PLANAPP

PLANEAMENTO | POLÍTICAS | PROSPETIVA

Artificial intelligence for better regulation (AI4IA@EU)

Relatório Síntese

Agosto 2024

FICHA TÉCNICA

Título
Artificial intelligence for better regulation (AI4IA@EU) – Relatório Síntese

Data
Agosto 2024

Autoria
Equipa Multidisciplinar de Avaliação de Políticas e Impacto Legislativo (EMAPIL) – PlanAPP

Revisão e *layout*
Equipa Multidisciplinar de Comunicação e Gestão do Conhecimento (EMCGC) – PlanAPP

Nota: Esta síntese é produzida com base no Relatório final do projeto "*Artificial intelligence for better regulation* (AI4IA@EU)", financiado pela União Europeia através da *DG Reform* (Direção-Geral de Apoio às Reformas Estruturais), no âmbito do Instrumento de Assistência Técnica (IAT). O projeto foi levado a cabo pela *NOVA IMS*, tendo como beneficiário final o PlanAPP.

PlanAPP – Centro de Competências de Planeamento, de Políticas e de Prospetiva da Administração Pública
Rua Filipe Folque, 44
1069-123 Lisboa
planapp@planapp.gov.pt
www.planapp.gov.pt

Índice

Índice de figuras, quadros e tabelas

Sumário executivo

O presente relatório corresponde a uma síntese de lições aprendidas a partir da execução do projeto *Artificial intelligence for Better Regulation* (AI4IA), adaptada do Relatório final do projeto, sintetizando igualmente as principais atividades realizadas.

São abordadas as vantagens e desvantagens da experiência relativamente à utilização da IA para realizar avaliações de impacto legislativo, enunciados os principais bloqueios enfrentados ao longo do projeto e evidenciadas as melhores práticas que surgiram da experiência efetuada.

Resumidamente, os objetivos do projeto AI4IA eram os seguintes:

- **Avaliação de impacto "escalável":** a principal funcionalidade da prova de conceito baseada em IA é fornecer informações sobre Diretivas comunitárias e outra legislação existente ou sobre propostas legislativas. No âmbito desta análise, o protótipo de IA produz um relatório de avaliação de impacto automático. Particularmente, através das capacidades avançadas de um modelo de processamento de linguagem natural (NLP), o sistema identifica parágrafos e trechos específicos nos quais estão presentes encargos administrativos. Além disso, é possível quantificar esses custos para as empresas, dessegregando a estimativa pelos diversos setores de atividade económica e por dimensão de empresa. Ao dispor de estimativas realizadas com recurso à IA, em tese, estamos a contribuir para capacitar os decisores políticos com evidências objetivas sobre potenciais impactos económicos, ajudando na tomada de decisão informada.

- **Comparação da legislação:** a prova de conceito desenvolvida ao longo do projeto introduz uma característica inovadora para comparar dois diplomas. Esta funcionalidade permite avaliar dois diplomas de origem nacional ou uma diretiva europeia comparando-a com o diploma de origem nacional que a transpõe para o ordenamento jurídico interno. Assim, ao identificar discrepâncias e situações de *gold plating* (a prática de acrescentar requisitos ou obrigações além do estipulado pela diretiva ou regulamento original), com a ajuda desta funcionalidade é possível reduzir tempos de análise, aumentar a transparência e facilitar uma compreensão mais profunda das variações de encargos entre dois diplomas.

- ***Feedback* de peritos e aprendizagem contínua:** a terceira característica relevante desta prova de conceito baseada em IA é que permite um mecanismo de *feedback* contínuo. Em particular, os peritos humanos mantêm a capacidade de corrigir as previsões do modelo. Este ciclo de *feedback* estabelece uma relação colaborativa entre os modelos de IA e os peritos. Em especial, o sistema aciona alertas quando se acumulam um número suficiente de encargos administrativos validados por humanos, assinalando a possibilidade de reconversão dos modelos. Assim, esta característica da prova de conceito garante a sua adaptabilidade à evolução dos processos legislativos.

Em suma, a importância da prova de conceito reside no seu potencial para transformar e racionalizar o processo de avaliação do impacto legislativo. Ao automatizar tarefas complexas, o projeto AI4IA

representa um marco significativo em direção a processos legislativos mais eficientes, transparentes e adaptativos, no qual os modelos de IA e peritos humanos colaboram ativamente.

O projeto AI4IA produziu resultados muito relevantes, demonstrando que a inteligência artificial pode melhorar de forma bastante significativa o paradigma da avaliação de impacto legislativo. Concretamente, a prova de conceito demonstrou:

- **Rigor e eficiência:** a prova de conceito revelou-se altamente eficaz na identificação precisa dos encargos administrativos presentes nos diplomas, permitindo uma análise rápida e a emissão automática de um relatório de avaliação de impacto.
- **Transparência e Precisão de Comparação:** a funcionalidade de comparação de diplomas revelou variações não detetadas em exercícios anteriores, garantindo uma compreensão mais transparente das alterações legislativas. Isso capacita os peritos e decisores a tomar decisões baseadas em análises mais abrangentes, permitindo detetar rapidamente situações de *gold plating*, com o potencial de evitar sobrerregulamentação das atividades económicas em Portugal.
- **Adaptabilidade e *feedback* contínuo:** o mecanismo de *feedback* contínuo demonstrou a capacidade do protótipo se adaptar a alterações nos padrões de elaboração de diplomas. A sinergia da IA com a experiência humana garante um processo de melhoria interativo, refinando e aprimorando, constantemente, o desempenho dos modelos.

No que respeita à aplicabilidade prática desta prova de conceito, a nível interno, estará em condições de ser utilizada pelo PlanAPP como suporte ao exercício de avaliação de impacto legislativo, podendo ser extensível, a nível nacional, a outras avaliações de impacto de medidas de simplificação e modernização administrativa nas quais, tipicamente, são avaliadas variações de custos administrativos.

Além disso, este documento poderá ainda servir como um ponto de partida para iniciativas que outros Estados-Membros da União Europeia pretendam desenvolver, com vista à implementação de projetos recorrendo à Inteligência Artificial, para apoio aos respetivos procedimentos de avaliação de impacto legislativo.

1. Introdução

O projeto AI4IA pretende definir um novo paradigma para a avaliação de impacto legislativo em Portugal, explorando o potencial da inteligência artificial. Genericamente, a avaliação de impacto corresponde a um exercício de análise sistemática dos efeitos das medidas públicas, resultantes de legislação primária ou derivada, sobre a sociedade, a economia e o ambiente. Trata-se de um instrumento que cria informação de apoio ao decisor público, disponibilizando dados sobre os impactos da intervenção e sem adotar qualquer posição relativa à intervenção pública em si.[1] Esta avaliação destina-se a assegurar que as propostas de intervenção são bem fundamentadas e transparentes e que os potenciais impactos são cuidadosamente examinados antes de serem tomadas decisões.

No entanto, a avaliação de impacto legislativo é, em regra, uma tarefa morosa que implica análise meticulosa das propostas de diploma. Implica, também, a colaboração de peritos de diversas áreas para prever os efeitos de uma medida proposta. O rigor necessário na avaliação dos potenciais resultados exige uma análise exaustiva das opções políticas alternativas e das eventuais mitigações.

Em Portugal, o PlanAPP (Centro de Competências de Planeamento, de Políticas e de Prospetiva da Administração Pública) é atualmente a entidade governamental responsável pela realização dos exercícios de avaliação de impacto legislativo. Não obstante, desde a adoção do modelo de AIL, primeiro como projeto-piloto, em 2017, e depois em definitivo, a partir de 2018, ter havido reforço significativo da equipa, dada a necessidade de uma realização atempada e eficaz dos exercícios de avaliação de impacto, o PlanAPP solicitou apoio à Direção-Geral de Apoio às Reformas Estruturais (*DG REFORM*) da Comissão Europeia, ao abrigo do Instrumento de Assistência Técnica (IAT), que, por sua vez, contratou a NOVA IMS (Nova *Information Management School*) para executar este projeto e, assim, aumentar a capacidade do PlanAPP no sentido da realização de exercícios de avaliação de impacto mais eficientes.

Deste modo, o projeto AI4IA teve como objetivo desenvolver uma prova de conceito baseada em inteligência artificial para apoiar o PlanAPP na automatização de parte do processo de avaliação de impacto. Este processo é hoje realizado por peritos humanos e requer uma quantidade significativa de tempo alocado. Além disso, trata-se de um processo caracterizado por alguma subjetividade. A prova de conceito, construída sobre um modelo de aprendizagem ajustado, é uma ferramenta inovadora para automatizar o processo de avaliação de impacto legislativo. Em particular, este projeto transformador pode representar ganhos de eficiência, precisão e transparência, para uma tarefa tradicionalmente complexa e demorada.

[1] Cf. https://diariodarepublica.pt/dr/lexionario/termo/avaliacao-impacto-legislativo-ail.

2. Prova de conceito de IA: desafios e lições aprendidas

2.1. Constrangimentos e limitações na utilização da IA

Apesar dos potenciais benefícios que a IA pode proporcionar, a literatura científica aponta várias questões e fatores limitativos para a sua adoção (consultar o anexo técnico para maior detalhe). Com efeito, foram identificados desafios ao longo de todo o projeto, que também resultaram das consultas às partes interessadas:

- **Falta de dados de treino para a construção dos modelos de IA**: verificou-se a ausência de diretivas anotadas ou legislações transpostas que atendam às necessidades específicas do projeto. Para colmatar este constrangimento, foi necessário efetuar um reforço de recursos humanos com competências jurídicas para proceder à anotação manual de diplomas para constituir um conjunto diversificado e extenso de legislação anotada para treinar os modelos.
- **Preocupação dos *stakeholders* com os resultados de um sistema automatizado**: outros os vários interessados frisaram a sua falta de confiança numa avaliação de impacto puramente baseada na IA. Esta limitação foi considerada no desenvolvimento da prova de conceito, uma vez que, além da análise realizada pela ferramenta, subsiste a validação humana efetuada pelos peritos.
- **Envolvimento de diversos *stakeholders***: os resultados do projeto devem ser facilmente partilháveis e fáceis de utilizar, uma vez que se espera que a prova de conceito também seja utilizada por utilizadores sem conhecimento especializado em avaliação de impacto. Esta limitação, previamente identificada, foi tida em conta, por exemplo na produção da interface da prova de conceito.

2.2. Melhores práticas/lições aprendidas sobre o uso da IA

Ao longo do desenvolvimento do projeto, foram várias as lições aprendidas. O Quadro 1 faz referência às mais importantes, que posteriormente são explicitadas.

Quadro 1 – Exemplos de aprendizagens ao longo do projeto

	LIÇÕES APRENDIDAS
1	Do laboratório ao mundo real, o principal desafio são os dados, não os algoritmos
2	Falta de dados rotulados
3	Formação de modelos de IA
4	Alinhamento de expectativas
5	Qualidade dos dados
6	Modelos amplamente utilizados
7	*Hardware* disponível para usar a prova de conceito
8	Solução de fácil utilização
9	Formação de recursos humanos para a utilização desta ferramenta
10	Envolvimento das partes interessadas

Fonte: relatório final do projeto.

1) Do laboratório ao mundo real, o principal desafio são os dados, não os algoritmos

A primeira lição aprendida com o projeto é que o principal desafio está relacionado com os dados, e não propriamente com os algoritmos. De facto, um projeto desta natureza deparou-se com a dificuldade de não dispor de um repositório de informação que pudesse servir de base para o desenvolvimento da prova de conceito.

2) Falta de dados rotulados

Em segundo lugar, além do desafio dos dados, identificamos uma falta de dados rotulados, que se traduziu num obstáculo significativo ao desenvolvimento do projeto. Com efeito, para que a prova de conceito realizasse a análise da avaliação de impacto legislativo, era necessária uma categorização/rotulagem da legislação existente que pudesse servir de base de comparação com a legislação destinada a essa avaliação.

3) Construção de modelos de IA

Em terceiro lugar, a lição aprendida com este projeto é a necessidade de treinar os modelos de IA, algo que provou ser uma tarefa exigente e que obriga a despender uma quantidade substancial de

tempo e recursos. Por este motivo, é importante confiar na utilização de fontes alternativas de dados de rotulagem e técnicas de anotação.

4) Alinhamento de expectativas

No desenvolvimento destes projetos, é importante avaliar cuidadosamente a situação existente da parte do beneficiário, seja em termos de disponibilidade de dados, seja em termos de recursos humanos. A expectativa da equipa de desenvolvimento, no caso deste projeto, era significativamente diferente da situação real verificada no PlanAPP, onde se encontrou uma quantidade muito limitada de dados rotulados. Assim sendo, a realização de uma avaliação adequada da situação inicial, com a correta identificação dos constrangimentos e limitações existentes (em termos de recursos humanos, de informação e materiais disponíveis), é crucial para o sucesso de projetos desta natureza.

5) Qualidade dos dados

Um dos aspetos críticos deste projeto prendeu-se com a qualidade da informação. De facto, embora a investigação do desempenho de diferentes métodos de *machine learning* possa ser importante, a qualidade dos dados é incomparavelmente mais relevante.

Neste projeto, foram investigadas várias técnicas de última geração que tentam melhorar a precisão e a recordação da prova de conceito. Não obstante, este esforço exigiu uma quantidade significativa de tempo e não produziu resultados significativos. Portanto, no âmbito deste projeto, dedicar algum esforço aos aspetos de qualidade dos dados revelou-se eficaz e produziu resultados significativos no que diz respeito ao desempenho da prova de conceito (precisão/*recordação*).

6) Modelos amplamente utilizados

Outra lição aprendida está relacionada com a confiança em modelos amplamente utilizados. De facto, para alcançar um desempenho próximo do excelente, é possível confiar em modelos amplamente utilizados, sem a necessidade de desenvolver algoritmos e modelos *ad hoc*.

7) *Hardware* disponível para usar a prova de conceito

É importante considerar a complexidade do modelo (ou seja, o número de parâmetros) ao fazer a transição de uma prova de conceito para um sistema real em ambiente produtivo. Desta forma, é necessário garantir o *hardware* adequado para a utilização eficaz desta solução de IA. Assim, uma das lições aprendidas é a importância de avaliar a disponibilidade de providenciar o *hardware* adequado e necessário pela parte do beneficiário.

8) Solução de fácil utilização

O desempenho do modelo — em termos de precisão ou outras métricas de avaliação — é relevante, mas outros indicadores de desempenho devem ser tidos em consideração. Um destes indicadores de desempenho está relacionado com o tempo de inferência, bem como com as restrições/requisitos informáticos. Isso leva-nos a uma questão central para o sucesso da implementação dessa prova de conceito: a facilidade de utilização do sistema para outros técnicos que não peritos em IA. Na verdade, o potencial desta solução será diretamente proporcional quanto mais for possível a sua plena utilização.

9) Formação de recursos humanos para a utilização desta ferramenta

Produzir uma prova de conceito que seja simples e intuitiva no uso é fundamental para a sua adoção e posterior desenvolvimento e passagem para a fase de produção.

Objetivamente, existe uma preocupação legítima na utilização de ferramentas automatizadas, especialmente devido ao desconhecimento generalizado sobre as capacidades e consequências da IA. Portanto, é natural que as organizações sejam relutantes em utilizar ferramentas inovadoras para apoiar tarefas que fazem parte do seu trabalho diário. Neste sentido, a integração dos sistemas de IA nas atividades das organizações deve ser o mais transparente possível, existindo um esforço de capacitação e formação dos recursos humanos para garantir uma utilização plena das ferramentas de inteligência artificial.

10) Envolvimento dos *stakeholders*

Considerando o impacto significativo que as soluções de inteligência artificial podem ter no dia-a-dia das operações das organizações e processos, uma das lições deste projeto é a importância da consulta contínua e do envolvimento com os diferentes *stakeholders*, desde aqueles que são envolvidos diretamente nos processos, até aos destinatários. O envolvimento constante do beneficiário e dos principais *stakeholders* foi fundamental para orientar o desenvolvimento da prova de conceito com o objetivo de atingir os resultados esperados.

3. Principais conclusões e desafios futuros

A principal conclusão do projeto, como já referido anteriormente, centra-se na importância da disponibilidade e qualidade dos dados, mais do que no desenvolvimento dos algoritmos. Para além da existência e qualidade dos dados, o tipo de informação disponível é crucial para a implementação de projetos desta natureza. Na verdade, a falta de dados rotulados foi um constrangimento encontrado, revelando-se um estrangulamento significativo no desenvolvimento do projeto. Além disso, o treino dos modelos de IA revelou-se de significativa importância para garantir elevados padrões de precisão.

Verificou-se que, para obter um desempenho próximo do excelente é possível recorrer a modelos amplamente utilizados, sem a necessidade de desenvolver algoritmos e modelos *ad hoc*.

Na prática, a criação desta prova de conceito terá os seguintes impactos a curto prazo:

- Contribuir para um exercício de avaliação de impacto legislativo mais eficiente e menos moroso, permitindo uma comparação mais rápida entre propostas legislativas ou entre uma proposta e a legislação existente.
- Facilitar análises da legislação da UE e da legislação transposta ou a transpor para o ordenamento jurídico interno, identificando obrigações sobrepostas para empresas ou cidadãos, assim como permitindo detetar situações de *gold plating*.

Espera-se que o projeto AI4IA contribua para reforçar a agenda Legislar Melhor durante os exercícios de avaliação de impacto legislativo e possa constituir um primeiro passo no desenvolvimento de uma plataforma mais abrangente que explore a utilização de tecnologias avançadas baseadas em dados, com vista a promover uma melhor regulamentação.

3.1. Próximos passos

O desenvolvimento da prova de conceito de IA comporta inúmeras vantagens, como já referido anteriormente. Para além da sua potencial utilização no processo de avaliação de impacto legislativo realizado pelo PlanAPP, poderá ser extensível, a nível nacional, a avaliações de impacto de medidas de simplificação e modernização administrativa nas quais, tipicamente, são avaliadas variações de custos administrativos.

A nível internacional, existem também muitas possibilidades de divulgação desta ferramenta, em especial junto de outros Estados-Membros da UE. No entanto, para a adoção da prova de conceito noutros Estados-Membros, será imperativo requalificar os modelos utilizados desde o início, personalizando-os para acomodar as particularidades linguísticas específicas de cada Estado-Membro, sendo necessário, simultaneamente, abordar potenciais disparidades nas características e forma dos textos legislativos.

No que diz respeito aos objetivos de médio e longo prazo do projeto, é possível destacar:

- Aumento da confiança dos *stakeholders* e, em particular, das empresas face aos legisladores a nível da UE e a nível nacional, resultante de um processo legislativo menos subjetivo e mais fundamentado com base em evidência suportada em dados. Além disso, o aumento da confiança tem também o potencial de se refletir no maior envolvimento dos *stakeholders* durante fases de consulta pública.
- Disseminação de boas práticas e da utilização de técnicas de IA nos Estados-Membros a nível da UE, e capitalização do seu potencial para gerar informações fiáveis e objetivas a partir de grandes quantidades de dados.
- Promoção da utilização da IA para avaliação de outros impactos, nomeadamente, na ação climática, na igualdade de género, entre outros.

Referências Bibliográficas

Assenta, Burr. 2012. "Aprendizagem Ativa". Synthesis Lectures on Artificial Intelligence and Machine Learning 18 (julho):1–111. https://doi.org/10.2200/S00429ED1V01Y201207AIM018.

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, e Ion Androutsopoulos. 2020. "LEGAL-BERT: Os Muppets Direto da Faculdade de Direito", outubro. http://arxiv.org/abs/2010.02559.

Clark, Kevin, Minh-Thang Luong, Google Brain, Quoc V Le Google Brain e Christopher D Manning. 2020. "ELECTRA: Codificadores de texto pré-treino como discriminadores em vez de geradores." Conferência Internacional sobre Representações de Aprendizagem (ICLR), março de https://doi.org/10.48550/arxiv.2003.10555.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google e A I Language. 2019. "BERT: Pré-Formação de Transformadores Bidirecionais Profundos para a Compreensão da Linguagem". Atas da Conferência do Norte de 2019, 4171–86. https://doi.org/10.18653/V1/N19-1423.

Gururangan, Suchin, Ana Marasovi'cmarasovi'c, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, Noah A Smith e ♦ † Allen. 2020. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks", julho, pp. 8342–60. https://doi.org/10.18653/V1/2020.ACL-MAIN.740.

Howard, Jeremy e Sebastian Ruder. 2018. "Ajuste fino do modelo de linguagem universal para classificação de texto". ACL 2018 - 56ª Reunião Anual da Association for Computational Linguistics, Atas da Conferência (Long Papers) 1:328–39. https://doi.org/10.18653/V1/P18-1031.

Johnson, Justin M. e Taghi M. Khoshgoftaar. 2019. "Pesquisa sobre Deep Learning com Desequilíbrio de Classe". Jornal de Big Data 6 (1): 1–54. https://doi.org/10.1186/S40537-019-0192-5/TABLES/18.

Li, Qian, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu e Lifang He. 2022. "Uma Pesquisa sobre Classificação de Texto: Do Tradicional ao Deep Learning". Transações ACM em Sistemas e Tecnologia Inteligentes (TIST) 13 (2). https://doi.org/10.1145/3495162.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, et al., 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach", julho. https://doi.org/10.48550/arxiv.1907.11692.

Loshchilov, Ilya e Frank Hutter. 2017. "Regularização do Decaimento de Peso Desacoplado". 7th International Conference on Learning Representations, ICLR 2019, novembro. https://arxiv.org/abs/1711.05101v3.

Northcutt, Curtis G., Lu Jiang e Isaac L. Chuang. 2021. "Aprendizagem Confiante: Estimando a Incerteza em Rótulos de Conjuntos de Dados". Journal of Artificial Intelligence Research 70 (abril):1373–1411. https://doi.org/10.48550/arxiv.1911.00068.

Pan, Sinno Jialin e Qiang Yang. 2010. "Um Inquérito sobre a Aprendizagem por Transferência". Transações IEEE sobre Conhecimento e Engenharia de Dados 22 (10): 1345–59. https://doi.org/10.1109/TKDE.2009.191.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee e Luke Zettlemoyer. 2018. "Representações de palavras profundamente contextualizadas". NAACL HLT 2018 - 2018 Conferência do Capítulo Norte-Americano da Association for Computational Linguistics: Human Language Technologies - Atas da Conferência 1:2227–37. https://doi.org/10.18653/V1/N18-1202.

Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai e Xuanjing Huang. 2020. "Modelos Pré-Treinados para Processamento de Linguagem Natural: Uma Pesquisa". Ciência China Ciências Tecnológicas 63 (10): 1872–97. https://doi.org/10.1007/s11431-020-1647-3.

Radford, Alec, Karthik Narasimhan, Tim Salimans e Ilya Sutskever. 2018. "Melhorar a Compreensão da Linguagem através da Pré-Formação Generativa." https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Ren, Pengzhen, Yun Xiao, Xiaojun Chang, Po Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen e Xin Wang. 2021. "Uma Pesquisa de Aprendizagem Ativa Profunda". Inquéritos Computacionais ACM (CSUR) 54 (9). https://doi.org/10.1145/3472291.

Vaswani, Ashish, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser e Illia Polosukhin. 2017. "Atenção é tudo o que você precisa". Avanços em Sistemas de Processamento de Informação Neural 30.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov e Quoc V. Le. 2019. "XLNet

Anexo Técnico

Objetivo

O projeto AI4IA@EU, visa melhorar a eficiência da avaliação do impacto legislativo através da automatização da identificação dos potenciais encargos administrativos presentes nos diplomas. Atualmente, o processo manual é moroso e necessita de muitos recursos humanos alocados, limitando o número de diplomas que podem ser analisados. O AI4IA@EU, procura automatizar este processo utilizando tecnologias avançadas como a IA, permitindo, assim, uma avaliação de impacto mais abrangente e sistemática.

Tarefas de *Machine Learning* e *pipeline* de inferência

O componente de IA da prova de conceito responde à necessidade de compreensão de texto em escala com recurso a técnicas de Processamento de Linguagem Natural (NLP). Concretamente, o objetivo é aliviar o perito em avaliação de impacto de rever, manualmente, cada passagem da legislação, especificamente da proposta de transposição e/ou adaptação no caso de diretivas e regulamentos comunitários e, em vez disso, treinar um modelo de NLP para identificar os encargos administrativos existentes como uma classificação de texto (Qiu et al. 2020); (Li et al. 2022).

DEFINIÇÃO DO CONCEITO

A Classificação de Texto é a tarefa de atribuir um rótulo ou classe a uma determinada passagem ou trecho de texto. Esse texto de entrada pode ser uma frase, um parágrafo ou até mesmo um documento.
Se a tarefa for:

- *binária*, significa que a variável alvo tem apenas dois rótulos para selecionar (por exemplo, positivo ou negativo);
- *multi-classe*, significa que ainda existe apenas uma variável alvo, mas com mais de dois valores dos quais o modelo só pode escolher um;
- *multi-label, refere-se* a configurações onde são consideradas várias variáveis e o modelo pode atribuir mais de que uma ao texto de entrada.

Neste sentido, foi desenvolvido uma espécie de "modelo de filtro" capaz de receber as passagens de texto de um diploma e classificar cada passagem como contendo ou não um encargo administrativo — ou seja, uma tarefa binária de classificação de texto. Com essa capacidade, o sistema consegue acelerar o trabalho manual habitualmente realizado pelos peritos. Posteriormente, como capacidades complementares, foram desenvolvidos dois modelos secundários para rotular os encargos administrativos identificados, desagregados por Obrigação de Informação (IO) e Atividade Administrativa mais adequados (ou seja, uma tarefa *multi-label*). A Figura 1 ilustra como a tarefa do modelo de filtro de identificar encargos (ou seja, obrigações de informação) e os classificadores de

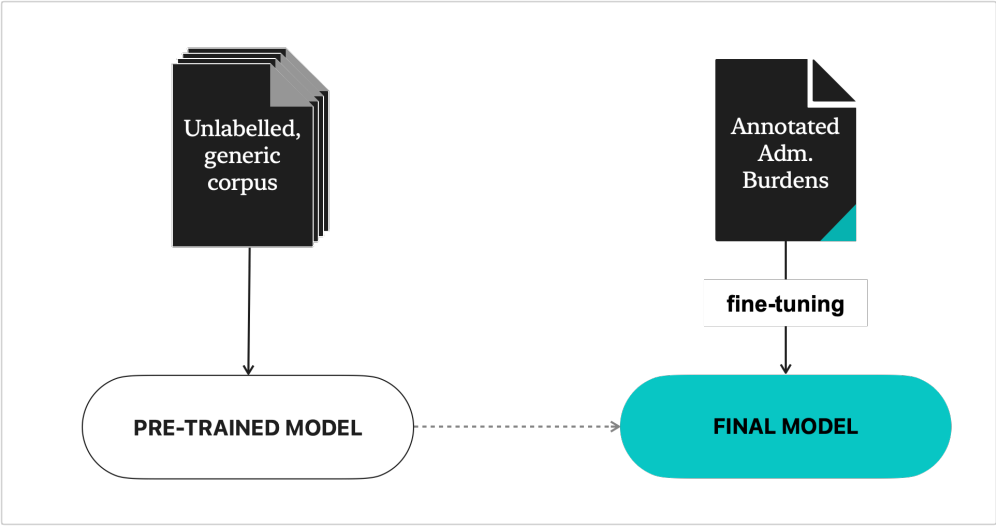
apoio poderiam simplificar ainda mais a rotina do perito — reforçando que o objeto de estudo para esta pesquisa e relatório é o principal modelo de "filtro".

Aprendizagem de transferência

A aprendizagem de transferência visa mitigar a escassez de rotulagem, aproveitando dados de outros domínios para treinar modelos que apresentem uma melhor capacidade de generalização. Tais modelos podem servir como ponto de partida para serem posteriormente personalizados para tarefas específicas. Recentemente, o campo da NLP tem capitalizado técnicas de aprendizagem de transferência com recurso a grandes quantidades de texto produzir modelos pré-treinados que melhoraram significativamente os resultados de última geração para várias tarefas de NLP. São alguns exemplos de confirmação empírica sobre o benefício de treinar uma rede neural inicial em dados não rotulados para serem posteriormente personalizados (ou ajustados, como comumente referido) em tarefas supervisionadas a jusante (Pan e Yang 2010) (2019), Howard e Ruder (2018), Peters *et al.* (2018) e Radford *et al*..

A maioria dos modelos pré-treinados publicados nos últimos anos para servir propósitos de aprendizagem de transferência em NLP apresentam duas características principais: uma rede neural profunda com um grande número de parâmetros e um *corpus* de dados maciço para treiná-la. Tais proporções permitem alcançar uma representação da própria língua — daí serem denominadas como modelos linguísticos. O modelo pré-treinado aprende primeiro uma representação da língua portuguesa e, em seguida, recebe uma nova camada de informação onde aprende a tarefa desejada, neste caso específico de identificar encargos administrativos em diplomas (Figura 1).

Figura 1 — Visualização conceptual da Aprendizagem de Transferência (através de afinação) dentro da NLP



Fonte: relatório final do projeto.

Aprendizagem Ativa

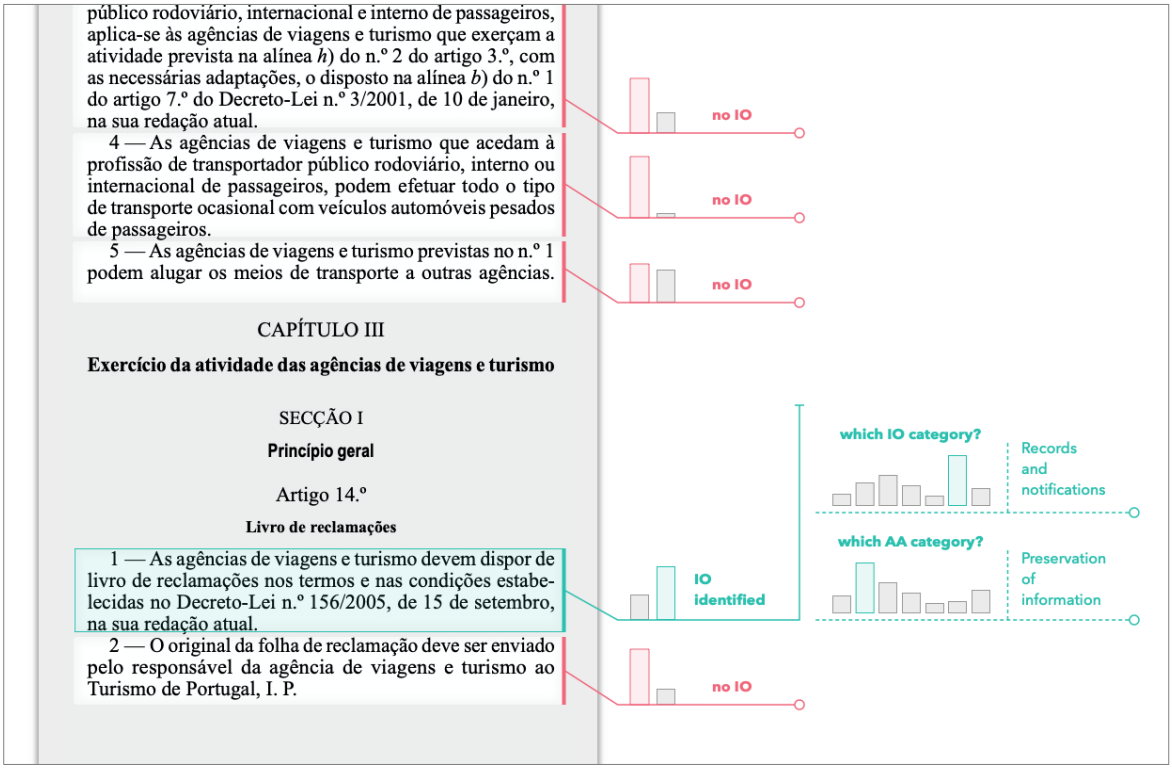
A aprendizagem ativa visa minimizar a quantidade de rotulagem manual necessária para alcançar resultados satisfatórios no modelo final. A partir de um conjunto inicial rotulado, o modelo é primeiro treinado com os dados supervisionados disponíveis. Depois, em vez de se efetuar uma amostragem aleatória de dados adicionais a serem anotados por peritos, as instâncias são estrategicamente selecionadas com base em um critério de incerteza que estima esses pontos de dados como sendo os mais valiosos para o modelo. Por exemplo, escolher amostras que estejam próximas do limite de decisão do modelo. Este processo é repetido até que um critério de rescisão seja cumprido. O racional deste método é que, servindo apenas exemplos informativos, o modelo aprende os padrões desejados com menos dados, consequentemente, exigindo menos tempo da parte de quem efetua a rotulagem (Liquida 2012; 2021).

Aprendizagem de confiança

Além de projetar os dados de treino e reduzir a necessidade de rotulagem manual, é crucial garantir que as anotações nas quais o desempenho do modelo está a ser avaliado são confiáveis. A aprendizagem existe no contexto dos dados, mas as noções de confiança, normalmente, concentram-se em previsões de modelos e não em rótulos de qualidade. O método de Aprendizagem Confiante não se concentra nas previsões do modelo para avaliar a confiança, mas sim na correção das anotações. Tem como objetivo caracterizar e identificar erros de rotulagem nos conjuntos de dados, sendo capaz de “filtrar” dados considerados como sendo de informação duvidosa ou incorreta e classificar exemplos. Este método de aprendizagem foi aplicado para validar a qualidade das anotações efetuadas e proteger o treino do modelo com rótulos de informação duvidosa ou incorreta (Northcutt, Jiang e Chuang 2021).

A Figura 2 retrata o processo de classificação efetuado em duas etapas. Ou seja, em primeiro lugar são filtradas passagens de texto com obrigações e, em seguida, classificadas dentro das diferentes categorias para as Obrigações de Informação e Atividades Administrativas.

Figura 2 — Ilustração do processo de classificação em duas etapas



Fonte: relatório final do projeto.

Algoritmo

Avanços em Redes Neurais Profundas, como a arquitetura *transformer* e grandes modelos de linguagem pré-treinados, são considerados o estado da arte em NLP preditiva e, portanto, foram explorados para a construção do modelo utilizado na prova de conceito. O objetivo central era tirar partido de uma configuração de Aprendizagem de Transferência na qual o conhecimento anteriormente aprendido pelo modelo a partir de uma vasta quantidade de dados genéricos é reaproveitado e ajustado para uma tarefa específica. (Vaswani et al. 2017) (Devlin et al. 2019; Liu et al., 2019; 2019; Clark et al., 2020) (Pan e Yang 2010)

Especificamente, três modelos pré-treinados foram avaliados para a prova de conceito AI4IA. Dois deles foram desenvolvidos pela Unicamp (Universidade Estadual de Campinas, Brasil) em parceria com a Universidade de Waterloo, no Canadá, e estão disponíveis no repositório aberto *Hugging Face*. "BERTimbau Base"[2] e "BERTimbau Large"[3] — como são chamados os modelos — alavancam a arquitetura original do BERT e foram ambos pré-treinados num *corpus* em português do brasil (ou seja, o conjunto de dados brWaC), que é composto por 2,7 mil milhões de *tokens*, durante um milhão de etapas. A principal diferença entre estes dois é o tamanho: "Base" tem 110 milhões de parâmetros,

[2] https://huggingface.co/neuralmind/bert-base-portuguese-cased
[3] https://huggingface.co/neuralmind/bert-large-portuguese-cased

enquanto “Large” tem 335 milhões de parâmetros. Mais informações podem ser encontradas no repositório *github* dos modelos (Devlin et al. 2019)[4].

O terceiro modelo pré-treinado alavancado para avaliação experimental foi a variante de 100 milhões de parâmetros da família de modelos “Albertina[5]”. “Albertina 100M PTPT”, como é referido, é um modelo linguístico para o português europeu. É um codificador da família BERT, desenvolvido sobre o modelo DeBERTa. Foi treinado ao longo de um conjunto de dados de 2,2 mil milhões de *tokens* que resultaram da recolha a partir das seguintes fontes:

- OSCAR: o conjunto de dados OSCAR inclui documentos em mais de cem línguas, incluindo o português, e é amplamente utilizado na literatura. É o resultado de uma seleção realizada sobre o conjunto de dados Common Crawl, rastreado da *web*, que retém apenas páginas cujos metadados indicam permissão para ser rastreado, que realiza desduplicação de dados e que remove alguns clichês, entre outros filtros. Considerando que não discrimina entre as variantes portuguesas, foi realizada uma filtragem adicional retendo apenas documentos cujos metadados indicam o código de país da Internet de domínio superior ao de Portugal. Foi utilizada a versão de janeiro de 2023 do OSCAR, que é baseada na versão de novembro/dezembro de 2022 do Common Crawl.
- DCEP: o Corpus Digital do Parlamento Europeu é um *corpus* multilingue que inclui documentos em todas as línguas oficiais da UE, publicados no sítio *web* oficial do Parlamento Europeu. A parte portuguesa foi mantida.
- Europarl: o Corpus Paralelo dos Trabalhos do Parlamento Europeu extraiu os trabalhos do Parlamento Europeu de 1996 a 2011. A parte portuguesa foi mantida.
- ParlamentoPT: o ParlamentoPT é um conjunto de dados obtidos através da recolha de documentos publicamente disponíveis com a transcrição dos debates na Assembleia da República.

A utilização de modelos pré-treinados de maior dimensão foi evitada devido a considerações sobre como a prova de conceito deve funcionar em ambiente produtivo. O principal objetivo da prova de conceito é racionalizar a análise efetuada no processo de avaliação de impacto legislativo, que, por conseguinte, deve traduzir-se num tempo de inferência razoavelmente rápido. Considerando que alguns documentos podem ser significativamente longos, modelos maiores podem levar até vários minutos para produzir previsões, o que não se afigurava muito producente.

[4] https://github.com/neuralmind-ai/portuguese-bert
[5] https://huggingface.co/PORTULAN/albertina-100m-portuguese-ptpt-encoder

Dados

A principal tarefa de *machine learning* no projeto AI4IA resume-se a uma classificação de texto binário da existência ou não de obrigações no âmbito das diretivas da UE e da legislação portuguesa transposta. No entanto, considerando o *pipeline* completo previsto de um modelo principal de "filtro" suportado por modelos complementares de "categorias", decidimos aproveitar a oportunidade para, simultaneamente, recolher dados para essas duas tarefas secundárias. Portanto, do ponto de vista do desenvolvimento de dados, as tarefas consideradas foram:

- classificação de texto binário com a presença de obrigações de informação (IO);
- classificação de texto multi-rótulo de categorias de IO;
- classificação de texto multi-rótulo de categorias AA.

TEORIA *SNIPPET*

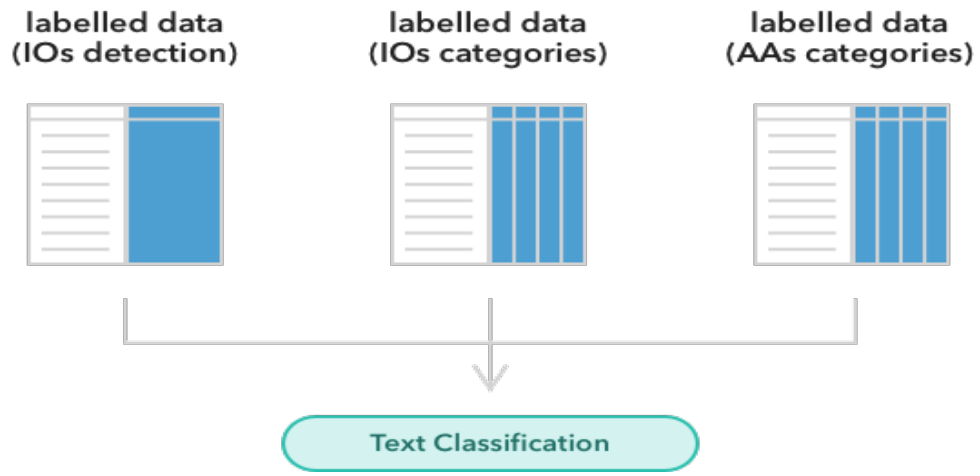
A Classificação de Texto encontra-se sob um paradigma de "aprendizagem supervisionada", caracterizando-se por exigir conjuntos de dados anotados (também conhecidos como "dados rotulados") para treinar o modelo. Ou seja, um conjunto de dados onde cada entrada deve ser acompanhada pelo valor da variável (ou variáveis) a ser prevista. A relação entre o texto e este valor é o que permite aprender semelhanças implícitas entre as entidades analisadas. Em resumo, o algoritmo aprende executando os dados várias vezes e, após cada iteração, compara suas previsões com o valor real esperado para a entrada. Dependendo de quão certo ou errado o modelo consegue ser, é ajustado o seu "conhecimento" (e.g., os seus "pesos") em conformidade. Uma vez aprendido o padrão implícito entre os textos e as variáveis dependentes, o modelo torna-se capaz de generalizar para novos cenários, destacando ocorrências da variável alvo aprendida em novos *inputs*.

A tradução dessas tarefas para os seus conjuntos de dados dedicados levou a equipa técnica a construir 3 conjuntos finais (Figura 3). Os peritos em avaliação de impacto anotaram diplomas para produzir:

i. um conjunto de dados supervisionado com um objetivo binário, em que documentos inteiros são concatenados e cada uma das suas passagens — frases ou parágrafos — recebe um rótulo positivo ou negativo (ou seja, para indicar se existe uma obrigação). Estes servem de exemplo para o modelo aprender a semântica que se correlaciona com os encargos administrativos e, portanto, tornar-se capaz de prever a sua presença (ou ausência) em documentos novos e invisíveis.
ii. um conjunto de dados supervisionado com um alvo multi-rótulo, em que todas as observações representam uma passagem de texto com um IO e atribuído a ele, uma ou várias das categorias de IO consideradas na versão portuguesa do SCM elencadas no Quadro 2.

iii. um conjunto de dados supervisionado também com passagens de texto de destino multi-rótulo e somente IO, mas agora com categorias AA atribuídas elencadas na Quadro 2.

Figura 3 — Conjuntos de dados recolhidos durante a investigação do projeto AI4IA.



Fonte: relatório final do projeto.

Quadro 2 – Categorias de obrigações de informação e atividades administrativas seguidas pelo Governo português ao empregar a avaliação de impacto

Obrigação de informação (IO)	Atividade Administrativa (AA)
Entrega de informações comerciais e fiscais	Familiarização com a obrigação
Disponibilização de manuais de procedimentos e planos de ação	Recolha e organização da informação
Registos e notificações	Processamento da informação
Candidatura a subsídios ou outros apoios	Tempos de espera
Pedidos de licenças, certidões, autorizações ou permissões	Deslocações
Cooperação com auditorias, fiscalizações e inspeções	Submissão de informação
Colocação de rótulos informativos e prestação de informação a consumidores e outras entidades	Preservação da informação

Fonte: relatório final do projeto.

Fontes de informação

Uma vez que não existiam dados previamente explorados ou processados para serem reutilizados neste projeto, foram tratados novos dados. Primeiro, com o objetivo de construir modelos capazes de analisar a legislação transposta, foram extraídos documentos do *website* do Diário da República

(DRE)[6]. Segundo, para ensinar aos modelos a identificar obrigações presentes nas diretivas da UE que devem ser transpostas, foram extraídos documentos do *website* EUR-Lex[7]. Ambas as fontes de dados são portais institucionais, com acesso universal e gratuito, incluindo a possibilidade de impressão, armazenamento, pesquisa e acesso integral ao conteúdo da legislação.

Extração

No que respeita à extração, os documentos foram obtidos através de programas personalizados e direcionados para esta tarefa.

O *website* do DRE não dispõe de uma interface de programação de aplicações (API) que facilite o acesso ao seu conteúdo. Neste caso, recorreu-se a um *Web Scrapper* que foi desenvolvido para extrair os documentos. Em grande medida, esta aplicação obtém informação através de um navegador automatizado (*Selenium*[8]), sendo o conteúdo do site carregado de forma dinâmica.

Já no que respeita à extração de ficheiros do EUR-Lex, existe a possibilidade de recorrer a uma API. No entanto, a equipa técnica optou por adotar uma abordagem semelhante a um *Web Scrapper*.

Formatação

Depois de selecionar as amostras de documentos necessários de ambas as fontes e extrair o texto dos ficheiros, foram analisadas num formato considerado adequado para treinar os modelos de IA. O mesmo formato de entrada é usado para as três tarefas de classificação de texto mencionadas, sendo que apenas diferem as variáveis-alvo entre elas. Uma vez que o objetivo é identificar as obrigações de informação específicas dentro do documento (permitindo quantificá-las), é preferível trabalhar no nível mais granular possível que descreve semanticamente uma obrigação individual. Por conseguinte, os textos dos diplomas são divididos em frases e a cada um é atribuído um rótulo adequado. Na prática, cada frase-rótulo representa uma ocorrência no conjunto de dados para classificação de texto.

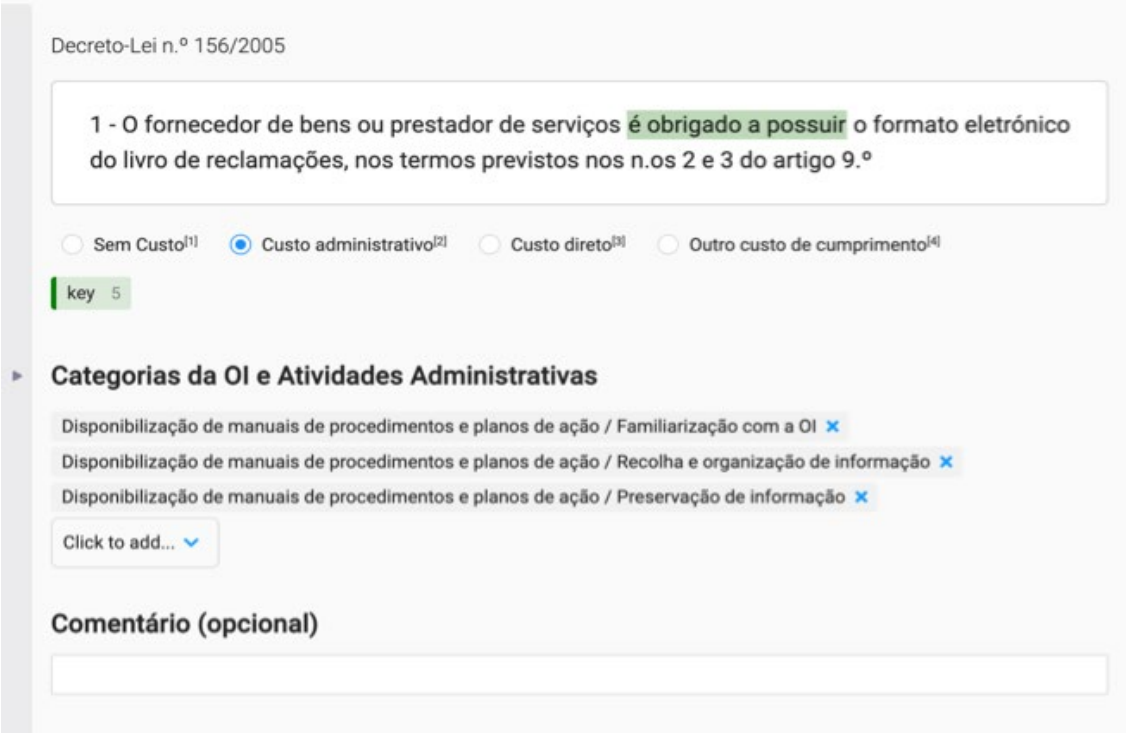
Criação de *ground-truth data*

Para o projeto AI4IA, a obtenção de dados rotulados não foi interpretada como uma etapa única. Em cada iteração, as previsões do modelo sobre dados fora da amostra — i.e., o conjunto de dados retido durante o treino — foram reaproveitadas como entrada para análise de erros. A partir disso, dados adicionais foram estrategicamente rotulados para complementar os casos em que o modelo teve um desempenho fraco, repetindo-se este ciclo de atividades até conseguir alcançar um resultado desejável.

[6] https://dre.pt/
[7] https://eur-lex.europa.eu/homepage.html?locale=pt
[8] https://www.selenium.dev/

O procedimento real de rotulagem de dados foi conduzido através de uma aplicação *web* onde tanto os especialistas em avaliação de impacto e juristas da equipa como os especialistas de ciência de dados conseguiram trabalhar de forma colaborativa. A Figura 4 mostra a interface de anotação para uma passagem de texto específica.

Figura 4 — Interface de anotação de rotulagem de dados baseada utilizada no projeto AI4IA



Fonte: relatório final do projeto.

Reduzir o ruído nos dados

Após várias dezenas de documentos terem sido rotulados, a contagem combinada de passagens de texto com encargos administrativos atingiu apenas 322 exemplos, enquanto o corpus combinado dos documentos totalizou 19 465 passagens de texto. Nestes cenários, com um desequilíbrio tão significativo – em média, apenas 1,68% das passagens dos documentos continham obrigações – os modelos podem exibir preconceito em relação à classe maioritária e ignorar completamente a classe minoritária. (Johnson e Khoshgoftaar, 2019)

Com tal constrangimento, foram exploradas técnicas para aumentar a qualidade dos dados e reduzir a necessidade de rotulagem manual. Na prática, foi construído o módulo redutor de ruído. Esta componente de pré-processamento do *pipeline* da prova de conceito concentra-se na remoção de passagens que anteriormente eram identificadas por não conterem encargos administrativos. É seguida uma série de regras definidas em conjunto com os peritos em avaliação de impacto, aproveitando o seu conhecimento prévio da estrutura da legislação. Depois, este módulo foi projetado

para funcionar como uma técnica automatizada. O funcionamento do algoritmo é resumido da seguinte forma:

- Identifica automaticamente — e rotula como não tendo obrigações — títulos, legendas, passagens que apenas contêm números de artigos e símbolos, entre outros trechos de texto com conteúdo irrelevante para a avaliação de impacto.
- Os preâmbulos dos documentos também não são tidos em consideração, uma vez que os encargos administrativos, em regra, apenas são indicados nos artigos do diploma.
- No caso de uma proposta de diploma que altera outro diploma existente, o módulo pode identificar apenas as novas passagens.
- Extrai artigos inteiros que indicam a ausência de encargos administrativos através do contexto semântico subjacente.

Como forma de comparação, durante a avaliação experimental no AI4IA, o módulo de redução de ruído foi capaz de mitigar o desequilíbrio inicialmente verificado, aumentando a percentagem média de passagens ou trechos que continham encargos administrativos de 1,68% para 5,89%.

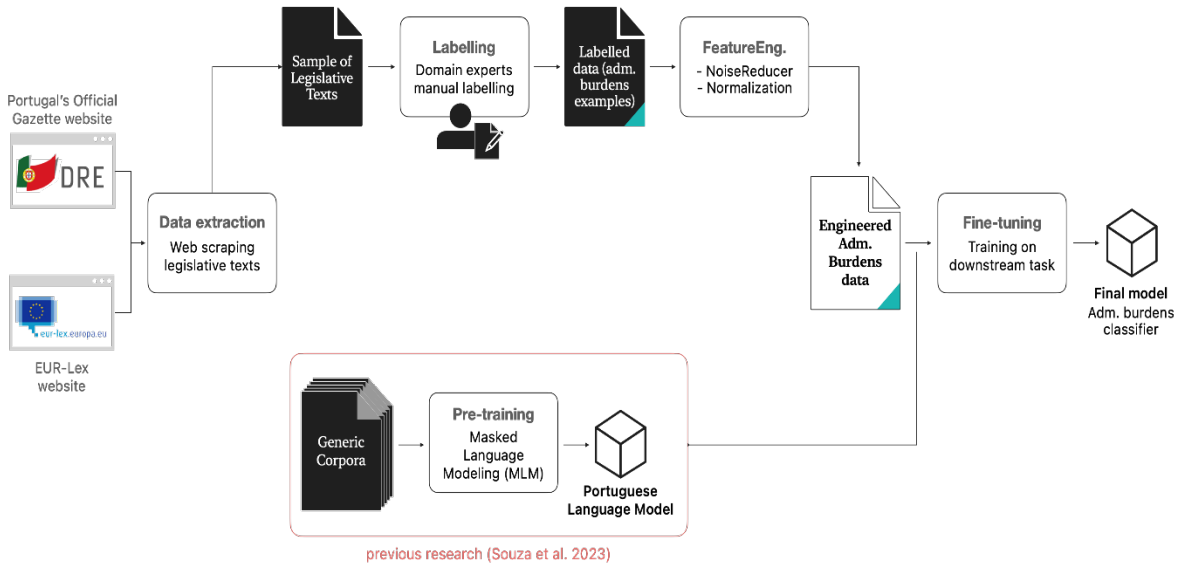
Normalização de texto

Juntamente com a rotulagem de passagens irrelevantes efetuada com recurso ao módulo Redutor de Ruído, uma etapa de normalização também foi incluída para padronizar o formato dos dados de entrada. Especificamente, todos os marcadores e prefixos de numeração foram removidos e espaços em branco foram desconsiderados.

Avaliação experimental

A configuração experimental para desenvolver a principal capacidade da prova de conceito AI4IA. Especificamente, a Figura 5 possibilita obter uma visão geral de como decorreu o processo de ajuste e treino do modelo de "filtro" de NLP para identificar encargos administrativos.

Figura 5 — Visão sistémica do processo de desenvolvimento do modelo



Fonte: relatório final do projeto.

Execução técnica

Em relação aos detalhes de implementação da fase experimental, o treino do modelo foi apoiado principalmente pelo ecossistema de ferramentas e bibliotecas *Hugging Face*. O código foi desenvolvido em módulos Python específicos, nomeadamente:

- ***Datasets***: esta biblioteca é utilizada para carregar e gerenciar conjuntos de dados. Ela fornece uma interface fácil de utilizar para aceder e pré-processar vários conjuntos de dados que são essenciais para o treino e avaliação do modelo.
- ***Evaluate***: esta biblioteca é utilizada para avaliação de modelos. Fornece várias métricas e funções de avaliação com vista a avaliar o desempenho do modelo treinado, garantindo que cumpre padrões de precisão e eficiência desejados.
- ***Transformers***: a biblioteca de transformadores da *Hugging Face* é um conjunto abrangente de ferramentas projetadas para trabalhar com modelos de processamento de linguagem natural de última geração. Os módulos específicos utilizados a partir desta biblioteca incluem:
    - ***AutoTokenizer***: Este módulo é usado para representar os dados de texto de entrada. A tokenização é uma etapa crucial na preparação de dados de texto para o modelo, pois converte texto bruto num formato que o modelo pode compreender.
    - ***DataCollatorWithPadding***: Este módulo ajuda no processamento em lote dos dados, garantindo que todas as sequências num lote tenham o mesmo comprimento. Na prática, preenche as sequências até o comprimento máximo dentro do lote, facilitando o processamento eficiente pelo modelo.

- ***AutoModelForSequenceClassification***: O módulo que aproxima os pesos pré-treinados, adicionando um cabeçalho para tarefas de classificação. É utilizado como modelo base que será ajustado no conjunto de dados específico.
- ***TrainingArguments***: Este módulo é utilizado para especificar vários parâmetros de treino, como a taxa de aprendizagem, tamanho do lote, número de época, etc. Ajuda na configuração do processo de treino de acordo com os requisitos da tarefa.
- ***Trainer***: Este módulo compõe o processo de formação. Integra o modelo que representa, e recolhe os dados, argumentos de treino e o conjunto de dados para treinar e avaliar o modelo de forma perfeita.

Resultados e discussão

Nesta secção, são avaliados os resultados alcançados durante a modelização. O projeto visa superar uma precisão de 0,50 e uma *recordação (recall)* de 0,5 para a tarefa binária de classificação de encargos. Juntamente com esse limiar que representa um desempenho melhor do que um classificador aleatório, no caso da identificação de encargos administrativos, esse objetivo torna-se difícil devido ao desequilíbrio extremo no conjunto de dados. Portanto, alcançar tal resultado valida a hipótese de um modelo de inteligência artificial ser capaz de aprender o contexto semântico que representa uma obrigação de informação.

No que respeita à comparação inicial entre os 3 modelos, o BERTimbau Base obteve o melhor resultado com uma *recordação (recall)* de 0,715, Precisão de 0,62 e F1 de 0,664. Esse resultado já supera o limite desejado para o projeto. No entanto, os hiperparâmetros da Base BERTimbau foram otimizados.

A partir dos resultados de ajuste de hiperparâmetros, o melhor modelo foi produzido com 3 épocas, um tamanho de lote de 32 amostras e uma taxa de aprendizagem de 5e-5 para o optimizador AdamW — i.e., algoritmo Adam com correção de "peso" (Loshchilov e ainda Hutter 2017). Estes foram os valores considerados para a construção do modelo final que alavancou todos os dados de desenvolvimento disponíveis. A Tabela 1 resume os resultados alcançados pelos modelos considerados. O modelo ajustado mostra uma *recordação* (recall) de 0,762, precisão de 0,611 e pontuação F1 de 0,678.

Tabela 1 — Resultados do modelo "F1", "P" e "R" especificam as pontuações F1, Precisão e recordação (*recall)*

Modelo	F1	P	R
Base BERTimbau [avaliação inicial]	**0,664**	**0,62**	**0,715**
BERTimbau Large [avaliação inicial]	0,519	0,470	0,575
Albertina 100M PTPT [avaliação inicial]	0,582	0,552	0,615
BERTimbau Base' [avaliação final]	**0,678**	**0,611**	**0,762**

Fonte: relatório final do projeto.

O sistema é atualmente capaz de identificar, em média, 76,2 % dos encargos administrativos presentes num diploma (i.e., a pontuação de *recordação*). Por outro lado, a pontuação de precisão indica que, de todas as passagens que o sistema devolve, 61,1% realmente contêm uma IO.

Não obstante, para o uso diário desta prova de conceito, pode-se argumentar que é preferível ter uma recordação (*recall*) maior, mesmo que isso signifique sacrificar alguma precisão. Isto deve-se ao facto de que é menos moroso analisar se as passagens identificadas contêm um encargo administrativo do que procurar as que o sistema não devolve/identifica em todo o texto do diploma.

www.planapp.gov.pt

PlanAPP

@planapp_

Newsletter

Parceiros

NOVA IMS Information Management School

PLANAPP
PLANEAMENTO | POLÍTICAS | PROSPETIVA

Financiamento

Financiado pela União Europeia