UTILIZAÇÃO DE TÉCNICAS DE MACHINE LEARNING PARA A ORGANIZAÇÃO DE DOCUMENTOS DIGITAIS NO FORMATO PDF

Thiago Serafina Possamai[1], Marlon de Matos de Oliveira[2]

Resumo: Este trabalho apresenta o desenvolvimento e a avaliação de um sistema automatizado para organização de documentos digitais no formato PDF utilizando técnicas de Machine Learning. O objetivo central é comparar diferentes abordagens de Inteligência Artificial, como Processamento de Linguagem Natural (PLN), Regressão Logística e o modelo BERT, aplicadas à classificação de documentos. Para isso, foram utilizadas duas bases de dados públicas, totalizando 4.901 documentos, divididos em até cinco categorias distintas, submetidas a etapas de pré-processamento, extração de texto e codificação de categorias. Embora existam estudos prévios aplicando IA à organização documental, muitos se concentram em domínios restritos ou em dados estruturados de forma simplificada. Os modelos foram treinados e avaliados com métricas de acurácia, precisão, *recall* e *F1-score*, além de terem sido integrados a uma API REST desenvolvida em Django. Este trabalho se diferencia ao propor uma comparação entre abordagens consolidadas de PLN aplicadas à classificação automatizada de arquivos PDF em larga escala, com aplicação prática em um sistema funcional. Os resultados demonstraram que, em cenários de maior complexidade textual, o modelo BERT atingiu acurácia de até 97% e F1-score acima de 0,95, enquanto a Regressão Logística obteve acurácia média de 92%, com quedas pontuais em algumas classes. Esses resultados destacam a robustez do BERT na classificação de documentos heterogêneos e sugerem que a utilização de técnicas avançadas de IA pode aprimorar significativamente a organização e a gestão de documentos digitais, proporcionando maior eficiência e confiabilidade aos processos.

Palavras-chave: aprendizado de máquina; inteligência artificial; classificação de documentos; processamento de linguagem natural; BERT; regressão logística.

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

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

ABSTRACT: This work presents the development and evaluation of an automated system for organizing digital documents in PDF format using Machine Learning techniques. The main objective is to compare different Artificial Intelligence approaches, such as Natural Language Processing (NLP), Logistic Regression, and the BERT model, applied to document classification. To achieve this, two public datasets were used, totaling 4,901 documents divided into up to five distinct categories, and subjected to preprocessing, text extraction, and category encoding steps. Although previous studies have applied AI to document organization, many are limited to specific domains or structured datasets with simplified content. The models were trained and evaluated using accuracy, precision, recall, and F1-score metrics, and were also integrated into a REST API developed with Django. This study stands out by offering a comparison of well-established NLP approaches applied to large-scale automated classification of PDF documents, with practical implementation in a functional system. The results showed that, in scenarios of greater textual complexity, the BERT model achieved up to 97% accuracy and an F1-score above 0.95, while Logistic Regression obtained an average accuracy of 92%, with occasional drops in some classes. These findings highlight the robustness of BERT in classifying heterogeneous documents and suggest that the use of advanced AI techniques can significantly improve the organization and management of digital documents, providing greater efficiency and reliability to the processes.

Keywords: machine learning; artificial intelligence; document classification; natural language processing; BERT; logistic regression.

1 INTRODUÇÃO

Nas últimas décadas, a sociedade passou por uma transformação significativa na forma como produz, armazena e consome informações. A popularização dos computadores pessoais, dispositivos móveis, sensores inteligentes e da Internet das Coisas (IoT) provocou um aumento exponencial na geração de dados digitais (Ribeiro, 2014; Muniswamaiah; Agerwala; Tappert, 2023). Esse fenômeno impulsionou a digitalização de documentos e processos em diversas áreas, tornando os arquivos digitais, como contratos, exames médicos, notas fiscais e registros jurídicos, elementos centrais no funcionamento de instituições públicas e privadas. Segundo Barreto (2005), a digitalização de documentos, como diários oficiais, ementas de jurisprudência e materiais de ensino, facilita o acesso à informação para um

público mais amplo. Estimativas da plataforma Statista (2021) indicam que o volume de dados digitais no mundo deve ultrapassar 180 zettabytes até 2025, evidenciando a magnitude do desafio enfrentado pelas organizações no que diz respeito à organização e gerenciamento dessas informações.

Neste contexto de transformação digital, surge a necessidade de soluções automatizadas para gerenciar grandes volumes de documentos digitais de forma eficiente, segura e automatizada. A crescente demanda por sistemas que organizem e classifiquem arquivos automaticamente justifica o interesse em explorar tecnologias baseadas em Machine Learning e Processamento de Linguagem Natural (PLN), que possibilitam o tratamento e a análise semântica de conteúdos textuais, promovendo uma gestão documental mais ágil, precisa e escalável. Conforme apontado por Reis e Rodolpho (2020), a desorganização documental pode gerar perdas significativas de produtividade e ocasionar falhas administrativas críticas.

A automação de tarefas como classificação e indexação de documentos permite não apenas otimizar processos internos, mas também reduzir custos operacionais e aumentar a confiabilidade dos sistemas de armazenamento. Ao propor uma análise comparativa entre diferentes abordagens de IA aplicadas à classificação de documentos PDF, este estudo pretende contribuir tanto para o avanço acadêmico na área quanto para a aplicação prática de soluções tecnológicas inovadoras. A escolha do tema deste trabalho se justifica, portanto, pela importância crescente da digitalização de documentos e pela oportunidade de aplicar técnicas avançadas de Inteligência Artificial (IA) na resolução de problemas reais e recorrentes relacionados à organização de dados.

Diante do cenário de crescente digitalização e do aumento exponencial no volume de informações produzidas, armazenadas e consumidas por indivíduos, empresas e instituições, a necessidade de soluções inteligentes para a organização e gestão de documentos digitais tornou-se um desafio central para a sociedade contemporânea. O avanço das tecnologias de Machine Learning e Processamento de Linguagem Natural (PLN) abre novas perspectivas para lidar com esse volume massivo de dados, promovendo maior eficiência, precisão e confiabilidade nos processos documentais.

Neste contexto, este trabalho tem como objetivo geral investigar e comparar diferentes técnicas de Inteligência Artificial, como o Processamento de Linguagem Natural (PLN) e abordagens clássicas de Machine Learning, aplicadas à classificação e organização automática de documen-

tos digitais no formato PDF. Busca-se, com isso, não apenas identificar as abordagens mais eficazes diante de bases de dados com diferentes níveis de complexidade textual, mas também contribuir para a consolidação de práticas inovadoras e escaláveis na gestão documental digital.

Para alcançar tal propósito, o estudo compreende a implementação de métodos de pré-processamento, normalização e extração de características textuais, bem como a análise do desempenho dos modelos em múltiplos cenários, utilizando métricas como acurácia, precisão, *recall* e *F1-score*. Ao propor um olhar comparativo sobre o potencial e as limitações de cada abordagem, espera-se oferecer subsídios tanto para o avanço acadêmico da área quanto para aplicações práticas capazes de transformar rotinas de armazenamento, busca e categorização de documentos digitais.

No contexto da organização automática de documentos digitais, destacam-se as contribuições do Processamento de Linguagem Natural (PLN) e das redes neurais, especialmente as convolucionais. O PLN viabiliza o tratamento de grandes volumes de dados textuais, oferecendo recursos para a extração de informações semânticas relevantes e a categorização automatizada de documentos, por meio de estratégias como *Bag-of-Words*, TF-IDF e modelos de linguagem modernos, a exemplo do BERT (*Bidirectional Encoder Representations from Transformers*). Esses avanços facilitam a organização de documentos, categorização e melhor compreensão do conteúdo de arquivos digitais, tornando o processo de classificação mais eficiente e preciso (Flayeh; Hamodi; Zaki, 2022).

Paralelamente, as redes neurais convolucionais (CNNs) se consolidaram como alternativas robustas para o reconhecimento de padrões em dados estruturados e não estruturados. Sua aplicação em tarefas relacionadas à análise documental abrange desde a identificação de estruturas e tabelas até a detecção de padrões textuais em arquivos PDF, potencializando a organização automática de grandes acervos (Prathyakshini; Shetty, 2024). Além disso, segundo Goodfellow, Bengio e Courville (2016), as CNNs possuem características distintas que as tornam adequadas para processar dados como uma topologia de grade, como imagens e séries temporais. Algumas dessas características incluem: as camadas convolucionais são responsáveis por aplicar filtros sobre imagens para extrair informações específicas; *Pooling* que reduzem a dimensionalidade dos dados, mantendo as características mais relevantes; e as camadas totalmente conectadas que realizam a classificação ou previsão final (Didatica, 2024; InsightLab, 2024). A integração dessas técnicas, portanto, configura-se

como um dos caminhos mais promissores para enfrentar os desafios impostos pela crescente digitalização e pelo volume de documentos a serem processados.

Entre os estudos revisados, destaca-se o trabalho de Antonio (2019), que propõe um protótipo baseado em OCR para o reconhecimento de rótulos de produtos a partir de imagens capturadas por dispositivos móveis. A metodologia adotada consiste na utilização de plataformas de visão computacional para realizar a extração automática das informações textuais presentes nos rótulos, permitindo a consulta a um banco de dados para facilitar a identificação e avaliação dos produtos. Os testes realizados demonstraram que a abordagem com OCR tradicional foi mais viável que o uso de classificadores baseados em redes neurais convolucionais (CNN), especialmente devido à limitação de dados para o treinamento das CNNs, obtendo bons resultados na identificação dos produtos, mas com restrição ao contexto de imagens e rótulos.

Outro trabalho relevante é o de Pajeú, Moura e Carvalho (2018), que apresenta a criação de um quadro de arranjo arquivístico voltado para documentos digitais pessoais. A proposta fundamenta-se em princípios teóricos da Arquivologia e é complementada pelo uso de serviços em nuvem para o armazenamento dos documentos. O estudo realizou etapas como identificação das tipologias documentais, análise do contexto de produção e aplicação de métodos arquivísticos para classificação e migração dos arquivos para suportes mais seguros. Como resultado, houve uma melhoria significativa na organização, na busca e na recuperação dos documentos, demonstrando a importância da aplicação de métodos arquivísticos mesmo em acervos pessoais.

No campo do aprendizado de máquina, destaca-se o trabalho de Lai et al. (2015), que desenvolveu uma rede neural recorrente convolucional (RCNN) para a classificação de textos. Diferentemente das abordagens tradicionais, o modelo proposto dispensa a necessidade de *features* projetadas manualmente, combinando estruturas recorrentes para captura de contexto e camadas convolucionais para identificar padrões-chave nos textos. Os experimentos, realizados em diferentes bases de dados, evidenciaram que a RCNN foi capaz de capturar informações contextuais relevantes, demonstrando desempenho superior em comparação com outros métodos, especialmente em tarefas de classificação em nível de documento.

O trabalho de Pereira (2021), focou no desenvolvimento de um agente conversacional (chatbot) para atendimento de primeiro nível, espe-

cificamente para responder a dúvidas relacionadas ao programa UNIEDU. Para isso, a pesquisa fez uso da plataforma IBM Watson Assistant, que emprega técnicas de aprendizado de máquina e aprendizado profundo para classificar intenções do usuário, realizar compreensão de leitura e extração de perguntas frequentes (FAQ). O agente conversacional foi treinado com uma base de conhecimento que incluía 46 intenções de atos, com uma média de 10 exemplos de sentenças para cada uma.

Por fim, o estudo de Silva, Ribeiro e Dezani (2023) analisou a aplicação do Processamento de Linguagem Natural (PLN) aliado a técnicas de Machine Learning para a classificação automática de documentos jurídicos nas áreas cível e criminal. Utilizando um corpus composto por ementas processuais, o estudo empregou a linguagem Python e bibliotecas especializadas para o pré-processamento dos textos e para a construção de modelos de classificação supervisionada. Os resultados alcançados demonstraram elevada acurácia na categorização dos documentos, evidenciando a eficácia das técnicas de IA para reconhecer e classificar automaticamente diferentes tipos de documentos jurídicos.

Esses estudos contribuem para delinear o estado da arte na área, evidenciando os avanços e limitações das abordagens atuais, além de inspirarem a estrutura metodológica adotada neste trabalho.

A estrutura deste trabalho está dividida em quatro seções. A primeira apresenta a introdução, com a contextualização do tema, os objetivos e a justificativa. A segunda, Materiais e Métodos, aborda os desafios e lacunas da área, descreve o ambiente de desenvolvimento, as ferramentas utilizadas, a metodologia adotada e as etapas da implementação. A terceira seção, Discussão e Resultados, traz os dados obtidos por meio dos experimentos, com análises comparativas e críticas. Por fim, a seção de Conclusão retoma os objetivos, destaca os principais resultados e sugere direções para pesquisas futuras.

2 MATERIAIS E MÉTODOS

A aplicação foi desenvolvida em Python 3.12, utilizando bibliotecas amplamente adotadas nas áreas de Processamento de Linguagem Natural (PLN) e aprendizado de máquina, como *transformers, datasets, torch, scikit-learn, pandas* e *pdfminer.six*. Também foi utilizada a GPU para o treinamento utilizando o CUDA; a placa de vídeo utilizada foi a Nvidia GeForce RTX 3050. A estrutura do projeto foi organizada em dois módulos principais: um dedicado à lógica de classificação e outro à interface via API. O

módulo de classificação foi isolado em um pacote denominado core, contendo os componentes de pré-processamento, treinamento do modelo e inferência. A API, por sua vez, foi desenvolvida com o framework Django e o pacote Django REST Framework, localizada na pasta api, e interage com o módulo de IA de forma desacoplada, promovendo modularidade e manutenção simplificada.

2.1 BASE DE DADOS

Para o treinamento e avaliação do modelo de classificação, foram utilizadas duas bases de dados públicas disponíveis na plataforma Kaggle. A primeira, intitulada *Company Documents Dataset* (Cherguelaine; Boubekri, 2024), é composta por 2.676 documentos corporativos em formato PDF, divididos em quatro categorias: faturas (830), pedidos de envio (809), pedidos de compra (830) e relatórios de inventário (207). A segunda base de dados, intitulada *Text Document Classification Dataset* (Sunil, 2023), é composta por 2.225 documentos de maior complexidade, distribuídos em cinco categorias: política (417 documentos), esporte (511), tecnologia (401), entretenimento (386) e negócios (510). Embora os documentos originais estejam em formato PDF, ambas as bases oferecem versões organizadas em arquivos CSV, nos quais cada linha representa um exemplo textual (campo *text*) e sua respectiva categoria (campo *label*). Esses arquivos CSV foram utilizados exclusivamente no treinamento dos modelos, uma vez que facilitam o manuseio dos dados ao estruturarem o texto e o rótulo em colunas distintas. Ressalta-se que, na aplicação final, os documentos classificados são arquivos PDF, cujo conteúdo textual é extraído previamente para que o modelo possa realizar a predição. A segunda base foi escolhida por apresentar maior diversidade temática e nuances textuais, o que contribui para uma avaliação mais robusta dos modelos.

O arquivo *dataset.csv*, contido no repositório do projeto, centraliza os textos e seus respectivos rótulos, servindo como fonte primária para as etapas subsequentes.

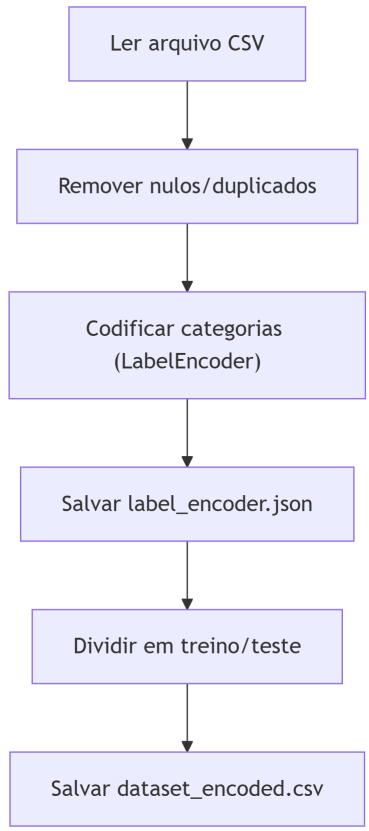
2.2 PRÉ-PROCESSAMENTO DOS DADOS

Os dados foram submetidos ao pré-processamento para codificação das categorias e preparação dos arquivos para o treinamento. Inicialmente, os dados foram lidos a partir de um arquivo .csv contendo os campos *text* (com o conteúdo textual) e o rótulo (com a categoria associada). As entradas foram verificadas para remoção de registros nulos ou

duplicados, assegurando a consistência do conjunto de dados.

Conforme apresentado na Figura 1, as categorias textuais foram transformadas em valores numéricos utilizando a classe *LabelEncoder* da biblioteca *scikit-learn*. Os mapeamentos entre as categorias originais e os valores numéricos atribuídos foram salvos no arquivo *label_encoder.json*, garantindo que o mesmo esquema de codificação pudesse ser utilizado posteriormente durante o processo de inferência.

Figura 1 - Fluxo de pré-processamento



Fonte: Elaborado pelo autor.

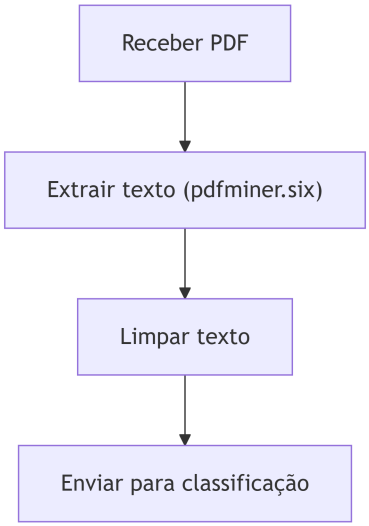
Após a codificação, os dados foram divididos em conjuntos de treino e teste utilizando a função *train_test_split* da biblioteca *scikit-learn*, com proporção de 70% dos dados destinados ao treinamento e 30% à avaliação. Essa divisão permitiu avaliar o desempenho do modelo de forma mais confiável. O conjunto final de dados codificados foi salvo em um novo arquivo denominado *dataset_encoded.csv*, o qual foi utilizado como base para o treinamento do modelo de classificação.

2.3 EXTRAÇÃO DE TEXTO DE DOCUMENTOS PDF

A extração de texto dos arquivos PDF foi realizada com a biblioteca *pdfminer.six*, capaz de converter documentos PDF em texto plano por meio da análise da estrutura interna dos arquivos, extraindo sequências de caracteres interpretadas diretamente a partir dos objetos do tipo texto. Esse processo permite obter o conteúdo dos documentos de forma programática, sem depender de interfaces visuais ou ferramentas de OCR.

Conforme a Figura 2, foi desenvolvida uma função específica para essa tarefa, a qual recebe o caminho de um arquivo PDF, processa o conteúdo por meio do *pdfminer.six* e retorna o texto extraído como uma única string. Essa função foi utilizada tanto na fase de construção do conjunto de dados quanto na operação da API, permitindo que arquivos PDF enviados pelo usuário sejam automaticamente convertidos em texto para posterior classificação. O texto extraído também passa por uma limpeza básica, removendo quebras de linha excessivas e espaços desnecessários, a fim de padronizar a entrada do modelo.

Figura 2 - Fluxo de extração textual do PDF



Fonte: Elaborado pelo autor.

2.4 TREINAMENTO DOS MODELOS DE CLASSIFICAÇÃO

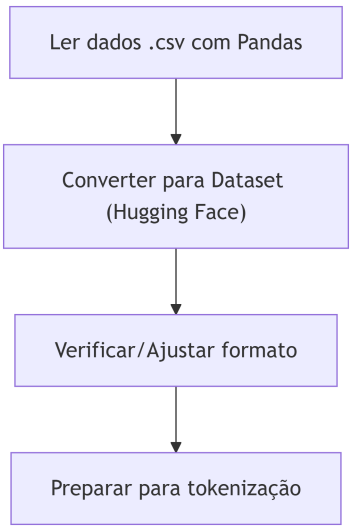
O primeiro modelo de classificação foi implementado utilizando Regressão Logística. Antes do treinamento do modelo, os textos pré-processados foram convertidos em uma representação numérica utilizando a técnica TF-IDF (*Term Frequency-Inverse Document Frequency*). Para esta finalidade, utilizou-se a classe *TfidfVectorizer* da biblioteca *Scikit-learn*. Fo-

ram utilizados os parâmetros padrão do *TfidfVectorizer*. O modelo foi instanciado utilizando a classe *LogisticRegression* da *Scikit-learn*. Os parâmetros utilizados foram os padrões, porém aqui modificamos o parâmetro de iterações máximas (*max_iter*), definido como 1000.

O segundo modelo de classificação textual foi implementado com base no modelo pré-treinado BERT (bert-base-uncased), disponibilizado pela biblioteca *Hugging Face Transformers*. Para adaptar o modelo à tarefa de classificação com múltiplas categorias, foi utilizada a classe *BertForSequenceClassification*, com o parâmetro *num_labels* definido conforme o número de categorias presentes na base de dados. Como os pesos da camada de classificação estão presentes no modelo pré-treinado, foram inicializados automaticamente.

No desenvolvimento, os dados de treinamento e avaliação foram carregados utilizando a biblioteca Pandas a partir de arquivos no formato .csv. Em seguida, esses dados foram convertidos para o formato *Dataset*, da biblioteca *datasets*, o qual é mais eficiente para integração com o *framework* de treinamento da *Hugging Face*, conforme ilustrado na Figura 3.

Figura 3 - Leitura e conversão em Dataset

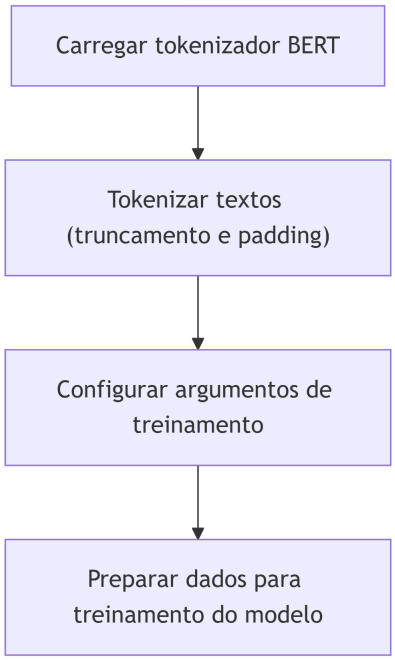


Fonte: Elaborado pelo autor.

Após a conversão, conforme a Figura 4, foi carregado o tokenizador pré-treinado correspondente ao modelo BERT. A função *tokenize*, definida no projeto, aplica truncamento e preenchimento (*padding*) nas sequências de entrada, garantindo comprimento padronizado de 512 *tokens* para processamento em lote pelo modelo. Na sequência, o modelo *BertForSequenceClassification* foi instanciado com o número de classes configurado

de acordo com os rótulos únicos presentes no conjunto de dados. Também foram definidos os argumentos de treinamento, como número de épocas em 3, *batch size* de 4 e estratégias de avaliação usando as épocas, a utilização do fp16 para otimizar o desempenho e a estratégia de salvar o melhor modelo ao final do treinamento.

Figura 4 - Tokenizador BERT e argumentos de treinamento

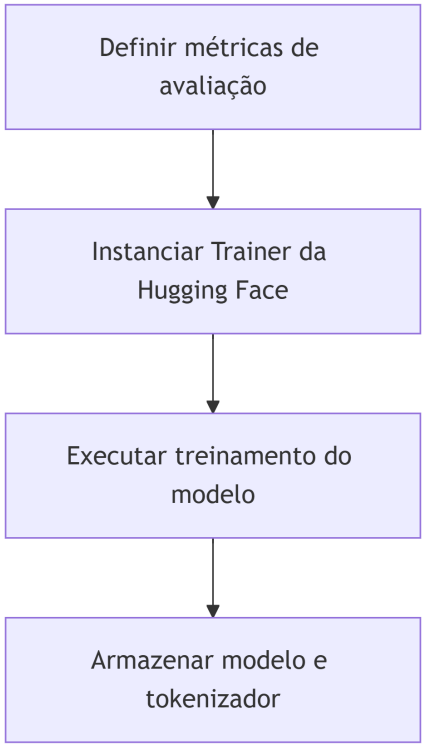


Fonte: Elaborado pelo autor.

Para o treinamento, foi utilizado o componente *Trainer*, que permite abstrair diversas etapas como a alimentação do modelo, cálculo das métricas e salvamento dos *checkpoints*. Ao final do treinamento, tanto o modelo ajustado quanto o tokenizador utilizado foram armazenados no diretório models/fine_tuned_bert, conforme ilustrado na Figura 5.

Para comparar o desempenho dos modelos, foram utilizadas as métricas de acurácia, precisão, *recall* (sensibilidade) e F1-score e implementadas em ambos os modelos com base nas funções da biblioteca *scikit-learn*. A acurácia mede a proporção de classificações corretas em relação ao total de amostras avaliadas, sendo calculada pelo número de acertos dividido pelo número total de exemplos. A precisão indica a proporção de exemplos classificados como positivos que realmente pertencem à classe positiva. O *recall*, ou sensibilidade, mede a capacidade do modelo de identificar todos os exemplos de uma determinada classe, sendo calculado como a razão entre os verdadeiros positivos e o total de exemplos realmente po-

Figura 5 - Métricas, Treinamento e armazenamento do modelo



Fonte: Elaborado pelo autor.

sitivos. Já o F1-score corresponde à média harmônica entre precisão e *recall*, sendo especialmente útil quando se busca um equilíbrio entre essas duas métricas, principalmente em situações em que as classes estão desbalanceadas.

Essas métricas foram utilizadas para avaliar e comparar o desempenho de diferentes modelos, visto que o objetivo do trabalho é identificar qual abordagem apresenta os melhores resultados na classificação de documentos digitais.

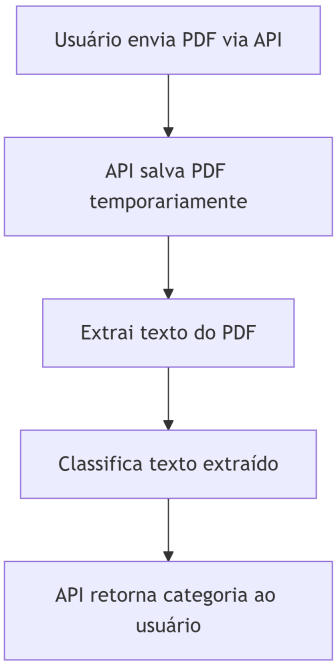
2.5 INTEGRAÇÃO COM A API EM DJANGO

A aplicação foi integrada com uma API REST desenvolvida com Django e Django REST Framework. Um endpoint do tipo POST foi implementado no caminho /api/classifier/classify/ permitindo o envio de arquivos PDF para classificação. Conforme a Figura 6, o fluxo da requisição inclui: o salvamento temporário do PDF, a extração do texto via *pdfminer*, a classificação do texto com o modelo BERT pré-treinado e o retorno da categoria prevista em formato JSON.

A *view* chamada *DocumentClassifierView* foi desenvolvida para lidar com essas solicitações POST na API. Seu funcionamento envolve, primeiramente, receber um arquivo, que é então encaminhado para uma

classe responsável por extrair o texto contido nesse arquivo. Posteriormente, esse texto extraído é processado por um classificador, que avalia o conteúdo e gera uma resposta classificando-o em uma categoria específica.

Figura 6 - Fluxo da API em Django



Fonte: Elaborado pelo autor.

3 DISCUSSÃO E RESULTADOS

Nesta seção, são apresentados e analisados os resultados obtidos pelos dois modelos de Machine Learning propostos para a classificação de documentos PDF: Regressão Logística (RL) com TF-IDF e a rede neural BERT. O objetivo principal foi comparar o desempenho de ambas as abordagens em cenários distintos: um utilizando um conjunto de dados mais homogêneo e limpo, e outro incorporando documentos com maior complexidade e nuances textuais, refletindo desafios mais próximos de aplicações reais.

3.1 DESEMPENHO NAS BASE DE DADOS

Na avaliação com a base de dados de menor complexidade (o Company Documents Dataset), que exibe características textuais mais diretas e menor variabilidade semântica, ambos os modelos atingiram 100% nas métricas de classificação dos documentos do conjunto de teste. Um

desempenho perfeito como este, embora indique alta eficácia em um ambiente controlado, levanta a discussão sobre o potencial de *overfitting*, onde o modelo pode ter se ajustado excessivamente às características dos dados de treinamento, especialmente quando as categorias são muito distintas e a complexidade textual é baixa. Contudo, isso também sugere que, para tarefas de classificação com documentos de estrutura textual simples e categorias bem definidas, modelos como a Regressão Logística podem ser altamente eficientes e oferecer uma solução com menor custo computacional e de treinamento. A performance máxima de ambos os modelos neste cenário de menor ruído e ambiguidade demonstra a clareza dos padrões identificados a partir das representações textuais empregadas.

A distinção no desempenho dos modelos se tornou mais clara ao introduzir um conjunto de dados mais complexo, caracterizado por uma maior diversidade de vocabulário, estruturas variadas e nuances textuais que exigem uma compreensão mais profunda da linguagem. As Tabelas 1 e 2 demonstram os resultados neste cenário. Observa-se que o modelo BERT demonstrou uma vantagem que pode ser atribuída à capacidade do modelo de entender as relações textuais e as nuances semânticas de forma mais eficaz do que abordagens como a Regressão Logística, que tipicamente opera sobre representações de texto (como *bag-of-words* ou TF-IDF) que podem não capturar totalmente essas complexidades contextuais.

Tabela 1 - Resultados do Modelo de RL

Classe	Precisão	Recall	F1-Score	Acurácia
Política	0.98	0.76	0.86	0.92
Esporte	0.93	0.99	0.96	0.92
Tecnologia	0.98	0.91	0.94	0.92
Entretenimento	0.96	0.92	0.94	0.92
Negócios	0.80	0.97	0.88	0.92

Fonte: Elaborado pelo autor.

Tabela 2 - Resultados do Modelo BERT

Classe	Precisão	Recall	F1-Score	Acurácia
Política	0.96	0.94	0.95	0.97
Esporte	1.00	0.97	0.98	0.97
Tecnologia	1.00	0.98	0.99	0.97
Entretenimento	0.93	1.00	0.96	0.97
Negócios	0.97	0.97	0.97	0.97

Fonte: Elaborado pelo autor.

Apesar dos resultados promissores, cabe ressaltar uma importante limitação deste estudo: o tamanho reduzido das bases de dados empregadas para o treinamento e a avaliação dos modelos. Modelos baseados em BERT, por exemplo, são reconhecidos por sua elevada capacidade de representação e, consequentemente, pela necessidade de grandes volumes de dados para expressar todo o seu potencial. Com um conjunto de dados limitado, o BERT pode sofrer tanto de *underfitting* — por não dispor de exemplos suficientes para aprender padrões complexos — quanto de *overfitting*, caso memorize padrões específicos do conjunto de treinamento sem conseguir generalizar para novos documentos. Esse cenário dificulta a observação de diferenças expressivas no desempenho entre BERT e modelos mais simples, como a Regressão Logística, principalmente quando a tarefa de classificação envolve categorias bem delimitadas ou poucos exemplos ambíguos.

Dessa forma, os resultados obtidos devem ser interpretados tendo em vista essa restrição. A ausência de uma grande diferenciação entre os modelos pode estar diretamente relacionada ao tamanho das amostras, limitando a capacidade do BERT de superar significativamente métodos mais tradicionais em contextos de baixa complexidade ou com pouca variabilidade textual. Estudos futuros com bases de dados mais extensas e diversificadas tendem a explorar melhor as vantagens do BERT, permitindo avaliar de forma mais robusta seu desempenho frente a diferentes desafios de classificação documental.

Em comparação com o estudo de Antonio (2019), que se concentrou na implementação de tecnologia OCR aliada a redes neurais convolucionais (CNNs) para o reconhecimento de rótulos em imagens, este trabalho adota uma abordagem distinta ao empregar técnicas de Processamento de Linguagem Natural (PLN) para a categorização automática de documentos digitais no formato PDF. Enquanto o foco de Antonio (2019) está no reconhecimento visual e recuperação de informações em rótulos, este estudo contribui para a organização documental baseada no conteúdo textual extraído dos arquivos.

Os estudos de Pajeú, Moura e Carvalho (2018) alinham-se à perspectiva arquivística clássica, com foco na classificação e acessibilidade de documentos pessoais por meio da aplicação de princípios teóricos da Arquivologia. Embora ambos os trabalhos tratem da organização de documentos digitais, diferem quanto aos objetivos e às metodologias: enquanto Pajeú, Moura e Carvalho (2018) priorizam a estruturação arquivística e a

organização manual, o presente estudo propõe uma solução automatizada baseada em Inteligência Artificial, avaliando seu desempenho com métricas computacionais.

O trabalho de Lai et al. (2015) propôs uma arquitetura baseada em RCNNs (Redes Neurais Convolucionais Recorrentes) voltada para classificação de textos, atingindo desempenho superior às abordagens anteriores em três das quatro bases utilizadas, com acurácias e F1-scores variando entre 47% e 96%. Embora ambos os trabalhos tratem de classificação textual, Lai et al. (2015) foca na otimização arquitetural de modelos, enquanto o presente trabalho realiza uma comparação entre abordagens consolidadas, como Regressão Logística e BERT, aplicadas à classificação documental em larga escala.

Os resultados obtidos no estudo de Pereira (2021) indicaram que o modelo final do agente conversacional alcançou uma precisão de 80% de acertos nas mensagens trocadas, demonstrando um desempenho satisfatório em relação à sua base de dados. Durante o estudo, o desempenho foi de 79,7% de acerto, e, após refinamento, atingiu 80,1% de precisão. Dessa forma, enquanto o trabalho de Pereira (2021) exemplifica o sucesso do PLN na otimização do atendimento ao usuário por meio de diálogo automatizado, este estudo explora o potencial do PLN e da IA na organização e gestão eficientes de grandes volumes de documentos digitais. Ambos os estudos destacam a habilidade dessas tecnologias de IA em resolver problemas complexos e repetitivos em diversos campos de aplicação.

O estudo de Silva, Ribeiro e Dezani (2023), por sua vez, apresenta a maior similaridade com este projeto, ao aplicar técnicas de PLN e aprendizado supervisionado na classificação de documentos jurídicos. No entanto, sua base de dados era mais restrita (228 ementas jurídicas), e o modelo empregado não foi explicitamente nomeado. Ainda assim, obteve resultados expressivos, com acurácia de 94% antes e 100% após o pré-processamento. Em contrapartida, este trabalho utiliza duas bases de dados com maior diversidade temática e volume (totalizando 4.901 documentos) e compara diretamente dois modelos de PLN. Os resultados obtidos indicam acurácia de 100% na base mais simples e 97% na base mais complexa com o modelo BERT, evidenciando a robustez da abordagem adotada.

4 CONCLUSÃO

Este trabalho objetivou desenvolver e avaliar um sistema de classificação automática de documentos PDF com Machine Learning. A metodologia envolveu o estudo de técnicas de IA, o pré-processamento dos arquivos para extração de texto e a comparação de desempenho dos modelos de Regressão Logística e BERT com base em métricas de acurácia e precisão, permitindo validar a eficácia da abordagem para a organização documental automatizada.

O modelo BERT apresentou desempenho superior e mais consistente em bases de dados complexas, superando a Regressão Logística. O BERT atingiu acurácia de 97% e F1-score superior a 0,95, enquanto a Regressão Logística obteve 92% de acurácia, com falhas em classes específicas (Política e Negócios), sugerindo que a aplicação de modelos avançados de Processamento de Linguagem Natural pode otimizar significativamente os processos de gestão documental. Contudo, a diversidade dos PDFs e o tamanho reduzido das bases de dados representam um desafio para a generalização e robustez dos modelos.

Como trabalhos futuros, propõe-se o uso de fine-tuning incremental, permitindo que os modelos se adaptem dinamicamente a novos conjuntos de dados e categorias documentais sem a necessidade de retreinamento completo. Recomenda-se também a curadoria de bases de dados maiores e mais diversas para aprimorar a generalização e a robustez dos classificadores em ambientes reais.

REFERÊNCIAS

ANTONIO, D. V. **Implementação de protótipo baseado na tecnologia ocr aplicada ao reconhecimento de rótulos para busca em banco de dados**. 2019, UNESC.

BARRETO, A. M. **Informação e conhecimento na era digital**. Transinformação, 2005, v. 17, n. 2, p. 1–12.

CHERGUELAINE, A.; BOUBEKRI, F. **Company Documents Dataset**. Kaggle, 2024. Acesso em: 07 maio 2025. Disponível em: <https://www.kaggle.com/datasets/ayoubcherguelaine/company-documents-dataset?resource=download>.

DIDATICA, T. **Introdução a Redes Neurais Convolucionais**. 2024. Acesso em: 13 out. 2024. Disponível em: <https://didatica.tech/introducao-a-redes-neurais-convolucionais/>.

FLAYEH, A. K.; HAMODI, Y. I.; ZAKI, N. D. **Text analysis based on natural language processing (nlp)**. 2022, Institute of Electrical and Electronics Engineers Inc., p. 774–778.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016. Acesso em: 13 out. 2024. Disponível em: <http://www.deeplearningbook.org>.

INSIGHTLAB, U. **Aprenda a Criar e Treinar uma Rede Neural Convolucional (CNN)**. 2024. Acesso em: 13 out. 2024. Disponível em: <https://www.insightlab.ufc.br/aprenda-a-criar-e-treinar-uma-rede-neural-convolucional-cnn/>.

LAI, S. et al. **Recurrent convolutional neural networks for text classification**. 2015, AAAI Press, p. 2267–2273.

MUNISWAMAIAH, M.; AGERWALA, T.; TAPPERT, C. C. **Big data and data visualization challenges**. 2023, p. 6227–6229.

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

PEREIRA, J. F. **Processamento de linguagem natural aplicada na construção de uma agente conversacional por meio do ibm watson assistant**. 2021, UNESC.

PRATHYAKSHINI; SHETTY, J. **Deeptext: Pioneering the future of text classification with innovative deep learning techniques**. 2024, Institute of Electrical and Electronics Engineers Inc., p. 911–917.

REIS, J. V. dos; RODOLPHO, D. **A saúde e segurança ocupacional na gestão de documentos empresariais**. Revista Interface Tecnológica, 2020, v. 17, n. 2, p. 838–848.

RIBEIRO, C. J. S. **Big data: os novos desafios para o profissional da informação**. Informação & Tecnologia, 2014, v. 1, n. 1, p. 96–105.

SILVA, G. S.; RIBEIRO, L.; DEZANI, H. **Processamento da linguagem natural para análise de documentos jurídicos**. 2023, Faculdade de Tecnologia de São José do Rio Preto.

STATISTA. **Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025**. 2021. Acesso em: 05 jun. 2024. Disponível em: <https://www.statista.com/statistics/871513/worldwide-data-created/>.

SUNIL, T. **Text Document Classification Dataset**. Kaggle, 2023. Acesso em: 10 maio 2025. Disponível em: <https://www.kaggle.com/datasets/sunilthite/text-document-classification-dataset>.