Desvendando Dados Corrompidos: Uma Jornada de Limpeza, Transformação e Geolocalização em Registros Ambientais *

Mateus A. G. Oliveira[1], **Aurora T. Pozo**[1], **Carmem S. Hara**[1]

[1] Departamento de Informática
Universidade Federal do Paraná (UFPR) – Curitiba, PR – Brasil

{mago23,aurora,carmem}@inf.ufpr.br

Abstract. *Data quality is essential for reliable analyses in all application domains and, in particular, in environmental studies. This paper addresses the cleaning and normalization of data on sightings of portuguese man o'war (Physalia physalis) along the Brazilian coast, collected from social media and scientific literature. Issues with date formats, character encoding, and geographic inaccuracies were corrected using standardization, automated fixes, and the Google Geocoding API. The results highlight the importance of data curation and integration to enhance the data quality and enable more accurate analyses.*

Resumo. *A qualidade dos dados é essencial para análises confiáveis em todas as áreas de conhecimento e, em particular, em estudos ambientais. Este artigo aborda a limpeza e normalização de dados sobre avistamentos de caravela-portuguesa (Physalia physalis) no litoral brasileiro, coletados de redes sociais e literatura científica. Foram corrigidos formatos de data, codificação de caracteres e imprecisões geográficas, utilizando padronização, correção automatizada e a API de Geocodificação do Google. Os resultados destacam a importância da curadoria e integração de dados para melhorar sua qualidade e viabilizar análises mais precisas.*

1. Introdução

A aquisição, extração e armazenamento de dados provenientes de uma variedade de fontes frequentemente introduz uma série de erros nos registros, como valores ausentes, erros de digitação, mistura de formatos, duplicatas e violações de regras, entre outros [Ilyas and Chu 2019]. O problema de dados "sujos" é um dos mais comuns na rotina de quem trabalha com dados, sendo uma tarefa exaustiva que pode consumir até 60% do tempo total dedicado ao projeto [Ilyas and Chu 2019]. A curadoria, unificação, preparação e limpeza dos dados são etapas essenciais para que as análises resultem em insights valiosos sobre um domínio específico. Assim, o desenvolvimento de soluções eficazes para a limpeza de dados é fundamental [Ilyas and Chu 2019].

O objetivo deste artigo é apresentar o processo de limpeza realizado em um conjunto de dados sobre avistamentos de caravelas-portuguesas (*Physalia physalis*) no litoral do Brasil. Este conjunto de dados foi criado pela Dra. Lorena Silva do Nascimento e utilizado em sua tese de doutorado [NASCIMENTO 2023], que avaliou a utilidade das mídias

*Este trabalho foi parcialmente financiado pelo CNPq (Processo 407644/2021-0), CAPES (Programa de Excelência Acadêmica - PROEX) e UFPR-TN (BOLSA PIBIC).

sociais como fonte de dados para a obtenção de observações da caravela-portuguesa no Brasil, além de examinar a distribuição dos riscos à saúde humana causados por esta espécie. A limpeza foi realizada com o objetivo de normalizar e padronizar os dados, facilitando a identificação de duplicatas e a exploração subsequente dos dados.

Os registros foram obtidos a partir de diversas fontes, como literatura, redes sociais, sites de notícias, e foram coletados manualmente ou através de técnicas de *Web Scraping*. Antes da limpeza, o conjunto de dados apresentava inconsistências nos formatos de datas, com diferentes padrões sendo usados. Além disso, alguns registros incluíam intervalos de tempo em vez de datas específicas. Em relação às localizações, alguns registros sofreram problemas de codificação de caracteres, resultando em símbolos no lugar de letras acentuadas. Havia também inconsistências na geolocalização, onde endereços eram representados como strings em vez de coordenadas geográficas precisas.

A motivação principal para a limpeza dos dados foi garantir a consistência e a precisão das informações de avistamento antes de integrá-las a um sistema de análise de grande escala. Dada a diversidade das fontes de coleta — como *Instagram*, *iNaturalist*[1] e Literatura —, foi essencial uniformizar e completar dados críticos de data e localização. Com isso, busca-se obter uma base sólida que permita análises geográficas e temporais mais confiáveis, oferecendo uma visão detalhada do comportamento dos avistamentos costeiros e assegurando uma base de dados padronizada para estudos futuros. Além disso, pretende-se desenvolver uma biblioteca contendo todas as técnicas de limpeza de datas e localizações utilizadas neste trabalho, de forma a atender às demandas da comunidade científica.

O restante do artigo está organizado da seguinte forma: a Seção 2 apresenta trabalhos relacionados ao tema de limpeza de dados e novos métodos na área. A Seção 3 descreve os desafios encontrados durante o processo de limpeza e detalha os métodos e técnicas utilizados na execução do processo. A Seção 4 apresenta os resultados obtidos, comparando o estado dos dados antes e depois da limpeza. Por fim, a Seção 5 conclui o artigo e discute possíveis trabalhos futuros.

2. Trabalhos Relacionados

O processo de limpeza de dados é reconhecidamente uma tarefa que demanda tempo e esforço significativo por parte dos profissionais da área. Para abordar esse desafio, a comunidade acadêmica tem se dedicado a desenvolver soluções que facilitem, otimizem e aumentem a precisão desse processo, frequentemente utilizando técnicas de aprendizado de máquina [Côté et al. 2024].

Um exemplo é o RLClean [Peng et al. 2024], um framework de limpeza de dados considerado pelos seus autores como o primeiro a empregar *deep reinforcement learning* (DRL) para essa finalidade. O RLClean é capaz de combinar múltiplas técnicas para detectar e reparar erros de forma integrada, aprendendo continuamente a maneira mais otimizada de executar seus métodos [Peng et al. 2024].

Além disso, existem abordagens que não dependem de técnicas avançadas de aprendizado de máquina, mas que ainda assim aprimoram significativamente o processo de limpeza de dados. Malek et al. [Malek and Jalil 2025], por exemplo, apresentam

[1] www.inaturalist.org

um pipeline projetado para tratar os erros mais comuns no processamento de dados, incluindo normalização, detecção e remoção de duplicatas, tratamento de valores faltantes, identificação e correção de *outliers*, além da transformação de dados. Essas técnicas são semelhantes às aplicadas em nosso processo de limpeza. No entanto, elas são aplicadas no contexto de detecção de fraudes, enquanto neste trabalho o foco é desenvolver scripts projetados para atender às particularidades da base de avistamentos de uma espécie.

3. Limpeza de Dados Temporais e de Localização

Em bases de dados de eventos ambientais, como a de avistamentos de caravelas-portuguesas, a precisão das informações temporais e de localização é fundamental. Esta seção apresenta os principais problemas identificados nesses atributos e as soluções aplicadas para mitigá-los.

3.1. Normalização de Datas

Durante a análise exploratória dos dados presentes nos atributos **`Post date`** (data da postagem sobre o avistamento) e **`Sighting date`** (data do avistamento), identificou-se que algumas datas estavam incompletas, sem a informação do dia específico. Como o formato adotado exigia a inclusão do dia, foi necessário definir uma abordagem para padronizar esses registros. Optou-se por preencher as datas incompletas utilizando o primeiro dia do mês correspondente, garantindo a consistência do conjunto de dados e evitando a exclusão dessas informações.

Além disso, alguns registros indicavam um intervalo de tempo para os avistamentos de caravelas-portuguesas, o que exigiu um ajuste na estrutura da base de dados. Para esses casos, foram criadas colunas separadas para representar o início e o fim do avistamento. Quando a data indicava apenas um único dia, considerou-se que o evento ocorreu no mesmo dia tanto para início quanto para término. Outro desafio encontrado foi a presença de múltiplas datas em uma única célula da planilha, o que dificultava a análise e consulta dos dados. Para evitar a perda de informações, decidiu-se criar novos registros individuais para cada data, mantendo as demais informações inalteradas.

A heterogeneidade nos formatos de data representou mais um obstáculo. Foram identificados diferentes padrões, como MM-DD-AAAA, AAAA-MM-DD e MÊS/DIA/ANO, além de registros que especificavam apenas um mês ou até mesmo uma estação do ano. Para viabilizar a pesquisa e padronizar a base de dados, todas as datas foram convertidas para o formato AAAA-MM-DD.

Para otimizar o agrupamento e a análise temporal dos dados, novos atributos foram criados. As datas de avistamento foram padronizadas como **`sighting_date_start`** e **`sighting_date_end`**, enquanto as datas de postagem foram estruturadas no atributo **`post_date`**, garantindo maior consistência e uniformidade.

Registros que continham apenas o mês e o ano da observação foram ajustados para incluir o primeiro dia do mês como data de início e o último dia como data de fim. Já aqueles que indicavam apenas a estação do ano foram tratados com as datas correspondentes ao início e fim desse período no respectivo ano da observação.

Para automatizar e garantir a precisão desse processo, desenvolveu-se um script em *Python* utilizando a biblioteca `datetime`. O código, disponível no repositório do

GitHub[2], define diversos padrões de data e converte automaticamente os registros para o formato padronizado. Caso um formato desconhecido seja encontrado, um erro é retornado, permitindo a revisão manual do dado.

3.2. Limpeza de Dados de Localização

Os dados de localização do avistamento estavam originalmente distribuídos em três atributos: **`Location`** (descrição textual da localização), **`Geolocation`** (coordenadas geográficas) e **`State`** (estado correspondente ao local do avistamento).

Grande parte dos registros extraídos do *iNaturalist* apresentou problemas de codificação, possivelmente devido à leitura equivocada de caracteres originalmente em UTF-8 como ISO-8859-1 (Latin-1), resultando na substituição de caracteres acentuados por símbolos incorretos. Tentativas de conversão direta não foram eficazes, sugerindo uma possível corrupção anterior dos dados. Para solucionar esse problema, utilizou-se Inteligência Artificial Generativa, por meio da API do GPT, que analisou os padrões nos símbolos corrompidos e auxiliou na restauração dos caracteres originais. Foi criado um *prompt* específico solicitando a identificação de padrões recorrentes entre os caracteres corrompidos. Com o padrão identificado, os dados puderam ser corrigidos automaticamente.

Os registros extraídos do *Instagram* apresentaram outro desafio: as informações de localização, inseridas manualmente pelos usuários, muitas vezes continham apenas nomes de praias ou apelidos regionais, dificultando a obtenção de coordenadas precisas. Exemplos como "Hostel Fran", "Aonde o mal não me atinge" e "vcgobuy com" ilustram a baixa confiabilidade desses dados. Inicialmente, utilizou-se a biblioteca Geopy, mas sua cobertura limitada para locais pouco conhecidos levou à adoção da Google Maps API, que ofereceu maior precisão, embora com restrições impostas pelo custo por requisição. Havia também ambiguidade na interpretação dos atributos **`Geolocation`** e **`Location`**. Enquanto **`Location`** continha descrições genéricas como nomes de cidades ou pontos turísticos, **`Geolocation`** deveria fornecer coordenadas específicas. Para uniformizar os dados, os valores de **`Geolocation`** foram transformados em dois novos atributos: **`Latitude`** e **`Longitude`**.

A obtenção de coordenadas precisas foi feita por meio da Geocoding API do Google, que converte endereços em latitude e longitude. Para otimizar o retorno, uma string de pesquisa foi construída combinando os campos **`Location`**, **`Geolocation`** e **`State`**, já existentes na base de dados, sempre adicionando "Brazil" ao final. Caso a API retornasse dados inválidos ou se as coordenadas já estivessem preenchidas, o script ignorava a entrada.

Outro problema foi a validação das coordenadas extraídas do Instagram, pois muitos usuários utilizam localizações fictícias ou imprecisas. Para mitigar esse problema, um script em *Python* foi desenvolvido para verificar se as coordenadas estavam dentro de um raio de 40 km do litoral brasileiro. Para isso, recorreu-se a um arquivo *shapefile* do Natural Earth Data[3], permitindo traçar a linha costeira do Brasil e validar as localizações. Se os dados fossem válidos, a resposta era processada e armazenada na base, junto com um novo atributo que atesta a proximidade ao litoral chamado de **`onCoastline`**.

[2]https://github.com/MateusPersonalProjects/PortugueseManOfWarIC

[3]naturalearthdata.com

Durante a limpeza dos dados de localização, criou-se um novo atributo chamado **City**, extraído do campo **Location**, para facilitar buscas e evitar redundâncias com os atributos **State** e **Location**. Inicialmente, tentou-se eliminar duplicações utilizando a API de Reverse Geocoding do Google, mas essa abordagem resultou em perda de informações. Como alternativa, desenvolveu-se um script em *Python* para comparar e remover duplicatas. No entanto, devido a variações ortográficas e erros nos dados do *Instagram*, foi necessária uma revisão manual para garantir a consistência.

Quando as colunas **City** e **State** não eram preenchidas pela Geocoding API, uma segunda tentativa utilizava a Reverse Geocoding API para obter essas informações a partir das coordenadas já registradas. Caso as coordenadas fossem inválidas, essa inconsistência era reportada ao usuário. O código completo desse processo pode ser encontrado no repositório do GitHub[4].

4. Resultados

A base de dados de avistamentos de caravelas-portuguesas contém 1.457 registros e originalmente continha 15 atributos. Os atributos eram os seguintes: **ID:** Identificação do registro; **Source:** Identificação da fonte do registro; **Geolocation:** Geolocalização do avistamento; **Location:** Localização do avistamento; **State:** Estado referente à localização do avistamento; **Sighting date:** Data do avistamento; **Post date:**Data da postagem sobre o avistamento; **Abund:** Quantidade aproximada de caravelas portuguesas avistadas.; **Size:** Tamanho aproximado das caravelas-portuguesas avistadas; **Ecological Relation/Other strandings:** Descrição da interação biológica; **Human interaction/Accident:** Descrição da interação com humanos e se houve acidente durante a interação; **Age:** Idade do indivíduo que teve a interação; **Sex:** Sexo do indivíduo que teve a interação; **Site:** Local do corpo onde houve contato; **URL:** Link ou informações sobre a proveniência dos registros. Após o processo de limpeza descrito na Seção 3, novos atributos foram definidos: **sighting_date_start**: data inicial do avistamento; **sighting_date_end**: data final do avistamento; **Latitude**: latitude do avistamento; **Longitude**: longitude do avistamento; **City**: cidade do avistamento; **onCoastLine**: *flag* indicando se as coordenadas estão no litoral brasileiro.

4.1. Transformação das datas

Os resultados para transformações do atributo **Sighting Date** variaram entre as fontes de dados. No *Instagram*, 684 dos 863 registros estavam vazios, enquanto 179 continham datas, das quais 109 foram normalizadas. Na Literatura, todos os 122 registros tinham datas, com 26 modificações. No *iNaturalist*, os 276 registros estavam corretos, sem necessidade de ajustes como mostra a Figura 1.

A normalização de **Post Date** variou entre as fontes: no *Instagram*, 347 das 863 datas precisaram ser modificadas, enquanto todos os registros da literatura estavam vazios. Os dados do *iNaturalist* estavam corretamente preenchidos, como apresentado na Figura 1.

Ao analisar os resultados para **Sighting Date**, percebe-se que as fontes de dados da Literatura e do *iNaturalist* forneceram registros mais consistentes e precisos em

[4]https://github.com/MateusPersonalProjects/PortugueseManOfWarIC

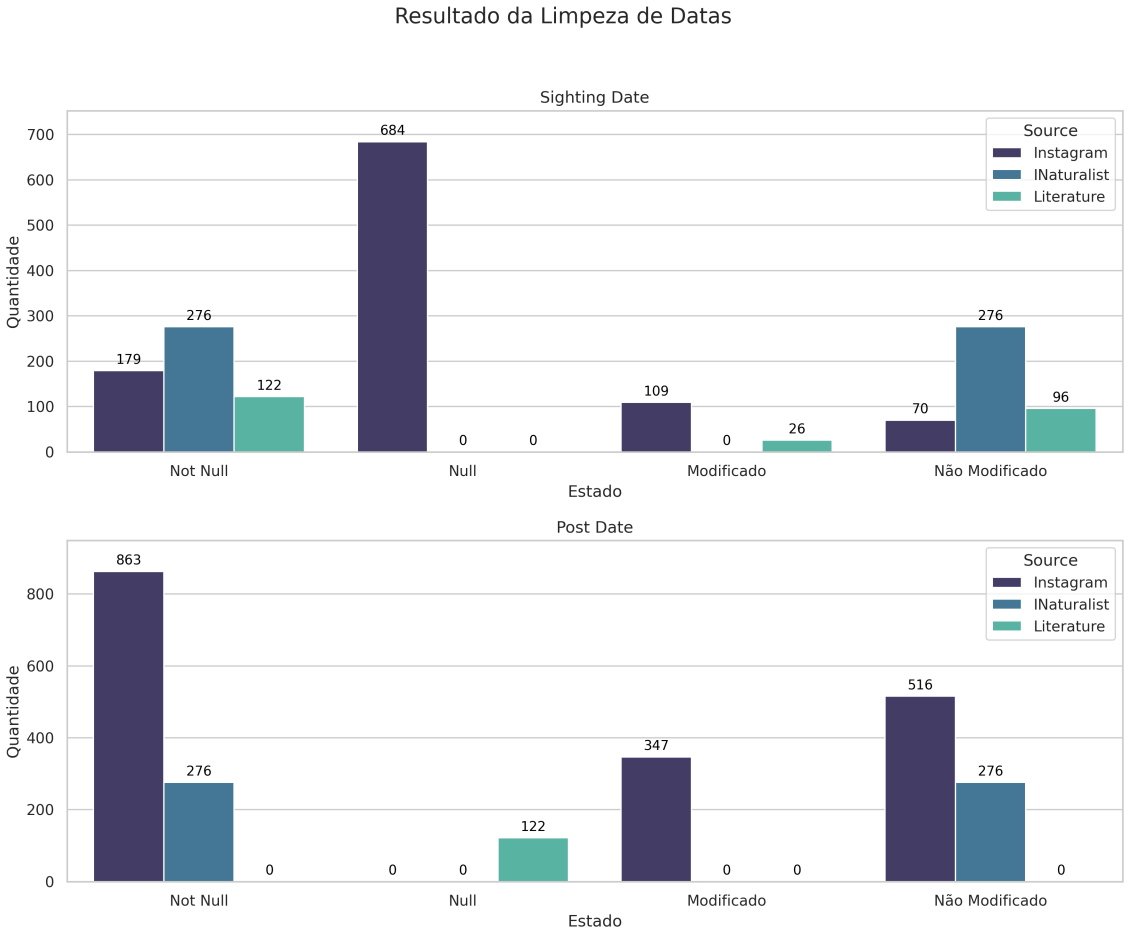

Figura 1. Distribuição dos resultados após a normalização dos atributos `Sighting Date` e `Post Date`. A figura ilustra o número de registros que foram modificados e os que permaneceram inalterados em cada fonte de dados analisada (*Instagram*, Literatura e *iNaturalist*). Ela destaca as diferenças no volume de dados processados e a qualidade das informações fornecidas por cada fonte, evidenciando as particularidades de cada conjunto de dados em termos de completude e necessidade de ajuste.

relação à data dos avistamentos. Por outro lado, os dados do *Instagram* apresentaram uma quantidade significativa de registros sem datas (Null), refletindo a natureza imprevisível dessa rede social, onde a data de publicação nem sempre coincide com a data real do evento. Isso justifica a maior necessidade de modificações nesses dados. Todas as datas presentes foram padronizadas.

Já para o atributo **`Post Date`**, o *Instagram* e o *iNaturalist* apresentaram todos os campos preenchidos, o que reflete sua natureza como redes sociais. No entanto, as modificações feitas no Instagram indicam a necessidade de ajustes manuais, provavelmente devido à coleta manual desses dados. O *iNaturalist*, por outro lado, mostrou maior consistência, com todos os dados já padronizados, o que reflete a qualidade dessa fonte em relação à precisão temporal. Todas as datas presentes foram padronizadas.

4.2. Localização e Geolocalização - Descoberta de Latitude e Longitude

A verificação das coordenadas de latitude e longitude, com base na Google Geocoding API, variou entre as fontes de dados: no *Instagram*, 312 dos 863 registros com dados de **`Location`** retornaram resultados válidos, e 289 estavam a até 40 km do litoral. No *iNaturalist*, o campo **`Location`** estava vazio, e na literatura, 113 dos 122 registros com dados retornaram resultados válidos, sendo 107 dentro do raio de 40 km do litoral como apresentado na Figura 2.

Quanto ao campo **`Geolocation`**, 585 dos 628 registros do *Instagram* retornaram resultados válidos, com 555 dentro do raio de 40 km. No *iNaturalist*, 260 dos 276 registros tiveram resultados válidos, com 218 dentro do raio de 40 km, e na literatura, 42 dos 108 registros válidos estavam dentro desse limite como apresentado na Figura 2.

Por fim, com a combinação dos campos **`Location`**, **`Geolocation`** e **`State`**, todos os 864 registros do *Instagram* e 276 do *iNaturalist* retornaram resultados válidos, com 832 do Instagram e 254 do iNaturalist dentro do raio de 40 km do litoral. Todos os 122 registros da literatura também estavam dentro desse raio como apresentado na Figura 2.

Os resultados mostram que a abordagem que combina os campos **`Location`**, **`Geolocation`** e **`State`** foi a mais eficaz, recuperando todas as localizações disponíveis e apresentando o maior número de registros dentro de um raio de 40 km do litoral. Em comparação com a base original, essa estratégia permitiu atribuir coordenadas precisas a todos os registros, sendo que 95,72% deles estão garantidamente localizados a até 40 km do litoral brasileiro. A Figura 2 compara os registros válidos retornados pela Google Geocoding API e quantos estavam dentro desse limite, evidenciando a melhora na precisão dos resultados com a nova abordagem.

5. Conclusão

Neste artigo, foi apresentado o processo de limpeza, transformação e enriquecimento dos dados, provenientes de três fontes principais: *Instagram*, *iNaturalist* e literatura. O foco esteve na normalização de datas e na recuperação de coordenadas geográficas. O trabalho enfrentou desafios como inconsistência de encodings, falta de precisão em localizações fornecidas por usuários. Houve a necessidade de combinar diferentes ferramentas e métodos, como a API do Google Maps, para alcançar a qualidade desejada nos dados.

A normalização de datas foi um passo essencial para padronizar as informações, especialmente com relação às fontes que não traziam as datas completas ou que as apresentavam em formatos variados. O problema das datas incompletas foi resolvido preenchendo os dias ausentes, e registros que continham mais de uma data foram separados para evitar perda de informação.

A recuperação das coordenadas geográficas foi realizada com sucesso por meio da Geocoding API do Google, uma vez que a biblioteca Geopy não se mostrou eficaz para lidar com localizações incompletas e apelidos regionais, frequentemente presentes nas postagens do *Instagram*. Apesar das limitações impostas pelo custo por requisição da API do Google, foi possível assegurar que as coordenadas obtidas estivessem dentro de um raio de 40 km do litoral brasileiro, o que é um fator relevante para a validação dos

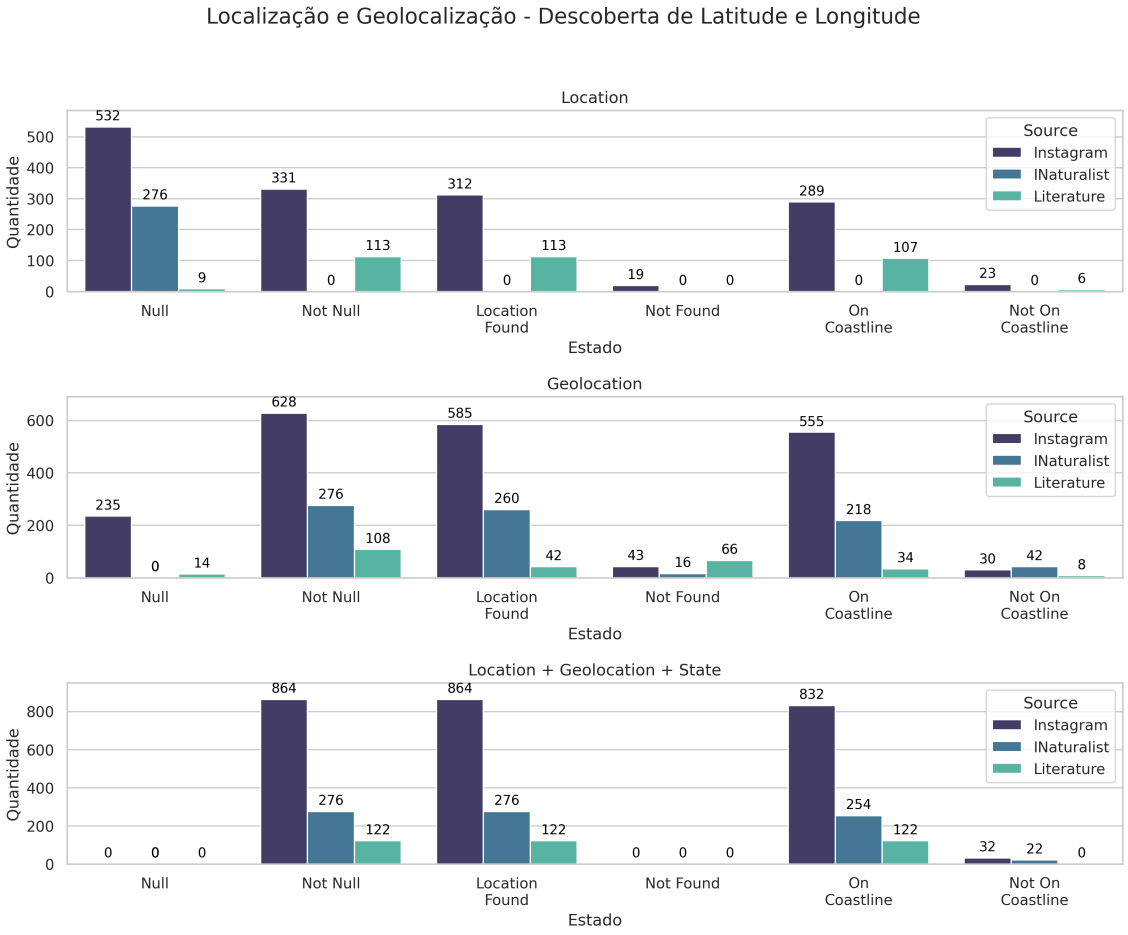

Figura 2. Resultados da obtenção de latitude e longitude para *Instagram*, *INaturalist* e *Literatura*. Cada gráfico exibe os resultados para três abordagens: utilizando os dados do campo `Location`, do campo `Geolocation`, e a nova abordagem que combina `Location`, `Geolocation` e `State`. As barras comparativas mostram a quantidade de registros válidos retornados pela Google Geocoding API e quantos desses registros estão a um raio de 40 km do litoral brasileiro. A figura evidencia a evolução na precisão dos resultados, com uma melhora significativa observada ao aplicar a nova abordagem.

dados de localização.

Além disso, o processo envolveu a verificação manual de duplicatas em localizações e a utilização de inteligência artificial para corrigir caracteres corrompidos em registros com problemas de encoding, o que foi crucial para recuperar dados que de outra forma seriam inutilizáveis.

Com esses ajustes e correções, os dados ficaram mais consistentes e confiáveis, permitindo que se tornem uma base sólida para futuras análises. A combinação de métodos automatizados e verificações manuais mostrou-se eficaz em superar os desafios da qualidade dos dados, demonstrando que, com as ferramentas certas, é possível lidar com dados complexos e imprecisos, e transformá-los em informação útil para investigação científica e outras aplicações.

Como trabalho futuro está planejada a inclusão de novas fontes de dados, como redes sociais adicionais e fontes científicas. Isso poderá aumentar a diversidade e representatividade geográfica e temporal das informações sobre avistamento de caravelas portuguesas no litoral brasileiro. Além disso, integrar modelos de predição e análise de tendências pode facilitar a previsão de padrões de avistamento com base em aspectos geográficos e sazonais, oferecendo insights úteis para estudos ambientais e de conservação marinha. Técnicas de análise de sentimento aplicadas a postagens de mídias sociais, usando NLP (Processamento de Linguagem Natural), poderiam ainda capturar percepções públicas sobre eventos ambientais. Por fim, para tornar a análise dos dados mais dinâmica e acessível a outros usuários e pesquisadores, o desenvolvimento de uma interface visual interativa seria útil para explorar avistamentos ao longo de diferentes períodos e regiões. Os scripts desenvolvidos neste trabalho foram projetados para abordar problemas específicos das planilhas e do contexto analisado. Embora eficazes para este caso, sua aplicação é limitada a situações similares. Por isso, seria relevante o desenvolvimento de uma biblioteca mais geral e abrangente, capaz de lidar com problemas de limpeza, normalização e análise de dados em diversos contextos e áreas do conhecimento. Essa biblioteca permitiria maior reutilização e flexibilidade, tornando o processo mais eficiente para futuros estudos ou projetos relacionados. Por fim, o uso de técnicas de *deep reinforcement learning* no processo de limpeza pode ser promissor, assim como demonstrado por [Peng et al. 2024], auxiliando na automatização e na portabilidade para outros projetos.

Referências

Côté, P. O., Nikanjam, A., Ahmed, N., Humeniuk, D., and Khomh, F. (2024). Data cleaning and machine learning: a systematic literature review. *Automated Software Engineering*, 31.

Ilyas, I. F. and Chu, X. (2019). *Data Cleaning*. ACM.

Malek, M. A. A. and Jalil, K. A. (2025). Enhancing data cleaning process on accounting data for fraud detection. *Indonesian Journal of Electrical Engineering and Computer Science*, 37:1014–1022.

NASCIMENTO, L. S. D. (2023). *REDES SOCIAIS COM FONTE DE DADOS ALTERNATIVA PARA MONITORAR ÁGUAS-VIVAS*. PhD thesis.

Peng, J., Shen, D., Nie, T., and Kou, Y. (2024). Rlclean: An unsupervised integrated data cleaning framework based on deep reinforcement learning. *Information Sciences*, 682.