Automação da Classificação de Documentos do Ministério Público de acordo com os Objetivos de Desenvolvimento Sustentável utilizando Processamento de Linguagem Natural

Pedro P. Berger[1], Matheus W. Souza[1], Heitor Quartezani[1], Guilherme Merisio[1], Iara A. Fazolo[1], Luciana G. F. de Andrade[1], Sandro T. Silva[1]
[1]Ministério Público do Estado do Espírito Santo (MPES)
Espírito Santo, Brasil
{pberger, mwsouza, ifazolo, landrade, stsilva}@mpes.mp.br

Resumo. *O presente trabalho aborda a classificação de procedimentos do Ministério Público do Estado do Espírito Santo (MPES) conforme os Objetivos de Desenvolvimento Sustentável (ODS) para promover transparência e eficiência. Utilizando técnicas de Processamento de Linguagem Natural (NLP) e engenharia de dados, a metodologia proposta envolve quatro etapas: classificação inicial, pré-processamento, extração de características e classificação. Os resultados preliminares indicam boa precisão na classificação dos documentos utilizando técnicas de simples aplicação.*

Abstract. *This paper addresses the classification of procedures of the Public Prosecutor's Office of the State of Espírito Santo (MPES) according to the Sustainable Development Goals (SDGs) to promote transparency and efficiency. Using Natural Language Processing (NLP) techniques and data engineering, the proposed methodology involves four stages: initial classification, preprocessing, feature extraction, and classification. Preliminary results indicate good accuracy in document classification using simple application techniques.*

1. Introdução

O Ministério Público (MP) desempenha um papel fundamental na sociedade brasileira, atuando como guardião da ordem jurídica e defensor dos direitos sociais e individuais. A classificação adequada dos procedimentos que tramitam no MP, de acordo com as ODS (Objetivos de Desenvolvimento Sustentável) relacionadas a cada documento, é essencial para mensurar resultados, definir estratégias e garantir a transparência de suas ações para a sociedade. A padronização na classificação facilita a análise e o acompanhamento das atividades do MP, promovendo uma gestão mais eficiente e uma prestação de contas mais clara e precisa à população.

Os Objetivos de Desenvolvimento Sustentável (ODS) são uma agenda global adotada pela ONU em 2015, composta por 17 objetivos e 169 metas a serem alcançados até 2030. Eles visam erradicar a pobreza, proteger o planeta e garantir que todas as pessoas desfrutem de paz e prosperidade. A classificação dos documentos do Ministério Público do Estado do Espírito Santo (MPES) de acordo com os ODS é crucial para alinhar suas ações com essa agenda global, promovendo transparência, eficiência e responsabilidade social. Isso facilita a identificação de áreas prioritárias, a medição de

impactos e a promoção de práticas sustentáveis, contribuindo para o desenvolvimento sustentável do estado [ONU 2015]

São muitos os benefícios da utilização de métodos de Processamento de Linguagem Natural (NLP) para a classificação automática de documentos de acordo com as ODS, como a redução do trabalho repetitivo, eficiência, maior precisão e consistência das classificações realizadas [Fux *el al* 2022].

Nesse contexto, o presente estudo visa o aprimoramento da qualidade e automação da classificação dos documentos do Ministério Público do Estado do Espírito Santo de acordo com as ODS, utilizando para tal, ferramentas de NLP e engenharia de dados, a fim de ganhar eficiência e reduzir os erros relacionados a essa classificação, que impactam diretamente na transparência e na qualidade dos dados utilizados para traçar estratégias institucionais.

2. Metodologia Proposta

A metodologia aplicada consiste em quatro etapas: classificação inicial dos procedimentos, pré-processamento, extração de características e classificação, conforme Figura 1.

2.1 Seleção dos documentos e classificação dos procedimentos para treino do modelo

De acordo com a Resolução 74/2011 do Conselho Nacional do Ministério Público (CNMP), os procedimentos que tramitam no Ministério Público do Estado do Espírito Santo (MPES) são classificados conforme o sistema de tabelas unificadas [CNMP 2011]. Este sistema categoriza os procedimentos em Classe e Assunto, permitindo a identificação do tipo de procedimento ou processo e a matéria tratada. Essa classificação foi utilizada como um filtro inicial para identificar os possíveis Objetivos de Desenvolvimento Sustentável (ODS) relacionados ao procedimento.

Foram selecionados para análise os de "Petição Inicial" contidos nos procedimentos, pois representam o momento em que o documento é ajuizado ao tribunal, carregando, assim, uma maior quantidade de informações relevantes sobre o teor do procedimento. Em seguida, esses procedimentos foram classificados utilizando um modelo de linguagem GPT-3.5 turbo, acessado via API [OpenAI 2023]. Não foram considerados nesse estudo procedimentos criminais, dado que em sua grande maioria se relacionarem apenas ao ODS 16.

Foram classificados 7.732 documentos, e destes foram descartados 680 que tiveram sua classificação limitada ao caírem no filtro de conteúdo. As classificações foram validadas com a análise manual de 773 documentos e comparação entre os ODSs e sua classificação taxonômica.

Inicialmente foi proposta a classificação humana dos documentos, mas além de dispendiosa, exigia o conhecimento aprofundado dos Objetivos do Desenvolvimento Sustentável e metas relacionadas a esse, o que levava a uma classificação genérica e de baixa precisão quando realizada por agentes não treinados. Foi abandonada a proposta então.

2.2 Pré-processamento

O pré-processamento dos dados envolveu várias etapas para garantir a qualidade e consistência dos textos utilizados no modelo de classificação. Inicialmente, foi realizada

a extração do texto bruto a partir de documentos em formato HTML, removendo tags e juntando as palavras de forma coerente, mantendo os espaços entre elas.

Em seguida, foi aplicada a tokenização dos textos, que envolveu a conversão das palavras para letras minúsculas e a remoção de stopwords, utilizando o conjunto de stopwords em português disponível na biblioteca NLTK [Bird, S., Klein, E., & Loper, E. 2009]. Tokens desnecessários, como pontuações e números, foram eliminados para reduzir o ruído no conjunto de dados.

O processo também incluiu a normalização de referências legais. Padrões específicos, como números de leis seguidos de ano (exemplo: "Lei nº 1234/2021"), foram identificados e normalizados, combinando o número da lei diretamente com a palavra "lei" para manter a coerência sem perder a informação essencial. Além disso, tokens de baixa frequência foram removidos para reduzir a dimensionalidade do conjunto de dados, e tokens muito curtos ou puramente numéricos foram descartados para evitar que informações irrelevantes influenciassem o modelo.

2.3 Extração de Características

A fase de extração de características foi conduzida utilizando a técnica de TF-IDF (Term Frequency-Inverse Document Frequency), que permite transformar o texto dos documentos em uma matriz numérica representativa [Jones, K. S. 1972]. Inicialmente, os tokens previamente gerados para cada documento foram convertidos em strings concatenadas, formando um corpus textual adequado para a aplicação do TF-IDF. Essa transformação envolveu a união dos tokens em uma única string por documento, garantindo que cada documento fosse tratado como uma entidade textual coesa.

Para gerar a matriz TF-IDF, utilizou-se o TfidfVectorizer da biblioteca scikit-learn [Pedregosa et al 2011], configurado com parâmetros específicos para otimizar a qualidade da representação textual. O vetor foi ajustado para considerar termos com frequência mínima de 50 documentos e máxima em até 50% dos documentos, filtrando assim termos raros e excessivamente comuns, que poderiam prejudicar a discriminação entre os diferentes textos. Além disso, foi aplicada a sublinearidade na contagem de termos, que reduz o impacto de termos extremamente frequentes. Para refinar ainda mais a análise, foram removidas as stopwords em português, e o texto foi convertido para minúsculas para garantir a consistência.

O resultado foi uma matriz numérica onde cada linha representa um documento e cada coluna, um termo relevante, ponderado pelo TF-IDF, permitindo capturar as nuances semânticas dos textos e facilitando a identificação das relações entre os procedimentos analisados e os Objetivos de Desenvolvimento Sustentável.

2.4 Classificação

A classificação dos procedimentos foi realizada utilizando o modelo XGBoost [Chen, T., & Guestrin, C. 2016], escolhido por sua alta precisão em comparação com outros modelos testados, como Regressão Logística e Máquinas de Vetores de Suporte (SVM). O processo de classificação envolveu a análise individual de ada uma das 17 ODSs, utilizando um pipeline de Machine Learning cuidadosamente estruturado.

Para cada ODS, os dados foram inicialmente divididos em variáveis independentes (X) e dependentes (y), seguidos por uma divisão em conjuntos de treino e teste com uma proporção de 70/30. Embora o XGBoost não exija estritamente a

padronização dos dados, aplicou-se uma normalização através do StandardScaler para melhorar o desempenho e garantir consistência nas entradas. A especificação dos hyperparâmetros do modelo incluiu o uso do parâmetro random_state=42 para garantir a reprodutibilidade dos resultados, e a métrica de avaliação logloss, que foi utilizada para otimizar o modelo durante o processo de treinamento.

Durante a etapa de avaliação, as previsões realizadas pelo modelo nos dados de teste foram comparadas com os valores reais, utilizando-se matrizes de confusão e relatórios de classificação para calcular métricas como precisão, recall e F1-score. Essas métricas permitiram uma análise detalhada do desempenho do modelo, confirmando sua superioridade em relação aos outros algoritmos testados, especialmente em termos de precisão na identificação correta das ODSs.

É importante destacar que foram desenvolvidos 17 modelos, cada um apresentando variações de precisão, principalmente devido à diferença na quantidade de dados classificados para cada ODS. No contexto da atuação ministerial e suas respectivas áreas, procedimentos relacionados aos ODS 3, 4 e 10 (Saúde e Bem-estar, Educação de Qualidade e Redução das Desigualdades) são significativamente mais comuns do que aqueles relacionados aos ODS 7 e 9 (Energia Acessível e Limpa e Indústria, Inovação e Infraestrutura), por exemplo.

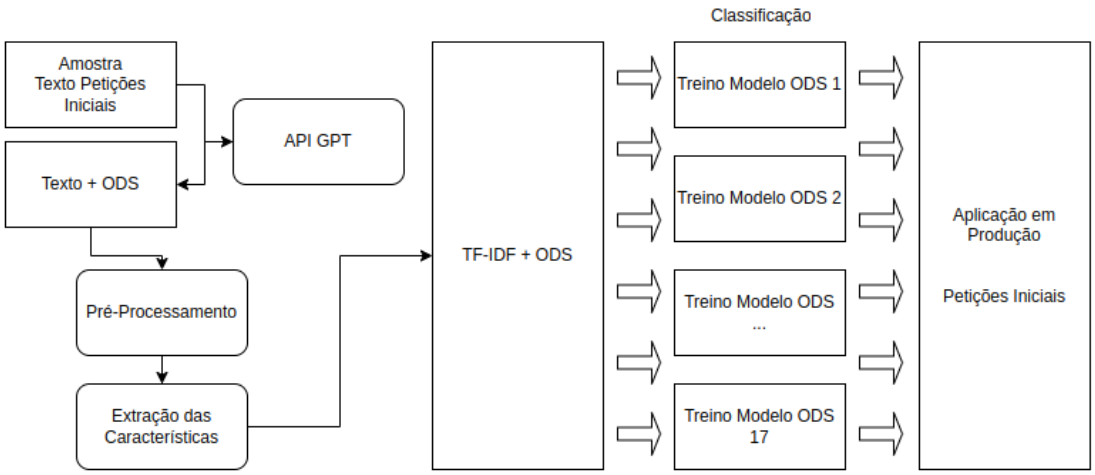

Figura 1. *Workflow* de trabalho

3. Resultados Preliminares

Os resultados preliminares da classificação para o ODS 3 mostraram um desempenho robusto do modelo XGBoost. A precisão geral do modelo foi de 92%, com uma taxa de acurácia semelhante. Especificamente, a classe 0.0 (ausência de associação com o ODS 3) alcançou uma precisão de 94% e um recall de 92%, resultando em um F1-score de 93%. Já a classe 1.0 (presença de associação com o ODS 3) obteve uma precisão de 89% e um recall de 91%, culminando em um F1-score de 90%. Esses resultados indicam que o modelo consegue distinguir eficazmente entre documentos associados e não associados ao ODS 3, com uma performance balanceada entre as classes, conforme refletido nas médias macro e ponderada de 92% para as três principais métricas (precisão, recall e F1-score). Esses valores sugerem uma alta confiabilidade do modelo para a tarefa de classificação automatizada dos procedimentos segundo o ODS 3.

4. Implementações Futuras

Apesar dos resultados promissores obtidos com o modelo XGBoost na classificação dos procedimentos relacionados ao ODS 3, há várias oportunidades para aprimoramentos e expansões futuras. Em primeiro lugar, pretende-se explorar a integração de técnicas de **ensemble learning** mais sofisticadas, que combinam múltiplos modelos de classificação para potencializar a performance geral [Opitz, D., & Maclin, R. 1999]. Essa abordagem pode ajudar a capturar diferentes padrões e melhorar a robustez das previsões, especialmente em classes minoritárias.

Outra linha de investigação envolve a **otimização dos hyperparâmetros** do XGBoost e de outros modelos candidatos. Embora os resultados atuais tenham sido satisfatórios, a aplicação de técnicas como busca em grade (grid search) e busca aleatória (random search) pode identificar configurações de hyperparâmetros que maximizem a acurácia e a generalização do modelo. Além disso, a implementação de algoritmos de otimização bayesiana pode oferecer uma busca mais eficiente e orientada.

A expansão do conjunto de dados para incluir documentos adicionais e a diversificação das classes de ODS também são áreas de interesse. Aumentar a representatividade do corpus documental pode não apenas melhorar a precisão das classificações, mas também permitir o desenvolvimento de modelos que abranjam uma gama mais ampla de ODSs, aumentando a aplicabilidade e utilidade do sistema automatizado no contexto do MPES.

Referências

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. *In Proceedings of the 24th International Conference on Neural Information Processing Systems* (pp. 2546-2554).

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Conselho Nacional do Ministério Público. (2011). Resolução nº 74, de 19 de julho de 2011. Diário Oficial da União, Seção 1, 19 ago. 2011. Disponível em https://www.cnmp.mp.br/portal/images/Resolucoes/Resoluo-0741.pdf

Fux, L., Santos, P. F. de O., Braga, A. C. D., Edokawa, P. S. D., & Castro, J. L. S. de. (2022). "Classificação de processos judiciais segundo Objetivos de Desenvolvimento Sustentável da Agenda ONU 2030". Revista da CGU, 14(26), 173-189. https://doi.org/10.36428/revistadacgu.v14i26.548

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11-21.

OpenAI. (2023). *GPT-3.5 Turbo: Advanced Language Models for Various Applications*. OpenAI. Disponível em https://openai.com

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11, 169-198.

Organização das Nações Unidas. (2015). Transformando Nosso Mundo: A Agenda 2030 para o Desenvolvimento Sustentável. Disponível em https://brasil.un.org/pt-br/sdgs

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). Butterworth-Heinemann.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.