Integração e Rotulação Automatizada de Dados sobre o Cnidário Physalia physalis, usando a Geolocalização como Referência - Pesquisa de Mestrado

Lisiane Reips[1], **Carmem Satie Hara**[1]

[1]Programa de Pós-Graduação em Informática
Universidade Federal do Paraná (UFPR)
Curitiba – PR – Brasil

lisiane.reips@ufpr.br, carmemhara@gmail.com

Data de Ingresso no Mestrado: abril de 2021
Data do exame de qualificação: outubro de 2022
Data de defesa do Mestrado: março de 2023

Abstract. *Classification techniques in machine learning models have been effectively applied to text and image recognition. But for any and every application, data need to be trained and tested. In order to achieve good performance in the classification process, these data need to be reliably labeled, which makes the process expensive and time-consuming. In this paper, we propose an approach to reduce the cost of manual labeling a database composed of Portuguese man of war (Physalia physalis) sightings on Brazilian beaches. The technique is based on integrating Instagram posts with newspaper articles based on their temporal and spatial proximity. The ultimate goal is to use these labeled data for training a classification technique in the machine learning process.*

1. Introdução

Caravelas-portuguesas são organismos vivos do grupo dos cnidários, que vivem em colônias nos oceanos. Seu nome científico é Physalia physalis, mas são mais conhecidas por caravelas-portuguesas devido ao seu formato parecer com as caravelas usadas pelos portugueses, entre os séculos XV e XVI. Estes organismos vivos possuem uma beleza natural e exótica e, diferentemente do que muitos pensam, eles não nadam, mas flutuam até as praias, levados pelo vento[1].

Instigando e atraindo a atenção das pessoas que por ali circulam, as caravelas-portuguesas são organismos que não atacam, mas se sentem ameaçadas ao entrar em contato com o ser humano. Sentindo-se ameaçados, esses organismos liberam uma toxina paralisante que provoca acidentes com queimaduras de até terceiro grau nos seres humanos. Nem todas as pessoas conhecem o cnidário e acabam tocando no animal, sofrendo lesões com possibilidades de sequelas, podendo até evoluir para a morte [Bochner and Struchiner 2002]. Mas acidentes assim podem ser minimizados, caso haja um melhor monitoramento dessas espécies e notificações sobre a sua existência nas praias do Brasil. O monitoramento de espécies também é útil para a manutenção e controle da biodiversidade.

Há projetos que fazem esse trabalho, disseminando a informação, e proporcionando acesso a bases de dados contendo notificações de espécies em todo o mundo. Uma destas iniciativas é a plataforma iNaturalist[2], que disponibiliza na Web informações sobre o monitoramento de espécies marinhas. Outro exemplo de rede social que pode ser útil ao acesso a informações é o Instagram, que a cada ano cresce nas postagens de imagens. Essas redes sociais podem ser grandes fontes de dados, contribuindo em notificações e disseminação do conhecimento. Por ser um animal exótico e com lindas cores, a caravela-portuguesa atrai registros fotográficos e prováveis postagens no Instagram. Essas postagens podem ser ricas fontes de dados de avistamentos do cnidário. Para auxiliar na classificação dos dados coletados do Instagram, pode ser utilizado o aprendizado de máquina. Os métodos de aprendizado de máquina comumente têm a capacidade de realizar tarefas de previsão (quando a variável de resultado é um valor) ou de classificação (quando a variável de resultado é uma classe) [Abhari et al. 2019] e o método de interesse deste trabalho é a classificação.

Mas há uma questão que precisa ser tratada para utilizar a rede social Instagram como fonte de dados: a detecção de verdadeiros avistamentos do animal caravela-portuguesa. Postagens do Instagram geralmente contêm hashtags e alguma informação, mas não garantem que esta informação seja a realmente procurada. Por exemplo: pode-se localizar uma hashtag seguida da expressão caravela-portuguesa e esta se referir à caravela que os portugueses usavam nos séculos passados. Denomina-se como rótulo, a informação que acompanha a imagem e, neste caso, mesmo contendo uma rotulação de acordo com a expressão buscada, a imagem não está de acordo para a pesquisa. Para uma constatação confiável, a verificação de um ser humano que saiba diferenciar as imagens é necessária. Contudo, rotular milhares de postagens é um processo custoso e demo-

[1]https://www.natgeo.pt/animais/2018/06/caravela-portuguesa-saiba-porque-e-tao-temida

[2]A plataforma iNaturalist é uma rede social e projeto científico que conecta naturalistas, biólogos e cientistas com a finalidade de compartilhar informações e observações sobre a biodiversidade em todo o mundo. Link para acesso: https://www.inaturalist.org/

rado. Para um bom desempenho do modelo de classificação em aprendizado de máquina é necessário uma base de dados com rótulos fortes e confiáveis. O objetivo geral deste trabalho é desenvolver uma abordagem motivada pela diminuição de custos na rotulagem manual de dados. Em particular, o objetivo é diminuir o custo da rotulação manual de uma base de dados contendo avistamentos de caravelas-portuguesas nas praias brasileiras. Estes dados, posteriormente, serão utilizados em uma técnica de classificação no processo de aprendizado de máquina.

Para atingir o objetivo geral, serão utilizados mais dados, provenientes de diferentes fontes. As diferentes fontes escolhidas envolvem a plataforma iNaturalist e diversos sites de notícias que informam sobre avistamentos de caravelas-portuguesas nas praias brasileiras, além da rede social Instagram. O trabalho contará com bases contendo dados dos anos de 2012 a 2022. Após as bases de dados de diferentes fontes serem coletadas, será necessário resolver o problema do cruzamento de dados de diferentes fontes. A ideia é identificar se um avistamento que encontra-se em uma base de dados corresponde a uma mesma postagem no Instagram, levando em consideração a proximidade geográfica e temporal. Considerando que as postagens do iNaturalist são curadas e que os jornais apresentam informações verdadeiras, as postagens do Instagram que coincidem no espaço e no tempo podem ser rotuladas como verdadeiras. As postagens rotuladas podem então ser utilizadas para o treinamento de um classificador, que pode posteriormente ser utilizado para rotular as demais postagens.

A rotulação manual é considerada forte e é realizada por uma pessoa especialista no assunto. Em aprendizado de máquina, a rotulação automática é denominada como supervisão fraca, como proposto neste trabalho. Mas além do problema de integração de dados, há a preocupação com dois problemas pontuais em Banco de Dados: a deduplicação dos dados, buscando técnicas para eliminar cópias redundantes de dados, reduzindo a sobrecarga do armazenamento; e a manutenção da integridade, fazendo uso de métodos para garantir a confiabilidade dos dados.

A solução proposta visa fazer uso de uma rede social, analisando sua utilidade como fonte de informação confiável. Também visa utilizar de técnicas existentes de integração de dados e de supervisionamento fraco. A relevância da pesquisa está na: aplicação de uma abordagem para reduzir o custo e o tempo de rotulação de dados de uma base contendo avistamentos de caravelas-portuguesas; criação de uma base de dados que poderá ser utilizada para outros modelos de classificação de caravelas-portuguesas; demonstração da eficácia na obtenção de informações de uma rede social, podendo estas informações serem úteis para notificar a população.

2. Trabalhos Relacionados

A revisão da literatura acerca das abordagens existentes foi dividida em três áreas: integração de dados de fontes diversas; uso de redes sociais como fonte de informações úteis e confiáveis à população; e técnicas de supervisão para rotulação de dados.

2.1. Integração de dados de diferentes fontes

Os autores [Kulkarni and Di Minin 2021] construíram um pipeline, com destaque para métodos de coleta e análise de dados de forma automatizada. Entre os métodos, os autores trabalham na deduplicação de dados e na integração de dados semelhantes. Para isso,

foram coletados 15.088 artigos, de 585 espécies listadas na Convenção sobre Comércio Internacional de Espécies Ameaçadas de Fauna e Flora Selvagens (CITES), no período de um mês, das plataformas digitais Google Notícias e Twitter. Para extrair informações como a localização dos dados, um modelo de Reconhecimento de Entidade Nomeada (processo de localizar e classificar palavras ou frases em um texto simples, em categorias como ”Pessoas”e ”Lugares”, por exemplo) foi utilizado. Muitos dados podem vir com conteúdos semelhantes ou idênticos. Para isso, um método de vetorização de texto foi aplicado, deduplicando dados, quando idênticos ou integrando-os, quando muito semelhantes e ambos relevantes. Os autores concluíram que o pipeline facilita a extração de dados digitais para análises posteriores, reduzindo a carga de trabalho manual de pesquisadores na coleta de grande quantidade de dados. O diferencial desta dissertação está no uso da geolocalização como referência para a integração dos dados.

2.2. Redes sociais como fonte de informações úteis e confiáveis

Dados para treinamento em aprendizado de máquina são cada vez mais consumidos e úteis nas técnicas de classificação e regressão linear. Porém, conseguir esses dados em grande quantidade é um desafio. Para suprir essa necessidade, diversas fontes da Web tornaram-se essenciais, incluindo as mídias sociais. Um exemplo de uso de redes sociais encontra-se no trabalho de [Daume 2016], no qual os autores desenvolveram uma avaliação da rede social Twitter, observando os pontos fortes e fracos na mineração dos dados para monitoramento ecológico e de espécies exóticas invasoras.

Os autores conseguiram comprovar que a rede social Twitter é uma rica fonte de geração e de observações sobre a biodiversidade, demonstrando percepções das comunicações públicas sobre as espécies. Mas essa comprovação contou com uma avaliação manual de 2.842 Tweets. A diferença para o presente trabalho é que serão usadas uma rede social e uma plataforma social (Instagram e iNaturalist) para verificar a relevância das postagens como fonte de informações úteis. Também será trabalhada a automatização de todo o processo.

2.3. Técnicas de supervisão para rotulação de dados

Um dos gargalos mais significativos no processo de aprendizado de máquina é a rotulação de dados [Bach et al. 2019]. Mas rótulos de alta qualidade são caros e muitas vezes indisponíveis [Tang et al. 2021]. Então, para diminuir o custo da rotulação manual, o supervisionamento fraco se torna útil. Um sistema que faz uso de supervisionamento fraco é o Snorkel [Ratner et al. 2017]. Com o sistema, os usuários escrevem funções de rotulagem, podendo ter precisões e correlações desconhecidas, combinando fontes de supervisão fraca para criar dados de treinamento. Para fins de comprovação na eficácia, o Snorkel foi comparado a outras abordagens. Especialistas no assunto constataram que o Snorkel consegue criar modelos 2,8 vezes mais rápido, com aumento do desempenho preditivo, em comparação com sete horas de rotulagem manual. O Snorkel colaborou com implantações ao Departamento de Assuntos de Veteranos dos EUA e o Hospital e Clínicas de Stanford, e a Food and Drug Administration dos EUA.

Além do Snorkel, o sistema Snuba [Varma and Ré 2018] gera heurísticas automaticamente, fazendo uso de um pequeno conjunto de dados rotulados para atribuir rótulos de treinamento a um conjunto grande de dados não rotulados, usando supervisão fraca.

O Snuba foi comparado ao aprendizado semi-supervisionado, ao aprendizado de transferência e ao desempenho de suas heurísticas com as heurísticas desenvolvidas pelos usuários, ficando mais próximo do desempenho de aprendizado semi-supervisionado. Em colaboração com radiologistas do Stanford Hospital and Clinics, nos Estados Unidos, tarefas de classificação de tumores ósseos e mamografias demonstraram que a heurística gerada pelo Snuba é comparável às desenvolvidas por especialistas do domínio. O presente trabalho visa aplicar estas ou técnicas similares para gerar um conjunto de dados de avistamentos de caravelas-portuguesas.

3. Metodologia

Nesta seção será apresentada a metodologia a ser utilizada e que é ilustrada na Figura 1.

3.1. Bases de Dados a serem Utilizadas

As bases de dados a serem utilizadas serão coletadas das seguintes fontes: plataforma iNaturalist, sites de notícias online e rede social Instagram. A plataforma iNaturalist é confiável para pesquisas sobre qualquer espécie, pois é utilizada por naturalistas, biólogos e cientistas para troca de informações. Já foi coletada, para este trabalho, uma base contendo 7.714 dados sobre o cnidário caravela-portuguesa. Também foi realizada uma busca inicial de notícias sobre caravelas-portuguesas, na qual 70 sites de notícias foram encontrados. Através de uma raspagem de dados, todos os sites serão analisados e as notícias que contêm a expressão ”caravela-potuguesa” serão coletadas. Uma análise será feita para identificar quais realmente se referem ao cnidário. A terceira coleta é na rede social Instagram. Através das expressões #caravelaportuguesa, #cnidario e #physaliaphysalis, já foram coletadas 2.226, 650 e 190 postagens, respectivamente.

3.2. Processo de Integração de Dados

Primeiramente será tomada como referência a geolocalização do dado de várias fontes. A combinação entre os dados terá como parâmetro um intervalo de tempo de 7 dias, anterior e posterior à data da postagem. A escolha de 7 dias como parâmetro se deu devido à avaliação preliminar realizada detectar que as datas de notícias que coincidiam com as datas de postagens do Instagram tinham uma diferença de 2 a 3 dias. Então para haver uma margem de segurança, foi realizada tal escolha. De forma automatizada, os dados serão combinados e agrupados quando estes estiverem em fontes diferentes, mas possuírem o mesmo intervalo de tempo e a mesma localização. Para esta automatização, pretende-se utilizar técnicas de Reconhecimento de Entidades Nomeadas e de vetorização de textos, como descritas em [Kulkarni and Di Minin 2021].

3.3. Processo de Rotulação de Dados

Após os dados combinados, haverá uma base de dados integrados, originada das fontes iNaturalist, sites de notícias e rede social Instagram. Os demais dados das postagens do Instagram, que não puderam ser integrados, por não haver a geolocalização ou rotulação correta, serão rotulados de forma automatizada. Para esta automatização, existem técnicas como Snorkel [Ratner et al. 2017] e Snuba [Varma and Ré 2018], que serão avaliadas para o uso neste caso, com a ideia de realizar um estudo comparativo entre elas, a fim de identificar qual seria mais eficiente para este problema.

A Figura 1 apresenta a abordagem proposta para o processo de integração levando em consideração a geolocalização, e a rotulação dos dados de forma automatizada.

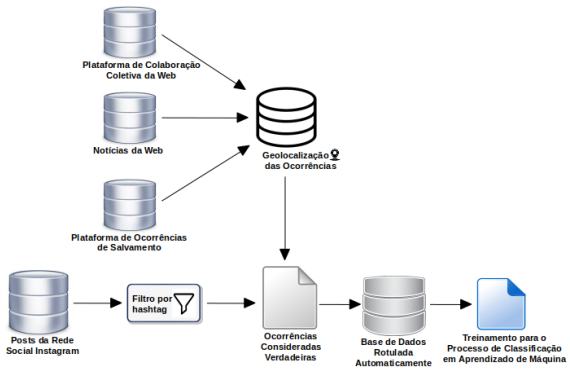

Figura 1. Ilustração da abordagem para diminuir o custo na rotulação manual de dados.

4. Avaliação da Proposta

Para avaliar a proposta deste trabalho, uma base de dados rotulada manualmente por especialistas será utilizada. Uma comparação na eficácia do uso dos dados rotulados manualmente, com a eficácia dos resultados adquiridos no uso dos dados sem rotulação manual (supervisionamento fraco) será realizada. Essa comparação utilizará a precisão, como ideia inicial de métrica. Será possível também analisar o percentual das postagens que se consegue rotular usando o supervisionamento fraco.

5. Resultados Esperados

Como resultados esperados, busca-se: verificar o quanto se obtém de informações consideradas verdadeiras, em redes sociais, mas que não se encontram em outras fontes de dados, o que demonstrará o quão útil a rede social pode ser no aspecto de informar à população com veracidade; comparar e avaliar a quantidade de rótulos considerados positivos verdadeiros, tanto no uso dos dados rotulados manualmente, quanto no uso dos dados não rotulados manualmente por especialistas; verificar qual será o impacto no modelo treinado, quando este for feito com o uso dos dados rotulados manualmente, comparando com o impacto no modelo, quando treinado com os dados não rotulados manualmente (supervisionamento fraco).

6. Avaliação Preliminar dos Resultados Obtidos

Em fase inicial do trabalho, uma raspagem de dados em sites de notícias que contêm dados sobre caravelas-portuguesas foi realizada. Através de um script desenvolvido na linguagem de programação Python e após a limpeza dos dados, foram coletadas 43 notícias e 96 postagens do Instagram sobre o animal. A coleta foi realizada somente com sites de notícias e postagens do Instagram da região Sul do Brasil, no período de 2019 a 2021. Mesmo com poucas fontes usadas por enquanto, percebe-se que a rede social retorna mais avistamentos sobre caravelas-portuguesas quando comparada aos sites de notícias.

Para uma breve identificação de notícias e postagens referentes ao mesmo período de tempo e localização, uma análise manual foi realizada. Apenas duas notícias com duas postagens do Instagram puderam ser cruzadas para integração. Além disso, quatro sites diferentes retornaram as mesmas notícias, no mesmo intervalo de tempo e, em dois casos, pela mesma colunista e com o texto exatamente igual. Percebe-se que é baixa a quantidade de dados integrados que se consegue realizar neste caso, mas ainda há variáveis

a serem trabalhadas, como o uso de outras hashtags na busca pelo Instagram, o uso dos outros dados de todas as fontes desejadas e o uso de um intervalo maior de tempo de análise. Essa técnica de coleta de dados e integração às postagens no Instagram, usando como referência o espaço-temporal, pode ser usada em outros contextos como em casos de monitoramento de outras espécies aquáticas ou em casos de poluição em praias, por exemplo.

7. Próximos Passos

Os próximos passos estão descritos na Tabela 1.

Atividades	2022						2023		
	Jul	Ago	Set	Out	Nov	Dez	Jan	Fev	Mar
Realizar o Exame de Qualificação				X					
Finalizar as coletas de bases de dados		X	X						
Aplicar as técnicas de integração nas bases coletadas			X	X					
Realizar os ajustes necessários para a padronização da base de dados				X	X				
Aplicar as técnicas de rotulação automática nos dados					X	X			
Utilizar a base rotulada, de forma automática, em um modelo de classificação de aprendizado de máquina e comparar com o desempenho de uma base rotulada manualmente						X	X		
Redigir e submeter artigo científico					X	X	X	X	
Elaborar a redação da dissertação de Mestrado					X	X	X	X	
Defender a dissertação de Mestrado									X

Tabela 1. Cronograma das atividades a serem desenvolvidas.

Referências

Abhari, S., Rostam Niakan Kalhori, S., Ebrahimi, M., Hasannejadasl, H., and Garavand, A. (2019). Artificial intelligence applications in type 2 diabetes mellitus care: Focus on machine learning methods. *Healthcare Informatics Research*, 25:248–261.

Bach, S. H., Rodriguez, D., Liu, Y., Luo, C., Shao, H., Xia, C., Sen, S., Ratner, A., Hancock, B., Alborzi, H., et al. (2019). Snorkel drybell: A case study in deploying weak supervision at industrial scale. In *Proceedings of the 2019 International Conference on Management of Data*, pages 362–375.

Bochner, R. and Struchiner, C. J. (2002). Acidentes por animais peçonhentos e sistemas nacionais de informação. *Cadernos de Saúde Pública*, 18:735–746.

Daume, S. (2016). Mining twitter to monitor invasive alien species—an analytical framework and sample information topologies. *Ecological Informatics*, 31:70–82.

Kulkarni, R. and Di Minin, E. (2021). Automated retrieval of information on threatened species from online sources using machine learning. *Methods in Ecology and Evolution*, 12(7):1226–1239.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, 11(3):269.

Tang, C., Yuan, G., and Zheng, T. (2021). Weakly supervised learning creates a fusion of modeling cultures. *Observational Studies*, 7(1):203–211.

Varma, P. and Ré, C. (2018). Snuba: Automating weak supervision to label training data. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, 12(3):223.