Classification of Court Lawsuits Pages using Multimodal Convolution Neural Networks

Caio C. R. Mota[1], **Andressa L. S. de Lima**[1], **André C.A. Nascimento**[1], **Péricles B.C. Miranda**[1], **Rafael F. L. de Mello**[1]

[1] Departamento de Computação
Universidade Federal Rural de Pernambuco (UFRPE), Recife – PE, Brazil

{andre.camara}@ufrpe.br

Abstract. *The classification and separation of documents is a crucial step in the analysis of court lawsuits. Deep learning algorithms have achieved promising results in this task, extracting relevant information from the texts of these documents. However, court documents have become increasingly heterogeneous, i.e., photos, receipts, text documents, etc., directly impacting classification accuracy. This work investigates the use of multimodal convolutional networks, combining characteristics extracted from texts and images to classify court lawsuits pages. Two multimodal approaches were compared with four single-mode. In terms of accuracy and kappa, all algorithms were evaluated in a database composed of 117 lawsuits. The results showed that the approach that achieved the best performance is multimodal, presenting effectiveness and efficiency in the classification of court lawsuits pages.*

Resumo. *A classificação e separação de documentos é uma etapa de extrema importância na análise de processos judiciais. Algoritmos de aprendizado profundo têm alcançado resultados promissores nesta tarefa, extraindo informações relevantes a partir dos textos destes documentos. No entanto, os documentos de processos judiciais têm se tornado cada vez mais heterogêneos, i.e. fotos, recibos, documentos de texto, etc., impactando diretamente a precisão na classificação. Este trabalho investiga o uso de redes convolucionais multimodais, combinando características extraídas de textos e imagens, para classificação de páginas de processos. Duas abordagens multimodais foram comparadas com quatro monomodais. Todos os algoritmos foram avaliados, em termos em acurácia e kappa, em uma base de dados composta por 117 processos judiciais. Os resultados mostraram que a abordagem que atingiu o melhor desempenho é multimodal, apresentando eficácia e eficiência na classificação de páginas de processos.*

1. Introdução

Nos últimos anos, com a crescente adoção dos tribunais de justiça brasileiros ao modelo de processo eletrônico, uma série de desafios tecnológicos tem sido identificados no gerenciamento de tal volume de documentos. De acordo com o relatório *Justiça em Números* do Conselho Nacional de Justiça – CNJ, em 2018, os 92 tribunais brasileiros receberam um total de 28 milhões de novos casos. Desse total, cerca de $79,7\%$ estão totalmente em meio eletrônico [Toffoli and Gusmão 2019].

No submissão de processos por meio eletrônico, além da própria petição inicial, também podem ser incluídos formulários, documentos de identificação, recibos, cópias de emails, dentre outros. Por questões de praticidade, é comum que esses diferentes documentos sejam digitalizados em lotes de páginas seguidas sem nenhuma separação manual. Deste modo, um único arquivo em formato .pdf é enviado via sistema. Vale salientar que a maioria dessas digitalizações são geradas pelos próprios usuários dos sistemas, utilizando qualquer meio que tenham acesso, como por exemplo, scanners, câmeras e smartphones. A Figura 1 apresenta um exemplo de submissão que inclui o processo, assim como outros diferentes tipos de documentos anexos.

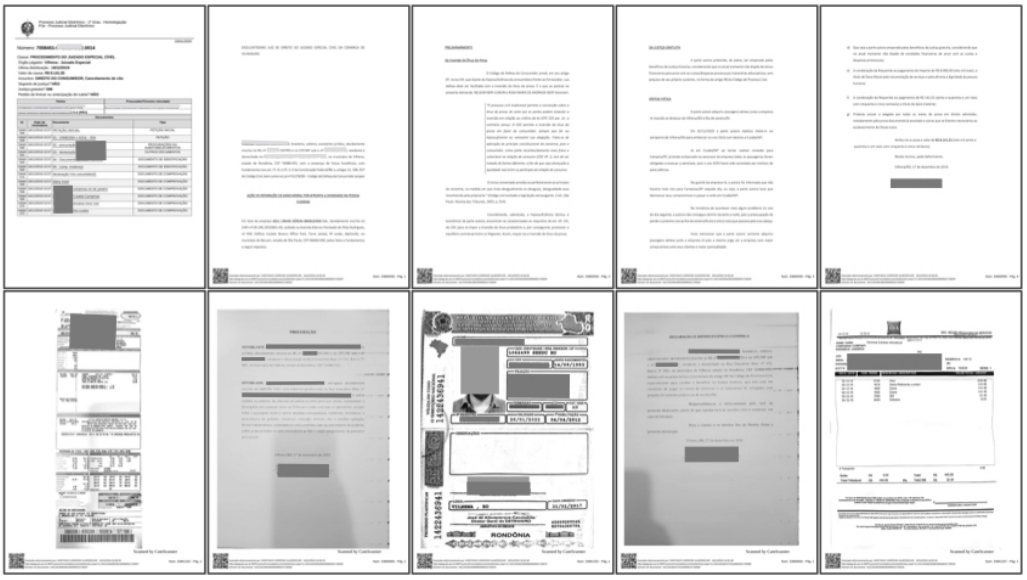

Figura 1. Exemplos de documentos da base de dados de petições iniciais consideradas. Tais documentos normalmente contém folhas de rosto do processo (primeira imagem), páginas da petição inicial (quatro documentos mais a direita na primeira linha), e anexos (segunda linha)

Nesse cenário, as imagens digitalizadas de documentos com várias páginas chegam a um sistema de gerenciamento de documentos como um fluxo ordenado de páginas únicas, mas sem informações sobre os limites dos documentos. A segmentação por fluxo de páginas (*Page Stream Segmentation* - PSS) é a tarefa de separar o fluxo contínuo de documentos em sequências de páginas que representam documentos físicos únicos. A aplicação de uma abordagem totalmente automatizada de PSS é fundamental em contextos de conjuntos de dados muito grandes[Wiedemann and Heyer 2019].

Em geral, o problema de PSS pode ser visto como uma instância da classificação de textos, a qual vem sendo abordada com sucesso utilizando métodos baseados em aprendizagem de máquina [Wiedemann and Heyer 2019, Audebert et al. 2019]. Mais especificamente, a tarefa que consideramos é a categorização automática de cada página de acordo com seu conteúdo (semântico). O conteúdo de cada página pode ser analisado tanto em seu formato de imagem, como também na forma do texto extraído, com o uso de sistemas de reconhecimento óptico de caracteres (OCR). A classificação de documentos baseada em imagens é um tópico de pesquisa bastante maduro, diferentes abordagens já propostas na literatura [Chen and Blostein 2007]. Outros realizam a classificação de páginas de documentos baseados numa representação textual do mesmo

[Kowsari et al. 2019]. Ambas as abordagens são em geral baseadas em técnicas de aprendizagem de máquina supervisionada, comuns em problemas de classificação (e.g., k-NN, árvores de decisão, SVM, redes neurais, etc.).

Aprendizado de máquina também tem sido aplicado com frequência na análise de documentos jurídicos [Undavia et al. 2018, Hammami et al. 2019, de Araujo et al. 2020]. De acordo com [de Araujo et al. 2020], algoritmos de aprendizagem clássicos (i.e. *Support Vector Machine*, *Naive Bayes*), bem como redes neurais convolucionais (*Convolutional Neural Networks* - CNN) têm sido usadas, e alcançado resultados promissores. Porém, os trabalhos identificados para o problema do PSS no contexto de processos judiciais consideram apenas as características textuais dos documentos. Como os documentos envolvidos no processo possuem naturezas heterogêneas (i.e. digitalizações, fotos, pdfs), a análise exclusiva do texto trás desafios para uma segmentação eficaz.

Mais recentemente, uma série de trabalhos [Jain and Wigington 2019, Gallo et al. 2016, Wiedemann and Heyer 2019, Audebert et al. 2019] propuseram a criação de modelos de segmentação que levam em consideração as informações que estão presentes tanto nas imagens quanto nos textos, em um processo de integração de informações. Os trabalhos supracitados consideraram documentos provenientes de arquivos de grandes empresas, como a base de dados Tobacco800 [Wiedemann and Heyer 2019]. Embora os trabalhos supracitados não tenham sido aplicados em documentos jurídicos de processos eletrônicos, a natureza dos documentos é similar. Os resultados mostraram que a adoção de redes neurais convolucionais multimodais são promissoras para o problema de PSS.

Este trabalho investiga o desempenho de redes neurais convolucionais (*Convolutional Neural Networks* - CNN) multimodais para o problema de PSS aplicado à documentos jurídicos de processos eletrônicos. De acordo com o nosso conhecimento, este seria o primeiro trabalho a abordar o problema de PSS em documentos jurídicos de forma automática através da integração de características visuais e textuais. Deste modo, foram selecionadas duas abordagens multimodais, e quatro monomodais, sendo duas focadas na classificação de textos, e outras duas especialistas em imagens.

As abordagens selecionadas foram avaliadas em uma base de dados composta por 117 processos judiciais, totalizando 2970 páginas. Os resultados mostraram que a melhor abordagem, em termos de acurácia e kappa, foi multimodal e combina o algoritmo FastText (especialista em texto) com o VGG16 (especialista em imagem). Além de eficaz, esta combinação também se mostrou eficiente, sendo capaz de atingir bons resultados com pouco tempo de treinamento.

O restante deste artigo está estruturado da seguinte forma: A seção 2 apresenta trabalhos relacionados ao problema de PSS. As Seção 3 destaca a pergunta de pesquisa deste trabalho. A Seção 4 detalha a proposta deste trabalho, e apresenta duas arquiteturas multimodais que foram adaptadas para a classificação de processos judiciais. Na seção 5, apresentamos a metodologia adotada para a realização dos experimentos. Em seguida, a seção 6 apresenta uma avaliação quantitativa dos resultados. Por fim, a conclusões e possíveis extensões a este trabalho são discutidas na seção 7.

2. Trabalhos relacionados

As principais estratégias de classificação de páginas de documentos podem ser categorizadas em sistemas baseados em regras (SBR) e sistemas baseados em aprendizagem de máquina (AM) [Gallo et al. 2016, Wiedemann and Heyer 2019]. Como o SBR em geral não generaliza bem para repositórios de documentos heterogêneos do mundo real, e também exige um esforço maior para projetar manualmente descritores relevantes, as abordagens baseadas em AM para o problema de PSS se tornaram mais populares [Wiedemann and Heyer 2019]. Dentre as abordagens baseadas em AM, algumas utilizam a estrutura da sequência de páginas para produzir um classificador, normalmente utilizando modelos probabilísticos sequenciais, como os modelos de Markov ocultos (HMM). Em [Frasconi et al. 2002], um HMM é treinado a partir de sequências de páginas rotuladas, cujas emissões são uma representação de *bag-of-words* extraída de cada página. Dado um novo documento (sem páginas rotuladas), o algoritmo gera uma sequência de categorias de páginas com maior probabilidade a posteriori.

Em [Rusiñol et al. 2014], o PSS é abordado como um problema de classificação multimodal para classificação de 70.000 páginas de documentos administrativos do domínio bancário. Os autores avaliaram estratégias de combinação prévia e tardia de descritores visuais (de imagem) de uma determinada página, juntamente com os recursos de texto. Ou seja, na combinação prévia, múltiplas representações (e.g., vetores) dos padrões (ou exemplos) são concatenados para formar um único vetor, o qual é utilizado para treinar um único classificador. A integração tardia por sua vez, consiste em treinar diferentes classificadores, um para cada representação, e posteriormente combiná-los em um outro classificador final. Neste trabalho, os autores utilizaram como descritores visuais uma representação hierárquica da distribuição de intensidade de pixel, enquanto os descritores textuais envolviam uma técnica de modelagem de tópicos, a análise semântica latente (LSA), para representar o conteúdo da página como uma mistura de tópicos. Os classificadores considerados foram o Naive Bayes, K-*nearest neighbors* (K-NN) e *Support Vector Machine* (SVM). Finalmente, o método proposto usa um modelo de *n* gramas sobre o fluxo de páginas, para obter uma classificação mais refinada das páginas. Os autores demonstraram que melhores resultados foram obtidos pela fusão tardia, obtendo $96,84\%$ de acurácia nos experimentos realizados.

A fim de segmentar um conjunto de aproximadamente um milhão de documentos retro-digitalizados do governo alemão, datados entre 1922 e 2010, [Wiedemann and Heyer 2019] utilizou uma abordagem de classificação binária com CNN, sobre textos extraídos por OCR e imagens. A CNN utilizada considerou tanto uma representação vetorial das imagens digitalizadas quanto do conteúdo de texto das mesmas (i.e., *embeddings*), os quais foram combinados posteriormente e fornecidos como entrada para uma rede totalmente conectada. Além da base de dados governamental (não disponibilizada publicamente), os autores avaliaram o desempenho na base de dados Tobacco800 [Wiedemann and Heyer 2019]. A arquitetura de CNN proposta alcançou acurácia de 95% com os documentos internos e 93% na base Tobacco800.

O problema do PSS é relevante e recorrente em diferentes domínios. No entanto, no contexto jurídico percebe-se a ausência de literatura relacionada. Embora os domínios dos documentos considerados nos trabalhos relacionados sejam diferentes, a natureza é similar ao domínio jurídico. Portanto, servem de inspiração para o desenvolvimento

de possíveis soluções. Trabalhos relacionados recentes [Wiedemann and Heyer 2019, Audebert et al. 2019], mostraram o potencial do uso de abordagens multimodais, levando em conta características visuais e textuais dos documentos envolvidos, para o problema de PSS. Inspirado nestes trabalhos, o presente artigo propõe um modelo de CNN multimodal para a segmentação automática de processos judiciais eletrônicos.

3. Pergunta de Pesquisa

Este trabalho investiga o desempenho de CNNs multimodais, que levam em conta características visuais e textuais, na classificação e segmentação automática de páginas de processos judiciais eletrônicos.

> **PERGUNTA DE PESQUISA 1:** *Algoritmos de CNN multimodais são mais eficazes que as abordagens monomodais para o problema de PSS em processos judiciais?*
> **PERGUNTA DE PESQUISA 2:** *Modelos multimodais são eficientes o suficiente, na classificação de páginas de documentos judiciais, para serem usados em ferramentas e aplicações reais da esfera judiciária?*

4. Proposta

Este trabalho trata o problema da segmentação de processos judiciais eletrônicos como um problema de classificação de documentos de múltiplas classes com fusão de características. Como se pode ver na Figura 2, cada documento, composto por diferentes páginas, é submetido a dois processos de extração de características, i.e., *embedding*: textos e imagens. As representações aprendidas são então combinadas, em um processo também chamado de *early fusion*, ou seja, os vetores correspondentes ao domínio textual são concatenados aos vetores do domínio de imagem, para em seguida serem apresentados a uma rede neural multicamadas. Duas arquiteturas multimodais foram adotadas para o problema em questão, e serão apresentadas a seguir.

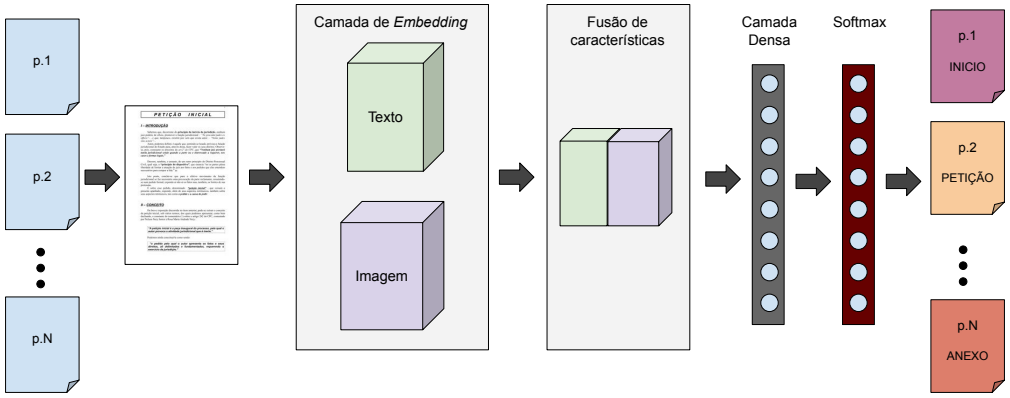

Figura 2. Ilustração da fusão de características multimodais (heterogêneas) de texto e imagem, para classificação de páginas de petições iniciais

4.1. Texto: CharCNN + Imagem: CNN2D

A arquitetura CharCNN+CNN2D é baseada em características aprendidas diretamente sobre o conjunto de treinamento, sem a utilização de transferência de aprendizado (*transfer*

learning). Nesse caso, a extração de características textuais é obtida por uma rede neural convolucional de caracteres (CharCNN) [Zhang and LeCun 2015]. Neste tipo de rede, a codificação é feita mediante um alfabeto de tamanho *m* para o idioma de entrada. A sequência de caracteres é transformada em uma sequência de vetores de tamanho *m* com comprimento fixo l_0, correspondente a quantidade de entradas da rede. Qualquer caractere que exceda o comprimento l_0 é ignorado e qualquer caractere que não esteja no alfabeto, incluindo caracteres em branco, é quantizado como vetor zero. O alfabeto usado consiste em 70 caracteres, incluindo 26 letras, 10 dígitos, 33 outros caracteres e o caractere de quebra de linha. Assim sendo, o alfabeto é composto pelos seguintes símbolos:

```
abcdefghijklmnopqrstuvwxyz0123456789
-,;.!?:'/_$%^*~`+-=<>()[]{}
```

A extração de características visuais foi obtida utilizando uma rede convolucional de duas dimensões, também conhecida como *shallow CNN* [Chollet 2018]. Esta arquitetura consiste em uma sequência de camadas de convolução 2D, com função de ativação 'relu', de dimensões 32, 64, 128 e 128, intercaladas com camadas de *max pooling* 2D 2x2. Ao final, é utilizada uma camada densa de 512 neurônios e um neurônio de saída, com função de ativação sigmoide.

4.2. Texto: FastText + Imagem: VGG-16 com Transferência de Aprendizado

Esta arquitetura multimodal foi proposta por [Wiedemann and Heyer 2019], e originalmente projetada para classificação binária (ou seja, para detectar continuidade ou ruptura de documentos). Neste trabalho, esta arquitetura foi adaptada ao contexto multi-classe. Este modelo utiliza princípios de transferência de aprendizado para aumentar o potencial de generalização da rede neural. Dessa forma, foram ajustados os parâmetros de função de custo (*loss*) da rede neural, de *binary_crossentropy* pra *categorical_crossentropy*, bem como foi modificada função de ativação (de sigmoid para softmax) e número de neurônios da camada de saída.

A arquitetura de extração de características textuais (FastText) utilizou a mesma arquitetura descrita em [Wiedemann and Heyer 2019]: a primeira camada consiste na conversão dos tokens em representações vetoriais, neste caso, utilizamos o conjunto pré-treinado de vetores de 300 dimensões `wiki-pt` do FastText[1]. A camada seguinte é composta por *Gated Recurrent Units* (GRU 's) com 128 células, seguidas por uma camada de convolução com janelas de $k = \{3, 4, 5\}$ tokens. A camada de saída deste módulo consiste em uma camada densa de 128 unidades. Vale salientar que o vocabulário de caracteres adotados nesta arquitetura foi o mesmo mencionado na seção 4.1.

A extração de características visuais utilizou a mesma arquitetura proposta em [Wiedemann and Heyer 2019], ou seja, considerando a representação binarizada e redimensionada de cada página (224x224 pixels), que por sua vez é submetida a um modelo VGG16 pré-treinado com a base ImageNet. Essa arquitetura contém 13 camadas de convolução agrupadas em cinco blocos sequencialmente encadeados. Cada bloco de convolução é finalizado com uma camada global de *max-pooling*. As camadas densas finais, assim como em [Wiedemann and Heyer 2019], foram substituídas por duas camadas, uma com 256 unidades com função de ativação 'Leaky RELU' e uma camada de

[1] `https://fasttext.cc/docs/en/pretrained-vectors.html`

predição, no nosso caso, com 14 neurônios de saída (um para cada classe) com função de ativação softmax. O treinamento foi feito utilizando usando a otimização de Adam com momento de Nesterov e uma taxa de aprendizado baixa (0, 00005) e mini-lotes de tamanho 32.

5. Metodologia Experimental

No presente trabalho é feita uma investigação do uso de CNNs multimodais para o problema de PSS no contexto de processos judiciais. Para isso, foram selecionados quatro abordagens: a CharCNN e a FastText, que levam em consideração informações textuais; e a CNN2D e a VGG16, que levam em consideração informações extraídas de imagens. Neste artigo, foram definidas duas versões multimodais (Texto+Imagem) a partir das abordagens adotadas: CharCNN+CNN2D e FastText+VGG16. As seis abordagens consideradas foram avaliadas quantitativamente em termos de eficácia e eficiência na classificação de páginas de processos judiciais.

5.1. Base de Dados

O conjunto de dados utilizado para avaliação consiste em um total de 117 documentos de processos judiciais de tribunais brasileiros, de acesso público, totalizando 2970 páginas. As páginas de cada documento foram submetidas ao processo de OCR (*optical character recognition*), realizado com o framework TesseractOCR.

Tabela 1. Características da base de dados.

# documentos	117
# classes	13
# páginas	2970
média de páginas por documento	25.38

A Tabela 2 apresenta algumas estatísticas da base, onde podemos observar que a número médio de páginas de documentos é considerável (> 20). A Figura 5.1 apresenta um histograma da distribuição de páginas por documento. As páginas foram classificadas em um total de 13 classes, distribuídas conforme apresentado na Figura 5.1.

Podemos observar o alto grau de desbalanceamento entre as classes. Muitas delas estão relacionadas ao processo de digitalização em si, como por exemplo, 'Erro', 'Página em branco', 'Movimentação' e 'Folha de rosto'. A classificação de tais páginas é de fundamental importância, como etapa de pré-processamento em sistemas de extração de informação, uma vez que pode reduzir drasticamente o número de páginas a serem processadas.

5.2. Medidas de Avaliação

O conjunto de dados de petições iniciais foi dividido em dois subconjuntos de treinamento (70%) e teste (30%), com 81 (2279 páginas) e 36 (621 páginas) documentos, respectivamente, selecionados aleatoriamente. Para cada método (com ou sem transferência de conhecimento) foram avaliadas três diferentes estratégias de classificação: utilizando apenas características textuais, visuais ou ambas (multimodal). Como medida de avaliação, foram utilizadas as métricas de acurácia e estatística Kappa. Esta última é uma medida de concordância, corrigida pelo acaso, entre as classificações e as classes verdadeiras.

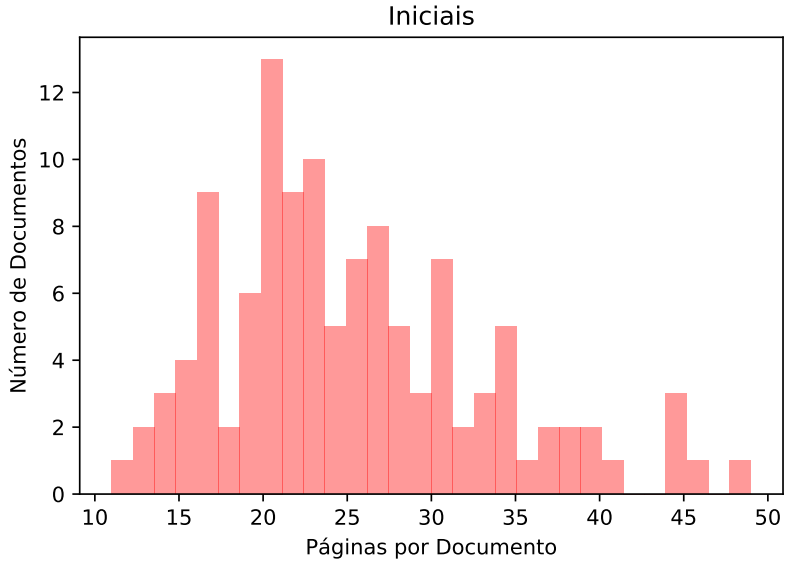

Figura 3. Distribuição do número de páginas por documento na base de petições iniciais.

É calculada considerando a concordância esperada quando as classes são atribuídas ao acaso, dividida pela maior concordância possível. Um valor maior que 0 significa que o classificador está se saindo melhor que o acaso. As métricas de acurácia (Acc) e Kappa são dadas por:

$$Acc = \frac{\textit{\#predições corretas}}{\textit{\#total de predições}}, \tag{1}$$

$$Kappa = \frac{(Acc – Acc_{aleatória})}{(1 - Acc_{aleatória})}, \tag{2}$$

onde, $Acc_{aleatória}$ é uma atribuição aleatória de classes.

5.3. Recursos de Hardware e Software

Os experimentos foram realizados em um computador com processador Ryzen 5 2600x, 16GB ram DDR4. As bibliotecas utilizadas foram, Tensorflow versão 1.15.0, Keras versão 2.3.1, Numpy versão 1.18.3, Pandas versão 1.0.3, Fasttext versão 0.9.1.

6. Resultados

Nesta seção são apresentados os resultados obtidos, sendo analisados sob os aspectos de eficácia e eficiência.

6.1. Análise da Eficácia

A Tabela 2 apresenta os resultados das abordagens monomodais e multimodais em termos de acurácia e kappa considerando-se 200 épocas. O número de épocas selecionado foi definido empiricamente, baseado no número de documentos utilizados neste trabalho. O valor escolhido extrapola o número de épocas necessário para o aprendizado dos algoritmos, garantindo, portanto, que nenhum algoritmo melhore seu desempenho com mais épocas.

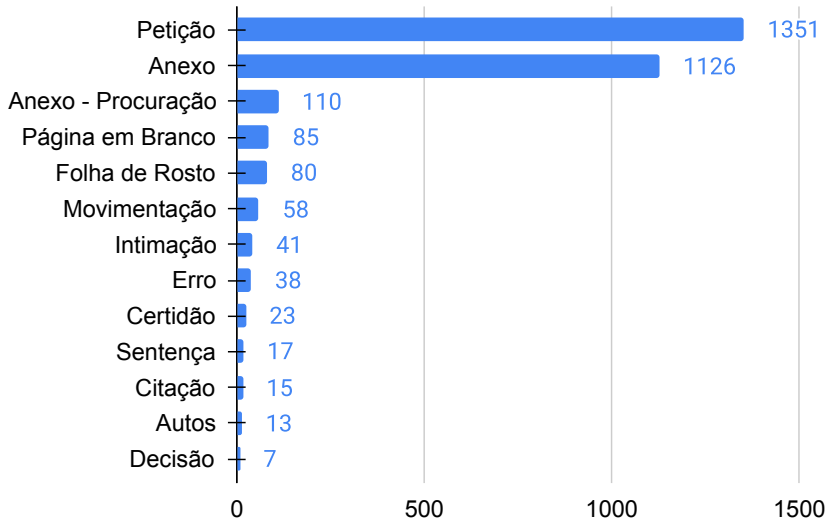

Figura 4. Distribuição de classes na base de dados.

Como se pode ver, o modelo CharCNN + CNN2D atingiu resultados promissores superando suas versões monomodais, as abordagens CharCNN e CNN2D, tanto em acurácia como em kappa. Vale salientar que a arquitetura utilizando CharCNN como entrada é eficiente para bases grandes, com milhões de exemplos para treino, como demonstrado por [Zhang and LeCun 2015]. Sendo assim, como este trabalho considera uma base de dados pequena de 117 documentos, o desempenho da CharCNN e de sua versão multimodal foi prejudicado. Dentre as abordagens monomodais, a FastText foi a que obteve o melhor resultado, superando, inclusive, o modelo CharCNN + CNN2D em acurácia e kappa. O modelo do FastText já conta com vetores de palavras pré treinados. Isto ajuda em bases de dados pequenas, pois trás um maior contexto sem que seja necessário aprender este contexto direto da base treinada. O FastText quando combinado com o VGG16 (modelo FastText+VGG16) teve seu desempenho melhorado ainda mais, alcançando os maiores valores médios de acurácia e kappa em relação a todas as demais abordagens.

Tabela 2. Desempenho médio das abordagens em termos de acurácia e kappa.

Modelo	**Acurácia**	**Kappa**
CharCNN (Texto)	0.8696	0.7901
FastText (Texto)	0.9483	0.9196
CNN2D (Imagem)	0.9173	0.8695
VCG16 (Imagem)	0.9376	0.8998
CharCNN + CNN2D (Texto + Imagem)	0.9189	0.8706
FastText + VCG16 (Texto + Imagem)	**0.9581**	**0.9335**

A Figura 6.1 apresenta a matriz de confusão obtida ao utilizar o FastText + VGG16 sobre a base de testes. Os resultados mostram que o classificador apresenta alta precisão, mesmo a base sendo desbalanceada. Vale ressaltar que esta precisão é ainda melhor quando se trata de classes mais frequentes no conjunto de dados analisado. Um ponto importante é que, dado que a divisão dos conjuntos de treinamento e teste é feita sobre os documentos, e não sobre as páginas, isso fez com que algumas classes não fossem contempladas no conjunto de testes, tais como: autos, citação, página de erro e sentença. Como essas classes estão presentes em poucos documentos (ver Figura 5.1), na divisão

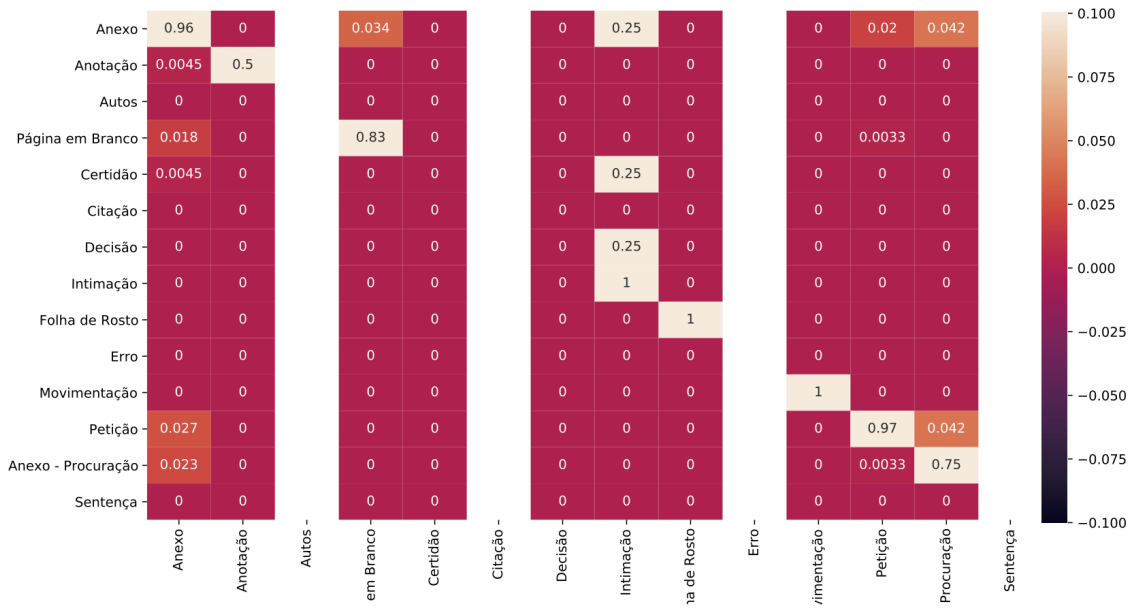

Figura 5. Matriz de confusão.

ficaram de fora do conjunto de teste. Em trabalhos futuros, com o aumento da base de dados, pretende-se evitar cenários como este.

Os resultados apresentados mostraram que a adoção de abordagens multimodais, que combinam as características textuais e visuais, podem melhorar o desempenho individual de cada algoritmo, e consequentemente alcançar melhores resultados em termos de classificação. Deste modo, a *Pergunta de pesquisa 1* é respondida positivamente, mostrando o potencial da combinação de modelos, principalmente no contexto deste trabalho, que é classificação de páginas em processos judiciais eletrônicos.

6.2. Análise da Eficiência

A Tabela 3 apresenta a complexidade arquitetural de cada modelo (segunda coluna); o custo de tempo de treinamento (este experimento considerou 200 épocas), em segundos (terceira coluna); e o custo de tempo de teste/validação (36 documentos, com 621 páginas), em segundos.

Tabela 3. Complexidade arquitetural, tempo de treino, e tempo de teste, obtidos pelas abordagens.

Modelo	**#Parâmetros**	**Tempo de treino (s)**	**Tempo de teste (s)**
CharCNN	1.023.206	4.000	1,34
FastText	1.012.996	4.000	2,95
CNN2D	116.526	5.800	0,98
VCG16	27.695.182	79.000	76,67
CharCNN + CNN2D	1.129.508	7.400	2,01
FastText + VGG16	28.872.598	84.200	78,75

Fazendo-se uma análise das abordagens monomodais, o algoritmo CNN2D apresenta o menor número de parâmetros, apresentando 5.800 segundos de tempo de execução. Os algoritmos CharCNN e FastText apresentam um número similar de

parâmetros, em torno de $1.000.000$, e consequentemente um custo de treino também parecido. Como se pode ver, a arquitetura com o maior número de parâmetros é a VGG16. A arquitetura da rede VGG-16 é uma arquitetura extremamente profunda de camadas convolucionais. Para o caso especifico da VGG16, o número de filtros por camada é alto e a rede possui entre duas e três camadas convolucionais consecutivas sem uso de camadas de agrupamento, contribuindo para o aumento do número de parâmetros treináveis. Este número elevado de parâmetros treináveis, impacta diretamente no tempo de trenamento.

O maior custo da VGG16 tem um impacto direto na versão multimodal FastText + VGG16, tornando-o o algoritmo com mais parâmetros, e o mais custoso, em tempo de treino, dentre todas as demais abordagens. Em tempo de teste, o FastText + VGG16 leva $78,75$ segundos (pouco mais de 1 minuto) para classificar 621 páginas de 36 diferentes documentos, levando em média $2,18$ segundos por documento. Embora seja um tempo de teste bem superior aos das demais abordagens, ainda sim é viável, seja em tarefas síncronas ou assíncronas. Com isso, a *Pergunta de Pesquisa 2* pode ser respondida positivamente.

7. Conclusão

Este trabalho propõe um estudo a cerca do uso de redes neurais convolucionais multimodais para o problema de classificação de páginas em processos eletrônicos. Devido à composição heterogênea de documentos em um processo eletrônico, a hipótese deste trabalho é que a fusão de características textuais e visuais pode contribuir para uma melhor classificação de páginas. Diferentes algoritmos monomodais e multimodais foram avaliados em uma base de dados composta por 117 documentos de processos judiciais, totalizando 2970 páginas.

Os resultados mostraram que a melhor abordagem, em termos de acurácia e kappa, é multimodal e combina o algoritmo FastText (especialista em texto) com o VGG16 (especialista em imagem). Além de atingir resultados superiores, esta abordagem apresenta viabilidade de ser usada tanto em aplicações síncronas como assíncronas. Como trabalhos futuros, acreditamos que outras formas de fusão, além da concatenação dos vetores de texto e imagem, como por exemplo a média ou outras medidas de agregação, possam melhorar os resultados de classificação. Além disso, também pretende-se ampliar o tamanho da base de processos judiciais, e incorporar a informação sequencial dos documentos analisados para aprimorar as classificações.

8. Agradecimentos

Os autores gostariam de agradecer a CAPES, CNPq e FACEPE pelo apoio financeiro. Também gostariam de registrar o agradecimento a NVIDIA Corporation, pela doação de uma placa GPU GeForce Titan XP, utilizada nos experimentos realizados.

Referências

Audebert, N., Herold, C., Slimani, K., and Vidal, C. (2019). Multimodal deep networks for text and image-based document classification.

Chen, N. and Blostein, D. (2007). A survey of document image classification: problem statement, classifier architecture and performance evaluation. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(1):1–16.

Chollet, F. (2018). *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*. MITP-Verlags GmbH & Co. KG.

de Araujo, P. H. L., de Campos, T. E., Braz, F. A., and da Silva, N. C. (2020). Victor: a dataset for brazilian legal documents classification. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1449–1458.

Frasconi, P., Soda, G., and Vullo, A. (2002). Hidden Markov models for text categorization in multi-page documents. *Journal of Intelligent Information Systems*, 18(2-3):195–217.

Gallo, I., Noce, L., Zamberletti, A., and Calefati, A. (2016). Deep Neural Networks for Page Stream Segmentation and Classification. *2016 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2016*, pages 1–7.

Hammami, E., Akermi, I., Faiz, R., and Boughanem, M. (2019). Deep learning for french legal data categorization. In *International Conference on Model and Data Engineering*, pages 96–105. Springer.

Jain, R. and Wigington, C. (2019). Multimodal Document Image Classification. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, volume 91, pages 71–77. IEEE.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4):150.

Rusiñol, M., Frinken, V., Karatzas, D., Bagdanov, A. D., and Lladós, J. (2014). Multimodal page classification in administrative document image streams. *International Journal on Document Analysis and Recognition*, 17(4):331–341.

Toffoli, J. A. D. and Gusmão, B. G. (2019). *Inteligência artificial na Justiça / Conselho Nacional de Justiça*.

Undavia, S., Meyers, A., and Ortega, J. E. (2018). A comparative study of classifying legal documents with neural networks. In *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 515–522. IEEE.

Wiedemann, G. and Heyer, G. (2019). Multi-modal page stream segmentation with convolutional neural networks. *Language Resources and Evaluation*.

Zhang, X. and LeCun, Y. (2015). Text Understanding from Scratch. pages 1–9.