CHAPTER 5
CONCLUSION AND FUTURE WORKS

5.1 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.

5.2 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
- MAE: 0.2670

As a classification task, the model achieved:

- Accuracy: 83%
- F1 Score (Optimal threshold = 0.2): 0.8523
- Recall (Similar class): 1.0000

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.

5.3 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

5.4 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.