UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ROGERIO RODRIGUES CARVALHO

Classificação de documentos da administração pública utilizando inteligência artificial

Goiânia
2024

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Rogerio Rodrigues Carvalho

3. Título do trabalho

Classificação de documentos da administração pública utilizando inteligência artificial

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [**X**] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Ronaldo Martins Da Costa**, **Professor do Magistério Superior**, em 14/06/2024, às 14:40, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Rogerio Rodrigues Carvalho**, **Discente**, em 14/06/2024, às 15:10, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4588665** e o código CRC **F73438D9**.

Referência: Processo nº 23070.016924/2024-85 SEI nº 4588665

ROGERIO RODRIGUES CARVALHO

Classificação de documentos da administração pública utilizando inteligência artificial

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Informática da Universidade Federal de Goiás, como requisito para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação.

Linha de pesquisa: Sistemas Inteligentes e Aplicações.

Orientador: Prof. Dr. Ronaldo Martins da Costa

Goiânia
2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Rodrigues Carvalho, Rogério
Classificação de documentos da administração pública utilizando inteligência artificial [manuscrito] / Rogério Rodrigues Carvalho. - 2024.
62 f.

Orientador: Prof. Dr. Ronaldo Martins da Costa .
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2024.
Bibliografia. Apêndice.
Inclui algoritmos, lista de figuras, lista de tabelas.

1. Classificação de documentos. 2. dados sensíveis. 3. organização pública. 4. inteligência artificial. I. , Ronaldo Martins da Costa, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **13/2024** da sessão de Defesa de Dissertação de **Rogerio Rodrigues Carvalho**, que confere o título de Mestre em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos trinta dias do mês de abril de dois mil e vinte e quatro, a partir das dez horas, via sistema de webconferência da RNP, realizou-se a sessão pública de Defesa de Dissertação intitulada **"**Classificação de documentos da administração pública utilizando inteligência artificial**"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor Ronaldo Martins da Costa (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Rodrigo Gonçalves de Souza (TRF1), membro titular externo; e Professora Doutora Nádia Félix Felipe da Silva (INF/UFG), membra titular interna. A realização da banca ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Ronaldo Martins da Costa, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos trinta dias do mês de abril de dois mil e vinte e quatro.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Nadia Felix Felipe Da Silva**, **Professor do Magistério Superior**, em 30/04/2024, às 11:54, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Ronaldo Martins Da Costa**, **Professor do Magistério Superior**, em 30/04/2024, às 11:54, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Rodrigo Goncalves De Souza**, **Usuário Externo**, em 30/04/2024, às 11:57, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Rogerio Rodrigues Carvalho**, **Discente**, em 30/04/2024, às 12:04, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4511480** e o código CRC **7E74E444**.

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador(a).

Rogerio Rodrigues Carvalho

Bacharel em Sistemas de Informação pelo Instituto Federal de Educação Ciência e Tecnologia de Goiás (IFG) tendo desenvolvido como TCC um sistema de predição educacional, capaz de prever a reprovação em disciplinas e a evasão do curso pelo aluno, além de identificar características comuns dos alunos classificados como reprovados ou evadidos. Especialista em Design de Sistemas e Soluções de Business Intelligence pelo Instituto de Informática da Universidade Federal de Goiás tendo desenvolvido como TCC um sistema de predição de desempenho das ações da bolsa de valores do Brasil. Atualmente é servidor público federal da Universidade Federal de Goiás, e tem como objetivo profissional oferecer soluções orientadas a dados para subsidiar gestores e partes interessadas em suas análises e tomadas de decisões.

Resumo

Carvalho, Rogerio R.. **Classificação de documentos da administração pública utilizando inteligência artificial**. Goiânia, 2024. 63p. Dissertação de Mestrado. Programa de Pós-Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

As organizações públicas enfrentam dificuldades para realizar a devida classificação e promover a transparência dos inúmeros documentos produzidos durante a execução de suas atividades. A classificação correta dos documentos é fundamental para prevenir o acesso público a informações sensíveis e proteger indivíduos e organizações contra o uso malicioso. Este trabalho propõe dois métodos para realizar a tarefa de classificação de documentos sensíveis, utilizando técnicas de inteligência artificial do estado da arte e melhores práticas encontradas na literatura: um método convencional, que utiliza técnicas de inteligência artificial e expressões regulares analisando o conteúdo textual dos documentos, e um método alternativo, que emprega a técnica de CBIR para a classificação de documentos quando a extração de texto não é viável, baseando-se nas imagens dos documentos. Utilizando dados reais do Sistema Eletrônico de Informações (SEI) da Universidade Federal de Goiás (UFG), os resultados alcançados demonstraram que a aplicação de expressões regulares como uma verificação preliminar pode melhorar a eficiência computacional do processo de classificação, apesar de apresentar um aumento modesto na precisão da classificação. O método convencional mostrou-se eficaz na classificação de documentos, com o modelo BERT destacando-se por seu desempenho com taxa de acerto de 94%. O método alternativo, por sua vez, ofereceu uma solução viável para cenários desafiadores, apresentando resultados promissores com taxa de acerto de 87% na classificação de documentos públicos.

Palavras–chave

Classificação de documentos, dados sensíveis, organização pública, inteligência artificial

Abstract

Carvalho, Rogerio R.. **Classification of public administration documents using artificial intelligence**. Goiânia, 2024. 63p. MSc. Dissertation. Programa de Pós-Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

Public organizations face difficulties in classifying and promoting transparency of the numerous documents produced during the execution of their activities. Correct classification of documents is critical to prevent public access to sensitive information and protect individuals and organizations from malicious use. This work proposes two approachs to perform the task of classifying sensitive documents, using state-of-the-art artificial intelligence techniques and best practices found in the literature: a conventional method, which uses artificial intelligence techniques and regular expressions to analyze the textual content of documents, and an alternative method, which employs the CBIR technique to classify documents when text extraction is not viable. Using real data from the Electronic Information System (SEI) of the Federal University of Goiás (UFG), the results achieved demonstrated that the application of regular expressions as a preliminary check can improve the computational efficiency of the classification process, despite showing a modest increase in classification precision. The conventional method proved to be effective in document classification, with the BERT model standing out for its performance with an accuracy rate of 94%. The alternative method, in turn, offered a viable solution for challenging scenarios, showing promising results with an accuracy rate of 87% in classifying public documents.

Keywords

Document classification, sensitive data, public organization, artificial intelligence

Sumário

Lista de Figuras **11**

Lista de Tabelas **12**

Lista de Códigos de Programas **13**

1 Introdução **14**
- 1.1 Objetivos 15
  - 1.1.1 Objetivo Geral 15
  - 1.1.2 Objetivos Específicos 16

2 Fundamentação Teórica **17**
- 2.1 Legislação Sobre Classificação de Documentos 17
- 2.2 Revisão da Literatura 18
- 2.3 Trabalhos Relacionados 21
- 2.4 Aprendizado de Máquina 24
  - 2.4.1 Redes Neurais 24
  - 2.4.2 CBIR 25
  - 2.4.3 Métricas de avaliação de desempenho 26

3 Método **27**
- 3.1 Método convencional 28
- 3.2 Método alternativo 29
- 3.3 Sistema web para uso 30

4 Experimentos **31**
- 4.1 Método convencional 31
- 4.2 Método alternativo 35
- 4.3 Sistema web 37

5 Resultados **39**
- 5.1 Método convencional 39
- 5.2 Método alternativo 44
- 5.3 Sistema web 45

6 Conclusão e trabalhos futuros **50**
- 6.1 Publicação 52

7 Agradecimentos **53**

Referências Bibliográficas **54**

A Códigos-fonte **59**

Lista de Figuras

3.1 Arquitetura do método convencional. 28
3.2 Arquitetura do método alternativo. 30

4.1 Exemplo de imagem gerada de um documento. 36

5.1 Evolução da acurácia dos modelos sem utilizar expressão regular. 43
5.2 Evolução da acurácia dos modelos com expressão regular de CPF e celular. 43
5.3 Evolução da acurácia dos modelos com expressão regular de CPF, celular e e-mail. 44
5.4 Matriz de confusão. 45
5.5 Tela para interação com as APIs do back-end. 47
5.6 Exemplo de retorno da API de classificação de texto livre. 47
5.7 Tela de listagem de documentos. 48
5.8 Tela de detalhe de um documento. 49

Lista de Tabelas

2.1 Visão geral dos trabalhos de revisão - parte 1. 19
2.2 Visão geral dos trabalhos de revisão - parte 2. 20
2.3 Técnicas utilizadas em trabalhos de detecção de informação confidencial. 20
2.4 Resumo dos trabalhos relacionados. 23

4.1 Palavras mais frequentes em documentos restritos. 34

5.1 Acurácia dos modelos em cada variação. 40
5.2 Resultado dos avaliadores de desempenho dos modelos sem utilização das expressões regulares. 41
5.3 Resultado dos avaliadores de desempenho dos modelos com expressão regular de CPF e celular. 41
5.4 Resultado dos avaliadores de desempenho dos modelos com expressão regular de CPF, celular e e-mail. 41
5.5 Comparação do desempenho do método convencional com a literatura. 44
5.6 Resultado dos avaliadores de desempenho no método alternativo. 45
5.7 Comparação do desempenho do método alternativo com a literatura. 46

Lista de Códigos de Programas

A.1 Pré-processamento dos documentos: conversão de html, normalização de texto e remoção de acentuação. 59
A.2 Pré-processamento dos documentos: remoção de metadados. 60
A.3 Pré-processamento dos documentos: remoção de stop words. 61
A.4 Geração de imagem do documento. 62
A.5 Classe para utilizar VGG16 e retornar as características de uma imagem. 63

CAPÍTULO 1

Introdução

As organizações públicas produzem inúmeros documentos durante a execução de suas atividades, os quais necessitam ser classificados e ter seu acesso e divulgação controlado com base nas informações que compõem o documento. Segundo a Lei 12.527 [3], que regula o Acesso à Informação, os documentos produzidos pelas organizações públicas devem ser de acesso público, mas são definidas situações específicas em que os documentos devem ser classificados como restritos ou sigilosos.

É dever das organizações públicas promover a transparência dos seus atos, e a não realização disto pode acarretar em sanções. Entretanto, a transparência deve ser realizada sempre à luz das legislações que regem sobre a proteção e sigilo de informações sensíveis, como a Lei Geral de Proteção de Dados Pessoais (LGPD) [4] e a Lei de Acesso à Informação [3]. A exposição pública de dados sensíveis pode acarretar em penalidades tanto para as organizações públicas quanto para os agentes públicos que permitem tal divulgação.

O vazamento de dados sensíveis não prejudica apenas as organizações públicas, mas também compromete as pessoas que foram expostas, tornando elas suscetíveis a serem vítimas de golpes financeiros como a criação e sequestro de contas bancárias, solicitação de empréstimos, financiamentos de veículos, entre outros. Isso demonstra a importância e relevância de uma correta classificação dos documentos produzidos e tramitados pelas organizações públicas, uma vez que o vazamento de dados sensíveis pode ocasionar danos irreparáveis para toda a sociedade.

Em 2020 o Tribunal de Contas da União (TCU), por meio dos Acórdãos 389/2020 e 484/2021, solicitou que todas as Instituições Federais de Ensino Superior concedam, de forma pública, acesso ao inteiro teor dos processos e documentos tramitados pelo Sistema Eletrônico de Informações (SEI). Portanto, a Universidade Federal de Goiás (UFG) necessita publicizar os seus processos e documentos, mas para isso é necessário analisar todos os documentos presentes nos processos concluídos e em andamento, com o intuito de identificar a presença de informações sensíveis que impossibilitam sua divulgação.

Para atender às exigências do TCU, a UFG necessita realizar um trabalho manual

de análise e classificação de todos os seus documentos antes de realizar a publicização. Entretanto, torna-se inviável a análise de todos esses processos e documentos por humanos em virtude da elevada quantidade de registros envolvidos. Assim, faz-se necessário a adoção de técnicas computacionais com o propósito de otimizar o esforço humano demandado para a execução dessa tarefa [21].

Grandes prejuízos podem ser evitados, tanto sociais quanto econômicos, caso fosse possível realizar a classificação automatizada de documentos. Com o avanço e evolução das técnicas de aprendizagem de máquina, diversos problemas reais podem ser resolvidos ou ao menos mitigados. Entre esses problemas, encontra-se a identificação de padrões que indicam a presença de informações sensíveis em um documento.

Modelos de redes neurais profundas (*deep learning*) como a *Bidirectional Encoder Representations from Transformers* (BERT) [10], *Convolutional Neural Network* (CNN) [32] e *Long short-term memory* (LSTM) [17] têm apresentado relevante desempenho no problema de processamento de linguagem natural, especialmente na descoberta de padrões em textos. Adicionalmente, a técnica de recuperação de imagem baseada em conteúdo (*Content-based image retrieval* - CBIR) também se destaca na tarefa de extração de características de forma automática, sendo aplicável na análise de imagens de documentos. Segundo Real [38] apud Ribeiro [39] os sistemas CBIR executam pesquisas em um banco de dados de imagens utilizando como critério a semelhança entre elas. Tais sistemas incorporam uma série de técnicas destinadas ao processamento das imagens com o propósito de converter suas representações em vetores de características. Após isso utilizam-se esses vetores em substituição às próprias imagens para a realização das pesquisas.

Este trabalho está estruturado em seis seções, sendo a Seção 1 uma introdução do problema enfrentado pelas organizações públicas e motivação do trabalho. A revisão bibliográfica da temática do estudo e os trabalhos relacionados estão na Seção 2. Os métodos propostos para a classificação de documentos são expostos na Seção 3. Os experimentos realizados estão presentes na Seção 4. Já a Seção 5 traz a apresentação dos resultados obtidos. Por fim, a conclusão do trabalho e perspectivas para as próximas etapas são apresentados na Seção 6.

1.1 Objetivos

1.1.1 Objetivo Geral

O projeto de pesquisa possui como objetivo identificar automaticamente documentos que foram classificados de forma incorreta, utilizando técnicas de aprendizado de máquina aplicadas ao texto e à imagem dos documentos. Pretende-se propor métodos para

classificar documentos sensíveis, fundamentado pelas experiências e lacunas identificadas na literatura.

Para facilitar a utilização dos resultados deste método pelos gestores documentais da UFG, propõe-se o desenvolvimento de uma solução web projetada para a realização da classificação dos documentos já existentes da UFG e realização da monitoração dos novos documentos, provendo recomendações de classificação.

1.1.2 Objetivos Específicos

- Revisar a literatura para compreender o estado da arte das técnicas de aprendizagem de máquina aplicadas ao tratamento de documentos;
- Coletar dados dos documentos gerados internamente no SEI da UFG;
- Aplicar e avaliar as técnicas de aprendizagem de máquina nos documentos existentes no SEI da UFG;
- Desenvolver solução que identifique documentos classificados como públicos que contém dados sensíveis e documentos classificados como restrito mas que deveriam ser públicos;
- Desenvolver solução web para o monitoramento e classificação dos novos documentos que serão produzidos;
- Publicar artigo científico apresentando os resultados obtidos;
- Escrever, apresentar e revisar o projeto final.

CAPÍTULO 2

Fundamentação Teórica

Os conceitos necessários para compreender a natureza do estudo bem como os trabalhos relacionados são apresentados nesta seção.

2.1 Legislação Sobre Classificação de Documentos

Segundo a Lei de Acesso à Informação [3], um documento é a "unidade de registro de informações, qualquer que seja o suporte ou formato". Desta forma, podemos compreender que todos os atos da administração pública que fiquem registrados de alguma forma podem ser considerados como documento.

Por obrigação, uma organização pública deve realizar transparência e prover acesso à informação dos documentos, conforme salienta Prade [37] apud Brasil [3] "todas as informações produzidas ou custodiadas pelo poder público e não classificadas como sigilosas são públicas e, portanto, acessíveis a todos os cidadãos". No entanto, alguns documentos podem conter informações que impossibilitem sua disponibilização de forma pública, e as definições acerca do que deve ser protegido são esclarecidas por Vieira [48] "A Lei 12.527/2011, além de prever todas as prerrogativas e informações que devem ser públicas, também prevê as exceções, que vem a ser informações de cunho pessoal (dados pessoais), e informações classificadas como "sigilosas", ou seja, informações que tem a guarda do Estado, e que dizem respeito à intimidade, honra, e imagem das pessoas".

A correta classificação do documento é fundamental para uma eficiente transparência pública, uma vez que a divulgação de informações sensíveis pode acarretar em sanções legais tanto para a organização pública quanto para o agente público responsável por este ato. A regulamentação e fiscalização dessa classificação e divulgação são estabelecidas pela Lei Geral de Proteção de Dados Pessoais [4].

Além de danos legais, a publicação de dados sensíveis também pode ocasionar danos sociais significativos, especialmente para pessoas cujos dados pessoais foram vazados. "Com os subsídios obtidos no vazamento, criminosos puderam enviar faturas falsificadas ou golpes como saques indevidos do Fundo de Garantia do Tempo de Serviço (FGTS)" [36].

2.2 Revisão da Literatura

Estudos recentemente publicados que realizaram revisões da literatura sobre técnicas de classificação de texto e detecção de dados sensíveis foram identificados. As Tabelas 2.1 e 2.2 apresentam uma visão geral destes trabalhos, incluindo resumos de suas respectivas descobertas e contribuições.

Na pesquisa realizada por Kobayashi et al. [21], a tarefa de classificação de texto foi descrita de forma detalhada e em várias etapas sequenciais, sendo: preparação de dados de treinamento, pré-processamento, transformação, aplicação de técnicas de classificação e validação, com recomendações específicas para cada uma das etapas. Adicionalmente, foram disponibilizados para cada uma das etapas códigos na linguagem de programação R e um tutorial, visando auxiliar os pesquisadores na criação de seus classificadores de texto.

A detecção de informações sensíveis e confidenciais é uma especialidade da classificação de texto. Sousa et al. [42] realizou uma revisão sistemática abrangendo mais de 60 técnicas de *deep learning* (redes neurais profundas) aplicados à preservação da privacidade no processamento de linguagem natural. Neste estudo, um tópico específico foi dedicado aos métodos de detecção de informação confidencial. Foram analisados as contribuições de oito trabalhos publicados referente ao uso de *deep learning* na identificação de informações sensíveis. Dentre esses, destaca-se os trabalhos realizados por:

- Neerbek et al. [29] propôs a realização da classificação de documentos em conteúdo sensível e não sensível através do uso de uma rede neural recursiva, treinada em um conjunto de documentos rotulados. Os autores evidenciaram que abordagens baseadas em palavras-chaves podem não ser eficazes na detecção de informações sensíveis complexas, uma vez que essa abordagem desconsidera o contexto das frases no documento;
- Battaglia et al. [1] introduziram uma nova abordagem denominada análise de sensibilidade de conteúdo, que visa atribuir pontuações entre -1 e 1 aos documentos, levando em conta o seu grau de sensibilidade.

As técnicas de classificação utilizadas nos 8 estudos examinados por Sousa et al. [42], sobre detecção de informação confidencial estão sumarizados na Tabela 2.3.

As limitações do uso de técnicas tradicionais de aprendizagem de máquina e o desempenho superior das técnicas de *deep learning* nas tarefas de classificação de texto foram discutidos nos trabalhos de Minaee et al. [27], Wu et al. [49], Kowsari et al. [22] e Patil et al. [34]. De acordo com os autores, as abordagens convencionais para classificação de texto, baseadas em técnicas tradicionais, possuem uma série de fraquezas, tais como a limitação da capacidade de generalização, explosão de dimensão, dispersão de dados,

Tabela 2.1: Visão geral dos trabalhos de revisão - parte 1.

Trabalho	**Ano**	**Contribuição**
How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing [42].	2023	O estudo revisou mais de 60 técnicas de *deep learning* aplicados a preservação de privacidade em processamento de linguagem natural. Os autores categorizaram as tarefas de preservação de privacidade em: métodos de salvaguarda de dados, métodos de confiança (*trusted methods*) e métodos de verificação. Os métodos para detecção de informação confidencial que são estudados neste trabalho foram descritos dentro da categoria de métodos de verificação.
Systematizing the State of Knowledge in Detecting Privacy Sensitive Information in Unstructured Texts using Machine Learning [24].	2023	Os modelos mais utilizados nos trabalhos sobre detecção de informações sensíveis foram sumarizados neste artigo. As tendências e *gaps* dessa tarefa foram elencadas e discutidos, tendo sido destacado a tendência de uso de modelos de redes neurais e o *gap* da qualidade e disponibilidade de dados para treinar os modelos.
A Survey on Text Classification: From Traditional to Deep Learning [23].	2022	Este artigo oferece uma análise abrangente dos modelos de classificação de texto desde modelos tradicionais até modelos de redes neurais. Também foi exposto as métricas de avaliação de desempenho e bases de dados disponíveis para essa tarefa.
A Complete Process of Text Classification System Using State-of-the-Art NLP Models [11] e A Survey on Text Classification Algorithms: From Text to Predictions [14].	2022	Os artigos oferecem uma revisão abrangente dos modelos tradicionais e de redes neurais aplicados à classificação de textos, fornecendo as vantagens e desvantagens de cada modelo.
Deep Learning–based Text Classification: A Comprehensive Review [27].	2021	O artigo discute como os modelos de *deep learning* superaram as abordagens tradicionais de aprendizado de máquina em tarefas de classificação de textos. Uma revisão de mais de 150 modelos baseados em *deep learning* foi realizada, destacando vantagens e desvantagens. Também proveram uma síntese de mais de 40 conjuntos de dados populares usados para classificação de texto.
Review of Text Classification Methods on Deep Learning [49].	2020	Os desafios das tarefas de classificação de texto e limitações dos modelos tradicionais de aprendizado de máquina foram discutidos. Também foi apresentado um extensivo estudo de modelos de *deep learning* para classificação de texto.

Fonte: Elaborado pelo autor

Tabela 2.2: Visão geral dos trabalhos de revisão - parte 2.

Trabalho	**Ano**	**Contribuição**
Text Classification Algorithms: A Survey [22].	2019	Este artigo fornece uma revisão abrangente de várias técnicas de extração de *features* de texto, métodos de redução de dimensionalidade, algoritmos existentes e técnicas de avaliação. Além disso, foi discutido as limitações de cada técnica e sua aplicabilidade em cenários do mundo real.
Comparing automated text classification methods [18].	2019	O estudo realizou uma comparação do desempenho de cinco abordagens baseadas em léxico e cinco algoritmos de aprendizado de máquina, aplicados a 41 conjuntos de dados de mídias sociais.
Text Classification Techniques: A Literature Review [45].	2018	O trabalho analisou diversas técnicas de classificação de texto empregadas na prática. Foi identificado os pontos fortes, as limitações e as tendências de pesquisa atuais na classificação de texto.
Text Classification for Organizational Researchers: A Tutorial [21].	2018	O artigo descreve o processo de classificação de texto em etapas sequenciais: preparação de dados de treinamento, pré-processamento, transformação, aplicação de técnicas de classificação e validação. Recomendações específicas foram fornecidas em cada etapa para orientar a implementação do processo.

Fonte: Elaborado pelo autor

Tabela 2.3: Técnicas utilizadas em trabalhos de detecção de informação confidencial.

Trabalho	**Rede neural utilizada**
Neerbek et al. [29]	RecNN, GloVe
Battaglia et al.[1]	MLP, GloVe
Zhao et al. [51]	ELMO, BERT, XL-Net, fasttext
Tan et al. [44]	GloVe, ELMO, BERT, GPT, GPT-2
Hutchinson et al. [19]	BERT
Gonen et al. [16]	Hard-Debiased, GN-GloVe, Glove
Nissim et al. [30]	word2vec
Sweeney et al. [43]	word2vec, GloVe, Conceptnet

Fonte: Adaptado de Sousa et al. [42].

entre outras. Ainda segundo os autores, as técnicas de *deep learning* conseguem superar as fraquezas das abordagens tradicionais, apresentando desempenho superior na realização dessas tarefas.

2.3 Trabalhos Relacionados

O trabalho realizado por McDonald et al. [25] propôs a detecção de informações confidenciais presentes em documentos governamentais do Reino Unido e dos Estados Unidos. Utilizou-se a metodologia de *sensitivity load of POS n-grams* para identificar sequências de textos sensíveis na base de dados de documentos revisados por especialistas no assunto. A metodologia proposta apresentou taxa de acurácia de 97%.

Já Occhipinti et al. [31] apresentou um *pipeline* para classificação de textos, composto pelas seguintes etapas:

- Divisão dos dados em 70% para treino e 30% para teste;
- Pré-processamento dos textos por meio das técnicas de remoção de *stop word*, *HTML tags* e *lemmatisation*;
- Extração de *features* dos textos para gerar um dicionário;
- Utilização de doze modelos de aprendizado de máquina para treinamento com os dados de teste e execução de predições em dados novos.

A base de dados utilizada para avaliar o desempenho do *pipeline* foi a Enron [26], conhecida por ser uma base de dados de spam. O resultado obtido por meio da métrica F-*Score* foi de 94%.

Um *pipeline* para classificação de textos sensíveis foi apresentado no trabalho de Geetha et al. [15]. Os passos realizados para classificar um texto sensível se assemelham com o trabalho de Occhipinti et al. [31] mas com o diferencial da inclusão de uma etapa adicional caso o texto seja classificado como não sensível. Essa etapa consiste na realização de uma nova classificação utilizando uma base de dados das palavras-chaves mais frequentes obtidas a partir de um conjunto de textos sensíveis revisados por especialistas. O *pipeline* foi aplicado em um conjunto de dados de *tweets* e apresentou um aumento de 9% no desempenho da classificação de *tweets* que contém texto sensível em comparação com o desempenho de um *pipeline* tradicional.

Um conjunto de dados de 10.000 postagens sensíveis e não sensíveis feitas nas redes sociais Twitter e Facebook foi publicado pelo trabalho de Bioglio et al. [2]. O trabalho também avaliou as abordagens do estado da arte que utilizam a análise léxica e anonimato no conjunto de dados anotado pelos especialistas, concluindo que essas abordagens apresentam um desempenho insuficiente em problemas reais. Com o objetivo de melhorar o desempenho das abordagens na classificação de postagens com textos

sensíveis, foi identificado a possibilidade de utilizar a técnica *Bidirectional Encoder Representations from Transformers* (BERT) [10] tendo em vista que o conjunto de dados contém palavras-chaves específicas, como *home*, *tomorrow* e *tonight*, que podem obter melhor desempenho com a utilização da técnica BERT. A abordagem proposta obteve um resultado superior em comparação com as abordagens do estado da arte avaliadas.

Jiang et al. [20] demonstrou a viabilidade da utilização da classificação de texto como forma de proteção de dados médicos sensíveis. Uma base de dados contendo 736 registros de textos sobre diabetes foram utilizadas em conjunto com a técnica de pré-processamento de dados TF-IDF para avaliar o desempenho dos algoritmos de classificação Naive Bayes, *K-nearest Neighbor* (KNN) e *Support Vector Machine* (SVM). Os resultados dos experimentos demonstraram a viabilidade da utilização de técnicas de classificação de texto para identificar dados médicos sensíveis, tendo obtido taxas de acerto acima de 90% e *recall* acima de 85% nos classificadores Naive Bayes e SVM. Já o classificador KNN apresentou o menor desempenho entre os três classificadores avaliados, obtendo taxa de acerto de menos de 85% nos avaliadores de desempenho.

O artigo de Gambarelli et al. [13] apresentou o SPEDAC (*Sensitive Personal Data Categories Corpus*), um novo conjunto de dados anotados destinado à identificação de categorias de dados pessoais sensíveis, contendo 10.615 registros. O SPEDAC foi criado para preencher a lacuna na falta de um padrão para avaliar modelos computacionais em tarefas de detecção de informação sensível. Este conjunto de dados permite a avaliação de modelos computacionais em três sub-tarefas com níveis crescentes de complexidade, variando de classificação binária, avaliando se o texto é sensível ou não, até a categorização detalhada em 61 categorias de dados pessoais. Os autores avaliaram o desempenho de diversos modelos do estado da arte na classificação dos textos do conjunto de dados, tendo o modelo RoBERTa obtido a maior acurácia, de 98%.

O crescente desafio da gestão da privacidade de dados pessoais em meio à digitalização em massa e a ausência de leis robustas de proteção de dados foi abordado pelo trabalho de Patil et al. [35]. A avaliação da melhor combinação de pré-processamento e modelos foi realizada, tendo o uso da técnica de pré-processamento TF-IDF, uma técnica de representação do texto em formato numérico, em conjunto com o modelo Support Vector Machine (SVM) obtido o melhor desempenho, com acurácia de 98%. Adicionalmente foi explicado as técnicas de limpeza e tratamento de textos: remoção de caracteres de nova linha, remoção de HTML, checagem de valores nulos, conversão do texto para minúsculo e remoção de *stop words*.

A Tabela 2.4 apresenta o resumo dos trabalhos relacionados.

Tabela 2.4: Resumo dos trabalhos relacionados.

Trabalho	**Pipeline**	**Técnicas de classificação**	**Desempenho**
Gambarelli et al. (2023) [13]	-	KNN, SVM, LR, LABSE, RoBERTa e DeBERTa	Acurácia de 98%
Bioglio et al. (2022) [2]	-	BERT	F1-Score de 80%
Geetha et al. (2022) [15]	*Pipeline* de Occhipinti com a adição de uma etapa para reclassificar, utilizando palavras-chaves, os textos classificados como não sensíveis	SVM, Naive Bayes, Gradient Boosting, Palavras-chaves	Acurácia de 66%, precisão de 81% e recall de 71%
Patil et al. (2022) [35]	Pré-processamento: remoção de caracteres de nova linha, remoção de HTML, checagem de valores nulos, conversão do texto para minúsculo e remoção de *stop words*. Representação do texto em formato numérico: *One-Hot Encoding*, *Bag-of-words*, *Count Vectorizer*, TF-IDF, *Co-Occurrence Matrix*, Word2Vec, CBOW, *Skip-Gram* e Glove	Naive Bayes, SVM, Random Forest, XGBoost, Shallow NN, CNN, LSTM, GRU, BI-RNN e RCNN	Acurácia de 98%
Occhipinti et al. (2021) [31]	Pré-processamento: stop word, HTML tags e lemmatisation.	KNN, MPNN, Regressão logística, Random Forest, XGBoost, Multinomial NB, Gaussian NB, Bernoulli NB, RBF SVM, Linear SVM, Poly SVM, Sigmoid SVM	F1-Score, precisão e recall de 94%
Jiang et al. (2019) [20]	Pré-processamento: TF-IDF	Naive Bayes, KNN e SVM	F1-Score de 90%, precisão de 92% e recall de 97%
McDonald et al. (2015) [25]	*sensitivity load of POS n-grams*	*Conditional Random Fields (CRF)*	Acurácia de 97%, precisão de 99% e recall de 45%

Fonte: Elaborado pelo autor

2.4 Aprendizado de Máquina

Buscando apoiar a organização pública no dever de cumprir o estabelecido em legislações e evitar a exposição de informações sensíveis, tecnologias podem ser utilizadas como mecanismos para dar suporte nessa situação. “As tecnologias disruptivas estarão cada dia mais presentes na vida das pessoas, não podendo a Administração Pública ficar alheia a essa realidade; conforme demonstram as experiências descritas, a Inteligência Artificial mostra-se como uma importante ferramenta para o aperfeiçoamento da gestão pública” [9].

Segundo Monard et al. [28] “Aprendizado de Máquina é uma área de Inteligência Artificial cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática.”. Essa característica do aprendizado de máquina permite que inúmeros problemas do mundo real possam ser estudados e solucionados por meio do desenvolvimento de soluções robustas e automáticas que se utilizem de técnicas de aprendizado de máquina.

Para Rossi [40] um dos problemas em que as técnicas de aprendizagem de máquina podem ser aplicadas é a classificação automática de textos, que constitui o objeto de estudo deste trabalho.

Nas seções posteriores, introduziremos os conceitos das técnicas de classificação e os avaliadores de desempenho propostos para o trabalho, fundamentando-se nos estudos presentes nas Seções 2.2 e 2.3.

2.4.1 Redes Neurais

Com base nas evidências encontradas nos estudos da literatura, propomos a utilização dessas três técnicas de redes neurais:

1. *Bidirectional Encoder Representations from Transformers* (BERT) foi apresentado por pesquisadores do Google no final de 2018. Diferentemente dos modelos de representação de linguagem tradicionais que analisam palavras isoladamente ou em apenas uma direção, o BERT leva em consideração o significado de uma palavra com base em seu contexto bidirecional, analisando da esquerda para a direita e da direita para a esquerda. Esta abordagem permite que o BERT tenha uma compreensão mais aprofundada do significado de uma palavra em um texto, resultando em um desempenho superior nas tarefas de processamento de linguagem natural (PLN). Resultados do estado da arte foram alcançados em várias tarefas de PLN pelo BERT [10];
2. *Convolutional Neural Network* (CNN) é amplamente empregada em tarefas de processamento de imagens, devido à sua capacidade de reconhecer padrões visuais

complexos. Além disso, a CNN também demonstrou sua eficácia em tarefas de processamento de linguagem natural, como classificação de texto. Uma de suas vantagens é a necessidade de menos recursos computacionais para o processamento, já que a CNN requer uma quantidade menor de parâmetros em comparação com outras redes neurais. Isso resulta em uma arquitetura mais leve e rápida para executar [42];

3. *Long Short-Term Memory* (LSTM) é um tipo específico de *Recurrent Neural Network* (RNN) que se destaca por possuir uma célula de memória capaz de lembrar valores relevantes ao longo de intervalos de tempo arbitrários [27]. Esta arquitetura se mostra vantajosa em contextos onde é importante preservar informações relevantes de um trecho específico do texto durante a análise de um documento completo.

2.4.2 CBIR

Convencionalmente, nas tarefas de classificação de texto são utilizadas técnicas que analisam o conteúdo textual dos documentos, porém, existem cenários em que a obtenção de textos dos documentos é inviável ou a aplicação de técnicas de reconhecimento ótico de caracteres (OCR) é ineficaz. Nesses cenários, torna-se necessário a aplicação de abordagens que utilizem a imagem dos documentos.

A recuperação de imagem baseada em conteúdo (*Content-based image retrieval* - CBIR) é uma técnica que tem apresentado relevante desempenho na extração de características de forma automática, podendo ser aplicada na imagem de documentos. Segundo Costa et al. [7], o CBIR funciona através de 3 passos:

- Extração de Características: Primeiramente a imagem é analisada e as características de interesse são extraídas através de algum descritor.
- Indexação de Características (Vetor de Características): Uma vez extraídas, é montado um vetor de características com todas as informações extraídas (tais como cor, dimensão das estruturas encontradas, formas identificadas, etc.) da imagem.
- Comparação: Ao final é realizada a comparação entre os diversos vetores de características que foram extraídos das imagens, é nessa etapa que são definidas as imagens semelhantes que devem ser recuperadas.

Redes Neurais Convolucionais (CNN) são utilizadas para realizar a extração de características de forma automática. Conforme evidenciado por Saraiva et al. [41], o modelo de CNN VGG-16, com os pesos definidos pelo conjunto de dados ImageNet, demonstra desempenho significativo na solução do problema de extração de características.

A utilização de um modelo de CNN que apresente um bom tempo de busca e recuperação de conteúdo, além de um baixo tamanho do vetor de características

gerado, são fundamentais para possibilitar o desenvolvimento do trabalho. Neste contexto, Cools et al. [6] realizou no seu estudo um comparativo entre os modelos DenseNet169, MobileNet, ResNet50, VGG-16 e Xception avaliando esses dois critérios. De acordo com os resultados obtidos por Cools, a VGG-16 apresentou o menor tempo de busca e menor vetor de características entre os modelos avaliados. Diante disso, o autor concluiu que a VGG-16 apresenta-se como a melhor opção para ser utilizada nesse contexto.

2.4.3 Métricas de avaliação de desempenho

Conforme definido pelos trabalhos de Minaee et al. [27], Kowsari et al. [22], [49], Wu et al. [23] e Kobayashi et al. [21], os principais avaliadores de desempenho utilizados na mensuração do desempenho de modelos de classificação de texto e que serão utilizados no trabalho são:

- Acurácia: É uma métrica de avaliação utilizada em base de dados balanceadas. Essa métrica identifica a proporção de predições corretas em relação ao total de predições realizadas;
- Área Sob a Curva ROC: Mede a capacidade do modelo em conseguir distinguir as classes positivas e negativas;
- F1 Score: A média entre a precisão e *recall* é representada pelo F1 Score;
- Precisão: É a proporção de predições positivas corretas pelo total de observações positivas. A precisão reflete a capacidade do modelo de identificar as observações positivas corretamente, sendo amplamente utilizada em contextos onde o custo de um falso positivo é alto. Essa métrica compara do total de predições positivas feitas, quantas foram corretas;
- *Recall*: Métrica que quantifica a proporção de predições positivas corretas feitas em relação ao total de observações positivas que poderiam ter sido realizada. Essa métrica verifica do total de observações positivas, quantas foram identificadas.

CAPÍTULO 3

Método

O projeto tem como objetivo o desenvolvimento de métodos para a realização da classificação de documentos produzidos no SEI da UFG como público ou restrito, e também o desenvolvimento de uma aplicação web para que os gestores documentais da UFG possam utilizar o método desenvolvido.

Segundo a Plataforma Analisa UFG [46], em 15 de março de 2024, a UFG havia produzido um total de 291.694 processos e 3.362.166 documentos em seu SEI. Para atender a solicitação do TCU, de tornar público o inteiro teor dos processos e documentos presentes no SEI, a UFG necessita realizar um trabalho manual de análise e classificação de todos os documentos antes de realizar a publicização.

Com o objetivo de apoiar a gestão arquivística da UFG nessa tarefa e com base nos trabalhos da literatura apresentados na Seção 2 e nas lacunas identificadas, propomos o desenvolvimento de dois métodos para resolver a tarefa de classificação de documentos sensíveis. O método convencional parte da premissa de que é viável coletar o texto dos documentos com qualidade. No entanto, nos cenários em que a coleta dos textos não seja possível ou não retorne textos com qualidade, então pode-se aplicar o método alternativo.

O desenvolvimento deste trabalho será orientado pelo processo de descoberta de conhecimento em base de dados, também conhecido como *Knowledge Discovery in Database* (KDD). A descoberta de conhecimento em bases de dados consiste no processo de realizar a identificação nos dados de padrões relevantes, novos, potencialmente úteis e compreensíveis. Esse processo é estruturado em cinco atividades: coleta e seleção dos dados, pré-processamento, transformação, mineração de dados e interpretação/avaliação [12].

Para a realização de todas as atividades do processo de descoberta de conhecimento após a coleta e seleção dos dados, será utilizado a linguagem de programação Python[1] na versão 3.10, dada a sua robustez e ampla disponibilidade de bibliotecas de aprendizagem de máquina.

[1] https://www.python.org/

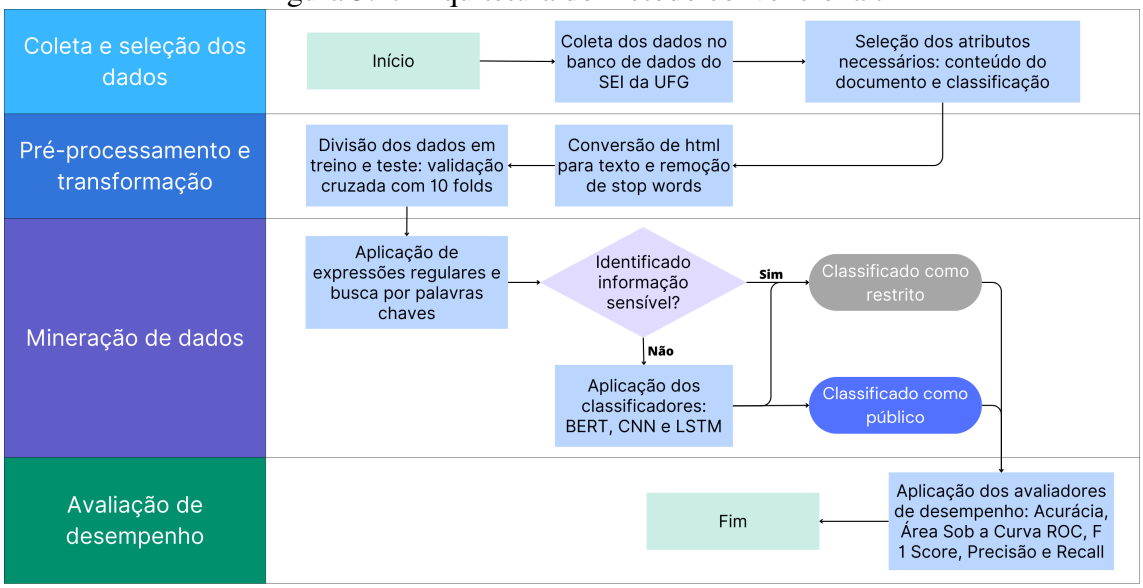
O projeto conta com apoio e parceria da Secretaria de Tecnologia e Informação (SETI) da UFG, por meio do Centro de Recursos Computacionais (CERCOMP) e Centro de Informação, Documentação e Arquivo (CIDARQ). Essa parceria garante a disponibilidade de todos os recursos necessários para a realização das atividades, incluindo o acesso ao banco de dados do SEI.

A autorização para a coleta e utilização dos dados foi submetida e aprovada pelo Comitê de Ética em Pesquisa (CEP) da UFG, através da Plataforma Brasil, sob o número CAAE: 60165422.7.0000.5083.

3.1 Método convencional

O método convencional consiste de atividades decorrentes do KDD, bem como dos *pipelines* descritos na literatura, como mostrado na Tabela 2.4. A Figura 3.1 ilustra cada uma das etapas do método convencional.

Figura 3.1: Arquitetura do método convencional.



Fonte: Elaborado pelo autor

No início, ocorre a coleta e seleção dos documentos. Após essa fase inicial, será realizado o pré-processamento e transformação dos dados, que inclui a conversão de HTML para texto e a remoção de *stop words*. Essa etapa é crucial para garantir a remoção de ruídos e informações desnecessárias dos documentos. Após isso, os dados serão divididos entre treino e teste utilizando a estratégia de validação cruzada (*cross validation*) com 10 dobras (*fold*). A quantidade de 10 dobras para a validação cruzada foi escolhida com base na recomendação de Kobayashi [21], que sugere o uso de 10 dobras para bases de dados menores, enquanto para bases de dados maiores, recomenda-se a utilização de 5 dobras.

Na mineração de dados, será realizado a análise do conteúdo para identificar palavras-chaves predominantes dentre os documentos restritos para construir um dicionário, que servirá como insumo para indicar que um documento deve ser restrito.

A próxima fase envolve na identificação de expressões regulares que podem ser aplicáveis neste contexto e que consigam realizar a detecção de padrões de dados nos documentos, os quais podem assinalar que um documento deve ser restrito. Para exemplificar, pode-se empregar expressões regulares para reconhecer uma sequência de onze dígitos e verificar se estes atendem a regra de um CPF válido.

Com as expressões regulares e palavras-chaves estabelecidas, a primeira checagem feita para a realização da classificação de um documento consiste na verificação se o conteúdo do documento atende alguma expressão regular ou palavra-chave. Se atender, o documento é imediatamente classificado como restrito. Caso o documento não se enquadre em nenhuma das expressões regulares e palavras-chaves, então ele irá passar pelo processo de classificação utilizando técnicas de aprendizagem de máquina, resultando na sua classificação final como público ou restrito.

Durante a mineração de dados, os avaliadores de desempenho serão empregados para mensurar a qualidade da classificação do método, fornecendo subsidio para ajustes de parâmetros e aperfeiçoamentos em cada etapa.

3.2 Método alternativo

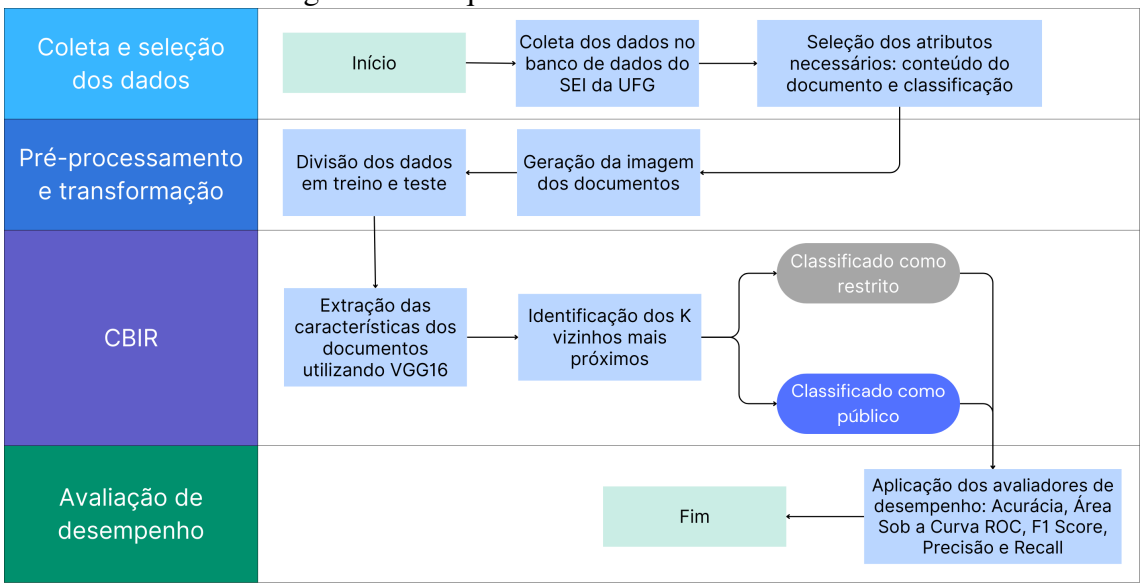
No cenário em que a obtenção de textos dos documentos é inviável ou a aplicação de técnicas de reconhecimento ótico de caracteres (OCR) não traz bons resultados, propomos um método alternativo. Esse método possibilita a realização da tarefa de classificação de texto nesses cenários por meio da utilização da imagem dos documentos em vez do seu texto.

O método alternativo distingue-se do método convencional em dois aspectos: a etapa de pré-processamento dos dados foi substituída pela geração de imagens dos documentos e apresenta um fluxo diferente durante a mineração de dados. Inicialmente, ocorre a extração de características dos documentos de treinamento e teste. Posteriormente, para cada documento de teste são identificados os K documentos mais semelhantes, ou seja, aqueles que tenham características mais próximas.

Após a identificação dos documentos mais semelhantes, então é utilizada a mediana para verificar qual categoria predomina entre esses documentos. O resultado da mediana é utilizado para a realização da classificação do documento em público ou restrito. Essa metodologia de classificação baseia-se no KNN, que foi descrito por Zhai [50].

A arquitetura do método alternativo pode ser observado pela Figura 3.2.

Figura 3.2: Arquitetura do método alternativo.



Fonte: Elaborado pelo autor

3.3 Sistema web para uso

Para assegurar que os gestores documentais da UFG possam utilizar os resultados do trabalho, um sistema web baseado na arquitetura de *back-end* e *front-end* será desenvolvido. O *framework* Django[2], escrito na linguagem de programação Python, será utilizado, dado que ele provê toda a estrutura necessária para um desenvolvimento ágil e que atenda as demandas dos usuários.

A arquitetura de *back-end* e *front-end* permite que o sistema tenha interoperabilidade e seja reutilizável por outras aplicações. A execução da classificação dos documentos será realizada por meio de interfaces de programação de aplicação (API), as quais estarão disponíveis para integração com outras aplicações.

[2]https://www.djangoproject.com/

CAPÍTULO 4

Experimentos

Os experimentos realizados para coletar o desempenho dos dois métodos propostos e para o desenvolvimento do sistema web são detalhados neste capítulo.

4.1 Método convencional

Para assegurar um treinamento eficaz dos modelos de aprendizado de máquina, faz-se importante o uso de bases de dados curadas e com grande quantidade de registros. Diante da inexistência de uma curadoria humana para elaborar uma base de dados para este trabalho, optou-se pela estratégia de utilizar os documentos gerados pelo Centro de Informação, Documentação e Arquivo (CIDARQ) da UFG. O CIDARQ, "órgão responsável pelo planejamento e desenvolvimento da gestão técnica da informação e de documentos de valor administrativo, acadêmico ou histórico"[47], dispõe de especialistas com conhecimentos necessários para avaliar e classificar documentos como públicos ou restritos. Portanto, a adoção de documentos provenientes do CIDARQ proporcionou uma base de dados próxima à ideal para o treinamento dos modelos.

A coleta dos dados resultou em uma base de dados de 21.398 documentos produzidos internamente no SEI, dos quais 20.459 são documentos públicos e 939 são restritos. Esses documentos foram gerados entre 26 de setembro de 2017 e 13 de fevereiro de 2024.

O pré-processamento dos dados coletados é uma etapa fundamental para assegurar a qualidade e a uniformidade dos dados por meio da limpeza de impurezas e ruídos. No método proposto, foram inicialmente adotados dois tratamentos específicos: a conversão de HTML para texto e a remoção de *stop words*, visando preparar adequadamente os documentos para as etapas subsequentes.

No entanto, durante a execução dos experimentos, constatou-se que esses procedimentos não eram suficientes para uma sanitização eficaz dos documentos. Ficou evidente a necessidade de implementação de tratamentos adicionais para remover ou mitigar ruídos e informações irrelevantes que ainda persistiam nos documentos. Esse reconhecimento levou à busca e aplicação de técnicas de pré-processamento mais sofisticadas e

específicas, a fim de aprimorar a qualidade dos dados e, consequentemente, melhorar o desempenho dos modelos na classificação dos documentos. Portanto, nesta etapa foram realizados os seguintes tratamentos:

- Conversão de HTML para texto: Os documentos do SEI são salvos no banco de dados em formato HTML. Para um melhor desempenho dos modelos de aprendizado de máquina é necessário extrair apenas o conteúdo textual dos documentos, descartando códigos HTML que podem interferir no treinamento do modelo (Código A.1);
- Normalização de texto: Todas as palavras foram convertidas para letras minúsculas, a fim de evitar que o modelo interprete variações de maiúsculas e minúsculas como palavras distintas (Código A.1);
- Remoção de acentuação: Acentos foram eliminados para simplificar o conjunto de caracteres e evitar a diferenciação entre palavras acentuadas e não acentuadas (Código A.1);
- Remoção de metadados específicos: Metadados irrelevantes para a análise, como número do processo, número do documento, e-mail padrão do SEI (noreply.sei@sistemas.ufg.br) e as palavras "sei/ufg", foram removidos (Código A.2);
- Remoção de *stop words*: Palavras comuns que não contribuem para o significado do texto, conhecidas como *stop words*, foram excluídas para focar o treinamento do modelo nas palavras que efetivamente podem representar informações relevantes (Código A.3).

Através destas etapas de pré-processamento, a base de dados foi preparada de forma a maximizar o desempenho do treinamento dos modelos de aprendizado de máquina.

Com a finalidade de identificar elementos que pudessem caracterizar um documento como restrito, foram elaboradas expressões regulares baseando-se nas diretrizes estabelecidas pelo Ministério da Economia e pela Controladoria Geral da União [8] sobre o que caracteriza um documento restrito:

- Identificação de CPF: Verificação de sequências numéricas de 6 a 11 dígitos e que atendem a regra de CPF válido;
- Identificação de e-mail pessoal: Uma lista de 73 provedores de e-mail pessoais foi elaborada, baseando-se nos mais utilizados pelos usuários da UFG. A seleção desses provedores foi feita a partir da análise de frequência em cadastros de usuários, seguida de uma triagem manual para identificar os provedores de e-mail pessoal. A expressão regular desenvolvida busca endereços de e-mail que correspondam a esses provedores. Após a validação dessa checagem, percebeu-se

uma característica na cultura de uso de e-mails da UFG. As pessoas criam e utilizam e-mails do provedor Gmail e colocam no inicio ou no final do nome a palavra "ufg", passando a utilizar esse e-mail como institucional. Então uma verificação adicional foi realizada para excluir os e-mails contendo a palavra "ufg"no nome;

- Identificação de números de telefone celular: Criou-se uma expressão regular capaz de identificar números de celular que atendam a regra: contenham o código de área (DDD) e comecem com o dígito 9. Para implementar isto, primeiro realizou-se a remoção de espaços, parênteses e traços, pois estes elementos são frequentemente utilizados para formatar um número de telefone. Depois foi identificado sequências numéricas de 10 a 11 dígitos, verificando se os dois primeiros dígitos representam um DDD válido e se o terceiro dígito é o número 9.

A busca por palavras-chave objetivou a identificação de termos predominantes em documentos restritos. Contudo, após feito esta ação, a análise revelou um conjunto de palavras amplamente genéricas, indicando a complexidade e a limitação de se utilizar palavras-chave para a identificação de documentos restritos. A Tabela 4.1 apresenta as 20 palavras mais frequentes identificadas, evidenciando a predominância de termos genéricos e não específicos, portanto, pouco úteis para a diferenciação entre documentos públicos e restritos.

Um desafio significativo encontrado na base de dados coletada foi o desbalanceamento acentuado entre as classes, com uma predominância de documentos públicos. Para garantir uma avaliação mais justa e precisa dos modelos de aprendizado de máquina, optou-se pela realização do balanceamento da base de dados. A estratégia adotada realizou a seleção aleatória de uma quantidade de documentos públicos igual à quantidade de documentos restritos disponíveis. Dessa forma, conseguimos compor uma base de dados balanceada, contendo 1878 registros, com 939 registros atribuídos a cada classe. Esse balanceamento é crucial para evitar viés nos modelos de aprendizado de máquina, permitindo que eles tenham a mesma capacidade de aprender a distinguir entre documentos públicos e restritos sem serem influenciados pela predominância de uma classe sobre a outra.

Para a classificação dos documentos, foram avaliados os modelos de redes neurais BERT, CNN e LSTM em suas arquiteturas padrão. A única alteração realizada na arquitetura padrão dos modelos foi o aumento no número de épocas (*epoch*) de treinamento, passando de 10 para 100. Adicionalmente, utilizou-se o mecanismo desenvolvido por Carvalho [5] para avaliar modelos tradicionais de aprendizado de máquina, visando uma comparação direta do seu desempenho em relação aos modelos de redes neurais.

A inclusão da avaliação de modelos tradicionais de aprendizado de máquina neste trabalho foi motivada pela necessidade de explorar alternativas de menor custo computacional que ainda pudessem oferecer desempenhos comparáveis aos avançados

Tabela 4.1: Palavras mais frequentes em documentos restritos.

Palavra	**Ocorrências**
trabalho	1910
sei	1214
sistema	1173
n	1162
nao	1111
cadastro	984
ufg	896
remoto	880
processo	868
usuario	835
servidor	684
eletronico	677
informacoes	670
atividades	647
data	639
ufgbr	623
federal	615
uso	578
documento	571
documentacao	567

Fonte: Elaborado pelo autor

modelos de redes neurais. Esta abordagem busca encontrar um equilíbrio considerando a importância de otimizar recursos em ambientes computacionais limitados ou em aplicações que demandam alta velocidade de processamento. A escolha de investigar modelos tradicionais baseou-se na premissa de que, embora menos complexos, estes podem ser suficientemente robustos para certas tarefas de classificação, apresentando um desempenho próximo ao de modelos mais sofisticados, mas com a necessidade de um menor poder computacional.

Neste contexto, o trabalho de Carvalho [5] emergiu como uma ferramenta adequada para o propósito, oferecendo um sistema de *auto machine learning* capaz de avaliar 27 diferentes modelos tradicionais de classificação e identificar aqueles com os melhores desempenhos. Por meio deste sistema, identificamos os três modelos tradicionais de aprendizado de máquina, que apresentaram melhor desempenho na tarefa de classificação dos documentos, para serem comparados com o desempenho dos modelos de redes neurais.

A validação cruzada com 10 dobras (*fold*) foi adotada como estratégia de avaliação, proporcionando uma análise robusta sobre o desempenho dos modelos. A

biblioteca StratifiedKFold[1] do Python foi utilizada para aplicar esta técnica.

4.2 Método alternativo

Para avaliar o desempenho do método alternativo, foi realizado a execução de todas as etapas propostas em um conjunto reduzido de documentos.

Para a aplicação da técnica de CBIR, primeiramente é necessário realizar a coleta e tratamento dos dados. Nesse sentido, foram coletados 20.000 documentos gerados internamente no SEI, entre 6 de abril de 2022 e 11 de agosto de 2022, sendo 10.000 documentos públicos e 10.000 documentos restritos. Esses documentos são armazenados no banco de dados do SEI como texto em formato HTML, e através da biblioteca imgkit[2] da linguagem de programação Python foi possível gerar a imagem dos documentos. A Figura 4.1 apresenta um exemplo de imagem gerada de um documento público, produzida pela biblioteca imgkit. Já o Código A.4 exibe o código fonte desenvolvido para realizar essa ação.

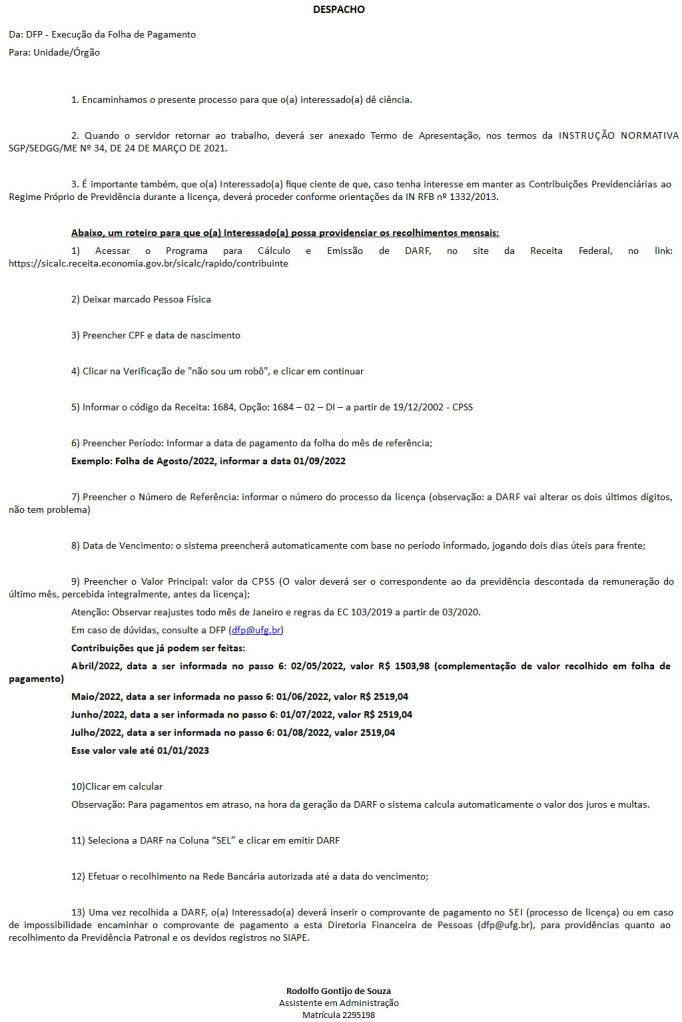
Após a geração das imagens de todos os documentos da base de dados, utilizou-se a CNN VGG-16[3], com os pesos pré treinados da ImageNet, para a realização da extração das características de cada documento e geração um vetor de características. O Código A.5 apresenta a classe usada para utilizar a VGG16 com os pesos pré treinados e retornar as características da imagem. Por meio dessa abordagem, a CNN foi capaz de gerar, de forma automática, 4096 características de cada documento. O tempo de execução da extração de características foi de aproximadamente 5 horas e 30 minutos, utilizando-se um computador equipado com processador core i5, 16GB de memória RAM e sem placa de vídeo.

Com o vetor de características de todos os documentos gerado, então iniciou-se a etapa de avaliação de desempenho da classificação. A base de dados foi dividida de modo que 80% dos dados foram destinados para ser usado como base de dados de características (treino) e os outros 20% foram designados para a avaliação de desempenho do método (teste).

A metodologia utilizada para classificar os documentos como públicos ou restritos foi baseada no algoritmo KNN. Conforme descrito por Zhai [50], a primeira etapa do KNN é a realização do cálculo de distâncias entre todos os dados, em seguida é escolhido o valor dos K vizinhos mais próximos que serão considerados e por fim é realizado a classificação com base na classe mais frequente entre os K vizinhos mais próximos.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
[2] https://pypi.org/project/imgkit/
[3] https://keras.io/api/applications/vgg/

Figura 4.1: Exemplo de imagem gerada de um documento.



DESPACHO

Da: DFP - Execução da Folha de Pagamento

Para: Unidade/Órgão

1. Encaminhamos o presente processo para que o(a) interessado(a) dê ciência.

2. Quando o servidor retornar ao trabalho, deverá ser anexado Termo de Apresentação, nos termos da INSTRUÇÃO NORMATIVA SGP/SEDGG/ME Nº 34, DE 24 DE MARÇO DE 2021.

3. É importante também, que o(a) Interessado(a) fique ciente de que, caso tenha interesse em manter as Contribuições Previdenciárias ao Regime Próprio de Previdência durante a licença, deverá proceder conforme orientações da IN RFB nº 1332/2013.

Abaixo, um roteiro para que o(a) Interessado(a) possa providenciar os recolhimentos mensais:

1) Acessar o Programa para Cálculo e Emissão de DARF, no site da Receita Federal, no link: https://sicalc.receita.economia.gov.br/sicalc/rapido/contribuinte

2) Deixar marcado Pessoa Física

3) Preencher CPF e data de nascimento

4) Clicar na Verificação de "não sou um robô", e clicar em continuar

5) Informar o código da Receita: 1684, Opção: 1684 – 02 – DI – a partir de 19/12/2002 - CPSS

6) Preencher Período: Informar a data de pagamento da folha do mês de referência;

Exemplo: Folha de Agosto/2022, informar a data 01/09/2022

7) Preencher o Número de Referência: informar o número do processo da licença (observação: a DARF vai alterar os dois últimos dígitos, não tem problema)

8) Data de Vencimento: o sistema preencherá automaticamente com base no período informado, jogando dois dias úteis para frente;

9) Preencher o Valor Principal: valor da CPSS (O valor deverá ser o correspondente ao da previdência descontada da remuneração do último mês, percebida integralmente, antes da licença);

Atenção: Observar reajustes todo mês de Janeiro e regras da EC 103/2019 a partir de 03/2020.

Em caso de dúvidas, consulte a DFP (dfp@ufg.br)

Contribuições que já podem ser feitas:

Abril/2022, data a ser informada no passo 6: 02/05/2022, valor R$ 1503,98 (complementação de valor recolhido em folha de pagamento)

Maio/2022, data a ser informada no passo 6: 01/06/2022, valor R$ 2519,04

Junho/2022, data a ser informada no passo 6: 01/07/2022, valor R$ 2519,04

Julho/2022, data a ser informada no passo 6: 01/08/2022, valor 2519,04

Esse valor vale até 01/01/2023

10)Clicar em calcular

Observação: Para pagamentos em atraso, na hora da geração da DARF o sistema calcula automaticamente o valor dos juros e multas.

11) Seleciona a DARF na Coluna "SEL" e clicar em emitir DARF

12) Efetuar o recolhimento na Rede Bancária autorizada até a data do vencimento;

13) Uma vez recolhida a DARF, o(a) Interessado(a) deverá inserir o comprovante de pagamento no SEI (processo de licença) ou em caso de impossibilidade encaminhar o comprovante de pagamento a esta Diretoria Financeira de Pessoas (dfp@ufg.br), para providências quanto ao recolhimento da Previdência Patronal e os devidos registros no SIAPE.

Rodolfo Gontijo de Souza
Assistente em Administração
Matrícula 2295198

Fonte: Elaborado pelo autor

Com a metodologia de classificação definida, o primeiro passo foi verificar quais documentos são mais semelhantes. Isso foi realizado calculando a distância euclidiana entre as características do documento que está sendo avaliado e todas as características presentes na base de dados de características, dessa forma é possível obter os documentos com maior semelhança.

A definição do valor dos K vizinhos mais próximos foi realizada usando o método *trial-and-error* explicado por Ougiaroglou et al. [33]. O valor foi definido com base em testes iniciando com o número 1 e incrementando de 2 em 2, sempre usando um número ímpar, até chegar ao valor de 15 que apresentou melhor desempenho. Por fim, aplicou-se a mediana para identificar a classe predominante entre os vizinhos mais próximos, adotando o seu resultado como a classificação predita para o documento avaliado.

4.3 Sistema web

Este projeto possibilitou o desenvolvimento de um sistema web destinado a utilização prática do método convencional.

No *back-end*, duas APIs foram desenvolvidas dentro da arquitetura do sistema:

- API "classificacao-documento": Responsável pela realização da classificação de documentos previamente armazenados na base de dados. Recebe o identificador de um documento específico, procede-se com a realização da classificação, entre público ou restrito, e armazena na base de dados o resultado juntamente com outras informações relevantes;
- API "classificacao-texto": Diferentemente da API anterior, esta permite que os usuários ou aplicações submetam um texto livre através de uma requisição HTTP. O sistema realiza o tratamento necessário do texto, executa a classificação e fornece uma resposta contendo os detalhes da classificação efetuada.

Para a interface de usuário (*front-end*), optou-se pela utilização do Django Admin, um componente do *framework* Django que simplifica a criação de áreas administrativas. Esse componente foi personalizado para atender às necessidades específicas do projeto, proporcionando:

- Visualização e gestão de documentos: Os usuários podem acessar a lista de documentos armazenados, visualizar informações detalhadas e executar ações como a solicitação de classificação dos documentos diretamente pela interface;
- Resultados da classificação: Após a realização de uma classificação, são exibidos informações detalhadas sobre o resultado. Isso inclui a classificação final (público ou restrito), o nível de atenção (Crítico, Alto, Médio ou Baixo), o classificador utilizado (Inteligência Artificial (IA) ou as expressões regulares específicas), a indicação da existência de divergência na classificação entre como está classificado no SEI e a classificação do método, e a identificação das expressões regulares atendidas.

O sistema atribui níveis de atenção para cada documento com base nos critérios:

- Documentos que atendem a alguma expressão regular são imediatamente classificados como restrito e com nível de atenção Crítico.
- Caso o documento não satisfaça alguma expressão regular, então a classificação é determinada pelo modelo preditivo que estima a probabilidade do documento ser restrito. Com base nessa probabilidade, o documento é classificado como Restrito se a probabilidade for acima de 50% e como Público se for 50% ou menos. O nível de atenção é atribuído como Baixo quando a probabilidade é menor que 70%, Médio com probabilidade entre 70% e 90% ou Alto quando for maior que 90%.

Visando enriquecer a experiência do usuário e facilitar a gestão documental, metadados foram incorporados aos documentos. Esses metadados incluem o tipo de documento, o número do documento no SEI e se o documento foi usado na fase de treinamento do modelo preditivo.

CAPÍTULO 5

Resultados

Este capítulo apresenta os resultados obtidos através da aplicação dos métodos de classificação de documentos propostos neste estudo. Buscamos investigar o desempenho de diferentes abordagens na tarefa de identificação de documentos contendo informações sensíveis. A avaliação dos modelos foi realizada em diversas configurações para entender a contribuição específica de expressões regulares e do volume de dados na melhoria do desempenho da classificação.

Além disso, apresentamos o resultado da implementação prática do método em um sistema web, descrevendo o *back-end* e *front-end* desenvolvidos com a finalidade de facilitar a interação do usuário com o resultado do trabalho.

5.1 Método convencional

A partir do método proposto na Seção 3.1 e nos experimentos realizados conforme descrito na Seção 4.1, obtemos resultados detalhados abrangendo diversas variações. Isso possibilitou uma investigação ampla da contribuição de cada uma das etapas propostas no método.

Os resultados da avaliação de desempenho foram obtidos nas seguintes variações:

- Modelo sem adição das expressões regulares;
- Modelo com a adição das expressões regulares para identificação de CPF e telefone celular;
- Modelo com a adição das expressões regulares para identificação de CPF, telefone celular e e-mail pessoal.

A acurácia para cada uma dessas variações é detalhada na Tabela 5.1.

A inclusão de expressões regulares para CPF e número de celular, de forma geral, contribuiu positivamente para o desempenho dos modelos, sugerindo que tais informações são indicadores relevantes para a identificação de documentos restritos.

Tabela 5.1: Acurácia dos modelos em cada variação.

Modelo	**Sem expressão regular**	**Com expressão regular de CPF e celular**	**Com expressão regular de CPF, celular e e-mail**
BERT	93,23%	**94,03%**	76,76%
CNN	92,81%	92,33%	75,64%
LSTM	90,14%	86,14%	72,34%
BaggingClassifier	92,59%	92,70%	76,23%
ExtraTreesClassifier	**93,60%**	93,76%	**77,24%**
RandomForestClassifier	92,91%	93,07%	76,76%

Fonte: Elaborado pelo autor

Entretanto, observou-se um padrão significativo referente ao desempenho dos modelos quando incorporado a expressão regular que realiza a identificação de e-mails pessoais. A aplicação dessa expressão regular não resultou na melhoria esperada no desempenho dos modelos, ao contrário, os resultados indicam que a inclusão dessa expressão regular degradou o desempenho da classificação em todos os modelos testados. Essa diminuição da acurácia indica que a metodologia aplicada para a identificação de e-mails pode não ser a ideal ou que a presença de e-mails nos documentos não constitui um critério discriminatório efetivo para a tarefa de identificação de documentos restritos. Este achado ressalta a importância de uma avaliação minuciosa dos critérios utilizados no pré-processamento dos dados, especialmente em contextos onde a precisão da classificação é crítica.

Apesar de ser modesto o aumento na acurácia alcançado pelas expressões regulares de CPF e número de celular, é importante ressaltar a significativa vantagem dessa abordagem quanto à eficiência computacional. O emprego das expressões regulares antes da etapa de predição pelos modelos de aprendizado de máquina configura um método eficaz e menos oneroso computacionalmente para a identificação de documentos contendo informações sensíveis. Este procedimento permite que apenas documentos que não correspondem aos padrões identificados pelas expressões regulares avancem para a etapa de classificação mais complexa e computacionalmente mais custosa. Desta forma, evita-se o emprego desnecessário de recursos computacionais em casos onde a presença de informação sensível é evidente, otimizando o processo de classificação e contribuindo para uma gestão mais ágil e eficiente dos recursos disponíveis. Assim, mesmo que o ganho em precisão seja modesto, o benefício no custo computacional representa uma vantagem considerável, especialmente em cenários com grandes volumes de documentos a serem processados.

Os resultados apresentados nas Tabelas 5.2, 5.3 e 5.4 proporcionam uma visão abrangente do desempenho dos modelos de aprendizado de máquina na tarefa de classificação de documentos em cada uma das variações. De forma geral, os modelos apresenta-

ram um desempenho robusto tendo todos alcançado acurácia, área sob a curva ROC, F1 *score*, precisão e *recall* acima de 90%.

Tabela 5.2: Resultado dos avaliadores de desempenho dos modelos sem utilização das expressões regulares.

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1 Score**	**Área Sob a Curva ROC**
BERT	93,23%	92,79%	93,81%	93,27%	96,95%
CNN	92,81%	91,79%	**94,14%**	92,91%	97,50%
LSTM	90,14%	90,46%	89,76%	90,07%	96,20%
BaggingClassifier	92,59%	91,62%	93,81%	92,68%	96,85%
ExtraTreesClassifier	**93,60%**	**93,57%**	93,71%	**93,61%**	**97,82%**
RandomForestClassifier	92,91%	92,53%	93,39%	92,94%	97,58%

Fonte: Elaborado pelo autor

Tabela 5.3: Resultado dos avaliadores de desempenho dos modelos com expressão regular de CPF e celular.

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1 Score**	**Área Sob a Curva ROC**
BERT	**94,03%**	**93,68%**	**94,56%**	**94,08%**	97,24%
CNN	92,33%	91,18%	93,82%	92,44%	97,32%
LSTM	86,14%	82,76%	91,58%	86,88%	87,60%
BaggingClassifier	92,70%	91,38%	94,35%	92,83%	96,89%
ExtraTreesClassifier	93,76%	93,41%	94,24%	93,80%	**97,70%**
RandomForestClassifier	93,07%	92,29%	94,03%	93,13%	97,45%

Fonte: Elaborado pelo autor

Tabela 5.4: Resultado dos avaliadores de desempenho dos modelos com expressão regular de CPF, celular e e-mail.

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1 Score**	**Área Sob a Curva ROC**
BERT	76,76%	69,30%	96,59%	80,65%	70,24%
CNN	75,64%	68,12%	96,91%	79,96%	70,84%
LSTM	72,34%	65,38%	95,21%	77,51%	64,09%
BaggingClassifier	76,23%	68,46%	**97,65%**	80,47%	76,77%
ExtraTreesClassifier	**77,24%**	**69,56%**	97,34%	**81,10%**	**76,78%**
RandomForestClassifier	76,76%	69,08%	97,23%	80,74%	74,76%

Fonte: Elaborado pelo autor

O modelo de rede neural profunda BERT quando combinado com expressões regulares para CPF e celular apresentou o melhor resultado, alcançando o melhor valor na acurácia, precisão, *recall* e F1 *score* entre os modelos avaliados. Na métrica área sob a curva ROC obteve desempenho satisfatório, aproximando-se do melhor desempenho registrado nesse avaliador. Portanto, a escolha do BERT como o melhor modelo nesta variação é atribuída à sua capacidade avançada de entender o contexto e as nuances linguísticas dos textos.

Já o CNN, tradicionalmente utilizado nas tarefas de processamento de imagens, demonstrou ser altamente eficaz também no processamento de texto. A inclusão de expressões regulares de CPF e celular resultou em uma leve diminuição dos resultados dos avaliadores de desempenho, indicando que essas expressões regulares podem não ser tão relevantes para o desempenho da CNN quanto para outros modelos.

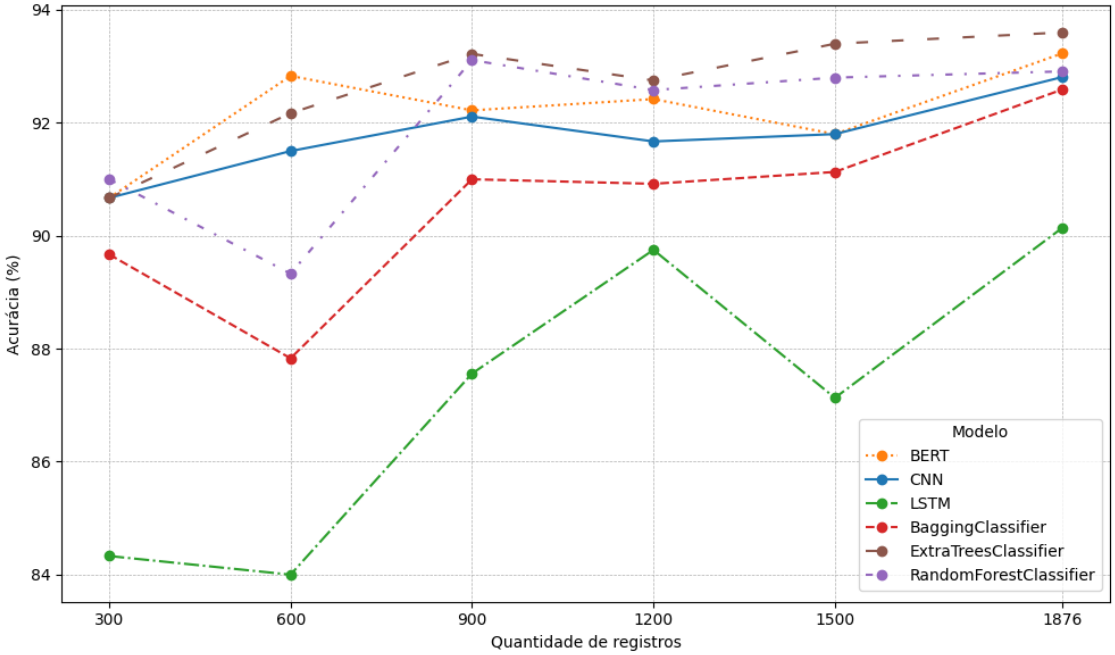
De forma semelhante ao CNN, a adição das expressões regulares de CPF e celular degradou o desempenho do LSTM, só que de forma considerável se comparado ao desempenho do modelo sem as expressões regulares.

Dentro os modelos tradicionais de aprendizado de máquina, o *Extra Trees Classifier* se destacou com o melhor desempenho, estando próximo ao desempenho alcançado pelo BERT. *Random Forest Classifier* e *Bagging Classifier* também obtiveram desempenhos satisfatórios, tendo ligeiras melhorias ao aplicar as expressões regulares de CPF e celular.

Os modelos de aprendizado de máquina necessitam de uma grande quantidade de dados para alcançarem um ótimo desempenho. Esse requisito é crucial pois permite que os modelos aprendam uma ampla variedade de padrões e nuances contidos nos dados, melhorando assim sua capacidade de generalização e precisão em tarefas de classificação. A base de dados utilizada para a avaliação dos modelos de classificação de documentos pode ser considerada relativamente pequena, o que despertou a necessidade de investigar a influência do volume de dados no desempenho dos modelos.

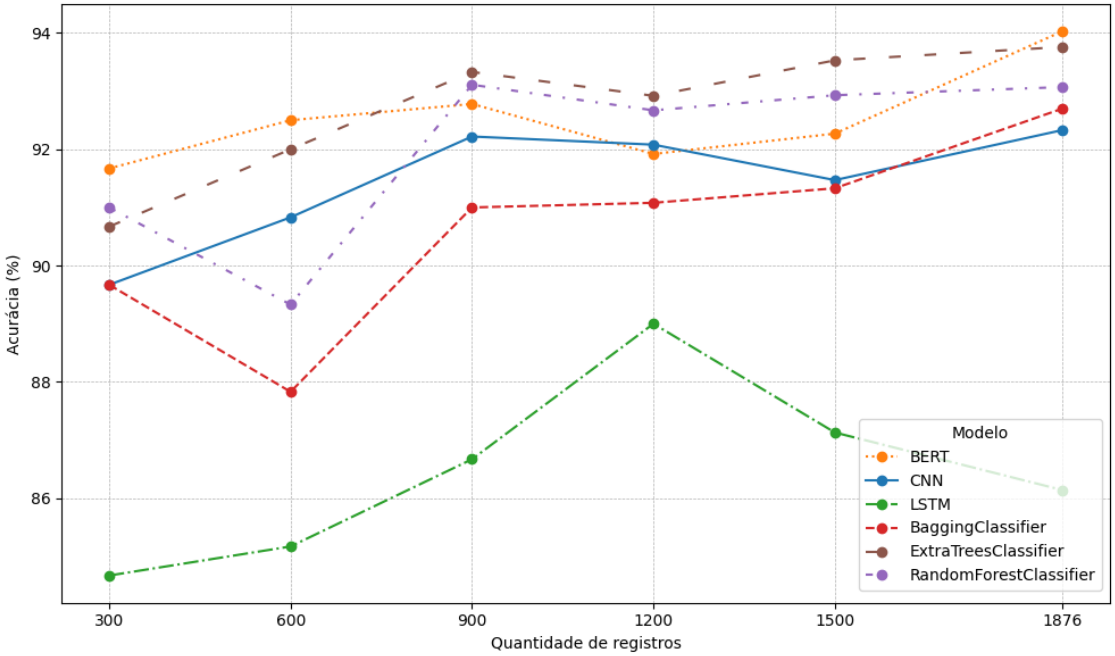
Para explorar essa relação e avaliar como o aumento da base de dados impacta o desempenho dos modelos, foram realizadas análises comparativas do desempenho com diferentes quantidades de registros, especificamente em conjuntos contendo 300, 600, 900, 1200, 1500 e 1876 registros. Essa estratégia permitiu observar a evolução do desempenho, particularmente avaliando a acurácia, à medida que a quantidade de registros aumenta. As Figuras 5.1, 5.2 e 5.3 ilustram essa evolução para os modelos avaliados, destacando como a expansão do volume de dados contribui significativamente para a melhoria do desempenho. Essas análises são fundamentais para compreender a relação entre o volume de dados disponíveis para treinamento e a eficiência dos modelos em tarefas de classificação, oferecendo percepções importantes para otimizar estratégias de treinamento em pesquisas futuras e aplicações práticas.

Figura 5.1: Evolução da acurácia dos modelos sem utilizar expressão regular.



Fonte: Elaborado pelo autor

Figura 5.2: Evolução da acurácia dos modelos com expressão regular de CPF e celular.



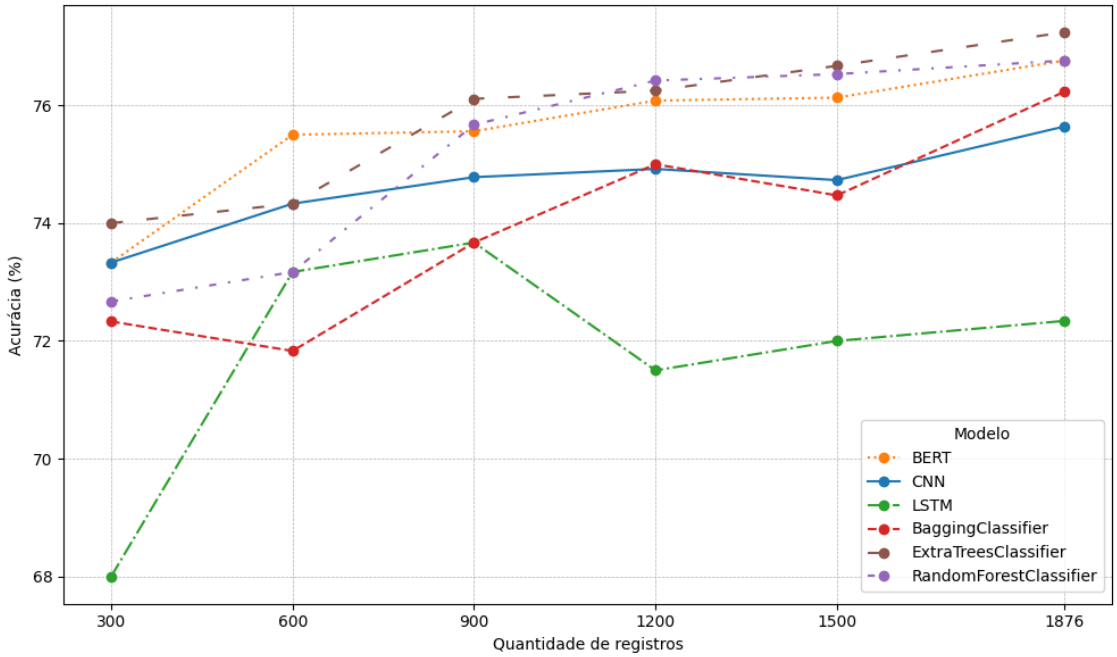
Fonte: Elaborado pelo autor

Tabela 5.5: Comparação do desempenho do método convencional com a literatura.

Trabalho/Método	**Acurácia**	**Natureza dos dados**
Trabalho de Gambarelli et al. [13]	98%	Textos coletados na internet
Trabalho de Patil et al. [35]	98%	Mensagens SMS
Trabalho de McDonald et al. [25]	97%	Documentos governamentais do Reino Unido e dos Estados Unidos
Método convencional: BERT e expressões regulares	94%	Documentos da UFG
Trabalho de Geetha et al. [15]	66%	Mensagens postadas no Twitter

Fonte: Elaborado pelo autor

Figura 5.3: Evolução da acurácia dos modelos com expressão regular de CPF, celular e e-mail.



Fonte: Elaborado pelo autor

A Tabela 5.5 oferece uma visão comparativa do desempenho, utilizando a acurácia, entre os trabalhos encontrados na literatura e o método convencional empregando BERT e expressões regulares. A comparação revela que o método convencional alcança uma acurácia significativa de 94%, destacando-se como uma abordagem relevante na tarefa em questão.

5.2 Método alternativo

Para mensurar o desempenho do método alternativo, foram utilizados os avaliadores de desempenho descritos na Seção 2.4.3. Os resultados obtidos para cada um dos

avaliadores de desempenho são detalhados na Tabela 5.6.

Tabela 5.6: Resultado dos avaliadores de desempenho no método alternativo.

Avaliador de desempenho	**Resultado**
Acurácia	83,52%
F1 Score	83,30%
Precisão	86,20%
Recall	80,58%
Área Sob a Curva ROC	83,58%

Fonte: Elaborado pelo autor

Adicionalmente, foi gerada uma matriz de confusão (Figura 5.4) que revelou resultados promissores para o problema. Levando em consideração que o objetivo da UFG é dar publicidade aos documentos públicos, o método alternativo alcançou uma taxa de acerto de 87% na classificação de documentos públicos. Ao realizar a comparação entre a taxa de acerto alcançada com os resultados dos trabalhos relacionados listados na Tabela 2.4, observamos que o método alternativo apresenta um resultado satisfatório, conforme evidenciado na Tabela 5.7.

Figura 5.4: Matriz de confusão.

Fonte: Elaborado pelo autor

5.3 Sistema web

O desenvolvimento do sistema web apresenta um marco significativo na aplicação prática do método convencional para a classificação de documentos sensíveis. Para a realização da classificação foi utilizado o *pipeline* do melhor resultado apresentado na Se-

Tabela 5.7: Comparação do desempenho do método alternativo com a literatura.

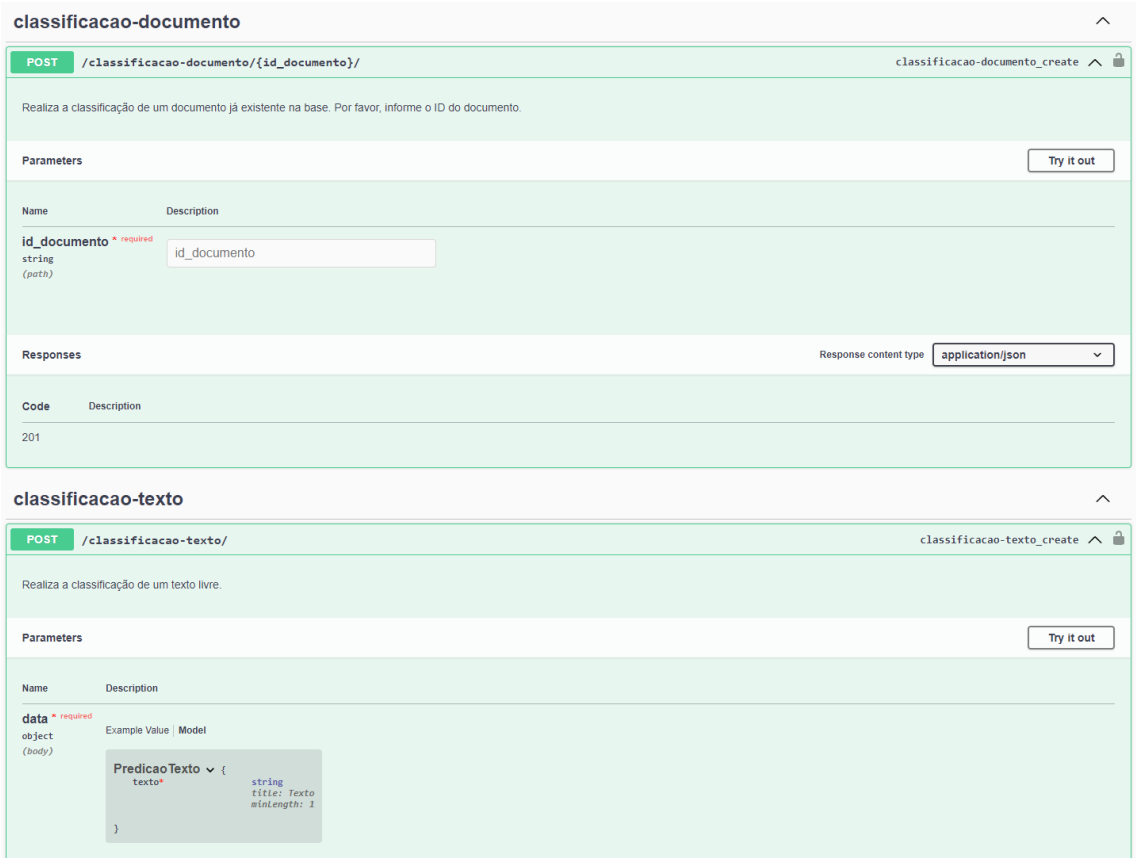
Trabalho/Método	**Acurácia**	**Natureza dos dados**
Trabalho de Gambarelli et al. [13]	98%	Textos coletados na internet
Trabalho de Patil et al. [35]	98%	Mensagens SMS
Trabalho de McDonald et al. [25]	97%	Documentos governamentais do Reino Unido e dos Estados Unidos
Método convencional: BERT e expressões regulares	94%	Documentos da UFG
Método alternativo: classificação de documento público	87%	Documentos da UFG
Método alternativo: geral	83%	Documentos da UFG
Trabalho de Geetha et al. [15]	66%	Mensagens postadas no Twitter

Fonte: Elaborado pelo autor

ção do método convencional (5.1): modelo BERT combinado com as expressões regulares que identificam CPF e telefone celular.

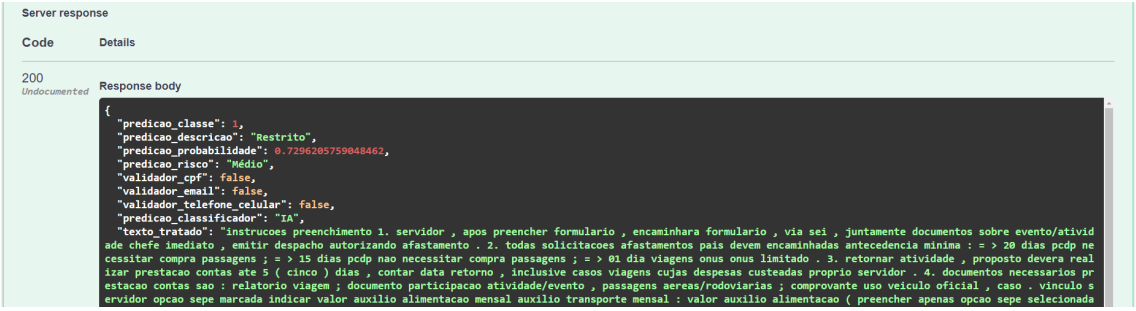
O *back-end* do sistema web é composto por duas APIs, dedicadas à classificação de documentos armazenados na base de dados e à submissão de textos livres para classificação imediata, respectivamente. Essa estrutura possibilita uma fácil integração entre aplicações diversas, permitindo que as instituições possam utilizar os resultados deste trabalho em suas aplicações próprias. A Figura 5.6 demonstra a interface de interação com as APIs do *back-end* e a Figura apresenta um exemplo de retorno de classificação de um texto livre.

Figura 5.5: Tela para interação com as APIs do back-end.



Fonte: Elaborado pelo autor

Figura 5.6: Exemplo de retorno da API de classificação de texto livre.

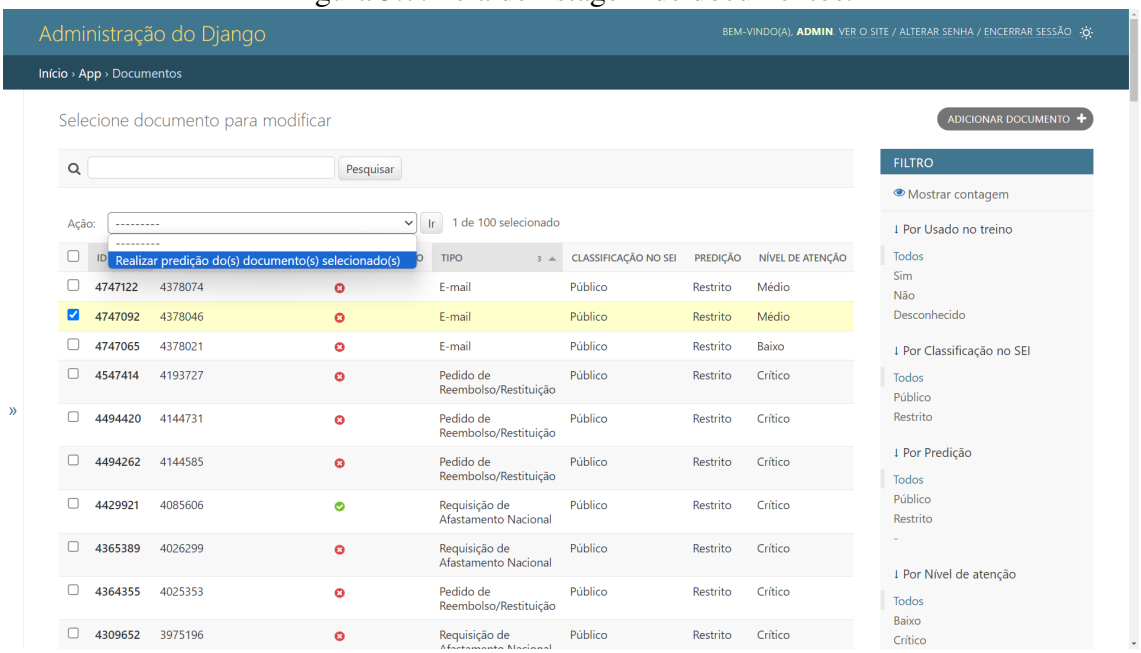


Fonte: Elaborado pelo autor

Para a interface do usuário (*front-end*), a escolha do Django Admin como base para a criação de áreas administrativas se mostrou acertada. Essa decisão permitiu uma fácil personalização do sistema para atender as necessidades específicas do projeto, oferecendo funcionalidades como a visualização e gestão de documentos armazenados, além da apresentação detalhada dos resultados das classificações realizadas.

A Figura 5.7 exibe a tela que oferece uma visão geral dos documentos armazenados, permitindo aos usuários uma navegação fácil e a possibilidade de realizar ações diretas, como solicitar classificação para os documentos selecionados.

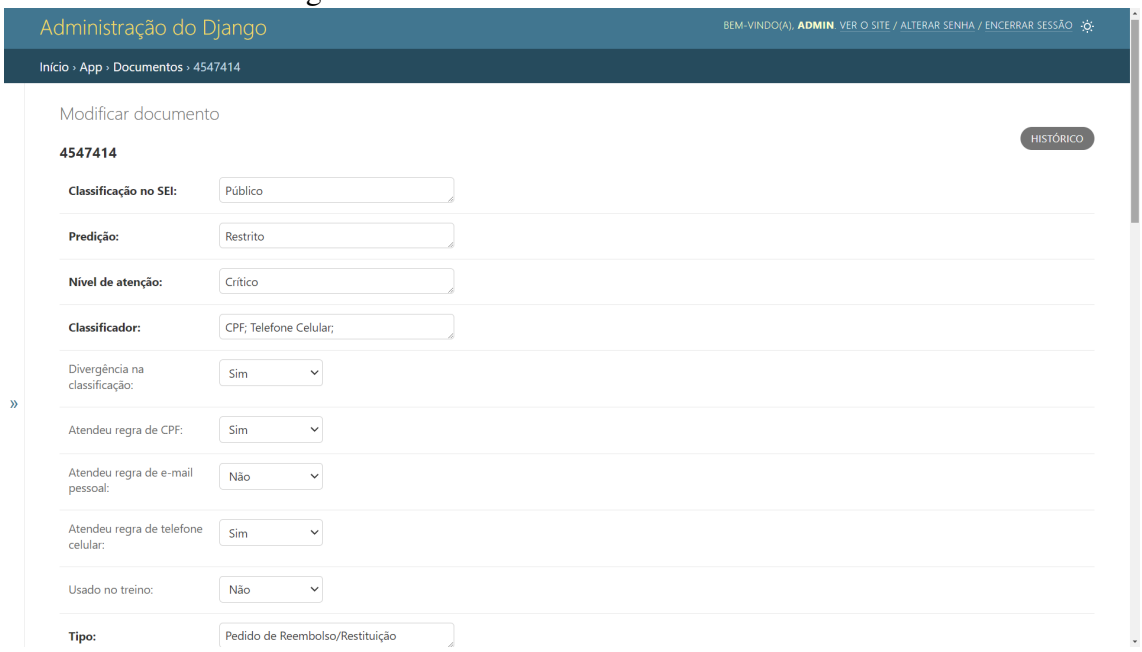
Figura 5.7: Tela de listagem de documentos.



Fonte: Elaborado pelo autor

Já a Figura 5.8 apresenta a tela com informações detalhadas sobre um documento específico, incluindo o resultado da sua classificação, o nível de atenção atribuído e os metadados relevantes.

Figura 5.8: Tela de detalhe de um documento.



Fonte: Elaborado pelo autor

CAPÍTULO 6

Conclusão e trabalhos futuros

Este trabalho teve como proposta o desenvolvimento e avaliação de métodos para a classificação de documentos do Sistema Eletrônico de Informações (SEI) da Universidade Federal de Goiás (UFG) como públicos ou restritos, abordando um desafio relevante demandado pelo Tribunal de Contas da União (TCU) que é a publicização dos documentos. O desenvolvimento de uma aplicação web para uso dos gestores documentais da UFG foi também uma contribuição significativa deste projeto, facilitando a aplicação prática dos métodos desenvolvidos.

O projeto foi guiado pelo processo de descoberta de conhecimento em bases de dados (KDD). A estratégia de utilizar os documentos gerados pelo CIDARQ como base de dados para o treinamento dos modelos de aprendizado de máquina mostrou-se eficaz, dada a expertise da equipe do órgão na avaliação e classificação de documentos. Entretanto, o desafio apresentado pelo desbalanceamento acentuado entre as classes de documentos demandou a implementação de técnicas de balanceamento para assegurar a imparcialidade dos modelos.

Dois métodos foram propostos: o método convencional e o método alternativo. O primeiro baseou-se na utilização de expressões regulares e técnicas de aprendizado de máquina para a classificação dos documentos, enquanto o segundo, destinado a cenários onde a coleta de textos é inviável, focou na utilização de imagens dos documentos.

Os experimentos conduzidos revelaram que, de modo geral, a inclusão de expressões regulares para CPF e número de celular contribuiu positivamente para o desempenho dos modelos, sugerindo a relevância dessas informações como indicadores para a identificação de documentos restritos. No entanto, a inclusão da expressão regular para identificação de e-mails pessoais resultou em uma queda no desempenho dos modelos, evidenciando a complexidade na utilização dessa informação como critério para a identificação de documentos restritos.

A análise dos resultados indicou que o modelo de rede neural BERT combinado com expressões regulares para identificação de CPF e números de celular apresentou o melhor desempenho, demonstrando a superioridade dessa abordagem na compreensão do contexto e das nuances linguísticas dos textos. No entanto, a avaliação dos modelos tra-

dicionais de classificação também apresentou resultados interessantes, que se aproximam dos resultados obtidos pelos modelos de redes neurais. Tal achado demonstra a viabilidade de utilização, sem redução significativa de desempenho, dos modelos tradicionais em cenários onde o custo computacional seja um fator crítico.

Por outro lado, o método alternativo, apesar de não alcançar o mesmo desempenho do método convencional, mostrou-se uma alternativa viável para a classificação de documentos em cenários desafiadores, com uma acurácia satisfatória que superou outro trabalho da literatura.

A comparação do desempenho entre os métodos propostos com os trabalhos existentes na literatura evidenciou a qualidade das abordagens desenvolvidas neste projeto, particularmente do método convencional que se mostrou competitivo com os melhores resultados da literatura. No entanto, observa-se que Gambarelli et al. [13], Patil et al. [35] e McDonald et al. [25] conseguiram obter um desempenho ainda superior, o que sugere que, embora o uso do BERT seja promissor, há espaço para otimizações que podem levar a melhorias adicionais. Importante salientar que embora os outros trabalhos tenham apresentado acurácias ligeiramente superiores, a natureza dos dados sobre os quais esses métodos foram avaliados difere daquela utilizada neste trabalho.

Em conclusão, este trabalho apresenta uma valiosa contribuição para a literatura na área de classificação de documentos, oferecendo métodos robustos e eficazes que podem ser aplicados não apenas na UFG, mas também em outras instituições que enfrentam desafios semelhantes. Ao abordar a necessidade imposta pelo TCU para a publicização dos documentos, este trabalho contribui para a transparência governamental e o acesso à informação, aspectos fundamentais para o fortalecimento da democracia e a promoção da cidadania. Ao aprimorar a classificação correta de documentos sensíveis, este trabalho também contribui para a proteção da privacidade das pessoas e a segurança de informações críticas, abordando um desafio real enfrentado por inúmeras organizações públicas e privadas.

Como trabalhos futuros, no método alternativo recomenda-se expandir a base de dados para contemplar todos os 3 milhões de documentos existentes no SEI da UFG. Além disso, propõe-se a avaliação de outras técnicas de CBIR além da VGG-16, visando explorar alternativas que possam oferecer resultados ainda melhores.

Quanto ao método convencional, sugere-se a exploração de novas abordagens para a identificação de e-mails pessoais, o uso de reconhecimento de entidade nomeada (NER) para enriquecer o pré-processamento e a análise de novas técnicas de redes neurais. Adicionalmente, propõe-se a generalização do método convencional para que possa permitir a classificação de documentos em categorias além de público ou restrito. Isso inclui, por exemplo, a capacidade de categorizar documentos por tipo ou natureza, como atas de reunião, relatórios técnicos, pareceres, entre outros. Essa expansão do método permitirá

uma aplicabilidade mais ampla, beneficiando não somente as atividades relacionadas à transparência, mas também outras necessidades oferecendo uma flexibilidade maior para a gestão documental.

6.1 Publicação

Durante o desenvolvimento deste trabalho, os resultados alcançados pelo método alternativo foram divulgados na seguinte publicação:

CARVALHO, Rogério Rodrigues; MACEDO, Sanderson Oliveira de; OLIVEIRA, Leandro Luis Galdino; COSTA, Ronaldo Martins da. Classificação de documentos sensíveis da administração pública utilizando CBIR. In: SIMPÓSIO BRASILEIRO DE SEGURANÇA DA INFORMAÇÃO E DE SISTEMAS COMPUTACIONAIS (SBSEG), 23. , 2023, Juiz de Fora/MG. Anais. Porto Alegre: Sociedade Brasileira de Computação, 2023 . p. 564-569. DOI: https://doi.org/10.5753/sbseg.2023.233683.

CAPÍTULO 7

Agradecimentos

Agradecemos a Universidade Federal de Goiás (UFG), por meio da Secretaria de Tecnologia e Informação, pelo total apoio na execução do trabalho e na disponibilização dos dados requisitados.

Esta pesquisa foi apoiada pelo Laboratório Multiusuário de Computação de Alto Desempenho da Universidade Federal de Goiás (LaMCAD/UFG).

Referências Bibliográficas

[1] BATTAGLIA, E.; BIOGLIO, L.; PENSA, R. G. **Towards content sensitivity analysis**. In: *Advances in Intelligent Data Analysis XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings 18*, p. 67–79. Springer, 2020.

[2] BIOGLIO, L.; PENSA, R. G. **Analysis and classification of privacy-sensitive content in social media posts**. *EPJ Data Science*, 11(1):12, 2022.

[3] BRASIL. **Lei nº 12.527, de 18 de novembro de 2011**. *Diário Oficial da República Federativa do Brasil*, 2011.

[4] BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. *Diário Oficial da República Federativa do Brasil*, 2018.

[5] CARVALHO, R. R. **Sistema de auto machine learning aplicado à predição de desempenho de ações para o mercado de valores**. *Trabalho de Conclusão de Curso. Especialização em Design de Sistemas e Soluções de Business Intelligence, Instituto de Informática, Universidade Federal de Goiás*, p. 52, 2018.

[6] COOLS, A.; BELARBI, M. A.; MAHMOUDI, S. A. **A comparative study of reduction methods applied on a convolutional neural network**. *Electronics*, 11(9):1422, 2022.

[7] COSTA, R.; JUNIOR, E.; NUNES, F.; OLIVEIRA, L.; SALVINI, R. **Analysis of techniques of the content-based image retrieval to construct an information system of the computer-aided diagnosis**. 2015.

[8] DA ECONOMIA E CONTROLADORIA GERAL DA UNIÃO, M. **Orientação conjunta nº 1/2021/ME/CGU**. Acessado em março de 2024. Disponível em: https://www.gov.br/economia/pt-br/acesso-a-informacao/sei/arquivos/orientacao-conjunta-no-1-2021-me-cgu/view.

[9] DESORDI, D.; DELLA BONA, C. **A inteligência artificial e a eficiência na administração pública**. *Revista de Direito*, 12(2):1–22, 2020.

[10] DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. **Bert: Pre-training of deep bidirectional transformers for language understanding**. *arXiv preprint arXiv:1810.04805*, 2018.

[11] DOGRA, V.; VERMA, S.; CHATTERJEE, P.; SHAFI, J.; CHOI, J.; IJAZ, M. F.; OTHERS. **A complete process of text classification system using state-of-the-art NLP models**. *Computational Intelligence and Neuroscience*, 2022, 2022.

[12] FAYYAD, U. **Knowledge discovery in databases: An overview**. In: *International Conference on Inductive Logic Programming*, p. 1–16. Springer, 1997.

[13] GAMBARELLI, G.; GANGEMI, A.; TRIPODI, R. **Is your model sensitive? SPEDAC: A new resource for the automatic classification of sensitive personal data**. *IEEE Access*, 11:10864–10880, 2023.

[14] GASPARETTO, A.; MARCUZZO, M.; ZANGARI, A.; ALBARELLI, A. **A survey on text classification algorithms: From text to predictions**. *Information*, 13(2):83, 2022.

[15] GEETHA, R.; KARTHIKA, S.; KUMARAGURU, P. **'Will I regret for this tweet?'—Twitter user's behavior analysis system for private data disclosure**. *The Computer Journal*, 65(2):275–296, 2022.

[16] GONEN, H.; GOLDBERG, Y. **Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them**. *arXiv preprint arXiv:1903.03862*, 2019.

[17] GRAVES, A.; GRAVES, A. **Long short-term memory**. *Supervised sequence labelling with recurrent neural networks*, p. 37–45, 2012.

[18] HARTMANN, J.; HUPPERTZ, J.; SCHAMP, C.; HEITMANN, M. **Comparing automated text classification methods**. *International Journal of Research in Marketing*, 36(1):20–38, 2019.

[19] HUTCHINSON, B.; PRABHAKARAN, V.; DENTON, E.; WEBSTER, K.; ZHONG, Y.; DENUYL, S. **Social biases in nlp models as barriers for persons with disabilities**. *arXiv preprint arXiv:2005.00813*, 2020.

[20] JIANG, H.; CHEN, C.; WU, S.; GUO, Y. **Classification of medical sensitive data based on text classification**. In: *2019 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, p. 1–2. IEEE, 2019.

[21] KOBAYASHI, V. B.; MOL, S. T.; BERKERS, H. A.; KISMIHOK, G.; DEN HARTOG, D. N. **Text classification for organizational researchers: A tutorial**. *Organizational research methods*, 21(3):766–799, 2018.

[22] KOWSARI, K.; JAFARI MEIMANDI, K.; HEIDARYSAFA, M.; MENDU, S.; BARNES, L.; BROWN, D. **Text classification algorithms: A survey**. *Information*, 10(4):150, 2019.

[23] LI, Q.; PENG, H.; LI, J.; XIA, C.; YANG, R.; SUN, L.; YU, P. S.; HE, L. **A survey on text classification: From traditional to deep learning**. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41, 2022.

[24] LÖBNER, S.; TESFAY, W. B.; BRACAMONTE, V.; NAKAMURA, T. **Systematizing the state of knowledge in detecting privacy sensitive information in unstructured texts using machine learning**. In: *2023 20th Annual International Conference on Privacy, Security and Trust (PST)*, p. 1–7. IEEE, 2023.

[25] MCDONALD, G.; MACDONALD, C.; OUNIS, I. **Using part-of-speech n-grams for sensitive-text classification**. In: *Proceedings of the 2015 International conference on the theory of information retrieval*, p. 381–384, 2015.

[26] METSIS, V.; ANDROUTSOPOULOS, I.; PALIOURAS, G. **Spam filtering with naive bayes-which naive bayes?** In: *CEAS*, volume 17, p. 28–69. Mountain View, CA, 2006.

[27] MINAEE, S.; KALCHBRENNER, N.; CAMBRIA, E.; NIKZAD, N.; CHENAGHLU, M.; GAO, J. **Deep learning–based text classification: a comprehensive review**. *ACM computing surveys (CSUR)*, 54(3):1–40, 2021.

[28] MONARD, M. C.; BARANAUSKAS, J. A. **Conceitos sobre aprendizado de máquina**. *Sistemas inteligentes-Fundamentos e aplicações*, 1(1):32, 2003.

[29] NEERBEK, J.; ASSENT, I.; DOLOG, P. **Detecting complex sensitive information via phrase structure in recursive neural networks**. In: *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22*, p. 373–385. Springer, 2018.

[30] NISSIM, M.; VAN NOORD, R.; VAN DER GOOT, R. **Fair is better than sensational: Man is to doctor as woman is to doctor**. *Computational Linguistics*, 46(2):487–497, 2020.

[31] OCCHIPINTI, A.; ROGERS, L.; ANGIONE, C. **A pipeline and comparative study of 12 machine learning models for text classification**. *Expert Systems with Applications*, 201:117193, 2022.

[32] O'SHEA, K.; NASH, R. **An introduction to convolutional neural networks**. *arXiv preprint arXiv:1511.08458*, 2015.

[33] OUGIAROGLOU, S.; EVANGELIDIS, G. **Dealing with noisy data in the context of k-nn classification**. In: *Proceedings of the 7th Balkan Conference on Informatics Conference*, p. 1–4, 2015.

[34] PATIL, D.; LOKARE, R.; PATIL, S. **An overview of text representation techniques in text classification using deep learning models**. In: *2022 3rd International Conference for Emerging Technology (INCET)*, p. 1–4. IEEE, 2022.

[35] PATIL, D.; LOKARE, R.; PATIL, S. **Privacy preserving document classification using convolution neural network-a deep learning approach**. In: *2022 5th International Conference on Advances in Science and Technology (ICAST)*, p. 215–219. IEEE, 2022.

[36] PEREIRA, A. J. Z. F.; SEVERO, L. L.; LANGOSKI, D. T. **Lei geral de proteção de dados no brasil: Breve análise sobre o direito à privacidade**. *Anais do Salão Internacional de Ensino, Pesquisa e Extensão*, 13(3), 2021.

[37] PRADE, A. M.; PEREZ, C. B. **A importância da gestão documental no contexto do acesso aos documentos e difusão dos arquivos**. *ÁGORA: Arquivologia em debate*, 27(54):226–253, 2017.

[38] REAL, L. G. D. S. **Uso de múltiplos descritores com condições de contorno e visualização hierárquica em cbir**. 2017.

[39] RIBEIRO, M. X. **Suporte a sistemas de auxílio ao diagnóstico e de recuperação de imagens por conteúdo usando mineração de regras de associação**. PhD thesis, Universidade de São Paulo, 2008.

[40] ROSSI, R. G. **Classificaçao automática de textos por meio de aprendizado de máquina baseado em redes**. PhD thesis, Universidade de São Paulo, 2016.

[41] SARAIVA, G. A.; GONÇALVES, A. D. A. C.; PINTO, R. A.; SOUTO, E.; DOS SANTOS, E. M. **Fusão de dados de raio-x com dados clínicos para detectar covid-19**. *Journal of Health Informatics*, 12, 2020.

[42] SOUSA, S.; KERN, R. **How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing**. *Artificial Intelligence Review*, 56(2):1427–1492, 2023.

[43] SWEENEY, C.; NAJAFIAN, M. **A transparent framework for evaluating unintended demographic bias in word embeddings**. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1662–1667, 2019.

[44] TAN, Y. C.; CELIS, L. E. **Assessing social and intersectional biases in contextualized word representations**. *Advances in neural information processing systems*, 32, 2019.

[45] THANGARAJ, M.; SIVAKAMI, M. **Text classification techniques: A literature review**. *Interdisciplinary journal of information, knowledge, and management*, 13:117, 2018.

[46] UFG. **Painel de indicadores do SEI-UFG**. Acessado em março de 2024. Disponível em: https://analisa.ufg.br/p/40356-sei-ufg.

[47] UFG. **Sobre o CIDARQ**. Acessado em março de 2024. Disponível em: https://cidarq.ufg.br/p/22834-sobre-o-cidarq.

[48] VIEIRA, C. D. **A importância da transparência e o acesso à informação para fiscalização e abrangência do governo municipal**. 2018.

[49] WU, H.; LIU, Y.; WANG, J. **Review of text classification methods on deep learning**. *Comput. Mater. Contin*, 63(3):1309–1321, 2020.

[50] ZHAI, H. **Improving KNN algorithm efficiency based on PCA and KD-tree**. In: *2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)*, p. 83–87. IEEE, 2022.

[51] ZHAO, J.; MUKHERJEE, S.; HOSSEINI, S.; CHANG, K.-W.; AWADALLAH, A. H. **Gender bias in multilingual embeddings and cross-lingual transfer**. *arXiv preprint arXiv:2005.00699*, 2020.

APÊNDICE A

Códigos-fonte

Código A.1 Pré-processamento dos documentos: conversão de html, normalização de texto e remoção de acentuação.

```
import pandas as pd
from bs4 import BeautifulSoup

dados = pd.read_csv('dados/dados_cidarq_20240213.csv', sep=';',
                    encoding = 'latin-1')

# Conversão de HTML para texto
dados['conteudo_documento_texto'] = dados['conteudo_documento'].apply(
    lambda x: BeautifulSoup(x, 'html.parser').get_text(
        separator=' ', strip=True
        )
    )

# Todas as palavras em minúsculo
dados['conteudo_documento_texto_tratado'] = \
    dados['conteudo_documento_texto'].apply(lambda x: x.lower())

# Remover acentuação
dados['conteudo_documento_texto_tratado'] = \
    dados['conteudo_documento_texto_tratado'].str.normalize('NFKD') \
        .str.encode('ascii', errors='ignore').str.decode('utf-8')
```

Código A.2 Pré-processamento dos documentos: remoção de metadados.

```
import re

# Função que remove metadados
def remover_metadados(row):

    # Remover número do processo
    pattern = re.escape(str(row['numero_processo_formatado']))
    texto = re.sub(pattern, '', row['conteudo_documento_texto_tratado'])

    # Remover número do documento
    pattern = re.escape(str(row['numero_sei']))
    texto = re.sub(pattern, '', texto)

    # Remover e-mail do SEI
    email_completo = r'ufg/noreply <noreply.sei@sistemas.ufg.br>'
    email = r'noreply.sei@sistemas.ufg.br'
    texto = re.sub(email_completo, '', texto)
    texto = re.sub(email, '', texto)

    # Remover termo sei/ufg
    sei = 'sei/ufg'
    texto = re.sub(sei, '', texto)

    return texto

# Aplicação da remoção de metadados
dados['conteudo_documento_texto_tratado'] = dados.apply(
    remover_metadados, axis=1
    )
```

Código A.3 Pré-processamento dos documentos: remoção de stop words.

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

# Remover stopwords
stop_words = set(stopwords.words('portuguese'))

def remove_stopwords(text):
    word_tokens = word_tokenize(text)
    filtered_text = ' '.join(
        [word for word in word_tokens if word.lower() not in stop_words]
        )
    return filtered_text

dados['conteudo_documento_texto_tratado'] = \
    dados['conteudo_documento_texto_tratado'].apply(remove_stopwords)
```

Código A.4 Geração de imagem do documento.

```
import pandas as pd
import imgkit

# Ler csv contendo o texto com HTML dos documentos
pub = pd.read_csv(
    'dados\\sei_documento_conteudo_all_publico_10000.csv',
    sep=';', encoding = 'latin-1'
    )

i = 1
# Percorrer cada linha
for index, row in pub.iterrows():

    # Definir nome da imagem como: numero_do_id_documento.jpg
    name_file = f'dados\\img\\publico\\{row["id_documento"]}.jpg'

    print(f'Execução {i}: {name_file}')

    try:
        # Coletar o campo contendo o html do documento
        html = row['conteudo']
        # Converter o html para imagem e salvar
        imgkit.from_string(html, name_file)
    except Exception as e:
        print(e) # Caso der erro, exibir o erro
        pass
    finally:
        i = i + 1
```

Código A.5 Classe para utilizar VGG16 e retornar as características de uma imagem.

```
class FeatureExtractor:
    # Fonte:
    # https://github.com/matsui528/sis/blob/master/feature_extractor.py
    def __init__(self):
        base_model = VGG16(weights='imagenet')
        self.model = Model(inputs=base_model.input,
                           outputs=base_model.get_layer('fc1').output)

    def extract(self, img):
        """
        Extract a deep feature from an input image
        Args:
            img: from PIL.Image.open(path) or
                tensorflow.keras.preprocessing.image.load_img(path)
        Returns:
            feature (np.ndarray): deep feature with the shape=(4096, )
        """
        # VGG must take a 224x224 img as an input
        img = img.resize((224, 224))
        # Make sure img is color
        img = img.convert('RGB')
        # To np.array. Height x Width x Channel. dtype=float32
        x = image.img_to_array(img)
        # (H, W, C)->(1, H, W, C),
        # where the first elem is the number of img
        x = np.expand_dims(x, axis=0)
        # Subtracting avg values for each pixel
        x = preprocess_input(x)
        feature = self.model.predict(x)[0]  # (1, 4096) -> (4096, )
        return feature / np.linalg.norm(feature)  # Normalize
```