Advancing Legal NLP: A Comparative Study of Transformer-Based Models for Ambiguous Clause Detection in Legal Documents

Vrishani Shah*

MIT World Peace University, Pune, Maharashtra, India

Abstract

The goal of this study is to comprehend several Natural Language Processing approaches for removing ambiguity from legal texts for readers who are not legal experts. Ambiguity refers to clauses having more than one meaning. This is significant in the legal field as words have more than one meaning in different contexts and they may defer from the general meaning in English Language. e.g. 'consideration' as opposed to the word 'considerate'. Ambiguity can cause misleading information like wrong interpretation or incorrect translation between languages. Various approaches are evaluated in order to determine which is the most effective way to find ambiguity in legal texts. We investigate sentence-context aware word transformers like BERT, Longformer and RoBERTa, which tell us the meaning of a word in a given sentence. Additionally, growing usage of Artificial Intelligence in the legal field will make it possible to train accurate models based on a legal document specific datasets. This study compares various NLP approaches to determine which one is most effective in identifying ambiguity in legal documents.

Keywords

Ambiguity detection, context understanding, Longformer, BERT, RoBERTa, legal documents,

1. Introduction

Ambiguity identification in legal documents is a critical activity with considerable impacts for persons and organizations, even those with no professional legal training. Legal documents, including contracts, statutes, and regulations, are lengthy and complicated, containing several technical jargon and phrases that can be challenging to comprehend. These words may have various meanings in English but signify something different in legal terms. Ambiguities in these agreements can cause misunderstanding, misinterpretation, and legal issues, which can take time and money to settle. Uncertainty about the scope and applicability of a law or regulation can lead to inconsistent interpretations and enforcement.

Ambiguity identification utilizing transformers [1] can aid with the resolution of ambiguities in legal documents by recognizing potential ambiguities and providing alternate interpretations. This can assist the public with minimal legal knowledge in comprehending possible difficulties before they escalate into disputes or legal challenges. One of the primary benefits of implementing transformers for ambiguity detection is their ability to analyze enormous amounts of text swiftly and properly.

ACI'23: Workshop on Advances in Computational Intelligence at ICAIDS 2023, December 29-30, 2023, Hyderabad, India

*Corresponding author.

Vrishanishah20@gmail.com (V. Shah)

Furthermore, transformer models may learn from massive amounts of data and be trained on a wide variety of legal documents to improve their accuracy and effectiveness. This means that the models can adapt to varied legal circumstances and detect nuances in documents that people may miss. BERT and Longformer are transformers that process and analyze natural language text. These models enable computers to comprehend language in context and carry out a variety of language-related tasks. These models' transformer architecture enables them to model relationships between words in a sentence or document, as well as capture context and meaning. This is particularly important in legal papers, where language is frequently complex and context is critical to understanding the meaning of a certain sentence.

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained transformer model that was trained on huge amounts of text data to achieve a thorough knowledge of natural language. BERT analyzes language in a bidirectional manner, enabling it to collect meaning from both left-to-right and right-to-left directions. Another transformer model is Longformer, which is specifically built to handle long documents such as legal contracts and regulations. It focuses on the most relevant parts of the content using an attention mechanism and can capture context over long distances. This makes it especially effective for detecting ambiguity in legal documents, where sections or provisions may be separated by substantial gaps in the text. Both BERT and Longformer can be used to detect ambiguities in legal documents by analyzing the document's context and structure and identifying potential ambiguities and conflicts in meaning. As a result, we investigate transformers like BERT, Longformer, RoBERTa, Albert, and others that grasp the context of a sentence and detect ambiguity, assisting us in determining whether a word has more than one meaning in a certain sentence.

This comparative study observes the performance of different transformer models in the detection of ambiguity in legal documents and contracts. The length of the contract being an important deciding factor. Longformer proves to outperform the other models due its ability to parse through a larger token size which proves crucial. This paper is organized as follows: Section 2 is a literature survey of existing surveys. Section 3.1 explores the Longformer transformer explaining the architecture, Section 3.2 talks about the BERT model. RoBERTa is explored in Section 3.3. Section 44discusses the resulting outcome. Section 5 is about the future scope and implementation of this study and Section 6 contains the conclusion.

2. Literature Survey

Natural language processing tasks, including question answering [2, 3], sentiment analysis [4], natural language translation [5, 6], fake news classification [7, 8, 9], and natural language inference, were evaluated by the authors using various computational approaches. BERT and its variants are now becoming popular because of it's robust mechanism. Legal-BERT [10], is a transformer which has shown considerable potential in detecting ambiguities in legal writings. To learn the complex vocabulary of legal language, it was trained on a vast dataset of legal documents, including case law, legislation, and legal papers. To illustrate the model's usefulness in legal applications, the authors provided a set of fine-tuning tasks for it, such as legal question answering and legal document classification. On these tests, they discovered that LEGAL-BERT beat various state-of-the-art models, obtaining an accuracy of 92% on legal document classification and 73% on legal question answering. Overall, LEGAL-BERT is a promising technique for legal text processing that takes advantage of the capabilities of BERT while focusing on legal-specific problems. Its great accuracy in tasks such as legal document classification and legal question answering illustrates its usefulness in actual legal applications.

The utility of fine-tuning pre-trained transformer models for processing long legal documents, such as LegalBERT and Longformer, has been studied when we compare them as indicated in [11]. The authors propose changes to these models to address the special issues provided by legal literature, such as extended sentences, complex syntax, and domain-specific terminology. The authors compared the performance of the improved LegalBERT and Longformer models to the original models on a dataset of legal documents. They discovered that the adjusted models performed better on a variety of classification tasks, such as document categorization, sentiment analysis, and legal topic identification.

The author introduces Longformer [12], a new transformer architecture that can efficiently process long sequences of up to tens of thousands of tokens. Because of their quadratic memory and computing complexity, the authors first highlight the limits of existing transformer models, such as BERT, in handling extended sequences. They then present Longformer, which integrates a sparse attention mechanism to lower the processing requirements of transformers' self-attention mechanism. Longformer is evaluated on many natural language processing tasks, including document categorization and question answering, and its performance is compared to that of existing transformer models. Longformer outperformed other models on tasks involving long sequences while maintaining comparable performance on shorter sequences, according to the researchers.

RoBERTa (Robustly Optimized BERT) [13] is an improved version of the well-known BERT model. The original BERT model has various shortcomings including inadequate pre-training procedures, small batch sizes, and insufficient training data. They then propose many changes to the BERT pre-training method, such as higher batch sizes, longer training sessions, dynamic masking, and no next sentence prediction. Furthermore, the authors undertook a thorough examination of the aspects influencing RoBERTa's performance, such as the effect of pre-training data size, batch size, and training duration. Overall, the RoBERTa model outperforms the original BERT model in terms of performance and resilience, and its changes to the pre-training approach can help inform the creation of future transformer models.

ALBERT [14], a new variation of the BERT model, provides state-of-the-art performance on various natural language processing tasks while lowering the number of parameters greatly when compared to BERT. The authors initially identify the original BERT model's enormous number of parameters as a key obstacle to training and deployment on resource-constrained devices. They then propose many BERT architecture adjustments, such as factorized embedding parameterization, cross-layer parameter sharing, and inter-sentence coherence loss. ALBERT was trained on a huge corpus of text data and its performance on several benchmark natural language processing tasks was evaluated by the authors. They discovered that ALBERT outperformed the original BERT model on various tasks while requiring much fewer parameters. Overall, the ALBERT model offers an appealing approach for lowering the number of parameters in transformer models while still performing well on natural language processing tasks. According to the research, we can observe the use of transformers to comprehend context and overcome the ambiguity problem in legal texts.

We can also explore the Legal-pretrained Longformer models [15], though their objective is different, it gives a good overview on how the Longformer can be trained of legal data. The standard Longformer uses a pre-trained RoBERTa multiplying the token size by 8 thus creating a new attention mechanism and similarly the Legal Longformer is warm-started from the Legal-BERT. There is no pretraining involved, only finetuning in addition to the warm starting. The Legal Longformer Extended also play a major role in increasing the token size and maintaining the window span at 128. The Legal Longformer Extra Global is another approach where the global pattern is replaced with the extra global pattern. As defined in [15] 'ExtraGlobal' approach is a periodic distribution of the global tokens across the input texts being processed during the finetuning phase instead of fewer global tokens.

According to the research, we can observe the use of transformers to comprehend context and overcome the ambiguity problem in legal texts. Overall, the application of NLP in legal documents can be highly advantageous, allowing those who do not have legal competence to decode the phrases, which can aid in the resolution of difficulty or misconceptions.

3. Methodology

3.1. Longformer

Longformer [12] is a modified version of the transformer architecture built to accommodate large input sequences common in documents. The authors begin by identifying the drawbacks of existing transformer models, such as their inability to handle large input sequences and the self-attention mechanism's quadratic time and space complexity. They then propose changes to the transformer architecture, such as a sliding window self-attention mechanism, global attention bias, and sparse attention. The sliding window self-attention mechanism [16] reduces the self-attention operation's time and space complexity from quadratic O(n*n) to linear O(n*W) (Figure 1), allowing the model to handle extended input sequences.

Sliding Window Attention Pattern

The Longformer proposes a technique that transitions from local to global attention over the full document without requiring a large amount of memory. It operates on the sliding window principle, but it employs a strategy in which querying is limited to its peer node and the key nodes adjacent to it. This may appear to be sacrificing information in order to gain context, but when layered, it establishes a pattern for gathering information from neighbouring nodes in different layers as shown in Figure 2. The global attention bias directs the model's attention to the beginning and end of the input sequence, whereas the local attention bias directs the model's focus to the middle of the input sequence.

Dilated Sliding Window Attention Pattern

To cover large documents, the number of layers increases, resulting in an O(n*W*L) memory demand. To address this, we must minimize L (the number of layers) so that n*W*L«n*n. To reduce layering, more nodes must be covered in each layer, which is accomplished by selecting the alternating node for each query. This may diminish the focus on local nodes. As a result, the authors of Longformer advise utilizing a mix of Sliding Window and Dilated Sliding Window, with the beginning layers using simple Sliding Window and increasing layers using Dilated Sliding Window.

Global Attention

The combination of sliding and dilated sliding windows in Longformer is not optimized for task-specific activities. When some nodes are given the authority to attend to all nodes in a layer at the same time, this is referred to as global attention. According to the user, this could be task specific.

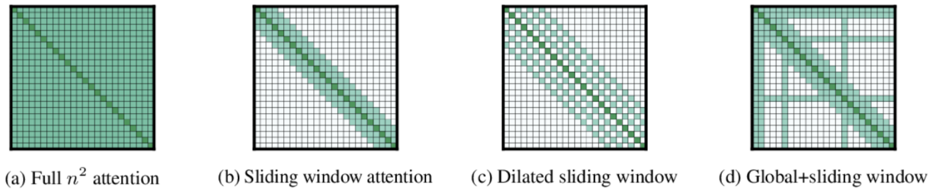
(a) Full n^2 attention
(b) Sliding window attention
(c) Dilated sliding window
(d) Global+sliding window

Figure 1: Longformer Attention Spans

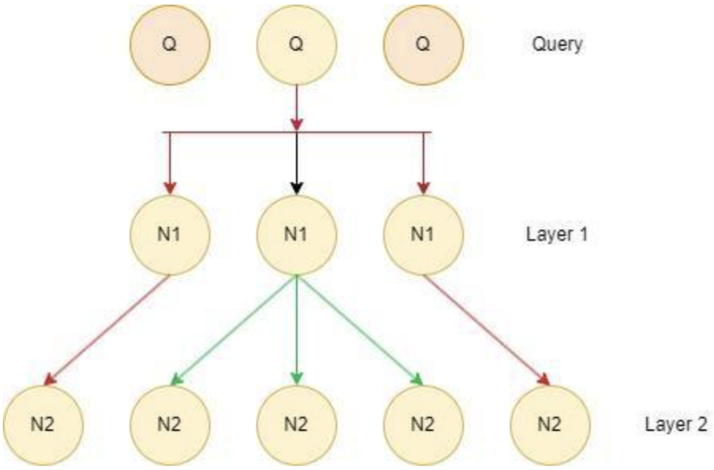

Figure 2: Layers of the Longformer

Longformer was trained on a variety of datasets, including text8 and enwik8. They attained very low BPC values such as 1.0 and 1.1, demonstrating their effectiveness in comparison to other transformers. They fine-tuned Longformer by training it on many natural language processing tasks, such as sentiment analysis and question answering, and evaluating its performance on a variety of long-document benchmarks. The document was trained on 4096 tokens, which is over 8 times the size of what BERT can handle. Longformer can be taught on any previously learned model and while training the Longformer for MLM, they continue pre-training from RoBERTa, an already trained BERT model, with just minor changes to fit the Longformer design. Except for gradient updates, sequence length, batch size, maximum learning rate, linear warmup, and polynomial decay, most parameters were retained the same as in RoBERTa. Different global attention for acceptable datasets were introduced while answering questions on datasets like WikiHOP and TriiviaQA. Before being run by Longformer, the questions were merged to make longer sequences. Most word pieces in document classification were longer than 512, which was a useful approach to gauge the skill of Longformers. Overall, the Longformer model provides a promising solution for handling long input sequences in natural language processing tasks and can inform the development of future transformer models for document processing.

3.2. BERT

BERT (Bidirectional Encoder Representations from Transformers) [17] is a pre-training method (Bidirectional Encoder Representations from Transformers) for deep neural network models based on the transformer architecture. The authors demonstrate that BERT can achieve state-of-the-art performance on a wide range of natural language processing tasks, including question answering, text classification, and language inference.

The authors note that previous pre-training methods for language models were unidirectional and lacked the ability to capture dependencies between words in both directions [17]. BERT addresses this limitation by introducing a bidirectional training objective that enables the model to capture context from both directions in the input sequence. BERT utilizes MLM (Masked Language Modelling), which masks some words and predicts them depending on the context of the sentence, as well as 'Next Sentence Prediction'.

Masked Language Modelling

Masked Language Modelling [18] is also known as Cloze Procedure. the essence of this process is masking a small number of the parameters and keeping the others the same and thus using context to predict the possible outcomes for the masked parameters. The most optimum option with highest probability is then chosen as the output. In BERT particularly, MLM is performed on unlabeled data. For context to be understood properly, the transformer should not take the meaning of just each word, but also the meaning of phrases in the tokens and the whole sentence as well. MLM then predicts the probability of words based on the meaning of the whole sentence. As stated in [18], Cloze Procedure definition is " Any single occurrence of a successful attempt to reproduce accurately a part deleted from a "message" (any language product) by deciding, from the context that remains, what the missing part should be. Cloze Procedure does not use isolated sentences to understand the perfect meaning, rather it tries to understand the meaning of the entire input and then give the best output based on what the author means. This can be very useful in legal documents because it is not just taking into account a particular phrase for eg. The phrase 'absolute discharge' in legal matter means a person has been found guilty but not convicted and the case is closed. [19] If taken separately it can mean 'allowed to flow' but when used as 'For example, a person charged with an offence may express sincere remorse and regrets toward the victim. In most cases, a charitable donation is imposed. The court considers this gesture as a compensation for the harm caused.', it means something completely different.

Next Sentence Prediction

Next Sentence Prediction is as the name suggests checking if a sentence follows another sentence. Usually during implementation, two sentences are considered where 50% of the time B does occur after A and the other 50% it is a random sentence with no contextual relation. This helps in training text pair relations. The main objective of NSP is to check if the context is valid or not based on the pair of sentences. If the NSP score is low, it may not lead to MLM thus making

BERT less sensitive to irrelevant information unlike RoBERTa, another transformer [20]. For example, if we have three sentences, 'This man was an accessory to this crime', 'The accessory enhances my outfit', 'He will be held responsible'. All sentences separately mean something, but sentence 2 is highly unlikely to follow sentence one whereas sentence 3 makes more sense. Taking a large group of sentences, separated by delimiters helps this method to then predict what fits in better and thus in understanding context of a sentence. Depending on the context of the next sentence, it can help the model to then detect ambiguity in legal documents if the sentence does not fit conventional English language meanings.

BERT delivers high performance mostly on sentence-level and token-level problems. BERT consists of two steps: pre-training and fine-tuning. Pre-Training occurs when unlabeled data is fed, and fine-tuning occurs when pre-trained labeled data from downstream tasks is fed.

They employed the same model size as OpenAI GPT when experimenting with BERT, however BERT used a bi-directional strategy, whilst the latter used constrained self-attention. BERT adopted the MLM technique during pre-training, but this caused schisms in pre-training and fine-tuning because the masked work occurred in pre-training but not fine-tuning. To reduce this, they chose to replace only 50% of the words with the masked term rather than all of them. They used texts from English Wikipedia and BooksCorpus for pre-training on Next Sentence Prediction, selecting 50% of the sentences as the next predicted one and 50% as random sentences. The authors also introduce several modifications to the transformer architecture to enable the bidirectional training objective and improve the model's performance. These modifications include adding a "segment embedding" to distinguish between the two sentences in the input, using a "position embedding" to capture the position of each token in the input sequence, and introducing a new pre-training task called "masked LM" to enable the model to handle masked input tokens.

Overall, the BERT model represents a breakthrough in the field of natural language processing and has become a widely used model for a variety of language tasks. Its success has inspired further research into pre-training methods for neural network models and has led to significant improvements in the state-of-the-art performance on many natural language processing benchmarks.

3.3. RoBERTa

RoBERTa [13] was made to improve upon BERT's performance by optimizing its pretraining objectives and hyperparameters. They achieved this by training the model on a larger corpus of text and using a new set of techniques to reduce overfitting and improve generalization.

The RoBERTa model is trained using a masked language modelling (MLM) objective, which involves randomly masking words in the input sequence and predicting them based on the context provided by the surrounding words. The authors also introduced a new training technique called dynamic masking, where the masking probability is increased as the training progresses, forcing the model to rely more on context and less on surface features. To further optimize the model, the authors trained it on a larger corpus of text compared to the corpus used to train BERT. They also introduced a set of additional pre-training tasks, such as next sentence prediction and document shuffling, to encourage the model to learn more about the structure and coherence of natural language text.

RoBERTa was also trained with a larger batch size and a longer training schedule compared to BERT, which helped reduce overfitting and improve generalization. The authors also applied a set of regularization techniques, such as dropout and weight decay, to further improve the model's robustness. The RoBERTa model (Figure 3 was evaluated on a range of benchmark NLP tasks, including question answering, natural language inference, and named entity recognition. The authors found that RoBERTa outperformed BERT on most of these tasks, demonstrating its improved ability to capture context and generalize well to new data.

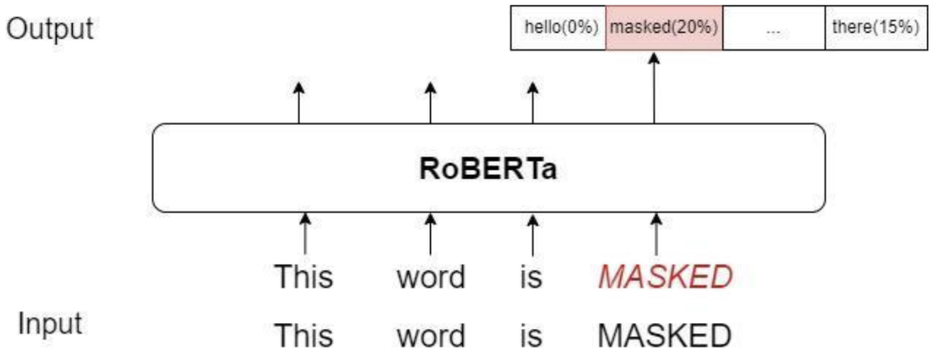

Figure 3: Working of RoBERTa

4. Results

Observing some of the best Transformers that can be used to detect ambiguity and understand context of text sequences, Longformer proves to be the best fit for detecting ambiguity in legal documents. Its receptiveness towards larger texts makes it more appropriate compared to other models available and usage of low memory and computational power gives it an edge over the other Transformers. Legal texts are frequently long and complex, making it difficult to detect ambiguity using traditional NLP models with a fixed attention span of 512 tokens, such as BERT and RoBERTa. Longformer, on the other hand, uses a sliding window attention method to analyze longer documents while preserving the same computational complexity as BERT and RoBERTa. Longformer's sliding window attention mechanism allows it to capture global context, which is crucial in detecting ambiguity in legal documents. By considering a larger context, Longformer can better understand the relationships between sentences and clauses in a document, which can help identify potential sources of ambiguity. Additionally, Longformer's pre-training objectives can also help improve its performance in detecting ambiguity. For example, the authors of the Longformer paper introduced a new pre-training objective called document token prediction, where the model is trained to predict which tokens belong to the same document. This objective encourages the model to learn more about the structure and coherence of long documents, which can help improve its ability to detect ambiguity. Overall, Longformer's ability to handle longer documents and capture global context makes it better suited for detecting ambiguity in legal documents. However, it's worth noting that the relative performance of these models may depend on the specific task and dataset, and it's always important to evaluate different models thoroughly before making a final decision.

5. Future Scope

In the future, we can use Longformer to train a model to detect ambiguity in legal documents. Longformer can handle a huge number of sequences using the Sliding Window approach, hence enormous documents can be employed. Using a suitable dataset can also aid in resolving the ambiguity by not only detecting it but also describing the meaning of the said sequence in relation to the content. This will help not only discover ambiguity, but also interpret the sentence based on its relation to the document.

6. Conclusion

In conclusion, Longformer has emerged as a promising solution for the problem of ambiguity detection in legal documents. Legal documents are often long, complex, and filled with technicalities, making it difficult for standard NLP models to capture the global context and identify potential sources of ambiguity. Longformer's ability to handle longer documents and its sliding window attention mechanism enables it to capture global context better than BERT and RoBERTa. Moreover, Longformer's pre-training objectives, such as the document token prediction task, improve its performance in detecting ambiguity. The results of recent studies demonstrate that Longformer outperforms BERT and RoBERTa in detecting ambiguity in legal documents.

Therefore, Longformer has the potential to enhance the accuracy of legal document analysis and improve the quality of decision-making by legal professionals and help those lacking legal expertise. As more research is conducted on Longformer and other transformer-based models, we can expect to see further advancements in NLP's application to the legal domain.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need (2017). arXiv:1706.03762.

[2] N. Limbasiya, P. Agrawal, Semantic textual similarity and factorization machine model for retrieval of question-answering, in: International Conference on Advances in Computing and Data Sciences (ICACDS'19), https://link.springer.com/chapter/10.1007/978-981-13-9942-8_19, 2019, pp. 195–206.

[3] N. Limbasiya, P. Agrawal, Bidirectional long short-term memory-based spatio-temporal in community question answering, Deep Learning-Based Approaches for Sentiment Analysis (2020) 291–310.

[4] C. Gupta, P. Agrawal, R. Ahuja, K. Vats, C. Pahuja, T. Ahuja, Pragmatic analysis of classification techniques based on hyperparameter tuning for sentiment analysis, in: International Semantic Intelligence Conference, volume 2786, http://ceur-ws.org/Vol-2786/Paper54.pdf, 2021, pp. 453–459.

[5] V. Madaan, P. Agrawal, Anuvaad: A hindi-sanskrit-hindi bilingual machine translation system using rule-based approach, International Journal of Social Ecology and Sustainable Development (IJSESD) 13 (2022) 1–14.

[6] P. Agrawal, L. Jain, Anuvaadika: Implementation of sanskrit to hindi translation tool using rule-based approach, Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science) 13 (2019) 1136–1151.

[7] P. K. Verma, P. Agrawal, V. Madaan, R. Prodan, Mcred: multi-modal message credibility for fake news detection using bert and cnn, Journal of Ambient Intelligence and Humanized Computing 14 (2023) 10617–10629.

[8] P. K. Verma, P. Agrawal, V. Madaan, C. Gupta, Ucred: fusion of machine learning and deep learning methods for user credibility on social media, Social Network Analysis and Mining 12 (2022) 54.

[9] K. Goel, C. Gupta, R. Rawal, P. Agrawal, V. Madaan, Fad-cods fake news detection on covid-19 using description logics and semantic reasoning, International Journal of Information Technology and Web Engineering (IJITWE) 16 (2021) 1–20.

[10] L. Zhao, X. Huang, C. Liu, J. Tang, LEGAL-BERT: The muppets straight out of law school, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 4665–4674.

[11] A. Bhatia, M. Kaur, Processing long legal documents with pre-trained transformers: Modding LegalBERT and longformer, in: Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval, 2021, pp. 31–39.

[12] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The Long- Document Transformer, 2020.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach (2019). arXiv:1907.11692.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS, 2019.

[15] P. Tsotsi, Exploration of Efficient Transformer Methods for Long Legal Document Processing, 2022.

[16] R. Eifler, Understanding longformer's sliding window attention mechanism, 2020.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, in: Proceedings of the 2019 Conference of the North, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019.

[18] W. L. Taylor, "cloze procedure": A new tool for measuring readability, Journalism Quarterly 30 (1953) 415–433.

[19] D. Avocats, Doyon avocats - discharge, https://www.doyonavocats.ca/en/discharge/, 2023. Accessed: 2024-5-2.

[20] F. Petroni, P. Lewis, A. Piktus, T. Rocktäschel, Y. Wu, A. H. Miller, S. Riedel, How context affects language models' factual predictions (2020). arXiv:2005.04611.