TRANSCRITOR-IA: TRANSCRIÇÃO INTELIGENTE APLICADA AO JORNAL HISTÓRICO O JAGUARIBE

TRANSCRITOR-IA: INTELLIGENT TRANSCRIPTION APPLIED TO THE HISTORICAL NEWSPAPER O JAGUARIBE

Alysson Henrique de Souza Pinheiro*
Raimundo Valter Costa Filho**

RESUMO

A digitalização de jornais históricos é fundamental para a preservação da memória cultural, porém, a transcrição é dificultada por layouts complexos e degradação do material. Este trabalho apresenta o Transcritor-IA, uma ferramenta que automatiza a extração de texto desses documentos. A abordagem integra um modelo de Visão Computacional para segmentação de layout, um sistema de OCR para extração de texto, e um LLM para pós-processamento. Como estudo de caso, a ferramenta foi validada em edições do jornal "O Jaguaribe". Os resultados demonstram que o modelo de segmentação alcançou 79,4% de mAP@50 e o pós-processamento com LLM melhorou a legibilidade, reduzindo a Taxa de Erro de Palavra em 9,7% em comparação com o OCR bruto.

Palavras-chave: OCR. Preservação Digital. Visão Computacional.

ABSTRACT

The digitization of historical newspapers is essential for the preservation of cultural memory, but transcription is hampered by complex layouts and material degradation. This work presents Transcritor-IA, a tool that automates text extraction from these documents. The approach integrates a Computer Vision model for layout segmentation, an OCR system for text extraction, and an LLM for post-processing. As a case study, the tool was validated on editions of the newspaper "O Jaguaribe." The results show that the segmentation model achieved 79.4% mAP@50 and post-processing with LLM improved readability, reducing the Word Error Rate by 9.7% compared to raw OCR.

Keywords: OCR. Digital Preservation. Computer Vision.

* Graduando em Ciência da Computação pelo Instituto Federal de Educação, Ciência e Tecnologia do Ceará – Campus Aracati. E-mail: [alysson.henrique.souza05@aluno.ifce.edu.br]

** Professor do Instituto Federal de Educação, Ciência e Tecnologia do Ceará – Campus Aracati. Doutor em Engenharia de Teleinformática. E-mail: [valter.costa@ifce.edu.br]

1 INTRODUÇÃO

Nas últimas décadas, os avanços em Inteligência Artificial (IA) e Visão Computacional têm impulsionado significativamente o desenvolvimento de sistemas para *Optical character recognition* (OCR), análise de *layout* e extração automatizada de informações em documentos. Essas tecnologias têm sido amplamente incorporadas em iniciativas de preservação histórica e pesquisa científica.

Nos Estados Unidos, o programa *Chronicling America* já digitalizou mais de 16 milhões de páginas de jornais históricos, disponibilizando imagens em alta resolução e textos completos gerados via OCR (LEE et al., 2020). Na Europa, a plataforma *Europeana* concentra mais de 60 milhões de itens digitalizados do patrimônio cultural europeu, promovendo o acesso unificado à informação (CAPURRO; PLETS, 2020).

No Brasil, uma pesquisa conduzida por Lima (2021) aponta que aproximadamente 98% das instituições culturais possuem algum tipo de acervo digital. No entanto, por limitações de financiamento ou de mão de obra especializada, mesmo nas instituições mais ativas o conteúdo digitalizado representa menos da metade do acervo físico.

A preservação dessa história é um dos temas de atuação do Museu Jaguaribano. Considerando que as edições físicas do periódico não se encontram em um acervo unificado, a criação de uma ferramenta que automatiza a conversão de seu conteúdo para um formato digital representa uma contribuição prática para a pesquisa de fontes documentais da cidade. Este trabalho se insere, portanto, como uma ferramenta de apoio a um projeto de pesquisa mais amplo sobre a transcrição de documentos históricos[1].

A metodologia descrita nesse trabalho foi aplicada em edições do jornal "O Jaguaribe", um periódico que circulou em Aracati entre os anos de 1930 e 1959. O jornal documentou eventos e o cotidiano da região do Vale do Jaguaribe, sendo, um objeto de estudo para a história local.

A estrutura deste artigo está organizada da seguinte maneira: fundamentação teórica necessária à compreensão das tecnologias adotadas, seguida pela revisão de trabalhos correlatos que contextualizam o estado da arte. São descritos a metodologia utilizada e o corpus de validação. Por fim são detalhados os resultados obtidos e perspectivas para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Nesta seção, é estabelecida a fundamentação teórica para o desenvolvimento do módulo de transcrição proposto. A abordagem inicia com a descrição do processo fundamental de OCR. Em seguida, são detalhadas as tecnologias complementares utilizadas no desenvolvimento do trabalho.

---

[1] Rodrigo C. et al. *Transcritor-IA: utilizando inteligência artificial para transcrição de manuscritos históricos*. In: Anais do LII Seminário Integrado de Software e Hardware. Maceió/AL: SBC, 2025. p. 133-144. DOI: 10.5753/semish.2025.7698

2.1 Reconhecimento Óptico de Caracteres

O OCR é uma tecnologia que converte documentos digitalizados ou imagens contendo texto impresso em arquivos de texto editáveis e pesquisáveis, podendo ser salvos em formatos como *ASCII* ou *UNICODE* (SHINDE; CHOUGULE, 2012). Essa tecnologia apresenta ampla aplicação em processos de digitalização de acervos, automação de entrada de dados e sistemas de arquivamento.

Conforme ilustrado na Figura 1, o processo de OCR pode ser dividido em três fases principais, pré-processamento, segmentação e classificação.

Figura 1 – Estrutura fundamental de um sistema de OCR destacando o fluxo sequencial de processamento desde a imagem de entrada até a geração do texto transcrito.

Fonte: Elaborado pelos autores (2025)

O pré-processamento é um estágio fundamental que segue para o estágio de extração de *features* ele regula a adequação dos resultados para os estágios consecutivos (KARTHICK et al., 2019). Processos comuns durante essa etapa incluem, mudança para tons de cinza, aplicação de técnica para binarização e ajuste de orientação na imagem, essas operações padronizam a entrada e otimizam o funcionamento das etapas seguintes.

Na sequência, a fase de segmentação divide a imagem pré-processada em partes contendo apenas *pixels* relevantes para análise. Essa fase serve para que o algoritmo de OCR possa focar em regiões específicas da imagem, reduzindo a complexidade do reconhecimento e aumentando a precisão.

Por fim, na etapa de extração de características cada letra é convertida para um vetor de onde são extraídas as suas características (PRASAD; JAYANTA, 2013). Esse processo converte cada letra em um conjunto de números que descrevem suas propriedades visuais como formato dos contornos, dimensões e padrões de pixel. A partir dessas características, o sistema compara cada letra da imagem com modelos previamente armazenados e determina o caractere correspondente.

O uso de OCR diretamente em páginas com *layouts* complexos tende a ser ineficiente. Para resolver isso, aplica-se a etapa de segmentação da imagem. No entanto, documentos como jornais apresentam uma grande variedade na forma como o texto pode estar organizado, o que exige a aplicação de métodos mais especializados para uma segmentação eficaz.

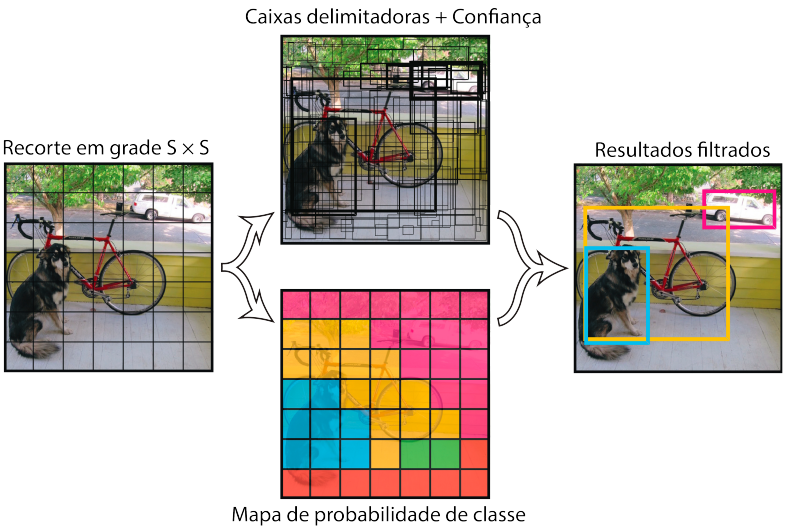
Nesse contexto, a arquitetura introduzida por Redmon et al. (2016) com o artigo *You Only Look Once: Unified, Real-Time Object Detection*, ou *You Only Look Once* (YOLO), apresenta-se como uma solução. Diferente de métodos tradicionais que seguem *pipelines* complexas ou aplicam redes neurais convolucionais de forma seletiva em múltiplas regiões da imagem, o

YOLO trata a detecção de objetos com operações de convolução sequenciais, esquadrinhando a imagem e localizando possíveis padrões.

O funcionamento de um modelo YOLO, pode ser separado em três estágios. Inicialmente, a imagem de entrada é dividida em uma grade de $S \times S$ células. Em seguida, a rede processa a imagem inteira de uma só vez e prevê caixas delimitadoras (*bounding boxes*) e probabilidades de classe para cada região simultaneamente (ZOU et al., 2023).

Esse processo gera um grande número de caixas candidatas, muitas delas sendo sobrepostas e redundantes, para refinar esses resultados, é selecionada a caixa com a maior pontuação de confiança, e são removidas as caixas que possuem uma sobreposição acima de um determinado limiar, resultando nas detecções finais.

Figura 2 – Exemplo de estágios de processamento do modelo YOLO.



Fonte: Adaptado de (REDMON et al., 2016))

2.2 Pós-processamento e Correção com Modelos de Linguagem

Após a extração dos textos do documento pelo sistema de OCR, os dados produzidos frequentemente ainda apresentam erros de transcrição. Para tratar esses problemas e melhorar a utilidade do conteúdo extraído, podem ser utilizadas abordagens de pós-processamento.

Os modelos de linguagem natural são sistemas computacionais capazes de compreender, processar e gerar texto de forma contextualmente relevante (JURAFSKY; MARTIN, 2025). Esses modelos são treinados em grandes volumes de dados textuais para aprender padrões linguísticos. Isso permite que realizem diversas tarefas, incluindo correção ortográfica e gramatical, tradução automática, sumarização de texto e geração de conteúdo.

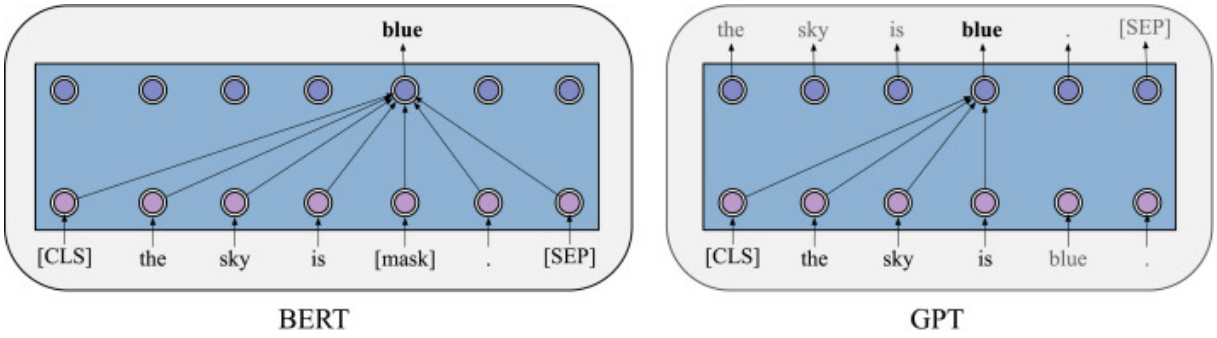
Historicamente a modelagem de linguagem natural baseava-se em métodos estatísticos. Um desses que ainda é utilizado até hoje é o N-gram, utilizado em sistemas como o *PyLaia* esse algoritmo modela a probabilidade de ocorrência de uma sequência de k itens baseando-se na

frequência com que sequências de $k-1$ itens são seguidas pelo k-ésimo item em um conjunto de treinamento.

Apesar de sua utilidade, especialmente em contextos com recursos limitados ou para erros mais simples, esses modelos linguagem enfrentam limitações como incapacidade de capturar dependências de longo alcance no texto, dificuldade em lidar com erros contextuais complexos e dependência de dicionários estáticos.

O marco transformador dessa área foi a introdução da arquitetura *Transformer* por Vaswani et al. (2017), que revolucionou o processamento de linguagem natural com o mecanismo de auto-atenção. Esse mecanismo é a base de arquiteturas de modelos como o *Bidirecional Encoder Representations from Transformers* (BERT) e o *Generative pre-trained transformer* (GPT).

Figura 3 – Demonstração das diferentes formas como o mecanismo de auto-atenção utiliza o contexto para construir a representação de um *token*, diferenciando a atenção bidirecional do BERT com a atenção unidirecional do GPT.



Fonte: (HAN et al., 2021)

O mecanismo de auto-atenção, como definido por (HAN et al., 2021) e ilustrado na Figura 3, permite ao modelo ponderar dinamicamente a importância de diferentes partes da sequência de entrada. Em arquiteturas como o BERT, a atenção é bidirecional, construindo representações a partir de tokens anteriores e posteriores. Em contraste, modelos como o GPT, empregam atenção unidirecional onde a previsão considera apenas os *tokens* anteriores.

A contínua evolução e o aprimoramento de arquiteturas baseadas no *Transformer*, como as exemplificadas pelo BERT e GPT, culminaram no desenvolvimento dos *Large Language Models* (LLM). Esses modelos, que herdam e expandem os mecanismos de atenção, são treinados com volumes massivos de dados, dando a eles capacidades ainda maiores de compreensão e geração de linguagem. A família *Large Language Model Meta AI* (LLaMA), desenvolvida pela *Meta*©, é um exemplo recente de modelos de código aberto que demonstram alta eficiência e desempenho competitivo em *benchmarks* linguísticos (TOUVRON et al., 2023).

3 TRABALHOS CORRELATOS

Esta seção apresenta estudos recentes relacionados à segmentação e extração de texto em documentos digitalizados. Os trabalhos a seguir têm foco em abordagens baseadas em redes

neurais profundas, modelos híbridos e a aplicação de ferramentas de OCR em documentos com *layouts* complexos.

Pagani (2023) propõe um sistema para extração de textos-chave em artigos científicos digitalizados, utilizando um *dataset* composto por artigos presentes nos Simpósios Brasileiros de Telecomunicações, publicados entre 1983 e 1998. Nesse trabalho, o autor testou duas abordagens. A primeira, utilizando a ferramenta *Tesseract* para classificar as regiões, não alcançou resultados satisfatórios. Na segunda abordagem, um modelo de *Faster R-CNN* pré-treinado com o *dataset PubLayNet* (ZHONG; TANG; YEPES, 2019) e ajustado com *transfer learning* foi usado para a segmentação. A *engine* de OCR *Tesseract* foi usada para extrair o texto das regiões identificadas.

Umer et al. (2021) apresentam uma técnica para segmentação de regiões texto em documentos com *layouts* complexos. A proposta consiste em realizar múltiplos recortes da imagem em diferentes resoluções, gerando regiões candidatas à presença de texto. Cada recorte é classificado por uma *Convolutional Neural Network* (CNN) em três categorias: texto, não texto e ambíguo. Para as regiões ambíguas, aplica-se um pós-processamento com particionamento recursivo, resultando na segmentação final. Os autores relatam desempenho superior a métodos tradicionais no *dataset ICDAR*, avaliados por precisão, *recall*, *F1-score* e acurácia de segmentação.

Sven e Matteo (2022) realizam uma análise comparativa entre diferentes estratégias de segmentação e classificação de seções em documentos históricos. Para isso o estudo utiliza o *dataset GT4HistCommentLayout*. Testando abordagens visuais com a arquitetura YOLOv5 nas variantes YOLO-Mono para detecção monoclasse e YOLO-Multi para multiclasse, textuais com o modelo *RoBERTa*, multimodal com *LayoutLMv3* e híbrida combinando YOLO-Mono com *LayoutLM* e *RoBERTa*.

Os resultados apresentados mostram que a abordagem visual com YOLO-Multi teve o melhor desempenho geral, superando o modelo multimodal *LayoutLMv3*. O *RoBERTa* teve o pior desempenho, e o *LayoutLMv3*, embora superior ao *RoBERTa*, mostrou depender das características visuais e de coordenadas em vez das textuais para documentos desse tipo.

Para facilitar a visualização das principais características e diferenças entre as abordagens discutidas e a proposta deste trabalho, a Tabela 1 apresenta um comparativo das abordagens.

Tabela 1 – Comparativo das abordagens de trabalhos relacionados e a proposta deste trabalho, destacando características relevantes.

	Pagani et al.	**Umer et al.**	**Sven et al.**	**Pinheiro et al.**
Processa layouts complexos		✓	✓	✓
Processa documentos degradados			✓	✓
Segmenta regiões de texto	✓	✓	✓	✓
Realiza extração de texto	✓			✓

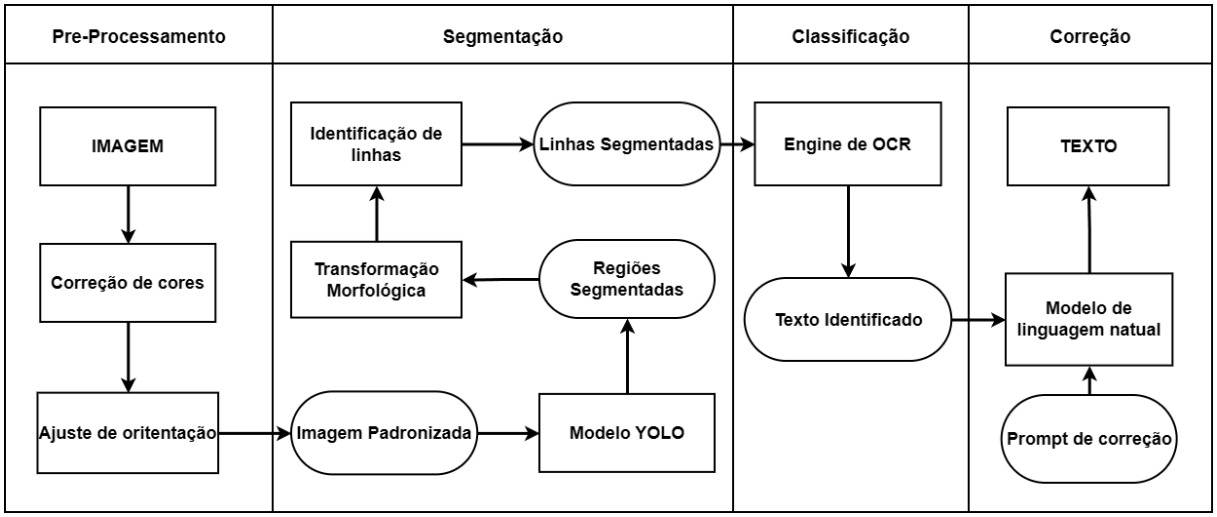
4 MATERIAIS E MÉTODOS

Esta seção apresenta os procedimentos e materiais utilizados no desenvolvimento e validação da ferramenta desenvolvida para a transcrição automática de jornais históricos.

O objetivo foi construir uma ferramenta capaz de detectar e segmentar regiões de texto presentes em imagens de jornal independente do *layout* e converter esses segmentos de imagens em texto digital, minimizando os erros de transcrição e preservando a estrutura do conteúdo original.

Para isso, foi desenvolvida uma solução modular onde cada estágio é responsável por uma tarefa específica do processo. A execução é sequencial e a saída de um módulo serve como entrada para o seguinte. A Figura 4 ilustra as quatro etapas principais da *pipeline*: pre-processamento, segmentação de regiões e linhas, extração de texto, e correção com modelo de linguagem.

Figura 4 – Operações dos módulos da *pipeline* de pré-processamento da imagem, reconhecimento de regiões, extração de conteúdo até correção do texto.



Fonte: Elaborado pelos autores (2025)

4.1 Segmentação de regiões de texto com *YOLO*

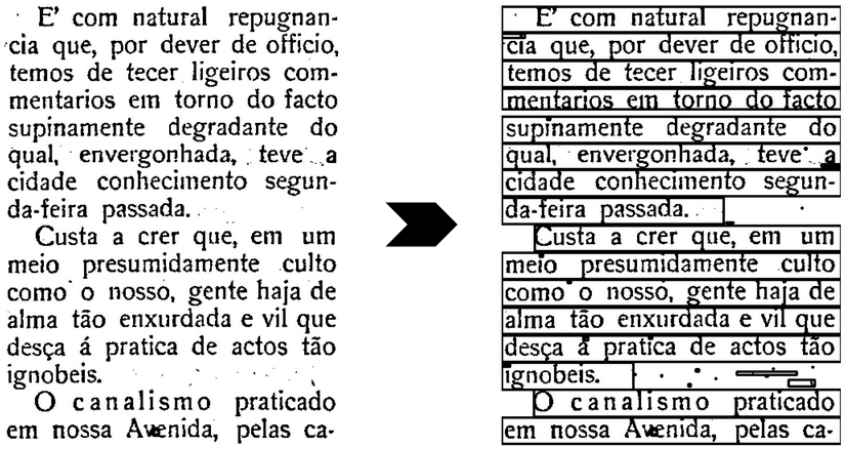
Para a identificação das regiões de texto foi utilizada uma CNN baseada na arquitetura YOLOv11 (JOCHER; QIU, 2024). O modelo de identificação foi treinado na plataforma *Colab*, utilizando uma *GPU NVIDIA T4* por 400 épocas.

O treinamento utilizou o *dataset* (SWEMPER, 2024), esse *dataset* originalmente composto por 3.051 imagens de jornais, panfletos e páginas de livros foi enriquecido com transformações aleatórias de exposição variando de -10% a $+10\%$ e adição de ruido em até 2% dos *pixels* resultando em um total de 5.076 imagens, sendo 80% usadas para o treinamento 10% para validação e 10% para testes.

4.2 Segmentação das Linhas de Texto

A segmentação das linhas dentro de cada região é realizada por meio de análise de projeção horizontal e vertical (O'GORMAN, 1993). Para aumentar a precisão, a imagem é convertida para escala de cinza, submetida à binarização por limiarização adaptativa gaussiana e a operações morfológicas de erosão e dilatação dos *pixels* pretos.

Figura 5 – Exemplo do resultado de segmentação de linhas em recorte de região de texto pre-processado.



Fonte: Elaborado pelos autores (2025)

A projeção horizontal identifica os limites superior e inferior de cada linha, descartando como ruído os segmentos com altura menor que 10 *pixels*. Em seguida, a projeção vertical determina os limites laterais, produzindo um par de coordenadas usados para gerar um recorte da linha.

4.3 Extração de Texto

O texto de cada recorte de linha é extraído com a *engine* de OCR *Tesseract* (SMITH, 2007), carregada através da biblioteca *pytesseract*. A *engine* foi configurada para o idioma português e com modo de segmentação *–psm 7*, uma otimização específica para imagens que contêm uma única linha de texto.

4.4 Correção com Modelo de Linguagem

O texto bruto extraído é pós-processado pelo LLM *LLaMA-3.3-70b*, acessado via *API* da *Groq*. A correção é realizada enviando o texto em blocos de 20 linhas, no formato *JSON*. O modelo é instruído por dois *prompts* de sistema sobre como realizar a correção.

O primeiro *prompt* contextualiza a tarefa, orientando o modelo a considerar o estilo do português da época dos documentos, as imperfeições típicas do OCR em material degradado e o formato de resposta esperado. O segundo *prompt* estabelece regras de correção, incluindo instruções para considerar os valores de confiança do OCR e utilizar o contexto das linhas adjacentes para realizar correções coerentes.

4.5 Dataset

Para avaliar a qualidade da transcrição, um gabarito de 846 linhas de texto transcritas manualmente foi criado a partir de uma amostra aleatória do corpus de validação. Para facilitar uso em trabalhos futuros e garantir a reprodutibilidade dos resultados o conjunto de dados do gabarito montado junto das imagens de onde o texto foi removido e suas respectivas coordenadas estão disponibilizados publicamente na plataforma *Kaggle*.[2]

O corpus utilizado para montagem do gabarito foi obtido através da Casa da Cultura de Aracati[3] ele consiste em edições digitalizadas do jornal "O Jaguaribe", publicadas entre os anos de 1930 e 1959.

Como é demonstrado na Figura 6 as imagens seguem o formato tradicional de artigos jornalísticos da época, com o conteúdo disposto em múltiplas colunas verticais.

Figura 6 – Exemplo de página frontal do jornal O Jaguaribe, edição de 24 de julho de 1932, apresentando múltiplas colunas de texto e elementos gráficos típicos da época.

O JAGUARIBE

SEMANARIO DE ORIENTAÇÃO CATHOLICA

Director — JOÃO FREIRE DE ANDRADE Gerente — JAIRE FREIRE DE ANDRADE

ANNO III | Ceará - Aracaty, 24 de Julho de 1932 | NUM. 106

A REBELLIÃO E O CEARÁ

PARA a Republica recem-nada, assente sob os melhores auspicios, num assomo de puro idealismo, pela juventude brasileira que, restaurando a gloria dos nossos maiores, realisou em Outubro de 1930 uma das paginas mais memoraveis da nossa historia politica; o momento que passa é a balisa que vae demarcar a sua existencia e decidir dos seus destinos.

Até agora, tudo nos persuade de que o Governo Provisorio, animado numa serena compenetração dos seus graves e instantes deveres, conduz os exercitos da Republica para o seu final triumpho.

Isto, porém, não teria tão sensivel relevancia, si a tropa victoriosa marchasse sobre os cadaveres dos nossos irreflectidos e bravos irmãos, si o pavilhão a cuja sombra combatem se ennodoasse na sangueira dum estremio sacrificio.

Felizmente, pelas noticias vehiculadas nos jornaes citadinos, o Governo se empenha e põe timbre em poupar a vida dos seus adversarios, nos mais acesos encontros procura assediá-los para lhes ensejar uma capitulação, que jamais pretenda diminuir os brios que lhes inspiram, nem desdoirar o civismo inflamado

A maldição dos seus antepassados cahiria sobre a sua cabeça.

Caim sempre foi repudiado pelo genero humano e, quando surge á nossa lembrança, deixa-nos um rastro negro na memoria.

Assim pensou e executou o Chefe das forças em operações na vanguarda.

E, por mais que São Paulo atarde a sua resolução, a garganta das metralhadoras ainda não vomitou a morte e o exterminio sobre o seu povo.

E', que a dictadura civil, ajudada na consecução de sua obra constructora por espadas que só se deslocaram do estojo para golpear fundo, num arremesso de são e santo patriotismo os apodrecidos fantoches da Republica esclerosada, os corsarios impudicos que assaltavam o liberalismo das suas instituições; não quer consentir em que ellas se voltem contra as incautas victimas do despeito alheio.

Porém, emquanto isso se esboça na téla da nossa vida republicana, o Ceará desventurado, mais uma vez, transforma o seu sertão escaldante e calcinado numa ribalta immensa, onde encena o seu heroismo sombrio e silencioso.

Posto as gambiarras projectem sobre nós as suas luzes, para destacar aos olhos do Paiz a apotheose da eschyliana tragedia, a impie-

Dr. JOÃO PESSOA

Fonte: (FREIRE, 2025))

Além dos blocos de texto as páginas incluem elementos variados como títulos em destaque, tabelas, anúncios publicitários e separações visuais entre seções. As digitalizações apresentam deteriorações típicas de documentos históricos, como manchas, descoloração, rasgos, e áreas com perda parcial ou total do conteúdo.

O conjunto completo é composto por 7.308 imagens digitais das páginas frontais do jornal, todas no formato *Joint Photographic Experts Group* (JPG) com resolução de 2000×3000 *pixels*.

[2] <https://www.kaggle.com/datasets/alyssonhenrique/o-jaguaribe>

[3] <https://www.casadaculturadearacati.org.br/hemeroteca>

5 DISCUSSÃO DOS RESULTADOS

Nesta seção, são apresentados os resultados da avaliação de desempenho dos dois principais componentes da ferramenta desenvolvida, o modelo de detecção de regiões de texto e o processo de correção de texto com LLM.

5.1 Modelo de detecção de regiões

A performance do modelo YOLO foi avaliada com base nas métricas de precisão, *recall* e *Mean Average Precision* (*mAP50*) com sobreposição de até 50%. A Tabela 2 demonstra a evolução dessas métricas durante o período de treinamento.

Embora o treinamento tenha durado 400 épocas, o modelo atingiu uma estabilidade de desempenho por volta da época 300, indicando que havia se aproximado de seu potencial máximo de aprendizado. O pico de performance ocorreu na época 392, onde o modelo alcançou $80,1\%$ de precisão, $78,2\%$ de *recall* e *mAP50* de $79,4\%$.

Tabela 2 – Evolução da precisão, *recall* e *mAP50* do modelo de identificação de regiões.

Época	**Precisão**	***Recall***	***mAP50***
50	75,6%	70,4%	72,8%
100	78,8%	76,6%	77,0%
200	80,2%	76,1%	78,0%
300	80,8%	76,8%	79,1%
400	80,4%	78,0%	79,4%
Melhor Modelo	**80,1%**	**78,2%**	**79,4%**

5.2 Transcrição do Texto

Para avaliar a qualidade do texto as imagens do gabarito foram transcritas pela ferramenta e os resultados gerados foram comparados ao gabarito em duas condições, o texto bruto extraído pelo OCR e o texto após a correção pelo LLM. As métricas utilizadas foram *Character Error Rate* (CER) e*Word Error Rate* (WER) que medem a porcentagem de erro em caracteres individuais e nas palavras completas, respectivamente.

Os resultados para cada abordagem são apresentados na Tabela 3.

Tabela 3 – Métricas de desempenho entre o texto bruto do OCR e o texto corrigido pelo LLM.

Abordagem	**WER**	**CER**
OCR	34,00%	12,35%
LLM	24,30%	13,16%

Os resultados obtidos mostram que o LLM foi eficaz em reduzir a taxa de erro de palavras apresentando uma redução de $9,7\%$. Isso demonstra sua capacidade de corrigir termos mal interpretados pelo OCR e reconstruir a semântica do texto, melhorando a legibilidade geral.

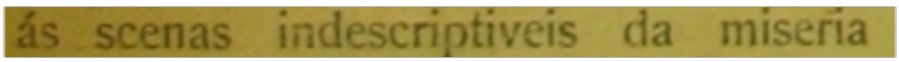
Figura 7 – Exemplo de correção positiva, em que o LLM aprimora a transcrição e mantém a fidelidade ao texto original.

ORIGINAL	
OCR	agora, que a Paz, o Direito e **-z_rgstiça** cons-
LLM	agora, que a Paz, o Direito e **a Justiça** cons-

Fonte: Elaborado pelos autores (2025)

Contudo, a melhoria não ocorreu de forma uniforme. Houve um aumento na taxa de erro de caracteres do texto corrigido, em relação ao texto bruto. Esse aumento é pode ser atribuído a uma tendência do LLM em realizar correções excessivas em texto já correto.

Figura 8 – Exemplo de correção negativa, em que o LLM insere novos erros em uma linha já transcrita corretamente pelo OCR.

ORIGINAL	
OCR	ás **scenas indescriptiveis** da miseria
LLM	ás **cenas indescritíveis** da miseria

Fonte: Elaborado pelos autores (2025)

Das linhas do gabarito, $19,6\%$ já haviam sido transcritas corretamente pelo OCR, em $30,7\%$ dessas linhas o LLM realizou correção excessiva, como demonstrado na figura 8, ao substituir uma palavra como "scenas" pela sua forma moderna "cenas", o LLM introduz erros a nível de caractere em relação ao texto original, o que afeta negativamente o CER.

6 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho apresentou o desenvolvimento e a validação do Transcritor-IA, uma ferramenta para a transcrição automática de jornais históricos que integra segmentação de layout com o modelo YOLO, extração de texto via OCR e pós-processamento com um LLM. Os resultados demonstram o potencial da abordagem, o modelo de segmentação alcançou $79,4\%$ de *mAP50*, confirmando sua eficácia na identificação de regiões de texto em documentos com vários graus de degradação.

No pós-processamento, a utilização do LLM se mostrou uma estratégia promissora, reduzindo a taxa de erro de palavra em $9,7\%$, o que representa uma melhora substancial na legibilidade e coerência do texto transcrito. Apesar disso, a avaliação revelou uma tendência do modelo em modernizar a ortografia original, o que, apesar de corrigir erros semânticos, introduziu novas imprecisões a nível de caractere.

Para mitigar essa limitação e preservar a fidelidade histórica dos documentos, os trabalhos futuros se concentrarão no refinamento da etapa de correção. As estratégias incluem a integração de um dicionário de época para guiar o LLM e o *fine-tuning* do modelo com um corpus de textos históricos, a fim de tornar o modelo sensível às particularidades do português antigo. A implementação dessas melhorias tem o potencial de consolidar o Transcritor-IA como um recurso valioso para pesquisadores, arquivistas e para o avanço da área de Humanidades Digitais.

REFERÊNCIAS

CAPURRO, C.; PLETS, G. Europeana, edm, and the europeanisation of cultural heritage institutions. **Digital Culture & Society**, transcript Verlag, v. 6, n. 2, p. 163–190, 2020.

FREIRE, K. M. **Hemeroteca**. 2025. Instituto José Freire d'Andrade. Visited on 2025-05-20. Disponível em: <https://www.casadaculturadearacati.org.br/>.

HAN, X. et al. Pre-trained models: Past, present and future. **AI Open**, Elsevier, v. 2, p. 225–250, 2021.

JOCHER, G.; QIU, J. Ultralytics yolo11. 2024. **URL https://github. com/ultralytics/ultralytics**, 2024.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models**. 3rd. ed. [s.n.], 2025. Online manuscript released January 12, 2025. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>.

KARTHICK, K. et al. Steps involved in text recognition and recent research in ocr; a study. **International Journal of Recent Technology and Engineering**, v. 8, n. 1, p. 2277–3878, 2019.

LEE, B. C. G. et al. The newspaper navigator dataset: extracting and analyzing visual content from 16 million historic newspaper pages in chronicling america. **arXiv preprint arXiv:2005.01583**, 2020.

LIMA, L. P. B. A digitalização de acervos no brasil segundo a pesquisa tic cultura. **Revista Brasileira em Humanidades Digitais**, v. 1, n. 1, p. 1–17, 2021.

O'GORMAN, L. The document spectrum for page layout analysis. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 15, n. 11, p. 1162–1173, 1993.

PAGANI, Y. N. Extração e classificação de textos-chave de artigos acadêmicos utilizando modelo de reconhecimento óptico de caracteres. **Repositório Institucional da UFSC**, 12 2023. Disponível em: <https://repositorio.ufsc.br/handle/123456789/253659>.

PRASAD, V.; JAYANTA, Y. A study on method of feature extraction for handwritten character recognition. **Indian Journal of Science and Technology**, v. 6, p. 174–178, 03 2013.

REDMON, J. et al. You only look once: Unified, real-time object detection. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 779–788.

SHINDE, A. A.; CHOUGULE, D. Text pre-processing and text segmentation for ocr. **International Journal of Computer Science Engineering and Technology**, Citeseer, v. 2, n. 1, p. 810–812, 2012.

SMITH, R. An overview of the tesseract ocr engine. In: IEEE. **Ninth international conference on document analysis and recognition (ICDAR 2007)**. [S.l.], 2007. v. 2, p. 629–633.

SVEN, N.-M.; MATTEO, R. Page Layout Analysis of Text-heavy Historical Documents: A Comparison of Textual and Visual Approaches. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2212.13924>.

SWEMPER. Open Source Dataset, **TrainingWS2 Dataset**. Roboflow, 2024. <https://universe.roboflow.com/swemper-annotation/training_ws_2>. Visited on 2025-08-28. Disponível em: <https://universe.roboflow.com/swemper-annotation/training_ws_2>.

TOUVRON, H. et al. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

UMER, S. et al. Deep features based convolutional neural network model for text and non-text region segmentation from document images. v. 113, p. 107917, 12 2021. ISSN 15684946. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1568494621008395>.

VASWANI, A. et al. Attention is all you need. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.

ZHONG, X.; TANG, J.; YEPES, A. J. **PubLayNet: largest dataset ever for document layout analysis**. 2019. Disponível em: <https://arxiv.org/abs/1908.07836>.

ZOU, Z. et al. Object detection in 20 years: A survey. **Proceedings of the IEEE**, IEEE, v. 111, n. 3, p. 257–276, 2023. Disponível em: <https://arxiv.org/abs/1905.05055#>.