Chapter 2: Literature Review

2.1 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.

2.2 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).

2.3 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.

2.4 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)

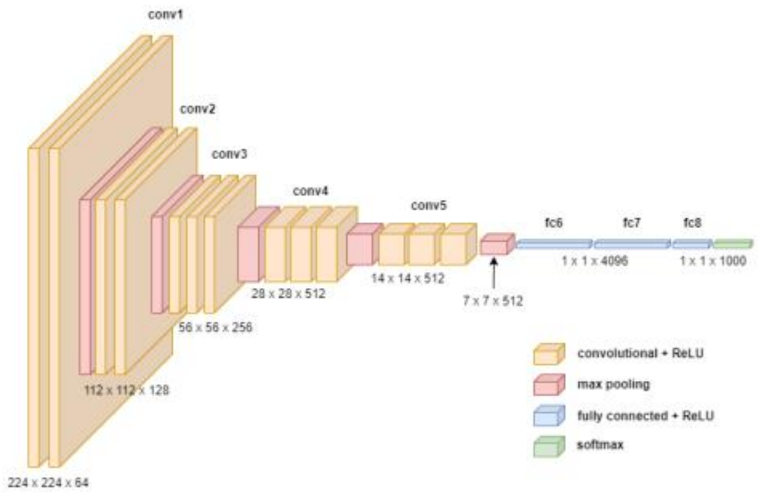

Figure 1. Architectures of traditional neural networks used in NLP: (a) Convolutional Neural Network (CNN). Source: Adapted from Sun (2023).

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)

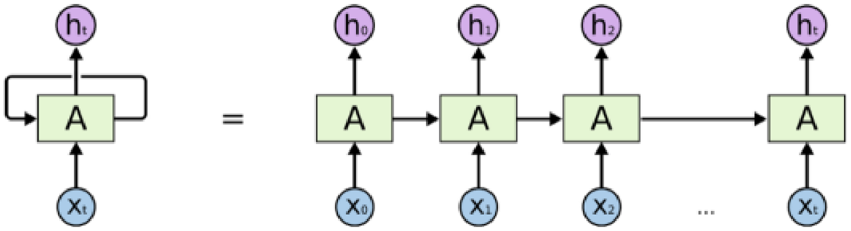

Figure 2. Architectures of traditional neural networks used in NLP: (b) Recurrent Neural Network (RNN). Source: Adapted from Sun (2023).

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)

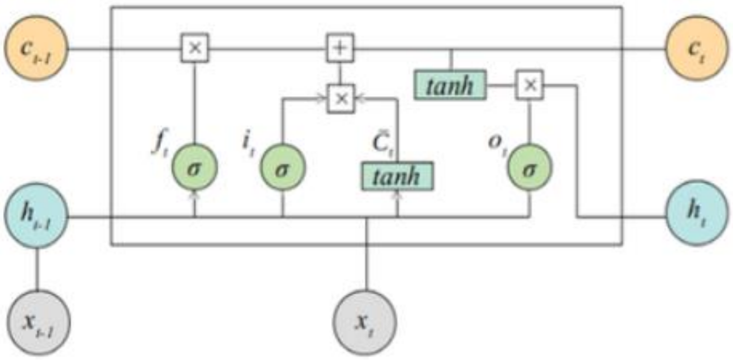

Figure 3. Architectures of traditional neural networks used in NLP: (c) Long Short-Term Memory (LTSM)*. Source: Adapted from Sun (2023).*

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).

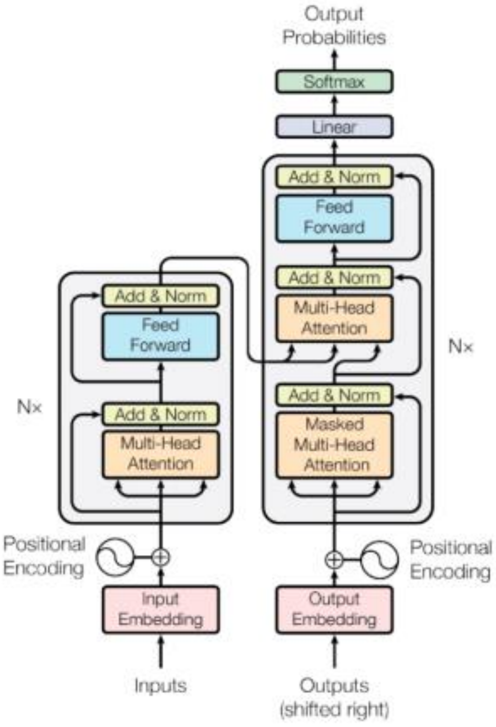

Figure 3. Transformer model. Source: Adapted from Sun (2023).

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 overcome RNN issues	Introduces memory cell and gates (input,	Effectively retains long-term	Still sequential; limits parallel training; may	Zhang & Woodland,

	using gating mechanisms to regulate information flow	forget, output); uses sigmoid functions to control flow	dependencies; addresses vanishing gradient issue	struggle with very long and hierarchically complex texts	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	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	benchmarks	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:

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

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:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

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:

$$Precision = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

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:

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

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:

$$F1\ Score = 2\ \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

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.

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 pages.

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

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. Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 862–868. Association for Computational Linguistics.

Jiajia Wang, Jimmy X. Huang, Xinhui Tu, Junmei Wang, Angela J. Huang, Md Tahmid Rahman Laskar, & Amran Bhuiyan. (2024). Utilizing BERT for information retrieval: Survey, applications, resources, and challenges. *ACM*, 33 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Kale, D., & Deshmukh, P. (2024). Abstractive text summarization: A transformer-based approach. Proceedings of the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), IEEE, 1–6. 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. Proceedings of the 2023 5th International Workshop on Artificial Intelligence and Education (WAIE), 128–133. IEEE. https://doi.org/10.1109/WAIE60568.2023.00036

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

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. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 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. *[Article source]*.

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, Yihang. (2023). The evolution of transformer models from unidirectional to bidirectional in natural language processing. *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 (IJPREMS), 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., et al. (2024). Dense passage retrieval techniques for legal document search. *Legal Informatics Review*, 9(1), 55–70.

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. *Proceedings of the ACM Web Conference 2024 (WWW '24)*, 9 pages. https://doi.org/10.1145/3589334.3645349

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