BERT-BASED SEMANTIC SIMILARITY OF MALAYSIAN LEGAL PRECEDENTS

MUHAMMAD HAZIQ BIN MOHAMAD

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name	:	MUHAMMAD HAZIQ BIN MOHAMAD			
Student's Matric No.	:	MCS241036	Academic Session	:	SEMESTER 2, 2024/2025
Date of Birth	:	04 NOVEMBER 2002	UTM Email	:	muhammadhaziqmohamad@graduate.utm.my
Title	:	BERT-BASED SEMANTIC SIMILARITY OF MALAYSIAN LEGAL PRECEDENTS			

I declare that this project report is classified as:

☒	**OPEN ACCESS**	I agree that my report to be published as a hard copy or made available through online open access.
☐	**RESTRICTED**	Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*
☐	**CONFIDENTIAL**	Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the project report belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this project report for academic exchange.

Signature of Student:

Signature : *Haziq*

Full Name: MUHAMMAD HAZIQ BIN MOHAMAD
Date : 28 JUNE 2025

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I ASSOC. PROF. DR. MOHD SHAHIZAN BIN OTHMAN	Full Name of Supervisor II
Date :	Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

"I hereby declare that we have read this project report and in my opinion this project report is sufficient in term of scope and quality for the award of the degree of Master in Data Science"

Signature	:	_____
Name of Supervisor I	:	ASSOC. PROF. DR. MOHD SHAHIZAN BIN OTHMAN
Date	:	

Signature	:	_____
Name of Supervisor II	:	
Date	:	

Signature	:	_____
Name of Supervisor III	:	
Date	:	

Declaration of Cooperation

This is to confirm that this research has been conducted through a collaboration Muhammad Haziq Bin Mohamad and Universiti Teknologi Malaysia (UTM)

Certified by:

Signature :
Name :
Position :
Official Stamp
Date

* This section is to be filled up for theses with industrial collaboration

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar **:**

Nama dan Alamat Pemeriksa Dalam **:**

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

BERT-BASED SEMANTIC SIMILARITY OF MALAYSIAN LEGAL PRECEDENTS

MUHAMMAD HAZIQ BIN MOHAMAD

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master in Data Science

School of Computing
Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2025

DECLARATION

I declare that this project report entitled *"BERT-Based Semantic Similarity of Malaysian Legal Precedents"* is the result of my own research except as cited in the references. The project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	.....HAZIQ.....
Name	:	MUHAMMAD HAZIQ BIN MOHAMAD
Date	:	28 JUNE 2025

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, ASSOC. PROF. DR. MOHD SHAHIZAN BIN OTHMAN, for encouragement, guidance, critics and friendship. Without his continued support and interest, this thesis would not have been the same as presented here.

To my parent that always supporting me to be able to complete this master's degree. I am very grateful for their full support towards me. Because of that, I able to take this Master's Degree. Therefore, through this journey, I hope it will make them proud.

My fellow postgraduate students must also be acknowledged for their support. My sincere appreciation also goes to all my colleagues and others who have provided assistance on various occasions. Their views and suggestions are indeed useful. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

The increasing volume of digital legal documents within the Malaysian judiciary presents a significant challenge for efficient precedent analysis and legal research. Traditional keyword-based search methods often fail to capture the semantic nuances and contextual complexities inherent in legal texts, leading to incomplete or irrelevant results. Therefore, this project addresses this critical gap by investigating, developing, and evaluating a sophisticated deep learning model for semantic similarity tailored to Malaysian legal precedents. Thus, the primary objective is the development and fine-tuning of a bespoke BERT-based model, leveraging the Sentence-BERT (SBERT) architecture, to accurately quantify the semantic relationship between passages of legal text. For instance, the methodology follows a structured pipeline, commencing with the sourcing and compilation of a specialized corpus of Malaysian legal cases. A rigorous data cleaning and preprocessing phase was executed, featuring text normalization tailored to local legal statutes and terminology. The "all-MiniLM-L6-v2" model was subsequently fine-tuned on this curated dataset using a cosine similarity loss function to optimize for the semantic similarity task. Furthermore, a comprehensive, multi-faceted evaluation framework was implemented to validate the model's performance. The assessment involved quantitative analysis using a suite of metrics including accuracy, precision, recall, F1-score, and Pearson correlation, alongside a qualitative domain-specific analysis to gauge effectiveness across different areas of law. Then, the performance of the fine-tuned model was compared with other baseline model such as Bag of Words (Bow) and Term Frequency-Inverse Document Frequency (TF-IDF). The evaluation was augmented with key visualizations, including confusion matrices and score distribution plots, to provide intuitive insights into the model's predictive behavior. The results demonstrate that the fine-tuned model effectively captures the complex semantic relationships within Malaysian legal discourse, showing significant promise over generalized models. Hence, this research contributes a valuable, domain-specific tool that can enhance legal discovery, support case law analysis, and ultimately improve the efficiency of legal practitioners in Malaysia.

ABSTRAK

Jumlah dokumen undang-undang digital yang semakin meningkat dalam badan kehakiman Malaysia memberikan cabaran besar untuk analisis duluan yang cekap dan penyelidikan undang-undang. Kaedah carian berasaskan kata kunci tradisional sering gagal menangkap nuansa semantik dan kerumitan kontekstual yang wujud dalam teks undang-undang, yang membawa kepada hasil yang tidak lengkap atau tidak relevan. Oleh itu, projek ini menangani jurang kritikal ini dengan menyiasat, membangun dan menilai model pembelajaran mendalam yang canggih untuk persamaan semantik yang disesuaikan dengan duluan undang-undang Malaysia. Oleh itu, objektif utama ialah pembangunan dan penalaan halus model berasaskan BERT yang dipesan lebih dahulu, memanfaatkan seni bina Sentence-BERT (SBERT), untuk mengukur secara tepat hubungan semantik antara petikan teks undang-undang. Sebagai contoh, metodologi mengikut saluran paip berstruktur, bermula dengan penyumberan dan penyusunan korpus khusus kes undang-undang Malaysia. Fasa pembersihan dan prapemprosesan data yang rapi telah dilaksanakan, menampilkan penormalan teks yang disesuaikan dengan statut dan istilah undang-undang tempatan. Model "semua-MiniLM-L6-v2" kemudiannya diperhalusi pada set data susun atur ini menggunakan fungsi kehilangan persamaan kosinus untuk mengoptimumkan tugasan persamaan semantik. Tambahan pula, rangka kerja penilaian yang komprehensif dan pelbagai aspek telah dilaksanakan untuk mengesahkan prestasi model. Penilaian melibatkan analisis kuantitatif menggunakan set metrik termasuk ketepatan, ketepatan, ingatan semula, skor F1 dan korelasi Pearson, bersama analisis khusus domain kualitatif untuk mengukur keberkesanan merentas bidang undang-undang yang berbeza. Kemudian, prestasi model yang diperhalusi dibandingkan dengan model garis dasar lain seperti Bag of Words (Bow) dan Term Frequency-Inverse Document Frequency (TF-IDF). Sehubungan itu, penilaian telah ditambah dengan visualisasi utama, termasuk matriks kekeliruan dan plot taburan skor, untuk memberikan pandangan intuitif ke dalam tingkah laku ramalan model. Keputusan menunjukkan bahawa model yang diperhalusi secara berkesan menangkap hubungan semantik yang kompleks dalam wacana undang-undang Malaysia, menunjukkan janji yang signifikan terhadap model umum. Oleh itu, penyelidikan ini menyumbang satu alat khusus domain yang berharga yang boleh meningkatkan penemuan undang-undang, menyokong analisis undang-undang kes, dan akhirnya meningkatkan kecekapan pengamal undang-undang di Malaysia.

Contents

LIST OF TABLES

TABLE **TITLE** **PAGE**

LIST OF FIGURES

FIGURE **TITLE** **PAGE**

LIST OF ABBREVIATIONS

Abbreviation	Full Name
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
SBERT	Sentence-BERT
LEGALBERT	Legal Bidirectional Encoder Representations from Transformers
BOW	Bag of Words

LIST OF FORMULAS

EQUATION **FORMULAS** **PAGE**

LIST OF APPENDICES

APPENDIX	TITLE	PAGE

Chapter 1: Introduction

1.0 Introduction

In Malaysia, the legal system is mainly rooted in common law. Therefore, Malaysian legal system relies heavily on the judicial precedents. For instance, legal precedents are the decisions that has been made in the past by the courts. This past decision made by the court is used as a reference or authority in deciding future cases that have similar legal issues. To further said, it is a key part of the doctrine of stare decisis, which means "to stand by things decided." For example, the outcome of the cases decides by the Malaysian Federal court will be follow by the lower court in the future cases if it has the same legal issues or facts. This is because the Federal Court is the highest court in Malaysia and its decision are binding on all lower courts.

Therefore, there are thousands of judgments that Malaysian legal system produces annually. According to Judicial Appointments Commission (2022), the Federal Court registered 1059 new cases and 949 cases were brought forward from 2021. The disposed case was 1358. Moreover, the Court of Appeal in 2022 recorded 11,526 appeal cases and disposed of 5226 cases. This issue takes a lot of time for legal practitioners to conduct legal research. Moreover, legal precedent is often relied by the legal practitioners to build arguments. For instance, reading and evaluating legal case reports is labor-intensive for judges and lawyers, who usually base their choices on report abstracts, legal principles, and commonsense reasoning (Moro et al., 2023). However, many legal documents are quite challenging in efficiently retrieving relevant cases. This requires effective tools to identify the semantically similar cases. Furthermore, with the advancement of digital legal repositories in Malaysia such as the elaw and CLJLaw, there is an opportunity to applying Natural Language Processing (NLP) approaches to increase the effectiveness of legal research.

To be highlighted, Natural Language Processing (NLP), specifically transformer-based model such as BERT is used to have better understanding in contextual meaning of text. According to Devlin et al. (2018), the transformer-based model which is BERT showed effective performance in semantic similarity task. Furthermore, according to Chalkidis et al. (2020), Legal-BERT has shown superior performance for the legal text processing task because it is able to capture unique linguistics characteristics of the legal documents.

Hence, this project is aiming to develop the BERT-based model that apply specifically for the Malaysian Legal text documents. Therefore, it will help to enhance the performance during legal research for the legal professionals.

1.1 Problem Background

A legal professional is often involved in legal research. This is a fundamental component of legal practice, especially in Malaysia, which follows a common law jurisdiction. Judicial precedent plays a critical role in decisions made by the court. However, it also requires an extensive amount of time to sift through large volumes of case law to find relevant precedent.

In Malaysia, the current legal information retrieval (IR) predominantly relies on keyword matching. For instance, platforms such as CLJLaw and government official websites are used by legal professionals in Malaysia to locate relevant precedents. However, there are many challenges in locating judicial precedents due to the increasing volume of digitized case law over time. These legal platforms rely on keyword methods that are less effective in searching for synonyms or contextual meanings. For instance, the traditional methods often use Boolean keyword matching. Therefore, the traditional keyword method often lacks in performance that will result in irrelevant search results. This can illustrate by developing BERT-based semantic similarity model that can help the legal professional to improve the quality of legal research. For instance, Chalkidis et al. (2020) stated that the application of Legal-BERT has been proven to improve performance in legal text processing tasks. In contrast, transformer-based models such as BERT are efficient tools for legal text processing because it can help to capture the deeper semantic relationships between words.

1.2 Problem Statement

This study seeks to address he current methods for retrieving legal documents that are based on keyword methods. This traditional method is less effective and insufficient to capture the semantic context of legal documents. Therefore, it is quite challenging for legal professionals to use these current approaches to identify relevant judicial precedents. To further said, the existing tools are also not adequate to capture the deeper semantic relationships between words. This issue leads to the oversight of important precedents that have the same meaning but different expressions. This gap in retrieval system restricts the legal professional to have better efficiency in legal research. Through the application of BERT-based model for semantic similarity, this project can help enhance the retrieval of precedent legal documents in Malaysia.

1.3 Objectives

- To investigate existing semantic similarity methods in the field of legal Natural Language Processing (NLP).
- To develop and fine-tune a BERT-based model tailored for Malaysian legal case texts.
- To evaluate and visualize model performance using comprehensive metrics and analysis.

1.4 Gap Analysis

The application of BERT- model in Natural Language Processing has been used widely in other countries. This tool is proven to understand the context of words in text. For instances, several studies have demonstrated the effectiveness of the BERT model on legal NLP tasks. According to Chalkidis et al. (2020), the studies has introduced Legal-BERT model that trained on a large corpus of United States and Europe Union legal text. It also demonstrated that the use of Legal-BERT is more effective and has better performance than the general BERT model. For example, the model excels in task related to legal question and answer, prediction of judgment and classification of statutes. This shows that the model has a superior ability to capture the specific legal context knowledge.

Next, Zhong et al (2020) has developed the LeCard dataset. The studies proposed a multi-stage deep retrieval framework that using BERT for Chinese legal documents. Besides, it also demonstrated that the application of semantic embeddings increases the effectiveness in retrieval performance if incorporate with domain aware language models. Furthermore, it also outperforms the keyword-based search and rule-based methods. Besides, the studies highlight the importance of fine-tuning transformer models on legal corpora to enhance their effectiveness in downstream legal tasks.

Furthermore, platforms such as elaw, LexisNexis, CLJ Law and government websites relies on the lexical retrieval mechanism. This includes Boolean keyword matching or basic data filtering, which is less effective to capture the deeper semantic relationship between cases. This may further lead to the overlooking of legal documents that use different terminology.

Hence, this project addressees the gaps by developing a BERT-based semantic matching model that apply specifically for Malaysian legal precedents. This project framework will be collecting and preprocessing the court judgments specifically Malaysian Court of Appeal and Federal Court judgments.

Then, generating sentence embedding and computing semantic similarity scores between cases. Therefore, this project aims to enhance the effectiveness of legal research in Malaysia.

1.5 Scope

This project will focus on the Malaysian appellate court judgments which are the Court of Appeal and Federal Court. Besides, the judgment focuses on English written. Primarily, this project is limited to the publicly available legal documents that can be found on government websites and legal databases such as Malaysia Judiciary's e-Court, CLJ Law and others. The scope involves developing and testing a BERT-based semantic similarity model. This can be achieved by meticulous data collection (legal documents), preprocessing, and exploratory analysis. Then, the model will be implemented such as applying pre-trained BERT-based models and fine-tuning. Additionally, the model will be evaluated based on the quantitative metrics such as cosine similarity. Lastly, this project will not cover legal interpretation or development of new legal theories.

1.6 Significance of Research

Hence, this project has important value that contributes to 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. Therefore, according to Chalkidis et al. (2020) and Zhong et al. (2020), BERT-based models have been implemented in other jurisdictions, such as United States, China and Europe for tasks relating to legal case retrieval, judgment prediction and entailment. Furthermore, this project addresses a practical need among legal professionals who spend an extensive amount of time on legal research. This may lead to missing semantically similar cases. 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.

Chapter 2: Literature Review

2.0 Introduction

This chapter reviews existing literature and explores academic research issues, highlighting research issues within the broad scope of global scientific understanding. The chapter begins with a brief introduction to AI and NLP in legal tech focusing on semantic similarity, Transformer models in NLP, and comparison between BERT-based models, as well as a brief overview of evaluation metrics for semantic similarity, which provides a foundation for understanding the research effort.

The number of legal judgements cases in Malaysia is on the rise each year. The conventional keyword-based retrieval system is inadequate due to the substantial volume of legal cases. The semantic similarity model that employs Transformer is necessary. This can expedite the legal research process by reducing the time required for it.

2.1 Implementation AI and NLP in legal Domain

In Malaysia, the implementation of AI has been widely utilized across sector, this also include the legal domain. AI help in many ways to reduce the workload needed by the workers. For instances, Rožman et al. (2023) stated that the implementation of AI will considerably reduce the perceived workload of employees, as well as support organizational culture, leadership, and training. This helps in enhancing engagement and company performance. Furthermore, Malaysian government being serious to implement the artificial intelligence (AI) as part of its national digital transformation agenda. This is demonstrated by the Malaysia National AI Roadmap (2021-2025), which has been implemented by the government. This roadmap is indicative of Malaysia's endeavors to establish AI as a driving force behind technological innovation, enhanced public services, and economic expansion.

Therefore, this initiative also aims to integrate AI across key sectors and legal services also being the parts. This can enhance the decision-making and operational efficiency across various sectors in Malaysia. Furthermore, as part of this efforts, the application of AI has been growing in interest and the field such as natural language processing (NLP) is used to improve the legal research and document analysis.

Besides, this project aims to improve the efficiency of legal research. Besides, with the increasing of court judgments, the conventional method of legal retrieval is no longer necessary. In Malaysia, the efficiency of legal research can be improved by interpreting and analyzing legal text with the assistance of NLP, particularly Legal NLP. This is because NLP has the capabilities to capture the intricate contextual meaning to the fact that the contextual meaning that is essential in legal interpretation. For instances, NLP has demonstrated better accuracy in the identification of relevant legal information, a critical component of accurate legal interpretation and application (Seyler et al., 2020).

2.2 Semantic Similarity

The degree of the two texts that share meaning is known as semantic similarity. Semantic similarity is employed to see how two sentences is similar between each other even they use differences words. This is a crucial aspect as Natural Language Processing (NLP) utilized it for applications such as information retrieval. Furthermore, semantic similarity is evaluated through different methodologies such as embedding techniques and similarity metrics to measure the words closeness. To be highlighted, the used of semantic similarity technique will improve the efficiency of information retrieval systems especially for the legal precedents retrieval. For instance, Shaharao et al. (2024) explained that embedding techniques such as word2vec, GloVe, and BERT turned the text into vector representations to be able to calculate the semantic similarity using metrics such as cosine similarity. Furthermore, Shaharao et al. (2024) stated the semantic similarity will improve the efficiency of learner performance by retrieving relevant information efficiently.

2.3 Traditional Methods in Semantic Similarity

Early methods for the semantic similarity tasks in Natural Language Processing (NLP) is relying on the statistical and lexical representations of text. For instances, among the traditional methods, BAG-of-words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) have been used widely for the semantic similarity tasks. This is because, these traditional models often used due to their simplicity and interpretability (Jurafsky & Martin, 2021; Turney & Pantel, 2010).

2.3.1 Bag-of-Words (BoW)

Therefore, this model represents text by converting the sentences into vector of word counts. For instances, it will ignore the grammar, word order, and context for the semantic similarity tasks. Moreover, each document is reduced to a collection of its individual words. This means that the semantic similarity is measured using the distances metrics such as cosine similarity or Euclidean distance.

However, these traditional methods often ignoring the semantic meaning between words. This is not suitable for the tasks that required the strong semantic similarity and contextual relationships between words. Furthermore, this method also facing difficulty with the vocabulary mismatch. For instances, it struggles to recognize the different terms with similar meanings.

2.3.2 Term Frequency-Inverse Document Frequency (TF-IDF)

Next, the other traditional methods are TF-IDF. This method is much better than the BoW because of its assigning weights to words based on their frequency in a document relative to their frequency across all documents. This means, that it facilitates reduce the importance of common word and highlights more unique words.

However, these methods also struggling with the semantic understanding and word order awareness. This limitation is not suitable for the tasks that required high understanding between the

semantically similar words. Besides, it also not able to capture polysemy which is the words with multiple meanings and synonymy which is the different words with similar meanings.

2.4 Overview of Transformer Model in NLP

The field of natural language processing has been introducing novel architecture that effectively captures long range dependencies in textual data. This transformer models have revolutionized NLP. For instances, unlike the traditional method such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), the transformer model has the effective mechanism that can know the importance of each word in a sentence relative to others. According to Sun (2023), The transition from conventional sequential models to attention-driven architectures has resolved the challenges that RNNs, LSTMs, and CNNs faced in managing intricate dependencies and lengthy text sequences.

2.4.1 Convolutional Neural Network (CNN)

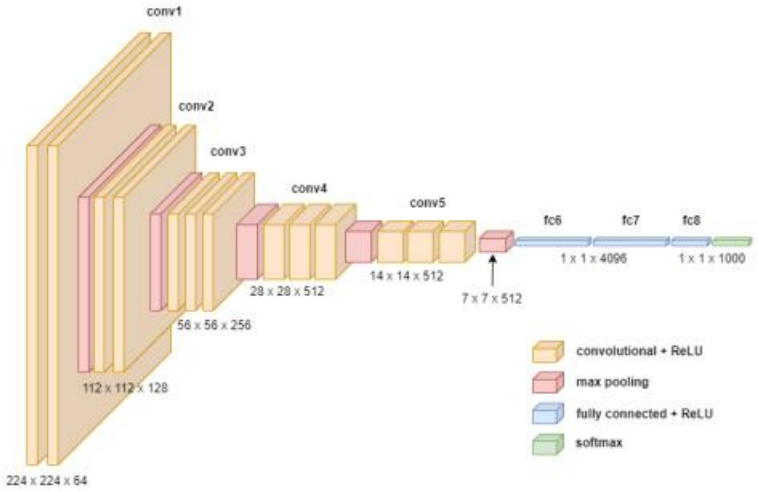

Figure 1: *Architectures of traditional neural networks used in NLP: (a) Convolutional Neural Network (CNN). Source: Adapted from Sun (2023).*

Figure 1 shows the Convolutional Neural Networks (CNNs). CNNs were originally designed for image recognition tasks but have been adapted for natural language processing. According to Iwasaki et al. (2018), CNNs trained on image data to be repurposed to text task as it has been integrated into NLP through transfer learning. For instances, as shown in Figure 1, a CNN architecture consists of multiple convolutional layers which are Conv1, Conv2, etc. This means the layers apply filters to input data to detect local features such as edges or patterns in images, in text, local phrase or n-gram patterns. Moreover, these layers are followed by pooling layers that reduce the dimensionality of the data. It summarizing the important features while maintaining spatial invariance. The Rectified Linear Unit (ReLU) activation function introduces non-linearity. It then enabling the network to learn complex mappings. However, CNNs process information

within fixed receptive field. This means, despite their strength in capturing localized features, it still has limits in their ability to capture long-range dependencies or the full context in sentences or documents. Therefore, according to Sun (2023), it becomes major limitation when dealing with complex and lengthy texts.

2.4.2 Recurrent Neural Network (RNN)

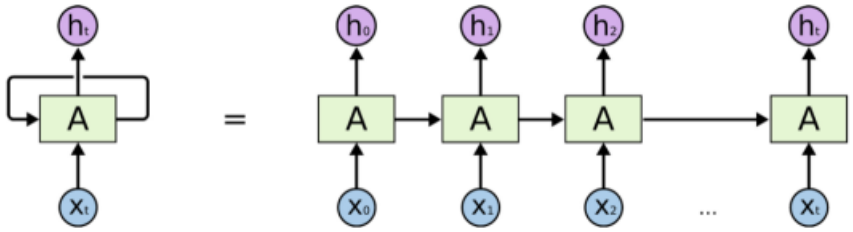

Figure 2: Architectures of traditional neural networks used in NLP: (b) Recurrent Neural Network (RNN). Source: Adapted from Sun (2023).

Next, figure 2 is an architecture of Recurrent Neural Network (RNNs) which is adapted from Sun (2023). This figure shows that RNNs are designed to handle sequential data by maintaining a hidden state (ht). This hidden state carries information from previous time steps. Moreover, the network process input which is xt step by step that allowing the model to sequence varying lengths such as sentences or paragraphs. However, RNNs have significance vanishing gradient problem, this leads for them to hard understand the dependencies across long sequences. This is critical to understand the complexity of the language. For instances, Graves (2012), stated that the influence the influence of earlier inputs diminishes rapidly, making it difficult for the network to learn from long sequences.

2.4.3 Long Short-Term Memory (LTSM)

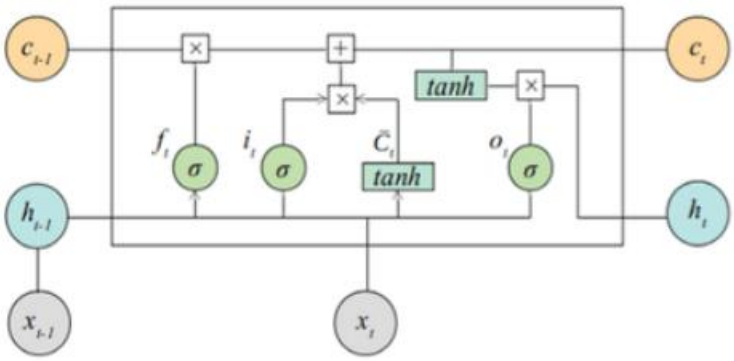

Figure 3: Architectures of traditional neural networks used in NLP: (c) Long Short-Term Memory (LTSM). Source: Adapted from Sun (2023).

Figure 3 indicated the Long Short-Term Memory (LSTM) networks. This architecture is designed to overcome the issues arise in RNNs. For instances, according to Zhang & Woodland (2018), LSTMs introduce memory cells and gating mechanisms that help retain information over longer periods, effectively addressing the vanishing gradient issue. LTSMs introduce special gating mechanism which are the input gate, forget gate and output gate. This helps to regulate the flow of information. Sun (2023), stated that LTSMs distinct from traditional RNNs because of its output mechanism. Besides, the gates use sigmoid functions that output values between 0 and 1 to decide how much old information to keep and how much new information to add to the cell state before passing it to the next step. This controlled flow of information makes LSTM well-suited for tasks involving long-term dependencies and memory retention (Sun, 2023). However, LSTMs still process sequences stepwise which leads to limiting parallel training. Hence, LTSMs may struggle with very long and hierarchically complex legal documents.

2.4.4 Transformer model

In contrast, the transformer models can eliminate recurrence and convolution by relying solely on the self-attention mechanism. Furthermore, the transformer architectural shift makes it efficient for processing of sequential data without the limitations of traditional recurrent neural networks (RNNs) or convolutional neural networks (CNNs).

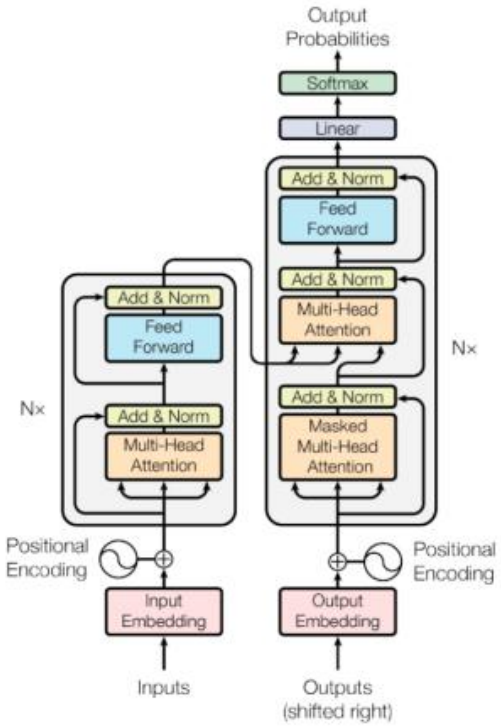

Figure 4: Transformer model. Source: Adapted from Sun (2023)

Figure 3 shows the transformer model. The Transformer model is a deep learning structure that is employed in natural language processing and various sequence-to-sequence tasks (Sun, 2023). The transformer as illustrate in the figure consist of an encoder and a decoder. This built from stacked layers of multi-head self-attention and feed-forward neural networks. The innovation of the Transformer lies in its extensive incorporation of the attention mechanism into neural network models, and it discarding the conventional LSTM and RNN architectures. As mentioned in "Attention Is All You Need," self-attention involves each word in the input sequence focusing on each other, and their separate contributions are combined to generate an output representation. The model is able to capture complex contextual information by weighing the importance of each word in the sequence relative to all other words, a process that is enabled by this mechanism. For instances, Islam et al. (2024), explained that the transformer model is capable of assessing the significance of each element in the sentences by capturing the relationships between all elements in a sequence by using a self-attention mechanism. Hence, the previous limitation mentioned in convolutional and recurrent models can be resolves.

Model	Description	Architecture Highlights	Strengths	Limitations	References
CNN	Originally designed for image recognition; adapted to NLP through transfer learning	Multiple convolutional layers (Conv1, Conv2, etc.) apply filters to detect local features; pooling layers reduce dimensionality; ReLU activation introduces non-linearity	Captures local phrase or n-gram patterns effectively	Limited receptive field restricts ability to capture long-range dependencies and overall context	Iwasaki et al., 2018; Sun, 2023
RNN	Designed for sequential data; maintains hidden state carrying past info	Processes input sequentially (xt), maintaining hidden state (ht); unfolds over time steps	Can handle sequences of varying length; captures short-term dependencies	Suffers from vanishing gradient problem; struggles with long dependencies; sequential processing limits parallelization	Graves, 2012; Sun, 2023
LSTM	Designed to	Introduces	Effectively	Still sequential;	Zhang &

	overcome RNN issues using gating mechanisms to regulate information flow	memory cell and gates (input, forget, output); uses sigmoid functions to control flow	retains long-term dependencies; addresses vanishing gradient issue	limits parallel training; may struggle with very long and hierarchically complex texts	Woodland, 2018; Sun, 2023
Transformer	Eliminates recurrence and convolution; relies on self-attention	Encoder-decoder architecture with stacked layers of multi-head self-attention and feed-forward networks; positional encoding	Captures global context efficiently; supports parallel computation; models complex dependencies	Complexity and high resource requirement; needs large datasets for effective training	Vaswani et al., 2017; Islam et al., 2024; Sun, 2023

Table 1: Summary of CNN, RNN, LTSM and Transformer model

2.5 Transformer Models in the Legal Domain

Therefore, transformer has showed superior performance in the field of natural language processing (NLP) because of it has the superior ability to understand the complex of legal texts. Transformer models are useful in applications such as sentiment analysis, spam detection, and information retrieval because they can efficiently manage contextual information as mentioned by (Tingare & Jangid, 2024)

Legal cases are typically lengthy documents with a complex structure. Therefore, it difficult for the traditional keyword retrieval model's ability to accurately represent the semantic relationship between the query and the candidate cases. For instances, legal documents frequently involve complex legal structure, case statutes and precedents which are difficult to model to comprehend the documents. Moreover, the needs for the retrieval model that can efficiently handle large-scale datasets is needed as the judgment cases in Malaysia is increasing yearly. For instances, retrieval models must be capable of managing large-scale datasets while maintaining accuracy, as legal databases are vast (Yang et al., 2023).

2.5.1 Legal BERT

Therefore, researchers have created Legal BERT, a BERT-based model that has been optimized for legal texts. To be highlighted, Legal-Bert is among effective model for Legal task and is a variant of BERT. For instances, Legal BERT has been pretrained on large corpora such as court decisions and statutes and primarily on a large corpus of United States and Europe Union legal text (Chalkidis et al., 2020). The studies indicated that the use of Legal BERT can overcome the scalability challenges of legal case retrieval.

For instance, Chalkidis et al. (2020) stated that the application of Legal-BERT has been indicated that it helps to improve performance in legal text processing tasks. Besides, Althammer et al. (2021) and Wang et al. (2024), stated that Legal BERT helpful in retrieving relevant cases from extensive legal databases by implementing efficient retrieval mechanisms, such as dense passage retrieval. However, it faces substantial challenges, including the necessity for extensive pre-training and domain-specific ambiguities.

2.5.2 Legal-longformer

Next, Legal-longformer has been introduced and utilized to manage longer legal documents. According to Lee & Lee (2023), claims that Legal-longformer can be mixed with the LTSM to manage the complex longer legal documents. Furthermore, Lee & Lee (2023) also clarified in legal documents, this approach successfully depicts local as well as worldwide dependencies. Therefore, this can be achieved by combining Longformer with LSTM. Moreover, it facilitates the retrieval of similarity in legal case documents to increase the efficiency of legal research. Besides, some studies show that this model performed well in handling long legal documents. Then, Hoang et al. (2023) have indicated that legal-longformer models have superior performance in the task related with lengthy documents. For instances, the studies found that this model achieves high rankings in tasks such as predicting court judgements. However, the effectiveness of this model in diverse legal contexts may be less effective by the intricacy of legal language and the need for more annotated datasets.

2.5.3 Sentence-BERT (SBERT)

Furthermore, SBERT is based on BERT architecture used to create semantically meaningful sentence embeddings ideal for applications such as semantic similarity. This model helps to reduce the computational burden associated with traditional BERT models. This can be achieved by employing the model with a Siamese and triplet network architecture. For instances, Reimers & Gurevych (2019) found that SBERT efficient for tasks such as semantic textual similarity and clustering. Furthermore, SBERT help to reduce the time required to identify comparable sentence pairs from approximately 65 hours to approximately 5 seconds (Reimers & Gurevych, 2019). In addition, it performed better than the other state-of-the-art sentence embedding methods while sustaining accuracy and efficiency. However, SBERT's practicality may be restricted in situations where labelled datasets are limited because of it’s dependent on high-quality data for training (Zhang et al., 2020).

2.5.4 Legal XLNet

Finally, Legal XLNet also has been used to operate within the legal domain. For example, this model helps to improve the efficiency of language understanding. Legal XLNet is a valuable tool than can help legal research for the legal professionals and researchers because of its ability to manage complex legal language and the documents that contain longer paragraphs. For instances, Legal XLNet has been employed to abstractly summarize legal documents, which shows its ability to effectively condense intricate

legal texts (Kale & Deshmukh, 2024). However, the computational demands and memory requirements of XLNet present obstacles in environments that are resource-limited. Therefore, it is necessary to optimize transformer models for specific legal applications.

Model	Problems / Issues	References	Algorithms / Policies / Strategies / Frameworks	Performance	Parameters / Notes	Simulation / Experimental Tools	Comparison	Results	Advantages	Disadvantage / Limitations
Legal BERT	Requires extensive pre-training; domain-specific ambiguities	Althammer et al., 2021; Wang et al., 2024	Dense Passage Retrieval; Domain-adapted BERT pretrained on legal corpora	Efficient retrieval of relevant cases from large legal databases	Transformer-based; pretrained on court decisions, statutes	Retrieval experiments on legal case databases	Compared to general BERT models, performs better on legal texts	Rapid retrieval of pertinent legal cases	Scalability in legal case retrieval; better domain adaptation	High computational cost; challenges with legal domain ambiguities
Legal Longformer	Complexity of legal language; limited labeled datasets	Lee & Lee, 2023; Hoang et al., 2023	Combines Longformer with LSTM for handling long documents	Excels at capturing local and global dependencies in lengthy legal texts	Uses sparse attention and LSTM layers	Legal judgment prediction; similarity retrieval tasks	Outperforms models like BERT in long-document handling	Top rankings in court judgment prediction and similarity tasks	Effective handling of long documents; captures hierarchical dependencies	Needs more labeled data; complexity challenges
Sentence-BERT (SBERT)	Dependence on large, high-quality	Reimers & Gurevych, 2019;	Siamese and triplet network architect	Highly efficient semantic textual similarity	Reduces computational overhead of	Semantic similarity and clustering benchmar	Outperforms other state-of-the-art sentence	Reduces time for comparable sentence	Efficient and accurate; suitable for	Performance limited by availabilit

	labeled datasets	Zhang et al., 2020	ures; Sentence embeddings	and clustering	traditional BERT	ks	embedding methods	pair identification from ~65 hours to ~5 seconds	sentence-level similarity tasks	y and quality of labeled datasets
Legal XLNet	High computational and memory requirements; resource-intensive for some environments	Kale & Deshmukh, 2024	Permutation-based pretraining; Domain fine-tuned XLNet	Effective summarization of complex legal documents	Transformer with autoregressive capabilities	Summarization tasks on legal document datasets	Compared favorably for summarization but resource-heavy	Capable of abstract summarization of legal documents	Strong comprehension of complex legal language; suitable for lengthy documents	High computational cost and memory demands; less efficient in low-resource environments

Table 2: Comparison of transformer-based models adapted for the legal domain

2.6 Hybrid Approaches

Expanding upon the transformer architectures previously discussed, hybrid approaches often integrate transformer model such as BERT with traditional keyword-based methods such as TF-IDF. Consequently, this combination is designed to improve the efficiency of conventional retrieval methods and to advance the comprehension of transformers like BERT. For instances, Wehnert et al. (2021) combined TF-IDF with BERT. The studies demonstrated that it enhanced the performance of statute law retrieval, underscoring the efficacy of ensemble methods in legal contexts. For example, TF-IDF or BM25 is used to perform an initial filtering of candidate documents and Bert-based models will re-rank these candidates by evaluating the semantic similarity. Nevertheless, this approach is limited by the bias associated with keyword-based filtering and necessitates additional and meticulous parameter refining. However, these methods have been extensively employed in the retrieval of legal information via extensive databases.

2.7 Fine Tuning of Transformer Models

Fine-tuning involves further training the transformer model to the legal specific knowledges or task with labeled data. For instances, in Legal NLP, this process is crucial to make the model able to adapt to the parameters to better understand domain-specific terminology, syntax and semantics. For instance, Su et al. (2023) introduced Caseformer, a pre-training framework that enables models to acquire legal knowledge without the need for human-annotated data. These can happen because of the employment of the unsupervised learning tasks to capture the complex language and document structures of legal cases. Therefore, it demonstrated that it able to achieve state-of-the arts results in both zero-shot and full-data fine-tuning settings. However, the constraints may restrict their applicability in a global context, as there are still challenges to guaranteeing that the models can generalize across a variety of legal systems and languages.

2.8 Evaluation Metrics

Subsequently, it is crucial to obtain precise results following the refinement of transformer models. Consequently, the assessment of semantic similarity models can assist in quantifying the extent to which they accurately represent genuine semantic closeness. Therefore, it requires certain metrics that need to be use to evaluates it such as cosine similarity, precision and recall, F1 score and accuracy score. According to Thenmozhi et al. (2017), cosine similarity facilitates the retrieval of older cases that are contextually similar to a current case, thereby guaranteeing consistent legal reasoning. For instances, cosine similarity calculates the angle between two sentence embeddings in vector space between -1 and 1. This means a smaller angle that closer to 1 in value indicates the greater similarity.

The formula for cosine similarity is as below:

$$Equation\ 1:\ Cosine\ Similarity = \frac{A \cdot B}{||A|| \times ||B||}$$

Besides, accuracy is another frequently used metric that quantifies the percentage of sentence pairs that are correctly predicted. For instance, the differentiation between similar and dissimilar based on a predetermined threshold. According to Owusu-Adjei et al. (2023), accuracy has been widely used in assessing the effectiveness of predictive algorithms. However, this metrics give limited insights if they are solely relying on accuracy which is not recommendable as it can obscure the true performance of models because it may not be enough to see the complexities of the data or the model's capabilities.

The formula for accuracy is as below:

$$Equation\ 2:\ Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Additionally, false positives can mislead legal reasoning. Therefore, the precision metric is used to measure retrieved cases that are actually relevant. Ebietomere & Ekuobase (2019), found that a semantic retrieval system for case law exhibited a precision of 94% and an F-measure of 84%, indicating a high level of efficacy in the retrieval of relevant cases.

The formula for precision is as below:

$$Equation\ 3:\ Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

Moreover, recall metrics measure the proportion of relevant cases that were successfully retrieved. Therefore, high recall means that critical precedents are not missed. According to Sivaranjani & Jayabharathy (2022), high recall is important in legal case retrieval because of multifaceted nature of legal queries, which often require comprehensive retrieval of similar cases.

The formula for recall is as below:

$$Equation\ 4:\ Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

Finally, The F1 score is a critical metric for evaluating legal case retrieval systems, as it balances precision and recall. Therefore, it helps to provide the insights of model's effectiveness. For instances, it is beneficial in datasets with imbalanced classes and maintains an equilibrium between the two metrics which are the precision and recall. According to Ye & Li (2024), F1 score is important for comprehend the trade-off between precision and recall in legal contexts.

The formula for F1 score is as below:

$$Equation\ 5:\ F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

2.9 Research Gaps

Consequently, the explanation of the progression in legal NLP through transformer-based models is provided in this chapter. Nevertheless, there are still numerous research openings, particularly in the context of Malaysian legal precedents. For example, the absence of domain-specific fine-tuning in Malaysian legal corpora. When the transformer models, such as Legal BERT, have been pre-trained on U.S. and Western legal documents, this is evident. Therefore, the documents and the structure of linguistic and jurisdictional components differ from those of Malaysian legal cases, which presents a challenge.

Additionally, there is a scarcity of annotated legal datasets that pertain to the Malaysian context. Additionally, the majority of legal retrieval benchmarks are derived from the English and Western legal systems. Consequently, the model training, fine-tuning, and evaluation in this project are significantly enhanced by the necessity of well-annotated datasets that accurately reflect the distinctive linguistics and legal characteristics of Malaysian legal judgement.

Furthermore, despite the transformer model's attainment of the most advanced performance in semantic tasks, there is still a lack of exploration of its integration, particularly in Malaysia. For example, computational efficiency remains an issue that has yet to be resolved. This is essential for the practical implementation of these systems.

Subsequently, the majority of the existing research concentrated on the similarity between sentences or paragraphs. This frequently disregards the necessity of document-level semantic analysis. Consequently, it is crucial to comprehend the entirety of legal arguments in order to achieve a precise semantic outcome. Additionally, the transformer model's application to lengthy legal documents yields a variety of results in the current models. This suggested that the models should be optimized for lengthy legal documents, such as Legal-Longformer or hierarchical attention frameworks.

Finally, the hybrid methods that combine transformer-based and traditional methods have been extensively investigated. Nevertheless, their implementation in local legal information systems, such as the Malaysian court archives or law libraries, has not been adequately evaluated.

2.10 Conclusion

This chapter includes a literature review of the ongoing project regarding semantic similarity and legal precedent retrieval using transformer-based models. This chapter presents an analysis of the similarities and differences between various models, algorithms, and evaluation metrics. Apart from that, this chapter also provides an in-depth discussion regarding transformer models adapted for the Malaysian legal domain. The next chapter will discuss the research methodology and the outlines of the main strategies used in this project.

Chapter 3: Research Methodology

3.0 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.

3.1 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:

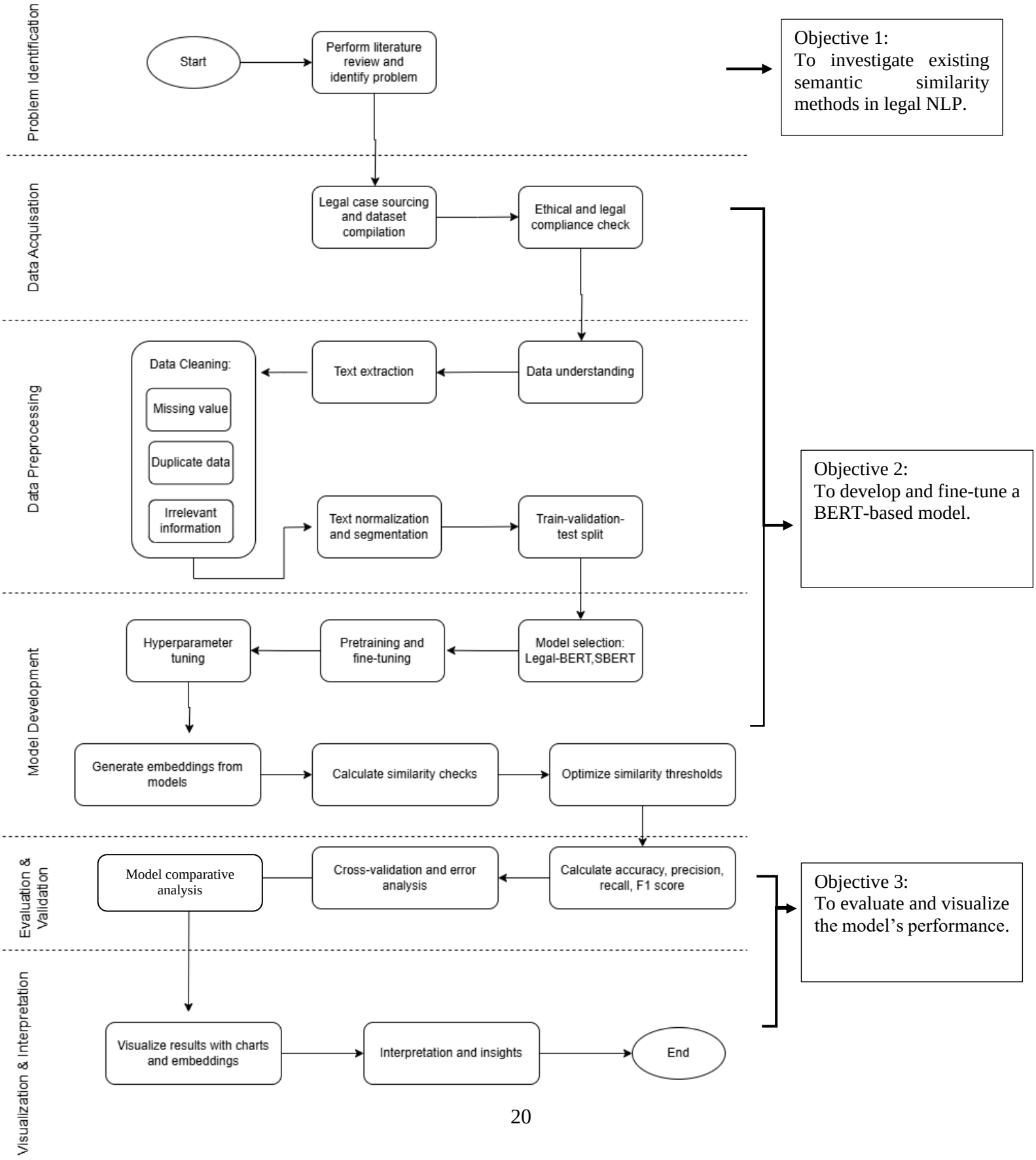

Figure 5: Flowchart of Research Framework For BERT-based Semantic Similarity of Malaysian Legal Precedents

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 Kaggle. 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.

3.2 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.

3.3 Phase 2: Data Acquisition

The legal dataset was obtained from Kaggle. It consists of a high-quality dataset of Malaysian legal documents to train and evaluate semantic similarity models. This data contains of 3000 sentences of legal cases across various domains. 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:

ID	Case 1 ID	Case 2 ID	Case 1 Text	Case 2 Text	Case 1 Domain	Case 2 Domain	True Label	Is Ambiguous	Complexity
1	ML-2024-2382	KL-2019-4065	Dispute arose over the interpretation of clauses in a partnership agreement.	Partnership dissolution case where partners disagreed on asset distribution.	Contract	Commercial	0	False	Expert
2	KL-2019-8604	PG-2020-7039	Dispute over ownership of land in Sandakan meant for agricultural use.	Strata property case involving common property usage among residents.	Property	Property	1	False	Expert
3	ML-2022-1625	KL-2019-8846	Adoption petition filed by {adoptive_parents} under the Adoption Act.	Divorce proceedings under the Law Reform (Marriage and Divorce) Act.	Family	Family	1	False	Expert
4	PG-2021-2570	ML-2020-5890	Criminal intimidation case where the accused threatened to harm the victim.	Criminal breach of trust case involving misappropriation of company funds.	Criminal	Criminal	1	False	Expert
5	PG-2023-9137	PG-2018-3088	Construction contract dispute involving {contractor_name} and {client_name}.	Religious law conflict with civil law regarding property inheritance.	Contract	Constitutional	0	False	Expert

Table 3: Example of Legal Cases Downloaded

3.4 Phase 3: Data Preparation

Stage	Task	Description
Stage 1: Data Preprocessing	**Text Extraction**	Extract legal case text from downloaded CSV file.
	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 3: Data Post-Processing	**Train-Test Split**	Split the data into training, validation, and testing datasets.

Table 4: 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.

3.5 Phase 4: Model Development

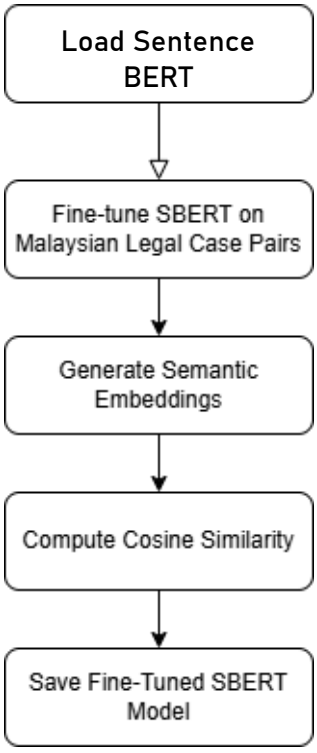

Figure 6: Pipeline of Model Development phase

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 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, SBERT architecture 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:

- **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.

3.5.1 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.

3.5.2 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

3.6 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).

3.6.1 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.

3.6.2 Comparative analysis

Therefore, the performance of the model also will be compared with the baseline traditional methods such BoW and TF-IDF. These models serve as foundational benchmark in text representation. Besides, these two traditional methods also commonly used in semantic similarity tasks due to their simplicity.

This comparison will be in terms of standard classification metrics such as accuracy, precision, recall and F1-score and the confusion matrix. Therefore, it will provide the meaningful insights of how well each model perform the task to distinguish between similar and dissimilar sentences.

3.6.3 Visualization and Reporting

In the last phase, the quantitative evaluation was visualized using techniques such as bar charts, other graphical charts, and a confusion matrix. This method of visualization projects the high-dimensional embeddings into a two-dimensional space. Therefore, 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.

3.7 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.

Chapter 4: Initial Findings

4.0 Introduction

This chapter introduced the initial results for BERT-based Semantic Similarity of Malaysia Legal Precedents study. It focuses on the presenting the comprehensive results from the development and evaluation of a BERT-based semantic similarity model for Malaysian legal texts. The objectives that had been discussed in the previous chapters are to investigate existing semantic similarity methods in legal NLP, develop and fine-tune a BERT-based model specifically for Malaysian legal contexts, and evaluate the model's performance through rigorous testing and visualization.

The analysis in this chapter following the complete pipeline from the data acquisition through model evaluation. It demonstrates that each phase of the methodology contributed to achieving a robust legal text similarity classification system. Therefore, the result shows the high effectiveness of transformer-based methods for understanding the semantic relationship in legal documents.

4.1 Data Sourcing and Acquisition Phase

4.1.1 Data Acquisition and Characteristics

This project used a Malaysian legal precedents dataset sourced from Kaggle, that included the comprehensive legal cases pairs for semantic similarity analysis:

Main characteristics of the Dataset:

- The dataset consists of Malaysian legal caselaw from various court of Malaysia.
- The dataset is pairs and have true label that annotated to give information about similarity between two cases.
- Each sentences have their own legal domain that help to train the model better.

This project used a Malaysian legal precedents dataset sourced from Kaggle, that included the comprehensive legal cases pairs for semantic similarity analysis:

- **Original dataset size**: 3,000 rows × 12 columns
- **Data source**: Kaggle legal text corpus
- **Format**: CSV file
- **Records successfully loaded**: 3,000 legal document pairs

Column Name	Description & Distribution
id	A unique identifier for each sentence pair (0 to 2999). Sequentially ordered, suggesting organized collection.
case1_id & case2_id	Unique case identifiers (e.g., ML-2024-2382). These follow a consistent alphanumeric format indicating state and year.
case1_text & case2_text	Legal sentence pairs, typically 237–430 characters in length. Sentences are descriptive and reflect real-world legal contexts.
case1_domain & case2_domain	Legal domains (e.g., criminal, property, family, contract). Help contextualize the source of each sentence.
true label	Ground truth for semantic similarity: 1 = similar, 0 = dissimilar.
is ambiguous	Boolean (True/False) indicating whether the similarity is ambiguous. Mostly False, used to mark edge cases.

Table 5: Analysis of each column in the dataset

Figure 1 below shows the distribution of the true labels which is 0 label is the higher with 1498 and 1 label is 1359 samples.

Figure 7: Distribution of True Labels

In Figure 2, the distribution of Case1 and Case2 domains shows that the number of each domain class is not far from each other, with the property law domain being the highest and the criminal domain being the lowest. This suggests a moderately uniform coverage of legal areas.

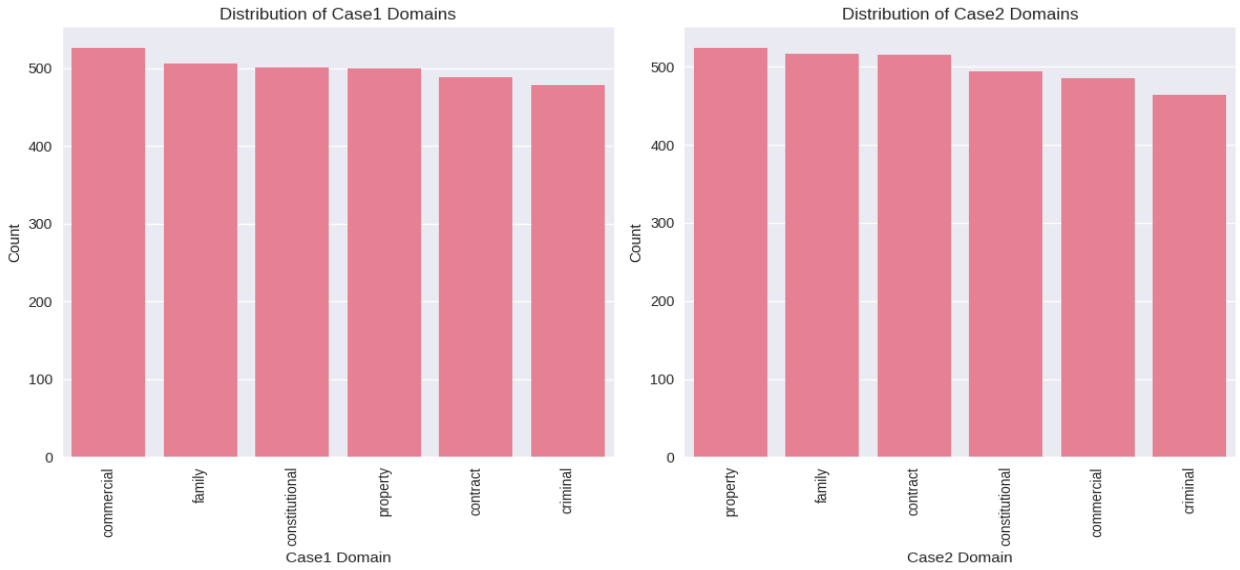

Figure 8: Distribution of Case1 and Case2 Domains

4.2 Data Cleaning and Preprocessing

Next, the process involved cleaning and preprocessing the dataset, where it went through a cleaning phase and the unnecessary columns were removed to gain better insights from the analysis

1. The legal text noise is cleaned while keeping useful punctuation, and filters out too-short or invalid entries.

```
def clean_text(text):
    """Clean legal text data"""
    if pd.isna(text) or text == '':
        return ""

    # Convert to string if not already
    text = str(text)

    # Remove extra whitespace
    text = re.sub(r'\s+', ' ', text).strip()

    # Keep legal punctuation and special characters
    # Only remove extreme special characters that might cause issues
    text = re.sub(r'[^\w\s\.,;:()\-\/&{}]', '', text)

    # Ensure minimum length
    if len(text) < 10:
        return ""

    return text
```

Figure 9: Cleaning process

2. The `standardize domain` function is used to clean and standardize legal domain names. First, the input is checked whether it missing or nan, and it return “unknown” if so. Then, it converts the input to lowercase and removes any extra spaces. The cleaned input is compared

with the predefined dictionary of know legal domains. Therefore, it returns the standardize domain if it found.

```python
def standardize_domain(domain):
    """Standardize legal domain names"""
    if pd.isna(domain):
        return "unknown"

    domain = str(domain).lower().strip()

    # Domain mapping for consistency
    domain_mapping = {
        'commercial': 'commercial',
        'contract': 'contract',
        'property': 'property',
        'family': 'family',
        'criminal': 'criminal',
        'constitutional': 'constitutional',
        'tort': 'tort',
        'employment': 'employment',
        'intellectual_property': 'intellectual_property',
        'banking': 'banking_finance',
        'finance': 'banking_finance',
        'insurance': 'insurance',
        'tax': 'tax',
        'administrative': 'administrative',
        'civil': 'civil',
        'company': 'company_corporate',
        'corporate': 'company_corporate',
    }

    # Try to match with known domains
    for key, value in domain_mapping.items():
        if key in domain:
            return value

    return domain.replace(' ', '_')
```

Figure 10: Standardizing process

3. After the cleaning and standardizing the dataset, the duplicates data is removed. This is to ensure that the dataset is in better condition for better result when the training model started. Firstly, each pair is normalized by sorting the two values and removing duplicates based on the reverse sorted appearing, (e.g., `(A, B)` and `(B, A)`). This ensures only one unique representation of each case pair remains in the dataset.

```
🔄 Removing semantic duplicates (reversed case pairs)...
   🔍 Found 229 rows that are part of duplicate pairs
   📄 Sample duplicate pairs:
     Pair 1 (ID: -6106279715210009412):
       Case1: Employment contract termination case where {employee} was dismissed without proper notice. The emplo...
       Case2: Sale and purchase agreement dispute regarding the {property_type} located at Kuala Lumpur. The purch...
     Pair 2 (ID: -1160084647589292402):
       Case1: Consumer protection claim under the Consumer Protection Act 1999. The consumer alleged {consumer_iss...
       Case2: Insurance claim dispute where the insurer denied liability for {claim_type}. The policy terms were i...
     Pair 3 (ID: 2491846763640360488):
       Case1: Religious law conflict with civil law regarding {religious_matter}. The court had to determine juris...
       Case2: Judicial review application against {authority} decision regarding {administrative_action}. The appl...
     Pair 4 (ID: -2562695766493108273):
       Case1: Federal-state jurisdiction dispute over {matter}. The case involved interpretation of the Ninth Sche...
       Case2: Federal-state jurisdiction dispute over {matter}. The case involved interpretation of the Ninth Sche...
✓ Removed semantic duplicates: 143 rows
   📊 Remaining rows: 2,857
```

Figure 11: Result for Removing Semantic Duplicates

4. After the above process, the process of balancing the data was begun. This process was done to ensure that the data is being proper balance for better result for the model training. Below are the graphs shows the imbalance class distribution before the balancing process. This will affect the result for the model training later. Therefore, this issue is addressed with proper balancing methods.

4.2.1 Initial Class Distribution (Before Balancing)

The dataset exhibited a moderate class imbalance requiring correction:

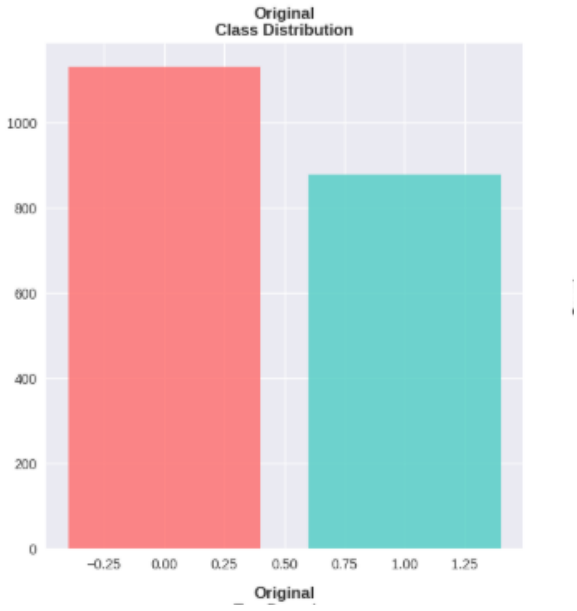

Figure 12: Original Class Distribution

Class Label	Sample Count	Percentage	Description
Label 0 (Dissimilar)	1,498	52.43%	Non-similar legal precedent pairs
Label 1 (Similar)	1,359	42.57%	Similar legal precedent pairs
Total	**2,857**	**100.00%**	**Complete dataset**

Table 6: Original Distribution of Class Label

4.2.2 Data Balancing

These steps have 3 methods which are:

- random oversampling
- random under sampling
- combined approach

Firstly, the random oversampling process is done by balancing the class distribution by duplicating samples from the minority class. Therefore, it used "RandomOverSampler" to generate a dataset with equal class representation. Hence, increasing model exposure to underrepresented cases without losing any original data. Below is the result after using the random oversampling technique.

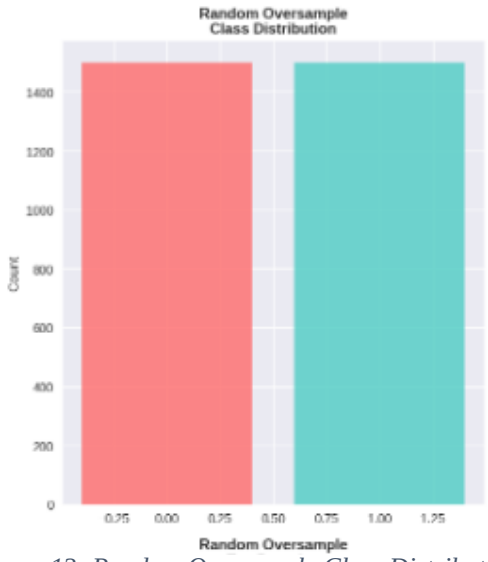

Figure 13: Random Oversample Class Distribution

Class Label	Sample Count	Percentage	Description
Label 0 (Dissimilar)	1,498	50%	Non-similar legal precedent pairs
Label 1 (Similar)	1,498	50%	Similar legal precedent pairs
Total	**2,996**	**100.00%**	**Complete dataset**

Table 7: Random Oversample Class Distribution

Besides, the random under sampling is another method that balancing the class distribution by duplicating samples from the majority class. Therefore, it used “RandomOverSampler” to generate a dataset with equal class representation. Therefore, the it reduced the dataset size while achieving class balance. This will result to speed up training but risks losing potentially valuable data. Below is the result after using the random under sampling technique.

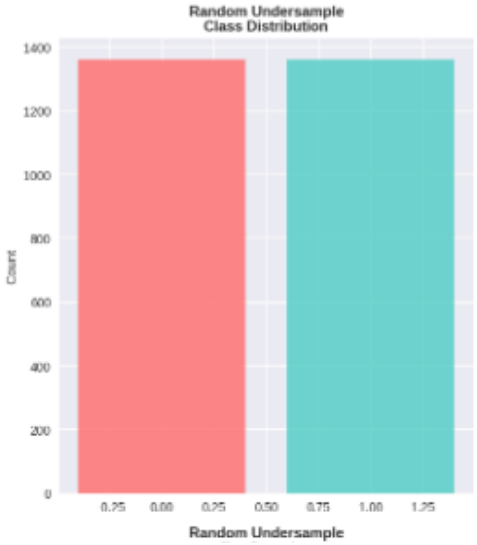

Figure 14: Random Undersample Class Distribution

Class Label	Sample Count	Percentage	Description
Label 0 (Dissimilar)	1,359	50%	Non-similar legal precedent pairs
Label 1 (Similar)	1,359	40%	Similar legal precedent pairs
Total	**2,718**	**100.00%**	**Complete dataset**

Table 8: Random Undersample Class Distribution

The last methods are combined approach. This technique is technically combining the combined random under sampling of the majority class with oversampling of the minority class. For instances, the majority class was reduced to twice the size of the minority, while the minority was increased to 75% of the majority size. This hybrid method is used to support the useful majority examples while boosting minority representation. Below is the result after using the combined approach technique.

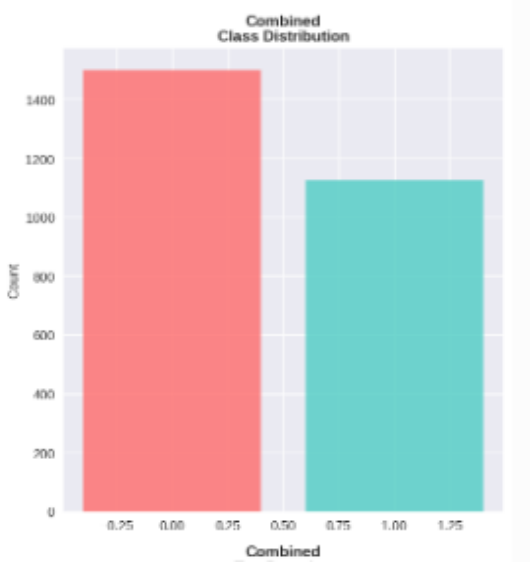

Figure 15: Combined Class Distribution

Class Label	Sample Count	Percentage	Description
Label 0 (Dissimilar)	1,480	50%	Non-similar legal precedent pairs
Label 1 (Similar)	1,211	40%	Similar legal precedent pairs
Total	**2,691**	**100.00%**	**Complete dataset**

Table 9: Combined Approach of Class Distribution

4.2.3 Method Ranking

Therefore, the 3 methods were compared and evaluated using a custom quality score metric. Based on this, both Random Oversampling and Random Under sampling achieved the highest performance (1.000), while the Combined Approach followed with a slightly lower score (0.850). The score shows that the Random Oversampling and Random Under sampling achieved excellent score. Hence, it was decided that the **Random Oversampling** method was selected as the preferred balancing method due to its top performance (quality score: 1.000). Besides, it proven to enhance the minority class without discarding any original data. This preserves the full diversity of the dataset while addressing class imbalance effectively.

Selected Class Distribution (After Balancing using Random Oversampling):

Class Label	Sample Count	Percentage	Description
Label 0 (Dissimilar)	1,498	50%	Non-similar legal precedent pairs
Label 1 (Similar)	1,498	40%	Similar legal precedent pairs
Total	**2,996**	**100.00%**	**Complete dataset**

Table 10: Selected class distribution (Random Oversample)

4.3 EDA

Exploratory Detail Analysis is conducted after the data cleaning and processing done. This is to understand the dataset for more accurate used in model development. For instance, EDA is the process of analyzing and summarizing the dataset to uncover the main characteristics. Therefore, it will be represented with visualization for more meaningful insight. Furthermore, it helps in understanding the data structure, spotting patterns and detecting the anomalies. Therefore, the quality of overall dataset will be focused on before the application of machine learning.

Therefore, the dataset that has been cleaned and processed earlier will be analyzed using EDA. The analysis is useful for further actions, specifically for the model development phase. Below is the visualization of the EDA that has been conducted on the cleaned dataset and after the balancing process.

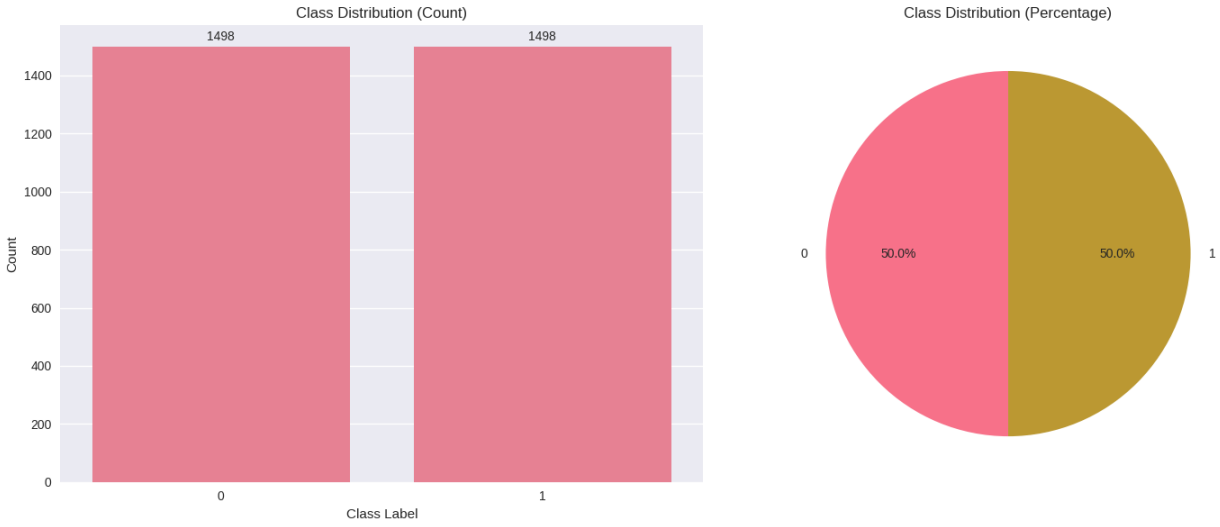

Figure 16: Finalized Class Distribution

The analysis of the class distribution count has been conducted, and the results show that the dataset is balanced, with 1,498 samples evenly distributed between the "similar" and "not similar" classes. Each class represents 50% of the data. This balance is important to avoid classification bias in machine learning models and to ensure more accurate results.

The figure 10 below shows the distribution of text lengths after the balancing process. It illustrates the distribution of text lengths for Case 1 and Case 2 after applying random oversampling. For instance, the lengths of both texts are relatively similar, with a peak around 270–280 characters. Additionally, the maximum length exceeds 400 characters, but only for a few samples.

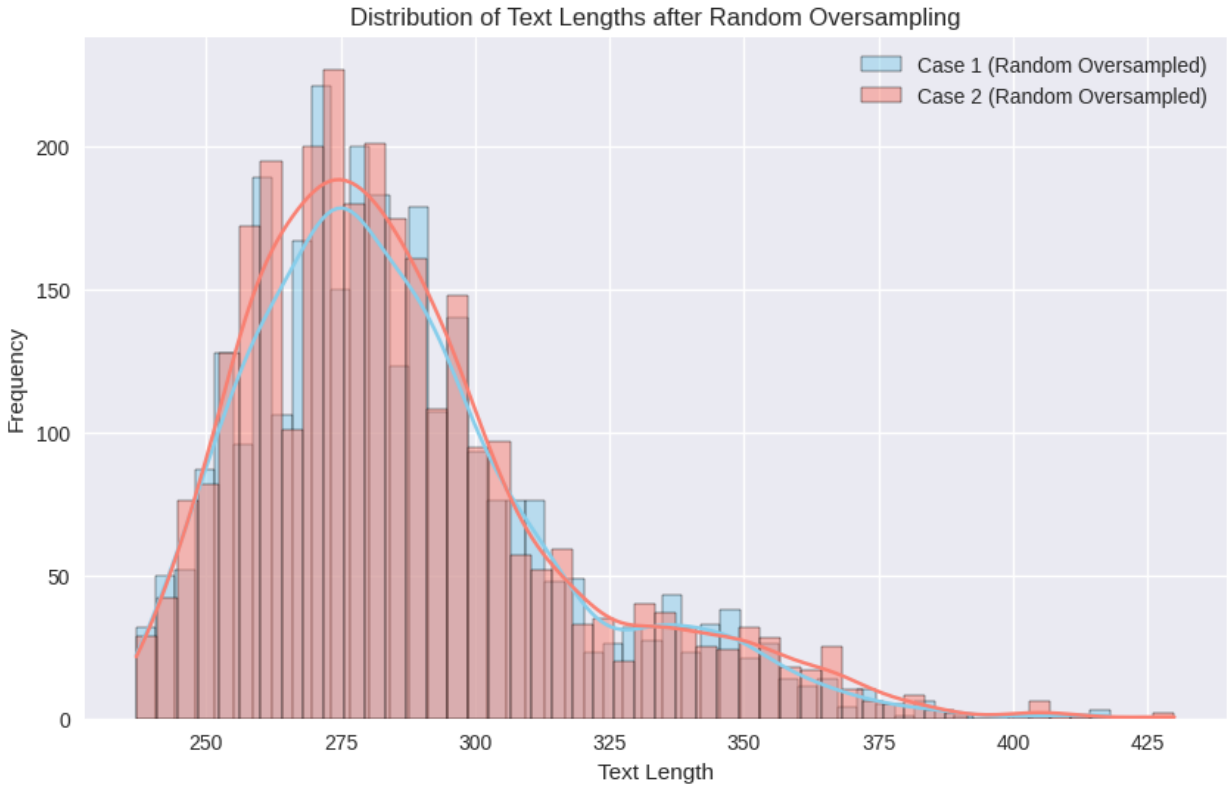

Figure 17: Distribution of Text Lengths After Random Oversampling

4.4 Model Development

The finalized dataset of 2,996 entries, which had been cleaned and preprocessed, was first loaded to be split into training, validation, and test sets. The proportion of each set were 2396 (80%) for training, and 300 (10%) each for validation and testing. Therefore, it will then be wrapped inside “InputExample” objects with the corresponding similarity labels 0 or 1. The DataLoader was set with a batch size of 16 that will be used during training and evaluation.

Next, the selected model which is a pre-trained Sentence-BERT model was fine-tuned using “CosineSimilarityLoss”. This tool is suitable for the semantic similarity tasks. Then the “EmbeddingSimilarityEvaluator” was employed on the validation set. This is to ensure the performance of the model being critically evaluated and the best performing model was saved based on the validation scores.

Moreover, the hyperparameter tuning was set to the model with three epochs. This epoch was evaluated with intermediate evaluations and it performed every 500 steps. Therefore, the best checkpoints which is the highest score will be saved as the final model.

Furthermore, the “EmbeddingSimilarityEvaluator” also conducted the evaluation on the test. For instances, the best model showed a Pearson correlation score of 0.6648 and Spearman correlation of 0.53304. The final model scored 0.6641 for Pearson correlation and 0.5362 for Spearman correlation. Therefore, the final model was selected based on the Pearson correlation as it is the primary metric for the semantic similarity tasks. Hence, the final model with best-performing checkpoint saved as Bert-sbert-legal that will be use for the evaluation’s visualization.

Hence, it concluded that the model demonstrated a good level of semantic correlation with a Pearson score of 0.665. It shows that the model is able to effectively capturing the semantic similarity between the legal sentence pairs. This making it suitable for the legal NLP tasks and application.

4.5 Model Evaluation and Visualization

After the training, the final evaluation was conducted on a test set that had been split earlier. The test set consist of 300 sentence pairs and was processed using the fine-tuned model. Therefore, the predictions were obtained by calculating the cosine similarity between embedded representations of each sentence pair. For instances, the predicted score was ranged from -.0.187 to 0.996. This range will be compared with the true label range from 0 to 1.

4.5.1 Regression Metrics

To assess the model's capability in capturing semantic similarity as a continuous measure, several regression metrics were computed:

- Pearson Correlation: 0.6641
- Spearman Correlation: 0.5362
- R^2 Score: 0.4278
- Root Mean Squared Error (RMSE): 0.3782
- Mean Absolute Error (MAE): 0.2670

Above result shows that the model demonstrates a moderate to good correlation. For instances, the model prediction was a reasonably low error, which means it is reliable for legal NLP tasks. Besides, the Pearson correlation of 0.6641 suggests that the model is able to know and capture the semantically similar legal sentences.

4.5.2 Classification Metrics

Then, it was evaluated as a binary classification problem. For instance, the sentence pairs such as similar or dissimilar were classified based on the thresholds. Below are the three thresholds that were tested:

Threshold	Accuracy	Precision	Recall	F1 Score
0.5	0.8267	0.7426	1.0000	0.8523
0.6	0.8100	0.7360	0.9667	0.8357
0.7	0.7600	0.7321	0.8200	0.7736

Table 11: Result of Classification Metrics

The result shows that the 0.5 thresholds showed the highest F1 score of 0.8523. This provides that the balance trade-off between the precision and recall. Besides, the recall of 1.0000 indicated that the model successfully identified all actual similar pairs.

4.5.3 Optimal Threshold Selection

Therefore, to obtain the optimal thresholds that effectively classified the pairs, a fine-grained search from 0.1 to 0.95 was conducted. This involved in steps of 0.05 to get the best result. Hence, it was found that the optimal thresholds were 0.200 which yielded the highest F1 score of **0.8523**.

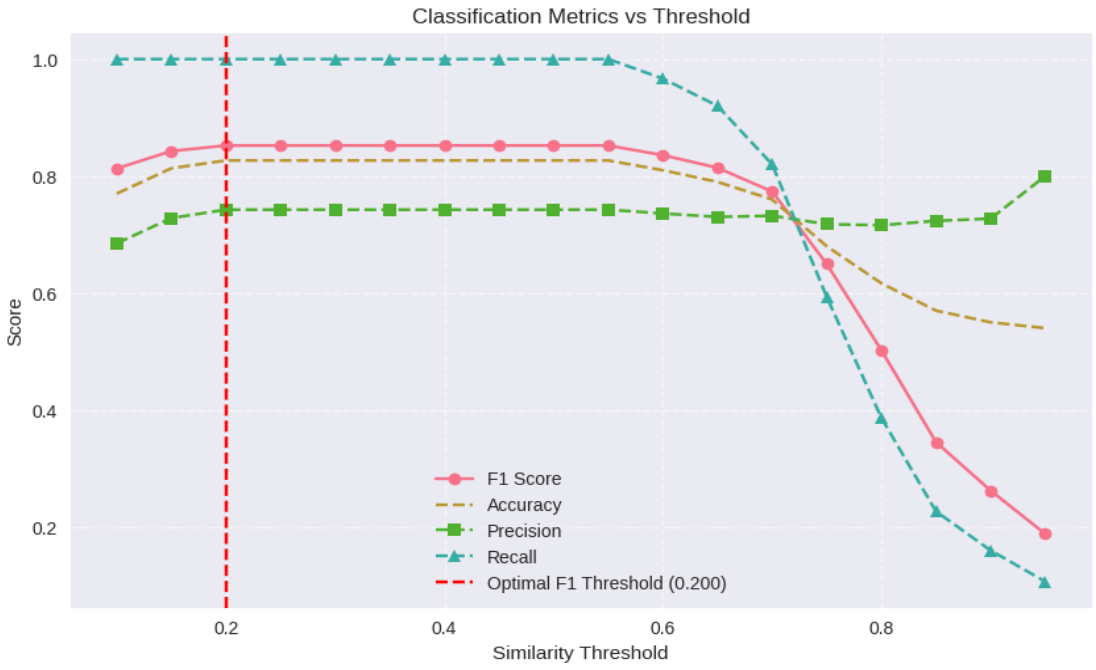

Figure 18: Graph of Classification Between Metrics and Threshold

Class	Precision	Recall	F1-Score	Support
Dissimilar	1.00	0.65	0.79	150
Similar	0.74	1.00	0.85	150
Accuracy			0.83	300
Macro Avg	0.87	0.83	0.82	300
Weighted Avg	0.87	0.83	0.82	300

Table 12: Classification Metrics vs Threshold

It shows that the performance confirms the model is well-tuned to favor recall. Moreover, it still maintains the high precision and making it highly suitable for Legal NLP tasks. This is because it requires a high recognition of true semantic similarity to avoid significant implications.

4.5.4 Confusion Matrix

Figure 12 Confusion Matrix illustrates the model's binary classification performance at the optimal threshold of 0.200. Therefore, it shows that the number of the correctly and incorrectly classified sentence pairs as ether similar or dissimilar:

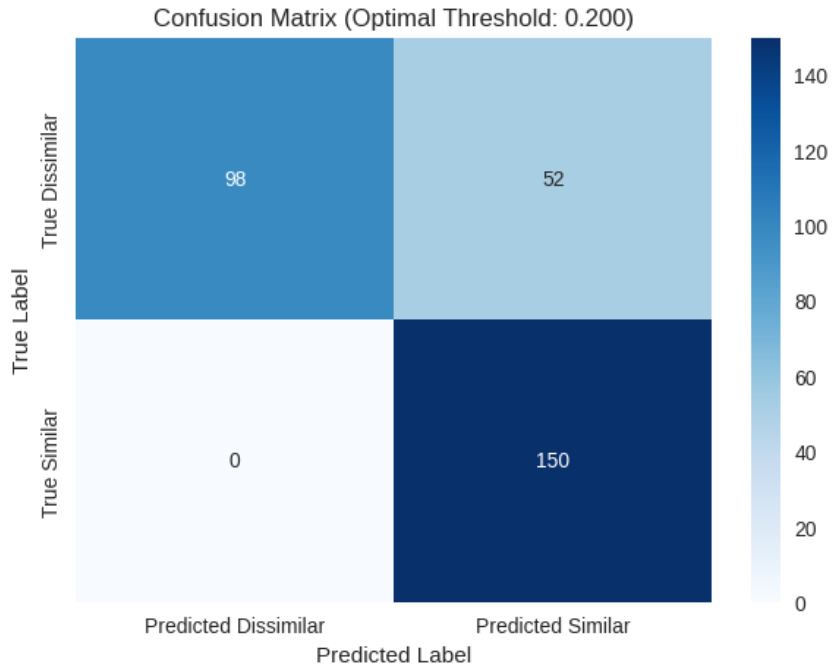

Figure 19: Confusion Matrix

	Predicted Similar	Predicted Dissimilar
True Similar	150 (TP)	0 (FN)
True Dissimilar	52 (FP)	98 (TN)

Table 13: Confusion Matrix

4.6 Model Performances Comparative Analysis

Therefore, to assess the performance of the proposed Sentence-BERT model, a comparative analysis was conducted against the traditional baseline methods. These methods were evaluated using the test dataset which contain the 300 legal text pairs. Hence, it compared across both the regression and classification metrics.

4.6.1 TF-IDF with Cosine Similarity

Table 14 below shows the performance for the TF-IDF with Cosine Similarity. It indicated that this traditional model has low Pearson correlation with only 0.1999 and Spearman correlation of 0.2225. Therefore, this means the model fails to capture the semantic similarity legal texts.

Metric	Value
Accuracy	56.67%
Precision	0.7632
Recall	0.1933
F1 Score	0.3085

Table 14: Performance of TF-IDF with Cosine Similarity

Confusion Matrix for TF-IDF with Cosine Similarity

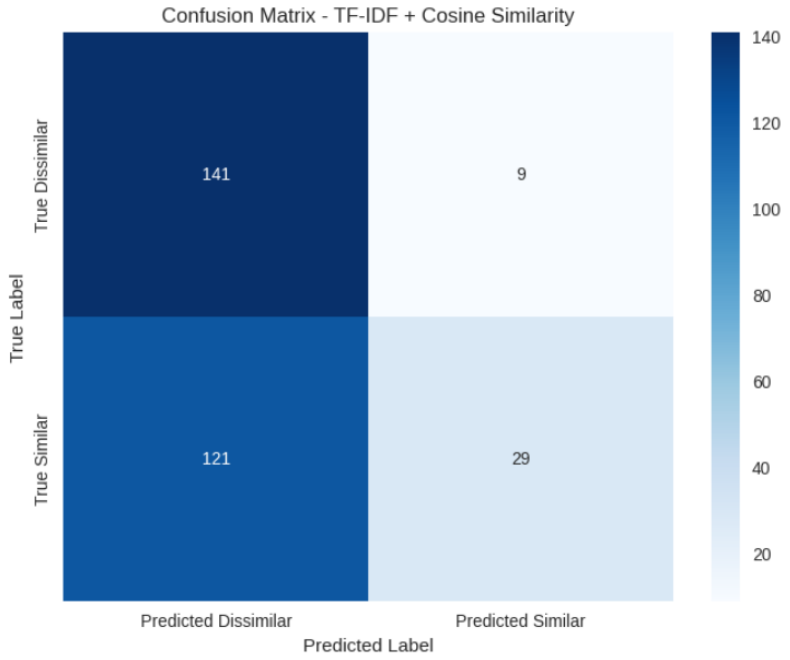

Figure 20: Confusion Matrix for TF-IDF with Cosine Similarity

4.6.2 Bag-of-Words with Cosine Similarity (BoW)

Next, BoW methods shoed slightly better result for the performance of both regression and classification tasks. For instances, the Pearson correlation score of 0.2015 and Spearman correlation of 0.2009. Besides for the classification tasks, it is as below:

Metric	Value
Accuracy	58.00%
Precision	0.5659
Recall	0.6867
F1 Score	0.6205

Confusion Matrix for BoW with Cosine Similarity

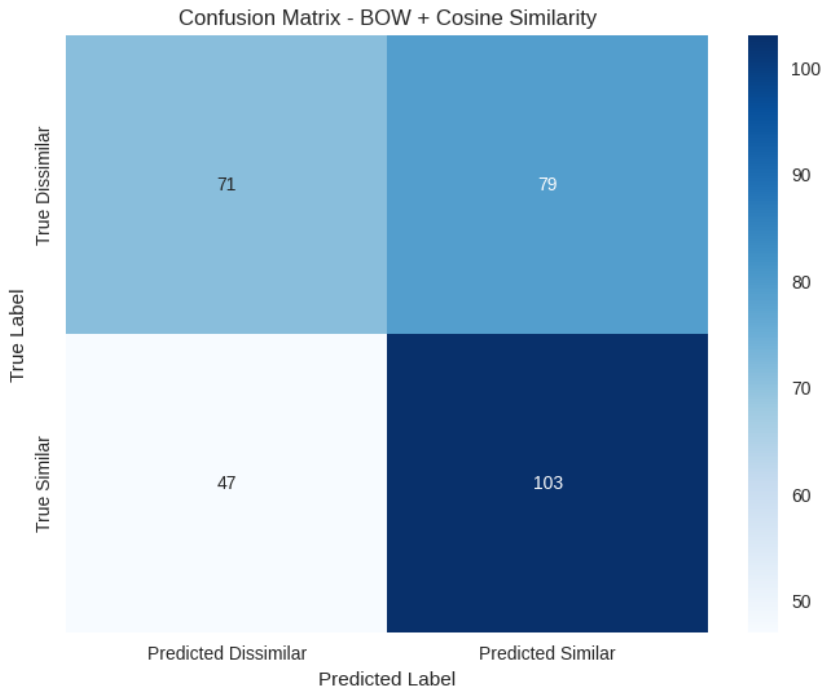

Figure 21: Confusion Matrix for BoW with Cosine Similarity

4.7 Summary of Model Performance

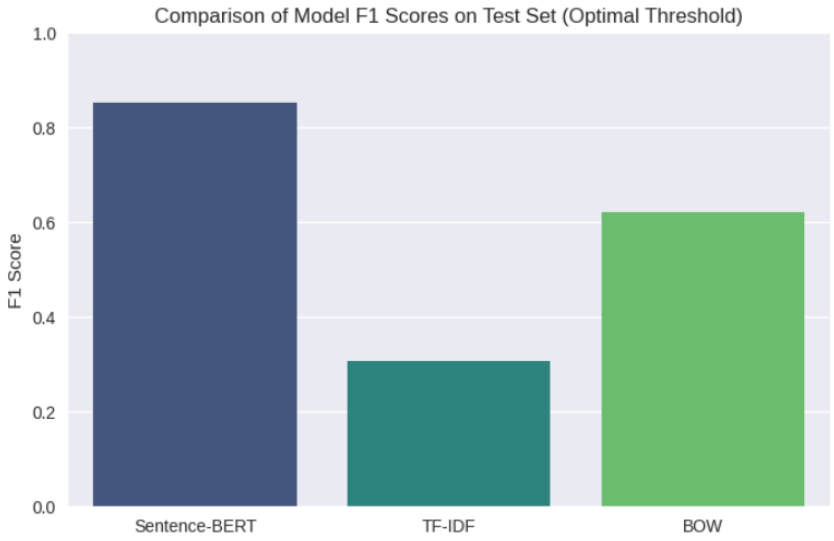

Figure 22: Comparison of Model F1 Scores

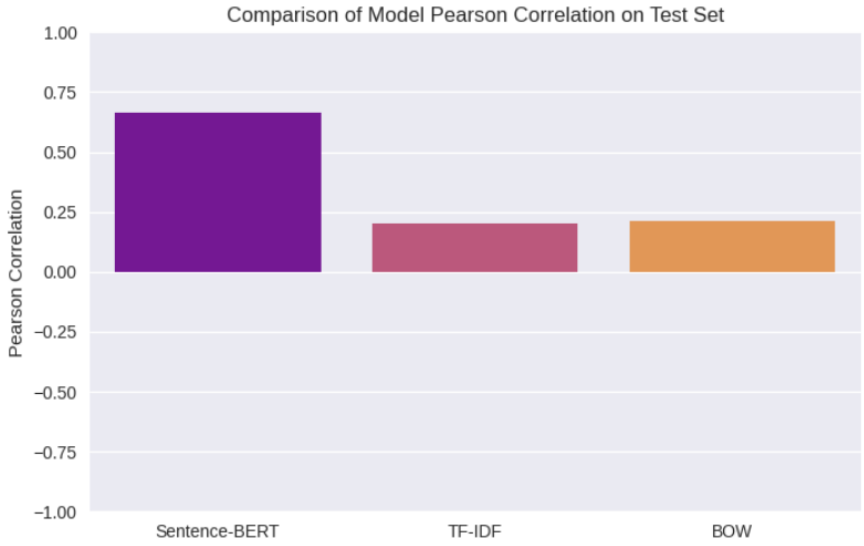

Figure 23: Comparison of Model Pearson Correlation

Model	F1 Score	Pearson Correlation
Sentence-BERT (all-miniLM6-v2)	**0.8523**	**0.6658**
BoW + Cosine Similarity	0.6205	0.2105
TF-IDF + Cosine Similarity	0.3085	0.1999

Table 15: Summary Table of Model Performance

As shown above the result strongly suggest that the traditional methods like TF-IDF and BoW are less effective in their ability for the semantic similarity tasks in legal texts. The finetuned SBERT model well performed with F1 score of 0.8523 (85.23%) and Pearson Correlation of 0.6658 (66.58%). Then, BoW with F1 score of 0.6205 (62.05%) and Pearson Correlation of 0.2105 (21.05%). Lastly, TF-IDF with 0.3085 (30.85%) and Pearson Correlation of 0.1999 (19.99%). This is because of the traditional methods' often shallow representation and lack of contextual awareness. Therefore, the result indicated that SBERT model with fine-tuned on legal sentence pairs is more reliable because it demonstrates the better performance compare to the others traditional methods.

4.8 Conclusion

The sentence-BERT model that was fine-tuned with Malaysian legal case pairs shows a strong performance in identifying semantic similarity against the traditional baseline methods. This can prove by the Pearson correlation of 0.6641 and the F1 score of 0.8523 at the optimal classification threshold. Next chapter will be the future work and conclusion.

Chapter 5: Conclusion and Future Works

5.0 Introduction

This chapter summarizes the result of the BERT-based semantic similarity of Malaysian Legal Precedents project through data acquisition from Kaggle. The results and insights obtained after going through a comprehensive data analysis phase, starting from data cleaning, and applying machine learning algorithms. Therefore, the result indicates a good performance of the BERT-based model in handling semantic similarity tasks. In addition, it also shows an overview of future project development. For instances, possibilities for making improvements in terms of quality and accuracy for better analysis. Thus, this study followed the methodology framework starting from data processing to model evaluation. Therefore, aiming to make a positive contribution by reducing the workload of legal researchers and enhancing legal research.

5.1 Summary

This project explores the application of Bidirectional Encoder Representations from Transformers (BERT) for measuring the semantic similarity between legal sentences pairs collected from Kaggle. The dataset comprising 3000 legal case sentence pairs and is used to further processed. This study motivated by the legal professionals that often face challenges in manually analyzing large volumes of legal text to identify the relevant cases. Therefore, the dataset was gone through comprehensive preprocessing such as text cleaning, domain standardization, semantic duplicates removal and class balancing using Random Oversampling.

Then, a pre-trained BERT model which is Sentence BERT was fine-tuned using the cleaned and preprocessed dataset. For instances, it was fine-tuned using CosineSimilarityLoss and evaluated with the EmbeddingSimilarityEvaluator. This process further with the model is evaluated to capture the semantic relationships between these legal sentence pairs. For instance, this study utilized regression-based methods to predict the similarity scores and evaluated them using various performance metrics.

Model performance metrics:

- Pearson Correlation: 0.6641
- Spearman Correlation: 0.5362
- R^2 Score: 0.4278
- RMSE: 0.3782 (Lower is better)
- MAE: 0.2670 (Lower is better)

As a classification task, the model achieved:

- Accuracy: 83% (High accuracy in classification)
- F1 Score (Optimal threshold = 0.2): 0.8523 (85.23%) (Excellent classification performance)
- Recall (Similar class): 1.0000 (All actual "similar" pairs were correctly identified)

Furthermore, the confusion matrix showed that 150 True Positives (Similar/Similar) pairs, 98 True Negatives) Dissimilar/Dissimilar) pairs, 52 False Positives, and 0 False Negatives. This model indicated that it is highly sensitive to capturing actual semantic similarity. This is important in legal applications, where it is beneficial to have high sensitivity.

5.2 Future works

While this study provided the valuable insight, there are some areas that need to be addressed for more accurate and excellent performance. Several suggestions for future works are as follow:

1. Larger Domain-Specific Pretraining

Instead of using a pretrained model, this project needs to train the model from scratch using a larger Malaysian legal precedents dataset. This will give the model to learn the semantic similarity of the sentences more accurately. Then, it will be fine-tuned for better performance result.

2. Sentence-Pair Expansion

Increasing the sentence pairs, such as add more diversity of the dataset. For instances, include more legal domains such as environmental law, tax law and others or integrating real-case court transcripts.

3. Contextual Embedding with Metadata

Adding more features in the dataset such as the court level, year of the judgement and jurisdiction. This will enrich the semantic context for getting excellent prediction accuracy.

4. Multilingual and Cross-Lingual Capabilities

Introduced the model that have the ability to understand multilingual sentences such as XLM-R. This can help to handle the bilingual Malaysian legal documents. Therefore, it able to capture the relationships between different language words more accurately

5.3 Conclusion

The above steps will allow further research to increase the scope of this project, improving the accuracy and relevance of the results. The current project has paved the way for the use of BERT-based semantic models as an effective tool in legal document analysis. Hence, further development will have greater implications in the future for legal research support and judicial decision-making.

References

Althammer, S., Askari, A., Verberne, S., & Hanbury, A. (2021). *DoSSIER@COLIEE 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval*. Proceedings of COLIEE 2021 Workshop: Competition on Legal Information Extraction/Entailment, 7.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). *LEGAL-BERT: The muppets straight out of law school*. arXiv preprint arXiv:2010.02559. https://doi.org/10.48550/arXiv.2010.02559

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1810.04805

Hoang, T. D., Bui, C. M., & Bui, K.-H. N. (2023). Viettel-AI at SemEval-2023 Task 6: Legal document understanding with Longformer for court judgment prediction with explanation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)* (pp. 862–868). Association for Computational Linguistics.

Judicial Appointments Commission. (2022). *The Malaysian Judiciary Yearbook 2022*. https://www.kehakiman.gov.my/sites/default/files/documents/Laporan_Tahunan/Yearbook2022.pdf

Kale, D., & Deshmukh, P. (2024). Abstractive text summarization: A transformer-based approach. In *Proceedings of the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT)* (pp. 1–6). IEEE. https://doi.org/10.1109/I2CT61223.2024.10544120

Lee, H., & Lee, H. (2023). Taiwan Legal Longformer: A Longformer-LSTM model for effective legal case retrieval. In *Proceedings of the 2023 5th International Workshop on Artificial Intelligence and Education (WAIE)* (pp. 128–133). IEEE. https://doi.org/10.1109/WAIE60568.2023.00036

Moro, G., Piscaglia, N., Ragazzi, L., & Italiani, P. (2023). Multi-language transfer learning for low-resource legal case summarization. *Artificial Intelligence and Law*, *32*(4), 1111–1139. https://doi.org/10.1007/s10506-023-09373-8

Ni, S., Li, Y., & Wang, J. (2024). *Pre-training, fine-tuning and re-ranking: A three-stage framework for legal question answering*. arXiv preprint arXiv:2412.19482. https://arxiv.org/abs/2412.19482

Owusu-Adjei, M., Hayfron-Acquah, J. B., Frimpong, T., & Abdul-Salaam, G. (2023). A systematic review of prediction accuracy as an evaluation measure for determining machine learning model performance in healthcare systems. *medRxiv*. https://doi.org/10.1101/2023.06.01.23290837

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

Putra, A. I., & Santika, R. R. (2020). Implementasi machine learning dalam penentuan rekomendasi musik dengan metode content-based filtering. *Edumatic, 4*(1), 121–130.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3982–3992).

Rosita, A., Puspitasari, N., & Kamila, V. Z. (2022). Rekomendasi buku perpustakaan kampus dengan metode item-based collaborative filtering. *Sebatik, 26*(1), 340–346.

Rosydah, S., & Widiyaningtyas, T. (2024). Collaborative filtering cosine similarity formula. *[Unpublished manuscript]*.

Rosyad, S., Mahendra, D., & Azizah, N. (2023). Sistem rekomendasi buku di perpustakaan daerah Jepara menggunakan metode item-based collaborative filtering. *Biner: Jurnal Ilmiah Informatika dan Komputer, 2*(2), 76–81.

Seyler, D., Bruin, P., Bayyapu, P., & Zhai, C. X. (2020). Finding contextually consistent information units in legal text. *CEUR Workshop Proceedings, 2645*, 48–51.

Sun, Y. (2023). The evolution of transformer models from unidirectional to bidirectional in natural language processing. In *Proceedings of the 2023 International Conference on Machine Learning and Automation*.

Tingare, B. A., & Jangid, A. (2024). Exploring the potential of transformers in natural language processing: A study on text classification. *International Journal of Progressive Research in Engineering Management and Science, 4*(8), 407–410.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Wang, X., Huang, J. X., Tu, X., Wang, J., Huang, A. J., Laskar, M. T. R., & Bhuiyan, A. (2024). Utilizing BERT for information retrieval: Survey, applications, resources, and challenges. *ACM*, 33. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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. *iFLYTEK Research; Harbin Institute of Technology*.

Widiyaningtyas, T., Ardiansyah, M. I., & Adji, T. B. (2022). Recommendation algorithm using SVD and weight point rank (SVD-WPR). *Big Data and Cognitive Computing, 6*(4), 1–15.

Ye, F., & Li, S. (2024). MileCut: A multi-view truncation framework for legal case retrieval. In *Proceedings of the ACM Web Conference 2024 (WWW '24)* (pp. 1–9). https://doi.org/10.1145/3589334.3645349

Zhang, Y., et al. (2020). Sentence embedding with limited labeled data. *Neural Computing and Applications, 32*(12), 8365–8376.

Zhong, H., Guo, Z., Tu, C., Feng, Y., & Zhang, T. (2020). Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(5), 1250–1257. https://doi.org/10.1609/aaai.v34i05.6203