Abordagem Cientométrica Orientada a Dados para Classificação Multi-Alvo dos Objetivos de Desenvolvimento Sustentável na Automação

Alexandre Dias [*] **Gisliany Alves** [*] **Germano Lima** [*] **Ariel Alsina** [**] **Ivanovitch Silva** [*,**]

[*] *Programa de Pós-Graduação em Engenharia Elétrica e de Computação, Universidade Federal do Rio Grande do Norte, RN, (e-mail: {alexandre.dias.105,gisliany.alves.094, germano.yoneda.362}@ufrn.edu.br)*
[**] *Departamento de Engenharia de Computação e Automação, Universidade Federal do Rio Grande do Norte, RN, (e-mail: ariel.alsina.110@ufrn.edu.br, ivanovitch.silva@ufrn.br)*

Abstract: The United Nations created the 17 Sustainable Development Goals (SDGs) to promote environmental protection, economic growth, and social justice. In this scenario, science is crucial to solving the challenges addressed by the SDGs. SciVal, for example, is a tool that tracks scientific publications related to the SDGs with the support of a team of experts. Aiming to reduce the need for specialized knowledge and to provide a more autonomous tool, this study proposes a multi-label classification model based on natural language processing and recurrent neural networks to map scientific publications to the SDGs. The proposed model is tested with the articles of the Brazilian Congress of Automatics (CBA) 2020. The data used to train the model comprises manuscript titles acquired from the Scopus database using the SciVal analytics tool, and they are related to 16 out of the 17 SDGs. Results have shown that the papers published in the CBA 2020 focused on SDGs 7, and 9, which are related to clean energy and industry innovation. Furthermore, all SDGs were associated with at least one publication, indicating that intelligent automation can contribute in a interdisciplinary way to the SDGs implementation.

Resumo: As Nações Unidas criaram os 17 Objetivos de Desenvolvimento Sustentável (ODS) visando promover a proteção ambiental, o crescimento econômico e a justiça social. Nesse cenário, a ciência é crucial para a resolução dos desafios abordados pelos ODS. O *SciVal*, por exemplo, é uma ferramenta que rastreia as produções científicas relacionadas aos ODS a partir de uma análise feita por especialistas. Com o intuito de reduzir a necessidade de conhecimento especializado e de prover uma ferramenta mais autônoma, este estudo propõe um modelo de classificação multi-alvo baseado em processamento de linguagem natural e redes neurais recorrentes, a fim de mapear publicações científicas aos ODS. O modelo proposto é testado sobre as publicações do Congresso Brasileiro de Automática (CBA) 2020. Os dados utilizados na etapa de treinamento compreendem títulos de publicações coletadas da *Scopus*, por meio da própria ferramenta *SciVal*, que abrange 16 dos 17 ODS. Resultados evidenciaram que os temas mais frequentes abordados no CBA 2020 estão vinculados aos ODS 7 e 9, sobre energia limpa e inovação industrial. Ademais, todos os ODS associaram-se a pelo menos uma publicação, indicando que a automação inteligente pode contribuir de forma interdisciplinar com a implementação dos ODS.

Keywords: Sustainable Development Goals; Scientometrics; Natural Language Processing; Recurrent Neural Networks; Multilabel Classification.

Palavras-chaves: Objetivos de Desenvolvimento Sustentável; Cientometria; Processamento de Linguagem Natural; Redes Neurais Recorrentes; Classificação Multi-alvo.

1. INTRODUÇÃO

Nos últimos anos, a comunidade global tem reunido esforços em prol de um presente e um futuro sustentáveis. Em 2015, os Estados-membros da Organização das Nações Unidas (ONU) adotaram a Agenda 2030, um plano de ação centrado em 17 Objetivos de Desenvolvimento Sustentável (ODS), 169 metas e 232 indicadores para guiar regulamentações e investimentos que garantam melhorias sociais, econômicas e ambientais (Sachs et al., 2019). Os ODS foram criados para serem universais e focados em abordagens integrativas que unam o desenvolvimento humano à sustentabilidade ambiental (ElAlfy et al., 2020).

Nesse cenário, a ciência é crucial na defesa de estratégias baseadas em evidências e na realização de pesquisas que permitam o cumprimento dos ODS (Allen et al., 2021). As soluções da ciência, tecnologia e inovação podem direcionar a produção científica para identificar as barreiras e oportunidades, bem como os avanços tecnológicos necessários em favor do alcance dos ODS, além de promover igualdade, inclusão e efeitos ambientais positivos (Walsh et al., 2020).

Fica evidente que as instituições de pesquisa e fomento têm papel vital nesse processo. Para contribuir nesse aspecto, a empresa *Elsevier*, por exemplo, firmou uma colaboração com universidades ao redor do mundo para estabelecer uma iniciativa de mapeamento da pesquisa científica aos ODS (Elsevier, 2021b). Para isso, a *Elsevier* elaborou e aplicou majoritariamente consultas avançadas criadas por especialistas para 16 dos 17 ODS, complementando-as com o uso de técnicas de Processamento de Linguagem Natural (PLN) instrumentadas com um modelo de regressão logística (Rivest et al., 2021). O resultado desse mapeamento foi incorporado ao *SciVal* — uma plataforma web da *Elsevier* baseada em dados da *Scopus*, com um portfólio de ferramentas para análise cientométrica da produção científica —, possibilitando o acompanhamento das publicações por ODS e a descoberta de lacunas e tópicos proeminentes de pesquisa (Elsevier, 2021a).

Face à necessidade em se priorizar a implementação dos ODS, trazer esse mesmo mapeamento cientométrico da produção científica para o contexto nacional brasileiro apresenta-se como tarefa indispensável. De forma particular, o Congresso Brasileiro de Automática (CBA), promovido pela Sociedade Brasileira de Automática (SBA) tem grande potencial para contribuir nesse aspecto. O CBA tem sido um congresso itinerante pelo país que tem registrado uma participação crescente e foca em áreas como automação, controle, eletrônica, ciência de dados, robótica, sistemas inteligentes e muitos outros campos (SBA, 2022) com impacto direto no desenvolvimento sustentável. De acordo com Khamis et al. (2019), a automação aliada à inteligência artificial tem sido componente central do desenvolvimento tecnológico e fator determinante por trás dos avanços em direção ao alcance dos ODS, além de auxiliar no combate a problemas das esferas humanitária, econômica e ambiental.

Dessa forma, a pesquisa científica e os avanços da automação precisam ser direcionados para os ODS, integrando também a produção científica brasileira a esse contexto. No entanto, a tarefa de mapeamento do *SciVal* requer um grupo de especialistas para atualizar e validar as consultas. Nesse cenário, para reduzir a necessidade de conhecimento especializado e de prover uma ferramenta mais autônoma e inteligente, este trabalho propõe o uso de técnicas de PLN aliadas a um modelo de Rede Neural Recorrente (do inglês, *Recurrent Neural Networks* ou RNN) com aprendizagem supervisionada para o mapeamento multi-alvo das publicações do CBA 2020 aos ODS. O estágio de treinamento do modelo utiliza dados de publicações coletados por meio da plataforma *SciVal* entre 2019 e 2022, considerando precisamente o título dos artigos e os ODS associados. Vale ressaltar que o *SciVal* disponibiliza artigos classificados apenas em 16 dos ODS, sendo esses os rótulos utilizados para o desenvolvimento deste trabalho.

Diferente da ferramenta *SciVal*, este trabalho emprega técnicas de aprendizagem profunda para realizar o mapeamento multi-alvo no contexto do desenvolvimento sustentável, sem depender de conhecimento especializado sobre os ODS, sendo essa sua principal contribuição. Além disso, a inferência dos ODS para o CBA 2020 contribuirá para o alinhamento dos avanços da automação inteligente no cenário brasileiro e, portanto, no contexto de uma nação em desenvolvimento, onde se supõe que o progresso dos ODS seja uma tarefa desafiadora. Por fim, a metodologia proposta pode ser adaptada para o mapeamento dos ODS em outras bases de dados de produção acadêmica.

Além desta introdução, este trabalho apresenta, na Seção 2, o estado da arte, desafios e limitações referentes ao uso de PLN e de inteligência artificial na implementação dos ODS. Na sequência, a Seção 3 trata da fundamentação teórica necessária para o entendimento da metodologia apresentada na Seção 4. Os resultados e suas implicações são relatados na Seção 5. Por fim, as conclusões são discutidas na Seção 6.

2. TRABALHOS RELACIONADOS

Uma das principais limitações relacionadas às tarefas de linguagem natural é a da interpretação de informações implícitas no texto ou utilizadas em contextos específicos. A habilidade da máquina em compreender contextos é nomeada de inferência de linguagem natural (Storks et al., 2019). Soma-se a isso o desafio de representar palavras de forma numérica. Visando resolver esse problema, inúmeros modelos de representação gerais foram desenvolvidos, sendo capazes de agrupar sintaticamente e semanticamente palavras similares sem utilizar rótulos (Yu et al., 2020). Além disso, palavras de um domínio específico, fora do vocabulário ou desconhecidas também podem aumentar a complexidade dos problemas de PLN.

Nesse cenário, alguns trabalhos relacionados aos ODS são descritos na literatura, sendo um deles proposto por Smith et al. (2021). Os autores utilizaram os relatórios da Organização das Nações Unidas para Economia e Conselho Social ao secretário geral da ONU de 2016 a 2020, totalizando 85 relatórios. Cada um desses documentos foram agrupados de acordo com as 17 iniciativas da ONU. A partir desse arranjo, foram implementadas técnicas de tokenização e lematização como uma forma de pré-processamento e em seguida aplicado um modelo de vetorização (Doc2Vec) para criar representações numéricas das palavras e permitir a geração de métricas de similaridade entre os ODS (Smith et al., 2021).

Outro estudo realizado por Hsu et al. (2022) utilizou dois modelos para classificar os ODS. Um modelo classificou os tópicos utilizando *Latent Dirichlet allocation* (LDA). O outro modelo aplicou ligação semântica sem depender de um conjunto de treinamento ao usar a *Semantic Web* para mensurar o grau de conexão entre as publicações. A fusão do resultado desses modelos é proposta de forma a melhorar a classificação geral.

Um estudo envolvendo classificação com múltiplos alvos foi realizado por Matsui et al. (2022), que analisou documentos em japonês relacionados aos ODS empregando *Bidirectional Encoder Representations from Transformers* (BERT). BERT foi projetado para treinar representações bidirecionais profundas de texto não rotulado, utilizando o contexto de ambos os lados em todas as camadas, e podendo ser aplicado em inúmeras tarefas apenas ajustando a última camada (Devlin et al., 2019). Técnicas de validação cruzada aninhada foram conduzidas para definir os parâmetros do modelo, que foi avaliado conforme as métricas de precisão, *recall* e *F1 Score*.

A metodologia desenvolvida neste trabalho se diverge das demais por utilizar:

- Uma arquitetura própria, sem modelos com transferência de aprendizado para classificação multi-alvo, diferente do realizado por Matsui et al. (2022);
- Uma metodologia de treinamento supervisionado, diferente de Smith et al. (2021) e Hsu et al. (2022);
- Apenas o título dos artigos para a classificação multi-alvo, demonstrando que com poucos *tokens* é possível construir um modelo que reconhece os ODS;
- Uma função de custo ponderada descrita na subseção 3.2.

3. FUNDAMENTAÇÃO TEÓRICA

Esta seção aborda a fundamentação teórica e a motivação por trás dos artifícios utilizados na implementação do classificador de ODS. Mais especificamente, ela aborda Redes Neuras Recorrentes, Função de Custo, e as métricas de avaliação de desempenho no contexto de classificação em múltiplos alvos.

3.1 Redes Neurais Recorrentes

As RNNs são especialmente desenhadas para processar informações sequenciais como sentenças de texto, séries temporais ou áudio. No entanto, sabe-se que RNNs simples têm duas principais limitações: a instabilidade dos gradientes e o problema de memória curta.

Por tratar-se de um modelo de aprendizagem profunda, os gradientes utilizados para atualizar os pesos das RNNs podem desestabilizar a aprendizagem do modelo. Algumas técnicas são utilizadas para contornar este problema, como a *recurrent dropout* e a *recurrent layer normalization*, introduzidas por Semeniuta et al. (2016) e Ba et al. (2016), respectivamente.

Por outro lado, no processamento de uma sequência longa, RNNs podem perder de vista as informações contidas no início da sequência. Dessa forma, características importantes das sequências são "esquecidas". Para lidar com esse problema, foram introduzidas às RNNs células mais eficientes. Primeiro, foram desenvolvidas as células *Long-Short Term memory* (LSTM), introduzidas por Hochreiter and Schmidhuber (1997) e posteriormente melhoradas por Sak et al. (2014) e Zaremba et al. (2014).

Com o sucesso das LSTM, surgiram suas variantes, incluindo a celebrada *Gated Recurrent Unit* (GRU) (Cho et al., 2014). As células GRU são consideradas como uma versão simplificada das LSTM. Não obstante, Greff et al. (2017) mostraram que não há diferença significativa de desempenho entre os dois tipos de células. Ainda assim, células GRU são mais rápidas ao processar dados, o que as torna atrativas no contexto de aprendizagem de máquina envolvendo um volume considerável de dados.

Por esses motivos, optou-se por utilizar células GRU na arquitetura da RNN implementada neste trabalho, juntamente com a aplicação da técnica de *recurrent dropout* nas camadas ocultas do modelo.

3.2 Função de custo ponderada

Em problemas de classificação binária, comumente utiliza-se a função de entropia cruzada (BCE) (do inglês, *binary cross entropy*) para medir o erro de classificação do modelo. Alternativamente, em um problema de classificação multi-alvo, um conjunto de dados $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ apresenta n instâncias em que cada instância i pode pertencer a k classes, isto é, $y_i = [y_i^{(1)}, \cdots, y_i^{(k)}] \in \{0,1\}^k$. Nesse cenário, considera-se também um classificador multi-alvo com saída $z_i = [z_i^{(1)}, \cdots, z_i^{(k)}] \in \mathbb{R}^k$, cuja função de custo médio, C_{BCE}, é definida em (1) como:

$$C_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}(y_i^{(j)}\log(p_i^{(j)}) + (1-y_i^{(j)})\log(1-p_i^{(j)})). \tag{1}$$

Onde $p_i^{(j)}$ é o resultado da aplicação da função sigmóide (σ) à saída do classificador, isto é, $p_i^{(j)} = \sigma(z_i^{(j)})$. Em particular, para uma instância, pode-se omitir o cálculo da média, e escrever:

$$C_{BCE} = \begin{cases} -\log(p_i^{(j)}), & \text{se } y_i^{(j)} = 1, \\ -\log(1-p_i^{(j)}), & \text{se } y_i^{(j)} = 0. \end{cases} \tag{2}$$

A C_{BCE} (2) padrão considera que todas as classes tem o mesmo peso sobre a classificação. Segundo Durand et al. (2019), se o conjunto de dados é desbalanceado com relação às classes, a C_{BCE} prioriza as instâncias das classes mais ocorrentes sobre aquelas menos ocorrentes.

Huang et al. (2021) e Ho and Wookey (2020) listam vários métodos para lidar com desbalanceamento de classes em problemas de classificação multi-alvo como, por exemplo, a BCE ponderada (WBCE, do inglês, *weighted binary cross entropy*). A WBCE para uma instância é definida por (3):

$$C_{WBCE} = \begin{cases} -w_1^{(j)}\log(p_i^{(j)}), & \text{se } y_i^{(j)} = 1, \\ -w_0^{(j)}\log(1-p_i^{(j)}), & \text{se } y_i^{(j)} = 0. \end{cases} \tag{3}$$

Onde $w_0^{(j)}$ e $w_1^{(j)}$ são os pesos das instâncias negativas e positivas da j-ésima classe, respectivamente. $w_0^{(j)}$ e $w_1^{(j)}$ podem ser escolhidos arbitrariamente.

O conjunto de dados utilizado neste trabalho tem 16 classes, cada uma com cerca de 12% de instâncias positivas. Logo, há uma desbalanceamento de aproximadamente 1:7 entre instâncias positivas e negativas. Para mitigar esse problema, utilizou-se a função C_{WBCE}, definindo $w_0^{(j)} = 1$ e $w_1^{(j)} = 2$, para dar o dobro de "atenção" para as instâncias positivas de cada classe durante o treinamento.

3.3 Métricas

O desempenho de modelos de classificação para problemas com múltiplas classes, segundo Grandini et al. (2020), pode ser avaliado através de algumas métricas clássicas, *e.g.*, acurácia, *recall*, precisão e *F1-score*. Nesses casos, apenas uma classe dentre as definidas no conjunto de dados é atribuída a cada amostra.

Contudo, o problema de classificação abordado neste trabalho distingue-se por atribuir um conjunto de classes a cada instância. Em outras palavras, trata-se de uma classificação multi-alvo. Desta forma, as predições, assim como é dito por Sorower (2010), podem assumir, além das noções de total corretude e incorretude, uma noção de corretude parcial: situação em que nem todas as classes foram adequadamente previstas.

Tomando-se as variáveis definidas na seção 3.2, e definindo $\hat{y}_i = [\hat{y}_i^{(1)}, \cdots, \hat{y}_i^{(k)}]$ como o conjunto de classes preditas para a instância y_i, algumas das métricas utilizadas e descritas por Sorower (2010) são:

Proporção exata (Exact Match Ratio, EMR)

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i), \tag{4}$$

$$I(y_i = \hat{y}_i) = \begin{cases} 1, & \text{se } y_i = \hat{y}_i, \\ 0, & \text{caso contrário.} \end{cases}$$

A proporção exata (4) quantifica uma proporção entre todas as amostras em que todas as classes estão corretamente previstas. Uma desvantagem dessa métrica é que ela não distingue predições totalmente incorretas e parcialmente corretas.

Acurácia Geral

$$\frac{1}{n}\sum_{i=1}^{n} \frac{|y_i \wedge \hat{y}_i|}{|y_i \vee \hat{y}_i|}. \tag{5}$$

A acurácia geral (5) é a média aritmética da proporção de predições corretas positivas em relação ao total de classes positivas, para cada amostra. Todavia, ela pode ser enviesada em conjuntos de dados que apresentam classes desbalanceadas.

Recall

$$\frac{1}{n}\sum_{i=1}^{n} \frac{|y_i \wedge \hat{y}_i|}{|y_i|}. \tag{6}$$

Recall (6) é a média da proporção de classes positivas corretamente inferidas e o número total de classes positivas por amostra.

Precisão

$$\frac{1}{n}\sum_{i=1}^{n} \frac{|y_i \wedge \hat{y}_i|}{|\hat{y}_i|}. \tag{7}$$

A precisão geral (7) do classificador pode ser entendida como a média da acurácia das predições positivas para todas as amostras.

F1-Score

$$\frac{1}{n}\sum_{i=1}^{n} \frac{2|y_i \wedge \hat{y}_i|}{|y_i| + |\hat{y}_i|}. \tag{8}$$

F1-Score (8) é a média harmônica entre *Recall* e a *Precisão*. Um F1-Score alto indica que as duas métricas que o compõe estão altas.

4. METODOLOGIA

Esta seção aborda a metodologia empregada para a aquisição de dados, implementação do classificador de ODS, e a etapa de inferência do modelo sobre as publicações do CBA 2020.

4.1 Aquisição e conjunto de dados

Os dados utilizados no treinamento do modelo foram obtidos manualmente da base de dados *Scopus*, através da ferramenta *SciVal*. Na coleta de dados, consideraram-se apenas artigos publicados entre 2019 e 2022, e relacionados com ao menos um ODS.

Ao todo, cerca de 1,3 milhões de publicações foram extraídas, contendo informações como autores, fator de impacto, etc. Para os propósitos deste trabalho, utilizou-se apenas a informação do título das publicações. A Tabela 1 ilustra o conjunto de dados, mostrando três dos registros utilizados para treinamento do modelo.

Tabela 1. Amostra dos dados coletados a partir da plataforma *SciVal*.

Título	**ODS**
A Social Vulnerability Index for Disaster Management	ODS 1
Assessing differential impacts of COVID-19 on black communities	ODS 3 | ODS 10 | ODS 14
Research on Irrigation Water Efficiency of Guizhou Province Based on SFA	ODS 6 | ODS 13

4.2 Preparação dos dados

A etapa de preparação dos dados é realizada imediatamente antes do pré-processamento textual ilustrado na Figura 1. Na preparação, é feita a binarização dos rótulos dos ODS que, como pode ser visto na Tabela 1, são obtidos originalmente em formato de texto separados pelo caractere barra vertical ou *pipe* (|).

Em seguida, é realizada uma subamostragem para balancear o conjunto de dados com relação aos ODS. Esse

processo é necessário pois o número de registros de alguns ODS excede em muito a quantidade média de registros por ODS, o que prejudica o treinamento do modelo e a avaliação das métricas.

Ainda, como os dados foram obtidos manualmente, alguns registros foram coletados repetidamente, gerando muitas entradas duplicadas. Os registros duplicados foram removidos. Ao final, o conjunto de dados foi separado em três subconjuntos para treinamento (com 201.892 registros), validação (com 22.432 registros) e teste (com 56.082 registros) do modelo.

4.3 Estágio de treinamento

O pré-processamento textual dos títulos das publicações é realizado logo após a preparação dos dados. Na Figura 1, pode-se ver que a camada de pré-processamento textual é composta por alguns módulos, que realizam operações tais como a conversão para letras minúsculas, remoção de acentos e caracteres especiais, lematização, remoção de *stopwords* (palavras que podem ser consideradas irrelevantes) e filtragem de registros vazios ao fim do *pipeline*. Todos esses passos foram empregados objetivando a normalização das sentenças a fim de garantir que o modelo seja alimentado com dados padronizados. O pré-processamento é aplicado tanto ao conjunto de dados obtido do *SciVal*, utilizado para o treinamento, quanto durante o estágio de inferência no conjunto de dados do CBA, que será melhor explicado na subseção 4.4.

Todos os componentes do modelo foram desenvolvidos utilizando o *framework* de código aberto, *Keras* (Chollet et al., 2015). O treinamento do modelo foi gerenciado e rastreado pela ferramenta *Weights & Biases* (Biewald, 2020), que conta com um módulo que realiza busca de parâmetros utilizando o método de validação cruzada de busca em *grid*, chamado *Sweeps*. Os melhores parâmetros para o modelo foram definidos segundo o seu desempenho de acordo com as métricas citadas na Seção 3.3. Posteriormente, algumas adaptações foram aplicadas ao modelo com a finalidade de estabilizar os resultados do treinamento e viabilizar o uso de unidades de processamento gráfico (GPU, do inglês, *graphics processing unit*). Essas adaptações são:

- A taxa de aprendizagem foi recondicionada a comportar-se com um decaimento exponencial. Esse método possibilita estabilizar os resultados do treinamento, pois evita que o modelo oscile em torno da solução ótima.
- A taxa de *dropout* das unidades da transformação linear do estado recorrente das células GRU foi redefinida de 0,3 para 0. Isso foi feito para satisfazer um pré-requisito do *Keras* para o uso de GPUs.
- Após o procedimento anterior, foi usado um artifício alternativo para a regularização dos pesos da RNN. A estratégia adotada foi a de restringir os pesos das camadas ocultas até o valor máximo de sua norma euclidiana (cinco). Resultados empíricos mostraram que essa abordagem obteve êxito, evitando o sobreajuste do modelo.

O modelo final é obtido ao término do estágio de treinamento. Suas métricas de desempenho e mais detalhes são comentados na próxima seção.

4.4 Aquisição de dados do CBA 2020

Uma vez que o modelo final é obtido, é realizada a etapa de inferência utilizando dados do CBA 2020 extraídos do próprio site do CBA via *web scraping*.

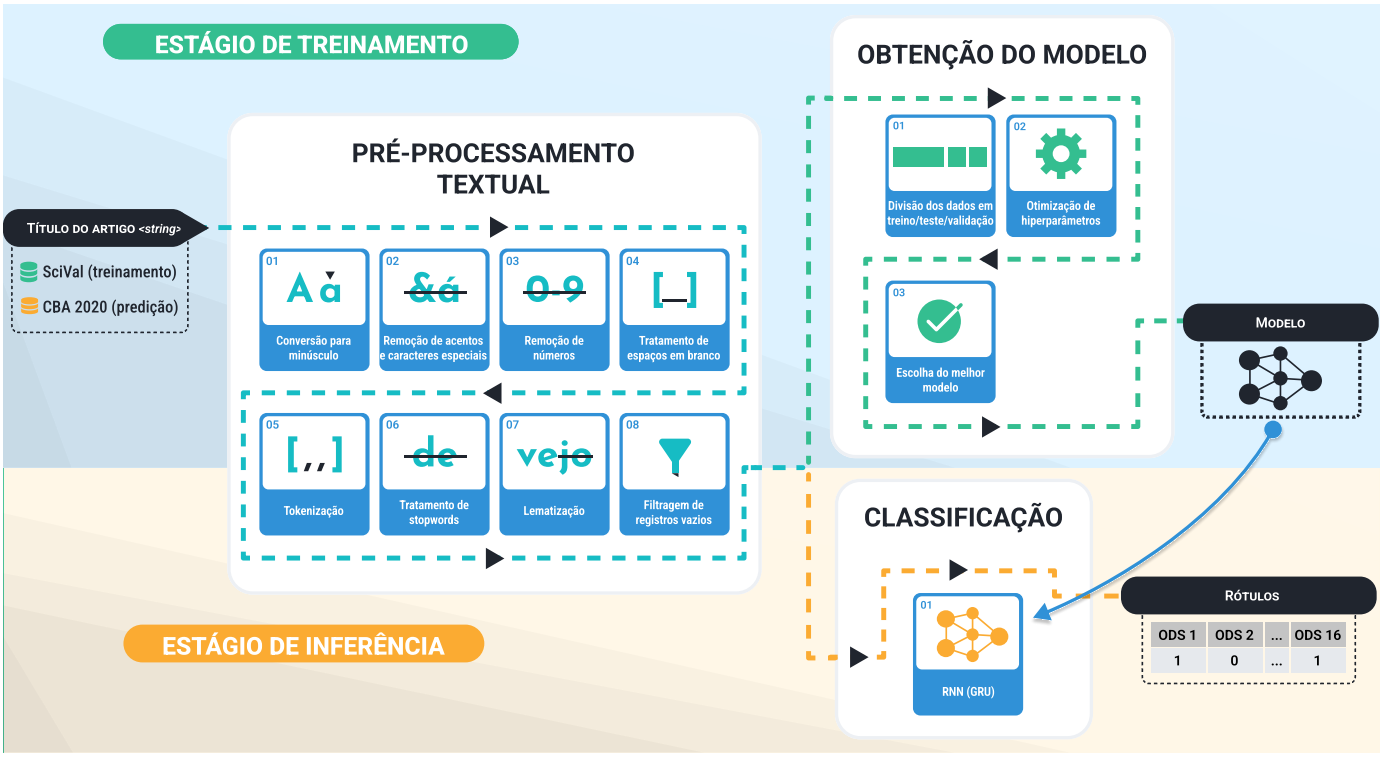

Figura 1. Estágios de treinamento e inferência do modelo.

O modelo foi treinado usando sentenças em língua inglesa. Portanto, foi necessário traduzir os títulos de publicações do CBA 2020. A tradução foi feita via uma interface de programação de aplicações (API, do inglês, *application programming interface*) capaz de enviar requisições de tradução à ferramenta online *Google Translate*.

Dessa maneira, foi possível obter e traduzir para o inglês os 777 títulos listados na página de anais do CBA 2020.

5. RESULTADOS

Os dados utilizados para treinar o modelo foram consumidos por lotes de 32 sentenças (títulos) cada. A arquitetura da RNN final é retratada na Figura 2. Ela conta com uma camada de vetorização textual, responsável por tornar sentenças textuais em vetores numéricos com 50 entradas. Em seguida, essa camada é conectada a uma camada de *embeddings*, que são vetores treináveis, isto é, passíveis de aprendizagem por parte do modelo, usados para representar uma palavra ou *token* em um espaço multidimensional. Em particular, a arquitetura do modelo usa *embeddings* de 50 dimensões para representar cada palavra das sentenças.

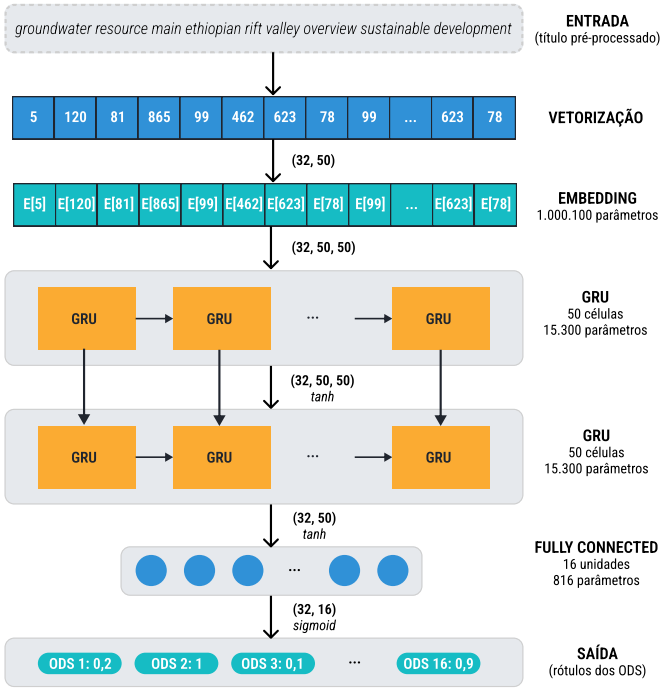

Figura 2. Arquitetura do modelo obtido.

Posteriormente, os *embeddings* são transmitidos para primeira camada oculta da RNN. As camadas ocultas da RNN são compostas por células GRU, com função de ativação tangente hiperbólica (*tanh*, do inglês, *hyperbolic tangent*). As saídas da segunda camada oculta passam por uma camada densa de 16 neurônios ativados pela função sigmóide. Cada saída do modelo representa a probabilidade de um título estar relacionado com uma das ODS.

Todos os parâmetros da arquitetura foram determinados empiricamente pelo processo de validação cruzada usando a ferramenta *Weights & Biases*, citada na Seção 4.2. Ao todo, a RNN conta com 1.031.516 parâmetros treináveis.

5.1 Métricas de desempenho

A avaliação do modelo abrange duas perspectivas distintas: A primeira aborda aspectos de desempenho gerais, enquanto a segunda destina-se a compreender a performance do modelo específica a cada ODS.

Em relação aos aspectos de desempenho gerais sobre o conjunto de dados de teste, o modelo apresentou uma proporção exata de $0,4047$, acurácia de $0,6413$, precisão de $0,7641$, *recall* de $0,7327$ e *F1-score* de $0,7184$.

Sob outra perspectiva, as métricas de desempenho por ODS são expressas na Figura 3, que exibe a acurácia, precisão, *recall*, *F1-Score* e a área sob a curva de características operacionais do receptor (ROC AUC, do inglês, *Receiver Operating Characteristic Area Under the Curve*).

Vale salientar que, neste caso, as métricas são definidas de maneira usual, isto é, como em um problema de classificação binária. A obtenção dessas métricas dá-se isolando as predições do modelo multi-alvo para cada ODS individualmente.

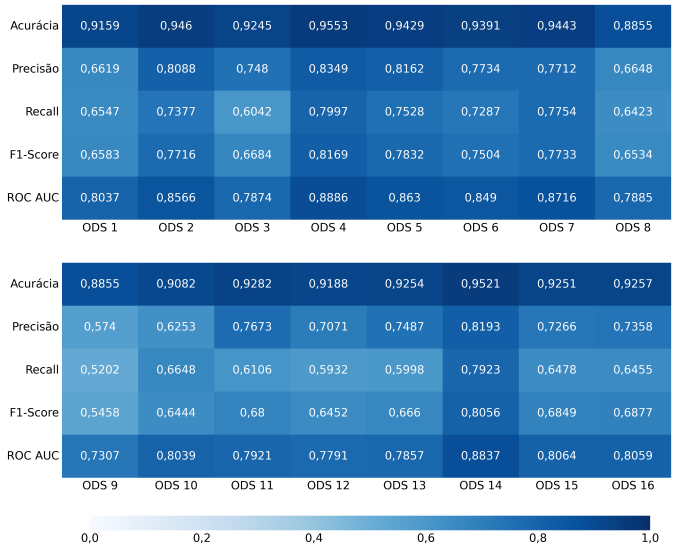

Figura 3. Métricas binárias para cada ODS.

5.2 Inferência sobre os dados do CBA 2020

Com o intuito de mapear as contribuições científicas apresentadas no CBA 2020, foram realizadas inferências sobre os títulos de artigos publicados naquela edição. Como resultado, o histograma ilustrado na Figura 4 mostra a frequência de ocorrências dos ODS nas publicações.

Percebe-se que os dois ODS mais frequentes são o ODS 7 (Energia Acessível e Limpa) e o ODS 9 (Indústria, Inovação e Infraestrutura). Em contraste, os ODS menos abordados foram os ODS 5 e 12, relacionadas aos temas de "Igualdade de Gênero" e "Consumo e Produção Responsáveis", cada um com apenas quatro ocorrências.

Além disso, foi analisada a co-ocorrência de ODS entre as publicações, cujas quantidades podem ser observadas na Tabela 2. Nota-se que 85 títulos não obtiveram nenhuma classificação, sendo considerados como "Indefinidos". Para

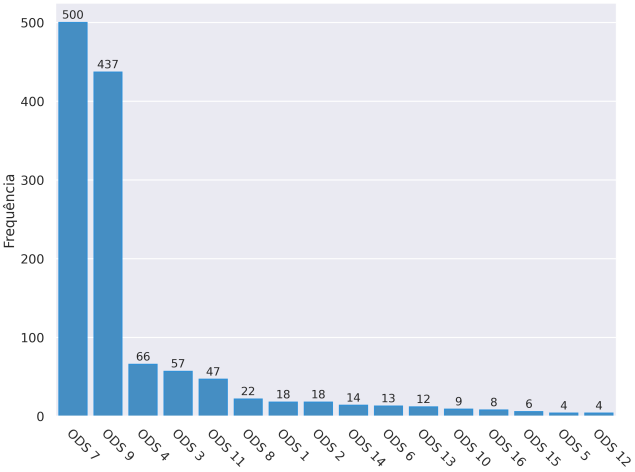

Figura 4. Histograma de recorrências de ODS.

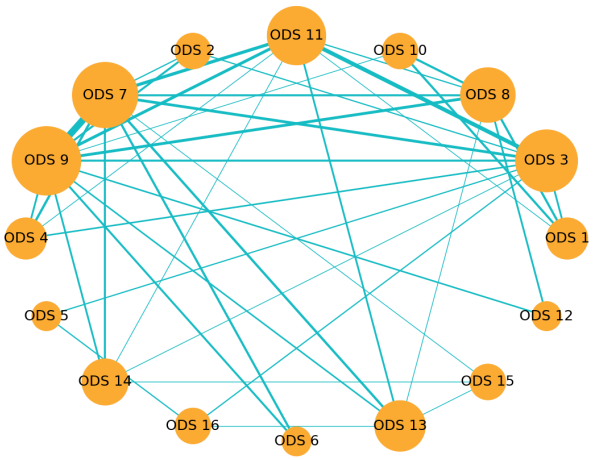

Figura 5. Grafo de relações entre os ODS.

compreender melhor esse fenômeno, algumas análises foram realizadas e chegou-se às seguintes hipóteses: (i) o vocabulário dos 85 títulos indefinidos contém termos genéricos e, por isso, o modelo não foi sensível o suficiente para rotulá-los ou as probabilidades de relação desses títulos com pelo menos um ODS ficaram próximas do limiar de decisão de $0,5$; (ii) existe a possibilidade de haver palavras nesses títulos que não fizeram parte do treinamento do modelo, tornando-o incapaz de classificá-los corretamente.

Tabela 2. Frequência de co-ocorrência dos ODS.

	Quantidade de ODS					
	Zero	**Um**	**Dois**	**Três**	**Quatro**	**Cinco**
Frequência	85	199	451	35	6	1

Por outro lado, percebe-se que mais da metade das publicações versa sobre pelo menos dois ODS. A Tabela 3 apresenta os cinco pares de ODS mais frequentes, sendo o primeiro par composto pelos ODS 7 e 9.

Tabela 3. Frequência de co-ocorrência de pares de ODS.

	Pares de ODS				
	7 & 9	**3 & 11**	**4 & 7**	**4 & 9**	**1 & 3**
Frequência	382	22	6	3	3

1 - Erradicação da Pobreza; 3 - Saúde e Bem Estar
4 - Educação de Qualidade; 7 - Energia Limpa e Acessível
9 - Industria, Inovação e Infraestrutura; 11 - Cidades e Comunidades Sustentáveis

Uma representação visual das co-ocorrências é exibida na Figura 5, que ilustra um grafo relacional. Nele, o tamanho dos nós no grafo representa o grau de relacionamento (centralidade) de um ODS com relação aos demais e a espessura da aresta remete à frequência (força) dessas relações.

6. CONCLUSÕES

A implementação dos Objetivos de Desenvolvimento Sustentável tem se tornado cada vez mais necessária globalmente. Somado a isso, entra em cena o papel fundamental da ciência em identificar lacunas e oportunidades de pesquisa que direcionem os esforços produtivos em prol de um futuro sustentável. Sistemas como o *SciVal* tem contribuído para o mapeamento da pesquisa científica aos ODS, mas ainda por meio de uma abordagem que requer especialistas para sua atualização e validação. A fim de reduzir a necessidade de tal conhecimento especializado e de prover uma ferramenta mais autônoma e inteligente, neste trabalho foi proposta uma abordagem de mapeamento das publicações científicas aos ODS. O modelo foi testado com trabalhos do Congresso Brasileiro de Automática 2020, um evento itinerante pelo país que foca em campos com impacto direto na tecnologia e no desenvolvimento sustentável. O método de mapeamento consistiu na aplicação de técnicas de Processamento de Linguagem Natural sobre o título dos artigos aliadas a um modelo de Rede Neural Recorrente, que foi treinado sobre um conjunto de títulos de publicações internacionais coletadas a partir da plataforma *SciVal*.

Como resultado, o melhor modelo apresentou acurácia de 64,13%, precisão de 76,41%, *recall* de 73,27% e *F1-Score* de 71,84% sobre o conjunto de teste. Na etapa de inferência, os ODS 7 e 9 foram os rótulos mais frequentes, versando sobre "Energia Acessível e Limpa"e "Indústria, Inovação e Infraestrutura". Esse resultado é condizente com o próprio foco do congresso em ciências elétricas, automação e controle. Além disso, observa-se que há publicações relacionadas a todos os 16 ODS treinados, o que reforça o potencial da automação em contribuir de forma polivalente e interdisciplinar para o desenvolvimento sustentável.

Este trabalho é ainda aplicado diretamente na pesquisa no Brasil que, enquanto nação em desenvolvimento, supõe-se que o alcance dos ODS seja uma demanda desafiadora. Quanto às limitações, possíveis falhas na tradução automática para o inglês do conjunto de dados do CBA podem ter ocorrido, já que os títulos são predominantemente em português. A coleta manual dos dados do *SciVal* e o desbalanceamento inicial dos ODS no conjunto de dados também representam ameaças. Por fim, os trabalhos futuros incluem o uso de mais dados dos artigos, como resumo e palavras-chave, para treinar o modelo, bem como investigar o impacto das traduções automáticas nos resultados. Outro trabalho inclui a aplicação de métodos distintos que aprimorem os *embeddings* gerados.

REFERÊNCIAS

Allen, C., Metternicht, G., and Wiedmann, T. (2021). Priorities for science to support national implementation of the sustainable development goals: a review of progress and gaps. *Sustainable Development*, 29(4), 635–652.

Ba, J.L., Kiros, J.R., and Hinton, G.E. (2016). Layer normalization. URL https://arxiv.org/abs/1607.06450.

Biewald, L. (2020). Experiment tracking with weights and biases. URL https://www.wandb.com/. Software available from wandb.com.

Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Chollet, F. et al. (2015). Keras. URL https://github.com/fchollet/keras.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics. URL https://doi.org/10.18653/v1/n19-1423.

Durand, T., Mehrasa, N., and Mori, G. (2019). Learning a deep convnet for multi-label classification with partial labels. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 647–657.

ElAlfy, A., Palaschuk, N., El-Bassiouny, D., Wilson, J., and Weber, O. (2020). Scoping the evolution of corporate social responsibility (csr) research in the sustainable development goals (sdgs) era. *Sustainability*, 12(14). URL https://www.mdpi.com/2071-1050/12/14/5544.

Elsevier (2021a). Scival. URL https://www.elsevier.com/about/partnerships/sdg-research-mapping-initiative. Acesso em: 24/04/2022.

Elsevier (2021b). SDG Research Mapping Initiative. URL https://www.elsevier.com/solutions/scival. Acesso em: 24/04/2022.

Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., and Schmidhuber, J. (2017). Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10), 2222–2232.

Ho, Y. and Wookey, S. (2020). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access*, 8, 4806–4813.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9, 1735–80. doi: 10.1162/neco.1997.9.8.1735.

Hsu, D.F., LaFleur, M.T., and Orazbek, I. (2022). Improving sdg classification precision using combinatorial fusion. *Sensors*, 22(3), 1067.

Huang, Y., Giledereli, B., Köksal, A., Özgür, A., and Ozkirimli, E. (2021). Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8153–8161. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. URL https://aclanthology.org/2021.emnlp-main.643.

Khamis, A., Li, H., Prestes, E., and Haidegger, T. (2019). Ai: A key enabler for sustainable development goals: Part 2 [industry activities]. *IEEE Robotics Automation Magazine*, 26(4), 122–127.

Matsui, T., Suzuki, K., Ando, K., Kitai, Y., Haga, C., Masuhara, N., and Kawakubo, S. (2022). A natural language processing model for supporting sustainable development goals: translating semantics, visualizing nexus, and connecting stakeholders. *Sustainability Science*, 1–17.

Rivest, M., Kashnitsky, Y., Bédard-Vallée, A., Campbell, D., Khayat, P., Labrosse, I., Pinheiro, H., Provençal, S., Roberge, G., and James, C. (2021). Improving the Scopus and Aurora queries to identify research that supports the United Nations Sustainable Development Goals (SDGs) 2021. *Mendeley Data*, 4.

Sachs, J.D., Schmidt-Traub, G., Mazzucato, M., Messner, D., Nakicenovic, N., and Rockström, J. (2019). Six transformations to achieve the sustainable development goals. *Nature sustainability*, 2(9), 805–814.

Sak, H., Senior, A.W., and Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *CoRR*, abs/1402.1128. URL http://arxiv.org/abs/1402.1128.

SBA (2022). Sociedade brasileira de automática. URL https://www.sba.org.br/. Acesso em: 24/04/2022.

Semeniuta, S., Severyn, A., and Barth, E. (2016). Recurrent dropout without memory loss. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1757–1766. The COLING 2016 Organizing Committee, Osaka, Japan. URL https://aclanthology.org/C16-1165.

Smith, T.B., Vacca, R., Mantegazza, L., and Capua, I. (2021). Natural language processing and network analysis provide novel insights on policy and scientific discourse around sustainable development goals. *Scientific reports*, 11(1), 1–10.

Sorower, M.S. (2010). A literature survey on algorithms for multi-label learning.

Storks, S., Gao, Q., and Chai, J.Y. (2019). Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. URL https://arxiv.org/abs/1904.01172.

Walsh, P.P., Murphy, E., and Horan, D. (2020). The role of science, technology and innovation in the un 2030 agenda. *Technological Forecasting and Social Change*, 154, 119957.

Yu, H., An, J., Yoon, J., Kim, H., and Ko, Y. (2020). Simple methods to overcome the limitations of general word representations in natural language processing tasks. *Computer Speech & Language*, 59, 91–113.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *CoRR*, abs/1409.2329. URL http://arxiv.org/abs/1409.2329.