Information Extraction from Financial Statements based on Visually Rich Document Models

Elioenai L. G. Alves[1], **Cecília Carvalho**[1], **Patrick Martins de Lima**[1], **Vládia Pinheiro**[1], **Vasco Furtado**[1]

[1]Universidade de Fortaleza (UNIFOR) – Programa de Pós-Graduação em Informática Aplicada
Fortaleza – CE – Brasil.

l.oenaialves@gmail.com, ceciliacarvalhoo@gmail.com,

patrick.ml.dev@gmail.com, vladiacelia@unifor.br, vasco@unifor.br

Abstract. *This paper presents an Information Extraction system for visually rich financial documents. The system takes pre-trained neural models from the LayoutXLM family and refines them for use in Financial Statements. Two post-processing steps were developed in order to adjust the results generated by the refined model. From empirical evaluations, it is concluded that the proposed system is effective in extracting information from financial documents and offers potential to automate and optimize the process of analysis and validation of financial statements.*

Resumo. *Este artigo apresenta um sistema de Extração de Informação para documentos financeiros visualmente ricos. O sistema utiliza modelos neurais pré-treinados da família LayoutXLM e os refina para uso em Demonstrações Financeiras. Duas etapas de pós-processamento foram desenvolvidas com o intuito de ajustar os resultados gerados pelo modelo refinado. A partir de avaliações empíricas comparativas, conclui-se que o sistema proposto é eficaz na extração de informações de documentos financeiros e oferece potencial para automatizar e otimizar o processo de análise e validação de demonstrações financeiras.*

1. Introdução

Apesar do grande avanço dos modelos de aprendizagem profunda baseados em transformers (e.g. BERT) para classificação, sumarização e traduções, algumas outras tarefas exigem mais do que a habilidade de compreender um texto e ser capaz de fazer inferências a partir dele. A compreensão de certos documentos requer mais do que a exploração exclusiva de textos. Muitas vezes se necessita de informação adicional sobre como o texto está diagramado e/ou estruturado. A existência e o formato de apresentação de tabelas e figuras, bem como a estrutura do texto, dividida em seções com títulos e subtítulos, são importantes para extração de informações em geral. Esses documentos, contendo textos, tabelas, figuras, títulos e subtítulos em fontes, *layout* e estilos variados são chamados na literatura de Documentos Visualmente Ricos (VrDs) [Sarkhel and Nandi 2019].

Em VrDs, o mero reconhecimento de um título de uma nota explicativa revela-se desafiador. Esse título pode se apresentar de diversas formas (e.g. “1. Contexto operacional” ou “1 Contexto operacional”). Abordagens que recorrem ao desenvolvimento de expressões regulares (RegEx) foram pensadas [Stubblebine 2003], mas o número de falsos positivos acaba sendo grande porque números, seguidos ou não de pontuação, espaços

em branco e textos, podem aparecer em várias partes de um documento. Percebe-se que é necessário desenvolver modelos que levem em conta o texto e as posições no Layout em que as palavras e/ou sentenças aparecem no documento.

Diante disso, vê-se o surgimento de modelos neurais pré-treinados para entendimento de VrDs [Wang et al. 2022], visando suportar tarefas de classificação textual, pesquisas textuais, classificação de documentos por meio de *layout*, segmentação de informações para manipulação textual utilizando outros modelos com tarefas como *Question-Answer*, NER, etc. Dentre esses modelos podemos citar os modelos da família LayoutLM (Layout LM, Layout LM2, Layout LM3 e LayoutXLM)[Xu et al. 2020a] e o modelo LILT [Wang et al. 2022] que combinam informações de *layout* e textuais. Em particular, o modelo LayoutXLM, por ser multilíngue, pode ser utilizado em documentos de língua portuguesa compreendendo melhor a relação entre *layout* e texto. As principais vantagens ao utilizar esse modelo, além de classificar textos com diferentes rótulos (cabeçalho, texto, tabela, rodapé, número da página, título) é que ele automatiza e acelera o processo de classificação e extração de informações.

Nessa pesquisa descrevemos a abordagem seguida para desenvolver um sistema de Extração de Informação para um tipo específico de VrD: as demonstrações financeiras de empresas. Demonstrações Financeiras (DFs) são documentos extensos contendo em média 90 páginas e compostos de (1) Demonstrações Contábeis - Balanço Patrimonial, Demonstração do Resultado do Exercício, Demonstração das Mutações do Patrimônio Líquido, Demonstração do Resultado Abrangente, Demonstração do Fluxo de Caixa, e Demonstração do Valor Adicionado; e (2) Notas Explicativas, que visam facilitar o entendimento das demonstrações contábeis, apresentando todas as informações que serviram como respaldo para elaboração das mesmas. Quanto à forma, as Demonstrações Contábeis são apresentadas em tabelas com linhas e colunas, de diversos formatos e estilos, e as Notas Explicativas são dispostas em seções (uma para cada nota explicativa) contendo título, textos e tabelas associadas, também com variações de formato e estilo.

No Brasil, empresas constituídas sob a forma de sociedade por ações, bem como as empresas de grande porte, são obrigadas a submeter as DFs a empresas de auditoria independente (tais como KPMG, Price WaterHouse, Ernst & Young e Deloitte). Essas, por sua vez, precisam realizar análises, verificações, correlações e validações nessas DFs, como verificar se existem as notas explicativas para alguns itens das demonstrações contábeis e se os valores citados em ambas (notas explicativas e demonstrações contábeis) estão consistentes. Desenvolver uma ferramenta computacional que facilite a estruturação, extração e análise automáticas destas informações foi foco prático de nossa pesquisa.

Neste trabalho, propomos um sistema de Extração de Informação de Demonstrações Financeiras baseado em um modelo pré-treinado para entendimento de VrD. O sistema proposto é composto de quatro módulos: (1) DF-OCR, que converte os textos presentes de uma imagem em um texto editável, possibilitando a manipulação textual; (2) LayoutDF, um modelo refinado a partir do LayoutXLM para entendimento de DFs que classifica cada token de uma DF nas *labels* - *text*, *title* e *table*; (3) DF-Filter, que realiza o pós-processamento dos tokens classificados pelo LayoutDF; e (4) o DF-Hierarchizer, que organiza o conteúdo textual conforme a hierarquia das partes presentes no VrD. As principais contribuições deste trabalho são o LayoutDF - um modelo refinado para entender VrD do tipo DFs, a CDFin - uma Coleção Dourada de DFs, e o desenvol-

vimento de módulos de pós-processamento e estruturação, necessários para a melhoria e otimização do processo de extração de DFs. Foram realizados experimentos em um conjunto de DFs de teste, visando comparar diferentes abordagens para a tarefa de reconhecimento de títulos, inclusive com modelos generativos baseados em LLM (*Large Language Model*) GPT. O sistema de Extração de Informação de DFs, aqui proposto, apresentou um desempenho promissor em todos os testes realizados.

O artigo está estruturado como descrito a seguir. A seção 2 apresenta a fundamentação teórica e os trabalhos relacionados. Na seção 3 detalham-se a arquitetura do sistema Extrator de DFs e a metodologia de refinamento do modelo LayoutXLM para a tarefa de classificação de tokens a partir de uma coleção dourada. A seção 4 apresenta os cenários dos experimentos e os resultados. Por fim, discorre-se sobre as conclusões e os trabalhos futuros.

2. Fundamentação Teórica

2.1. Modelos Multimodais para Leitura de Documentos Visualmente Ricos

Os modelos da família Layout LM (LayoutLM, LayoutLM2, LayoutLM3 e LayoutXLM) foram desenvolvidos seguindo uma arquitetura Transformer multimodal e pré-treinados para extrair informações de Documento Visualmente Ricos - VrDs [Xu et al. 2020a]. Tais modelos são multimodais, pois mesclam informações textuais (utiliza o WordPiece [Wu et al. 2016] para tokenizar as palavras), visuais (utiliza a arquitetura ResNeXt-FPN [Xie et al. 2017] para codificar a imagem) e de *layout* (utiliza as informações das *bounding box* de cada token para gerar uma sequência de informações posicionais 2D (xmin, xmax, ymin, ymax, largura, altura)). Ou seja, o codificador multimodal concatena os embeddings visuais e textuais em uma sequencia única e soma com as infomações de *layout*, formando uma camada de entrada. Em seguida outras camadas de Self-Attention são adicionadas. A premissa destes modelos é que a tarefa de Entendimento de VrDs não depende somente da informação textual, mas também de informações visuais e de *layout* [Xu et al. 2020b], pois incluem informações que normalmente são perdidas nos modelos comuns, tais como: formatação de textos (estilo de fonte - negrito e itálico, tamanho de fonte, etc.), posicionamento de textos específicos, dentre outras. Esses modelos foram pré-treinados utilizando um número grande de documentos não rotulados de diferentes domínios, em tarefas que visam aprender a integração entre texto, imagem e seu *layout*. São elas: modelagem de linguagem visual mascarada (em que alguns textos foram aleatoriamente mascarados e pediu-se para o modelo recuperar o texto original, fazendo com que o modelo aprendesse a linguagem); alinhamento Texto-Imagem (para ajudar o modelo a entender a localização espacial imagem e as *bounding box* dos textos); e, por fim, a tarefa de correspondência de texto-imagem (que é aplicada para o modelo aprender a correspondência entre a imagem do documento e o conteúdo textual) [Xu et al. 2020b]

Outro modelo que também faz uso de informações textuais e de *layout* é o LILT [Wang et al. 2022] o qual pode ser treinado em um idioma e facilmente ajustado para outros, sendo essa a sua grande vantagem. Segundo [Wang et al. 2022] os modelos existentes, por exemplo, o Layout LM2, tem a desvantagem de lidar apenas com documentos na língua em que o modelo foi pre-treinado. Para este trabalho vamos utilizar o modelo LayoutXLM [Xu et al. 2021], a versão multilíngue do LayoutLM2, que foi pré-treinado com 53 idiomas diferentes, dentre eles o português.

2.2. Trabalhos Relacionados

[Keocheguerian and Martins 2021] apontou que as principais empresas de auditoria e consultoria, conhecidas como *Big Four* (KPMG, Price WaterHouse, Ernst & Young e Deloitte), têm realizado significativos investimentos no uso da IA em diversas atividades, inclusive para a auditoria de documentos, cujas tarefas são estruturadas e repetitivas. Algoritmos de *Machine Learning* (ML) têm sido empregados para a detecção de fraudes, outliers e agrupamentos de empresas [Hooda et al. 2018].

Em [Cho et al. 2023] tem-se a proposta de um framework IDP que combina o uso do um modelo de aprendizado profundo pré-treinado LayoutXLM com técnicas de RPA (automação de processos robóticos), para extrair informações do tipo chave-valor em documentos financeiros coreanos e capturar dados em sistemas externos. O uso do LayoutXLM visou flexibilizar a aplicação do *Framework* em vários tipos de documentos. Os resultados demonstraram que o LayoutXLM obteve o melhor desempenho na tarefa de extração de informações do tipo chave-valor.

[Déjean et al. 2022] investiga a tarefa de extração de relações em documentos comparando dois modelos diferentes de rede neural: o modelo de linguagem multimodal (LayoutXLM) e uma *Graph Neural Network*: *Edge Convolution Network* (ECN). Esta tarefa visa vincular dois elementos textuais relacionados, como uma pergunta e uma resposta em um formulário. O grafo é construído usando os elementos da página como nós e as arestas são criadas usando a estratégia de linha de visão (dois nós são vizinhos se eles se virem). Nos experimentos realizados neste trabalho, os resultados mostram que ambos os modelos funcionam de maneira semelhante, sendo que o LayoutXLM é melhor quando o conjunto de dados aumenta.

Em [Ylisiurunen et al. 2022], tem-se três modelos para a extração de informação em recibos, incluindo nome comercial, endereço, data de compra, valor e imposto. Os modelos LayoutLM2 e LayoutXLM, baseados em Transformer, e o modelo PICK [Yu et al. 2021], que utiliza uma rede convolucional de grafos como base. Nos experimentos realizados em um conjunto de recibos finlandeses coletados manualmente, a pontuação F1-média obtida no SROIE foi de 92,92%, inferior ao desempenho relatado pelos autores do LayoutLM2 (96,25%). Um experimento adicional foi realizado com a adição de rótulos de Imposto sobre o Valor Agregado (IVA) nos recibos finlandeses. No entanto, nesse caso, F1-Score diminuiu (= 45,0%). Essa queda no desempenho pode ser atribuída às variações textuais significativas na informação do IVA e aos poucos dados rotulados para treinamento.

É importante destacar que, até a elaboração deste artigo, não foram encontradas publicações descrevendo o uso de modelos para entendimento de VrD, escritos em português.

3. Extração de Informações de Demonstrações Financeiras

Demonstrações Financeiras podem ser consideradas um tipo de Documento Visualmente Rico (VrD), pois os estilos, formatos e posições dos textos, títulos e tabelas em um *layout* são vitais para o entendimento das informações nelas contidos.

Para exemplificar, as Figuras 1 e 2 apresentam partes de uma mesma Demonstração Financeira (DF) da Empresa Taurus Armas S.A., que por ser de capital

aberto é obrigada a publicizar suas demonstrações contábeis e notas explicativas. A Figura 1 ilustra o Balanço Patrimonial de 2021 da empresa no formato de uma tabela com linhas e colunas. Nesta tabela, tem-se, por exemplo, que duas colunas possuem o mesmo título "Consolidado" e subtítulos com as datas 31-12-2021 e 31-12-2020; e linhas referentes a itens do ativo e passivo da empresa como "Clientes" e "Estoque", e outras linhas contendo apenas subtotais de valores. É comum nestes documentos existir uma coluna intitulada "Nota" contendo referência numérica a uma nota explicativa específica, que deve constar como uma seção dentro das Notas Explicativas. No exemplo deste Balanço Patrimonial, a linha "Clientes", em sua coluna "Nota", faz referência à nota explicativa de número 9, transcrita na Figura 2. Por sua vez, esta nota explicativa de No. 9 possui o título "9. Clientes" e partes-filha contendo textos e uma tabela que detalha os valores do item "Clientes" do Balanço Patrimonial.

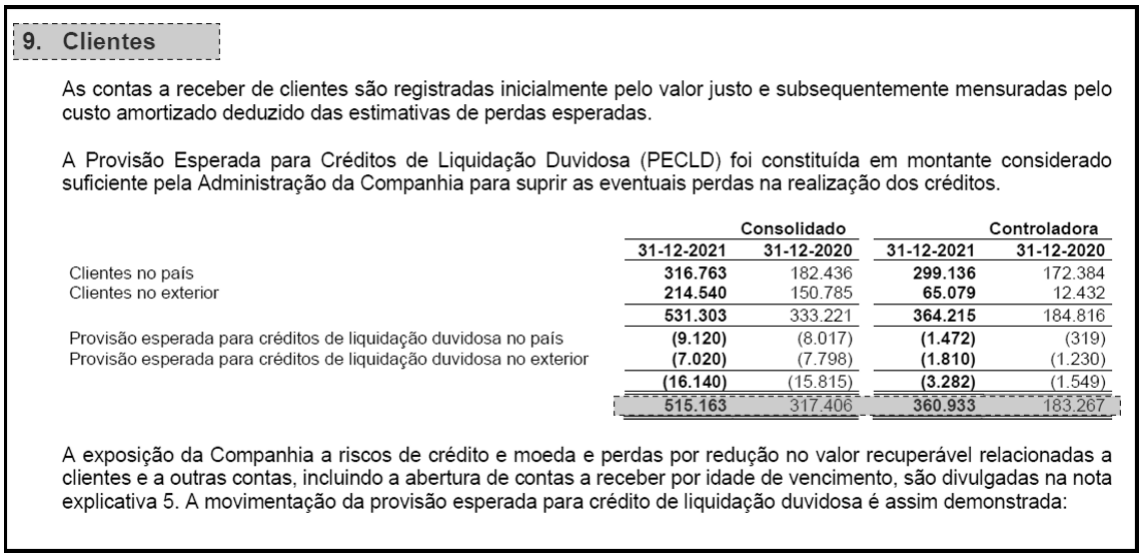
Taurus Armas S.A.
Balanço patrimonial em 31 de dezembro de 2021
Valores expressos em milhares de Reais – R$

		Consolidado		Controladora	
	Nota	31-12-2021	31-12-2020	31-12-2021	31-12-2020
Ativo					
Circulante					
Caixa e equivalentes de caixa	7	**185.764**	91.231	**65.399**	34.623
Aplicações financeiras e contas vinculadas	8	**70.778**	-	**70.778**	-
Clientes	9	**515.163**	317.406	**360.933**	183.267
Estoques	10	**491.864**	298.343	**274.370**	204.894
Impostos a recuperar	11	**65.261**	33.319	**53.471**	28.987
Pagamentos antecipados		**30.985**	22.222	**7.265**	4.793
Outras contas a receber	12	**29.779**	34.488	**29.850**	28.848
Ativos mantidos para venda	14	**66.396**	133.850	**-**	-
		1.455.990	930.859	**862.066**	485.412
Não circulante					
Impostos a recuperar	11	**5.627**	-	**4.886**	-
Imposto de renda e contribuição social diferidos	13	**121.380**	188.580	**101.951**	166.291
Crédito com empresas ligadas	24	**-**	-	**40.681**	29.661
Outras contas a receber	12	**24.809**	14.541	**13.160**	13.132
		151.816	203.121	**160.678**	209.084
Investimento em controladas	15	**4.420**	2	**683.822**	462.148
Imobilizado	16	**379.023**	233.355	**204.027**	130.012
Intangível	17	**102.371**	93.313	**26.213**	18.666
		485.814	326.670	**914.062**	610.826
Total do Ativo		**2.093.620**	1.460.650	**1.936.806**	1.305.322

Figura 1. Demonstração Contábil "Balanço Patrimonial de 2021" da Taurus Armas S.A.

Considerando a necessidade de empresas de auditoria realizarem a verificação e validação das informações constantes em uma DF (verificação de notas explicativas e checagem de valores e somas), é notória a necessidade de ferramentas computacionais que facilitem a estruturação e a extração destas informações de forma automática. Em especial, o reconhecimento de títulos de seções de uma DF é crítico, dada as variações de formatação, a seguir exemplificadas:

- **1. Contexto operacional -** título contendo número, seguido de ponto ("."), caractere(s) em branco e texto;
- **1 Contexto operacional -** título contendo numero, seguido de caractere(s) em branco e texto;
- **1.Contexto operacional -** título contendo número, seguido de ponto (".") e texto (sem caractere(s) em branco);



9. Clientes

As contas a receber de clientes são registradas inicialmente pelo valor justo e subsequentemente mensuradas pelo custo amortizado deduzido das estimativas de perdas esperadas.

A Provisão Esperada para Créditos de Liquidação Duvidosa (PECLD) foi constituída em montante considerado suficiente pela Administração da Companhia para suprir as eventuais perdas na realização dos créditos.

	Consolidado		Controladora	
	31-12-2021	31-12-2020	31-12-2021	31-12-2020
Clientes no país	**316.763**	182.436	**299.136**	172.384
Clientes no exterior	**214.540**	150.785	**65.079**	12.432
	531.303	333.221	**364.215**	184.816
Provisão esperada para créditos de liquidação duvidosa no país	**(9.120)**	(8.017)	**(1.472)**	(319)
Provisão esperada para créditos de liquidação duvidosa no exterior	**(7.020)**	(7.798)	**(1.810)**	(1.230)
	(16.140)	(15.815)	**(3.282)**	(1.549)
	515.163	317.406	**360.933**	183.267

A exposição da Companhia a riscos de crédito e moeda e perdas por redução no valor recuperável relacionadas a clientes e a outras contas, incluindo a abertura de contas a receber por idade de vencimento, são divulgadas na nota explicativa 5. A movimentação da provisão esperada para crédito de liquidação duvidosa é assim demonstrada:

Figura 2. Nota Explicativa 9, referenciada na coluna "Nota" na Figura 1 (na linha "Clientes").

- **1.1 Oferta pública de ações -** subtítulo com numeração (dois níveis), seguido de caractere(s) em branco e texto;
- **2.4.12 Intangível -** subtítulo com numeração (três níveis), seguido de caractere(s) em branco e texto.

Neste sentido, este trabalho propõe um sistema que realiza a estruturação de Documentos Visualmente Ricos (VrD), especificamente, de Demonstrações Financeiras (DFs) visando a extração e correlação de informações inter e intra-documentos. O cerne do sistema proposto é o módulo refinado para documentos DFs a partir do modelo LayoutXLM, no entanto, módulos adicionais de pós-processamento se mostraram relevantes para melhoria do desempenho. A Figura 3 apresenta o sistema composto de quatro módulos:

- **DF-OCR (*Optical Character Recognition*)** - que realiza o reconhecimento de caractere por processamento de imagem e recupera as informações textuais (*words*) e posicionais (*boundingBox*) dos elementos de um documento, o qual está disponível, geralmente, digitalizado e em formato PDF. No caso de tabelas, este módulo provê informações complementares para a reconstrução das tabelas com suas linhas e colunas. Podem ser usadas soluções livres como Tesseract Open Source OCR Engine, ou proprietárias (Ex. OCR Azure e Cloud Vision), a depender da complexidade do VrD. Neste trabalho, utilizou-se o OCR Azure, por apresentar melhores resultados na extração de tabelas;
- **LayoutDF -** este módulo identifica e classifica as partes de uma DF, conforme a necessidade. Por exemplo, pode ser treinado para classificar o que é texto, tabela, título, cabeçalho e nota de rodapé, presentes em cada página do documento. Nesse módulo, utilizou-se modelos de aprendizado multimodais para leitura de VrD, como o modelo LayoutXLM (explanado na seção 2.1), o qual foi refinado (*fine-tuning*) para classificação e separação das partes de DFs;
- **DF-Filter -** este módulo é responsável por realizar o pós-processamento dos tokens classificados, visando filtrar, normalizar e corrigir imprecisões na saída do LayoutDF. Por exemplo, separação de títulos concatenados, que em alguns casos,

por estarem um seguido do outro no documento o LayoutDF entende se tratar de único título. Outro exemplo, são títulos repetidos em páginas da DF;

- **DF-Hierarchizer-** este módulo é responsável por gerar a estrutura hierárquica da DF a partir da saída do módulo anterior. Cada DF é organizada em uma hierarquia de partes e subpartes (*childs*), contendo títulos, textos e tabelas. Formalmente, a saída desse módulo é um documento no formato JSON contendo as partes da DF caracterizadas pelos atributos: página (*page*); identificador (*title-id*); texto (*text*); coordenadas posicionais na página, geradas pelo módulo de OCR (*boundingBox*); classe do token, que pode ser um título, texto ou tabela (*token-type*); e suas partes-filha (*child*[...]). Esta estruturação servirá a diversas aplicações que visem realizar a conciliação de valores e demais validações.

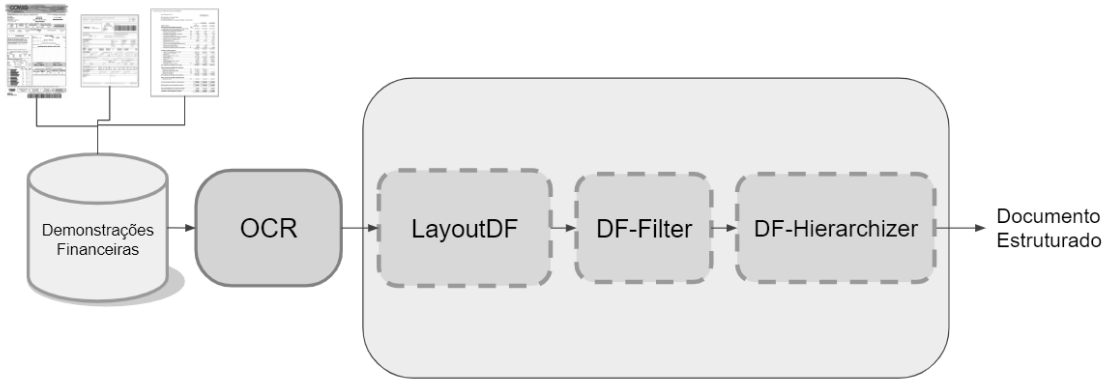


Figura 3. Arquitetura do Sistema de Extração de Informação de Demonstrações Financeiras (DFs)

3.1. Modelagem e Treinamento do LayoutDF

No treinamento do módulo LayoutDF foi aplicado um processo de *Fine-Tuning* do modelo LayoutXLM, utilizando-se a abordagem de *token-classification*. Nesta versão, o LayoutDF foi treinado para classificar cada token presente em uma DF nas seguintes classes: **título** (*title*), **texto** (*text*) ou **tabela** (*table*). Esta classificação é básica para a estruturação da DF, principalmente a classificação dos títulos, essenciais para delimitação das notas explicativas.

A Figura 4 apresenta o diagrama do processo de treinamento do LayoutDF. Um conjunto de treinamento de DFs anotadas (Coleção Dourada) é enviado como entrada do modelo LayoutXLM, com as seguintes informações: texto e *BoundingBox*, recuperadas de um processo de OCR; Imagem e Layout, processadas pelo módulo do Layout LM2 (LayoutLMv2Processor); e a Matriz de *labels*, a ser explanada na seção 3.1.1. O modelo aprende, a partir das DFs, as variações de símbolos, de espaçamentos entre o número e o texto do título, de tamanhos das fontes, de mudanças na espessura das letras, e de posicionamento do texto na página e espaçamento entre os textos. Por fim, o módulo LayoutDF *Fine-Tuned* (refinado) é gerado e está ajustado às características de DFs.

LayoutDF foi treinado por 14 épocas, usando o PyTorch nativo do ambiente Google Colab, com uma GPU de 15 Gb e parou o treinamento após atingir o valor de *Patience* estabelecido no *EarlyStopping*. Utilizou-se também o otimizador AdamW com taxa de

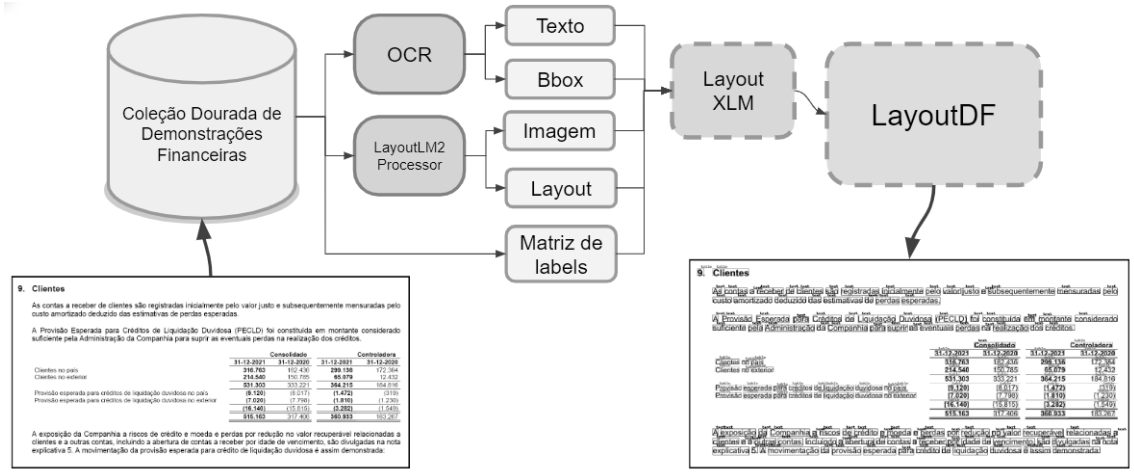


Figura 4. Diagrama de Treinamento do LayoutDF.

aprendizado = 5e-5 (valor padrão para modelos baseados em Transformer). A Figura 5 ilustra um exemplo de classificação dos tokens realizada pelo LayoutDF na Nota Explicativa No.9.

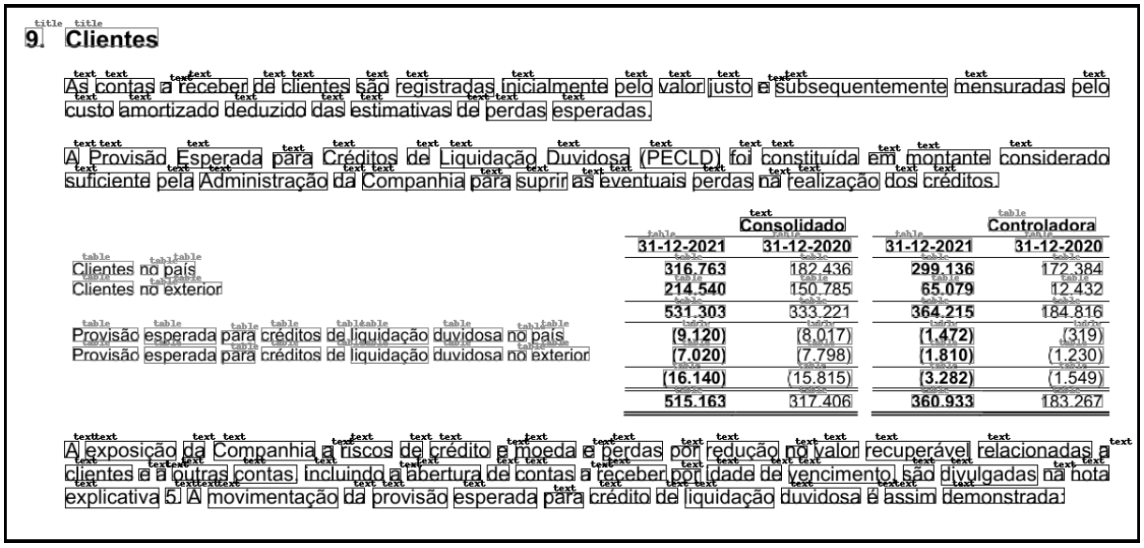


Figura 5. Classificação dos tokens da Nota Explicativa No.9, realizada pelo módulo LayoutDF.

3.1.1. Coleção Dourada de DFs para Treinamento do LayoutDF

Para o processo de treinamento e validação do LayoutDF, foi anotado um conjunto de Demonstrações Financeiras de várias empresas, buscando gerar uma Coleção Dourada (CDFin) representativa da diversidade de formatos e estilos de títulos, textos e tabelas. Dois especialistas em auditoria financeira e contábil selecionaram 22 empresas de capital aberto, e, a partir deste *pool* de empresas, a CDFin foi construída com 23 DFs (totalizando 189.963 tokens) - 14 DFs para treino, 3 DFs para validação e 6 DFs para teste (dados não-vistos), a seguir identificadas:

- **Treino -** Delloite4, 2019Vale, 2020Mizuho, 2020Enel, 2020Pagseguro, 2020Sicredi, 2021Aacd, 2021Telefonica, BahiaInveste, BancoVolks, ConglomeradoPrudential, DFP, 2019XPInvestimento, Neon.
- **Validação -** 2021Fleury, 2021Correios e BancoCargil
- **Teste -** Delloite1, DTVM, BradescoSaude, BradescoSeguros, Cury e BSA.

Para a criação da CDFin foram testados dois modelos de OCRs - Tesseract Open Source OCR Engine (v 5.3.1) e o OCR da Microsoft Azure (v3.0). O OCR Microsoft Azure foi o escolhido por apresentar uma qualidade superior no reconhecimento de tabelas complexas.

Para auxiliar o processo de anotação de cada token com os *labels* - *text*, *title* e *table* - foi desenvolvida uma ferramenta de anotação que percorre cada token da DF e solicita ao anotador que indique o *label* a ser associado àquele *token*, e, ao final, é gerada a Matriz de Labels correspondente. Na Figura 6, apresenta-se um exemplo da Matriz de *label* gerada para a Nota Explicativa No. 9, onde os dois primeros tokens, receberam *label* 1 (*tokens* de título), e os demais receberam *label* 0 e 2 indicando *tokens* de texto e tabela, respectivamente.

```
[1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0]
```

Figura 6. Exemplo da Matriz de Labels gerada para a Nota explicativa No.9 (Figura 2)

A Matriz de *Labels*, juntamente com informações sobre a imagem e o *layout* das páginas, foram associadas aos textos e boundingbox de cada página da DF na CDFin. A imagem e *layout* das páginas foram recuperadas utilizando o módulo LayoutLMv2Processor do modelo LayoutLM2. Adicionalmente, à cada DF foram anotadas as seguintes informações: **dataset-name -** nome do documento; **data -** onde os dados são armazenados; **page -** número da página do PDF de onde as informações estão sendo extraídas; **path-image -** indicador do nome do arquivo de imagem dessa página; **ner-tags -** a Matriz de *Labels* para os *tokens* da página; **words -** palavras presentes na página; **boxes -** as *bounding boxes* de cada palavra na página.

3.2. DF-Filter e DF-Hierarchizer

Os módulos DF-Filter e DF-Hierarchizer são responsáveis pelas tarefas finais, mas não menos importantes, do Extrator de DFs. O primeiro realiza o pós-processamento, aplicando regras para normalização e correção da classificação dos *tokens* da DF, e o segundo é responsável pela hierarquização das seções presentes nas DFs com seus títulos, textos e tabelas.

3.2.1. DF-Filter - Regras de Pós-processamento de tokens

Durante o processo de construção do sistema, alguns experimentos preliminares com o LayoutDF foram realizados e algumas imprecisões foram identificadas. Por exemplo, concatenação de títulos e repetição de títulos. Como estes erros foram recorrentes, as seguintes regras foram modeladas e implementadas neste módulo visando a filtragem dos títulos falsos-positivos:

- **Regra F1 - Validação dos *tokens* classificados como títulos -** por padrão, os títulos das notas explicativas em DFs devem vir numeradas, pois com essa numeração há o cruzamento de informações. O padrão de RegEx utilizado para filtrar os títulos corretos, da lista de *tokens* identificados como títulos foi `\d+(?:\.\d+)*(?:\.\s)?[\w\s]+`. Ele reconhece um ou mais números, seguido de ponto ou não, com algumas strings no final. Por exemplo: "9. Clientes" ou "9 Clientes". Essa regra é responsável por eliminar os subtítulos identificados erroneamente pelo LayoutDF;
- **Regra F2 - Remoção de títulos repetidos -** algumas DFs apresentam a repetição do título da nota explicativa em cada página que contém a continuação da nota. Por exemplo, se a Nota Explicativa No.9 se estende da página 24 até a página 30, no início de cada página, a partir da página 25, a DF apresentava o título "9. Clientes - Continuação". Nestes casos, o LayoutDF identificava 7 (sete) títulos distintos, dos quais 6 (seis) era falsos-positivos, pois, de fato, não são títulos de notas explicativas distintas. Esta regra realiza a exclusão destes títulos repetidos por um padrão RegEx que utiliza o token "Continuação" e suas variações;
- **Regra F3 - Separação de títulos concatenados -** pode acontecer de um títulos vir imediatamente seguido por um subtítulo. Por exemplo, "9. Clientes" e na linha seguinte "9.1 Clientes Ativos". Esses casos ocorrem quando não há texto ou tabela entre um título e seu subtítulo. Nesses casos, o LayoutDF identifica os dois títulos como sendo um só. Por exemplo: "9. Clientes 9.1 Clientes Ativos". Assim, para separá-los utilizou-se o padrão de RegEx que procura um ou mais números seguido de ponto e espaço. O padrão de RegEx foi `\s(?=\d\.\d\.?\s)`.

3.2.2. DF-Hierarchizer - Hierarquizador de seções

Nesta etapa, cada página da DF é gerada para um arquivo (em JSON) seguindo a ordem hierárquica encontrada em cada seção. Na Figura 7, apresenta-se o Documento Estruturado de uma parte da DF da empresa Taurus, no caso, a Nota Explicativa No.9 (transcrita na Figura 2). No exemplo, a parte mais alta na hierarquia está na página 41, com text = "9. Clientes", *token_type* = "*title*", e possui como partes-filha (childs), uma parte do tipo texto (text = "As contas a receber de clientes são registradas ...") e uma tabela, referenciada num arquivo csv. No caso, uma aplicação a partir da estruturação da referida DF, pode validar os valores 515.163, 317.406, 360.933, 183.267, que aparecem na Linha "Clientes"do quadro Balanço Patrimonial, a qual referencia a Nota Explicativa No.9 (Figura 1, ver coluna "Nota" = 9), com os valores que aparecem na tabela da nota explicativa.

4. Avaliação Experimental

Para o entendimento de VrDs, especificamente de Demonstrações Financeiras (DFs), e a extração das informações nelas contidos, é essencial realizar a identificação correta de cada seção do documento. Geralmente, as seções em documentos são demarcadas por títulos, como mencionado anteriormente. Caso uma seção não seja identificada, a hierarquia é quebrada e os textos e tabelas pertencentes a uma seção serão associados a outra seção de maneira incorreta, prejudicando o entendimento da DF. Portanto, esta avaliação experimental concentrou-se em analisar várias abordagens para a tarefa de reconhecimento de títulos de seções de DFs. Neste sentido, foram modelados e implementados os seguintes cenários para avaliação:

Figura 7. Arquivo JSON que representa a estruturação de parte da DF da empresa Taurus

- **CENÁRIO 1 - Baseline -** foi desenvolvido um algoritmo baseline para reconhecimento de títulos que utiliza padrões de expressões regulares. Alguns padrões RegEx implementados nesta solução foram: (1) sequências de dígito seguido de ponto; (2) sequência de dígito numérico (de 0 a 9), um ponto (.) ou um espaço em branco, seguidos de letras maiúscula ou minúscula (de A a Z) (ou acentuadas), a qual esteja em mesma linha (verificado via *BoundingBox*).
- **CENÁRIO 2 - LayoutDF -** abordagem que utiliza apenas a saída do módulo LayoutDF, que é primordialmente baseado no treinamento do modelo multimodal LayoutXLM;
- **CENÁRIO 3 - Extrator DFs -** neste cenário utilizou-se toda a arquitetura proposta neste trabalho - LayoutDF + DF-Filter + DF-Hierarchizer, visando avaliar a relevância do módulo DF-Filter no processamento e aplicação de regras para normalização e correção dos tokens;
- **CENÁRIO 4 - LLM GPT -** neste cenário foi implementado um *prompt* de consulta ao LLM GPT 3.5 para a tarefa de reconhecimento de título, usando uma abordagem *few-shot learning*.

As Tabelas 1, 2 e 3 apresentam os resultados comparativos de cada cenário, dois a dois. Na Tabela 1, tem-se os resultados dos Cenários 1 e 2 para cada DF do conjunto de teste. Embora o Sistema Baseline (CENÁRIO 1) tenha apresentado melhor precisão, o LayoutDF apresentou melhor cobertura na identificação dos reais títulos, resultando em melhor estruturação das DFs. Considerando-se a média geral de *F1-Score*, tem-se importante melhoria do módulo LayoutDF em relação ao Sistema *Baseline*.

Dois especialistas de empresas de auditoria analisaram manualmente os títulos reconhecidos no CENÁRIO 2 para a DF BRADESCO SEGUROS (pior caso em termos de precisão no CENÁRIO 2 - LayoutDF). Eles reportaram que a maioria dos títulos reconhecidos incorretamente (falsos-positivos), são, na verdade, subtítulos que poderiam ser tratados pelas regras do módulo DF-Filter. Na Figura 8 estão listados os títulos reconhecidos no CENÁRIO 2 para a DF BRADESCO SEGUROS. Por exemplo, "e.Segregação entre circulante e não circulante" e "g.Aplicações e instrumentos financeiros" são, na verdade, subtítulos.

Tabela 1. Resultados da tarefa de Reconhecimento de Títulos de Seções de VrDs nos CENÁRIO 1 (Baseline) e CENÁRIO 2 (LayoutDF)

	CENÁRIO 1			CENÁRIO 2		
Demonstração Financeira	**Precisão**	**Cobertura**	**F1 Score**	**Precisão**	**Cobertura**	**F1 Score**
(1)Delloite	50,00%	90%	64,29%	80,95%	85%	82,92%
DTVM	95,45%	100%	97,67%	69,56%	95,23%	72,72%
BRADESCO SAÚDE	75,00%	12,00%	20,69%	47,05%	96%	63,15%
BRADESCO SEGUROS	83,33%	18,51%	30,30%	38,46%	92,59%	54,34%
CURY	100,00%	94,44%	97,14%	90,56%	92,59%	89,71%
BSA	100,00%	42,85%	60,00%	64,10%	89,28%	74,62%
MÉDIA GERAL	89,39%	66,43%	62,85%	66,83%	92,59%	73,67%

Fonte: Construção do Autor.

RETORNO DO MODELO - DF BRADESCO SEGUROS
2 Resumo das principais políticas contábeis
1 Contexto operacional
Base de consolidação
e. Segregação entre circulante e não circulante
f. Classificação dos contratos de seguros
g. Aplicações e instrumentos financeiros
j. Intangível
h. Redução ao valor recuperável (impairment) de recebíveis
i. Imobilizado
k. Bens à venda - Salvados
l. Ativos de resseguro e retrocessão
m. Custos de aquisição diferidos
n. Redução ao valor recuperável (impairment) de ativos não financeiros
p. Ativos e passivos contingentes e obrigações legais - fiscais e previdenciárias
r. Imposto de renda e contribuição social
q. Benefícios a empregados
t. Novas normas e interpretações ainda não adotadas
CPC 48 – Instrumentos Financeiros
3 Gerenciamento de riscos

Figura 8. Títulos reconhecidos pelo LayoutDF na DF da BRADESCO SEGUROS

A seguir, na Tabela 2, são apresentados os resultados obtidos pelo módulo LayoutDF (CENÁRIO 2) e pelo sistema Extrator de DFs, proposto neste trabalho (CENÁRIO 3). Observa-se que houve um aumento expressivo na precisão do CENÁRIO 3, destacando-se a melhora da DF BRADESCO SEGUROS, que apresentava uma precisão de 38,46% no CENÁRIO 2, e que alcançou precisão de 96,15% após o pós-processamento executado no DF-Filter, indicando a eficácia das regras implementadas neste módulo.

Com relação ao CENÁRIO 4, o objetivo foi avaliar modelos de IA generativos atuais, como o GPT3.5-Turbo (*Generative Pre-Training Transformer* 3 (GPT-3))

Tabela 2. Resultados da tarefa de Reconhecimento de Títulos de Seções de VrDs nos CENÁRIO 2 (LayoutDF) e CENÁRIO 3 (Extrator de DFs)

	CENÁRIO 2			CENÁRIO 3		
Demonstração Financeira	**Precisão**	**Cobertura**	**F1 Score**	**Precisão**	**Cobertura**	**F1 Score**
(1)Delloite	80,95%	85%	82,92%	89,47%	85%	87,17%
DTVM	69,56%	95,23%	72,72%	95,23%	95,23%	95,23%
BRADESCO SAÚDE	47,05%	96%	63,15%	96%	96%	96%
BRADESCO SEGUROS	38,46%	92,59%	54,34%	96,15%	92,59%	94,33%
CURY	90,56%	92,59%	89,71%	98,03%	92,59%	95,23%
BSA	64,10%	89,28%	74,62%	92,59%	89,28%	90,90%
MÉDIA GERAL	66,83%	92,59%	73,67%	95,62%	92,59%	94,78%

Fonte: Construção do Autor.

[Brown et al. 2020] na tarefa de reconhecimento de títulos, usando uma abordagem few-shot learning [DAIR.AI 2023]. Neste cenário, foi utilizada a API (OpenAIApi) com uma variedade de *prompts*. Foram testados mais de 10 prompts diferentes, como, por exemplo:

- **Prompt 1 -** 'Você é um auditor e precisa transcrever os títulos das notas explicativas de uma demonstração financeira. O texto que você está recebendo é parte de uma demonstração financeira. Os títulos das notas explicativas sempre iniciam com um número. Não modifique o número, apenas transcreva. Texto: `<conteúdo da DF>`'
- **Prompt 2 -** 'Você é um auditor e precisa transcrever os títulos das notas explicativas de uma demonstração financeira. O texto que você está recebendo é parte de uma demonstração financeira. Transcreva os títulos e o seu identificador numérico. Texto: `<conteúdo da DF>`'
- **Prompt 3 -** 'Você é um auditor sucinto que precisa extrair as seções das Notas explicativas presentes nesse trecho de Demonstração Financeira. Faça como no exemplo a seguir identificando as seções das notas explicativas: LECCA DISTRIBUIDORA DE TÍTULOS E VALORES MOBILIÁRIOS LTDA. CNPJ: 07.138.049/0001-54 Notas explicativas às demonstrações contábeis (...) aos mesmos períodos do exercício social anterior para as quais foram apresentadas. A resposta esperada é a seguinte: 1. CONTEXTO OPERACIONAL; 2. APRESENTAÇÃO DAS DEMONSTRAÇÕES CONTÁBEIS; 2.1. Base de apresentação; Faça a mesma coisa para o texto a seguir: `<conteúdo da DF>`'

A estrutura do *prompt* que gerou os melhores resultados neste CENÁRIO 4 foi a do **Prompt 3**, a qual possui os seguintes elementos: "Instrução" a respeito da tarefa; o "Contexto" em que a tarefa está inserida: os "Dados de Entrada" e um "Indicador de Saída", ou seja, alguns exemplos de resposta esperada. Este *prompt* segue a abordagem *few-shot learning*, onde exemplos são dados para o LLM de forma a orientar o modelo em um contexto específico. Este tipo de abordagem tem apresentado melhores resultados em relação à abordagem *zero-shot learning* [DAIR.AI 2023].

Na Tabela 3, são apresentados os resultados do CENÁRIO 4 em comparação com o CENÁRIO 3 (Extrator de DFs). Pode-se observar que o Modelo GPT3.5 apresentou baixa precisão, retornando elevado número de falsos-títulos, resultando em um F1-Score médio de apenas 35,12%. Esses resultados ocorreram, principalmente, devido as variações de formato dos títulos presentes nas Demonstrações Finaceiras. Além disso,

por diversas vezes, o modelo GPT3.5 criou títulos que não existiam, prejudicando os resultados.

Tabela 3. Resultados da tarefa de Reconhecimento de Títulos de Seções de VrDs nos CENÁRIO 3 (Extrator de DFs) e CENÁRIO 4 (LLM GPT 3.5)

	CENÁRIO 3			CENÁRIO 4		
Demonstração Financeira	**Precisão**	**Cobertura**	**F1 Score**	**Precisão**	**Cobertura**	**F1 Score**
(1)Delloite	89,47%	85%	87,17%	100%	100%	100,00%
DTVM	95,23%	95,23%	95,23%	100%	100%	100,00%
BRADESCO SAÚDE	96%	96%	96%	13,70%	68%	22,81%
BRADESCO SEGUROS	96,15%	92,59%	94,33%	7,10%	51,85%	12,49%
CURY	98,03%	92,59%	95,23%	20,99%	70,37%	32,34%
BSA	92,59%	89,28%	90,90%	26,86%	64,28%	37,89%
MÉDIA GERAL	95,62%	92,59%	94,78%	23,93%	69,19%	35,12%

Fonte: Construção do Autor.

5. Conclusão

O presente trabalho propõe um sistema inovador e eficiente para a extração de informações em Demonstrações Financeiras, que são exemplos típicos de Documentos Visualmente Ricos (VrDs). A solução desenvolvida contempla um modelo de Inteligência Artificial refinado, denominado LayoutDF, que identifica a classe de cada token - textos, títulos e tabelas, além de dois módulos adicionais - DF-Filter e o DF-Hierarchizer - que desempenham um papel importante no pós-processamento e na estruturação de cada parte do documento.

Para treinar e avaliar a eficácia do Extrator de DFs, foi criada uma Coleção Dourada de Demonstrações Financeiras, composta por uma amostra de 22 empresas selecionadas criteriosamente por auditores especializados. Os resultados obtidos na identificação das seções dos documentos, com base em seus títulos, foram satisfatórios e suplantaram sistemas baseline e de IAs generativas baseadas em LLM GPT 3.5, permitindo a aplicações externas realizar verificações e validações automáticas nas Demonstrações Financeiras.

Como trabalhos futuros, nosso objetivo é avançar no estudo e desenvolver uma solução mais genérica, capaz de extrair e estruturar informações em diversos tipos de VrDs como contratos, relatórios anuais e outros documentos relevantes para análise e tomada de decisões, beneficiando outros domínios - jurídico, médico, acadêmico e governamental. Outra linha de investigação será refinar LLMs com um corpus de VrDs e avaliar se, mais ajustados, tais modelos de linguagem largos podem ser usados para a tarefa de estruturação e extração de informações destes documentos.

Referências

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cho, S., Moon, J., Bae, J., Kang, J., and Lee, S. (2023). A framework for understanding unstructured financial documents using rpa and multimodal approach. *Electronics*, 12(4):939.

DAIR.AI (2023). Few-shot prompting. Disponível em: https://www.promptingguide.ai/pt/techniques/fewshot. Acesso em: 29 de Junho 2023.

Déjean, H., Clinchant, S., and Meunier, J.-L. (2022). Layoutxlm vs. gnn: An empirical evaluation of relation extraction for documents. *arXiv preprint arXiv:2206.10304*.

Hooda, N., Bawa, S., and Rana, P. S. (2018). Fraudulent firm classification: a case study of an external audit. *Applied Artificial Intelligence*, 32(1):48–64.

Keocheguerian, I. B. and Martins, V. F. (2021). A utilização da inteligência artificial nos trabalhos de auditoria independente. *Revista Científica e-Locução*, 1(20):21–21.

Sarkhel, R. and Nandi, A. (2019). Visual segmentation for information extraction from heterogeneous visually rich documents. In *Proceedings of the 2019 international conference on management of data*, pages 247–262.

Stubblebine, T. (2003). *Regular expression pocket reference*. "O'Reilly Media, Inc.".

Wang, J., Jin, L., and Ding, K. (2022). Lilt: A simple yet effective language-independent layout transformer for structured document understanding. *arXiv preprint arXiv:2202.13669*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M. (2020a). Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., and Wei, F. (2021). Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*.

Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al. (2020b). Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Ylisiurunen, M. et al. (2022). Extracting semi-structured information from receipts.

Yu, W., Lu, N., Qi, X., Gong, P., and Xiao, R. (2021). Pick: processing key information extraction from documents using improved graph learning-convolutional networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4363–4370. IEEE.