ANÁLISE COMPARATIVA DE MODELOS PARA PROCESSAMENTO E CLASSIFICAÇÃO AUTOMÁTICA DE DOCUMENTOS PESSOAIS NO SETOR IMOBILIÁRIO

Matheus Machado Laureano[1], Marlon de Matos de Oliveira[2]

Resumo: Este trabalho apresenta o desenvolvimento e a análise comparativa de dois modelos voltados à classificação automática de documentos utilizados em processos comerciais de construtoras. O primeiro modelo adota uma abordagem tradicional, baseada em regras e expressões regulares, enquanto o segundo aplica técnicas de Inteligência Artificial supervisionada, utilizando vetorização TF-IDF e regressão logística. Ambos os modelos foram testados sobre um conjunto de 52 documentos reais, abrangendo certidões de nascimento, certidões de casamento, CPFs, CNHs e comprovantes de residência. O objetivo central foi avaliar a precisão, escalabilidade e eficiência de cada abordagem frente às variações de formato e qualidade dos arquivos. Os resultados demonstraram que o modelo baseado em IA obteve desempenho superior, alcançando acurácia média de 92% e *F1-score* de 0,91, em comparação aos 81% de acurácia do modelo baseado em regras. Conclui-se que a automação documental mediada por IA representa um avanço significativo na eliminação de processos manuais e na otimização do fluxo de recebimento de propostas comerciais no setor da construção civil.

Palavras-chave: inteligência artificial; processamento de linguagem natural; classificação de documentos; OCR; construção civil.

[1]Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), matheus030702@unesc.net

[2]Orientador. Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), marlon.oliveira@unesc.net

ABSTRACT: This paper presents the development and comparative analysis of two models aimed at the automatic classification of documents used in the commercial processes of construction companies. The first model adopts a traditional approach based on rules and regular expressions, while the second applies supervised Artificial Intelligence techniques, using TF-IDF vectorization and logistic regression. Both models were tested on a set of 52 real documents, covering birth certificates, marriage certificates, ID cards (CPF), driver's licenses (CNH), and proof-of-residence documents. The main objective was to evaluate the accuracy, scalability, and efficiency of each approach in handling variations in format and image quality. The results showed that the AI-based model achieved superior performance in terms of accuracy and recall, although it requires greater computational power and larger training datasets for robust learning. It is concluded that document automation driven by AI represents a significant advancement in eliminating manual processes and optimizing the flow of commercial proposal handling in the construction industry.

Keywords: artificial intelligence; natural language processing; document classification; OCR; construction industry.

1 INTRODUÇÃO

O avanço da transformação digital tem modificado profundamente a forma como organizações produzem, armazenam e utilizam informações. No setor da construção civil, que historicamente enfrenta processos administrativos complexos e intensivos em documentação, essa mudança se traduz em uma pressão crescente por eficiência e precisão na gestão de documentos.

Propostas comerciais, contratos, comprovantes de renda e cadastros de clientes compõem um volume expressivo de informações que precisam ser analisadas e validadas em prazos cada vez mais curtos. A execução manual dessas atividades torna-se um gargalo, pois está sujeita a erros humanos, atrasos e retrabalho, comprometendo diretamente a confiabilidade dos dados e a competitividade das empresas (Davenport; Ronanki, 2018).

Nesse cenário, a Inteligência Artificial (IA) apresenta-se como uma tecnologia capaz de transformar a gestão documental, oferecendo velocidade, precisão e escalabilidade no processamento de dados. Segundo Russell e Norvig (2021), a IA já demonstrou resultados consistentes em tarefas repetitivas e baseadas em regras, superando a intervenção humana

em acurácia e eficiência. Subcampos como o aprendizado de máquina (Machine Learning) permitem treinar modelos capazes de reconhecer padrões em documentos, enquanto o Processamento de Linguagem Natural (PLN) possibilita interpretar e organizar informações textuais não estruturadas (Jurafsky; Martin, 2021). Complementarmente, o Reconhecimento Óptico de Caracteres (OCR) viabiliza a conversão de documentos digitalizados em texto processável (Smith, 2007), funcionando como etapa inicial em fluxos de automação.

Diversos estudos reforçam o potencial dessas tecnologias em contextos distintos. Possamai (2024) investigou a aplicação de modelos de Processamento de Linguagem Natural, comparando técnicas tradicionais com modelos de última geração como o BERT. Seu trabalho demonstrou ganhos significativos de precisão na classificação de documentos em PDF, evidenciando que modelos mais robustos conseguem lidar melhor com a complexidade de textos longos e não estruturados. Esses resultados mostram que o uso de algoritmos avançados pode aumentar a confiabilidade no reconhecimento e na categorização de documentos.

No campo jurídico, Silva, Ribeiro e Dezani (2023) aplicaram técnicas de PLN para a classificação automática de documentos cíveis e criminais. O estudo apresentou resultados expressivos de acurácia, mostrando a viabilidade da automação em domínios que exigem elevado rigor no tratamento da informação. Embora direcionado ao direito, esse trabalho reforça que o PLN é capaz de lidar com documentos de linguagem formal e estruturalmente variados, o que se aproxima do desafio enfrentado em propostas comerciais.

Antonio (2019) apresentou um protótipo baseado em OCR e visão computacional para reconhecimento de rótulos de produtos. Seu trabalho destacou a importância da etapa inicial de extração de dados a partir de imagens ou arquivos digitalizados, apontando como a qualidade da captura pode influenciar diretamente no desempenho dos modelos posteriores. Embora aplicado ao varejo, o estudo evidencia que a integração entre OCR e algoritmos de classificação é essencial em qualquer processo que dependa de documentos não estruturados.

Pajeú et al. (2018) exploraram métodos arquivísticos para organização de documentos digitais pessoais em nuvem. O estudo evidenciou a relevância de estratégias de classificação estruturada, mostrando que a simples digitalização não resolve os desafios de organização e acesso eficiente à informação. Apesar de focar no nível individual, o trabalho traz

contribuições conceituais que podem ser estendidas ao ambiente corporativo, em especial para empresas que lidam com alto volume de documentos digitais.

Complementarmente, Lai e Xu Liheng e Liu (2015) propuseram o modelo RCNN (Recurrent Convolutional Neural Network) para classificação de textos, buscando solucionar uma limitação comum em métodos tradicionais, que tratavam as palavras de forma isolada e perdiam o contexto global das sentenças. O RCNN combinou camadas recorrentes (responsáveis por capturar dependências de longo alcance no texto) com camadas convolucionais, que destacam os padrões mais relevantes. Essa arquitetura permitiu eliminar a necessidade de criação manual de atributos e, ao mesmo tempo, preservar informações semânticas importantes. Os experimentos realizados em quatro bases de dados distintas mostraram que o modelo obteve desempenho superior a abordagens anteriores, reforçando o potencial das redes neurais híbridas para tarefas de classificação textual em contextos complexos e heterogêneos.

Apesar dos avanços, a literatura ainda apresenta limitações quando trata-se de aplicações direcionadas ao setor da construção civil, especialmente no âmbito comercial, onde o grande volume e a heterogeneidade dos documentos exigem soluções robustas e específicas. É nessa lacuna que este trabalho busca atuar: a escassez de estudos voltados à aplicação prática de IA para reconhecimento, classificação e validação de documentos comerciais no ramo imobiliário. Embora existam pesquisas que tratam de automação documental em domínios jurídicos, corporativos ou pessoais, faltam abordagens direcionadas às necessidades de construtoras e incorporadoras, em que a documentação representa a porta de entrada para novos negócios e impacta diretamente a geração de receita.

Diante disso, o objetivo geral deste trabalho é comparar métodos de aprendizado de máquina aplicados à leitura, extração de dados e classificação de documentos digitais, avaliando seu potencial para otimizar o processamento e a análise documental em empresas do setor da construção civil. Especificamente, busca-se aplicar técnicas de OCR na extração de informações, utilizar visão computacional na avaliação de qualidade documental, empregar algoritmos de aprendizado de máquina na classificação e realizar testes comparativos de eficiência. Ao alinhar tecnologia e gestão, este estudo pretende contribuir para a redução de falhas, a agilidade nos processos comerciais e a consolidação da competitividade organizacional.

Este trabalho está estruturado em quatro seções principais. A

Introdução apresenta a contextualização do tema, a definição do problema, os objetivos e a justificativa da pesquisa. Em seguida, em Materiais e Métodos, descrevem-se as ferramentas utilizadas, a metodologia aplicada e as etapas de desenvolvimento do estudo. A seção de Discussão e Resultados reúne a análise crítica e comparativa dos dados obtidos a partir dos experimentos realizados. Por fim, a Conclusão retoma os objetivos, destaca os principais achados do trabalho e sugere caminhos para pesquisas futuras.

2 MATERIAIS E MÉTODOS

O presente trabalho constitui um estudo comparativo entre duas abordagens de classificação automática de documentos digitais, aplicadas ao contexto comercial de construtoras. Ambas as abordagens partem de uma mesma etapa inicial de extração e tratamento textual, realizada por meio de técnicas de leitura de PDFs (pdfminer) e Reconhecimento Óptico de Caracteres (OCR), que converte os arquivos em texto legível para posterior análise.

A primeira abordagem representa um modelo baseado em regras fixas, utilizando expressões regulares e condições lógicas para identificar padrões específicos e determinar o tipo de documento analisado. Já a segunda abordagem faz uso de técnicas de Inteligência Artificial supervisionada, fundamentadas em Processamento de Linguagem Natural (PLN) e aprendizado de máquina, com a aplicação de algoritmos como Regressão Logística para reconhecer e classificar automaticamente os documentos a partir de exemplos previamente rotulados.

A comparação entre as duas metodologias visa avaliar o desempenho, a precisão e a adaptabilidade de cada uma frente a diferentes tipos de documentos e níveis de complexidade textual, destacando tanto a eficiência dos métodos tradicionais quanto o potencial de modelos inteligentes na automação de processos documentais.

2.1 BASE DE DADOS

A base de dados utilizada neste trabalho foi composta a partir de documentos reais, provenientes do ambiente corporativo de uma construtora, aos quais o autor já possuía acesso autorizado. Ressalta-se que nenhuma informação sensível, pessoal ou sigilosa foi ou será exposta neste estudo, sendo todos os arquivos tratados exclusivamente para fins de análise técnica e acadêmica.

Os documentos estavam originalmente armazenados em doze

arquivos PDF, cada um contendo o conjunto completo de documentos referentes a um cliente distinto. Para viabilizar o processamento individual e a classificação automática, os arquivos foram separados manualmente em documentos específicos, de acordo com seu tipo. Essa separação foi realizada de forma segura, utilizando ferramentas de edição de PDF, de modo a preservar a integridade visual e textual dos arquivos originais.

Após a segmentação, os documentos foram organizados em pastas numeradas de “01” a “12”, correspondendo aos doze clientes analisados. Em seguida, todos os arquivos foram reunidos em uma pasta denominada “tests”, localizada na estrutura principal do projeto (TCC3/tests). Nessa etapa, foi aplicada uma padronização de nomenclatura, combinando o tipo de documento e o número do cliente, conforme o formato “tipo_doc” + “nº” (por exemplo: cpf01.pdf, nascimento03.pdf, cnh05.pdf, casamento08.pdf, residencia10.pdf).

A base final resultou em 52 documentos individuais, distribuídos entre cinco categorias documentais: certidões de nascimento (8), certidões de casamento (4), CNHs (14), CPFs (14) e comprovantes de residência (12). Essa diversidade de tipos documentais reflete um cenário realista de análise, contemplando as principais classes de documentos exigidas na formação de uma proposta comercial e representando, portanto, um conjunto adequado para a avaliação dos modelos de classificação.

Para fins de organização e rastreabilidade, foi criado um arquivo de referência denominado labels.csv, contendo a correspondência entre o nome de cada arquivo e sua respectiva classe documental. Esse arquivo foi utilizado como base de apoio para a execução dos testes comparativos de classificação nos dois modelos propostos: modelo 01 (baseado em regras) e modelo 02 (baseado em aprendizado supervisionado).

2.2 EXTRAÇÃO DE TEXTO

A etapa de extração de texto foi responsável por converter o conteúdo dos documentos em formato PDF em dados textuais legíveis e processáveis pelos modelos de classificação. Essa fase foi essencial para uniformizar a leitura dos arquivos, considerando que a base de dados utilizada continha tanto documentos com texto selecionável quanto imagens escaneadas.

O processo foi implementado em um script dedicado denominado extract_text.py, executado de forma independente em cada modelo. Em ambos os casos, o funcionamento é idêntico: o script é chamado pelo

arquivo principal, antecedendo as etapas de tratamento e classificação.

A extração textual adota uma abordagem híbrida, combinando simultaneamente duas técnicas complementares: a leitura direta via biblioteca pdfminer.six, utilizada para interpretar e extrair o texto embutido em PDFs com conteúdo selecionável, e o Reconhecimento Óptico de Caracteres (OCR) com o motor Tesseract, aplicado às regiões de imagem. Essa execução paralela garante que informações presentes em qualquer tipo de representação, textual ou visual, sejam capturadas integralmente, evitando perdas comuns em documentos híbridos, como digitalizações que contêm texto parcialmente reconhecível.

O OCR foi configurado para o idioma português (lang="por") e operou diretamente sobre as imagens rasterizadas de cada página, retornando uma sequência textual unificada. O resultado da execução conjunta das duas técnicas é consolidado em um único bloco de texto, representando a totalidade do conteúdo do documento. Nas execuções manuais, cada modelo gera automaticamente dois arquivos de saída: um contendo o texto bruto extraído e outro com o texto final tratado, obtido após a aplicação do processo de normalização descrito na subseção seguinte.

Essa estratégia integrada de extração permite alcançar maior abrangência e confiabilidade na etapa de leitura, reduzindo a dependência de um único método e assegurando que tanto os textos embutidos quanto as imagens escaneadas sejam igualmente considerados no processo de classificação automática dos documentos.

A Figura 1 apresenta um trecho do código responsável pela extração híbrida de texto, que combina leitura direta via pdfminer.six e OCR com Tesseract, garantindo abrangência tanto em PDFs pesquisáveis quanto em documentos digitalizados.

2.3 TRATAMENTO E NORMALIZAÇÃO DO TEXTO

Após a extração do conteúdo textual dos documentos, os dados resultantes foram submetidos a uma etapa de tratamento e normalização, implementada no script decypher.py. Essa fase teve como objetivo aprimorar a legibilidade e a consistência dos textos extraídos, eliminando ruídos e padronizando o conteúdo antes de sua utilização nos modelos de classificação.

Figura 1 - Código do script extract_text.py

```python
import pytesseract
from pdf2image import convert_from_path
from pdfminer.high_level import extract_text

pytesseract.pytesseract.tesseract_cmd = r"C:\Program Files\Tesseract-OCR\tesseract.exe"

def extract_pdf_text(file_path):
    # Extrai texto de PDF mesclando camada selecionável e OCR das imagens
    native_text = ""
    ocr_text = ""

    # Extrai texto selecionável
    try:
        native_text = extract_text(file_path) or ""
    except Exception:
        native_text = ""

    # Extrai texto das imagens (OCR)
    try:
        images = convert_from_path(file_path, dpi=300)
        for img in images:
            ocr_text += pytesseract.image_to_string(img, lang="por") + "\n"
    except Exception:
        ocr_text = ""

    # Junta ambos os textos, removendo duplicatas simples
    combined = (native_text.strip() + "\n" + ocr_text.strip()).strip()
    lines = combined.splitlines()
    seen = set()
    deduped = []
    for line in lines:
        l = line.strip()
        if l and l not in seen:
            seen.add(l)
            deduped.append(l)
    return "\n".join(deduped)

```

Fonte: Do autor.

O decypher.py atua diretamente sobre o texto bruto retornado pelo extract_text.py, realizando uma série de operações de limpeza que incluem a remoção de quebras de linha desnecessárias, espaçamentos excessivos, caracteres invisíveis ou especiais (como \u200b e símbolos de controle) e normalização de acentuação por meio do módulo unicodedata. Também são aplicadas substituições de caracteres irregulares como hífens longos (“–”, “—”) e aspas diferenciadas por equivalentes padronizados em formato ASCII, garantindo uniformidade no corpus textual.

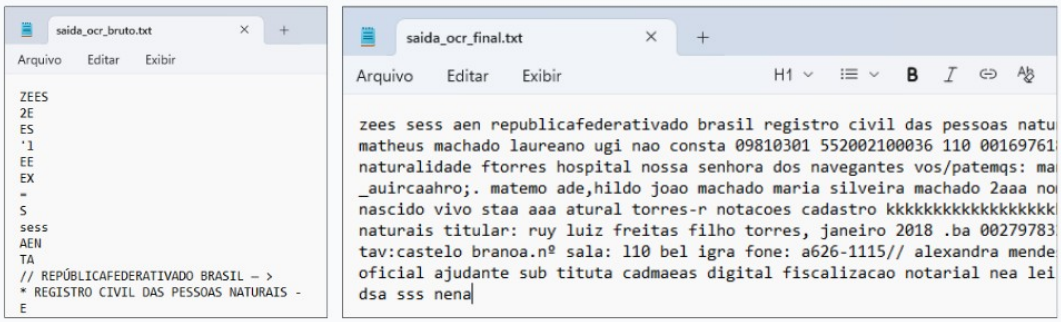
Além dessas correções estruturais, o script executa uma redução e normalização lexical, convertendo todo o conteúdo para letras minúsculas e padronizando variações ortográficas e espaciais. Esse processo é essencial para minimizar divergências semânticas durante a vetorização e classificação, uma vez que garante que termos como “CPF”, “cpf” e “C.P.F.” sejam tratados de forma equivalente pelos algoritmos.

O resultado final dessa etapa é um texto limpo, contínuo e semanticamente consistente, armazenado em um arquivo complementar que representa o texto final tratado. Esse material é então encaminhado para os módulos de classificação dos modelos 01 e 02, servindo como entrada direta para as etapas de inferência e avaliação.

Ao aplicar uma rotina de normalização unificada, o processo contribui para a redução de ruídos gerados pela diversidade de layouts e formatos presentes nos documentos comerciais, garantindo que as diferenças de formatação não interfiram na interpretação do conteúdo. Dessa forma, o decypher.py constitui uma etapa intermediária fundamental entre a extração e a classificação, assegurando que o texto analisado pelos modelos represente de forma precisa e padronizada as informações contidas em cada documento.

Para ilustrar o impacto da normalização e tratamento implementados no script decypher.py, a Figura 2 apresenta um exemplo real extraído da base de testes, comparando o texto bruto retornado pelo OCR com o texto final tratado. Essa visualização evidencia como a remoção de ruídos, padronização de acentuação, unificação de caracteres e redução de variações tipográficas contribuem para gerar um corpus mais consistente para a classificação automática.

Figura 2 - Comparação entre texto bruto e texto tratado/final



Fonte: Do autor.

2.4 CLASSIFICAÇÃO DE DOCUMENTOS

A etapa de classificação de documentos constitui o núcleo funcional deste projeto, sendo responsável por identificar automaticamente o tipo de cada arquivo com base no texto previamente extraído e normalizado. Essa fase representa a transição entre o processamento de dados e a análise inteligente, permitindo avaliar a eficiência de diferentes estratégias de categorização aplicadas a documentos comerciais.

O processo foi projetado de forma a permitir a comparação entre duas abordagens complementares: uma baseada em regras e expressões regulares (Modelo 01) e outra fundamentada em aprendizado supervisionado (Modelo 02). Ambas operam sobre o mesmo conjunto de documentos e compartilham o mesmo fluxo de pré-processamento, diferenciando-se apenas na lógica de decisão empregada para classificar os textos. A seguir, são detalhados os princípios de funcionamento e a metodologia de cada modelo.

2.4.1 Modelo 01 - Abordagem Baseada em Regras

O Modelo 01 foi desenvolvido como uma abordagem determinística, fundamentada em regras explícitas e expressões regulares (regex) para a identificação dos tipos documentais. Após o processamento de extração e tratamento textual, o texto final de cada documento é analisado pelo script principal, que percorre padrões previamente definidos para reconhecer palavras-chave e estruturas características de cada classe.

Entre os exemplos de regras aplicadas, estão a busca por termos como “certidão de nascimento”, “número do CPF”, “carteira nacional de habilitação” e “comprovante de residência”. A presença desses elementos no texto, em diferentes variações e grafias, direciona a classificação do documento para sua respectiva categoria: certidão de nascimento, certidão de casamento, CPF, CNH ou comprovante de residência.

O funcionamento do modelo é orientado por uma sequência de condições lógicas que priorizam a correspondência mais específica encontrada no texto. Por se tratar de uma abordagem baseada em matching direto, o modelo não realiza inferências sobre o contexto, o que pode levar a classificações equivocadas quando o texto apresenta ruídos, trechos ilegíveis ou formatos fora do padrão.

Apesar dessas limitações, o Modelo 01 se destaca pela simplicidade, velocidade de execução e interpretabilidade. Seu comportamento é totalmente previsível e facilmente ajustável, o que o torna uma alternativa viável em ambientes controlados ou em aplicações onde as estruturas documentais seguem padrões rígidos e pouco variáveis.

2.4.2 Modelo 02 - Abordagem com Aprendizado Supervisionado

O Modelo 02 foi desenvolvido com base em técnicas de Inteligência Artificial e Processamento de Linguagem Natural (PLN), visando explorar a capacidade de algoritmos supervisionados de identificar padrões linguísticos e contextuais de forma autônoma.

Nesta abordagem, o modelo é treinado previamente a partir de um conjunto de exemplos rotulados contidos no arquivo dataset.csv, que reúne expressões textuais representativas de cada classe documental. Esse conjunto funciona como base de aprendizado supervisionado, permitindo que o algoritmo associe padrões de palavras e estruturas linguísticas aos respectivos tipos de documentos. O treinamento é realizado pelo script train_baseline.py, responsável por gerar e armazenar o modelo ajustado.

Durante o treinamento, cada exemplo do dataset.csv é transformado em uma representação numérica utilizando a técnica TF-IDF (Term Frequency–Inverse Document Frequency), que quantifica a importância de cada termo em relação ao conjunto de textos. Esses vetores são utilizados como entrada para o algoritmo de Regressão Logística, implementado com a biblioteca scikit-learn, o qual aprende a distinguir as classes com base nas regularidades observadas no corpus.

Após o treinamento, o modelo é utilizado para inferência e classificação dos documentos reais processados pelo sistema. Nessa etapa, o texto extraído e tratado de cada arquivo é convertido em um vetor TF-IDF compatível com o formato aprendido durante o treinamento. Esse vetor é então submetido ao modelo previamente ajustado, que calcula a probabilidade de pertencimento do texto a cada uma das classes conhecidas e retorna como resultado a categoria com maior valor preditivo. Dessa forma, o Modelo 02 atua como um decodificador semântico, capaz de interpretar o conteúdo textual e identificar o tipo de documento de forma automática, com base nos padrões de linguagem assimilados durante o processo de aprendizado supervisionado.

2.4.3 Garantia de Paridade Experimental entre os Modelos

Durante o desenvolvimento e a execução dos testes, foi adotado um cuidado metodológico rigoroso para assegurar a paridade experimental entre as duas abordagens. Ambos os modelos utilizam exatamente as mesmas entradas textuais extraídas e tratadas pelos scripts extract_text.py e decypher.py, bem como o mesmo conjunto de documentos de teste e o mesmo arquivo de referência labels.csv. Dessa forma, qualquer variação observada nos resultados se deve exclusivamente à diferença entre as estratégias de classificação, e não a fatores externos ou de processamento.

Além disso, os dois modelos foram executados sob as mesmas condições computacionais, utilizando o mesmo ambiente virtual, bibliotecas, hardware e sequência de execução controlada pelo script run_tests.py.

Essa estrutura garante que ambos os métodos sejam avaliados de forma equitativa, assegurando uma comparação justa quanto ao desempenho e à eficiência. Assim, a distinção entre os modelos está restrita unicamente à abordagem lógica empregada (regras determinísticas no Modelo 01 e aprendizado supervisionado no Modelo 02), preservando a integridade e a validade dos resultados apresentados na seção seguinte.

2.5 RECURSOS COMPUTACIONAIS

Para o desenvolvimento deste estudo, foram utilizados recursos de hardware e software de uso pessoal do acadêmico, aliados a ferramentas de código aberto amplamente disponíveis na internet. O ambiente de execução foi configurado em um computador desktop equipado com processador AMD Ryzen 5 5600, placa de vídeo Nvidia GeForce RTX 3060 (12 GB), 16 GB de memória RAM e armazenamento SSD de 1,5 TB, operando com o sistema Windows 11. Essa configuração possibilitou o processamento local de todos os experimentos, garantindo desempenho adequado para o treinamento, testes e execução dos modelos de classificação.

O desenvolvimento do projeto foi realizado integralmente em Python 3.12, utilizando o editor Visual Studio Code e um ambiente virtual dedicado (venv) para o gerenciamento das dependências. Entre as principais bibliotecas empregadas destacam-se: pdfminer.six e pytesseract, responsáveis pela extração de texto de arquivos PDF e imagens; pandas e numpy, utilizadas na manipulação e organização dos dados; scikit-learn, aplicada na implementação e avaliação do modelo supervisionado de Regressão Logística; e transformers, empregada na integração de técnicas de Processamento de Linguagem Natural (PLN). O pacote joblib foi utilizado para armazenamento e carregamento de modelos treinados, enquanto o Tesseract OCR, configurado para o idioma português, executou o reconhecimento óptico de caracteres em documentos digitalizados.

Para o controle de versões e rastreabilidade das modificações no código, foi utilizado um repositório no GitHub, garantindo organização, histórico de alterações e segurança na gestão dos arquivos. Todos os scripts, modelos e bases de dados foram processados localmente, sem o uso de serviços externos de computação em nuvem, assegurando independência e reprodutibilidade dos experimentos.

2.6 MÉTRICAS DE AVALIAÇÃO

Para avaliar o desempenho dos modelos propostos, foram empregadas métricas quantitativas amplamente utilizadas em tarefas de classificação supervisionada, permitindo mensurar a precisão e a consistência dos resultados obtidos. A avaliação foi conduzida a partir das predições geradas automaticamente pelo script run_tests.py, que percorre toda a base de testes, executa os modelos 01 e 02 sobre cada documento e compara as classificações obtidas com os rótulos de referência presentes no arquivo labels.csv. Os resultados individuais e consolidados são exibidos no terminal durante a execução e simultaneamente armazenados em arquivo no formato .csv, possibilitando posterior análise comparativa.

As métricas adotadas (acurácia, precisão, *recall* e *F1-score*) foram calculadas por meio das funções disponíveis na biblioteca scikit-learn. A acurácia expressa a proporção de documentos corretamente classificados em relação ao total de exemplos avaliados, sendo uma medida global de desempenho. A precisão indica a proporção de classificações positivas corretas entre todas as predições positivas realizadas pelo modelo, enquanto o *recall* (ou sensibilidade) mede a capacidade do modelo de identificar corretamente todos os exemplos pertencentes a uma determinada classe. Já o *F1-score* representa a média harmônica entre precisão e *recall*, sendo especialmente útil em cenários com classes desbalanceadas, por equilibrar a influência de falsos positivos e falsos negativos.

Essas métricas foram aplicadas uniformemente a ambos os modelos, de forma a assegurar condições equivalentes de comparação. A escolha desse conjunto de indicadores fundamenta-se em sua ampla utilização em estudos de aprendizado de máquina e Processamento de Linguagem Natural (PLN), permitindo uma análise objetiva do desempenho de cada abordagem frente à variabilidade textual dos documentos avaliados. Os valores obtidos são apresentados e discutidos na Seção 3 – Discussão e Resultados, acompanhados de análises comparativas entre os modelos.

3 RESULTADOS E DISCUSSÃO

Esta seção apresenta a análise dos testes realizados sobre os dois modelos propostos para o reconhecimento e classificação automática de documentos. O objetivo central foi comparar o desempenho da abordagem tradicional baseada em regras e expressões regulares (Modelo 01) com a abordagem baseada em Inteligência Artificial supervisionada (Modelo 02), avaliando sua eficácia, precisão e robustez frente a um conjunto

real de documentos utilizados em processos comerciais de construtoras.

3.1 VISÃO GERAL DOS EXPERIMENTOS

Os testes foram executados sobre um conjunto composto por 52 documentos reais, distribuídos entre os seguintes tipos: certidão de nascimento, certidão de casamento, CPF, CNH e comprovante de residência. Esses documentos representam situações comuns no processo de análise de propostas comerciais de uma construtora, onde diferentes formatos e qualidades de imagem são frequentes.

Ambos os modelos foram alimentados pelo mesmo pipeline de processamento, que compreende as etapas de extração de texto via OCR (Tesseract), normalização e limpeza textual, seguidas pela classificação automática em uma das cinco categorias. Os resultados foram avaliados por meio das métricas de acurácia, precisão, *recall* e *F1-score*, amplamente utilizadas em problemas de classificação supervisionada.

3.2 DESEMPENHO DOS MODELOS

Esta subseção apresenta os resultados obtidos nos testes realizados com os dois modelos propostos, analisando seu desempenho quanto à precisão e consistência das classificações. Em seguida, é feita uma comparação entre as abordagens, destacando diferenças relevantes em acurácia, eficiência e adaptabilidade.

3.2.1 Modelo 01 - Abordagem Baseada em Regras

O Modelo 01 utilizou expressões regulares, vocabulários de palavras-chave e condições lógicas para identificar o tipo de documento com base em padrões textuais recorrentes. Trata-se de um método determinístico, sem aprendizado de máquina, cujo desempenho depende diretamente da qualidade do texto extraído e da cobertura das regras definidas.

O desempenho do Modelo 01, baseado em regras determinísticas e expressões regulares, pode ser observado na Tabela 1. A tabela apresenta as métricas de classificação obtidas para cada tipo de documento, evidenciando os pontos fortes e as limitações da abordagem tradicional.

Tabela 1 – Desempenho do Modelo 01

Classe	Precisão	Recall	F1-score
Certidão de Casamento	1,00	0,75	0,86
Certidão de Nascimento	0,73	1,00	0,84
CNH	1,00	0,71	0,83
Comprovante de Residência	1,00	0,58	0,74
CPF	0,74	1,00	0,85
Acurácia geral		**0,81**	

Fonte: Elaborado pelo autor.

Observa-se que o modelo alcançou acurácia de 81%, com melhor desempenho em documentos com padrões textuais bem definidos, como certidões e CPFs. Por outro lado, apresentou falhas em documentos mais heterogêneos, como comprovantes de residência, cujo layout e estrutura textual variam amplamente entre concessionárias. Essas inconsistências indicam que a ausência de estrutura semântica limita a generalização das regras. Ainda assim, trata-se de uma solução leve, interpretável e de fácil manutenção, o que a torna eficiente para cenários controlados e de baixa variabilidade.

3.2.2 Modelo 02 - Abordagem Baseada em IA

O Modelo 02, por sua vez, utilizou técnicas de aprendizado supervisionado com vetorização TF-IDF e algoritmo de Regressão Logística, permitindo que o sistema aprendesse padrões linguísticos a partir dos textos extraídos dos documentos. Essa abordagem busca generalizar melhor diante de ruídos, erros de OCR e variações de formatação.

A Tabela 2 apresenta os resultados do Modelo 02, construído com técnicas supervisionadas de aprendizado de máquina. As métricas demonstram a capacidade do modelo em generalizar padrões textuais e lidar melhor com variações estruturais entre os documentos.

Tabela 2 – Desempenho do Modelo 02

Classe	Precisão	Recall	F1-score
Certidão de Casamento	1,00	0,25	0,40
Certidão de Nascimento	1,00	1,00	1,00
CNH	1,00	0,93	0,96
Comprovante de Residência	0,75	1,00	0,86
CPF	1,00	1,00	1,00
Acurácia geral		**0,92**	

Fonte: Elaborado pelo autor.

O modelo baseado em IA apresentou acurácia geral de 92%, superando o modelo de regras em praticamente todas as métricas. As maiores melhorias ocorreram em CNH e comprovante de residência, documentos com maior variação de estrutura e ruído textual, o que reforça a capacidade do modelo de aprender padrões sem depender de palavras exatas.

Apesar da melhora global, a classe certidão de casamento manteve *recall* reduzido (0,25), possivelmente devido à baixa representatividade dessa categoria no *dataset* (apenas quatro amostras). Esse comportamento indica a necessidade de uma base mais equilibrada para capturar melhor as variações linguísticas específicas de cada tipo de documento.

A comparação consolidada entre os dois modelos avaliados é exibida na Tabela 3. O objetivo é evidenciar diferenças de desempenho e destacar as melhorias proporcionadas pela abordagem baseada em Inteligência Artificial em relação ao método tradicional.

Tabela 3 – Comparativo geral entre os modelos

Métrica / Modelo	**Modelo 01 (Regras)**	**Modelo 02 (IA)**
Acurácia média	0,81	**0,92**
Precisão média	0,89	**0,94**
Recall médio	0,81	**0,92**
F1-score médio	0,82	**0,91**
Tempo médio por documento	**~0,6 s**	~1,9 s

Fonte: Elaborado pelo autor.

De forma geral, o Modelo 02 apresentou ganhos consistentes em acurácia, robustez e sensibilidade, demonstrando melhor adaptação a variações textuais e erros de OCR. Já o Modelo 01, embora menos preciso, mostrou-se competitivo em velocidade e simplicidade, sendo uma alternativa viável em contextos onde o custo computacional e a interpretabilidade são prioridades.

Nos testes práticos, observou-se que o Modelo 02 foi capaz de corrigir confusões frequentes no Modelo 01, como a diferenciação entre CPF e comprovante de residência, mantendo resultados consistentes mesmo com textos ruidosos ou incompletos. No entanto, o tempo de execução foi maior, reflexo do processamento vetorial e da etapa de inferência do classificador.

Em comparação com o estudo desenvolvido por Possamai (2024), observa-se uma convergência metodológica significativa, especialmente na

utilização de técnicas de Processamento de Linguagem Natural e aprendizado supervisionado para a classificação automática de documentos digitais. Esse alinhamento também aparece em trabalhos como o de Silva, Ribeiro e Dezani (2023), que demonstraram a eficácia de métodos de PLN na organização e categorização de documentos jurídicos, reforçando que abordagens supervisionadas tendem a superar regras estáticas em cenários de alta variabilidade textual.

Além disso, os achados de Antonio (2019), ao destacar a influência direta da qualidade do OCR no desempenho de sistemas automatizados, dialogam com os resultados obtidos neste estudo, que igualmente evidenciam a importância da etapa de extração e limpeza textual para reduzir erros de classificação. Somado a isso, pesquisas como a de Lai e Xu Liheng e Liu (2015), que apresentam modelos capazes de capturar dependências contextuais mais profundas, sugerem caminhos para trabalhos futuros, nos quais técnicas baseadas em redes neurais poderiam superar as limitações observadas, especialmente em classes com baixa representatividade. Assim, os resultados aqui obtidos se inserem de forma coerente no conjunto de evidências presentes na literatura, reforçando a viabilidade e o potencial da automação documental mediada por IA.

3.3 DESAFIOS E LIMITAÇÕES

Os resultados reforçam o potencial da Inteligência Artificial na automação documental, mas também evidenciam limitações práticas:

- A **qualidade do OCR** influencia diretamente a precisão das classificações, uma vez que textos truncados ou distorcidos afetam o reconhecimento de padrões;
- O **Modelo 01** demanda manutenção manual das regras, dificultando a escalabilidade;
- O **Modelo 02**, embora mais preciso, requer maior poder computacional e bases mais extensas para treinar de forma robusta.

Essas limitações evidenciam que a eficácia da automação documental depende não apenas da escolha do modelo de IA, mas também da qualidade dos dados e da infraestrutura disponível. Além disso, demonstram o equilíbrio necessário entre precisão, custo operacional e capacidade de generalização dos modelos, fatores que influenciam diretamente a aplicabilidade prática em ambientes corporativos.

3.4 CONSIDERAÇÕES FINAIS DA DISCUSSÃO

A comparação entre as abordagens evidencia que a IA oferece ganhos expressivos em termos de desempenho e flexibilidade, especialmente quando aplicada a documentos com estruturas diversas e sujeitos a ruído. Em contrapartida, o modelo de regras ainda se mostra útil como solução de baixo custo para validações rápidas em cenários previsíveis.

Os resultados obtidos indicam que a integração dessas tecnologias — combinando a eficiência lógica de regras determinísticas com a capacidade adaptativa dos algoritmos de aprendizado — pode representar o caminho ideal para sistemas de automação documental mais completos, confiáveis e escaláveis no setor da construção civil.

4 CONCLUSÃO

O presente trabalho teve como objetivo desenvolver e comparar duas abordagens distintas para o reconhecimento e a classificação automática de documentos comerciais no contexto da construção civil, buscando avaliar o potencial da Inteligência Artificial (IA) em relação a métodos tradicionais baseados em regras e expressões regulares. A partir dos resultados obtidos, foi possível observar avanços significativos em termos de acurácia, flexibilidade e capacidade de generalização quando se empregam técnicas de aprendizado supervisionado.

O **Modelo 01**, fundamentado em regras determinísticas, apresentou desempenho satisfatório em cenários com textos bem estruturados, alcançando acurácia média de 81%. Sua principal vantagem reside na simplicidade de implementação, baixo custo computacional e interpretabilidade, fatores que o tornam adequado para aplicações pontuais e ambientes de baixa variabilidade documental. Contudo, mostrou limitações em documentos heterogêneos, nos quais pequenas alterações no texto ou na formatação comprometem a eficiência das regras manuais.

O **Modelo 02**, baseado em vetorização TF-IDF e Regressão Logística, superou a abordagem anterior em todas as métricas principais, atingindo uma acurácia média de 92% e *F1-score* de 0,91. Seu desempenho superior reflete a capacidade dos algoritmos de aprendizado de máquina em identificar padrões linguísticos mesmo diante de ruído, erros de OCR e variações de layout. Apesar de demandar maior tempo de processamento e poder computacional, demonstrou ser uma alternativa mais robusta e escalável, especialmente para ambientes que lidam com grande volume de documentos e diferentes formatos.

A comparação entre os modelos permitiu concluir que a automação documental por meio de IA representa um avanço concreto em relação a métodos puramente baseados em regras. No entanto, o estudo também evidenciou que a eficácia desses sistemas depende fortemente da qualidade dos dados de entrada, da curadoria das amostras utilizadas para treinamento e do equilíbrio entre precisão e custo operacional. Esses fatores são determinantes para viabilizar a adoção prática das soluções em contextos corporativos.

Em síntese, este trabalho demonstrou que a aplicação de técnicas de Inteligência Artificial pode elevar substancialmente a eficiência e a confiabilidade no processamento documental, reduzindo o tempo de análise e minimizando erros humanos. A experiência obtida evidencia o potencial dessas tecnologias para transformar processos administrativos não só na construção civil, mas também em outras áreas de mercado, tornando-os mais ágeis, precisos e inteligentes.

Como trabalhos futuros, recomenda-se expandir a base de documentos utilizados nos testes, de modo a aumentar a representatividade de cada classe e a diversidade de layouts. Também se sugere a implementação de modelos híbridos que combinem regras heurísticas e técnicas de IA supervisionada, aproveitando as vantagens de ambos os métodos. Além disso, propõe-se a integração do sistema com plataformas empresariais, como ERPs ou sistemas de gestão comercial, possibilitando a automação completa, desde o recebimento até a validação das propostas. Por fim, recomenda-se a avaliação de técnicas mais avançadas de Processamento de Linguagem Natural, como modelos baseados em *Transformers*, com o objetivo de aprimorar ainda mais a interpretação semântica dos documentos.

REFERÊNCIAS

ANTONIO, D. V. **Implementação de protótipo baseado na tecnologia ocr aplicada ao reconhecimento de rótulos para busca em banco de dados**. Repositório Institucional da Unesc, 2019, v. 1, n. 1, p. 1–94.

DAVENPORT, T. H.; RONANKI, R. **Artificial intelligence for the real world**. Harvard Business Review, 2018. Acesso em: 22 maio 2025. Disponível em: <https://hbr.org/2018/01/artificial-intelligence-for-the-real-world>.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 3. ed. Stanford University, 2021.

Acesso em: 06 fev. 2025. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>.

LAI, S.; XU LIHENG E LIU, K. e. Z. J. Recurrent convolutional neural networks for text classification. In: AAAI PRESS. **Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)**. Austin, Texas, USA, 2015. p. 2267–2273.

PAJEú, H. M.; MOURA RHAYZA RODRIGUES E CARVALHO, D. O. d. **Organização e classificação de documentos digitais de arquivos pessoais nas nuvens**. Ciência da Informação em Revista, 2018, Maceió, v. 5, n. 3, p. 58–70.

POSSAMAI, T. S. TCC, **Utilização de Técnicas de Machine Learning para a Organização de Documentos Digitais no Formato PDF**. Criciúma: [s.n.], 2024. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade do Extremo Sul Catarinense, Criciúma, SC.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**. 4. ed. [S.l.]: Pearson, 2021.

SILVA, G. S.; RIBEIRO, L.; DEZANI, H. **Processamento da Linguagem Natural para análise de documentos jurídicos**. 2023. Trabalho de Conclusão de Curso (TCC) - Faculdade de Tecnologia de São José do Rio Preto.

SMITH, R. An overview of the tesseract ocr engine. In: IEEE. **Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)**. Curitiba, Brazil, 2007. p. 629–633.