SCHOOL OF COMPUTING
Faculty of Engineering

Project Proposal Form MCST1043
Sem: 2 Session: 2024/25

SECTION A: Project Information.

Program Name:	**Masters of Science (Data Science)**
Subject Name:	**Project 1 (MCST1043)**
Student Name:	Muhammad Haziq bin Mohamad
Metric Number:	MCS241036
Student Email & Phone:	muhammadhaziqmohamad@graduate.utm.my & 601110667854
Project Title:	BERT-based Semantic Similarity of Malaysian Legal Precedents
Supervisor 1:	
Supervisor 2 / Industry Advisor(if any):	

SECTION B: Project Proposal

Introduction:

Malaysian law is based on common law. The Malaysian legal system relies significantly on judicial precedents. For instance, legal precedents are earlier court rulings. In similar legal cases, the court uses this ruling as a precedent. It is essential to stare decisis, which means "to stand by things decided." The subordinate court will follow the Malaysian Federal Court's rulings in future instances with similar legal concerns or facts. Because the Federal Court is Malaysia's top court, its decisions bind all lesser courts. The Malaysian judicial system produces thousands of judgements annually. According to the Judicial Appointments Commission (2022), the Federal Court received 1059 new cases and 949 from 2021. 1358 cases were disposed. In 2022, the Court of Appeal heard 11,526 appeals and disposed of 5226 cases. It takes lawyers a long time to research this topic. Also, lawyers use legal precedent to develop arguments. Judges and lawyers use report abstracts, legal concepts, and common sense to evaluate legal case reports, which is laborious (Moro et al., 2023). Many legal documents are difficult to retrieve for relevant instances. Effective methods are needed to discover semantically comparable scenarios. With the advent of digital legal libraries in Malaysia like elaw, CLJLaw, and LexisNexis, NLP can be used to improve legal research. To emphasize, Natural Language Processing (NLP) and transformer-based models like BERT are used to grasp text context. The transformer-based BERT model performed well in semantic similarity tasks, according to Devlin et al. (2018). Furthermore, according to Chalkidis et al. (2020), Legal-BERT performs better in legal text processing because it can capture legal documents' specific linguistics. Thus, this project seeks to provide a BERT-based paradigm for Malaysian legal text documents. Hence, it will improve legal researchers' performance.

Problem Background:

Legal research often involves lawyers. This is essential to legal practice, especially in common law Malaysia. Judicial precedence influences court judgements. However, searching through vast case law volumes for relevant precedent takes time. Malaysian legal information retrieval (IR) relies heavily on keyword matching. Malaysian legal professional use LexisNexis, CLJLaw, and government websites to find precedents. However, as digitized case law grows, finding judicial precedents becomes difficult. These legal platforms use keyword approaches that are ineffective for synonym and contextual meaning searches. Traditional methods include Boolean keyword matching. Traditional keyword methods often work poorly, resulting in irrelevant search results. Developing a BERT-based semantic similarity model can help legal professionals improve legal research. Chalkidis et al. (2020) found that Legal-BERT improves legal text processing performance. In contrast, transformer-based models like BERT are effective legal text processing tools because they capture deeper semantic links between words.

Problem Statement:

Current keyword-based legal document retrieval strategies are examined in this paper. This typical method fails to capture legal texts' semantic context. Therefore, legal practitioners find it difficult to uncover relevant judicial precedents using these present methods. Additionally, current techniques cannot capture deeper semantic links between words. This issue causes key precedents with the same meaning but different phrasing to be overlooked. Inefficient retrieval system limits legal professionals' research efficiency. This project can improve Malaysian precedent legal document retrieval using BERT-based semantic similarity model.

Aim of the Project:

This project aims to develop and evaluate a BERT-based semantic similarity model for Malaysian legal precedents, leveraging Natural Language Processing methods to enhance legal research and explore key linguistic features influencing case similarity.

Objectives of the Project:

1. To develop a semantic similarity model based on BERT for the Malaysian legal precedents that can help enhance the legal research among legal professionals.
2. To identify the key linguistics characteristics that influence semantic similarity in judicial precedents.
3. To evaluate the effectiveness of the BERT-based model in retrieving contextually similar legal cases.

Scopes of the Project:

1. Focus on Malaysian Federal Court and Court of Appeal judgments.
2. Only English-written judgments will be considered.
3. Legal documents will be sourced from publicly available platforms, such as Malaysia Judiciary's e-Court portal, LexisNexis, CLJ Law and other legal databases or official government websites.
4. Involves developing and testing a semantic similarity model based on BERT.

Expected Contribution of the Project:

The expected contribution of this project is to benefit the field of legal natural language processing (Legal NLP). It holds significant value in both academic and practical areas. For instances, it introduces a semantic similarity model that can help for retrieving Malaysian legal precedents. Then, by enhancing the retrieval of cases, it can help to reduce oversight and improve efficiency. In summary, this project aligns with national initiatives like the Malaysia Judiciary's e-Court to incorporate artificial intelligence into legal context.

Project Requirements:

Software:	Python
Hardware:	Intel i5, 16GB Ram
Technology/Technique/ Methodology/Algorithm:	Machine learning, Natural Language Processing, BERT, Semantic

Type of Project (Focusing on Data Science):

[/] Data Preparation and Modeling

[/] Data Analysis and Visualization

[ ] Business Intelligence and Analytics

[/] Machine Learning and Prediction

[ ] Data Science Application in Business Domain

Status of Project:

[/] New

[ ] Continued

If continued, what is the previous title?

SECTION C: Declaration

I declare that this project is proposed by:

[/] Myself

[ ] Supervisor/Industry Advisor ( )

Student Name: MUHAMMAD HAZIQ BIN MOHAMAD

Haziq — **Signature**

8/5/2025 — **Date**

SECTION D: Supervisor Acknowledgement

The Supervisor(s) shall complete this section.

I/We agree to become the supervisor(s) for this student under aforesaid proposed title.

Name of Supervisor 1:

	Signature	Date
Name of Supervisor 2 (if any):		
	Signature	**Date**

SECTION E: Evaluation Panel Approval

The Evaluator(s) shall complete this section.

Result:

[ ] FULL APPROVAL [ ] CONDITIONAL APPROVAL (Major)*
[ ] CONDITIONAL APPROVAL (Minor) [ ] FAIL*

***** Student has to submit new proposal form considering the evaluators' comments.

Comments:

Name of Evaluator 1: .....

Signature **Date**

Name of Evaluator 2: .....

Signature **Date**