NAME: MUHAMMAD HAZIQ BIN MOHAMAD MATRIC NO.: MCS241036

Chapter 1: Introduction

BERT-based Semantic Similarity of Malaysian Legal Precedents

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, CLJLaw and LexisNexis, 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.

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 LexisNexis, 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.

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.

Objectives

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

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.

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, LexisNexis, 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.

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.

References

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

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 (Vol. 34, No. 05, pp. 1250–1257). https://doi.org/10.1609/aaai.v34i05.6203

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

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

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

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