Chapter 3

Research Methodology

3.1 Introduction

In this chapter, the main discussion is to explain in detail the research methodology that is used for the BERT-based Semantic Similarity of Malaysian Legal Precedents project. For instance, the discussion explains the method used step-by-step from the start until the end of this project. This will be further discussed in the research framework part, which includes the cycles of data science projects. For example, problem background, process of data collection, data pre-processing, data modeling, model evaluation, and finding and visualization. This portion also will be integrated with the research framework to further align it with research goals. This study aims to help the legal professional by enhancing their legal research through BERT-based semantic similarity of legal precedents.

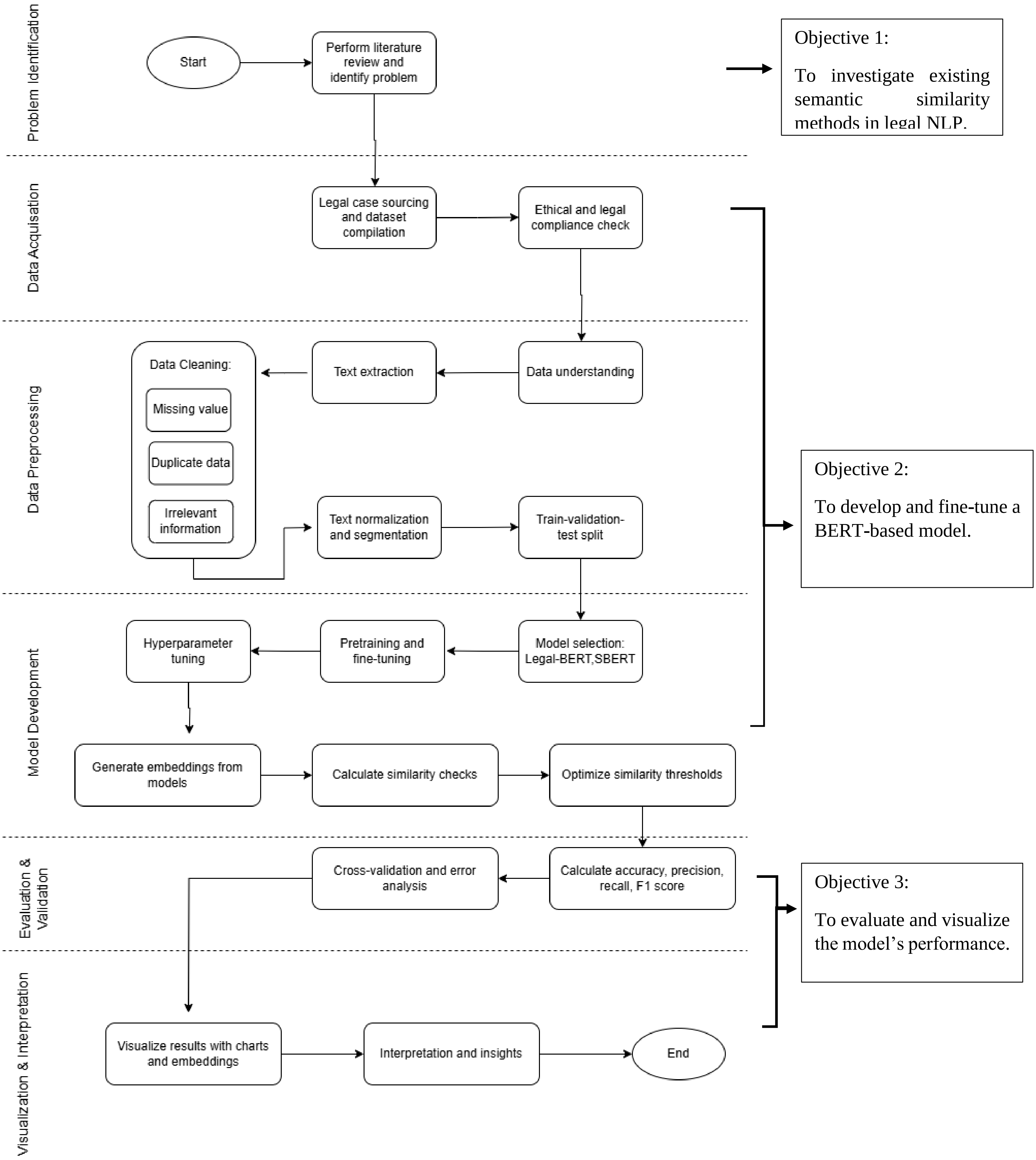
3.2 Research Framework

Firstly, this chapter will introduce the research framework of this project. These are important steps to get the clear picture of the process from the beginning until the end. According to Salinas-Atausinchi et al. (2023), research frameworks are important as they provide a basis for interpreting data and findings, allowing researchers to connect their results to existing theories and knowledge. Therefore, it served as the basis for the lenses through which researchers design, conduct, and analyze their studies. Then, to ensure that the research questions and methodologies are aligned.

Therefore, the research framework that used in this project includes the following steps:

1. Problem Identification and Literature Review
2. Data Collection
3. Data Preprocessing
4. Model Development
5. Evaluation and Validation
6. Visualization and Interpretation

Diagram for the Research Framework:


Problem Identification
Start
Perform literature review and identify problem
Data Acquisition
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
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
End

Objective 1:

To investigate existing semantic similarity methods in legal NLP.

Objective 2:

To develop and fine-tune a BERT-based model.

Objective 3:

To evaluate and visualize the model's performance.

Figure 1: Flowchart of Research Framework For BERT-based Semantic Similarity of Malaysian Legal Precedents

This project research framework contains 6 phases, and each phase contributed to a milestone. Firstly, it started with the problem identification and literature review. This phase is important as it served as the first step of this project. The problem identified is used to further analyze by conducting the literature review. Then, the second phase is data collection from trusted sources such as the government legal website and business-oriented legal websites such as LexisNexis and other open data sources. This phase is challenging, as the data collected will be the main reason for this project to succeed. Next, the third phase is data development and organizing. This includes data preprocessing, data processing, and data post-processing. This phase is divided into 3 steps to ensure the data is ready and in perfect condition for the next phase. Then, the next phase is model development. Once the data development is complete, the model utilizes the data. Moreover, the model will be pretrained and fine-tuned using the dataset to ensure the result is accurate. The next phase is semantic similarity computation to calculate the similarity from the embeddings generated by the model. Furthermore, the model will be evaluated using key metrics such as cosine similarity, accuracy, precision, and F1 score to ensure the model's performance is excellent. Lastly, the findings will be visualized to get the meaningful insight from the model.

3.3 Phase 1: Problem Identification and Literature Review

The problem identification is crucial for the whole research framework. This problem identification is a critical step that ensures the smoothness of the formulation of research objectives and the development of a systematic methodology. Therefore, in this project, problem identification is the first step to ensure an overall understanding of the interesting topics. In this study, the research begins with a review of existing literature reviews to gain an understanding of the research domain, which are the current advancements and limitations in the application of Natural Language Processing (NLP), particularly in semantic similarity tasks. For instance, the previous studies had introduced models and solutions for the research domain. Chakidis (2020) found that the BERT that trained with legal corpora has shown superior performance in legal-related tasks, in which the model is named Legal BERT. However, the model is trained on other countries' jurisdictions and differences from the Malaysian legal structure, which has the bilingual format and is embedded in a distinct legal tradition influenced by both civil and common law. Besides, the integration of NLP in legal has been widely used in other countries such as China and the United States. According to Paul, Mandal, Goyal, and Ghosh (2023), the development of models like Legal BERT and CaseLawBERT has improved performance in various legal tasks, indicating a growing sophistication in legal NLP applications. Furthermore, Wang et al. (2019) stated that the legal practices in China have employed advanced NLP techniques to understand legal context and provide tailored services, whereby the integration of those technologies marks a significant step. However, despite the advancement of using NLP in the legal domain, the Malaysian legal field has not yet widely adopted semantic NLP models. Therefore, this opens an opportunity for this project to address this gap. Hence, it may assist in improving the retrieval and interpretation of legal documents that will improve the current system.

3.4 Phase 2: Data Acquisition

The legal dataset was obtained from various sources, including the official Malaysian legal database, LexisNexis, a business-oriented legal database website, and other public data sources. It consists of a high-quality dataset of Malaysian legal documents to train and evaluate semantic similarity models. Furthermore, the documents were in the form of lengthy files that could be downloaded in PDF format from a variety of sources. This dataset is primarily considered from LexisNexis, a legal platform to which Universiti Teknologi Malaysia (UTM) has an active subscription. This legal platform contains a large corpus of Malaysian case law, and potentially up to 100,000 cases may be accessed for academic research. This data necessitated additional extraction and cleaning, which will be elaborated upon in the subsequent phase.

Nevertheless, there are constraints, such as copyright and usage restrictions, that are associated with the collection of significant amounts of data. Consequently, the project's ethical and legal compliance was thoroughly reviewed, and additional authorization is required to ensure its successful completion. Consequently, this research also investigated supplementary sources, including the official Malaysian legal database platforms, CommonLII, and other publicly accessible sources.

Example of Legal Cases downloaded:

Case Name	Court	Date	Citation	Legal Issues	Decision	Source
MENTERI HAL EHWAL LUAR NEGERI, MALAYSIA & ORS v SUNDRA RAJOO A/L NADARAJAH	Court of Appeal (Putrajaya)	6 October 2020	[2020] MLJU 1567; [2021] 2 MLJ 787	Immunity, AG’s discretion, judicial review	Appeal allowed; immunity not granted; judicial review not applicable.	LexisNexis

Datin Seri Rosmah bt Mansor v Public Prosecutor and another appeal	Court of Appeal	2022	[2022] 3 MLJ 601	Evidentiary rules, judicial conduct, high-profile corruption	Decision not fully stated (appeal likely dismissed or ongoing).	LexisNexis
IKI PUTRA BIN MUBARRAK v KERAJAAN NEGERI SELANGOR & ANOR	Federal Court	2021	[2021] 2 MLJ 323; [2021] MLJU 211, 212, 213	Constitutionality, legislative power, Islamic law	Appeal allowed; section invalidated as ultra vires the Constitution.	LexisNexis
SUNDRA RAJOO A/L NADARAJAH v MENTERI LUAR NEGERI, MALAYSIA & ORS	Federal Court	9 June 2021	[2021] MLJU 943; [2021] 5 MLJ 209	Functional immunity, AG's discretion, judicial review	Appeal allowed; appellant entitled to immunity.	LexisNexis
PUBLIC PROSECUTOR v DATO' SRI MOHD NAJIB BIN HJ ABD RAZAK	High Court	2020	[2020] MLJU 1254; [2020] 11 MLJ 808	Corruption, criminal breach of trust, abuse of power	Found guilty on all charges.	LexisNexis
IKI PUTRA BIN MUBARRAK v KERAJAAN NEGERI	Federal Court	2021	[2021] MLJU 213	Constitutional supremacy, state vs federal conflict	Section declared unconstitutional.	LexisNexis

SELANGOR & ANOR						

Figure 2: Example of Legal Cases Downloaded

3.5 Phase 3: Data Preparation

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: Data Processing	**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 2: Data Post-Processing	**Train-Test Split**	Split the data into training, validation, and testing datasets.

Figure 3: Table of Data Preparation phase

There are 3 steps involved in this phase, which are data preprocessing, data processing, and data post-processing. The reason behind this is to ensure that the dataset is adequately cleaned, structured, and transformed for the next phase, which is the model development.

Initially, the legal precedent cases that were downloaded from a variety of sources were converted to.txt format. This is to ensure compatibility with transformer models like Legal BERT, which process raw text input. At the data preprocessing stage, the missing values, duplicate entries, and irrelevant noise that appeared in the dataset were identified and addressed. This is to ensure that the data is a high-quality dataset that is reliable and consistent. Therefore, there are a few things to be considered when handling the missing and duplicate data. Firstly, the type of missing data should be defined at the earlier step before the cleaning process. The reason behind this is to ensure that the missing data could be considered for whether to remove it if it was irrelevant or replace it with other values when it carried the important features in the dataset. Besides, the data normalization involved in this phase converts text to lowercase, removes punctuation, and removes unnecessary whitespace. After that, the tokenization process is done to turn text into smaller units such as words, sentences, or paragraphs. Tokenization is an important step in the text preprocessing pipeline because the legal document is lengthy and complex. Therefore, tokenization helps to convert the raw legal documents into smaller and meaningful units. Besides, it makes it easier to manage when working with transformer-based language models like Legal-BERT and SBERT in the model development phase.

Next, for the data processing stage, including text, segmentation, feature extraction, and handling class imbalance. This stage focusing on enhancing the structure and semantic of the cleaned data.

Lastly, the final stage which is data post-processing to ensure that the dataset is ready for model training and evaluation. Therefore, in this stage, the processed data was split into training and testing parts for enable unbiased evaluation of model performance.

3.6 Phase 4: Model Development

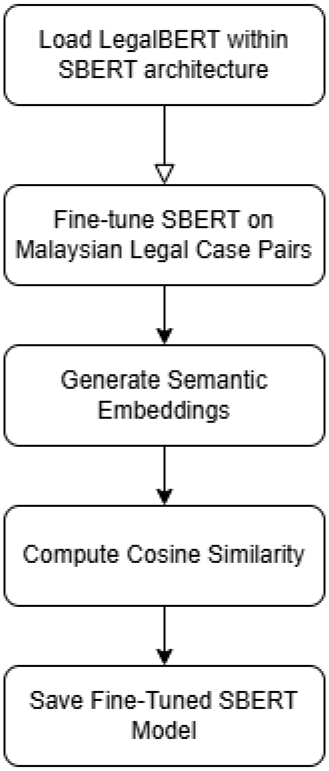

Load LegalBERT within SBERT architecture
Fine-tune SBERT on Malaysian Legal Case Pairs
Generate Semantic Embeddings
Compute Cosine Similarity
Save Fine-Tuned SBERT Model

Figure 4: Pipeline of Model Development phase

In this phase, the processed data that was cleaned and tokenized legal texts were transformed into dense vector representations using Bidirectional Encoder Representations from Transformers (BERT). Therefore, it enables measuring the semantic similarity between Malaysian legal precedents. Besides, it also enables the model to capture deep contextual meaning that can understand the relationship between words in both directions, which are left and right, at the same time. Furthermore, this project employed Legal-BERT and SBERT for greater domain relevance. This pretrained model is fine-tuned on legal corpora and able to represent more accurate legal terminology and phrasing. However, this model was pretrained on general legal corpora and not able to capture the complex meaning of Malaysian legal texts. Therefore, in this project, this model was fine-tuned on a curated dataset of Malaysian legal precedent cases that is paired to enhance the understanding of the contextual meaning of Malaysian legal texts.

To be further explained, the dataset of cleaned and processed Malaysian legal case documents was curated and transformed into sentences or pairs. Then, the pairs were assigned a similarity score ranging from 0 to 1, in which 0 is completely unrelated and 1 is highly similar. For instance, the cases with overlapping legal principles are considered similar pairs, and it's the same with the same statutory provisions or factual scenarios. Hence, this labeled dataset served as the training and validation set for the fine-tuning process.

Next, Legal BERT was adopted as the base model because of its exposure to legal language. This base model was integrated with SBERT architecture to begin the fine-tuning process using Malaysian legal precedent pairs. This is to ensure that the model was able to compute the semantic similarity in a domain-specific context. After that, the fine-tuning was performed using SBERT, which is designed for the sentence's similarity tasks. The input pair was encoded by the model into dense vector representations for more optimization and to increase the distance between dissimilar pairs.

Below are key fine-tuning settings:

- **Base Model:** Legal-BERT (for legal domain adaptation)
- **Architecture:** SBERT (Siamese structure with mean pooling)
- **Loss Function:** Cosine Similarity Loss or Triplet Loss
- **Batch Size:** 16–32
- **Epochs:** 3–5 (with early stopping based on validation loss)
- **Learning Rate:** 2e-5 to 5e-5
- **Tokenizer:** Pretrained Legal-BERT tokenize

Hence, this process was executed using the sentence-transformers Python library because it provided seamless integration for training. Therefore, the better-aligned model with the linguistic nuances of Malaysian legal texts was prepared for the next phase, which is the evaluation phase.

3.6.1 Semantic Embedding Generation

This process was done after the fine-tuning process to convert the full legal cases or extracted summaries into fixed-length dense vector representations. This embedding was then capturing the semantic content of the legal text into numerical form. This phase is a crucial step, as it allowed the comparison of legal documents beyond the keyword matching. Therefore, this embedding is generated by passing the text through the SBER architecture and applying a pooling strategy to produce a single vector per document. Hence, it served as compact and high-dimensional representations of the legal content.

3.6.2 Similarity Scoring

Therefore, the obtained embeddings were evaluated using cosine similarity to measure the closeness between the two legal documents or segments. For instance, cosine similarity with 1 is considered high, and -1 is considered low. The scoring mechanism enables the system to identify and retrieve precedent cases with deeper contextual meaning. Hence, the model offers a meaningful and interpretable method for identifying legally relevant documents.

Cosine Similarity was defined as:

$$Cosine\ Similarity = \frac{\mathrm{A} \cdot \mathrm{B}}{||\mathrm{A}|| \times ||\mathrm{B}||}$$

Where:

- A·B is the dot product of the two vectors
- ‖A‖ and ‖B‖ are the Euclidean norms (magnitudes) of the vectors

3.7 Phase 5: Evaluation and Validation

In this phase, the performance of the model was evaluated using the evaluation metrics. For instance, it was employed based on the nature of the annotated labels of the combination of classification and ranking metrics. For the classification metrics, the accuracy, precision, recall, and F1-score were evaluated. Besides, for the continuous similarity scores, which were the ranking or regression metrics it included, Pearson Correlation Coefficient, Spearman Rank Correlation, and Mean Squared Error (MSE).

3.7.1 Error Analysis

This analysis was conducted to identify the specific instances where the model misjudged the semantic similarity. For instance, the attention was given to the pairs that the model predicted with high similarity but unrelated cases, and vice versa. Therefore, it provides meaningful insights that will be used for potential improvements.

3.8 Phase 6: Visualization and Interpretation

In the last phase, the quantitative evaluation that had been done was visualized using techniques such as t-SNE (t-Distributed Stochastic Neighbor Embedding) or UMAP (Uniform Manifold Approximation and Projection). This method of visualization projects the high-dimensional embeddings into a two-dimensional space. There, the result helped the interpretation of how the model grouped semantically similar cases. Consequently, the visualization highlights potential clustering patterns among the various categories of legal matters. Hence, the result of the visualization will be reported to be ready for the next chapter.

3.9 Conclusion

In conclusion, this chapter explained the research framework as well as the steps that needed to be carried out to ensure this project is done smoothly. Therefore, the first objective, which is to investigate existing semantic similarity methods in legal NLP, is done in chapter 2. The next chapter will discuss the research design and implementation.

References:

Paul, S., Mandal, A., Goyal, P., & Ghosh, S. (2023). Pre-trained language models for the legal domain: A case study on Indian law (arXiv:2209.06049v5). arXiv. https://arxiv.org/abs/2209.06049v5

Wang, Z., Wang, B., Duan, X., Wu, D., Wang, S., Hu, G., & Liu, T. (n.d.). IFlyLegal: A Chinese legal system for consultation, law searching, and document analysis. State Key Laboratory of Cognitive Intelligence, iFLYTEK Research; Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology; iFLYTEK AI Research (Hebei).