Identificação de Ocorrências do Cnidário *Physalia physalis* em Dados Extraídos de Mídias Sociais

Heloisa F. Rocha[1], **Carmem S. Hara**[1]

[1]Departamento de Ciência da Computação
Universidade Federal do Paraná, Curitiba – PR – Brasil

heloisarocha@ufpr.br, carmem@inf.ufpr.br

Resumo. *As necessidades de conhecimento da biodiversidade são constantes, enquanto recursos para pesquisa, sejam financeiros, de tempo e humanos são escassos. Por outro lado, a Internet oferece um enorme volume de dados que podem ser explorados em favor da ciência da conservação. As caravelas-portuguesas (Physalia physalis) oferecem risco à população, e dados sobre sua ocorrência nem sempre estão disponíveis para estudo da espécie. Neste trabalho é proposto o treinamento de modelos de aprendizagem de máquina como ferramenta para classificar dados extraídos de uma mídia social e assim possibilitar a geração de uma base de dados sobre ocorrências de caravelas-portuguesas no litoral brasileiro.*

Nível: Mestrado

Ingresso: Abril de 2021

Previsão de término: Abril de 2023

Programa: Programa de Pós-Graduação em Informática (PPGInf)
Universidade Federal do Paraná

Etapas concluídas: disciplinas concluídas, revisão da literatura, definição do problema

Defesa da qualificação: Outubro de 2022 (estimada)

1. Introdução

A caravela-portuguesa (*Physalia physalis*) é um organismo pluricelular, cujos tentáculos possuem células urticantes que podem liberar toxinas nocivas. Essa espécie ocorre por toda a costa brasileira e acidentes com esses animais têm sido reportados com frequência [Cavalcante et al. 2020].

Apesar de ser de notificação compulsória, os dados do SINAN (Sistema de Informação de Agravos de Notificação) não apresentam indicação específica para envenenamento por caravela-portuguesa. Além disso, há evidências de subnotificação de casos [Cavalcante et al. 2020]. Ou seja, a informação limitada acerca desses animais dificulta o uso dos dados públicos oficiais no trabalho de pesquisadores interessados na espécie.

Uma alternativa para conseguir mais dados é a coleta realizada por profissionais. Entretanto, métodos tradicionais de coleta consomem muito tempo e recursos, além de muitas vezes terem baixa cobertura [Edwards et al. 2021].

Outra possibilidade seria a utilização de campanhas de ciência cidadã. Entretanto, organizar esse tipo de campanha requer esforço para recrutar, treinar e incentivar voluntários [Edwards et al. 2021].

Por fim, uma alternativa é a utilização de dados involuntários de mídias sociais. Chamada de ciência cidadã passiva, ela consiste no uso de dados gerados por não profissionais, coletados e compartilhados na Internet, principalmente em mídias sociais e, que não estão conectados a nenhum programa específico de ciência cidadã [Edwards et al. 2021].

Entre as vantagens resultantes da utilização de dados de mídias sociais, estão: conteúdo abundante e quase sempre público [Edwards et al. 2021]; abordagem menos trabalhosa, menos demorada e com custo menor, especialmente se for automatizada, em comparação a métodos tradicionais ou mesmo a campanhas de ciência cidadã [Edwards et al. 2021]; dados disponíveis quase em tempo real e de maneira contínua [Edwards et al. 2021]; e aquisição de dados com amplo escopo geográfico [Morais et al. 2021]. Além disso, o uso de ferramentas de tecnologia da informação, como a Aprendizagem de Máquina (AM), permite aproveitar o potencial dos dados de mídias sociais e transformá-los em informações úteis.

No projeto de [Nascimento 2020], aprovado pelo comitê de ética da UFPR, um dos seus objetivos específicos é a compilação de dados sobre ocorrências de *Physalia physalis* no litoral brasileiro a partir de dados extraídos de mídias sociais. No caso do Instagram, é permitida a busca de conteúdo por meio de hashtags. Entretanto, faz-se necessária a avaliação desse conteúdo a fim de verificar se trata-se ou não de uma ocorrência legítima de *Physalia physalis*. Tal verificação tem sido realizada manualmente.

Um trabalho anterior relacionado ao projeto [Carneiro et al. 2022] já avançou no sentido de classificar as imagens das postagens como sendo ou não de *Physalia physalis*. Em seus experimentos, foi obtida uma acurácia de 0,94. Mas ainda não é o ideal. Análises preliminares revelaram a existência de postagens que mencionam nomes comuns e imagens da espécie, mas não são observações diretas de vida selvagem. Por exemplo, algumas postagens alertam sobre o risco que a *Physalia physalis* oferece aos banhistas, o que não representa uma ocorrência de acidente ou uma observação direta da espécie. Além disso, a repostagem de conteúdo, seja do próprio usuário ou copiada de outro perfil, sem

a análise do texto é de difícil diferenciação. Notou-se também que em casos de acidentes, as imagens postadas ora são de partes do corpo atingidas, ora são imagens da própria *Physalia physalis*, fazendo-se necessária a análise do texto e imagem para diferenciar um acidente de outras situações.

Essas características indicam que para obter o refinamento desejado entre as várias situações encontradas (i.e, avistamento, acidente, alerta e repostagem) é necessária a análise conjunta do texto e imagem da postagem, ou seja, um caso para aplicação de AM multimodal. Além disso, essa variação de situações é um indicativo para utilização de classificação multiclasse.

Dessa forma, a principal motivação para essa pesquisa se dá pelas vantagens apresentadas para o uso de dados de redes sociais com apoio de técnicas de AM. Tais técnicas permitirão o treinamento de um modelo que classifique o conteúdo das postagens, identificando assim ocorrências legítimas de *Physalia physalis*. Tal classificador pode ser inserido dentro de um sistema de busca, classificação e armazenamento de dados, permitindo desta forma o monitoramento contínuo da espécie.

2. Trabalhos Relacionados

Diversos trabalhos ressaltam as vantagens e possibilidades para a utilização de dados involuntários para tarefas relacionadas à ciência da conservação [Edwards et al. 2021, Morais et al. 2021]. Também há um número de estudos que aplicam a AM para detectar aspectos do meio-ambiente, tais como [Foglio 2019]. Pode-se também encontrar estudos que utilizaram dados involuntários para a detecção de vida selvagem, porém sem o uso de ferramentas automatizadas [Taklis et al. 2020].

Entretanto foram poucos os trabalhos encontrados que aplicaram a AM em dados involuntários para classificação de vida selvagem, tais como: [Mazars-Simon 2019, Kulkarni and Di Minin 2021, Edwards et al. 2022]. Além disso, em seus trabalhos [Kulkarni and Di Minin 2021, Edwards et al. 2022] chamam a atenção para a escassez de trabalhos usando texto e a inexistência de trabalhos usando AM multimodal em tarefas relacionadas à ciência da conservação. Tal escassez também foi notada durante as buscas por trabalhos relacionados a esta proposta. Há também poucos trabalhos fazendo comparações entre diferentes arquiteturas como de [Edwards et al. 2022].

Também foi realizada uma busca por trabalhos que treinaram modelos utilizando texto e imagem para tarefas de classificação. Na maioria das tarefas testadas por [Giachanou et al. 2020, Ofli et al. 2020, Hu et al. 2021] os modelos multimodais tiveram desempenho superior aos dos modelos unimodais. Os resultados obtidos pelos autores reforçam a ideia que o uso combinado de imagem e texto pode melhorar o desempenho de tarefas de classificação.

Como principal diferencial ante aos trabalhos que propõem identificar vida selvagem em dados de mídia social, neste trabalho é proposto realizar comparações entre diferentes abordagens de treinamento, i.e., modelos clássicos, redes neurais (RN) e aprendizagem por transferência (AT), e a sua adequação para pequenos datasets extraídos de mídias sociais. São planejados experimentos com classificação binária e multiclasse. Além disso, planeja-se realizar uma análise comparativa dos resultados obtidos com o treinamento utilizando apenas os dados textuais e utilizando multimodalidades (texto e imagens das postagens).

3. Metodologia

Nesta seção é apresentada uma visão geral de como os dados que serão utilizados no desenvolvimento da dissertação estão sendo coletados e rotulados (Subseção 3.1). Enquanto nas Subseções 3.2, 3.3 e 3.4 são listadas as atividades propostas para a continuidade deste trabalho.

3.1. Coleta e Rotulação

Para o treinamento e validação dos modelos de AM serão utilizados dados extraídos do Instagram filtrados por hashtags com nomes científicos e populares da caravela-portuguesa, tais como: #caravelaportuguesa, #physaliaphysalis e #aguaviva. O procedimento de coleta de dados foi executado utilizando-se estratégias de *scraping*. Foram baixadas 7300 postagens datadas entre 2004 a 2021.

Os dados estão sendo rotulados pela autora e por uma especialista, e a eles estão sendo atribuídos dois rótulos: a) considerando apenas o texto da postagem; b) considerando texto e mídias da postagem. Os rótulos atribuídos são: AVISTAMENTO, quando a postagem é um avistamento de *Physalia physalis*; ACIDENTE, quando ela se refere a um acidente com *Physalia physalis*; ALERTA, quando é um alerta ou uma postagem educacional sobre *Physalia physalis* e não corresponde a um avistamento ou acidente; REPOST, quando é um avistamento ou acidente com *Physalia physalis*, porém é uma repostagem de uma ocorrência anterior, feita pelo próprio usuário ou copiada de outro perfil; NADA, quando não é sobre *Physalia physalis*; e DÚVIDA, para os casos cujos dados não são suficientes para atribuir-se um rótulo.

Além da classificação multiclasse, os modelos desta proposta serão treinados para realizar classificação binária, na qual os rótulos: avistamento, acidente, alerta e repost serão considerados como positivo para *Physalia physalis* e nada considerado como negativo para *Physalia physalis*.

Entre as características dos dados que podem representar um desafio para treinamento dos modelos está a variação linguística apresentada na escrita das postagens. Embora algumas postagens apresentem o uso correto das normas gramaticais, o texto de mídia social é informal por natureza, com muitos erros gramaticais, gírias, abreviações, neologismos e mistura de idiomas. Além disso, existem postagens que são formadas quase exclusivamente por hashtags e emojis.

Na sequência, são detalhadas as atividades planejadas para o desenvolvimento desta dissertação.

3.2. Atividade 1: Análise do Texto

Treinamento de modelos de AM unimodal para identificar, por meio apenas do texto, ocorrências de *Physalia physalis*. A Tabela 1 apresenta as abordagens que serão utilizadas para o treinamento dos modelos.

3.3. Atividade 2: Análise Multimodal

Treinamento de modelos de AM multimodal para identificar, por meio do texto e imagem, ocorrências de *Physalia physalis*. Para cumprir esta atividade o modelo obtido por [Carneiro et al. 2022] será utilizado como extrator de características das imagens. A Tabela 2 apresenta as abordagens que serão utilizadas para treinamento dos modelos.

Tabela 1. Abordagens para Treinamento dos Modelos com Texto

Abordagem	Descrição
Clássica	Selecionar um método de representação e modelo clássico de AM para ser utilizado como baseline. Exemplos de métodos de representação: TF-IDF, BERTimbau [Souza et al. 2020] e *Word Embeddings* (WE). Exemplos de WE: Word2Vec, FastText, Wang2Vec e Glove. Exemplos de modelos clássicos de AM: Regressão Logística (RL), Árvores de Decisão e Naive Bayes.
RN	Selecionar e treinar uma RN (e.g., Multi-layer Perceptron (MLP), Redes Neurais Convolucionais (CNN) e *Long Short-Term Memory*).
AT	Selecionar modelos de linguagem pré-treinados em Português (e.g., BERT [Devlin et al. 2019] e BERTimbau) e ajustar para tarefa de classificação de texto.

Tabela 2. Abordagens para Treinamento dos Modelos Multimodais

Abordagem	Descrição
Clássica	Fusão dos vetores de características de texto e imagem e, treinamento de algoritmo clássico de AM. Por exemplo: extrair características usando BERT e CNN e usar como entrada para uma RL.
RN	Fusão dos vetores de características de texto e imagem e, treinamento de RN. Por exemplo: extrair características usando BERT e CNN e usar como entrada para um MLP.
AT	Selecionar modelos pré-treinados em texto e imagem (e.g., VilBERT [Lu et al. 2019]) e ajustar para tarefa de classificação

3.4. Atividade 3: Análise Comparativa

Nesta atividade os modelos treinados nas Atividades 1 e 2 serão avaliados e comparados a fim de responder às seguintes perguntas:

- Um modelo treinado com texto e imagem (multimodal) é melhor para reconhecer postagens legítimas sobre *Physalia physalis* do que modelos treinados apenas com texto? Para responder a esta pergunta os modelos de AM serão treinados com bases rotuladas considerando apenas o texto e com bases rotuladas considerando texto e imagem. Apoiados nos resultados obtidos pelos autores [Giachanou et al. 2020, Ofli et al. 2020, Hu et al. 2021] é esperado que os modelos multimodais tenham melhor desempenho.
- Qual abordagem apresentou melhor desempenho para classificar as postagens? Para responder a esta pergunta, neste trabalho é proposta uma variedade de abordagens de treinamento que incluem modelos clássicos, RN e modelos pré-treinados. Em projetos utilizando visão computacional, a AT é largamente utilizada, principalmente para pequenos datasets. No projeto de [Carneiro et al. 2022] essa foi a abordagem com melhor desempenho. Embora a literatura mostre sucesso na utilização de AT em tarefas de processamento de linguagem natural [Souza et al. 2020], análises preliminares do texto (seção 4) mostraram que modelos clássicos permanecem competitivos.

Para avaliação dos modelos serão utilizadas as seguintes métricas: acurácia, precisão, revocação e F1.

4. Análises Preliminares

Nesta proposta foram realizadas análises preliminares utilizando-se uma amostra balanceada com 304 postagens rotuladas como positivo ou negativo para *Physalia physalis*.

Foram treinados o BERTimbau, o qual foi refinado para tarefa de classificação, e modelos de RL, os quais receberam como entrada vetores de características gerados por meio de TF-IDF e WE. Para os experimentos com WE foram escolhidos os modelos: Word2Vec, FastText, Wang2Vec e Glove, pré-treinados em Português do Brasil por [Hartmann et al. 2017], com 100 dimensões. Para treinamento dos modelos foi utilizada validação cruzada com 5 subconjuntos de dados. A RL foi treinada utilizando-se hiperparâmetros padrão do modelo, enquanto o BERTimbau foi treinado por 30 épocas com uma taxa de aprendizagem de 2e-5. Além disso, os dados de treinamento e validação não receberam nenhum tipo de pré-processamento, além da tokenização.

Como resultado, o modelo treinado com TF-IDF obteve o melhor desempenho com F1 de 0,91, seguido do BERTimbau com F1 de 0,90. Quanto aos modelos treinados com WE, o melhor resultado foi de F1 de 0,82 obtido pelo modelo treinado com FastText. Embora os resultados mostram-se interessantes, ainda há espaço para a exploração dos hiperparâmetros dos modelos e de técnicas de pré-processamento do texto.

5. Conclusão

As mídias sociais representam uma fonte importante para obtenção de dados sobre ocorrências de *Physalia physalis*. A automatização do processo de curadoria desses dados passa pela correta atribuição de rótulos.

Dessa forma, é proposto o treinamento de modelos de AM capazes de identificar ocorrências legítimas de *Physalia physalis*. Como principal contribuição deste trabalho é esperado obter um modelo que possa ser utilizado como parte de um processo automatizado de ETL (*Extract-Transform-Load*) de uma base de dados sobre ocorrências de *Physalia physalis* no litoral brasileiro a partir dados extraídos de redes sociais.

Espera-se também que a abordagem desenvolvida neste trabalho possa ser posteriormente aplicada para a automatização da identificação de outras espécies.

Uma outra contribuição é uma comparação entre o uso e desempenho de diferentes modais como recurso para treinamento de modelos de AM, no contexto de ciência da conservação e dados de mídias sociais.

A Tabela 3 mostra as atividades já realizadas para alcançar os objetivos e contribuições citadas nesta proposta, além dos próximos passos a serem executados.

Tabela 3. Cronograma de atividades

Atividades	2022			2023	
	1-9	10	11-12	1-2	3-4
Definição e contextualização do problema de pesquisa	X				
Coleta e rotulação dos dados	X	X			
Pesquisa bibliográfica	X				
Busca por trabalhos relacionados	X				
Exame de Qualificação		X			
Atividade 1 Análise do Texto			X	X	
Atividade 2 Análise Multimodal			X	X	
Atividade 3 Análise Comparativa					X
Escrita da dissertação	X	X	X	X	X
Defesa da dissertação					X

Referências

Carneiro, A., Nascimento, L., Noernberg, M., Hara, C., and Pozo, A. (2022). Portuguese man-of-war image classification with convolutional neural networks. *arXiv preprint arXiv:2207.01171*.

Cavalcante, M. M. E., Rodrigues, Z. M. R., Hauser-Davis, R. A., Siciliano, S., Haddad Júnior, V., and Nunes, J. L. S. (2020). Health-risk assessment of portuguese man-of-war (physalia physalis) envenomations on urban beaches in são luís city, in the state of maranhão, brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, 53.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Edwards, T., Jones, C. B., and Corcoran, P. (2022). Identifying wildlife observations on twitter. *Ecological Informatics*, 67:101500.

Edwards, T., Jones, C. B., Perkins, S. E., and Corcoran, P. (2021). Passive citizen science: The role of social media in wildlife observations. *PLOS ONE*, 16(8):e0255416.

Foglio, M. (2019). Animal Wildlife Population Estimation Using Social Media Images Collections. Master's thesis, University of Illinois, Chicago, Illinois, USA.

Giachanou, A., Zhang, G., and Rosso, P. (2020). Multimodal multi-image fake news detection. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 647–654.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Hu, C., Yin, M., Liu, B., Li, X., and Ye, Y. (2021). Detection of Illicit Drug Trafficking Events on Instagram. In *Proceedings of the 30th ACM International Conference on Information Knowledge Management*, pages 3838–3846, New York, NY, USA. ACM.

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.

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Mazars-Simon, A. E. (2019). The Wild in Live Project: A Human/Algorithm learning network to help citizen science in wildlife conservation. Master's thesis, Universidade de Coimbra.

Morais, P., Afonso, L., and Dias, E. (2021). Harnessing the Power of Social Media to Obtain Biodiversity Data About Cetaceans in a Poorly Monitored Area. *Frontiers in Marine Science*, 8.

Nascimento, L. (2020). Monitoring jellyfish population by social media. Technical report, Pós-Graduação em Sistemas Costeiros e Oceânicos, Universidade Federal do Paraná.

Ofli, F., Alam, F., and Imran, M. (2020). Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response. *arXiv preprint arXiv:2004.11838*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Taklis, C., Giovos, I., and Karamanlidis, A. A. (2020). Social media: a valuable tool to inform shark conservation in Greece. *Mediterranean Marine Science*, 21(3):493–498.