep Learning, ChatGPT Generated Code Detection, Code Analysis.