German BERT Model for Legal Named Entity Recognition

Harshil Darji[a], Jelena Mitrović[b] and Michael Granitzer[c]

Chair of Data Science, University of Passau, Innstraße 41, 94032 Passau, Germany

Abstract: The use of BERT, one of the most popular language models, has led to improvements in many Natural Language Processing (NLP) tasks. One such task is Named Entity Recognition (NER) i.e. automatic identification of named entities such as location, person, organization, etc. from a given text. It is also an important base step for many NLP tasks such as information extraction and argumentation mining. Even though there is much research done on NER using BERT and other popular language models, the same is not explored in detail when it comes to Legal NLP or Legal Tech. Legal NLP applies various NLP techniques such as sentence similarity or NER specifically on legal data. There are only a handful of models for NER tasks using BERT language models, however, none of these are aimed at legal documents in German. In this paper, we fine-tune a popular BERT language model trained on German data (German BERT) on a Legal Entity Recognition (LER) dataset. To make sure our model is not overfitting, we performed a stratified 10-fold cross-validation. The results we achieve by fine-tuning German BERT on the LER dataset outperform the BiLSTM-CRF+ model used by the authors of the same LER dataset. Finally, we make the model openly available via HuggingFace.

1 INTRODUCTION

In NLP, NER is the automatic identification of named entities in unstructured data. These named entities are assigned to a set of semantic categories (Grishman and Sundheim, 1996), for example, for German Wikipedia and online news, such semantic categories are, *Location* (**LOC**), *Organization* (**ORG**), *Person* (**PER**), and *Other* (**OTH**) (Benikova et al., 2014). However, these named entities are not compatible with the legal domain because the legal domain also contains some domain-specific named entities such as judges, courts, court decisions, etc. A NER model fine-tuned on such domain-specific data improves the efficiency of researchers or employees working on such documents.

In the past couple of decades, there have been many improvements in terms of approaches being used for NER. From standard linear statistical models such as Hidden Markov Model (Mayfield et al., 2003; Morwal et al., 2012) to CRFs (Lafferty et al., 2001; Finkel et al., 2005; Benikova et al., 2015), RNNs (Chowdhury et al., 2018; Li et al., 2020), and

[a] https://orcid.org/0000-0002-8055-1376
[b] https://orcid.org/0000-0003-3220-8749
[c] https://orcid.org/0000-0003-3566-5507

BiLSTMs (Huang et al., 2015; Lample et al., 2016). However, the introduction of Transformers (Vaswani et al., 2017) gave rise to more efficient tools for NLP, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), etc. With these improved language models, there has been a significant improvement in terms of research in NER. Nowadays, many BERT language models take advantage of their underlying transformer approach to produce a specific BERT model fine-tuned for NER tasks in different languages (Souza et al., 2019; Labusch et al., 2019; Jia et al., 2020; Taher et al., 2020). There is also research done for BERT in the legal domain that uses BERT for various legal tasks such as topic modeling (Silveira et al., 2021), legal norm retrieval (Wehnert et al., 2021), and legal case retrieval (Shao et al., 2020).

However, when it comes to NER in the legal domain, it remains to be explored in detail, mainly due to the lack of uniform typology of named entities' semantic concepts and the lack of publicly available datasets with named entities annotations (Leitner et al., 2020). In 2019, Leitner et al. (Leitner et al., 2019) published their work on this concept that employs CRFs and BiLSTMs and achieved above 90% F1 scores. Later, they also published the dataset on which they managed to achieve these results. How-

ever, there is very minimal research done on using more efficient and improved language models, such as BERT or RoBERTa for the task of NER in the legal domain. In this paper, we use the same dataset to fine-tune a pre-trained German BERT model (Chan et al., 2020). This German BERT model is trained on 6 GB of German Wikipedia, 2.4 GB of OpenLegalData [1], and 3.6 GB of news articles. We make this fine-tuned model publicly available in the HuggingFace library[2].

2 RELATED WORK

Named Entity Recognition and Resolution on legal documents from US courts was performed by (Dozier et al., 2010). These documents consist of US case laws, depositions, pleadings, and other trial documents. The authors used lookup, context rules, and statistical models for the NER task. The lookup method simply creates a list of required named entities and tags all mentioned elements in the list as entities of the given type. This method is susceptible to false negatives due to a lack of contextual cues in lookup taggers. The contextual rules method takes into account the contextual cues, for example, in the legal context, a word sequence followed by "§" represents a law reference. However, this method requires a large dataset with manual annotations. Statistical models assign weights to cues based on their probabilities and statistical concepts. However, as with the contextual rules method, this also requires a large amount of manually annotated data. The authors developed taggers for Jurisdiction, Court, Title, Doctype, and Judge with F1 of 91.72, 84.70, 81.95, 82.42, and 83.01, respectively.

For German legal documents, legal NER work was performed by (Leitner et al., 2019). For this purpose, the authors created and published their own open-source dataset consisting of 67,000 sentences and 54,000 annotated entities. The authors used this dataset to train Conditional Random Fields (CRFs) and bidirectional Long-Short Term Memory Networks (BiLSTMs). Experimental results showed that BiLSTMs achieve an F1 score of 95.46 and 95.95 for the fine-grained and coarse-grained classes, respectively. In the same year, (Luz de Araujo et al., 2018) published their work on Named Entity Recognition in Brazilian Legal Text using LSTM-CRF on the LeNER-Br dataset and reported an F1 score of 97.04 and 88.82 for legislation and legal case entities, respectively. The authors created the LeNER-Br

dataset by collecting a total of 66 legal documents from Brazilian courts. To have a baseline performance, the authors first performed experiments on the Paramopama corpus (Júnior et al., 2015).

Based on the success of LSTM-CRF models, many researchers conducted experiments in different languages and reported state-of-the-art performance in their respective languages (Pais et al., ; Çetindağ et al., 2022). As can be seen in these works, a lot of research focused on NER tasks in the legal domain is done by BiLSTMs and CRFs. However, there is only a handful of research is done using transformer-based language models. The following research shows that in most cases these transformer-based language models outperform LSTM-CRF models.

The impact of intradomain fine-tuning of deep language models, namely ELMo (Sarzynska-Wawer et al., 2021) and BERT, for Legal NER in Portuguese was studied by (Bonifacio et al., 2020). The authors evaluated language models on three different NER tasks, HAREM (Freitas et al., 2010), LeNER-Br (Luz de Araujo et al., 2018), and DrugSeizures-Br[3]. As for the methodology of their experiments, the authors fine-tuned deep LMs pretrained on general-domain corpus on a legal-domain corpus, and supervised training was done on a NER task. The baseline for these experiments is achieved by skipping the fine-tuning process. Based on the experimental results, the authors conclude that legal-domain language models outperform general-domain language models in the case of LeNER-Br and DruSeizures-Br. It reduces the performance in the case of HAREM.

BERT-BiLSTM-CRF model, proposed by (Gu et al., 2020), first used a pre-trained BERT model to generate word vectors and then fed these vectors to a BiLSTM-CRF model for training. The dataset used by the authors consisted of over 2 million words with legal context. As stated by the authors, this data is collected from the People's Procuratorate case information disclosure network, the judgment document network, the Supreme People's Court trial business guidance cases, the public case published by the Supreme People's Court Gazette, and the judicial dictionary "Compilation of China's Current Law". The experiment results show their model outperforms BiLSTM, BiLSTM-CFR, and Radical-BILSTM-CRF with 88.86, 87.49, and 87.97 precision, recall, and F1-score, respectively.

(Aibek et al., 2020) developed a prototype of the "Smart Judge Assistant", SJA, recommender system. While developing this prototype, the authors faced the challenge of hiding the personal data of concerned

---

[1]https://de.openlegaldata.io/

[2]https://huggingface.co/

[3]The public agency for law enforcement and prosecution of crimes in the Brazilian state of Mato Grosso do Sul.

parties. To solve this problem, they used several NER models, namely CRF, LSTM with character embeddings, LSTM-CRF, and BERT, to extract personal information in Russian and Kazakh languages. Out of all five NER models used, BERT shows the highest F1 score of 87.

In addition to the above-mentioned research works, there also exist works in different languages (Souza et al., 2019; Zanuz and Rigo, 2022). However, to the best of our knowledge, there is yet to be any work done when it comes to developing a transformer-based language model, BERT, for the legal domain in the German language. Therefore, as stated in section 1, in this paper, we aim to use the dataset from (Leitner et al., 2020) to fine-tune a pretrained German BERT model.

3 DATASET

We use the Legal Entity Recognition (LER) dataset published by Leitner et al. in 2020. This dataset was constructed using texts gathered from the XML documents of 750 court decisions from 2017 and 2018 from "Rechtsprechung im Internet"[4]. It includes 107 documents from the following seven federal courts: Federal Labour Court (**BAG**), Federal Fiscal Court (**BFH**), Federal Court of Justice (**BGH**), Federal Patent Court (**BPatG**), Federal Social Court (**BSG**), Federal Constitutional Court (**BVerfG**), and Federal Administrative Court (**BVerwG**). This data was collected from *Mitwirkung*, *Titelzeile*, *Leitsatz*, *Tenor*, *Tatbestand*, *Entscheidungsgründe*, *Gründen*, *abweichende Meinung*, and *sonstiger Titel* XML elements of corresponding XML documents. As shown in Table 1, it contains a total of 66,723 sentences with 2,157,048 tokens, including punctuation.

Table 1: Dataset statistics (Leitner et al., 2020).

Court	Documents	Tokens	Sentences
BAG	107	343,065	12,791
BFH	107	276,233	8,522
BGH	108	177,835	5,858
BPatG	107	404,041	12,016
BSG	107	302,161	8,083
BVerfG	107	305,889	9,237
BVerwG	107	347,824	10,216
Total	**750**	**2,157,048**	**66,723**

This dataset comprises seven coarse-grained classes: *Person* (**PER**), *Location* (**LOC**), *Organization* (**ORG**), *Legal norm* (**NRM**), *Case-by-case regu-*

lation (**REG**), *Court decision* (**RS**), and *Legal literature* (**LIT**).

These seven coarse-grained classes are then further categorized into 19 fine-grained classes: *Person* (**PER**), *Judge* (**RR**), *Lawyer* (**AN**), *Country* (**LD**), *City* (**ST**), *Street* (**STR**), *Landscape* (**LDS**), *Organization* (**ORG**), *Company* (**UN**), *Institution* (**INN**), *Court* (**GRT**), *Brand* (**MRK**), *Law* (**GS**), *Ordinance* (**VO**), *European legal norm* (**EUN**), *Regulation* (**VS**), *Contract* (**VT**), *Court decision* (**RS**), *Legal literature* (**LIT**).

Table 1 in (Leitner et al., 2020) shows the distribution of both classes in the dataset. As shown in that table, classes related to the legal domain, namely *Legal norm*, *Case-by-case regulation*, *Court decision*, and *Legal literature* make up a total of 39,872 annotated NEs, which 74.34% of the total annotated NEs.

This dataset is publicly available[5] in CoNLL-2002 format (Sang and De Meulder, 2003). It follows IOB-tagging, where prefix *B-* denotes the beginning of the chunk, prefix *I-* denotes the inside of the chunk, and prefix *O-* denotes the outside of the chunk. In the legal context, consider Table 2:

Table 2: An example of IOB-tagging in the legal context. Here, GRT stands for *Court* and GS stands for *Law*.

Chunk	IOB-Tag
Das	O
Bundesarbeitsgericht	B-GRT
ist	O
gemäß	O
§	B-GS
9Abs.	I-GS
2Satz	I-GS
2ArbGG	I-GS
iVm.	O

4 EXPERIMENT

4.1 German BERT

The German BERT model was published by deepset[6] in 2019. As mentioned by the authors of this state-of-the-art BERT model for the German language, it "*significantly outperforms Google's multilingual BERT model on all 5 downstream NLP tasks we've evaluated*", namely, *germEval18Fine*[7],

---

[4]http://www.rechtsprechung-im-internet.de/

[5]https://github.com/elenanereiss/Legal-Entity-Recognition/tree/master/data

[6]https://www.deepset.ai/

[7]https://github.com/uds-lsv/GermEval-2018-Data

germEval18Coarse, *germEval14*[8], *CONLL03*[9], and *10kGNAD*[10]. Figure 1 shows the relative performance of all five downstream tasks on seven different model checkpoints for up to 840k training steps.
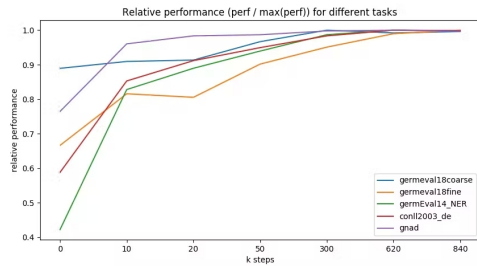

Figure 1: Relative performance of 5 different downstream tasks on 7 models checkpoints(deepset, 2020).

4.2 Fine-Tuning and Results

We compare the performance of our model with the BiLSTM-CRF+ model from (Leitner et al., 2019). To ensure that our model is generalized, i.e.that it does not rely on a portion of a dataset, we performed a stratified 10-fold cross-validation. During each cross-validation loop, we use one fold of the dataset as a validation set, while the remaining nine are used for training purposes. In each loop, we fine-tune the German BERT model for seven epochs. This cross-validation also helps in confirming that our model is not over-fitting.

Table 3 compares the individual performance scores of our model and BiLSTM-CRF+ model for each fine-grained class in the dataset. The reason for choosing the BiLSTM-CRF+ model to compare our performance to is because it has been proven to achieve better performance than CRFs for NER tasks on German legal documents (Leitner et al., 2019). The higher performance of our fine-tuned model can also be attributed to the fact that one of the datasets used for training the underlying German BERT model comes from OpenLegalData.

4.3 Published Fine-Tuned Model

Due to the satisfying performance of our fine-tuned German BERT model, we decided to open-source it on HuggingFace[11]. This model can be used with both popular frameworks, i.e. PyTorch[12] and Tensor-

---

[8]https://sites.google.com/site/germeval2014ner/data
[9]https://github.com/MaviccPRP/ger_ner_evals/tree/master/corpora/conll2003
[10]https://github.com/tblock/10kGNAD
[11]https://huggingface.co/harshildarji/gbert-legal-ner
[12]https://pytorch.org/

Flow[13]. HuggingFace also provides a "Hosted inference API"[14] that allows users to load and test a model in the browser. Figure 2 shows an example output of our model via this hosted interface API service.
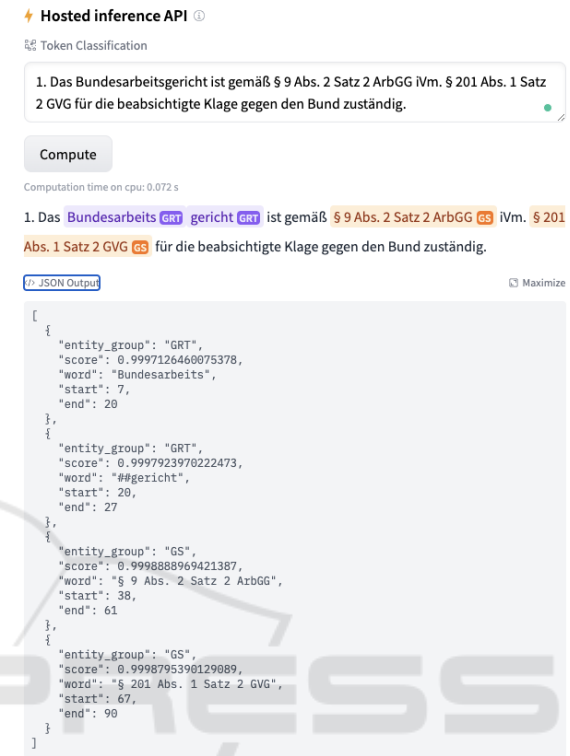

Figure 2: An example output of our German BERT for Legal NER model.

5 CONCLUSIONS

In order to fill the gap of having a proper Legal NER language model in the German language, we fine-tuned a state-of-the-art German BERT on the Legal Entity Recognition dataset on an Nvidia GeForce RTX GPU with a batch size of 64. It took 7 epochs for the fine-tuned model to achieve a very good performance.

If we look at the performance of individual fine-grained entities, in most cases, it outperforms the BiLSTM-CRF+ model used by the authors of the LER dataset. The only classes where our model significantly lags behind are *Country*, *Brand*, and *Landscape*. The performance on these classes can further be improved by having more examples of such instances in the dataset, as currently, only a couple of

---

[13]https://www.tensorflow.org/
[14]https://huggingface.co/docs/api-inference/index

Table 3: The performance of our fine-tuned German BERT model and the BiLSTM-CRF+ model for each individual fine-grained class.

Class	Our model			BiLSTM-CRF+		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Person	91.48	91.09	91.29	90.78	92.24	**91.45**
Judge	98.72	99.53	**99.12**	98.37	99.21	98.78
Lawyer	96.49	85.94	**90.91**	86.18	90.59	87.07
Country	92.51	94.2	93.34	96.52	96.81	**96.66**
City	88.21	89.92	**89.06**	82.58	89.06	85.60
Street	85.57	81.37	**83.42**	81.82	75.78	77.91
Landscape	68.49	68.49	68.49	78.50	80.20	**78.25**
Organization	89.11	92.22	**90.64**	82.70	80.18	81.28
Company	97.16	97.37	**97.27**	90.05	88.11	89.04
Institution	94.05	94.05	**94.05**	89.99	92.40	91.17
Court	97.3	98.02	97.66	97.72	98.24	**97.98**
Brand	81.86	54.57	65.49	83.04	76.25	**79.17**
Law	99.36	99.23	**99.29**	98.34	98.51	98.42
Ordinance	94.46	96.72	**95.58**	92.29	92.96	92.58
European legal norm	95.36	98.13	**96.73**	92.16	92.63	92.37
Regulation	89.94	87.99	**88.95**	85.14	78.87	81.63
Contract	96.52	95.08	**95.79**	92.00	92.64	92.31
Court decision	99.25	99.52	**99.39**	96.70	96.73	96.71
Legal literature	96.91	95.57	**96.24**	94.34	93.94	94.14

hundred of them exist compared to *Law* or *Court decisions*.

ACKNOWLEDGEMENTS

REFERENCES

Aibek, K., Bobur, M., Abay, B., and Hajiyev, F. (2020). Named entity recognition algorithms comparison for judicial text data. In *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–5. IEEE.

Benikova, D., Biemann, C., and Reznicek, M. (2014). Nosta-d named entity annotation for german: Guidelines and dataset. In *LREC*, pages 2524–2531.

Benikova, D., Muhie, S., Prabhakaran, Y., and Biemann, S. C. (2015). C.: Germaner: Free open german named entity recognition tool. In *In: Proc. GSCL-2015*. Citeseer.

Bonifacio, L. H., Vilela, P. A., Lobato, G. R., and Fernandes, E. R. (2020). A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In *Brazilian Conference on Intelligent Systems*, pages 648–662. Springer.

Çetindağ, C., Yazıcıoğlu, B., and Koç, A. (2022). Named-entity recognition in turkish legal texts. *Natural Language Engineering*, pages 1–28.

Chan, B., Schweter, S., and Möller, T. (2020). German's next language model. *arXiv preprint arXiv:2010.10906*.

Chowdhury, S., Dong, X., Qian, L., Li, X., Guan, Y., Yang, J., and Yu, Q. (2018). A multitask bi-directional rnn model for named entity recognition on chinese electronic medical records. *BMC bioinformatics*, 19(17):75–84.

deepset (2020). Open sourcing german bert model.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., and Wudali, R. (2010). Named entity recognition and resolution in legal text. In *Semantic Processing of Legal Texts*, pages 27–43. Springer.

Finkel, J. R., Grenager, T., and Manning, C. D. (2005). Incorporating non-local information into information

extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pages 363–370.

Freitas, C., Carvalho, P., Gonçalo Oliveira, H., Mota, C., and Santos, D. (2010). Second harem: advancing the state of the art of named entity recognition in portuguese. In *quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta 17-23 May de 2010) European Language Resources Association*. European Language Resources Association.

Grishman, R. and Sundheim, B. M. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Gu, L., Zhang, W., Wang, Y., Li, B., and Mao, S. (2020). Named entity recognition in judicial field based on bert-bilstm-crf model. In *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*, pages 170–174. IEEE.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Jia, C., Shi, Y., Yang, Q., and Zhang, Y. (2020). Entity enhanced bert pre-training for chinese ner. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6384–6396.

Júnior, C. M., Macedo, H., Bispo, T., Santos, F., Silva, N., and Barbosa, L. (2015). Paramopama: a brazilian-portuguese corpus for named entity recognition. *Encontro Nac. de Int. Artificial e Computacional*.

Labusch, K., Kulturbesitz, P., Neudecker, C., and Zellhöfer, D. (2019). Bert for named entity recognition in contemporary and historical german. In *Proceedings of the 15th conference on natural language processing*, pages 9–11.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Leitner, E., Rehm, G., and Moreno-Schneider, J. (2019). Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*, pages 272–287. Springer.

Leitner, E., Rehm, G., and Moreno-Schneider, J. (2020). A dataset of german legal documents for named entity recognition. *arXiv preprint arXiv:2003.13016*.

Li, J., Zhao, S., Yang, J., Huang, Z., Liu, B., Chen, S., Pan, H., and Wang, Q. (2020). Wcp-rnn: a novel rnn-based approach for bio-ner in chinese emrs. *The journal of supercomputing*, 76(3):1450–1467.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov,

V. (2019). Roberta: A robustly optimized bert pre-training approach. *arXiv preprint arXiv:1907.11692*.

Luz de Araujo, P. H., Campos, T. E. d., de Oliveira, R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). Lener-br: a dataset for named entity recognition in brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, pages 313–323. Springer.

Mayfield, J., McNamee, P., and Piatko, C. (2003). Named entity recognition using hundreds of thousands of features. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 184–187.

Morwal, S., Jahan, N., and Chopra, D. (2012). Named entity recognition using hidden markov model (hmm). *International Journal on Natural Language Computing (IJNLC) Vol*, 1.

Pais, V., Mitrofan, M., Gasan, C. L., Ianov, A., Ghit, C., Coneschi, V. S., and Onut, A. Legalnero: A linked corpus for named entity recognition in the romanian legal domain.

Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., and Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

Shao, Y., Mao, J., Liu, Y., Ma, W., Satoh, K., Zhang, M., and Ma, S. (2020). Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In *IJCAI*, pages 3501–3507.

Silveira, R., Fernandes, C., Neto, J. A. M., Furtado, V., and Pimentel Filho, J. E. (2021). Topic modelling of legal documents via legal-bert. *Proceedings http://ceur-ws org ISSN*, 1613:0073.

Souza, F., Nogueira, R., and Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Taher, E., Hoseini, S. A., and Shamsfard, M. (2020). Beheshti-ner: Persian named entity recognition using bert. *arXiv preprint arXiv:2003.08875*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wehnert, S., Sudhi, V., Dureja, S., Kutty, L., Shahania, S., and De Luca, E. W. (2021). Legal norm retrieval with variations of the bert model combined with tf-idf vectorization. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 285–294.

Zanuz, L. and Rigo, S. J. (2022). Fostering judiciary applications with new fine-tuned models for legal named entity recognition in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 219–229. Springer.