International Journal on Document Analysis and Recognition (IJDAR)
https://doi.org/10.1007/s10032-022-00406-7

ORIGINAL PAPER

Sequence-aware multimodal page classification of Brazilian legal documents

Pedro H. Luz de Araujo[1] · Ana Paula G. S. de Almeida[2] · Fabricio Ataides Braz[3] · Nilton Correia da Silva[3] · Flavio de Barros Vidal[1] · Teofilo E. de Campos[1]

Received: 8 April 2021 / Revised: 31 December 2021 / Accepted: 3 June 2022

Abstract
The Brazilian Supreme Court receives tens of thousands of cases each semester. Court employees spend thousands of hours to execute the initial analysis and classification of those cases—which takes effort away from posterior, more complex stages of the case management workflow. In this paper, we explore multimodal classification of documents from Brazil's Supreme Court. We train and evaluate our methods on a novel multimodal dataset of 6510 lawsuits (339,478 pages) with manual annotation assigning each page to one of six classes. Each lawsuit is an ordered sequence of pages, which are stored both as an image and as a corresponding text extracted through optical character recognition. We first train two unimodal classifiers: A ResNet pre-trained on ImageNet is fine-tuned on the images, and a convolutional network with filters of multiple kernel sizes is trained from scratch on document texts. We use them as extractors of visual and textual features, which are then combined through our proposed fusion module. Our fusion module can handle missing textual or visual input by using learned embeddings for missing data. Moreover, we experiment with bidirectional long short-term memory (biLSTM) networks and linear-chain conditional random fields to model the sequential nature of the pages. The multimodal approaches outperform both textual and visual classifiers, especially when leveraging the sequential nature of the pages.

Keywords Multimodal page classification · Document classification · Legal domain · Sequence classification · Portuguese language processing

1 Introduction

The Brazilian court system is burdened by a large number of lawsuits. In 2019, there were 77.1 million lawsuits awaiting judgment—almost one lawsuit for every three Brazilians. Some of these lawsuits will stay in the system for a long time, with average processing times that can reach more than six years. All of this contributes to raising the legal system cost: that same year, Brazil spent about R$100 billion in expenses with the judiciary, about 25 billion dollars considering the average exchange rate in 2019 [35].

Natural language processing (NLP) and machine learning (ML) techniques can improve this scenario by enabling faster and more efficient document analysis. Brazil's Supreme Court receives roughly 42 thousand cases each semester, which takes about 22 thousand hours for humans to sort through [40]. This time could be better spent on more com-

✉ Pedro H. Luz de Araujo
pedrohluzaraujo@gmail.com

Ana Paula G. S. de Almeida
anapaula.gsa@gmail.com

Fabricio Ataides Braz
fabraz@unb.br
https://www.github.com/fabraz

Nilton Correia da Silva
niltoncs@unb.br
http://lattes.cnpq.br/5916642485883241

Flavio de Barros Vidal
fbvidal@unb.br
https://www.cic.unb.br/professores/77-fbvidal

Teofilo E. de Campos
t.decampos@oxfordalumni.org
https://teodecampos.github.io

[1] Department of Computer Science, Universidade de Brasília, 70910-900 Brasília, Brazil

[2] Department of Mechanical Engineering, Universidade de Brasília, 70910-900 Brasília, Brazil

[3] Gama Faculty, University of Brasilia, Campus Gama - Setor Leste - Gama, 72444-240 Brasília, Brazil

Published online: 12 July 2022

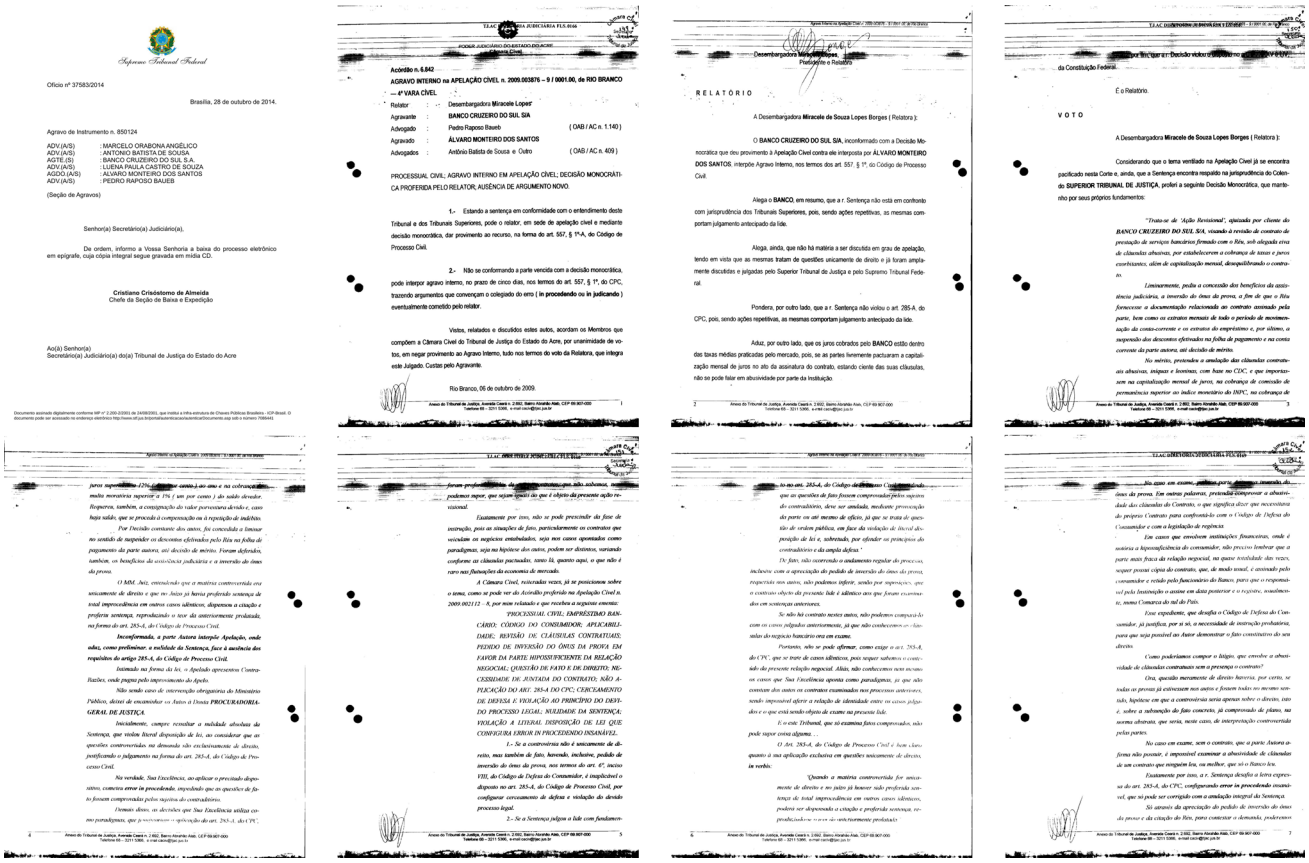


Fig. 1 The first eight pages of a lawsuit. While the first page is clean, the others come from an older document and contain ink stains, stamps, handwritten signatures and other artefacts

plex stages of the workflow, such as those requiring legal reasoning. The cases reach the court as mostly unstructured and unindexed PDF files of raster-scanned documents [27]. Intra-class diversity and document quality are the main challenges: The documents range from petitions and evidence to rulings and orders, originate from different Brazilian courts and often contain visual noise such as handwritten annotation, stamps and stains (Fig. 1).

Therefore, our goal is to explore and evaluate methods that automatically classify document pages by combining different sources of information. Though previous work [27,28] has examined Brazilian legal document classification, we are the first to combine visual, textual and sequential data to train better performing models. Our main contributions are:

1. SVic+, a multimodal dataset of lawsuits composed of ordered document images and corresponding texts.
2. Proposing and evaluating an early fusion multimodal combination method (Fusion Module) and a sequence learning method (BiLSTM-F) that leverages textual, visual and sequential information to improve Supreme Court document classification. We contrast the performance of our methods with late fusion and conditional random field (CRF) postprocessing baselines and with LayoutXLM [44], a state-of-the-art method for multimodal document classification
3. Training and evaluating our methods on Tobacco800 [1] to compare with the state-of-the-art results for that dataset.
4. Outperforming the state-of-the-art results on the small version of the VICTOR dataset [27].

The rest of this paper is organised as follows. In Sect. 2, we examine previous work on multimodal document classification. In Sect. 3, we describe the data we used to train and evaluate our models, which we describe, along with their experimental setting, in Sect. 4. We then discuss the results obtained and conclude the paper in Sects. 5 and 6.

2 Related work

2.1 Multimodal document classification

Textual and visual content are two of the four document aspects listed by Chen and Blostein [9] as possible feature sources. Image features range from fixed descriptors such as pixel density at different locations and scales [32] to approaches based on convolutional neural networks [2,15,22,28,42] such as VGG-16 [36] and MobileNetV2 [34].

Text features range from traditional methods such as latent semantic analysis [12] to pre-trained word embeddings (e.g., Fasttext [6]) and deep learning approaches [2,15,42].

These feature modalities may be used by themselves or combined to improve classification performance. This can happen at the feature level (early fusion) or at the prediction level (late fusion). Rusinol et al. [32] experiment with both options, trying different methods to combine predicted probabilities for late fusion (summing, multiplying, taking the maximum and logistic regression). Jain and Wigington [22] compare a spatially aware early fusion method with four alternative methods of feature combination: concatenation, addition, compact bilinear pooling and gated units. The spatially aware fusion underperformed simple feature combination, with concatenation, addition and bilinear approaches performing similarly. Engin et al. [15] explore late and early fusion for the classification of Turkish banking documents, finding that both outperform unimodal methods. Mota et al. [28] investigate multimodal classification of Brazilian court documents, concluding that multimodal approaches compare favorably with unimodal ones.

Text-only methods have had a significant evolution in the last years due to the advent of Transformers [4,13,29,41]. These methods have been extended to work in multimodal settings. Xu et al. [43] include document layout information by equipping a transformer [41] with two-dimensional position embeddings that model the relative spatial position of the words in a document. This work is later extended [45] with the addition of two pre-training tasks that integrate visual and textual information. Bakkali et al. [3] achieved SOTA results in the RVL-CDIP dataset [17] by combining image features from off-the-shelf visual models and text features extracted using BERT [13].

Fewer works have explored incorporating sequential information. Rusinol et al. [32] use an n-gram model of the page stream that conditions page predictions on the types of the $n-1$ previous pages to capture their sequential nature. Wiedemann and Heyer [42] use as a feature of the target page the encoding of its predecessor. Luz de Araujo et al. [27] feed the predictions of a text classifier to a linear-chain conditional random field (CRF) [25] to jointly predict pages of lawsuits.

In this paper, we focus on the sequentially aware combination of visual and textual features for legal document classification. To the best of our knowledge, this is the first work that considers both visual and textual modalities and sequential dependencies when classifying documents from the legal domain and in Portuguese—Luz de Araujo et al. [27] do not use visual features, while Mota et al. [28] do not leverage sequential information.

2.2 Document classification datasets

Previous works have presented a variety of document classification datasets. Table 1 compares SVic+ with 10 existing datasets. We examined the following characteristics:

1. Dataset domain;
2. Dataset language;
3. Number of pages (images) in the dataset;
4. Presence of multi-page documents;
5. Whether there is annotation for document type classification (DTC);
6. Whether the dataset is available online;
7. And whether text data is available online.

Based on this analysis we have found three main points that distinguish our dataset: size, domain and language, and data availability.

Table 1 Comparison of the examined datasets. PSS: has Page Stream Segmentation annotation. DTC: has Document Type Classification annotation. Available online: is freely available online. Text data available: is made available with text data

Dataset	Domain	Language	No. of pages	PSS	DTC	Available online	Text data available
Ai.Lab.Splitter [7]	Legal proceedings	Portuguese	31,784	✓		✓	
Archive26k [42]	Nuclear waste disposal	German	26,887	✓			
Court lawsuits [28]	Legal proceedings	Portuguese	2,970	✓	✓		
MARG [16]	Medical papers	English	1,553		✓	✓	✓
NIST Special Database 2 [14]	Tax forms	English	5,590	✓	✓	✓	
RVL-CDIP [17]	Tobacco industry	English	400,000		✓	✓	
Spanish banking docs [32]	Banking	Spanish	69,737	✓	✓		
Tobacco3482 [24]	Tobacco industry	English	3,482		✓	✓	
Tobacco800 [1]	Tobacco industry	English	1,290	✓		✓	
Turkish banking docs [15]	Banking	Turkish	$\approx 27,000$	✓	✓		
SVic+ (ours)	Legal proceedings	Portuguese	339,478	✓	✓	✓	✓

2.2.1 Size

From the datasets examined, the only one that compares to our 339,478 pages is the RVL-CDIP [17] dataset, which contains 400,000 document images originating from the US tobacco industry. That said, since RVL-CDIP does not contain annotation for multi-page documents, it does not support page stream modelling and page stream segmentation (PSS). It also differs from our data when considering the following two points.

2.2.2 Domain and language

Since all datasets in the Portuguese language we examined belong to the domain of legal proceedings, we consider both domain and language as one single distinguishing characteristic. Two datasets share the language and domain of our data: Ai.Lab.Splitter [7] and a dataset of court lawsuits developed by Mota et al. [28]. Another common feature is that all three datasets contain annotation for multi-page documents and therefore support methods that involve sequences of pages. On the other hand, our dataset has 10x and 100x more pages than Ai.Lab.Splitter and the court lawsuit dataset, respectively. In addition, the former has no annotation for DTC; while the latter is not available online[1].

2.2.3 Data availability

Although the cited datasets have been used for multimodal experiments, most of them do not contain explicit text data (e.g. in csv format or as raw text files). In those cases, one would have to either ask the authors for the data they used or run an optical character recognition (OCR) system. With different works using different versions of the text modality, the reproducibility and fair comparison of methods are compromised. We, on the other hand, make available both document images and corresponding text data.[2]

In summary, among the examined body of work, our dataset is the only publicly available dataset with multipage documents and hundreds of thousands of labelled pages that supports text, image, multimodal and sequence classification.

3 Data

3.1 SVic+

We propose SVic+, a dataset of 6510 Extraordinary Appeals comprising 339,478 pages. Each instance is a lawsuit as it is received by the Brazilian Supreme Court, before it is processed and judged. Each lawsuit contains different documents (petitions, rulings, orders) and is represented as an ordered sequence of pages containing text.

This dataset is an extension of Small VICTOR [27], which we expanded to include, in addition to textual data, the document images. Every page in the expanded corpus is stored in at least one of two formats. First, as text extracted through optical character recognition [39], with the following additional preprocessing steps: lower-casing, removal of stop words and alphanumeric tokens, e-mail and URL tokenisation (e-mails and URLs are replaced by the tokens "email" and "link"), and special tokenisation of legislation references (e.g., Lei (law) 11.419 to LEI_11419). Second, as JPEG images converted from the original PDF files, with mean width and height of 1664 and 2322 pixels respectively.

When members of the Court's staff upload documents to their internal database, they assign their corresponding categories. We use those manually attributed labels as the ground truth for our samples by assigning to each page the type of the document it belongs to. Since we use the annotation employed internally by the Court's staff and executed in the context of the Court's ordinary workflow, we are not able to compute inter-annotator agreement measures. That said, the fact that the labels were chosen by experts provides a fair degree of trustworthiness to the annotation..

There are six possible classes:

1. *Acórdão*, for lower court decisions under review;
2. *Recurso Extraordinário* (RE), for appeal petitions;
3. *Agravo de Recurso Extraordinário* (ARE), for motions against the appeal petition;
4. *Despacho*, for court orders;
5. *Sentença* for judgements; and
6. *Others* for documents not included in the previous classes.

Most of the samples contain both textual and visual sources of information, except for 33,849 images with no corresponding text and 4 texts with no corresponding image. We discuss our strategy for dealing with missing data when training fusion models in Sect. 4.3. The corpus is divided into train, validation and test splits containing 70%, 15% and 15% of all suits, respectively. Table 2 presents the number of text and image samples across data splits and classes. Due to the nature of the data, a document may appear more than once in a lawsuit, so we present both raw and deduplicated counts. That said, given that the corpus has been split by lawsuits, there is no sample intersection between splits.

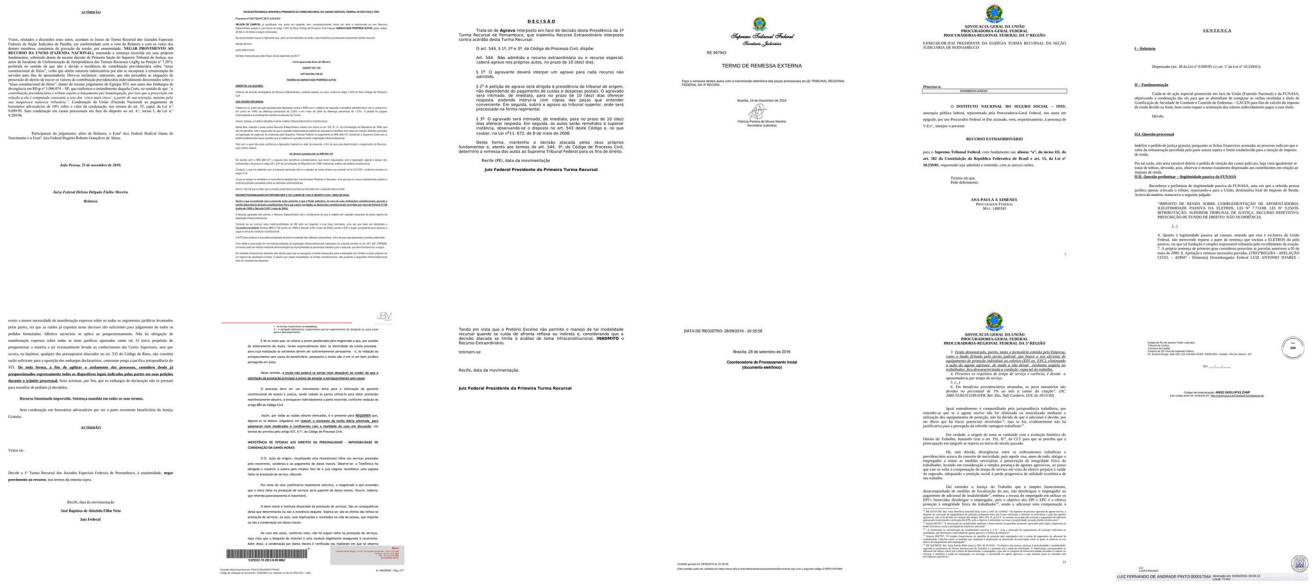
Human agents find the first page of documents easier to classify when compared to interior pages. This is true considering both visual and textual aspects, since first pages contain highly informative cues, such as headers and titles. Figure 2 compares first page and interior page samples for each class. We validate this intuition in Sect. 5.5.

[1] To the best of our knowledge.

[2] http://ailab.unb.br/victor/lrec2020/.

Table 2 Class counts per split, showing the number of page images and text extracted through optical character recognition (OCR). Between parentheses, the deduplicated counts

Class	Training set		Validation set		Test set	
	Image	Text	Image	Text	Image	Text
Acórdoão	583 (583)	553 (553)	320 (314)	299 (293)	287 (285)	273 (271)
ARE	4258 (4220)	2546 (2508)	2798 (2650)	2149 (2001)	2655 (2537)	1841 (1723)
Despacho	361 (361)	346 (346)	189 (183)	183 (177)	199 (198)	198 (197)
Others	144,583 (140,786)	134,134 (130,337)	95,602 (91,434)	84,104 (79,936)	92,529 (87,902)	85,408 (80,781)
RE	10,225 (10,181)	9509 (9465)	6987 (6803)	6364 (6180)	6386 (6177)	6331 (6122)
Sentença	2177 (2177)	2129 (2129)	1681 (1613)	1636 (1568)	1503 (1478)	1475 (1450)



(a) *Acórdão* **(b)** ARE **(c)** *Despacho* **(d)** Others **(e)** *Petição* **(f)** *Sentença*

Fig. 2 First page (top row) and interior page samples for each class

3.2 Tobacco800

To compare our methods with previous approaches, we also train and evaluate our classification pipeline on Tobacco800 [1], a dataset of 1,290 document images. We chose it due to its widespread use in the document classification community, with previously published multimodal sequential-aware results for comparison by Wiedemman and Heyer [42]. The dataset is a subset of the Illinois Institute of Technology Complex Document Information Processing (IIT-CDIP) Test Collection [26], a corpus of more than 14 million documents from seven US tobacco industry organisations with different types of documents such as letters, advertisements and reports.

Tobacco800 does not have annotation for document type, but can be used for PSS since it contains multi-page documents. We treat the task as binary classification between first and interior pages of a document. As the text data, we use the same OCR-extracted data used by Wiedemann and Heyer [42].

4 Methods

In this section we describe our methods for visual and textual page classification, feature fusion and sequence learning. We also describe the corresponding experimental settings. Unless stated otherwise, we optimise the cross-entropy loss using Adam [23] and mini-batches of 64 samples and training samples are randomly shuffled at the start of each epoch..

We use the F_1 score as the evaluation metric, defined as:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{1}$$

whereas precision and recall are defined as follows: let tp, fp and fn be the number of true positives, false positives and

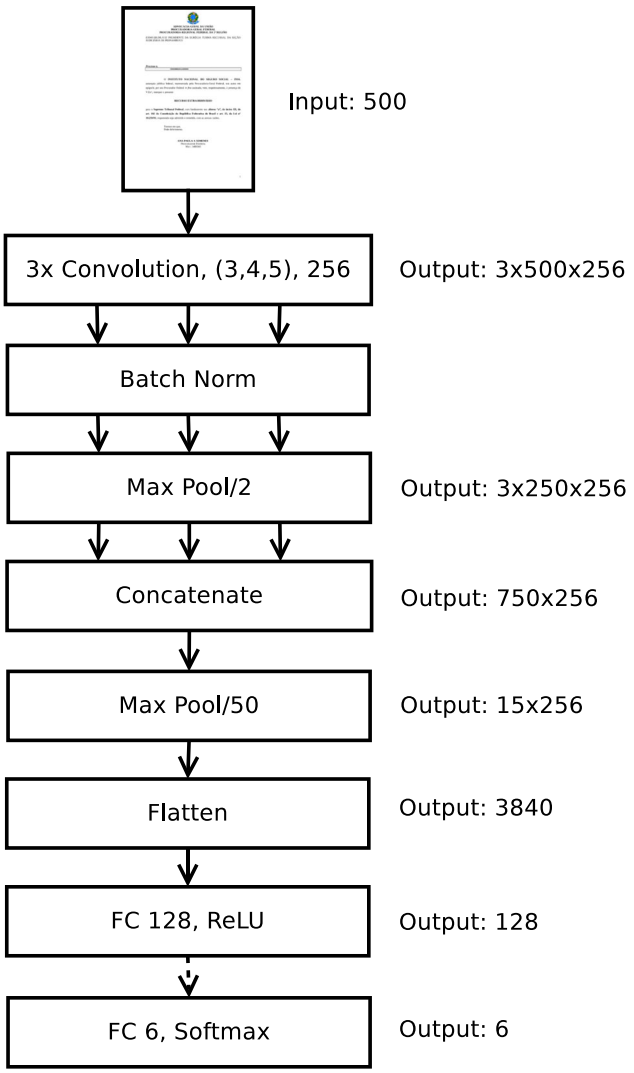


Fig. 3 The network architecture we used for text classification. Dashed lines indicate dropout was applied

false negatives, respectively. Then,

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}, \tag{2}$$

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}. \tag{3}$$

To aggregate scores for all classes, we report average and weighted F_1 scores, which correspond to the unweighted average and the average weighted by the number of samples in each class, respectively.

We evaluate the models using the parameters with the best average F_1 score computed on the validation set. That is, after each epoch, we only save model parameters if the validation performance is the highest up to that point.

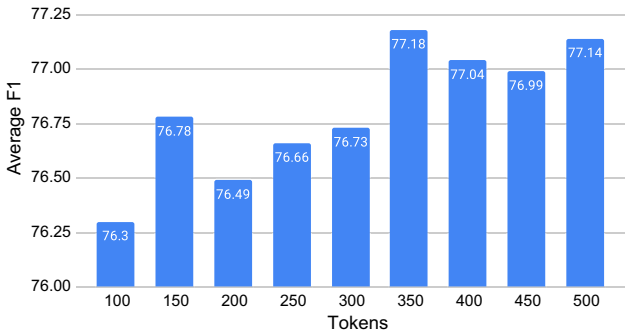


Fig. 4 Text classification: average F_1 scores of the CNN text model evaluated on the validation set varying the maximum number of tokens

4.1 Text classification

We use the convolutional neural network (CNN) architecture described by Luz de Araujo et al. [27] as the method for text classification. It is a shallower version of the one proposed by Conneau et al. [11] and works on the word level instead of the character level. Figure 3 summarises the CNN structure. To process a sample, the network takes as input the document's tokens and embeds them into 100-dimensional vectors of randomly initialised parameters that are updated during gradient descent. The vectors are fed to three convolutional blocks composed of a convolutional layer with three sets of 256 one-dimensional filters with different sizes (3, 4 and 5) followed by batch normalisation and max-pooling of size 2. The resulting vectors are concatenated and fed to a max pool layer of size 50. The output is flattened and processed by two fully connected (FC) layers, with the softmax function producing the prediction. Dropout is applied to the output of the first FC layer with a dropping probability of 50%.

Following [27], we truncate documents longer that 500 tokens (less than 0.6% of the data) and pad shorter documents, so that the input length is always 500. We have trained multiple versions of the text model with different maximum number of tokens to assess the impact of that hyperparameter. Figure 4 summarises the results. Although the model trained with a maximum input length of 350 tokens showed a slight improvement in 0.04 percentage points when compared with the one with a maximum of 500 tokens, we chose to follow previous work [27] and keep 500 tokens as the maximum. Since one of our goals is to evaluate the impact of the combination of different modalities by comparing with the previous unimodal [27] approach, we do not wish to introduce confounding factors that could affect model performance.

As a strategy to deal with class imbalance, we train a variant of the CNN model, which we call CNN-w, that weighs each sample contribution to the loss by a factor inversely proportional to its class frequency. Let c be the number of classes and $\mathbf{w} = [w_1, w_2, \ldots, w_i, \ldots, w_c]$ a c-dimensional vector

whose component i is the factor for class i. Then the factors are computed by the following equation, as implemented in the scikit-learn library [8]:

$$w_i = \frac{n}{c \cdot f_i}, \tag{4}$$

where n is the number of training samples and f_i is the number of samples from class i.

4.2 Image classification

To classify document images, we fine-tune a ResNet50 [18] model pre-trained on ImageNet [33]. We first train only the head of the model for one epoch, employing a cyclic learning rate with cosine annealing [38]. Then, we train all layers for one cycle of 6 epochs with discriminative fine-tuning [20]. As we did for the text classifier, we train a variant of the model with factors inversely proportional to class frequency: ResNet50-w.

To choose learning rates, we use the learning rate range test [37]. That is, we train the model for a few iterations, starting from a low learning rate value and increasing it after each mini-batch, plotting the loss against the learning rates. We then pick a learning rate close to the point where the loss starts to increase—high enough for quick learning, but not so high as to impede learning.

4.3 Image and text combination strategies

In this section we describe our proposed method for early fusion of visual and textual features, and our baselines for comparison.

4.3.1 Hybrid classifier

As a baseline method that fuses visual and textual data we use a hybrid classifier (HC) that works as follows: if textual data is available, use the best text classifier; otherwise, use the best image classifier. The intuition is that this approach would be at least as good as using only text data, which better discriminates the document classes when compared with visual data. Figure 5 illustrates the method.

4.3.2 Late fusion

As a second baseline method we use the best performing late fusion strategy evaluated by Rusiñol et al. [32]. It consists in multiplying the power-weighted predictions of the text and image models. That is, given the text and image probability vectors $\mathbf{p}_{text}, \mathbf{p}_{img} \in \mathbb{R}^c$, the late fusion combines them as

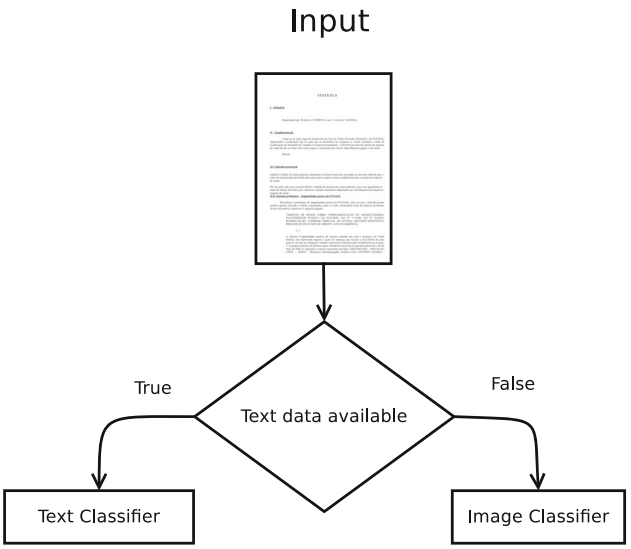


Fig. 5 The hybrid classifier (HC): a baseline fusion classifier that only uses visual information if text data is not available

follows:

$$\mathbf{p}_{lf} = \mathbf{p}_{text}^{j} \odot \mathbf{p}_{img}^{k}, \text{ with } j, k \in [0, 1], \tag{5}$$

where $\odot$ denotes element-wise multiplication. We choose the values for j and k by running a grid-search and evaluating on the validation set.

When a sample has only one modality available, we experiment with two alternatives to represent the missing modality: a vector of either uniform probabilities or class frequencies. That is, let $\mathbf{v}_u, \mathbf{v}_f \in \mathbb{R}^c$ identify the two representation choices, and $u_i, f_i, i \in [1, \ldots, c]$ be the ith component of $\mathbf{v}_u$ and $\mathbf{v}_f$, respectively. Then:

$$u_i = \frac{1}{c} \tag{6}$$

$$f_i = \frac{n_i}{\sum_i n}, \tag{7}$$

where n_i is the number of samples in the validation set that belong to the ith class. We denote the strategies as LF-u and LF-f, respectively.

4.3.3 Fusion module

For the early fusion of visual and textual data we first compute representations using the trained text and image classifiers. As text embeddings, we extract the 3840-dimensional activations of the flatten layer of the CNN (Fig. 3). As image embeddings, we take the activations of the last convolutional block in the ResNet and apply global average and global max pooling. Then we concatenate and flatten the result, obtaining 4096-dimensional vectors.

The pre-computed representations are concatenated and fed to a batch normalisation layer [21], followed by an FC layer with d units, batch normalisation, and a final FC layer. The softmax function produces the predictions. Figure 6 illustrates our proposed fusion module (FM).

In case of missing data, when only the document image or text is available—but not both—we experiment with two options. We try a simple baseline that simply uses a vector of zeroes in such cases (FM-zero). The second approach uses two learnable embeddings: one for missing text data and the other for missing image data. That is, we randomly initialise two vectors, $\mathbf{e}_{\text{image}} \in \mathbb{R}^{4096}$ and $\mathbf{e}_{\text{text}} \in \mathbb{R}^{3840}$, which are used to represent missing image and text, respectively. These vectors are updated during training, with the intuition that gradient descent should find better a representation for missing data than the one used in the FM-zero baseline. We examine the impact of these learnable embeddings in Sect. 5.3.

We run preliminary experiments with one cycle of 10 epochs for each of four configurations, varying the number of hidden units d (128 or 512) and the use of learnable embeddings or zero vectors for missing data. We then train the model that obtained the highest average F_1 score from scratch for one cycle of 20 epochs. Learning rates are chosen using the range test.

4.3.4 LayoutXLM

We also train and evaluate a LayoutXLM (LXLM) [44] model on SVic+ data. It is a state-of-the-art multimodal method for document classification. The architecture consists in a transformer-based [41] encoder that takes as input the document text data and visual feature maps extracted using an off-the-shelf computer vision model. This is the same architecture used in the LayoutLMv2 framework, which is pre-trained on the ITT-CDIP Test Collection [26], composed of documents in English. LayoutXLM, on the other hand, is pre-trained on a corpus of documents ranging from 53 languages, Portuguese among them, which should help with our data. We fine-tune the model on SVic+ for 15 epochs with mini-batches of six pages each.

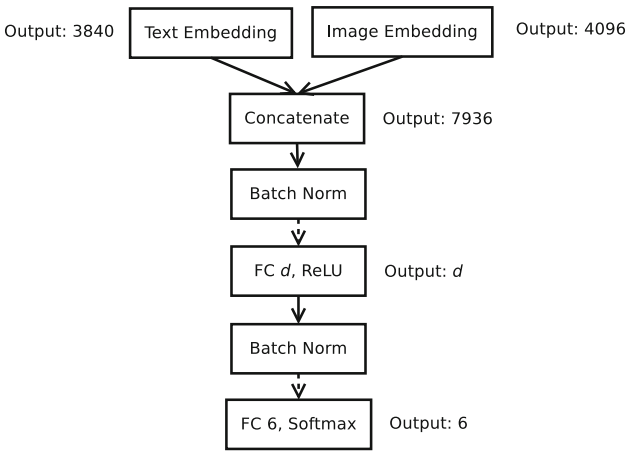


Fig. 6 Fusion Module (FM): the proposed method for early fusion of textual and visual information. Dashed lines indicate dropout was applied. The hyperparameter d is the number of units in the first fully connected layer

4.4 Sequence classification

Given that a lawsuit is composed of an ordered series of pages, one can, instead of classifying each page by itself, leverage the sequential nature of the data by treating the problem as a sequence labelling task. That is, rather than having a page and a class prediction as input and output, the sequence classification approach outputs a sequence of class predictions, given a sequence of input pages. We employ the inside-outside-beginning (IOB) tagging scheme [30] to better leverage the sequential information: we prepend "B-" to the ground truth of first-page samples and "I-" otherwise. For example, if a suit begins with a RE of three pages followed by an ARE of equal length, the label sequence would start with B-RE, I-RE, I-RE, B-ARE, I-ARE, I-ARE.

4.4.1 CRF postprocessing

As a baseline method for sequence classification, we use the CRF postprocessing approach described in [27], which obtained the best results for Small VICTOR, the text-only counterpart of the data we use in this work. It consists in using the predictions of a trained model to train a linear-chain conditional random field (CRF). While in [27] the predictions of the CNN described in Sect. 4.1 were used (CNN+CRF), we instead use the predictions of the FM (FM+CRF): we first save its predictions for all samples in the data; then, we use these six-dimensional vectors as features to train the linear-chain CRF. Figure 7 illustrates the method.

4.4.2 BiLSTM

As an alternative method for sequence classification of pages, we use a bidirectional long short-term (biLSTM) [19] layer to capture sequential dependencies at the feature extraction level—as opposed to the FM+CRF baseline, which only does so at the prediction level. We experiment with two different kinds of input: the activations of the first FC layer of the FM (128-dimensional vectors) and the concatenation of the pre-computed image and text embeddings (7936-dimensional vectors), obtained as described in Sect. 4.3.

The network consists of a biLSTM layer with 128 units for each direction followed by batch normalisation, dropout and an FC layer. When using concatenated image and text

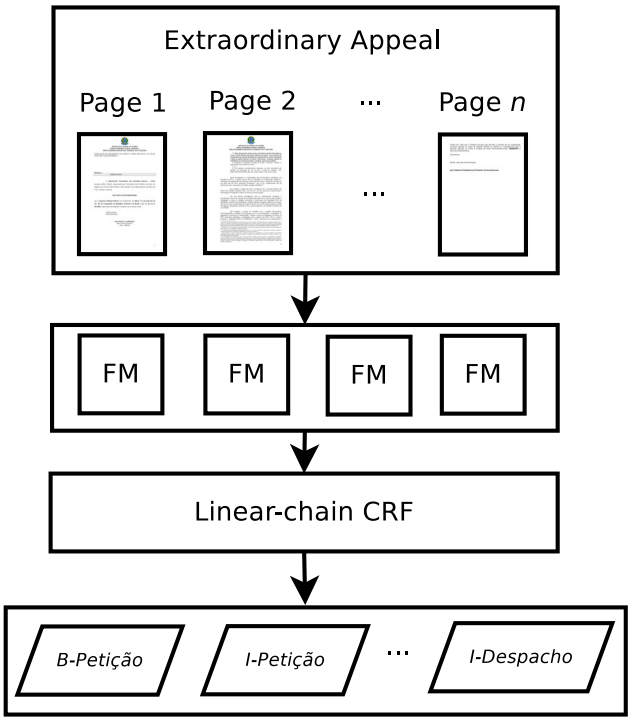


Fig. 7 Baseline sequence classification method (FM+CRF). We feed the (pre-computed) predictions of the fusion module (FM) to a linear-chain CRF to jointly predict the class of each page in an extraordinary appeal

embeddings as input, we first apply batch normalisation and dropout, followed by an FC layer with 512 units.

We train four model variants:

1. BiLSTM, which uses fusion activations as input;
2. BiLSTM-CRF, with the same input and a CRF head on top of the described network;
3. BiLSTM-F, which uses concatenated image and text embeddings as input; and
4. BiLSTM-F-CRF, with the same input and a CRF head.

Due to the memory footprint of BiLSTM-F and BiLSTM-F-CRF, we use mini-batches of eight lawsuits when training them. All models are trained for one cycle [38] of 20 epochs. We use the range test to choose learning rates.

5 Results and discussion

Table 3 exhibits the F_1 scores of the best performing models, categorised by whether they use textual (CNN), visual (ResNet50-w and ResNet50), textual and visual (FM), or sequential (FM+CRF and BiLSTM-F) information.

All models beat majority class classifiers considering both weighted and average F_1 scores—except for ResNet50-w, whose weighted F_1 score is 25.98 p.p. (percentage points) lower. The models with textual data performed much better than those with only visual information available, which is not surprising given that text content is more discriminative than visual aspects when considering the dataset documents—most of them are similar white pages with blocks of text (Fig. 2).

Regarding fusion and sequence classification results, each additional information source contributed to classification metrics: considering average F_1 scores, the FM and the LF-F surpassed the CNN by 0.89 p.p. and 1.22 p.p., respectively, while the multimodal sequential methods (LF-F+CRF, FM+CRF and BiLSTM-F) surpassed all fusion, image and text methods. Surprisingly, LayoutXLM, a state-of-the-art method, performed consistently worse than the text model. LayoutXLM is pre-trained on multimodal tasks where text, layout and image information are equally important; on the other hand, our unimodal results show how the text modality is paramount in SVic+. This discrepancy may cause negative transfer [31] between pre-training and target tasks. We leave to future work the training and evaluation of a randomly initialised LayoutXLM architecture on our data to assess that hypothesis. Furthermore, LayoutXLM is much more time-intensive than our approaches: each epoch takes about 14 hours (11 for training and 3 for evaluating) on a NVIDIA Tesla V100, while our full fusion (FM) and sequential (BiLSTM-F) pipelines take less than 3.5 h each per epoch. It also requires pre-computation of layout items (textual bounding boxes), which is not required by the other methods.

The multimodal sequence models outperformed the variant with only textual information, CNN+CRF, which further supports the importance of leveraging various input modalities: the BiLSTM-F beat the CNN+CRF by 2.03 p.p. when considering average F_1 scores.

The only class whose F_1 score did not improve with sequential information was *Despacho*: the CRF post-processing caused a reduction on F_1 score of 3.78 p.p and 0.99 for the LF-F and FM, respectively. This may be due to the small size of *Despacho* documents. They have an average size of 1.30 pages with 0.77 standard deviation. Furthermore, 78 % of the documents from that class have only one page, while 95 % have two or less pages.

In the paragraphs below we will further examine the results of each category (text, image, fusion and sequence) and perform an ablation analysis of the fusion module.

5.1 Text classification results

Table 4 compares the validation performance of our approaches for text classification. Using class frequency penalty weights to help with data imbalance improved recall on all classes, except for the most frequent one (Others). This increase came at the expanse of a sharp decrease in precision for those classes. As a result, the average and weighted

F_1 scores for the CNN-w were 12.90 and 3.39 p.p. lower than its counterpart with no penalty weights.

Despite using the same architecture, we achieved better results than the ones reported by Luz de Araujo et al. [27] and presented in Table 3. This is probably because we save model parameters only on validation metric improvement when training.

5.2 Image classification results

Table 5 compares the validation performance of the image classification models. Similarly to the text classification results, using class frequency penalty weights resulted in a decrease in recall for the majority class and correspondent increase in that metric for the other classes. As in the previous case, we see a big decrease in precision for the *Acórdão*, RE and *Sentença* pages, though there is a slight increase in precision for the ARE and *Despacho* pages. This is explained by the fact that the ResNet50 with no explicit strategies for class imbalance did not assign samples to those two classes at all.

The ResNet50 achieved higher average and weighted scores than its counterpart that uses class frequency penalty weights. That said, since the ResNet50 scores for *Acórdão*, ARE and *Despacho* were zero or close to zero, the ResNet50-w scores were more equally distributed across the different classes. With the intuition that this could lead to more discriminative features, we experiment with both models' activations when fusing textual and visual data. Considering both early and late combination methods, ResNet50-w worked better. The FM trained with its activations achieved an average F_1 score on the validation set 0.63 p.p. higher than its counterpart trained with ResNet50 activations. When using ResNet50 predictions as input to the LF, the method simply ignored them and always followed the text module predictions. Thus, all FM and LF results we report were obtained using ResNet50-w as the image module.

Table 3 Test set F_1 scores (in %) of the main approaches for image, text, fusion and sequence classification

Class	Text		Image		Fusion			Sequence			
	CNN [27]	CNN	Res Net50-w	Res Net50	LF-F	FM	LXLM	CNN +CRF [27]	LF-F +CRF	FM +CRF	BiLSTM-F
Acórdão	86.43	89.96	18.45	06.78	90.81	90.74	60.13	90.60	**92.46**	91.56	88.97
ARE	55.92	55.72	11.33	00.00	57.92	57.92	23.91	59.54	59.96	60.74	**61.16**
Despacho	59.88	62.94	08.44	00.00	**65.64**	63.98	23.72	56.69	61.86	62.69	64.07
Others	97.30	97.31	61.72	95.02	97.21	97.24	96.54	**97.68**	97.61	97.67	97.46
RE	76.23	75.59	32.59	34.96	75.60	75.47	72.25	78.77	77.39	78.43	**79.67**
Sentença	79.29	80.53	43.52	48.67	82.19	82.04	66.68	81.13	84.52	83.42	**85.26**
Average	75.84	77.01	29.34	30.91	78.23	77.90	57.21	77.40	78.97	79.09	**79.43**
Weighted	94.72	94.72	58.09	87.67	94.70	94.72	92.81	95.33	95.25	**95.38**	95.30

Image results are reported for the image test set; all the others, for the text test set. We also report the CNN and CNN+CRF results presented in [27] and the LayoutXLM [44] baseline (LXLM). A majority class baseline achieves average/weighted F_1 scores of 15.73/84.41 and 15.71/84.07 on the text and image test sets respectively.
Bold values indicate the highest score for each class and aggregate measure

Table 4 Text classification: comparison between validation set precision, recall and F_1 scores (in %) of the different approaches

Class	CNN-w			CNN		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score
Acórdão	43.46	**82.27**	56.88	**91.16**	75.92	**82.85**
ARE	46.95	**67.05**	55.23	**75.04**	47.42	**58.11**
Despacho	37.20	**75.41**	49.82	**77.27**	65.03	**70.62**
Others	**97.55**	91.73	94.55	96.29	**97.97**	**97.12**
RE	55.61	**80.47**	65.77	**76.45**	74.56	**75.49**
Sentença	52.82	**78.55**	63.16	**92.85**	68.22	**78.65**
Average	55.60	**79.25**	64.24	**84.84**	71.52	**77.14**
Weighted	92.53	90.13	90.98	**94.37**	**94.60**	**94.37**

The suffix *-w* indicates the use of class frequency penalty weights.
Bold values indicate the highest precision, recall and F_1 score for each class and aggregate measure

Table 5 Image classification: comparison between validation set precision, recall and F_1 scores (in %) of the different approaches

Class	ResNet50-w			ResNet50		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score
Acórdão	09.91	**81.85**	**17.68**	**66.67**	01.27	02.50
ARE	**06.27**	**73.17**	**11.56**	00.00	00.00	00.00
Despacho	**03.99**	**68.31**	**07.53**	00.00	00.00	00.00
Others	**98.27**	46.37	63.01	90.47	**99.50**	**94.77**
RE	21.93	**62.65**	32.48	**78.08**	20.95	**33.03**
Sentença	30.99	**72.72**	43.46	**93.73**	33.35	**49.20**
Average	28.56	**67.51**	29.29	**54.82**	25.85	**29.92**
Weighted	**89.37**	48.69	59.13	87.14	**90.24**	**87.09**

The suffix *-w* indicates the use of class frequency penalty weights.
Bold values indicate the highest precision, recall and F_1 score for each class and aggregate measure

Table 6 Fusion Module: impact of number of hidden units and learnable embeddings for missing data on average validation set F_1 scores (in %)

Method	Average F_1
FM-512	74.49
FM-512-zero	68.02
FM-128	**75.70**
FM-128-zero	72.95

The suffix *-zero* indicates the use of vector of zeros for missing data (as opposed to using learnable embeddings).
Bold value indicates the highest value

5.3 Image and text combination results

Table 6 shows the performance of the FM trained for 10 epochs with different hyperparameter configurations. Using learnable embeddings for missing textual or visual data proved to be fundamental, improving average F_1 scores by 6.47 and 2.75 p.p. for the models with 512 and 128 hidden units, respectively. While the smaller model performed best, we hypothesise that with further parameter tuning and longer training the bigger model would surpass it.

Table 7 compares the scores of the alternative fusion approaches with the ones from the FM. All of them performed much worse, with decreases in average F_1 score ranging from 2.16 to 14.52 p.p. These results signal how the increase in performance seen by the FM is due to the fusion of data sources, not to different training conditions or model capacity—combining visual and textual data helps.

Table 8 compares the early and late fusion methods. On the text test split, where all samples have both image and text data, with the exception of four samples with no image data, the late fusion methods outperformed the FM by 0.33 p.p. (average F_1 score). On this test split, LF-u and LF-f achieve identical scores, since there are only 4 samples with a missing modality. When evaluating on the text + image test split, which has 8037 samples with no text data, the FM surpasses LF-f by 0.11 p.p. and LF-u severely underperforms when compared with LF-f, with a drop in average F_1 score greater than 13 p.p. These results show that using a representation for missing modality that accounts for the unbalanced nature of

Table 7 Fusion Module ablation, comparing the test set F_1 scores (in %) of the hybrid classifier and of a version of the fusion module that ignores image activations (w/o img acts), that is, always uses the missing image embedding

Class	Text test split		Text + image test split		
	FM	fusion w/o img acts	FM	HC-w	HC
Acórdão	90.74	88.27 (-2.47)	88.50	41.36 (-47.14)	87.68 (-0.82)
ARE	57.92	54.09 (-3.83)	56.60	49.02 (-7.58)	43.91 (-12.69)
Despacho	63.98	62.01 (-1.97)	63.79	42.71 (-21.08)	61.85 (-1.94)
Others	97.24	97.27 (+0.03)	97.03	95.80 (-1.23)	97.02 (-0.01)
RE	75.47	73.26 (-2.21)	75.05	72.11 (-2.94)	75.00 (-0.05)
Sentença	82.04	79.58 (-2.46)	81.21	74.07 (-7.14)	79.68 (-1.53)
Average	77.90	75.74 (-2.16)	77.03	62.51 (-14.52)	74.19 (-2.84)
Weighted	94.72	94.47 (-0.25)	94.32	92.58 (-1.74)	93.95 (-0.37)

For the hybrid classifier, we report the results using both image classifiers: with (HC-w) and without (HC) class frequency penalty. Between parentheses, the difference in performance compared with using the original fusion module (FM)

Table 8 Early and late fusion: comparison between test set F_1 scores (in %) of the fusion module (FM), late fusion methods (LF-u and LF-f) and LayoutXLM (LXLM)

Class	Text test split				Text + image test split		
	FM	LF-u	LF-f	LXLM	FM	LF-u	LF-f
Acórdao	90.74	**90.81**	**90.81**	60.13	**88.50**	41.52	88.48
ARE	**57.92**	**57.92**	**57.92**	23.91	**56.60**	50.68	54.96
Despacho	63.98	**65.64**	**65.64**	23.72	63.79	46.46	**65.31**
Others	**97.24**	97.21	97.21	96.54	**97.03**	95.70	96.96
RE	75.47	**75.60**	**75.60**	72.25	**75.05**	72.16	74.58
Sentença	82.04	**82.19**	**82.19**	66.68	81.21	75.70	**81.24**
Average	77.90	**78.23**	**78.23**	57.21	**77.03**	63.70	76.92
Weighted	**94.72**	94.70	94.70	92.81	**94.32**	92.56	94.19

Bold values indicate the highest score for each class in each modality (text or text + image)

Table 9 Sequence classification: comparison between average and weighted by class frequencies validation set F_1 scores (in %) of the different approaches

Method	Average F_1	Weighted F_1
BiLSTM	77.16	94.25
BiLSTM-CRF	78.45	94.46
BiLSTM-F	**79.03**	**94.81**
BiLSTM-F-CRF	78.87	94.58

Bold values indicate the highest Average and Weighted F_1 scores

the data is essential for late fusion approaches. Furthermore, the FM seems to handle samples with missing modalities better, possibly because of the learnable embeddings for those cases.

5.4 Sequence classification

Table 9 compares the validation performance of the LSTM models. To ensure a fair comparison to the other approaches, though we use IOB tagging scheme during training, when reporting results we consider only the original classes. If a given page is an ARE, for example, the predictions B-ARE and I-ARE would both be considered correct, regardless of the position of the page in its lawsuit.

The variants that use as input the image and text embeddings (BiLSTM-F and BiLSTM-F-CRF) outperformed the ones that use the FM activations (BiLSTM and BiLSTM-CRF). This suggests that it is beneficial to jointly learn how to consider sequential dependencies and how to combine multimodal information. Surprisingly, the CRF layer helped the BiLSTM model, with an increase in 1.29/0.25 average/weighted F_1 scores, but not the BiLSTM-F model. This may be an artefact of our training settings, with its limited number of training epochs.

5.5 First page evaluation

Table 10 shows the difference in classification performance of samples that are the first page of a document versus those that are interior pages, considering all levels of data availability (text, image, and fusion).

Table 10 Comparison of first page and not first page of a document classification performance

Class	Text		Image		Fusion	
	First page	Not first page	First page	Not first page	First page	Not first page
Acórdão	**92.47 (199)**	83.66 (74)	**34.28 (197)**	08.24 (88)	**93.40 (199)**	77.19 (88)
ARE	47.65 (213)	**56.74 (1,628)**	06.71 (203)	**12.10 (2,334)**	**59.95 (213)**	56.28 (2,442)
Despacho	**71.54 (147)**	40.43 (51)	**12.59 (146)**	03.68 (52)	**71.81 (147)**	40.45 (52)
Others	**99.02 (25,744)**	96.58 (59,664)	**78.19 (24,193)**	54.29 (63,709)	**99.04 (25,744)**	96.26 (66,789)
RE	74.45 (312)	**75.65 (6,019)**	18.28 (301)	**33.72 (5,876)**	**75.50 (312)**	75.03 (6,074)
Sentença	**81.47 (265)**	80.32 (1,210)	26.61 (262)	**49.71 (1,216)**	**83.11 (265)**	80.78 (1,238)
Average	**77.77 (26,880)**	72.23 (68,646)	**29.44 (25,302)**	26.96 (73,275)	**80.47 (26,880)**	71.00 (76,683)
Weighted	**97.96 (26,880)**	93.46 (68,646)	**75.65 (25,302)**	51.13 (73,275)	**98.11 (26,880)**	93.00 (76,683)

We report test set F_1 scores (in %) for image, text, and fusion classification using as models the CNN, the ResNet50-w and the FM, respectively. We show the number of samples between parentheses.
Bold values indicate the highest score for each class in each modality (text, image or fusion)

The first page sample set obtained average/weighted F_1 scores 5.54/4.50, 2.48/24.52 and 9.47/5.11 p.p. higher than its complement, for the text, image and fusion levels, respectively. These results confirm our hypothesis that the first pages are more informative from the point of view of both textual and visual data. Therefore, one possible improvement for page classification of the legal documents is training under a multitask setting that jointly learns to classify pages and establish document boundaries.

5.6 Error analysis

To examine the differences in performance between the classes, we first built confusion matrices for the BiLSTM-F predictions (Fig. 8). We identified that the main source of confusion is the "Others" class. When distinguishing first and interior pages we observe a equivalent behaviour, with first pages and interior pages being misclassified as B-Others and I-others, respectively.

To better understand such confusion we executed the following error analysis. For each class, we manually inspected two samples from the test set: the most confident false positive (i.e. the one with the highest loss), and a randomly sampled false positive. In total, we sampled 12 pages, exhibited in Fig. 9. We see that for all (non-Others) wrong predictions, the ground truth was "Others", which is not surprising when we consider the aforementioned confusion matrices. What was unexpected was that in some cases the model was actually correct—the problem was the ground truth label. Of the six most confident supposed errors, four were wrongly labelled samples whose actual class was correctly predicted. When considering the randomly sampled errors, four or them were also wrongly labelled. Of those, three were correctly predicted by the model.

Once we identified those labelling mistakes, we randomly sampled five pages from each class, 30 pages in total. We manually examined each of them to assess the correctness of the ground truth labels. We found all non-others labels to be correct. Of the five documents labelled as "Others", two were found to be incorrect. We have also found that many documents wrongly labelled as "Others" originated from the Brazilian state of *Paraíba*. For that reason, we examined all pages containing the token "paraíba" and found all of them to be labelled as "Others". Thus, we hypothesise that documents from certain lower courts are reaching the Supreme Court with no document type annotation and being wrongly identified as "Others". This would also explain the excellent classification performance of that class: the model may have learned to identify cues (e.g. lower court of origin) that identify wrongly labelled documents. Surprisingly, even with wrong labels the model learned the task well—we could only identify the labelling issues because of supposed wrong predictions that we found to be actually correct. We leave to future work identifying the sources of label noise.

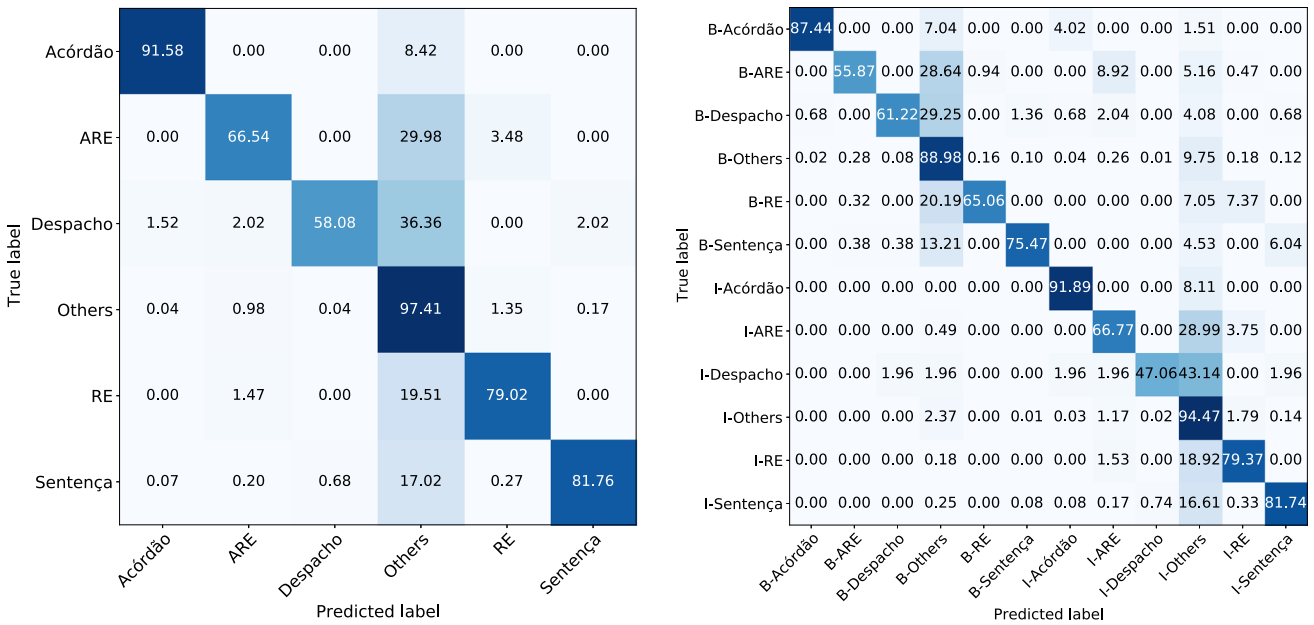
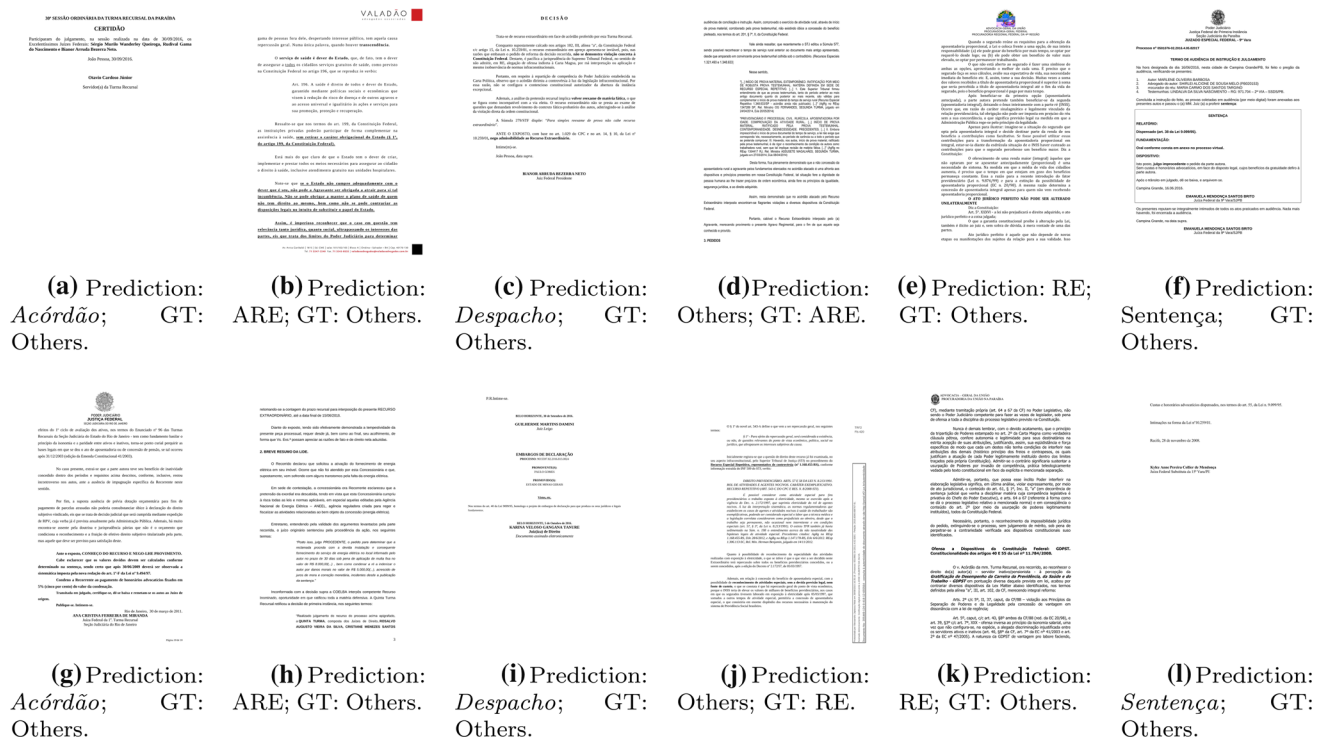


Fig. 8 Confusion matrices of the BiLSTM-F predictions for the test set. Values have been normalised by row and represent the percentage of samples from the row class that were classified as the column class. The entries in the main diagonal are the recalls for each class. To the left is the confusion matrix for class prediction; to the right, the confusion matrix for tag prediction, which distinguishes first and interior pages



(a) Prediction: *Acórdão*; GT: Others. **(b)** Prediction: ARE; GT: Others. **(c)** Prediction: *Despacho*; GT: Others. **(d)** Prediction: Others; GT: ARE. **(e)** Prediction: RE; GT: Others. **(f)** Prediction: Sentença; GT: Others.

(g) Prediction: *Acórdão*; GT: Others. **(h)** Prediction: ARE; GT: Others. **(i)** Prediction: *Despacho*; GT: Others. **(j)** Prediction: Others; GT: RE. **(k)** Prediction: RE; GT: Others. **(l)** Prediction: *Sentença*; GT: Others.

Fig. 9 Wrongly predicted examples for each class. Top row: most confident wrong prediction. Bottom row: randomly chosen wrong prediction. GT: ground truth label

5.7 Tobacco800 evaluation

Table 11 exhibits the accuracy of our main methods for each modality when trained and evaluated on Tobacco800. We report the accuracy instead of the F_1 score because the dataset is more balanced than ours (58% first page and 42% interior page samples) and the accuracy is the metric used in previous works [7,42].

Table 11 Tobacco800 test set performance (in %) across the different modalities

Method	Accuracy
Majority baseline	57.9
Text	
Text CNN	84.6
Image	
VGG16 [7]	87.8
ResNet50	81.9
Fusion	
Late Fusion	85.7
FM	87.6
Sequence	
Image + Text + Topic [42]	91.9
VGG16 (3-page window) [7]	**92.0**
BiLSTM-F	67.8

Using the FM represented an increase in accuracy of 3 p.p. when compared with our stronger unimodal approach (Text CNN).
Bold value indicates the highest accuracy

We compare our models with the best performing ones presented by Wiedemann and Heyer [42] and Braz et al. [7]. The former is based on the late fusion of image, text and topic-based classifier predictions of inputs that include the predecessor page, a way to include sequential information. The text module uses pre-trained FastText [6] word vectors followed by a bidirectional layer of Gated Recurrent Units [10], 1D convolutions of multiple filter sizes and a fully connected layer. The image module is based on VGG16 [36] with a modified linear block on top. Topic features are obtained through latent Dirichlet allocation [5]. The best performing model proposed by Braz et al. [7] is similar to the one used by Wiedermann and Heyer, but also includes the successor page as input.

Surprisingly, the best-performing method for tobacco800 is a VGG16 model with windows of 3 pages as input, which is a unimodal approach that uses only image features. That said, Braz et al. [7] have not explored combining features from multiple modalities, so we would expect that integrating their model in a multimodal pipeline such as the ones we present in this paper would further improve the accuracy.

The FM got comparable results without any hyperparameter tuning for tobacco800 apart from the learning rate—we reused the same configurations employed when processing

SVic+. In addition, our multimodal approaches show a clear improvement over the unimodal ones, especially our FM, which surpassed the late fusion model. This is contrary to previous works' findings [32,42], where late fusion performed best.

Since the image model from Braz et al. [7] outperformed ours by 5.9 p.p., the performance of the FM may be hampered by weak image and text models. Unfortunately, we cannot compare our text and fusion models with neither Braz et al. [7] nor Wiedemann and Heyer [42] methods. Braz et al. only explore image models. Wiedemann and Heyer [42] report the performance of their image and text models only on their validation set, which is not made available. That said, our fusion module is agnostic to its inputs and should improve with more robust image and text representations.

Excluding the majority baseline, the BiLSTM-F was the least accurate method. In fact, its performance was worse than the image and text models whose activations it uses as input. This may be due to the nature of the Tobacco800 dataset: while SVic+ is composed of lawsuits of related ordered documents, Tobacco800 is an unordered collection. Also, despite the presence of multi-page documents, single-page documents comprise the majority of the dataset. Therefore, the results suggest that the BiLSTM-F is more appropriate for sequences of pages with longer-term dependencies, while modelling windows of pages is the better method for corpora such as Tobacco800.

6 Conclusion

In this paper, we presented SVic+, a novel dataset of Brazilian lawsuits with visual and textual data, and proposed a method for sequence-aware multimodal classification of pages from legal documents. Our proposed fusion module combines visual and textual features extracted from convolutional neural networks trained separately on image and text data. We experiment with two approaches for sequence classification: post-processing the predictions of the multimodal methods using a linear-chain conditional random field and training bidirectional LSTM models that alternatively use as input fusion module activations or the concatenation of image and text embeddings. Our fusion module outperformed the unimodal models, with an ablation analysis confirming that improvement is due to the combination of modalities. It also surpassed LayoutXLM, a much more computating and time-intensive method, possibly due to the negative transfer between LayoutXLM's pre-training and target tasks. We find that learning embeddings for missing visual or textual input is much better than using a vector of zeroes for such cases.

Sequence classification of pages brought further improvements, with the best performing model jointly learning how to combine modalities and consider sequential dependencies. That said, the results on Tobacco800 show that the BiLSTM-F is better suited to model sequences of pages with longer-term dependencies, as opposed to unordered collections of documents. Finally, through our error analysis we have found data labelling errors that seem to originate from lower courts that submit documents with no type annotation.

Therefore, future work would include identifying the sources of label noise and correcting the annotations. It would also be interesting to examine the end-to-end training of the full pipeline: image and text feature extractors, fusion module and sequence modelling. Moreover, it is worthwhile to explore if transformer-based [41] text encoders such as BERT [13] and T5 [29] can further raise classification metrics by improving the text module.

Funding This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. TdC received support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grant PQ 314154/2018-3. We acknowledge the support of "Projeto de Pesquisa & Desenvolvimento de aprendizado de máquina (machine learning) sobre dados judiciais das repercussões gerais do Supremo Tribunal Federal - STF". We are also grateful for the support from Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF, project KnEDLe, convênio 07/2019) and Fundação de Empreendimentos Científicos e Tecnológicos (Finatec). TdC is currently on a leave of absence from the University of Brasilia and works at Vicon Motion Systems, Oxford Metrics Group.

Data availability Data used in this work is available at http://ailab.unb.br/victor/lrec2020/.

Code Availability Code used in this work is available at https://github.com/peluz/victor-visual-text.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Agam, G., Argamon, S., Frieder, O., Grossman, D., Lewis, D.: The Complex Document Image Processing (CDIP) test collection project (2006). http://ir.iit.edu/projects/CDIP.html
2. Audebert, N., Herold, C., Slimani, K., Vidal, C.: Multimodal deep networks for text and image-based document classification. CoRR **abs/1907.06370** (2019). http://arxiv.org/abs/1907.06370
3. Bakkali, S., Ming, Z., Coustaty, M., Rusinol, M.: Visual and textual deep feature fusion for document image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2020)
4. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. CoRR **abs/2004.05150** (2020). https://arxiv.org/abs/2004.05150
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003). http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguistics **5**, 135–146 (2017). https://doi.org/10.1162/tacl_a_00051.
7. Braz, F.A., da Silva, N.C., Lima, J.A.S.: Leveraging effectiveness and efficiency in page stream deep segmentation. Eng. Appl. Artif. Intell. **105**, 104394 (2021). https://doi.org/10.1016/j.engappai.2021.104394.
8. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp. 108–122 (2013)
9. Chen, N., Blostein, D.: A survey of document image classification: problem statement, classifier architecture and performance evaluation. Int. J. Document Anal. Recogn. (IJDAR) **10**(1), 1–16 (2007). https://doi.org/10.1007/s10032-006-0020-2
10. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR **abs/1412.3555** (2014). http://arxiv.org/abs/1412.3555
11. Conneau, A., Schwenk, H., Barrault, L., Lecun, Y.: Very deep convolutional networks for text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 1107–1116. Association for Computational Linguistics, Valencia, Spain (2017). http://www.aclweb.org/anthology/E17-1104
12. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inform. Sci. **41**(6), 391–407 (1990)
13. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018). http://arxiv.org/abs/1810.04805
14. Dimmick, D., Garris, M., Wilson, C., Flanagan, P.: Nist special database 2 - structured forms database users' guide (2017). https://doi.org/10.6028/NIST.NSRDS.2-2017
15. Engin, D., Emekligil, E., Oral, B., Arslan, S., Akpınar, M.: Multimodal deep neural networks for banking document classification. In: International Conference on Advances in Information Mining and Management, pp. 21–25 (2019)
16. Ford, G., Thoma, G.R.: Ground truth data for document image analysis. In: Symposium on document image understanding and technology (SDIUT), pp. 199–205 (2003)
17. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: International Conference on Document Analysis and Recognition (ICDAR), pp. 991–995. IEEE (2015)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
19. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
20. Howard, J., 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), pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1031
21. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning - Volume 37, pp. 448–456. JMLR.org (2015). http://proceedings.mlr.press/v37/ioffe15.html
22. Jain, R., Wigington, C.: Multimodal document image classification. In: International Conference on Document Analysis and Recognition (ICDAR), pp. 71–77 (2019). https://doi.org/10.1109/ICDAR.2019.00021
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optmisation. In: International Conference on Learning Representations (ICLR) (2015). Preprint available at https://arxiv.org/abs/1412.6980
24. Kumar, J., Ye, P., Doermann, D.: Structural similarity for document image classification and retrieval. Pattern Recogn. Lett. **43**, 119–126 (2014)
25. Lafferty, J.D., Andrew, M., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML, pp. 282–289. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
26. Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., Heard, J.: Building a test collection for complex document information processing. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, p. 665-666. Association for Computing Machinery, New York, NY, USA (2006). https://doi.org/10.1145/1148170.1148307
27. Luz de Araujo, P.H., de Campos, T.E., Ataides Braz, F., Correia da Silva, N.: VICTOR: a dataset for Brazilian legal documents classification. In: Proceedings of The 12th Language Resources and Evaluation Conference (LREC), pp. 1449–1458. European Language Resources Association, Marseille, France (2020). https://www.aclweb.org/anthology/2020.lrec-1.181
28. Mota, C., Lima, A., Nascimento, A., Miranda, P., 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, pp. 318–329. SBC, Porto Alegre, RS, Brasil (2020). https://doi.org/10.5753/eniac.2020.12139
29. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020). http://jmlr.org/papers/v21/20-074.html
30. Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Natural language processing using very large corpora, pp. 157–176. Springer (1999). https://doi.org/10.1007/978-94-017-2390-9_10. Preprint available at http://arxiv.org/abs/cmp-lg/9505040
31. Rosenstein, M.T., Marx, Z., Kaelbling, L.P., Dietterich, T.G.: To transfer or not to transfer. In: In NIPS'05 Workshop, Inductive Transfer: 10 Years Later (2005)
32. Rusiñol, M., Frinken, V., Karatzas, D., Bagdanov, A.D., Lladós, J.: Multimodal page classification in administrative document image streams. Int. J. Document Anal. Recogn. (IJDAR) **17**(4), 331–341 (2014). https://doi.org/10.1007/s10032-014-0225-8
33. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Visi. (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
34. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
35. Secretaria de Comunicação Social do Conselho Nacional de Justiça: Sumário executivo do relatório justiça em números 2020 (2018). https://www.cnj.jus.br/wp-content/uploads/2020/08/WEB_V2_SUMARIO_EXECUTIVO_CNJ_JN2020.pdf
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)

37. Smith, L.N.: Cyclical learning rates for training neural networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 464–472 (2017). https://doi.org/10.1109/WACV.2017.58
38. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. CoRR **abs/1708.07120** (2017). http://arxiv.org/abs/1708.07120
39. Smith, R.: An overview of the Tesseract OCR engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR), vol. 2, pp. 629–633. IEEE (2007)
40. Supremo Tribunal Federal: Ministra Cármen Lúcia anuncia início de funcionamento do Projeto Victor, de inteligência artificial (2018). http://www.stf.jus.br/portal/cms/verNoticiaDetalhe.asp?idConteudo=388443
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017). http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
42. Wiedemann, G., Heyer, G.: Multi-modal page stream segmentation with convolutional neural networks. Language Res. Evalu. (2019). https://doi.org/10.1007/s10579-019-09476-2
43. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-Training of Text and Layout for Document Image Understanding, p. 1192–1200. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3394486.3403172
44. Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., Wei, F.: Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding (2021)
45. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., Zhou, L.: LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 2579–2591. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.201