JurisBERT: Transformer-based model for embedding legal texts

1st Charles F. O. Viegas
Research and Development (R&D)
Juridics
Campo Grande, MS, Brazil
charles@juridics.com

2nd Bruno Catais Costa
Research and Development (R&D)
Juridics
Campo Grande, MS, Brazil
bruno@juridics.com

3rd Renato Porfirio Ishii
Faculty of Computing – FACOM
UFMS
Campo Grande, MS, Brazil
renato.ishii@ufms.br

***Abstract*—We propose in this paper a new extension of BERT (Bidirectional Encoder Representations from Transformers), called JurisBERT. It is applied in Semantic Textual Similarity (STS) and there is a considered improvement in fastness, in precision and it requires less computational resources than other approaches. JurisBERT was trained from scratch with specific domain texts to deal with laws, treatises, and precedents, and has better precision compared to other BERT models, which was our main finding of this work. Furthermore, our approach considers the concept of sublanguage, i.e., a model pre-trained in a language (Brazilian Portuguese) passes through refining (fine-tuning) to better attend to a specific domain, in our case, the legal field. JurisBERT includes $24,000$ pairs of `ementas` with degrees of similarity varying from 0 to 3. We extract these `ementas` from search mechanisms available on the courts' websites, in order to validate the approach with real data. Our experiments showed JurisBERT is better than other models in four scenarios: multilingual BERT and BERTimbau without fine-tuning in around 22% and 12% precision (F_1), respectively; and with fine-tuning in around 20% and 4%. Moreover, our approach reduced 5 times the training steps, besides using accessible hardware, i.e., low-cost GPGPU architectures. This result demonstrates that not always pre-trained models, such as BERT Multilingual and BERTimbau, which are heavy, require specialized and expensive hardware, are the best solution. So, we have proven that training the BERT from scratch with domain-specific texts has greater accuracy and shorter training time than large and general pre-trained models. The source code is available at https://github.com/juridics/brazilian-legal-text-dataset.**

***Index Terms*—Retrieving Legal Precedents, Semantic Textual Similarity, Sentence Embedding, BERT**

I. Introduction

Searching legal precedents is very important for legal professionals. They use it as a means for either supporting and strengthening their points or exposing opposing arguments. In Brazil, data from the *Conselho Nacional de Justiça*[1] [1] shows a large growth of legal proceedings, confirming that the Brazilian Judiciary System is overly congested, with a big amount of workload, and with an annual influx of millions of proceedings. In this scenario, an approach for efficiently retrieving precedents is very relevant for the Brazilian legal area.

In this context, several information retrieving applications are using methods for evaluating semantic similarities, a process that is in the Natural Language Processing (NLP) field and involves determining the similarity between two text segments. Recently, models based on Transformers [2] networks and big unlabeled datasets (e.g., BERT [3], RoBERTa [4]), are raising the bar in a lot of NLP tasks, including evaluating semantic textual similarity. Among some proposed approaches, the ones that stand out are Sentence BERT (sBERT) [5], that puts forward a change in the pre-trained BERT network and uses siamese and triple network structure. This is used to derive semantically relevant sentence embedding, which can be compared by cosine similarity. This approach does sentence embedding combined with indexing techniques such as FAISS [6]. It can deal with great amounts of data fastly and without losing the precision of transformers models.

However, most of the studies on transformers models are focused on the English language. This presents a challenge to bring technological advances to other languages, like Brazilian Portuguese. Even with popular models, such as BERT [3], having multi-language versions, models trained specifically with Brazilian Portuguese beat them, as shown by BERTimbau [7]. Besides, several works have revealed better results when pre-trained with domain specialized corpus [8], [9]. The BERTLaw study [10] shows that pre-training BERT with legal field specialized vocabulary have better results than using BERT Base. It exposes significant differences in unit vectors and that the intersection between both vocabularies is lower than half of the total vocabulary of each model separated. These differences directly affect the interpretation of a text segment by each model.

In this study, we propose an approach called JurisBERT, and we divided it on 3 steps: 1) pre-training BERT for Masked Language Modeling (MLM) starting from scratch and using legal field specific texts, containing laws, decisions, court votes, besides several legal treatises of the Brazilian law; 2) experiments with the sBERT network using as base pre-trained BERT networks, i.e., multilingual BERT (mBERT) and BERTimbau; and 3) fine-tuning sBERT models over our own dataset, which was prepared for evaluating the effectiveness of

Sponsored by Fundect, Fapesp, CNPq and CAPES Brazilian funding agencies.

[1]The Conselho Nacional de Justiça is a public institution that aims to help the Brazilian judiciary. It maintains administrative and procedural control and transparency.

our experiments and the similarities between the `ementas` of `acordãos`, these concepts will be explained in Section II-A. The developed dataset has 24k pairs of `ementas` with a degree of similarity ranging between 0 to 3 got from court websites. [2].

Our experiments showed JurisBERT is better than other models in four scenarios: multilingual BERT and BERTimbau without fine-tuning in around 22% and 12% precision (F_1), respectively; and with fine-tuning in around 20% and 4%. Moreover, our approach reduced 5 times the training steps, besides using accessible hardware, i.e., low-cost GPGPU architectures. This result demonstrates that not always pre-trained models, such as BERT Multilingual and BERTimbau, which are heavy, require specialized and expensive hardware, are the best solution. So, we have proven that training the BERT from scratch with domain-specific texts has greater accuracy and shorter training time than large and general pre-trained models.

We also noticed that sBERT with BERTimbau outperforms mBERT, validating that models with specific languages do better than the multilingual ones. Also, we could not find other public works of domain specialized corpus for evaluating text similarity in the Brazilian legal field. In that case, we believe that our research is the first to provide this data publicly and that our contributions will help the development of new studies in this field.

We organized this paper in the following sections: in Section II, we present the main concepts used in the other sections; in Section III, we discuss about other similar techniques that inspired us; in Section IV, we describe the steps of pre-training and dataset construction; in Section V, we cover the fine-tuning of the models and discuss the results; finally, in Section VI, we present the main contributions of our research and point out future works.

II. Background

A. Legal Precedent Retrieval

Legal precedents are used to substantiate arguments by attorneys and judges. They use it to reinforce the justification of their decisions [11]. In Brazil, after the Constitution amendment of 1998, precedents became more important since they started to have binding force over decisions made by the Brazilian Supreme Court [12]. For that reason, Brazilian courts have to provide public access to all its decisions over judged proceedings (except the ones that are classified as confidential). However, as the number of proceedings is big and keeps growing every year, more efficient solutions to retrieve precedents are in very high demand.

Worldwide, the retrieval of legal precedents is a very popular theme in the literature, especially the techniques for exploring semantic retrieval. They assist the better understanding of concepts related to contexts and the treatment of linguistic phenomenons, which affects the quality of the retrieval.

[2]STF, STJ, TJRJ, TJMS

In Brazil, the main document used as precedent is named `acordão`. It shows the decisions made by the judging court. Even though it does not have a standard format, most of the times it has the following sections: identification of the concerned parties, the judge who wrote the opinion, the discussed objects, the given facts, the court votes and the `ementa`, which is similar to the syllabus in the United States law. In Figure 1 we show an example of `ementa`. We can see a standard on the writing. In the superior part, the text is written in capital letter and in entry, while in the other parts, the text is written in enumerated paragraphs.

AGRAVO INTERNO – ART. 1.030, § 2º, CPC – ACÓRDÃO ESTADUAL QUE COINCIDE COM A ORIENTAÇÃO FIRMADA PELO SUPERIOR TRIBUNAL DE JUSTIÇA EM SEDE DE RECURSO REPRESENTATIVO DE CONTROVÉRSIA – LIMITAÇÃO DOS JUROS REMUNERATÓRIOS À TAXA MÉDIA DO MERCADO SOMENTE SE VERIFICADA ABUSIVIDADE – CAPITALIZAÇÃO MENSAL DOS JUROS PERMITIDA – EMBARGOS PROTELATÓRIOS – MULTA APLICADA – RECURSO IMPROVIDO. 1–As questões de direito enfrentadas e decididas nos recursos representativos da controvérsia guardam plena identidade ao posicionamento do Tribunal de Origem, pois somente será aplicável a limitação dos juros à taxa média do mercado em caso de comprovada desvantagem ao consumidor, demonstrando abusividade do fornecedor. 2–A capitalização mensal de juros é permitida nos contratos bancários desde que expressamente pactuada e celebrada após após 31.3.2000, ou que haja previsão de taxa de juros anual superior ao décuplo da mensal. 3–Recurso improvido.

Fig. 1. The `ementa` used for measuring similarity.

Empirically, we see a lot of legal professionals using only the `ementa` section of the `acordão` to decide which precedents to choose. This probably happens because the `ementa` summarizes the decision and can be enough to understand the whole `acordão`. Besides, it is very common to find in lawsuits the full transcription of the `ementa` used to reference the precedent. Such observations helped us make the choice to use the `ementa` as the source material for similarity comparison, which is the goal of our study.

B. Semantic Textual Similarity (STS)

To make precedent retrieving systems is necessary to use semantic similarity comparison techniques. In NLP, is proposed to use the STS task, which can be considered a regression task, to calculate the *score* that represents the similarity between two sentences. Given a collection of sentences and two sentences, the *score* will have higher values when the similarity is higher and lower values when the similarity is lower. In this area, the *International Workshop on Semantic Evaluation* (SemEval) [13] stands out promoting a series of researches on the NLP field to advance the semantic analysis and the creation of high quality datasets.

In Brazil, the first collection of public data that included semantic similarity between sentences in Portuguese was the ASSIN [14]. Years later, the ASSIN 2 [15] suggested a new data collection based on the SICK-BR [16] collection. However, neither of those collections specializes in legal texts.

C. BERT

BERT or *Bidirectional Encoder Representations from Transformers* is a language model used to pre-train deep bidirectional representations from unlabeled texts. It uses a bidirectional approach to model the context to the left and right of the entered sequence tokens. As a result, the pre-trained BERT model is adjusted with only one additional out layer to make models for NLP downstream tasks.

In Figure 2 we show the relations between BERT versions with pre-training and fine-tuning. The pre-trained version is the base for the fine-tuning versions that are adjusted to perform downstream tasks (e.g., STS, Entity Named Recognition, Text classification, etc.). The pre-training uses unlabeled data, while the fine-tuning uses labeled ones.

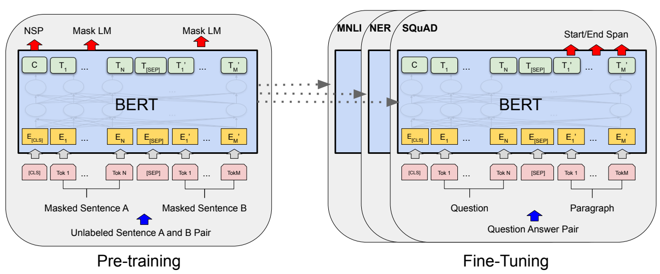

Fig. 2. BERT training process [3].

In the original BERT paper, there are two goals during pre-training: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). Within MLM, the number of random tokens in the input sequence are replaced by the special tokens [MASK] that are predicted using cross-entropy loss, 15% of the input tokens are evenly selected for possible replacements, of these tokens, 80% are in fact replaced by the token [MASK], 10% are unchanged and the remaining 10% are replaced by another random token of the vocabulary. In NSP, is used a binary classification loss to predict if two text segments follow one another in the original text. Positive examples are made with consecutive sentences and negative ones by paring text segments of different documents, both are created in equal proportions.

D. Transformers

The BERT architecture is based on the encoder part of the transformers [2] network architecture, which is considered a neural network of encoder-decoder type. The transformer network does not use neither Recurrent Neural Networks structures nor Convolution ones. Its main characteristics are being capable of reading sequential entries (from the left to the right or from the right to the left) in a single time. This characteristic allows for considering contexts from the right and the left. Also, it promotes better parallelization and requires less training time.

The transformer network is composed of an encoder that receives as input a sequence of words and transforms it into a vector sequence (internal representation). Next, a decoder, out-of the internal representation, makes a sequence of words, one by one. To learn more about the transformer architecture, read the original paper [2].

E. Sentence BERT

Even though BERT models have been raising the bar in evaluating semantic similarity, they use a cross-encoder where both text sentences are sent to the transformer network. This produces a huge computational cost. For example, it would be necessary roughly 50 million computational inferences (65 processing hours) to find the most similar sentences in a collection of $10,000$. This turns BERT into an unviable option for information retrieval systems.

A way to address this type of problem is to map each sentence in a vector space, where semantic similar ones lay close together. For this, fixed sentence embeddings can be gotten by sending individual sentences to the BERT network. The most used technique gets the fixed vector embeddings from the average of values generated in the output layer of BERT (known as BERT embeddings) or using the output of the first token (or [CLS] token), though both techniques have lower performances than older ones like GloVe embeddings [17].

To fix this problem, sBERT was developed. It uses siamese network architecture, which means using two identical networks with shared weights. As shown in Figure 3, sBERT allows for deriving fixed size vectors that, using a similarity measurement (e.g., cosine similarity and Manhattan distance), can calculate the similarity between two sentences. To make sure that the generated embeddings have fixed sizes, there is a pooling [3] operation on the BERT output.

sBERT is computationally efficient. In the previously discussed example, the authors say that sBERT can reduce the computational cost to find the most similar pair of sentences, from 65 hours to approximately 5 seconds.

III. Related works

sBERT is basically a sentence embedding technique, a field with several studies and methods proposed. Unsupervised methods based on encoder-decoder techniques looks to be dominating recent researches. Skip-Thought [18] proposes a model with encoder-decoder architecture that uses neighbour sentences to codify the embedding. However, this model needs a corpus made with continuous texts for training. The Universal Sentence Encoder [19] suggests a general model that uses the encoder part of the transformer network, it utilizes the attention mechanism to compute the words that are sensitive to the context, that way, getting the sentence embedding from the average of the internal state of the codified tokens. It is also suggested that transferring knowledge in the level of sentence is more efficient than in the level of words. More recently,

[3]It is an operation that reduces the dimensionality of data by applying an aggregation of type max average.

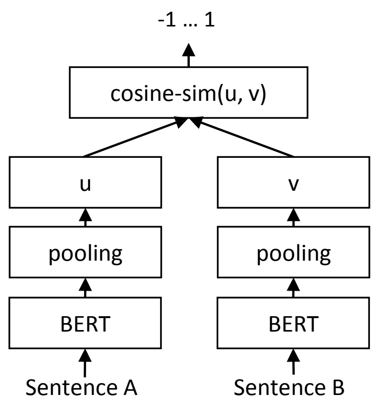

Fig. 3. sBERT architecture [5].

PromptBERT [20] applies prompts to reduce biases found in the internal state of the tokens for making the original BERT layers more effective.

In [21], the authors conclude transformers model pre-trained with domain specific contexts performs better than general models. This suggests that the best approach is merging both general and domain specific models, by continuing pre-training using the last checkpoint of the general model in domain specific corpus. The corpus size was compared and concluded that it makes little difference between themselves.

IV. JurisBERT

In our work, we propose an approach for semantic textual similarity of the `ementas` of `acordãos` with a sBERT network, called JurisBERT. It puts forward a pre-training BERT with Brazilian legal field domain-specific texts. For this, we constructed two corpus (for training and fine-tuning), discussed in more detail in Sections IV-A and V. For evaluating the experiments we have two steps: first, we iterate over each model with all the analysed `ementas` to create the embedding vectors [4]. Second, we calculate the degree of similarity between each pair of embeddings through the method of cosine similarity (as suggested by the authors of sBERT [5], [22]). The score got from the cosine similarity can only have values between 0 and 1, the closer to 1, the higher the similarity and the closer to 0, the lower the similarity. We stabilised a threshold number to optimize the division into two groups. That way, the score values over the threshold are similar and the ones under are not similar.

The approach chosen in this work is based on the sBERT model that needs a pre-trained BERT model to generate the embedding vectors. The choice of pre-trained models to be used in the performance comparison with JurisBERT, mainly took into account the compatibility with the Brazilian Portuguese language, which is the language of the `ementas` of `acordãos`. In addition, we only use publicly available models on huggingface [5] website, which is a very popular open source community that provides a repository for publishing pre-trained models, as well as offers a set of tools for training, evaluating and publishing transformers models. Therefore, considering these criteria, models RoBERTa [4], BERTLaw [10] and LegalBERT [23] were excluded from the evaluation, as all of these are specifically intended for the English language, thus leaving the multilingual BERT (mBERT) and BERTimbau.

According to the authors, these two models chosen for evaluation were pre-trained with general domain texts and with a much larger corpus than JurisBERT. The mBERT was pre-trained with 104 different languages, provided by the authors of BERT. BERTimbau, in turn, was pre-trained in Brazilian Portuguese with data from BrWac [24], which is a corpus got from Brazilian websites. Our approach made two variations of pre-training: one from scratch, called JurisBERT, and the other, coming from the BERTimbau checkpoint, we apply further pre-training with legal texts called BERTimbau further. The goal is to determine if there are significant differences between both methods.

All BERT models are based on the $\text{BERT}_{\text{BASE}}$ architecture (L=12, H=768, A=12, Total parameters=110M). We chose this architecture because, even though it is not the most performative, it has a lightweight model that is more suitable for our hardware.

A. *Pre-training corpus*

In order to pre-train JurisBERT, we had to create our own corpus, because we could not find one done with the amount of data and variety needed. Then, the first step was developing web scrappers to read and retrieve public documents from many sources, for example: court websites, public agencies and universities. During the construction of this corpus, we aimed for the primary sources of the legal field, such as laws, precedents, treatises, analogies, general principles of law and equity [25]. So, among the retrieved documents are laws and federal decrees, `súmulas`[6], decisions, `acordãos` and court votes, besides treatises of different legal fields. In Figure 4, we show a chart with the participation of each type of legal source. The law predomination happens because of the plentifulness of this kind of document on governmental websites.

After retrieving the documents, we pre-processed them to remove special characters, excesses of spaces and blank lines, though we maintained the accentuation and the letter casing, because the networks we used are case sensitive. Also, we split the documents in paragraphs and dispose of the ones with less than 10 tokens and the duplicated ones. We made this choice, intending to make the paragraphs more similar to the average size of the `ementa`. The result we got was 1.5 million of sentences(paragraphs), 99% of those with less than 384 tokens, as shown in Figure 5, repeating a proportion like the one in the corpus of the *fine-tuning*, detailed in Section V.

[4]It is a dense vector of floating points that aims to capture the semantic of the text in the vector space.

[5]https://huggingface.co/

[6]The `súmulas` summarizes the dominant precedent of a given court.

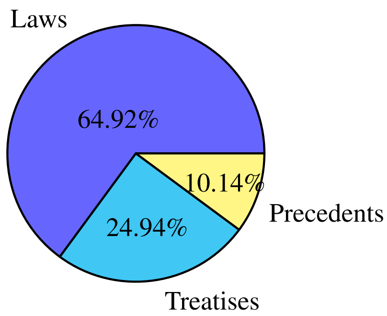

Fig. 4. Participation of each type of legal source in the pre-training corpus.

After the pre-processing, the corpus reached 410MB of raw text, a significantly lower number than the ones of BERTimbau (17GB) and RoBERTa (160GB). Then, we divided the corpus in two parts: training (95%) and evaluation (5%).

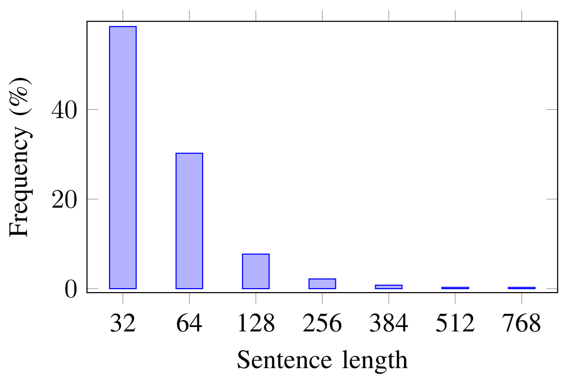

Fig. 5. Frequency of each sentence length.

B. Vocabulary generation

For training JurisBERT, we generated an uncased vocabulary of 30k units of sub-words using the training corpus. The chosen tokenizer was WordPiece [26] of the huggingface library.

C. Pre-training

We used only the MLM goal in the pre-training, since recent papers [4] have suggested that the NSP goal is not effective. During training, we optimized JurisBERT with Adam [27] using the following parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-6$ e L_2 weight decay of 0.01. The learning rate in the first 10,000 steps with apex of $1e-4$ and linear decay. We also used a *dropout* of 0.1 in all levels, alongside GELU's [28] activation function. The model was pre-trained with 20 epochs and 220k steps, 5 times smaller than the BERTimbau, i.e., 1M steps, with batch size of 128 and sequences with maximum length of 384 tokens. We used two `Nvidia GeForce RTX 3080` with 12 GB `GDDR6X` each in a `ThinkServer RD350` server equipped with 2x processors `E5 2620 V3`, 128 GB RAM, and 5 TB HD NAS.

V. Evaluating Semantic Textual Similarity

We evaluated the performance of the models after their fine-tuning. This step is essential for getting better results. For this, we also needed to develop our own dataset composed by many paired sentences showing whether they were similar. The term sentence, in our study, means the content of the `ementa` of an `acordão`, which usually is made of few paragraphs.

We constructed our dataset through search mechanisms available in the websites of the courts: STF, STJ, TJRJ and TJMS. There, the courts provide one or more `acordãos` for each theme considered stable, which means that it has a standard understanding in the legal context. They also provide notorious cases such as racial quotas, use of embryonic stem cells and party infidelity.

In particular, STF's search mechanism has an interesting characteristic to define degrees of similarity, it is organized hierarchically at three levels of grouping. The first level is divided by **fields** of law, such as: Administrative, Civil, Constitutional, Electoral, Criminal, Retirement, Civil Procedural, Criminal Procedural and Tax. The second by **themes**, for example, in the field of Administrative Law, the divided themes are Public Job Applications, Pharmaceutical Assistance Programs, Liability of Public Administration and Public Sector Employees. The third and final level is divided by **legal discussions**, for example, in the Public Sector Employees theme we got the following discussions: Teacher's Special Retirement, Public Employee Payment Discount for Striking, Vested Right of Probationary Period, Judgment Deadline for Legality of Retirement.

So, to STF's search mechanism, we applied a scale of similarity between `acordãos` considering the group hierarchy, because `acordãos` inside the same discussion are more similar than `acordãos` of different themes and fields. Thus, we automatically annotated `acordãos` of the same discussion with similarity of 3, `acordãos` of different discussion but of the same theme with similarity of 2, `acordãos` of different themes but of the same field with similarity of 1 and `acordãos` of different fields with similarity of 0.

Applying this strategy with STF, we got a total of 24,926 pairs of `acordãos`, 80% being dedicated to training and 20% to evaluate training. The other search mechanisms (STJ, TJRJ and TJMS) got only a single level of grouping, we only used those for testing and comparing models. The strategy used for automatically annotating these searches was considering `acordãos` of the same group as similarity of 1 and the rest as similarity of 0. Therefore, we generated 19,027 pairs from the STJ, 8,626 from the TJRJ and 6,305 from the TJMS.

A. Fine-tuning

In the beginning, we had defined the maximum sequence length after analysing the corpus; it showed that 91.33% of the `ementa` texts had length lower than 384 tokens. However, we trained models with low epochs varying only the maximum sequence length to validate this hypothesis. As shown in Table I, the value 384 really had better results.

TABLE I
EXPERIMENTS VARYING THE MAXIMUM SEQUENCE LENGTH.

Sequence length	$F_1(\%)$
64	71.30
128	76.72
256	81.93
384	**83.27**
512	82.70

After having defined the maximum sequence length, we made the fine-tuning training with 3 epochs, hyper-parameters, batch size and maximum sequence length of 8 and 384, respectively. The cosine similarity loss was used. The training took roughly 3 hours using the same hardware as described in Section IV-C. To measure and compare our experiments, we considered only the checkpoint of the models that got the best performance during the training epochs.

B. Threshold definition

An important step to maximize the F_1 metric is to define the threshold value that separates which examples should be considered similar or not. The strategy for this choice is basically to test as a threshold value the score obtained by comparing the cosines similarity of each pair of samples contained in the dataset, i.e., a dataset D, containing `ementas` E_n, organized in pairs $\{(E_1, E_2), (E_3, E_4), (E_5, E_6)\}$, annotated with the labels $\{1, 0, 1\}$, and the cosines similarity $\{0.95, 0.25, 0.57\}$, when testing every previous values we have the following measures of F_1 $\{0.66, 0.40, 1.0\}$, therefore, for this example, the best threshold value is 0.57.

C. Discussions

We evaluated the models considering the $F_1(\%)$ metric with their respective threshold (Thr), as shown in Table II. In general, we can see that all the models that went through fine-tuning performed better than the ones that did not. These results suggest that our approach to construct the dataset and make the fine-tuning was, in fact, effective. Also, we can notice that models pre-trained with specific domain texts of a sublanguage, in other words, texts specialized in a specific domain, performed better than the other ones. Models BERTimbau further and JurisBERT showed better general results. This means that there are advantages in training from scratch and in doing more training to pre-existing models. Our training approach proved effective even with a corpus size 42 times smaller and with 5 times less pre-training steps than other methods.

In that regard, we can say that our work address a very common problem in NLP, dealing with domain specific topics (healthcare, law and others). These have plenty of data pulverized through different mediums, but few of those are annotated for a specific task. Even without fine-tuning, our approach proved a viable option, beating even fine-tuned mBERT and BERTimbau. As shown in Table II, JurisBERT without fine-tuning obtained 68.49% of F_1, outperformed by

TABLE II
COMPARING $F_1(\%)$ RESULTS WITH THRESHOLD (THR) ACCORDINGLY.

Model	TJMS		TJRJ		STJ		Mean
	F_1	Thr.	F_1	Thr.	F_1	Thr.	F_1
No Fine-tuning							
mBERT	65.05	0.72	63.78	0.47	30.39	0.94	53.08
BERTimbau	71.20	0.94	63.88	0.67	35.55	0.95	56.88
Further	75.48	0.92	64.06	0.65	41.75	0.94	60.43
JurisBERT	**80.25**	0.79	**73.21**	0.83	**52.00**	0.87	**68.49**
Fine-tuning							
mBERT	69.51	0.26	64.21	0.11	39.49	0.69	57.73
BERTimbau	81.77	0.53	71.34	0.56	50.75	0.66	67.95
Further	84.39	0.56	73.46	0.55	47.37	0.68	68.41
JurisBERT	**85.16**	0.60	**77.65**	0.66	**50.95**	0.71	**71.25**

10.76% the fine-tuned mBERT and 0.54% more than fine-tuned BERTimbau. These results infer it is also possible to have good performances in other downstream tasks.

Experiments evidence the JurisBERT is better than other models considering four scenarios: multilingual BERT and BERTimbau without fine-tuning in around 22% and 12% precision (F_1), respectively; and with fine-tuning in around 20% and 4% as showed in Table II. So, we have proven that training the BERT from scratch with domain-specific texts has greater accuracy and shorter training time than large and general pre-trained models.

Our pre-training had a hardware cost lower than the other models, the training time was lower and the graphics card used is inferior to those used in BERTimbau and mBERT. The `P100` graphics card, by `Nvidia`, is more expensive than the `RTX` used in our study. So, we showed it is technically and financially viable to replicate these experiments in other sublanguage domains, considering the accessible price of current `RTX` models. For instance, the price of a graphic card used in our study is about $\$1,000$ dollars.

VI. CONCLUSION

This work's main contribution is the creation and availability of a corpus for unsupervised pre-training, including many laws, treatises, and decisions of several branches of Brazilian law. Furthermore, our approach used 5 times less steps than other approaches for pre-training, as well as, it can be deployed in a low-cost environment considering a usual graphics card, i.e., with a price of less than $\$1.000$. Also, we made a dataset for evaluating the similarity between legal decisions. Using this dataset alongside the data from `acordãos`, it contributes to many other goals, like clustering, topic modeling and classification tasks. Further, the source code of the web-scrappers and parsers we used in the construction of the datasets is available in a public repository[7].

Finally, this study shows that the sBERT model pre-trained and refined in other language and in specific domain performs better than general domain ones, either multilingual or native language. Thus, it confirms the hypothesis that pre-training

[7]https://github.com/juridics/brazilian-legal-text-dataset

with domain-specific texts has better performances in evaluating semantic similarity. Also, we can see that, even without fine-tuning, the models pre-trained with a domain specific corpus performed better than general domain fined-tuned ones. Besides, we showed the viability of using accessible hardware for pre-training, opening the path for other possibilities in this type of approach in other sub-language domains.

References

[1] CNJ, "Justiça em números 2020: ano-base 2019," Conselho Nacional de Justiça, Tech. Rep., 2020. [Online]. Available: https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[5] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[6] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," 2017. [Online]. Available: https://arxiv.org/abs/1702.08734

[7] F. Souza, R. Nogueira, and R. Lotufo, "Bertimbau: Pretrained bert models for brazilian portuguese," in *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 403–417.

[8] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction," 2020. [Online]. Available: https://arxiv.org/abs/2005.12833

[9] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *EMNLP*, 2019.

[10] H.-T. Nguyen and L.-M. Nguyen, "Sublanguage: A Serious Issue Affects Pretrained Models in Legal Domain," *arXiv e-prints*, p. arXiv:2104.07782, Apr. 2021.

[11] R. Weber, "Intelligent jurisprudence research: A new concept," in *Proceedings of the 7th International Conference on Artificial Intelligence and Law*, ser. ICAIL '99. New York, NY, USA: Association for Computing Machinery, 1999, p. 164–172. [Online]. Available: https://doi.org/10.1145/323706.323791

[12] H. Júnior, *Curso de direito processual civil*. Editora Forense, 2019, vol. I.

[13] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "SemEval-2012 task 6: A pilot on semantic textual similarity," in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, 7-8 Jun. 2012, pp. 385–393. [Online]. Available: https://aclanthology.org/S12-1051

[14] E. R. Fonseca, L. Borges dos Santos, M. Criscuolo, and S. M. Aluísio, "Visão geral da avaliação de similaridade semântica e inferência textual," *Linguamática*, vol. 8, no. 2, pp. 3–13, Dez. 2016. [Online]. Available: https://linguamatica.com/index.php/linguamatica/article/view/v8n2-1

[15] L. Real, E. Fonseca, and H. Gonçalo Oliveira, "The assin 2 shared task: A quick overview," in *Computational Processing of the Portuguese Language*, P. Quaresma, R. Vieira, S. Aluísio, H. Moniz, F. Batista, and T. Gonçalves, Eds. Cham: Springer International Publishing, 2020, pp. 406–412.

[16] L. Real, A. Rodrigues, A. Vieira, B. Albiero, B. Thalenberg, B. Guide, C. Silva, G. Lima, I. Câmara, M. Stanojević, R. Souza, and V. De Paiva, *SICK-BR: A Portuguese Corpus for Inference: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*. Springer, Cham, 01 2018, pp. 303–312.

[17] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[18] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, "Skip-thought vectors," 2015. [Online]. Available: https://arxiv.org/abs/1506.06726

[19] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil, "Universal sentence encoder for English," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 169–174. [Online]. Available: https://aclanthology.org/D18-2029

[20] T. Jiang, S. Huang, Z. Zhang, D. Wang, F. Zhuang, F. Wei, H. Huang, L. Zhang, and Q. Zhang, "Promptbert: Improving bert sentence embeddings with prompts," 2022. [Online]. Available: https://arxiv.org/abs/2201.04337

[21] G. Zhang, D. Lillis, and P. Nulty, "Can Domain Pre-training Help Interdisciplinary Researchers from Data Annotation Poverty? A Case Study of Legal Argument Mining with BERT-based Transformers," in *Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*. Association for Computational Linguistics, 2021, pp. 121–130. [Online]. Available: https://rootroo.com/downloads/nlp4dh_proceedings_draft.pdf

[22] H. Choi, J. Kim, S. Joe, and Y. Gwon, "Evaluation of bert and albert sentence embedding performance on downstream nlp tasks," 2021. [Online]. Available: https://arxiv.org/abs/2101.10642

[23] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho, "When does pretraining help? assessing self-supervised learning for law and the casehold dataset," in *Proceedings of the 18th International Conference on Artificial Intelligence and Law*. Association for Computing Machinery, 2021.

[24] J. A. Wagner Filho, R. Wilkens, M. Idiart, and A. Villavicencio, "The brwac corpus: A new open resource for brazilian portuguese," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[25] A. S. Cunha, *Introdução ao estudo do direito*. Saraiva, 2012.

[26] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou, "Fast wordpiece tokenization," 2020. [Online]. Available: https://arxiv.org/abs/2012.15524

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[28] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2016. [Online]. Available: https://arxiv.org/abs/1606.08415