CHAPTER 4

LEGALEPH: A HEBREW LEGAL LANGUAGE MODEL

In this chapter, we present the development of a language model for the legal domain in Hebrew, obtained by further pre-training the recently published AlephBERT model [Seker et al., 2021], on a carefully curated corpus from several sources of legal texts. While BERT-based models [Devlin et al., 2019] achieved state-of-the-art results in several downstream NLP tasks on generic benchmark datasets such as GLUE Wang et al. [2018], SQUAD Rajpurkar et al. [2016], and RACE Lai et al. [2017], relatively little work has been invested in adapting them to specialized domains, particularly in low-resource settings. One such domain is the legal domain, which differs greatly from other natural language domains, and has its own promising downstream NLP tasks, such as anonymization and masking of sensitive or private information.

This approach was recently proven effective in English. Chalkidis et al. [2020] released a family of BERT-based models specialized in the legal domain by training them over an ensemble of English legal-domain datasets. By using the annotations already available in these datasets, they found that their model achieves better results than $\text{BERT}_{\text{BASE}}$ model [Devlin et al., 2019] in binary and multi-label text classification of EU laws and cases from the European Court of Human Rights, as well as named entity recognition on US contracts. They also published their work so following research can leverage it to ask more complex questions.

Encouraged from these findings and the release of AlephBERT, we seek to develop a parallel Hebrew language model which can be applicable for domain-specific tasks which are currently done manually requiring excessive manual labour of experts, hence saving time and money. AlephBERT was pre-trained on Hebrew texts taken from Twitter, Wikipedia, and OSCAR [Ortiz Suárez et al., 2020], and while anecdotally we found evidence that OSCAR contains some legal texts, we are not aware of a systemic analysis of their extent or effect on downstream performance in relevant tasks.

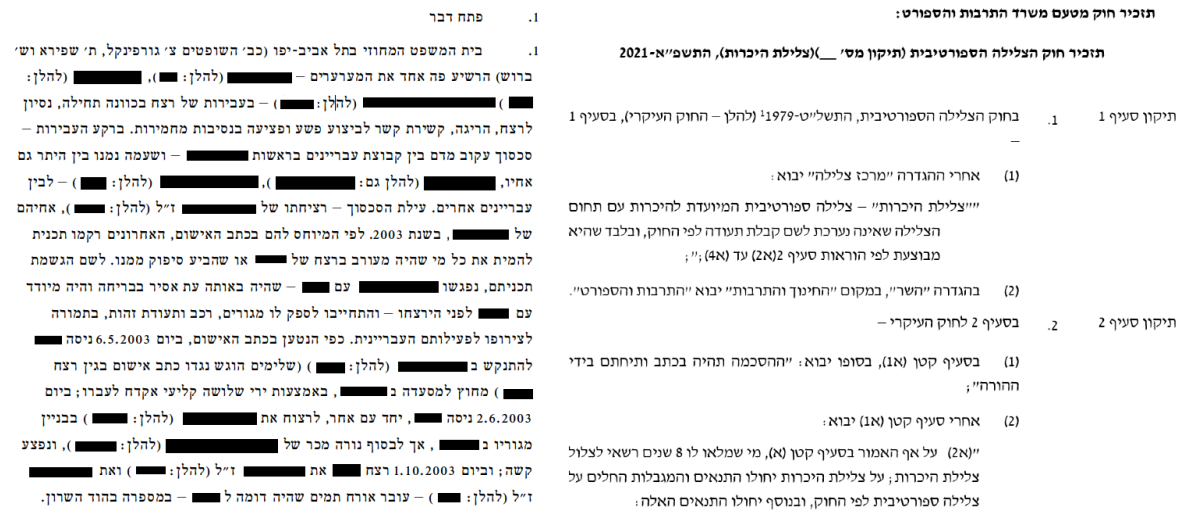

Figure 4.1: An excerpt from a court pleading (left) versus an excerpt from a memorandum of law (right). While court pleadings are composed of document-length narratives, legislation documents are mostly split across articles and sections, when one sentence may be split across multiple sections.

Dataset	# Texts	# Words	# Tokens
ISCD	16K	26M	38M
Open Laws Book	2k	5.5M	7.5M
Laws Book	650	1M	1.4M
Secondary Legislation	2350	2M	3M
MISHMORET	72K	30M	49M
Total	94K	34.5M	90M

Table 4.1: The number of texts, words, and tokens in our preprocessed version of the datasets. The number of tokens was calculated according to the tokenizer of AlephBERT.

In the following, we describe the curation of a first large-scale dataset for the legal domain in Hebrew (Section 4.1), several downstream tasks which can be derived from these datasets (Section 4.2), our trained model, and initial findings on some of these tasks (Section 4.3), and our plans for future work (Section 4.4).

4.1 Dataset Collection

To form a Legal Hebrew dataset, we gathered legal texts from several sources. From the Israeli Supreme Court Database (ISCD),[1] we collected DOC files of court pleadings from the years 2010-2017 including dozens of classification labels, such as the disposition and outcome of the final decision, the most important legal subject matter or issue of the case. From the Israeli Open

[1] http://iscd.huji.ac.il

Laws Book initiative,[2] we received XML files in AKN format of official laws. From the Israeli Ministry of Justice, we received amendments of laws, and secondary legislation with several classification labels such as the type and the main topic of each legislation document, as well as a summary of the legislation. We also scraped a vast amount of DOC files of protocols from the court for custody of illegals (MISHMORET).[3] Table 4.1 provides further information about these sources.

We unified and preprocessed all of these sources into a standard format amenable for NLP model development and testing. For example, we concatenated only the texts in the XML files in AKN format as we found this format inconsistent across multiple files. This unified format and the dataset is made publicly available,[4] and will hopefully ease future development of models and applications for the legal domain in Hebrew.

4.2 Downstream tasks

In the legal domain, there are several real-world applications that can be addressed with NLP modelling, this includes anonymization of legal documents, masking sensitive or private information (as illustrated on the left side of Figure 4.1), predicting the outcome of a trial based on evidence and protocols, identifying biases in judicial decisions and more. These tasks require extensive annotation and constitute research effort on their own. So, as a first step to enable such research projects, we follow the approach taken by Legal BERT and use the labels that are already present in the data in the form of metadata to estimate the quality of our legal language model.

After examining the labeled data in the different datasets, we found several optional downstream tasks which are depicted in Table 4.2. First, we split between court pleadings or protocols and legislation, as they have different characteristics: while court pleading and protocols have longer paragraphs containing multiple sentences, legislation documents are split across multiple articles, Figure 4.1 illustrates the difference. Second, while the datasets vary widely in their label, the classification of Court Issues in the ISCD and the classification by topic in secondary legislation documents are comparable and might be interesting to extrinsically evaluate the performance of the model across the two sub-domains.

One major challenge is the varying length of each document since both pleading and legislation documents can contain hundreds of pages. Therefore, we can either use models designed to handle longer documents [Beltagy et al., 2020, Zaheer et al., 2020] or reduce our finetuning data to relatively short documents. Alternatively, we can first focus on sentence-level tasks, such as author recognition, Legal and court issue classification, topic modelling.

[2] https://www.openlaw.org.il

[3] https://data.gov.il/dataset/mishmoret/resource/ee9cc077-763f-44da-aea6-235ffaf72d3a

[4] https://github.com/korenLazar/hebrew-legal-dataset

Dataset Name	Metadata	Optional Downstream Tasks
ISCD	Dozens columns	Author recognition Judicial decision prediction Legal Issue classification Legal Issue Area classification Legal Procedure classification Court Issue classification
MISHMORET	None	None
Open Laws Book	Date	Dating
Laws book	Date	Dating
Secondary Legislation	10 columns	Topic Modelling Summarization Document Type classification
Mutual	None	Dataset classification

Table 4.2: The metadata and some of the more interesting possible downstream tasks we found in each dataset with the labeled data. Worth noting that the Issue Court classification in the ISCD and the Topic Modelling in the Secondary Legislation dataset are quite similar, therefore comparable.

4.3 Model

To develop *Legaleph*, we further pre-trained AlephBERT for 20 epochs (similarly to Chau et al. [2020]) with 1 NVIDIA Tesla M60 GPU for 6 days. We did not train on the MISHMORET dataset as it was collected recently, and should be treated carefully since it contains a relatively large amount of similar texts. As a first comparison, in Figure 4.2 we evaluated the accuracy of Legaleph, AlephBERT, and M-BERT, in classifying by the datasets names depicted in Table 4.1 during training over 10% of the training data. First, we can see that Legaleph is more accurate than the AlephBERT and M-BERT. However, the other models seem to learn the domain gradually since the difference in performance decrease as they receive more training examples from the domain. These results may indicate that Legaleph is making better use of the labeled training data as it does not need to learn the domain from the annotated examples, and therefore it requires less annotated training data.

4.4 Future Work

In future research, we would like to test whether finetuning AlephBERT in a legal-domain dataset helps it to fine-tune faster while requiring less training data in different downstream tasks in the legal domain, as described in Table 4.2. To this end, we will measure the classification accuracy of AlephBERT before and after further pre-training as a function of epochs and the number of annotated instances the models were trained. Moreover, we can follow the systematic investigation scheme proposed by Chalkidis et al. [2020] to compare the Hebrew and English

Figure 4.2: Accuracy of classification by datasets over the development set when training over 10% of the training dataset as a function of training steps. Legaleph (further pre-trained AlephBERT) is marked in yellow line, AlephBERT is marked in purple line, and M-BERT marked in Cyan.

legal language models and examine the differences between the general language model and the specialized legal language model in each language.