Text Classification in Law Area: a Systematic Review

V. S. Martins[1], C. D. Silva[1]

Universidade Federal do Pará (UFPA)
Programa de Pós-graduação em Computação Aplicada (PPCA)
Campus Universitário de Tucuruí - Pará - Brazil
{victorsm,cleison}@ufpa.br

Abstract. This article is an extension of the KDMile 2021 accepeted submission. Automatic Text Classification represents a great improvement in law area workflow, mainly in the migration of physical to electronic lawsuits. A systematic review of studies on text classification in the legal context from January 2017 up to February 2021 was conducted. The search strategy identified 20 studies, that were analyzed and compared. The review investigates from research questions: what are the state-of-art language models (LM); LM applications on text classification in English and Brazilian Portuguese datasets from legal area; if there are available language models pre-trained on Brazilian Portuguese; and datasets from the Brazilian judicial context. It concludes that there are applications of automatic text classification in Brazil, although there is a gap on the use of language models when compared with English language dataset studies, also the importance of language model in domain pre-training to improve results, as well as there are two studies making available Brazilian Portuguese language models, and one introducing a dataset in Brazilian law area.

Categories and Subject Descriptors: I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: Law, Text Classification

1. INTRODUCTION

The application of Natural Language Processing (NLP) in the Legal area has promoted several gains for Justice Systems, which in Brazil gains huge contours given the number of processes in progress, which, according to sources from the National Council of Justice (*Conselho Nacional de Justiça* in Portuguese) in its annual report "Justice in Numbers"[1], referring to the base year of 2019, was 77.1 million. The possibility of including in the workflow of the courts, law firms, police stations, and Public Prosecutors, tools that help in the process of analysis, classification, and search of these documents, brings relative gains of speed, even more, when considering the public sector and the decreasing availability of human resources, in contrast to the growing increase in procedural demand.

Among the applications in the legal context, there is the classification of documents to migrating from physical to electronic processes [Silva and Maia 2020], or even for the correct direction of demands filed electronically, as analyzed in [Noguti et al. 2020] and [Mota et al. 2020]. These activities require human resources and working hours that can be allocated to other areas with more cognitive needs, enabling the promotion of greater speed in the Brazilian Justice System. Figure 1 shows a general flow of the textual classification process.

In this context, this work analyzes publications between January 2017 and February 2021 that apply

[1]https://www.cnj.jus.br/wp-content/uploads/2020/08/WEB-V3-Justiça-em-Números-2020-atualizado-em-25-08-2020.pdf

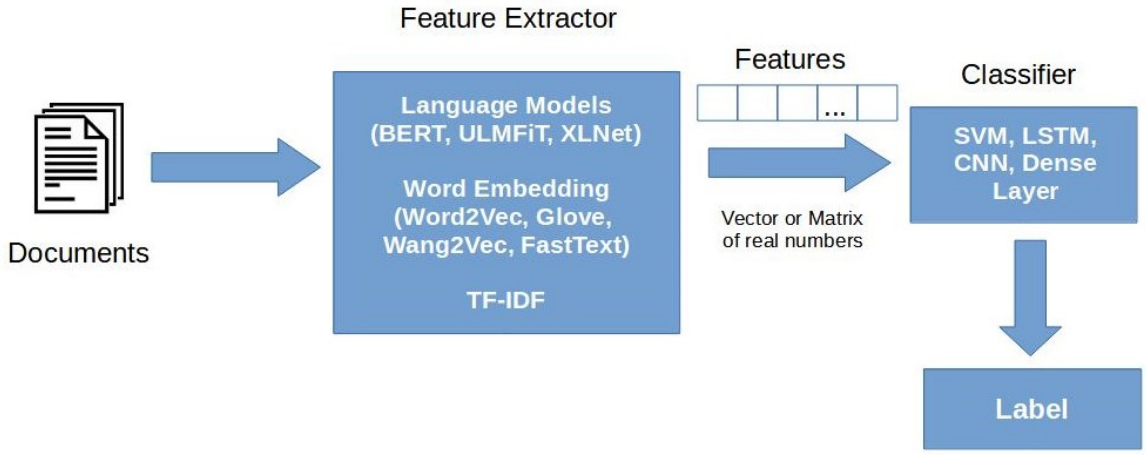

Fig. 1. General flow of a textual classification process. Source: Authors.

NLP techniques for textual classification in the legal contexts of Brazil and other countries, through a systematic review of the topic, bringing essential elements for researchers in the area to start studying the application of deep learning techniques in the task of textual classification in the legal context, being a starting point for the understanding of what is being studied and applied in it.

Section 2 shows the research protocol used in the review. In Sections 3 to 6.4, critical analyzes and comparisons between the selected works, according to the respective review questions. Finally, in Sections 7 and 8, the main gaps and conclusions on the topic after the criticisms and comparisons, respectively.

2. REVIEW PROTOCOL

The review protocol was constructed to answer the following review questions:

Q.1 What is the state of the art in the task of textual classification in NLP using pre-trained models?

Q.2 Are there scientific publications with pre-trained language models used in the legal domain?

Q.3 What NLP techniques in the Portuguese language domain have been applied to classify legal texts in Brazil?

Q.4 Are there public data repositories with texts in the Brazilian legal context?

Q.5 Is there a repository of pre-trained language models for Brazilian Portuguese?

2.1 Search Strings

To collect the analyzed papers, four search strings were defined with the keywords from Table I. The construction of each string aims to answer a specific question from the research questions.

As a reference for **Q.1**, the site "NLP-Progress: Tracking Progress in Natural Language Processing (Text Classification)" was used specifically in its Textual classification area[2], which maintains a ranking

[2]http://nlpprogress.com/english/text_classification.html

Table I. Keywords grouped to search in academic databases. Source: Authors.

Grouping	Keywords	String
1	Classification, labeling, label	classification OR label
2	Machine Learning, Deep Learning, Neural Network	"machine learning" OR "deep learning" OR "neural network"
3	Language Model, pre-trained, transfer learning	"language model" OR pre-trained OR "transfer learning"
4	denunciation, crime report, petition, legal, law	denunciation OR "crime report" OR petition OR legal OR law
5	Dataset, corpus	dataset OR corpus
6	Portuguese, Brazil	portuguese OR brazil
7	Document, text	document OR text
8	Natural Language	"natural language"

that compares the results in this task for different machine learning models and architectures, to bring together the state of the art in the area. Thus, the three studies that presented the best results in the month of collection, March 2021, were selected for review.

To answer the questions **Q.2** to **Q.2**, Table I shows the keywords used to assemble the search strings applied in the following academic databases: Capes Journal Portal, IEEE Xplore, Science Direct – Elsevier, Scopus, Google Scholar, and ACL-Web (Association for Computational Linguistics).

The searches included only complete articles in English or Portuguese, using the following strings for each question, assembled from the groupings of keywords in Table I:

—**Q.2**: 1 AND 3 AND 4 AND 7 AND 8

—**Q.3**: 1 AND 2 AND 4 AND 7 AND 6 AND 8

—**Q.4**: 5 AND 6 AND 8 AND 4

—**Q.5**: 3 AND 6 AND 8

All the results of the searched databases were collected and stored in spreadsheets, except for Google Scholar, which after limiting the period and defining the ordering of the result by importance, the first 20 pages of search results were collected, which ended in around 200 articles per string.

The articles found were gathered in a list, totaling 1,432, which after removing the duplicates, resulted in 1,142 papers.

2.2 Selection of articles for review

After reading the titles, and considering the inclusion, exclusion, and quality criteria, according to Tables II to 2.2, 85 articles resulted. Among these 85, those that had titles that raised doubts about the content were briefly examined, which generated 30 articles, and of these, after a more detailed reading, and once again applying the criteria, the 20 titles used in the review, indicated with an asterisk (*) in the bibliographic references. Figure 2 shows the article selection process flow.

The criteria ensure that the review meets the Research Questions, while at the same time restricting the search on NLP models that involve Textual Classification or similar tasks such as: sentiment analysis and named entity recognition, which use deep learning with or without pre-trained language

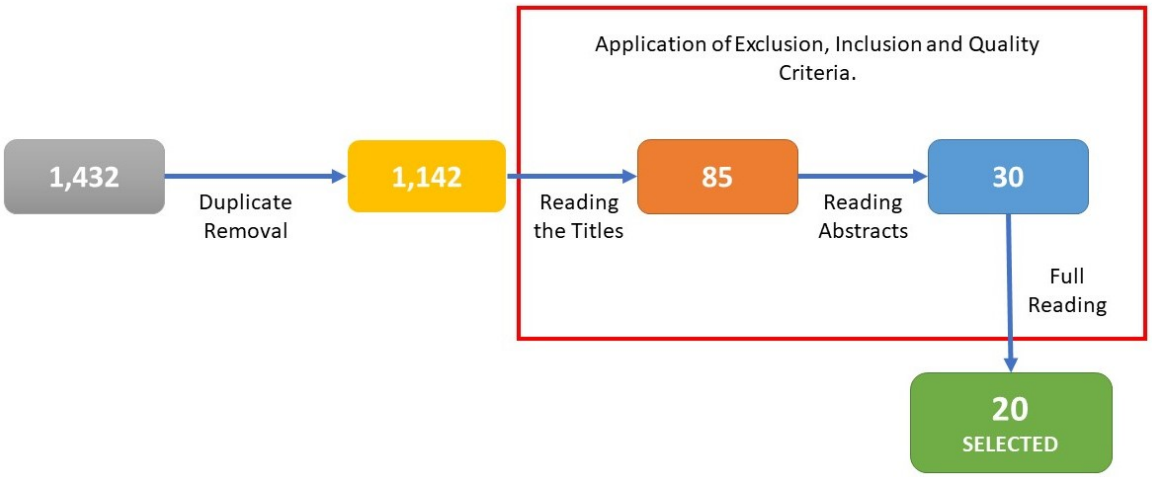

Fig. 2. Selection flow of analyzed articles. The boxes indicate the number of articles in each stage. Source: Authors.

Table II. Quality Criteria applied in the selection of articles.

Quality Criteria
The base language model used in the article is listed and among the top 5 on the NLP-Progress website.
Clear Description of datasets/corpus collection procedures
Comparison with previous state-of-the-art models
Include a study of legal documents classification in Portuguese using deep learning models. To answer questions **Q.3** and **Q.5**.

Table III. Inclusion Criteria applied in the selection of articles.

Inclusion Criteria
Text Classification, Sentiment Analysis, or Recognition of Named Entity in English or Portuguese.
Studies published in English or Portuguese from 2017 .
Present comparison data with one of the indices: accuracy, precision, recall, or F1-score.
It uses Datasets from the legal domain, for the questions **Q.2**, **Q.3**, and **Q.4**, or benchmarks widely used for tests, for **Q.1**.

Table IV. Exclusion Criteria applied in the selection of articles.

Exclusion Criteria
It does not use any deep learning technique in the studied models.
Bibliographic Abstracts.
Articles with duplicate studies.

models for studies with *datasets* in Portuguese, and in those with documents in English, only those that proposes an architecture including LM.

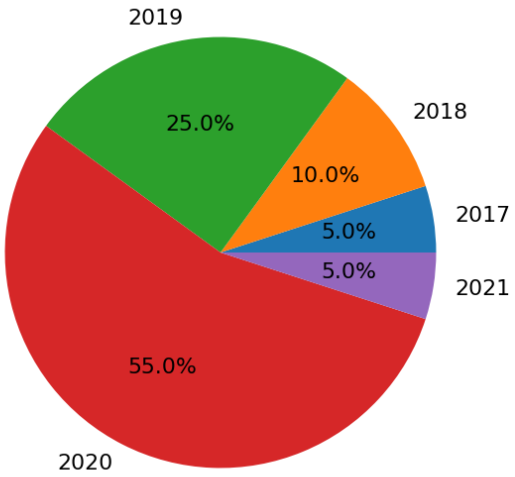

Fig. 3. Number of Articles Published per year. Source: Authors.

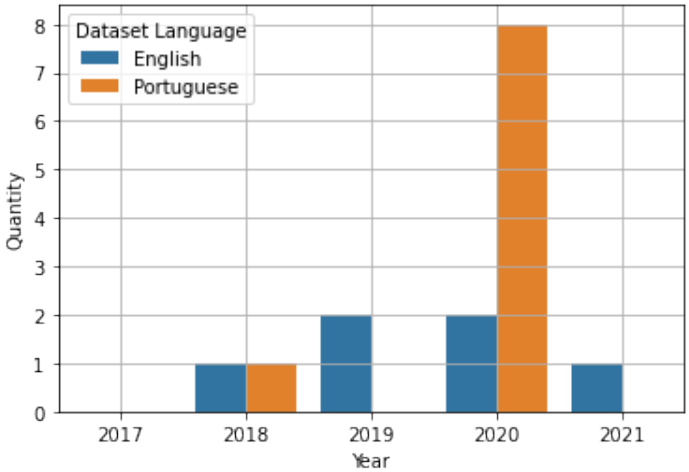

Fig. 4. Number of Articles Published per year in the legal field. Source: Authors.

For a better understanding of the selected articles, as well as the recent application of deep learning techniques in the legal field, Figures 4 and 4 show, respectively, the number of articles published by year among those selected for review, and the proportion that specifically address the legal domain by language. In Sections 3 to 6 these studies are analyzed in order to answer each of the research questions.

3. STATE-OF-THE-ART LANGUAGE MODELS IN TEXTUAL CLASSIFICATION (Q.1)

This Section is intended to answer Question 1 of the review questions: "What is the state of the art in the task of textual classification in NLP using pre-trained models?"

The articles [Howard and Ruder 2018], [Devlin et al. 2019] and [Yang et al. 2019], selected based on the NLP-Progress ranking, present the following language models: ULMFiT (Universal Language

Table V. Comparison between the use or not LM fine-tuning in domain final task documents. Source: Authors.

Modelo	Dataset	Fine-tuning	No Fine-tuning
ULMFiT[Howard and Ruder 2018]	IMDb [3]	5.00 (Error rate)	6.99 (Error rate)
BERT[Devlin et al. 2019]	CoNLL-2003[4]	96.4 (F1)	95.5 (F1)

Model Fine-tuning for Text Classification), BERT (Bidirectional Encoder Representations from Transformers), and XLNet, respectively.

The studies in [Howard and Ruder 2018], [Devlin et al. 2019] and [Yang et al. 2019], present proposals for architectures and training algorithms for the generation of language models with transfer learning, representing the state of the art at the date of their publications, in chronological order: [Howard and Ruder 2018], [Devlin et al. 2019] and more recently, in 2019, [Yang et al. 2019].

The study carried out in [Howard and Ruder 2018] focuses on the classification of texts and how to perform the fine-tuning for this task in the target problem domain by applying recurrent neural networks AWD-LSTM, and its great contribution was the proposal of a new algorithm of training in 3 stages: 1) pre-training in final task dataset language; 2) followed by unsupervised training with the dataset from the final task; 3) and finally the adjustment of the model parameters for the textual classification.

In the BERT and XLNet models, techniques based on Autoencoders and attention mechanisms [Vaswani et al. 2017], and Transformers-XL [Dai et al. 2019], are used respectively. These models proved to be more accurate in textual classification tasks when compared to ULMFiT, according to the NLP-Progress ranking, with XLNet currently leading this ranking.

The models and architectures analyzed in this Section, represent sentences or words (tokens) according to the context in which they are applied, i.e., their numerical representation through vectors or matrices in a continuous space, also called *word embeddings* [Mikolov et al. 2013], has different values according to the position and text in which the words are found in the sentence, different from what is found in *word embeddings* techniques used in the work [Hartmann et al. 2017], and analyzed in Section 4, which have a fixed vector representation value within the language vocabulary for which they were generated, which does not imply that they are less important, however, some characteristics related to the context of the application of the words in the text do not are well represented in the latter case.

For the works [Howard and Ruder 2018] and [Devlin et al. 2019], in addition to the experiments with fine-tuning of the model, others were also carried out to evaluate the possibility of their application as feature extractors of the texts without previous adjustment for the domain final task data, i.e., the parameters of the layers that represent the language are unchanged, and only the layers referring to the final task are adjusted. The tests showed good results when compared with the models that received the adjustment in domain final task documents, according to data from Table V formulated from the results contained in these works.

When considering the BERT and XLNet models, which presented a paradigm shift in the field of Natural Language Processing, precisely because they use attention mechanisms, Autoencoders, and Transformers-XL in their architectures, they manage to generate representations of deeper languages and are said to be bidirectional, as they can capture both context directions of a sentence.

The evolution that XLNet presented about the other two models analyzed, was to be able to combine the positive characteristics of the previous ones, and to mitigate those that did not bring benefits to the construction of a bidirectional language model. While BERT is an architecture based on Autoencoders

[3][Maas et al. 2011]
[4][Tjong Kim Sang and De Meulder 2003]

Table VI. Comparison of state-of-the-art models in classification tasks. Source: [Yang et al. 2019].

Model	IMDB	Yelp-2[5]	Yelp-5	DBPedia[6]	AG[7]	Amazon-2 [8]	Amazon-5 [9]
ULMFiT[Howard and Ruder 2018]	4.6	2.16	29.98	0.80	5.01	-	-
BERT[Devlin et al. 2019]	4.51	1.89	29.32	0.64	-	2.63	34.17
XLNet[Yang et al. 2019]	**3.20**	**1.37**	**27.05**	**0.60**	**4.45**	**2.11**	**31.67**

(AE)[Ponti et al. 2017], and ULMFiT is supported on an Autoregressive (AR) architecture, XLNet, through Transformer-XL, combines the best features of AR and AE, while managing to capture longer contexts of sentences in both directions.

An important point that XLNet solves compared to the BERT model, is the limitation in the number of tokens at the entrance of the network, which is characteristic of the training of the latter, consequently, it solves the difficulty in dealing with longer contexts of the AEs, a very common situation in legal document classification tasks. For BERT to be able to overcome this difficulty, it is necessary to implement mechanisms external to its architecture defining the best parts of the text to be used in the classification task or even to include a hierarchical attention layer, as proposed in [Chalkidis et al. 2019], with the HIER-BERT architecture.

Another feature that sets BERT and XLNet apart is that in the first, LM pre-training consists of randomly masking some tokens of the input sentence in the network with [MASK], which must be predicted by the model, is not a very common task for a final task application. Therefore, this distance in the way the LM and final task adjustment is performed can cause the model "catastrophic forgetting" [McCloskey and Cohen 1989], i.e., changes in the model parameters in such a way that all that knowledge previously learned in unsupervised training is lost.

In contrast, XLNet's pre-training task consists of predicting a token within a sentence, considering several combinations of the words, thus reinforcing the bidirectional property, without directing to include a specific token [MASK] during the training. Table VI, transcribed from [Yang et al. 2019], compares results (error rate) of textual classification of the 3 language models in benchmark datasets.

3.1 BERT in Text Classification Task

Even with the limitation of tokens in the model input, and therefore the difficulty of learning longer contexts, BERT is still an excellent alternative for textual classification, as shown in Table VI. In this vein, [Sun et al. 2019] makes a comparative study of how to use the model for this task, analyzing different techniques in the implementation of fine-tuning and pre-training on the problem domain data. For fine-tuning, it studies how to pre-process the texts, mainly regarding the BERT sentence size limitation feature, and goes to tests regarding the learning rate to avoid the "catastrophic forgetting".

For pre-training, [Sun et al. 2019] analyzes the impact of training the model with data from the domain of the final task, data from the general domain, and also those that present similar context with the final task, and that can be used to increase the volume of the corpus at this stage of the adjustment.

As general conclusions of the article specifically analyzed in this item, it became evident to the authors that: 1) the higher layers of BERT are more useful for the classification of texts; 2) With appropriate learning rates, BERT can overcome the problem of "catastrophic forgetting"; 3) pre-training on the final problem data and those of the same domain improves the model's performance;

[5] `https://www.yelp.com/dataset`
[6] `https://wiki.dbpedia.org/`
[7] `http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html`
[8] [Zhang et al. 2015]
[9] [Zhang et al. 2015]

4) pre-training in classification multitasking may improve the final task result, but the performance gain may not justify the addition of this step; 5) the BERT model can be used in tasks with small amounts of data available for training.

These results in [Sun et al. 2019] show the real possibility of applying the BERT model in the classification task, and are important to direct other works that use it for this purpose, which can save much of the initial search for the better architecture and pre-training technique and fine-tuning.

As can be seen in Section 5, most of the articles analyzed had the best results with BERT. For ULMFiT, applications were found in legal documents in Brazilian Portuguese, and it was also used as a comparison in studies that used datasets in English.

XLNet still needs further study and application in the legal context, since it was used as a test only in [Shaheen et al. 2020], and even so with restrictions due to the computational cost of the model, not being possible fine-tuning in the final task data.

4. PRE-TRAINED LM IN BRAZILIAN PORTUGUESE (Q.5)

This Section aims to answer question 5 of the review questions: "Is there a repository of pre-trained language models for Brazilian Portuguese?"

Only two works, [Souza et al. 2020] and [Hartmann et al. 2017], present pre-trained language models for Brazilian Portuguese. In [Hartmann et al. 2017], these templates are *word embedding* techniques, such as Word2Vec [Mikolov et al. 2013], FastText [Bojanowski et al. 2017], Glove [Pennington et al. 2014], and Wang2Vec [Ling et al. 2015]. In [Souza et al. 2020], the pre-trained model is BERT, as proposed in [Devlin et al. 2019]. The basic difference between what is trained in [Souza et al. 2020] and [Hartmann et al. 2017], is that the former uses recent techniques, notably proposed from 2018, to build language models for transfer of learning, as already explained in Section 5, and the latter uses deep learning architectures and statistical models to train the vectors that represent each word/token within the vocabulary.

In [Hartmann et al. 2017], an intrinsic evaluation was performed for each of the models, as proposed in [Mikolov et al. 2013], according to syntactic and semantic analogies of the token representations generated, in addition to evaluations in similarity tasks, sentences, and grammatical class labeling, defined as extrinsic evaluations. To carry out the training of these models, the work used datasets in Brazilian and European Portuguese, among them: articles from Wikipedia, scientific texts, news sites, e-books, children's books, etc., to form a corpus in Portuguese, which, after preprocessing, resulted in 714,286,683 tokens.

As a result of the work, the authors of [Hartmann et al. 2017] show: 1) regarding the intrinsic assessment for Brazilian Portuguese, the Glove method produced the best-aggregated results (syntactic and semantic), with an index of 46.7 for a vector of size 300, however, was the worst in NLP (extrinsic analysis) tasks together with FastText; 2) Wang2Vec presented the best accuracy indices for the extrinsic analysis, above 95%, in addition to showing very interesting indices in the intrinsic analysis, of 45.1, very close to the Glove, proving to be a good alternative for NLP tasks in Brazilian Portuguese. Furthermore, the tests also show that the representations of words with vectors greater than 300 did not obtain performance gains proportional to the increase in the use of computational memory.

After training and analysis, the study in [Hartmann et al. 2017] made available to the community all the models generated through the NILC (Inter-Institutional Nucleus of Computational Linguistics) repository[10].

Otherwise, in [Souza et al. 2020] the pre-trained model is BERT, which used the brWaC corpus [Wagner Filho et al. 2018] with 3.53 million documents for training. Two BERT network architectures

[10] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

that vary in size were trained, LARGE (7 days of training) and BASE (4 days of training), and the result was evaluated in two NLP tasks: textual link recognition, with the ASSIN2 dataset, and named entity recognition (NER), with the "First HAREM" and MiniHAREM datasets. The results show that the generated Brazilian Portuguese language models outperform other network architectures in the aforementioned NLP tasks, with differences in terms of F1-score that reaches 12 points when compared to a convolutional architecture, and 4 points with the Pre-trained multilingual BERT, which shows, in the latter case, that specific training in a single language brings benefits to the model.

In addition to these results, the authors of [Souza et al. 2020] also tested the models as feature extractors for the final NER task, i.e., they were not adjusted together with the target task, and remained with the static parameters, providing the vectors of features as input to a BiLSTM-CRF network. In this case, the proposed models also showed better results when compared to the multilingual BERT model. This study is important because it tests the use of BERT in an architecture similar to what occurs when *word embeddings* trained in [Hartmann et al. 2017] are used, which can be an alternative, especially if the computational cost is a limiting factor.

In the end, the authors of [Souza et al. 2020] made the parameters of the models[11] available to the community, and they are an important starting point for studies that apply pre-trained language models for Portuguese, especially when considering the computational and financial cost of generating them, in current values more than R$7,000.00 (seven thousand reais) for the use of a TPU v3- 8 from Google Cloud[12] for 7 days, which could make several studies unfeasible.

As explained in Section 6, many works in Portuguese in the legal area use *word embeddings* techniques applied in [Hartmann et al. 2017], especially [Silva and Maia 2020] and [Mota et al. 2020] that use the vectors available as part of the textual classification architecture, which shows that vectorization based on statistics and deep learning still has great importance and application in the real world, mainly due to its ease of use, the possibility of adjustment within the final task domain, and lower computational cost when compared to BERT or XLNet.

5. TEXT CLASSIFICATION IN THE LEGAL DOMAIN WITH LANGUAGE MODELS (Q.2)

This Section is intended to answer question 2 of the review questions: "Are there scientific publications with pre-trained language models used in the legal domain?"

Of the 7 articles analyzed in this Section, it appears that the majority ([Song et al. 2021],[Wang et al. 2020],[Chalkidis et al. 2020],[Shaheen et al. 2020], and [Chalkidis et al. 2019]) use the BERT model for the textual classification tasks, and the other two, [Soh et al. 2019] and [Campos et al. 2020], have the ULMFiT as their main model. Among the works, only [Wang et al. 2020] and [Campos et al. 2020] use documents in Brazilian Portuguese, the others in English, and one of them [Shaheen et al. 2020], presents a multilingual model that, in addition to the English, includes French and German legal documents. Thus, the analysis of items 5.1 and 5.2 divide the studies with ULMFiT and BERT, respectively, for better didactics in the comparison between the works.

5.1 ULMFiT Applications

The work [Campos et al. 2020] uses a dataset with documents from the Official Gazette of the Brazilian Federal District, and compares architectures with ULMFiT and SVM in terms of accuracy and computational cost. It is important to note that the SVM classifier employs the TF-IDF (term frequency – inverse document frequency) technique as vectorization of documents, which is a purely statistical approach, based on the relative frequency of words in a single document, and also on the set of data, different from those used in [Hartmann et al. 2017], or in the language models analyzed in Section 5.

[11] https://github.com/neuralmind-ai/portuguese-bert
[12] https://cloud.google.com/tpu/pricing

The task of [Campos et al. 2020] consists of classifying 717 labeled documents into 19 different classes, out of a total of 2,652 documents collected. The remainder of the unlabeled documents (1928) were used in the unsupervised pre-training phase in the problem domain. The results showed that the ULMFiT achieved better accuracy and F1-score than the SVM, but with very small differences in absolute terms of percentages, 0.85 and 0.57, respectively, in each index. Furthermore, the computational cost for training the ULMFiT was equivalent to training 1000 SVM models.

[Soh et al. 2019] uses a dataset with 6,227 judgments in English from the Supreme Court of Singapore, and performs an extensive comparison between judgment classification models, from those purely based on statistics, through topic modeling with LSA (Latent Semantic Analysis), *word embeddings* with the Glove technique in conjunction with SVM, and pre-trained BERT and ULMFiT language models. The tests also considered variations in dataset sizes, with proportions of 10%, 50%, and 100%. The results indicate that the language models achieve better indices for smaller sets, and more classical approaches for larger sets. As the language models were not adjusted for the problem domain, as in [Campos et al. 2020], this may have been a factor that implied a less significant improvement in their performance with the increase in the dataset compared to the other approaches, allowing that in the set with 100% of the documents, the SVM + LSA obtains the best indexes.

As the documents classified in [Soh et al. 2019] and [Campos et al. 2020] are relatively large in terms of number of words, it is interesting to compare the studies about the number of tokens that are used in the model input. In [Soh et al. 2019] there were 5 thousand, and in [Campos et al. 2020] it is not explicit, which suggests that all the words in the document that are in the initial vocabulary of the pre-trained model were used. These approaches are possible directly with ULMFiT but not with BERT for example, which has the maximum input sentence size limited by the pre-training process.

5.2 BERT Applications

Articles [Song et al. 2021], [Soh et al. 2019], [Chalkidis et al. 2020], [Shaheen et al. 2020], and [Chalkidis et al. 2019] presented the best results for classifying legal documents with models and architectures based on BERT. Among them, only the works [Soh et al. 2019] and [Chalkidis et al. 2019] did not use the EUROLEX57K [Chalkidis et al. 2019] dataset, and the latter proposed and made the ECHR (European Convention of Human Rights) available. In this wake of proposing data sets for the community, [Song et al. 2021] made available the POSTURE50K, which contains 50,000 cases annotated by experts in the legal field. Given the complexity of these sets, many of them were used in multilabel classification tasks, when the same document can contain more than one label.

In numerical terms, and considering only the F1-score index of those studies that used the EUROLEX57K dataset, the best result was achieved by the model proposed in [Chalkidis et al. 2020], which performed, in addition to training in the global domain, pre-training in data from the same problem domain from a 12 GB base of legal documents in English, and even in the specific data of the problem, this differential in the adjustment probably explains the better result obtained, as observed in Table VII. In [Shaheen et al. 2020], which obtained the second best result, it started from the initial parameters of ROBERTa [Liu et al. 2019], with subsequent adjustment to the problem data. On the other hand, [Song et al. 2021] obtained the worst result among them, the BERT language model was pre-trained from scratch only with data from the classification final task. This difference in the results may indicate that this previous knowledge acquired even in a general corpus, improves the results. Furthermore, [Song et al. 2021] used a feature vector of size 768 to represent the document, while the others did so with 1024, which may indicate that for classification tasks, increasing this representation size does not necessarily bring benefits.

[Wang et al. 2020] deals with a study of Named Entities Recognition in the Portuguese legal domain, and uses the LM available in [Souza et al. 2020], which, despite not being a textual classification task, is an important study to consider, as it was the only one found for the legal domain in Brazilian Portuguese that uses the BERT model in architecture, and indicates a possibility of overcoming the

Table VII. F1-score index comparison of articles that used EUROLEX57K dataset and BERT model. Source: Authors.

Article	Vector Size	Model Tuning	F1-score
[Song et al. 2021]	1024	Specific	0.745
[Chalkidis et al. 2020]	768	General + Same Domain + Specific	**0.824**
[Shaheen et al. 2020]	1024	General + Specific	0.758

Table VIII. Compares the F1-score of BERT and ULMFiT models applied with and without the language model fine-tuning. Source: Authors.

Article	LM	Tuning	no Tuning
[Song et al. 2021]	BERT	0.8030	0.7900
[Chalkidis et al. 2020]	BERT	0.8240	0.8320
[Campos et al. 2020]	**ULMFiT**	**0.8974**	**0.4724**

limitation in terms of sentence size in it, proposing an architecture in conjunction with a recurrent layer.

Regarding the number of tokens of each document used for classification, [Song et al. 2021],[Chalkidis et al. 2020],[Shaheen et al. 2020] use the first 512, and [Chalkidis et al. 2019] an attention mechanism after the BERT layer, to include all text tokens that are part of the model's initial vocabulary, and thus generate a feature vector for each document through an architecture that the authors called HIER-BERT. This format proposed in the article [Chalkidis et al. 2019] resulted in a performance gain, showing that it is a possible approach when there are documents with an average amount of 2,500 words, and like the architecture proposed in [Wang et al. 2020], present a solution to this limitation of BERT.

5.3 Experiments with Different Architectures

Among those studies that implemented these tests, the objective was to identify the importance of each structure within the architecture and training algorithm of the implemented language models. In this sense, the works [Song et al. 2021], [Campos et al. 2020],[Chalkidis et al. 2020], and [Shaheen et al. 2020], show the importance of fine-tuning the language model in problem domain data for BERT and ULMFiT, as well as, during this process, the importance of considering the gradual unfreezing of layers (gradual unfreezing [Howard and Ruder 2018]) during training, i.e., the number of layers that adjusts parameters is gradually increased, as shown in [Campos et al. 2020] and [Shaheen et al. 2020], to avoid losing previous learning.

Still concerning these studies, [Song et al. 2021], [Chalkidis et al. 2020], and [Campos et al. 2020] considered the use of the respective models as feature extractors, that is, without performing an adjustment to the problem domain. Table VIII compares the results and shows that ULMFiT suffered a much greater degradation than BERT in this application architecture.

Both approaches with language models based on Autoregressors or Autoencoders, show efficiency in classifying documents in the law area, obtaining F1-score values above 0.75, at the same time, comparing with classic machine learning tools, show that they still have application in this context, presenting satisfactory results in the comparison indexes, and have lower complexity and computational cost, as shown in [Campos et al. 2020], where the absolute difference in terms of F1-score between ULMFiT and SVM was only 0.02 more for the former.

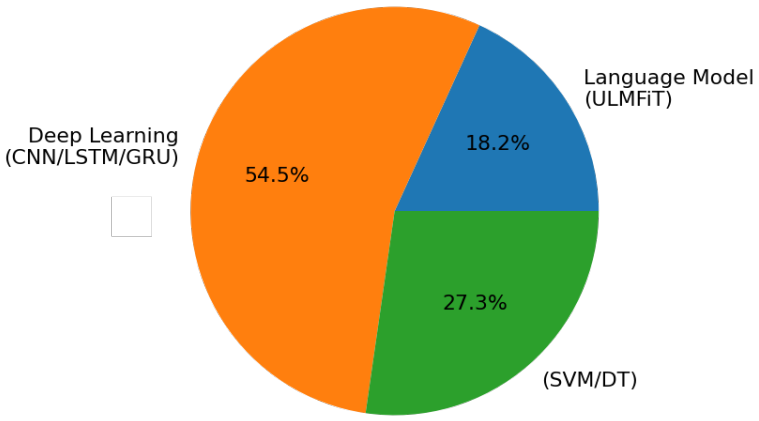

Fig. 5. The proportion of each strategy used in the classifiers of Brazilian Portuguese studies. Source: Authors.

6. TEXTUAL CLASSIFICATION IN THE BRAZILIAN PORTUGUESE LEGAL DOMAIN (Q.3 3 Q.4)

This Section is intended to answer questions 3 and 4 of the review questions: "What NLP techniques in the Portuguese language domain have been applied to classify legal texts in Brazil?" and "Are there public data repositories with texts in the Brazilian legal context?".

The studies in [Noguti et al. 2020], [Luz de Araujo et al. 2020],[Silva et al. 2018],[Silva and Maia 2020], [Raulino Dal Pont et al. 2020], [Wang et al. 2020], [Bertalan and Ruiz 2020], and [Mota et al. 2020], show the application of deep learning, classical and statistical models, to classify documents in the Brazilian legal context. The article [Campos et al. 2020], analyzed in Section 5, also studies the Brazilian legal context, but uses the ULMFiT as one of the test models. Figure 5 shows the proportion of use of each technique among them.

Most studies use deep learning models for classification, among them: CNN and LSTM. Two studies, [Luz de Araujo et al. 2020] and [Bertalan and Ruiz 2020], obtained F1-score results above 0.89 with XGBoost and decision trees, respectively. Regarding [Luz de Araujo et al. 2020], the authors used XGBoost for the task of classifying processes, which have an average of 47 pages, because in tests with other classifiers they obtained lower results, which shows that the size of the document or input text, is an important factor in the performance of the architecture used for textual classification.

In items 6.1 and 6.2, comparisons of the works are carried out, grouping those that are used in the classification architecture as an extractor of text features, *word embeddings* and TF-IDF , respectively.

6.1 Word Embeddings Applications

In the works [Noguti et al. 2020], [Silva and Maia 2020], [Raulino Dal Pont et al. 2020], and [Mota et al. 2020], the studies employ *word embedding* models for Portuguese, as proposed and available in [Hartmann et al. 2017], which shows the importance of applying language model techniques in the legal context of Brazil. Within these studies, the possibility of performing the adjustment of the vectors that represent the words in the problem domain was investigated, and in only one of them, [Silva and Maia 2020], this adjustment did not present better results when compared to the model without adjustment, which can be explained mainly due to the size of the dataset used in this study, while it has 111,343 examples, the other works have a maximum of 17,740, and [Mota et al. 2020]

has only 117, which highlights the importance of this adjustment in the domain of problem when the data available for training is smaller, but which may not be as effective when the data are available in larger quantities.

Among the works that apply the adjustment in the words embeddings, it is worth mentioning that [Raulino Dal Pont et al. 2020] did it from scratch, i.e, it used a corpus with legal documents from several courts in Brazil, and trained a Glove model focused on this domain of knowledge, obtaining an accuracy of 78.5%, in this case, and 76% when using the models already trained and available in [Hartmann et al. 2017].

In [Silva et al. 2018] and [Luz de Araujo et al. 2020], the concept of embedding layer is used to add a specific layer to the proposed model architecture for this task, which starts with random weights, being adjusted during the training for the final classification task, which is similar to training a Word2Vec from scratch, but the parameter adjustments are performed during the training of the complete architecture.

6.2 TF-IDF Applications

It is important to highlight that in addition to the vectorization of words with pre-trained models, statistical techniques are also applied in the analyzed works. In [Luz de Araujo et al. 2020], [Bertalan and Ruiz 2020], and [Campos et al. 2020] the TF-IDF was successfully implemented, obtaining F1-scores above 0.87, notably when the classification occurs in larger documents, or even entire law suits, such as [Luz de Araujo et al. 2020].

In terms of documents size, those works that use the TF-IDF technique ([Luz de Araujo et al. 2020], [Bertalan and Ruiz 2020], and [Campos et al. 2020]) have greater freedom, normally using the entire content. In approaches that use *word embeddings* together with convolutional networks, or even an embedding layer, this value was limited to 300 words in [Noguti et al. 2020], and 1200 in [Silva and Maia 2020], for example.

6.3 Results in Textual Classification

Works in Portuguese achieved excellent results using classical and deep learning models, usually above 0.90 in terms of F1-score, showing the efficiency of these approaches, especially when considering the computational cost compared to the language models presented in Section 3.

Anyway, compared to the studies presented in Section 5, which apply BERT and ULMFiT, it is necessary to consider that most perform multi-label classification in a dataset with at least 66 classes, and [Song et al. 2021], [Chalkidis et al. 2020] and [Shaheen et al. 2020], apply the classification to 4,271 classes, and in the case of studies with legal documents in Brazilian Portuguese, the one with the most classes is [Luz de Araujo et al. 2020], with 29.

In both cases, in Brazilian Portuguese and English documents, the performance gain when adjusting the models to the problem data, for LM or word embedding, is notorious. Figure 6 reinforces this statement and shows that more than 70% of the analyzed articles only use data from the final task domain, or a combination of these with the general domain corpus.

6.4 Public Repositories of Datasets in Brazilian Legal Domain

As an answer to Q.5, [Luz de Araujo et al. 2020] is the only one that provides a Dataset from the Brazilian legal context. It is a very broad set of data, with 45,532 extraordinary appeals from the Federal Supreme Court (STF), and within each of these processes, there are several documents, totaling 692,966. Of these processes, 44,855 are labeled, as are their 628,820 documents.

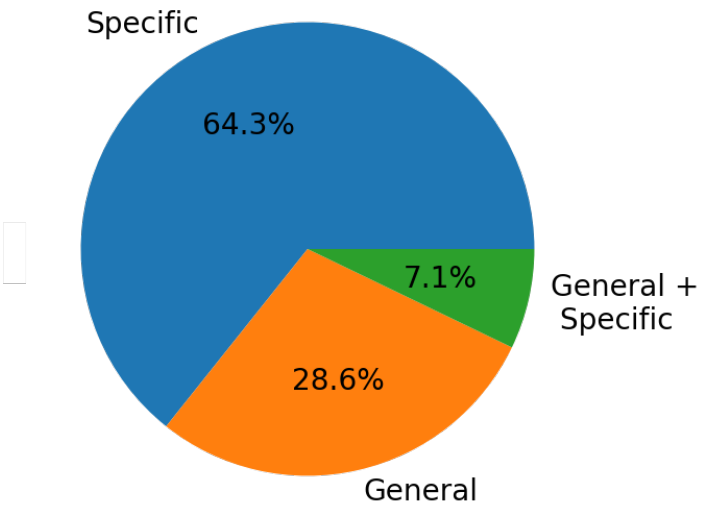

Fig. 6. Training strategy for language models or word embeddings. Source: Authors.

The availability of this dataset is essential for the studies of NLP application in law area, bringing new research opportunities, not only for the labeled data itself, but also as a source for the training of language models specific to the Brazilian legal context.

7. GAPS IN THE ANALYZED STUDIES

After analyzing the articles, specifically those that employ datasets of Brazilian legal documents, it became clear that there is still a need for further study of the application of state-of-the-art language models, despite 2 works, [Wang et al. 2020] and [Campos et al. 2020], make use of this technique, only the last one deals with textual classification. In addition, none of the studies with Brazilian legal documents presented multilabel classification studies.

Possibly, the gap described in the previous paragraph is the result of another, which is the low availability of pre-trained language models in Brazilian Portuguese for BERT, XLNet, and ULMFiT. Only one work [Souza et al. 2020] presented a proposal for BERT. Anyway, this gap can also be explained by the high computational cost for training these models in a large corpus of words, which leads to excessive financial costs, and therefore, makes it difficult to carry out research in the area.

For articles with legal documents in English, there are few that use XLNet, among those analyzed, only one made use of the model, which may be explained by the fact that it was recently proposed, in 2019, as well as the computational cost, difficulty reported in [Shaheen et al. 2020].

8. CONCLUSION

The need to apply machine learning and NLP methods in the legal context is evident, as justified in most of the analyzed works that applied this knowledge in the area. Comparisons show that there are studies in Brazil and Europe aimed at this application of this to the task of textual classification, an essential activity in the courts in any country, in addition, they show the technical feasibility of using the techniques, with non-inferior F1-score results to 80% in the comparative studies.

In English legal documents, notably in datasets from the European legal domain, it was possible to observe that there are already applications of the language models proposed in [Howard and Ruder

2018] and [Devlin et al. 2019], while in works with datasets in Brazilian Portuguese, this is still restricted to only two articles out of the nine selected in this category.

In most works with legal documents in the English language, the use of the technique of fitting the model with data from the context of the problem showed that this approach improves their performance, especially when there is less availability of data from the final task for training, which it was also evident in studies in Brazil in the use of pre-trained *word embeddings* models.

Therefore, 5 major conclusions can be drawn from this systematic review study: 1) there are few Brazilian Portuguese language models available; 2) the fine-tuning of language models and *word embedding* to the problem domain improves the classification result, even if it is done with other data that do not belong to the same knowledge domain of the final task, but keep a semantic relationship with it; 3) classical machine learning models still have their place in the textual classification task when compared to the language models proposed from 2018, even more, when comparing the computational cost of both; 4) pre-trained language models in the general domain of a language can be used without adjustment for the problem domain, acting in the classification architecture as feature extractors; 5) state-of-the-art language models have evolved from architectures based on recurrent neural networks, to those that use Transformers-XL, Autoencoders, and attention mechanisms.

REFERENCES

Bertalan, V. G. F. and Ruiz, E. Predicting judicial outcomes in the brazilian legal system using textual features. In *Digital Humanities and Natural Language Processing*, 2020*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* vol. 5, pp. 135–146, 2017.

Campos, T., Sousa, M., and Luz de Araujo, P. H. pp. 76–86. In , *Inferring the Source of Official Texts: Can SVM Beat ULMFiT?* pp. 76–86, 2020*.

Chalkidis, I., Androutsopoulos, I., and Aletras, N. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pp. 4317–4323, 2019*.

Chalkidis, I., Fergadiotis, E., Malakasiotis, P., and Androutsopoulos, I. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pp. 6314–6322, 2019.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, pp. 2898–2904, 2020*.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pp. 2978–2988, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186, 2019*.

Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. Portuguese word embeddings evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. SBC, Porto Alegre, RS, Brasil, pp. 122–131, 2017*.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, pp. 328–339, 2018*.

Ling, W., Dyer, C., Black, A. W., and Trancoso, I. Two/too simple adaptations of Word2Vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pp. 1299–1304, 2015.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019.

Luz de Araujo, P. H., de Campos, T. E., Ataides Braz, F., and Correia da Silva, N. VICTOR: a dataset for Brazilian legal documents classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pp. 1449–1458, 2020*.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pp. 142–150, 2011.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation, vol. 24. Academic Press, pp. 109–165, 1989.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Y. Bengio and Y. LeCun (Eds.), 2013.

Mota, C., Lima, A., Nascimento, A., Miranda, P., and de Mello, R. Classificação de páginas de petições iniciais utilizando redes neurais convolucionais multimodais. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*. SBC, Porto Alegre, RS, Brasil, pp. 318–329, 2020*.

Noguti, M. Y., Vellasques, E., and Oliveira, L. S. Legal document classification: An application to law area prediction of petitions to public prosecution service. In *2020 International Joint Conference on Neural Networks (IJCNN)*. pp. 1–8, 2020*.

Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543, 2014.

Ponti, M. A., Ribeiro, L. S. F., Nazare, T. S., Bui, T., and Collomosse, J. Everything you wanted to know about deep learning for computer vision but were afraid to ask. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*. pp. 17–41, 2017.

Raulino Dal Pont, T., Sabo, I., Hübner, J., and Rover, A. Impact of text specificity and size on word embeddings performance: An empirical evaluation in brazilian legal domain, 2020*.

Shaheen, Z., Wohlgenannt, G., and Filtz, E. Large scale legal text classification using transformer models. *Computer Science ArXiv* vol. abs/2010.12871, 2020*.

Silva, A. C. and Maia, L. C. G. The use of machine learning in the classification of electronic lawsuits: An application in the court of justice of minas gerais. In *Intelligent Systems*, R. Cerri and R. C. Prati (Eds.). Springer International Publishing, Cham, pp. 606–620, 2020*.

Silva, N., Braz, F., and de Campos, T. Document type classification for brazil's supreme court using a convolutional neural network. pp. 7–11, 2018*.

Soh, J., Lim, H. K., and Chai, I. E. Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2019*. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 67–77, 2019*.

Song, D., Vold, A., Madan, K., and Schilder, F. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Information Systems*, 2021*.

Souza, F., Nogueira, R., and Lotufo, R. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, R. Cerri and R. C. Prati (Eds.). Springer International Publishing, Cham, pp. 403–417, 2020*.

Sun, C., Qiu, X., Xu, Y., and Huang, X. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu (Eds.). Springer International Publishing, Cham, pp. 194–206, 2019*.

Tjong Kim Sang, E. F. and De Meulder, F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. pp. 142–147, 2003.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Vol. 30. Curran Associates, Inc., 2017.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

Wang, Z., Wu, Y., Lei, P., and Peng, C. Named entity recognition method of brazilian legal text based on pre-training model. *Journal of Physics: Conference Series* vol. 1550, pp. 032149, 05, 2020*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Vol. 32. Curran Associates, Inc., 2019*.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. MIT Press, Cambridge, MA, USA, pp. 649–657, 2015.