AraLegal-BERT: A pretrained language model for Arabic Legal text

Muhammad AL-Qurishi, Sarah AlQaseemi and **Riad Soussi**
Elm Company, Research Department, Riyadh 12382, Saudi Arabia

Abstract

The effectiveness of the bidirectional encoder representations from transformers (BERT) model for multiple linguistic tasks is well documented. However, its potential for a narrow and specific domain, such as legal, has not been fully explored. In this study, we examine the use of BERT in the Arabic legal domain and customize this language model for several downstream tasks using different domain-relevant training and test datasets to train BERT from scratch. We introduce AraLegal-BERT, a bidirectional encoder transformer-based model that has been thoroughly tested and carefully optimized with the goal of amplifying the impact of natural language processing-driven solutions on jurisprudence, legal documents, and legal practice. We fine-tuned AraLegal-BERT and evaluated it against three BERT variants for the Arabic language in three natural language understanding tasks. The results showed that the base version of AraLegal-BERT achieved better accuracy than the typical and original BERT model concerning legal texts.

1 Introduction

The impressive performance of bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) inspired numerous authors to try and improve the original BERT. Such follow-up research progresses in several directions, including the development of specific solutions for various thematic domains. This is necessary because the vocabulary used in some fields significantly differs from the language used for everyday purposes and may contain the specific meanings of certain phrases or atypical relationships between contextual elements. This problem can be partially resolved through domain-specific adjustments to the training process. A good example of this approach is demonstrated in (Chalkidis et al., 2020), who created Legal-BERT specifically for mining legal text in English, improving the

output of the standard transformer algorithm in this domain. Another form of the BERT concept was successfully adapted by (Beltagy et al., 2019; Lee et al., 2020), who created models that were pretrained on a compilation of scientific and biomedical data from various fields, achieving significantly better performance on scientifically related natural language processing (NLP) tasks. These examples show that BERT is currently far from a finished model and that its effectiveness could be further enhanced, at least for relatively narrowly defined tasks.

For the Arabic language, several BERT-based models have consistently demonstrated superior performance on numerous linguistic tasks requiring semantic understanding, outperforming all benchmarks on public datasets such as the Arabic NER corpus (ANERcorp) or the Arabic Reading Comprehension Dataset (ARCD), such as the works presented by (Antoun et al., 2020; Abdul-Mageed et al., 2020) and mBERT by Google (Devlin et al., 2018). This is largely a consequence of the efficient transfer learning inherent in this model, involving a high computational cost because this approach requires huge collections of training examples, followed by fine-tuning for specific downstream tasks. A significant advantage of BERT is that the training phase can be skipped because a pretrained version of the model can be used and trained further. However, (Chalkidis et al., 2020; Beltagy et al., 2019; Lee et al., 2020) has shown that a generic approach to pretraining does not work well when BERT must be used in a domain with highly specific terminology, such as legal, science, or medicine. There are two possible responses to this issue: continue specializing in a pretrained model or train a model from scratch with relevant materials from the domain. In this study, we built a language model from scratch based on the original BERT (Devlin et al., 2018), which is specific to Arabic legal texts. The aim was to improve the performance on most state-

of-the-art language understanding and processing tasks, especially related to Arabic legal texts. We believe that the specific nature of legal documents and terminology needs to be considered because it affects the way sentences and paragraphs are constructed in this field. The extent of formal and semantic differences is such that some authors describe the linguistic content used for legal matters as almost a language of its own (Zhang et al., 2022; Silveira et al., 2021).

By focusing on a single domain, the Arabic legal text, this study attempts to reveal means of adapting an NLP model to fit any thematic domain. Based on our experiments, we can confirm that pretraining BERT with examples from the Arabic legal domain from scratch provides a better foundation for working with documents containing Arabic legal vocabulary than using the vanilla version of the algorithm. We introduce AraLegal-BERT, a transformer-based model that has been thoroughly tested and carefully optimized with the goal of amplifying the impact of NLP-driven solutions on jurisprudence, legal documents, and legal practice. We fine-tuned AraLegal-BERT and evaluated it against three BERT variants for Arabic in three natural language understanding (NLU) tasks. The results show that the base version of AraLegal-BERT achieved better accuracy than the general and original version of the BERT model, in regard to legal text. AraLegal-BERT is a particularly efficient model that can keep up with the output of computationally intensive models while producing its findings faster and using far fewer resources. Consequently, the base version of the model was observed to have the ability to achieve comparable accuracy to larger versions of the large general and original version of the BERT model when they were trained with domain-relevant examples similar to those used to test our model.

2 Related work

In a very short time, the transformer (Vaswani et al., 2017) architecture has become the gold standard for machine learning methods in the field of linguistics (Wolf et al., 2020; Su et al., 2022). The unprecedented success of BERT combined with its flexibility has led to a proliferation of tools based on it, built with a more narrowly defined vocabulary (Young et al., 2018; Brown et al., 2020). AraBERT (Antoun et al., 2020; Abdul-Mageed et al., 2020) is an example of such specialization

and could have considerable practical value, given the number of Arabic speakers worldwide. Because the model is trained for some of the most common NLP tasks and has proven effective across regional variations in morphology and syntax, this language model has the potential to become a standard tool for analyzing Arabic text. The pretraining and fine-tuning procedures described in this work may not be optimal; however, the output of the localized model clearly indicated that the initial approach was correct. With further refinement, the model can become sufficiently reliable for a wide range of real-world applications. However, these models are based on data, most of which were collected from Modern Standard Arabic terms, and these language models may fail when the language switches to colloquial dialects (Abdelali et al., 2021). In addition, the performance of these language models can be affected when dealing with a language for a specific domain with special terms, such as scientific, medical, and legal terms (Yeung, 2019).

The majority of domain-specific BERT models are related to scientific or medical texts, and legal texts; however, the texts are all in English (Beltagy et al., 2019; Lee et al., 2020; Chalkidis et al., 2020). In a study by (Alsentzer et al., 2019), the main area of interest was clinical practice; therefore, the authors developed two different variants by pretraining basic BERT and BIOBERT with examples from this domain, with positive results in both cases. Another interesting related project was conducted by (Beltagy et al., 2019), resulting in the creation of SCIBERT, a whole branch of variations optimized for use with scientific documents. In this case, two different optimization strategies, including additional training and training from scratch using documents comprising scientific terminology, were tested, with both approaches yielding measurable improvements. A study by (Chalkidis et al., 2020) involved pretraining transformer models for English legal text by comparing three possible approaches to adapt BERT to thematic content niches: 1) using the vanilla model without any modifications, 2) introducing pretraining with datasets that contain examples from the target domain in addition to the standard training, and 3) using only domain-relevant samples to train BERT from scratch.

Essentially, all the adaptations of the standard BERT model that involve fine-tuning use the same approach to select hyperparameters as outlined in

the original BERT formulation, without even questioning it. Another research gap is observed regarding the possibility of using shallow models to perform domain-specific tasks. The impressive generalizability of deep models with several layers could be argued as wasted when the model operates in a narrowly defined field where linguistic rules are more streamlined and vocabulary volume more limited. Despite BERT being the most successful deep learning model for various tasks related to the legal sphere, there have been no published attempts to develop a unique variation for this type of content, especially in Arabic, inspiring this study. Therefore, this is the first study to build a BERT-based language model for legal texts in Arabic.

3 AraLegal-BERT: Transformer Model Pretrained with Arabic Legal Text

To optimize BERT to work with Arabic legal documents, we followed the same procedures in the original BERT model (Devlin et al., 2018); however, for the Arabic language, we followed the same procedure in AraBERT (Antoun et al., 2020).

3.1 Dataset

Due to the relative scarcity of publicly available resources containing legal text in Arabic, the training dataset had to be manually collected from numerous sources and included several regional variations. All the collected documents were in the Arabic language and related to different subfields of legal practice, such as legislative documents, judicial documents, contracts and legal agreements, Islamic rules, and Fiqh. All data were collected from public sources, and the final size of the dataset was 4.5 GB. The final size of the training set after removing duplicates was approximately 13.7 million sentences. Table 1 lists the dataset used in this study.

3.2 AraLegal-BERT

This version of the model was created by following the original pretraining process with additional steps involving textual material from the Arabic legal domain. The authors of the original BERT model indicated that 100,000 steps would be sufficient; however, in our implementation, the model was trained with up to half a million steps to determine the impact of extended pretraining with narrowly focused data samples on the performance of various linguistic tasks. The pretraining of the

BERT base model (Devlin et al., 2018) with general content involves significantly more steps; therefore, the model tends to be the most proficient, with a vocabulary containing approximately 30,000 words, found in everyday speech. With extended training using domain-focused examples, this tendency was presumed to have the ability to be partially reversed with a positive impact on model accuracy.

Before we started the training, the data was preprocessed, and in this phase, we followed the same procedure as in (Antoun et al., 2020). To account for the uniqueness of Arabic prefixes, subword segmentation was performed to separate all tokens into prefixes, stems, and suffixes, as explained in (Abdelali et al., 2016). This resulted in a vocabulary of approximately 64,000 words used to pretrain the model and create AraLegal-BERT. Subsequently, we trained our model using the masked language modeling (MLM) task, where 15% of the words in an entire input sequence were used as tokens because 80% of them were masked, 10% were replaced with a random token, and only 10% were left in their natural state. This procedure allows the algorithms to derive conclusions based on whole words and not just linguistic elements; this procedure is better suited for the Arabic language.

4 Experimental procedure

4.1 Pretraining stage

AraLegal-BERT was trained for approximately 50 epochs, involving a total of half a million steps, which is similar to the original BERT pretraining procedure. We trained our model at the Elm Research Center using NVIDIA DGX1 with eight GPUs. The batch size was set to 8 per GPU; therefore, the total training batch size (w. parallel, distributed & accumulation) was 512. The maximum sequence length was 512 tokens, and the learning rate ranged from $1e-5$ to $5e-5$.

4.2 Fine-tuning

The authors of BERT (Devlin et al., 2018) proposed an approach for determining the optimal parameters for fine-tuning based on a search within a limited range. In this concept, the learning rate, training duration, size of the training stack, and dropout rate are either fixed or can be one of a few possible discreet values. Although no particular reason was provided for this approach, it has been widely replicated in studies dealing with BERT derivatives (Wehnert et al., 2022; Rogers et al., 2020).

Table 1: Dataset used to train AraLegal-BERT

Type	Sample Size	Desc
Books	6K	Master and PhD theses, research papers, magazines, dictionaries and Fiqh books
Cases	336K	Legal Cases in KSA and Gulf countries which consists of copy rights, design rights, facts and appealing
Terms and laws	3K	Laws and regulations in KSA and Gulf countries
others	5K	Reports and studies, academic courses, forms, reports, contracts

Because these parameters do not always produce the best results, and their use can still leave a model undertrained, an alternative strategy was adopted to choose the upper limit of training epochs that tracks the loss of validation and terminates training only when the conditions are met.

4.2.1 Legal text classification

The samples used in the experimental dataset were collected from two main portals. The first dataset was collected from the Scientific Judicial Portal(SJP)[1], operated by the Ministry of Justice in Saudi Arabia. The SJP is the largest specialized information database in the field of justice in the Kingdom of Saudi Arabia. It is the ideal solution for specialists, including judges, lawyers, trained lawyers, academics, prosecutors, and graduate students, in the justice and legal domain. The second dataset collected was from the Board of Grievances(BoG) portal[2], where the following is stated in their website: *"The Board tested a judiciary academic series in the name of (judge library) and its distribution among the Board judges (hard copy and soft copy) to increase cognitive formation with them, a state which its effect shall be reflected on the judiciary verdicts they issue, including academic references in administrative, commercial and penal judiciary formed of 32 volumes in addition to judiciary verdicts"*.

Because existing documents in both datasets can belong to multiple categories depending on the submission details, they are suitable for the task of classifying lengthy legal documents. Three different classes of documents were selected from the SJP dataset and ten classes from the BoG dataset. Because all documents from certain classes are essentially headlined in the same manner, the classification task required that the parts of the document containing easily identifiable indicators of the class were trimmed. Owing to this omission, the model needs to analyze the entire content instead of deriving the conclusion based on just the first few lines. This modification was implemented for all classes.

4.2.2 Keyword Extraction

Unfortunately, compared with the data available for research in English and a few Latin languages, there are no ready-made and well-prepared data for research purposes in Arabic, especially for understanding the natural languages of legal texts. Therefore, we built our dataset for this task with the help of professionals in the Arab legal domain. This dataset consists of approximately 8,000 legal documents containing the most important keywords manually extracted by these professionals. We preprocessed and cleaned the data and extracted approximately 37640 sentences containing keywords and other words that formed the sentences. The average length of the sentences was no more than 65 words because we performed a sentence segmentation process to ensure that each sentence did not lose its meaning or was not trimmed. We tagged the keywords in the sentence with the number 1 and the others with the number 0.

4.2.3 Named Entity Recognition

This dataset was also generated in the research department of Elm, Saudi Arabia. It contains more than 311,000 sentences, including thousands of distinct entities of 17 different sequence tags manually labeled by multiple human annotators as a part of our CourtNLP project at Elm research[3]. All the classes used are shown in Figure 1. The main objective of the NER procedure is to assign a label belonging to a particular class to each of the included words. Furthermore, some complex named entities can span multiple words; however, they are always contained within a single sentence. The IOB format (short for inside, outside, beginning) is predominantly used to represent the sentences in this field, with words starting with the name of an entity marked as B, internally located words as I, and other tokens as O.

---

[1] https://sjp.moj.gov.sa/
[2] https://www.bog.gov.sa/en/ScientificContent/Pages/default.aspx

[3] https://www.elm.sa/en/research-and-innovation/Pages/Research.aspx

Table 2: Overall results of all fine-tuned models in the legal text classification task on BoG Dataset

Model / Macro-Average	Precision	Recall	F1-score
Arabertv2-Large (Antoun et al., 2020)	0.850387	0.810795	0.827078
ARBERT (Abdul-Mageed et al., 2020)	0.802514	0.821973	0.812820
mBERT (Devlin et al., 2018)	0.702017	0.635928	0.598267
AraLegal-BERT (base)	**0.89276**	**0.89173**	**0.89098**

Table 3: Overall results of all fine-tuned models in the legal text classification task on SJP Dataset

Model / Macro-Average	Precision	Recall	F1-score
Arabertv2-Large (Antoun et al., 2020)	0.885678	0.886816	0.884516
ARBERT (Abdul-Mageed et al., 2020)	0.837714	0.834804	0.843827
mBERT (Devlin et al., 2018)	0.814763	0.780226	0.782501
AraLegal-BERT (base)	**0.92395**	**0.92133**	**0.92210**

English Tag	Arabic Tag	Tag count
accusation	التهمة	19974
document	مستند	91058
evidence	دليل	67456
Hadith	حديث نبوي	10094
Qura'n	قرآن كريم	8019
job	وظيفة	46415
jurisprudence	فقه	15648
law	قانون	41694
location	موقع	9127
organization	منظمة/مؤسسة	73171
person	شخص	15685
Nationality	جنسية	26651
Currency	عملة	13456
Amount	مبلغ	17982
Date time	وقت وتاريخ	36548
citation	اقتباس	36133
verdict	حكم	38189

Figure 1: Main Arabic Legal Named Entity Tags

5 Results

5.1 Impact of pretraining

We trained two models from scratch: the first was a base model that contains 12 layers, and the second was a large model that consists of 24 layers, similar to the original BERT. As anticipated, the full-sized 24-layer model trained from scratch had a significantly better ability than that of the base model with 12 layers, to meet the pretraining objectives. However, after the completion of the pretraining stage, the base model displayed a level of loss similar to that of the original BERT model trained with general datasets. In particular, a model's ability to adapt to narrowly defined niches is faster, which can be a significant advantage for domain-focused

applications such as those used for the legal domain. Therefore, the content of the training set plays a role in choosing the appropriate training method. We are yet to perform experiments on the large model, and all fine-tuning results were based on the base model because we found that it provides significantly higher accuracy than the general Arabic BERT models in the three defined NLU tasks.

5.2 Results and discussion

We divided the datasets for all three tasks into training, validation, and testing sets. In this section, we discuss the test results. The evaluation was conducted using standardized hyperparameters, such as batch size and sequence length, with different datasets suitable for legal text classification, keyword extraction, and named entity recognition.

The best option for the first task of legal text classification is determined based on experimental results. For example, using this method, we found that multiple strategies could be used to bypass BERT's sequence length limitation of 512 tokens; however, the "head & tails" strategy, where only the first 128 and the last 382 tokens are retained, exhibits the best performance, such as the work in (Sun et al., 2019). Tables 2 and 3 summarize the overall results of our model compared with those of the three BERT variants for the Arabic language on the classification task with two datasets, namely SJP and BoG. On the BoG dataset, AraLegal-BERT outperformed all models by 0.7% in terms of F1-Macro average, which is higher than ARABERT-v2large. Similarly, our model also outperformed

the other models on the SJP dataset; it achieved approximately 0.4% higher F1-Macro average than that of ArabBERTv2-large.

Furthermore, for the tasks related to named entity recognition and keyword extraction, we followed the same procedure that was performed in our previous work (Al-Qurishi and Souissi, 2021); considering that no new layers were added to the model, a linear layer was used to make the words and sequence-level tagging. The results were extremely different for these two tasks; furthermore, there was a significant difference between the performance of AraLegal-BERT and the other models; AraLegal-BERT achieved 21% higher F1-Macro average than ARABERT-v2large (Antoun et al., 2020) in extracting named entities, as shown in Table 5. In addition, the difference was significantly higher in the keyword extraction task, where AraLegal-BERT outperformed the highest model, ARBERT (Abdul-Mageed et al., 2020), with a significant difference of almost 26% in F1-Macro average, as shown in Table 4.

We denote that the general BERT models exhibited not only a low F1-Macro score but also a low recall-macro score, where they were not able to retrieve most of the required words compared with those retrieved by AraLegal-BERT. Finally, we would like to highlight that AraLegal-BERT is the base version; yet, it outperformed the rest of the models in all three defined tasks, with low memory requirement, faster performance, and good accuracy.

6 Conclusions and future work

Our experimental results show that a BERT model pretrained for a specific domain is better than the typical language models, for specific NLU tasks. Therefore, we present AraLegal-BERT, which was trained exclusively for Arabic legal texts and is capable of making highly accurate predictions on three main NLU tasks: legal text classification, named entity recognition, and keyword extraction. Essentially, the level of difficulty of a task is correlated with the gains from choosing the right training strategy as the importance of domain-specific vocabulary and semantics becomes more pronounced. The tested version of AraLegal-BERT is the base, cost-efficient version suitable for a broad range of Arabic legal text applications. Our future work will focus on additional possibilities for pretraining other models, such as the Electra, Roberta and

XLM-R models for several NLU tasks in the Arabic legal domain with small, base, and large versions.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pages 11–16.

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Muhammad Saleh Al-Qurishi and Riad Souissi. 2021. Arabic named entity recognition using transformer-based-crf model. In *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 262–271.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Table 4: Overall results of all fine-tuned models in the Keywords Extraction Task

Model / Macro-Average	Precision	Recall	F1-score
Arabertv2-Large (Antoun et al., 2020)	0.765979	0.470098	0.582625
ARBERT (Abdul-Mageed et al., 2020)	0.789460	0.524715	0.630420
mBERT (Devlin et al., 2018)	0.721356	0.375075	0.493533
AraLegal-BERT (base)	**0.93481**	**0.84449**	**0.88736**

Table 5: Overall results of all fine-tuned models in the Named Entity Recognition Task

Model / Macro-Average	Precision	Recall	F1-score
Arabertv2-Large (Antoun et al., 2020)	0.891934	0.450073	0.598261
ARBERT (Abdul-Mageed et al., 2020)	0.889916	0.413266	0.564423
mBERT (Devlin et al., 2018)	0.886825	0.326475	0.477254
AraLegal-BERT (base)	**0.89848**	**0.73644**	**0.80943**

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Raquel Silveira, CG Fernandes, João A Monteiro Neto, Vasco Furtado, and José Ernesto Pimentel Filho. 2021. Topic modelling of legal documents via legal-bert. *Proceedings http://ceur-ws org ISSN*, 1613:0073.

Xing Su, Shan Xue, Fanzhen Liu, Jia Wu, Jian Yang, Chuan Zhou, Wenbin Hu, Cecile Paris, Surya Nepal, Di Jin, et al. 2022. A comprehensive survey on community detection with deep learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sabine Wehnert, Shipra Dureja, Libin Kutty, Viju Sudhi, and Ernesto William De Luca. 2022. Applying bert embeddings to predict legal textual entailment. *The Review of Socionetwork Strategies*, 16(1):197–219.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Chin Man Yeung. 2019. Effects of inserting domain vocabulary and fine-tuning bert for german legal language. Master's thesis, University of Twente.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75.

Gechuan Zhang, Paul Nulty, and David Lillis. 2022. Enhancing legal argument mining with domain pre-training and neural networks. *arXiv preprint arXiv:2202.13457*.