Classificação de documentos sensíveis da administração pública utilizando CBIR

Rogério Rodrigues Carvalho[1], **Sanderson Oliveira de Macedo**[1], **Leandro Luis Galdino Oliveira**[1], **Ronaldo Martins da Costa**[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)

{rogerior,sanderson_macedo,leandroluis,ronaldocosta}@ufg.br

Abstract. *Public organizations face difficulties in classifying and promoting transparency of their documents. Correct classification is critical to prevent public access to sensitive information and protect individuals and organizations from malicious use. This paper presents an ongoing research that proposes approaches to perform the task of classifying sensitive documents using machine learning techniques. Real data from the Electronic Information System (SEI) of UFG was used, and preliminary results demonstrate the potential and viability of the project, having already achieved an accuracy of 87% in the classification of public documents.*

Resumo. *As organizações públicas enfrentam dificuldades para realizar a devida classificação e promover a transparência dos seus documentos. A classificação correta é fundamental para prevenir o acesso público a informações sensíveis e proteger indivíduos e organizações contra o uso malicioso. Este trabalho apresenta uma pesquisa em andamento que propõe métodos para realizar a tarefa de classificação de documentos sensíveis utilizando técnicas de aprendizagem de máquina. Foram utilizados dados reais do Sistema Eletrônico de Informações (SEI) da UFG e os resultados preliminares demonstram o potencial e viabilidade do projeto, tendo já alcançado uma taxa de acerto de 87% na classificação de documentos públicos.*

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 [Brasil 2011], 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) [Brasil 2018] e a Lei de Acesso à Informação [Brasil 2011].

Em 2020 o Tribunal de Contas da União (TCU), por meio dos Acórdãos 389/2020 e 484/2021, solicitou que 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 essas exigências, a UFG deve realizar um trabalho de análise e classificação de todos os seus documentos antes de realizar a publicização. Entretanto, é inviável essa análise 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 [Kobayashi et al. 2018].

A técnica de recuperação de imagem baseada em conteúdo (*Content-based image retrieval* - CBIR) destaca-se na tarefa de extração de características de forma automática, sendo aplicável na análise de imagens de documentos. Segundo [Costa et al. 2015] a etapa inicial do CBIR consiste do processamento das imagens visando a extração de características. Posteriormente, essas características são consolidadas em vetores, os quais são utilizados na realização de consultas orientadas por critérios de semelhança.

O projeto de pesquisa em desenvolvimento 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.

Este trabalho apresenta um método alternativo para a classificação de documentos e os resultados preliminares alcançados. Convencionalmente, 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, sendo isso o enfoque deste trabalho.

2. Trabalhos Relacionados

A detecção de informações sensíveis e confidenciais é uma especialidade da classificação de texto. [Sousa and Kern 2023] realizou uma revisão sistemática abrangendo mais de 60 técnicas de *deep learning* 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. 2018] 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. 2020] 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.

O trabalho realizado por [McDonald et al. 2015] 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. A metodologia apresentou taxa de acurácia de 99% e *recall* de 45%.

Já [Occhipinti et al. 2022] 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 *stop word*, *HTML tags* e *lemmatisation*; Extração de *features* dos textos para gerar um dicionário; Aplicação de 12 modelos tradicionais de aprendizado de máquina na base de dados de spam Enron. O resultado obtido por meio da métrica F-*Score* foi de 94%.

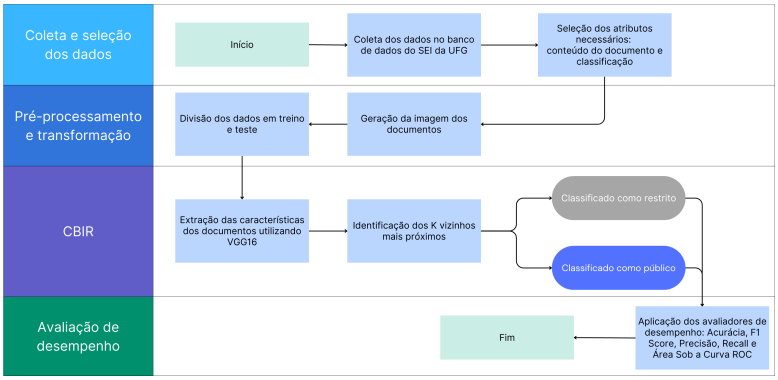
Um pipeline para classificação de textos sensíveis foi apresentado no trabalho de [Geetha et al. 2022]. Os passos realizados se assemelham com o trabalho de [Occhipinti et al. 2022] mas com o diferencial da inclusão de uma etapa caso o texto seja classificado como não sensível. Essa etapa consiste na 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 em comparação com o desempenho de um pipeline tradicional.

3. Método

Nas circunstâncias em que a obtenção de textos dos documentos é inviável ou a aplicação de técnicas de OCR não traz bons resultados, propomos um método alternativo. O método alternativo proposto distingue-se de métodos convencionais e da literatura encontrada por analisar a imagem dos documentos, em substituição da análise do texto.

No método, que pode ser observado pela Figura 1, inicialmente ocorre a extração de características dos documentos de treinamento e teste utilizando a técnica de CBIR nas imagens dos documentos. Posteriormente, para cada documento de teste são identificados os K documentos mais semelhantes, ou seja, aqueles com características mais próximas.

Figura 1. Arquitetura do método alternativo.



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 realizar a classificação do documento em público ou restrito. Essa abordagem de classificação baseia-se no KNN, que foi descrito por [Zhai 2022].

3.1. Atividades Realizadas

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.

Durante o desenvolvimento da pesquisa, foi realizado o estudo de viabilidade do projeto por meio da aplicação do método alternativo em um conjunto reduzido de documentos. Segundo a Plataforma Analisa UFG [UFG 2023], em 10 de junho de 2023, a UFG havia produzido um total de 243.204 processos e 2.817.138 documentos em seu SEI.

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[1] da linguagem de programação Python foi possível gerar a imagem dos documentos.

Após a geração das imagens de todos os documentos do conjunto de dados, utilizou-se a CNN VGG-16[2], com os pesos pré treinados da ImageNet, para realizar a extração das características de cada documento e gerar um vetor de características. Por meio dessa abordagem, a CNN foi capaz de gerar 4096 características de cada documento. O tempo de execução dessa extração 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.

O conjunto de dados foi dividido de modo que 80% dos dados foram destinados para ser usado como base de dados de características e os outros 20% para a avaliação de desempenho do método.

A forma utilizada para classificar os documentos como públicos ou restritos foi baseada no KNN. Logo, 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.

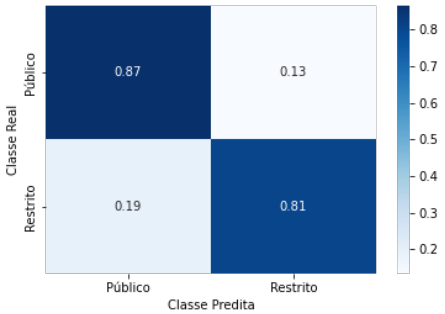
A definição do valor dos K vizinhos mais próximos foi realizada usando o método *trial-and-error* explicado por [Ougiaroglou and Evangelidis 2015]. O valor foi definido com base em testes iniciando com o número 1 e incrementando de 2 em 2, sempre usando um número ímpar, até chegar ao valor de 15 que apresentou melhor desempenho. Por fim, a mediana foi utilizada para verificar a classe predominante entre os vizinhos mais próximos e atribuir o resultado da mediana como valor predito para o documento avaliado.

[1]https://pypi.org/project/imgkit/
[2]https://keras.io/api/applications/vgg/

4. Resultados Preliminares

Para mensurar o desempenho do estudo de viabilidade do projeto, foram utilizados os avaliadores de desempenho mencionados nos trabalhos relacionados. Os resultados obtidos para cada um dos avaliadores de desempenho são: acurácia de 83,25%, F1 Score de 83,20%, 86,20% de precisão, *recall* de 80,58% e área sob a curva ROC de 83,58%.

Figura 2. Matriz de confusão.



Adicionalmente, foi gerada uma matriz de confusão (Figura 2) que revelou resultados promissores para o problema. Levando em consideração que o objetivo da UFG é dar publicidade aos documentos públicos, a metodologia alternativa 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 da acurácia dos trabalhos relacionados listados na Seção 2, observamos que o método alternativo apresenta um resultado satisfatório, conforme evidenciado na Tabela 1.

Tabela 1. Comparação do desempenho obtido com a literatura.

Trabalho/Abordagem	**Acurácia**
Trabalho de [McDonald et al. 2015]	97%
Método alternativo: classificação de documento público	87%
Método alternativo: classificação de documento restrito	81%
Trabalho de [Geetha et al. 2022]	66%

5. Conclusão, limitação e próximas etapas

Esta pesquisa visa propor métodos para realizar a classificação de documentos como públicos ou restritos. Os resultados preliminares são promissores e sinalizam o potencial da pesquisa. O estudo de viabilidade da metodologia alternativa apresentou um desempenho satisfatório, alcançando uma taxa de acerto de 87% na classificação de documentos públicos. Além disso, a metodologia mostrou-se como uma abordagem alternativa para o problema de classificação de documento, aplicável em situações onde a extração do texto dos documentos não é factível, dessa forma pode-se analisar a imagem do documento.

Como limitação, o conjunto de dados utilizado no estudo de viabilidade consiste de documentos nato-digitais que podem não refletir todos os cenários de documentos digitalizados, como distorções e manchas presentes em documentos físicos deteriorados.

Para as próximas etapas, no método alternativo planeja-se expandir a base de dados para contemplar todos os 2 milhões de documentos existentes no SEI da UFG, adicionar documentos físicos digitalizados na base de dados e avaliar outras técnicas de CBIR

além da VGG-16. Quanto aos métodos tradicionais de classificação de documentos, que realizam uma análise dos textos, será proposto e avaliado um método que adota essa abordagem, comparando o seu desempenho com o do método alternativo.

Referências

Battaglia, E., Bioglio, L., and Pensa, R. G. (2020). 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*, pages 67–79. Springer.

Brasil (2011). Lei nº 12.527, de 18 de novembro de 2011. *Diário Oficial da República Federativa do Brasil*.

Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018. *Diário Oficial da República Federativa do Brasil*.

Costa, R., Junior, E., Nunes, F., Oliveira, L., and Salvini, R. (2015). Analysis of techniques of the content-based image retrieval to construct an information system of the computer-aided diagnosis.

Geetha, R., Karthika, S., and Kumaraguru, P. (2022). 'Will I regret for this tweet?'Twitter user's behavior analysis system for private data disclosure. *The Computer Journal*, 65(2):275–296.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihok, G., and Den Hartog, D. N. (2018). Text classification for organizational researchers: A tutorial. *Organizational research methods*, 21(3):766–799.

McDonald, G., Macdonald, C., and Ounis, I. (2015). Using part-of-speech n-grams for sensitive-text classification. In *Proceedings of the 2015 International conference on the theory of information retrieval*, pages 381–384.

Neerbek, J., Assent, I., and Dolog, P. (2018). 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*, pages 373–385. Springer.

Occhipinti, A., Rogers, L., and Angione, C. (2022). A pipeline and comparative study of 12 machine learning models for text classification. *Expert Systems with Applications*, 201:117193.

Ougiaroglou, S. and Evangelidis, G. (2015). Dealing with noisy data in the context of k-NN classification. In *Proceedings of the 7th Balkan Conference on Informatics Conference*, pages 1–4.

Sousa, S. and Kern, R. (2023). 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.

UFG, A. (2023). Painel de indicadores do SEI-UFG. Acessado em junho de 2023.

Zhai, H. (2022). Improving KNN algorithm efficiency based on PCA and KD-tree. In *2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)*, pages 83–87. IEEE.