Chapter 3: Methodology

BERT-BASED SEMANTIC SIMILARITY OF MALAYSIAN LEGAL PRECEDENTS

NAME: MUHAMMAD HAZIQ BIN MOHAMAD

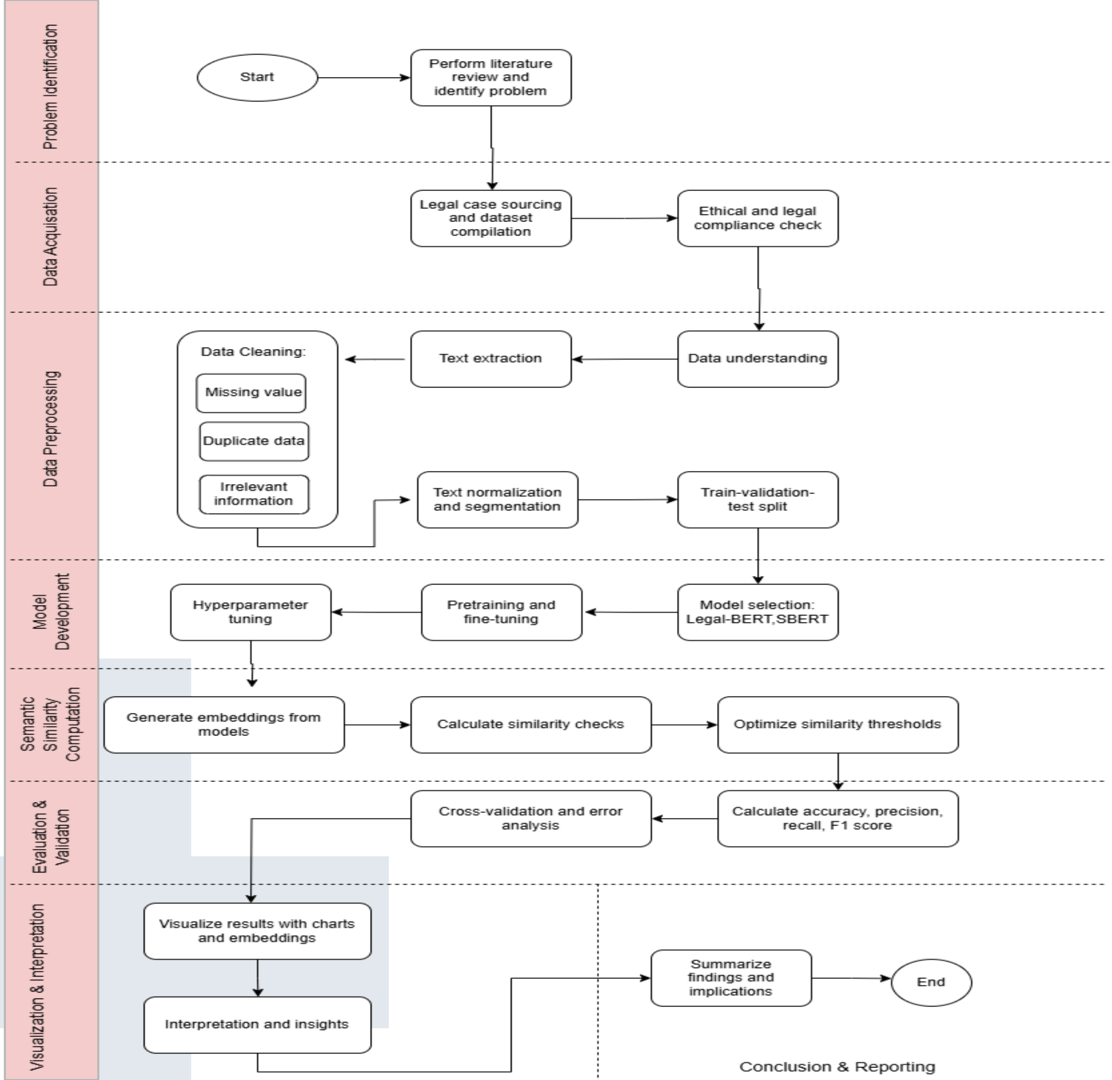
Chapter 3: Research Methodology

This chapter outlines the methodology used to develop the **BERT-based semantic similarity model** for Malaysian legal precedents. The focus is on the systematic approach taken to enhance legal research efficiency and accuracy.

Research Objectives and Research Questions

Research Objective (RO)	Research Question (RQ)
RO1: To investigate existing semantic similarity approaches in legal NLP tasks and identify gaps in the context of Malaysian legal texts.	**RQ1**: What are the challenges and limitations in applying semantic similarity models to Malaysian legal texts?
RO2: To develop, fine-tune, and compute the semantic similarity of Malaysian legal precedents using a BERT-based model.	**RQ2**: How can a BERT-based model be fine-tuned and applied to compute the semantic similarity of Malaysian legal precedents?
RO3: To evaluate and visualize the model's performance using key evaluation metrics.	**RQ3**: How can the model's performance metrics and similarity results be visualized to enhance understanding and interpretation by legal professionals?

Research Framework



Problem Identification

Start

Perform literature review and identify problem

Data Acquisation

Legal case sourcing and dataset compilation

Ethical and legal compliance check

Data Preprocessing

Data understanding

Text extraction

Data Cleaning:

- Missing value
- Duplicate data
- Irrelevant information

Text normalization and segmentation

Train-validation-test split

Model Development

Model selection: Legal-BERT,SBERT

Pretraining and fine-tuning

Hyperparameter tuning

Semantic Similarity Computation

Generate embeddings from models

Calculate similarity checks

Optimize similarity thresholds

Evaluation & Validation

Calculate accuracy, precision, recall, F1 score

Cross-validation and error analysis

Visualization & Interpretation

Visualize results with charts and embeddings

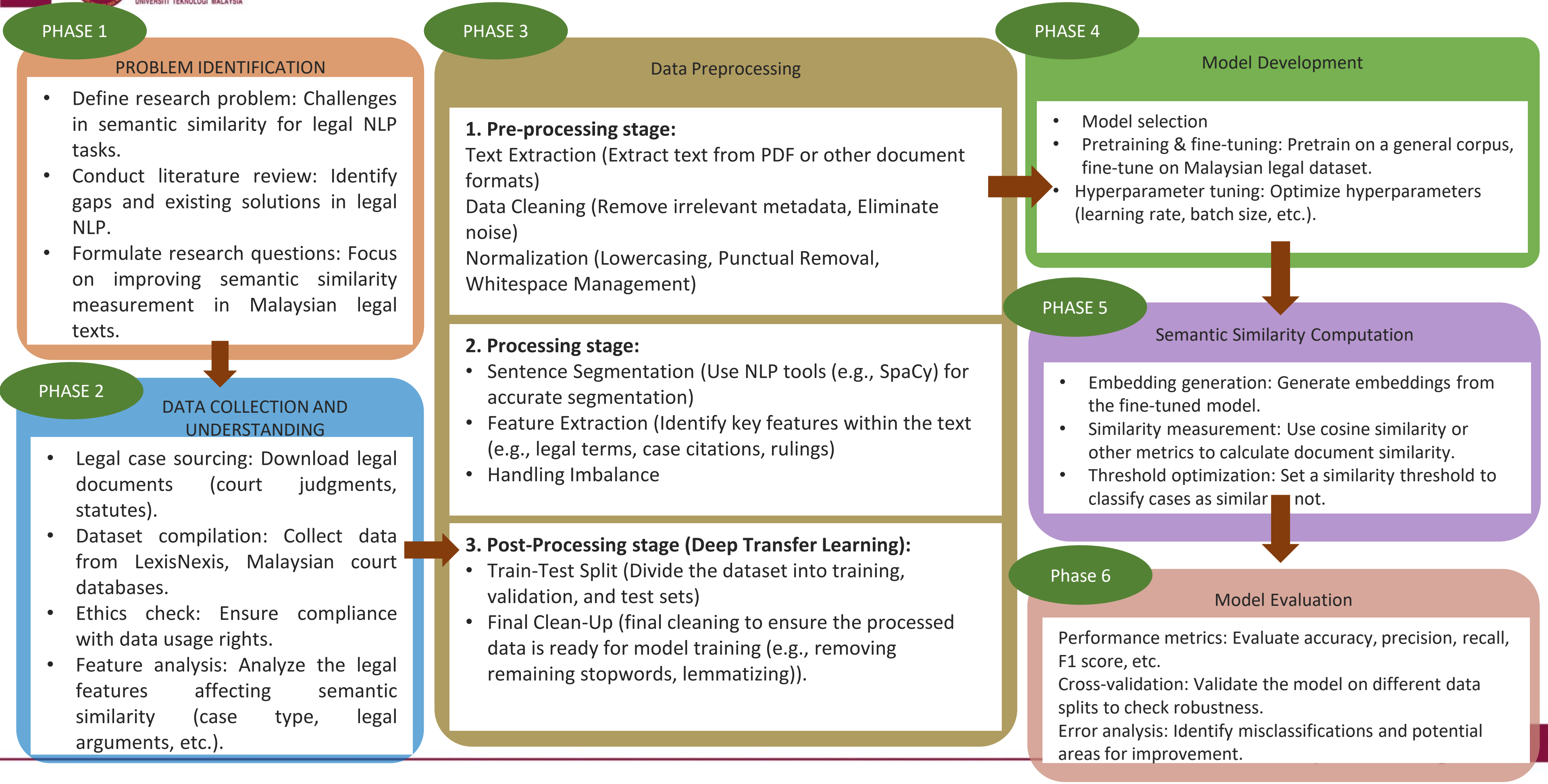
Interpretation and insights

Conclusion & Reporting

Summarize findings and implications

End

RESEARCH FRAMEWORK



PHASE 1

PROBLEM IDENTIFICATION

- Define research problem: Challenges in semantic similarity for legal NLP tasks.
- Conduct literature review: Identify gaps and existing solutions in legal NLP.
- Formulate research questions: Focus on improving semantic similarity measurement in Malaysian legal texts.

PHASE 2

DATA COLLECTION AND UNDERSTANDING

- Legal case sourcing: Download legal documents (court judgments, statutes).
- Dataset compilation: Collect data from LexisNexis, Malaysian court databases.
- Ethics check: Ensure compliance with data usage rights.
- Feature analysis: Analyze the legal features affecting semantic similarity (case type, legal arguments, etc.).

PHASE 3

Data Preprocessing

1. Pre-processing stage:
Text Extraction (Extract text from PDF or other document formats)
Data Cleaning (Remove irrelevant metadata, Eliminate noise)
Normalization (Lowercasing, Punctual Removal, Whitespace Management)

2. Processing stage:

- Sentence Segmentation (Use NLP tools (e.g., SpaCy) for accurate segmentation)
- Feature Extraction (Identify key features within the text (e.g., legal terms, case citations, rulings)
- Handling Imbalance

3. Post-Processing stage (Deep Transfer Learning):

- Train-Test Split (Divide the dataset into training, validation, and test sets)
- Final Clean-Up (final cleaning to ensure the processed data is ready for model training (e.g., removing remaining stopwords, lemmatizing)).

PHASE 4

Model Development

- Model selection
- Pretraining & fine-tuning: Pretrain on a general corpus, fine-tune on Malaysian legal dataset.
- Hyperparameter tuning: Optimize hyperparameters (learning rate, batch size, etc.).

PHASE 5

Semantic Similarity Computation

- Embedding generation: Generate embeddings from the fine-tuned model.
- Similarity measurement: Use cosine similarity or other metrics to calculate document similarity.
- Threshold optimization: Set a similarity threshold to classify cases as similar not.

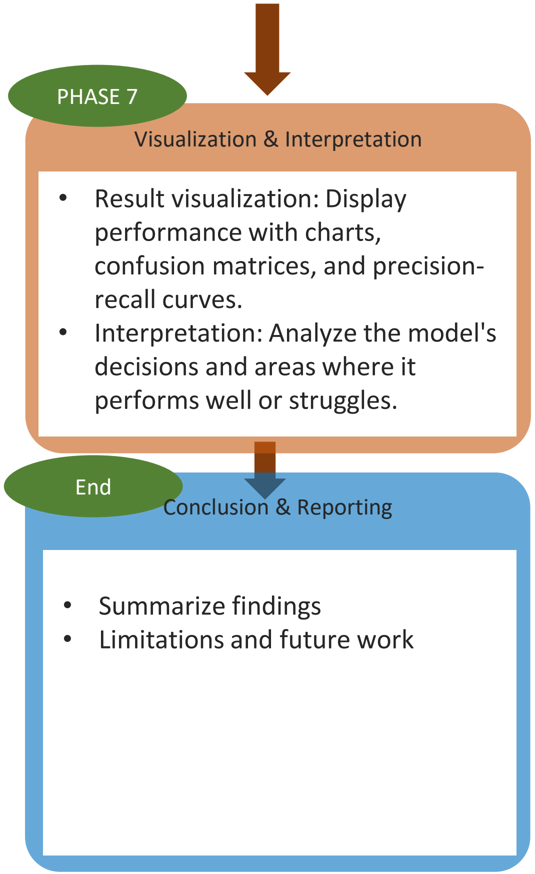
Phase 6

Model Evaluation

Performance metrics: Evaluate accuracy, precision, recall, F1 score, etc.
Cross-validation: Validate the model on different data splits to check robustness.
Error analysis: Identify misclassifications and potential areas for improvement.

RESEARCH FRAMEWORK

Cont.



PHASE 7

Visualization & Interpretation

- Result visualization: Display performance with charts, confusion matrices, and precision-recall curves.
- Interpretation: Analyze the model's decisions and areas where it performs well or struggles.

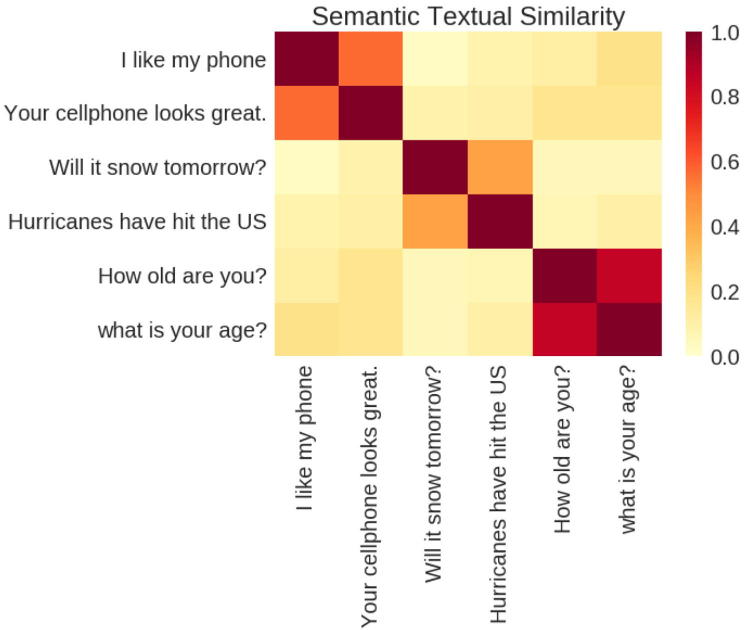
End

Conclusion & Reporting

- Summarize findings
- Limitations and future work

Problem Identification

The challenge of **semantic similarity** in Malaysian legal precedents is significant. The motivation behind this research is to enhance legal research efficiency and accuracy, allowing legal professionals to find relevant precedents more effectively.



RESEARCH FRAMEWORK

Phase 2 : Data Collection And Data Understanding

Dataset	Features	Cases/Instances	Legal Domain	Up-to-date Relevance	Legal Text Representation	Distinctive Characteristics
Malaysian Legal Dataset	60+ (e.g., case type, court, judgment)	**5,000+** (selected for training)	Malaysian Law	Focuses on **recent cases** in Malaysian law	Full case texts, judgments, rulings	Includes **citations, legal references**, and **metadata**
LexisNexis Malaysian Case Law	50+ (e.g., citation, case type, year)	**10,000+** (subset of 100,000+ cases)	Malaysian law	Covers **current legal trends in Malaysia**	Full judgments and summaries (PDF and other formats)	**High-quality**, authoritative content used in legal research
Court Case Corpus (Malaysia)	45+ (e.g., court, case facts, issues)	**2,000+** (random sample)	Criminal, Civil, Constitutional	Publicly available **recent judgments**	Text of judgments, rulings	Focus on **common legal issues** in Malaysia
Public Malaysian Legal Data	40+ (e.g., court level, case year)	**1,000+** (public cases)	Civil, Criminal, Administrative	Available through **open government channels**	Legal documents, summaries, statutes	Emphasizes **open access** and **government transparency**

RESEARCH FRAMEWORK

PHASE 3: Data Processing

Stage	Task	Description
Stage 1: Data Preprocessing	**Text Extraction**	Extract legal case text from downloaded PDF or other formats (e.g., LexisNexis).
	Data Cleaning	Remove irrelevant metadata, OCR artifacts, and noise from the text.
	Normalization	Convert text to lowercase, remove punctuation, extra spaces, etc.
	Tokenization	Split the text into tokens (e.g., words, sentences, or paragraphs).
Stage 2: Feature Engineering and Segmentation	**Text Segmentation**	Divide the text into meaningful sections (e.g., facts, issues, judgments).
	Feature Extraction	Extract legal features such as case type, legal terms, and citations.
	Handling Class Imbalance	Apply techniques (e.g., SMOTE, random oversampling) to balance dataset.
Stage 3: Post-Processing	**Train-Test Split**	Split the data into training, validation, and testing datasets.
	Text Vectorization	Convert text to machine-readable embeddings using models like BERT.
	Threshold Optimization	Set similarity thresholds to classify legal case relevance.

RESEARCH FRAMEWORK

PHASE 4: Model Development

Task	Description
Model Selection	**BERT-based model** (**Legal-BERT**, **SBERT**) for semantic similarity tasks.
Pretraining and Fine-Tuning	Pretrain the model on a generic corpus, and then **fine-tune** it with the **Malaysian legal dataset** to adapt it to the domain.
Hyperparameter Tuning	Tune hyperparameters such as **learning rate**, **batch size**, **number of epochs**, to improve model performance.

PHASE 5: Semantic Similarity Computation

Task	Description
Embedding Generation	Generate **semantic embeddings** for the legal documents using the fine-tuned model.
Similarity Measurement	Measure semantic similarity using metrics like **cosine similarity**, **Manhattan distance**, or other suitable metrics.
Threshold Optimization	Set an appropriate similarity threshold to classify legal precedents as **relevant** or **irrelevant**.

RESEARCH FRAMEWORK

PHASE 6: Evaluation and Validation

Task	Description
Performance Metrics	Evaluate the model’s performance using metrics like **accuracy**, **precision**, **recall**, and **F1-score**.
Cross-Validation	Perform **k-fold cross-validation** to ensure model robustness and generalization.
Error Analysis	Analyze misclassifications to identify specific errors or weaknesses in the model’s predictions.

RESEARCH FRAMEWORK

PHASE 7: Visualization & Interpretation

Task	Description
Result Visualization	Visualize model performance using **charts**, **graphs**, and **confusion matrices**.
Interpretation of Results	Interpret the model’s results to understand why certain cases were deemed similar or dissimilar.
Insights Generation	Generate insights based on the results, including the model's strengths and weaknesses in legal case retrieval.

Summary

Research Phase	Research Objectives (RO)	Deliverables
Phase 1: Problem Identification and Literature Review	**RO1**	Chapter 1 & 2
Phase 2: Data Acquisition and Dataset Preparation	**RO2**	Chapter 4
Phase 3: Data Preprocessing	**RO2**	Chapter 4
Phase 4: Model Development and Fine-Tuning	**RO2**	Chapter 4
Phase 5: Semantic Similarity Computation	**RO2**	Chapter 4
Phase 6: Model Evaluation and Validation	**RO3**	Chapter 5
Phase 7: Visualization and Interpretation	**RO3**.	Chapter 6
Phase 8: Conclusion and Reporting		Chapter 6