Computer Science and Information Systems 21(4):1887–1911 https://doi.org/10.2298/CSIS240301064R

Identifying Occurrences of the Cnidarian *Physalia physalis* in Social Media Data*

Heloisa F. Rocha[1], Lorena S. Nascimento[2], Leonardo Camargo[1],
Mauricio A. Noernberg[2], Aurora T. Ramirez Pozo[1], and Carmem S. Hara[1]

[1] Universidade Federal do Paraná, Departamento de Informática,
Curitiba-PR, Brazil
heloisaferr@gmail.com
{camargo.s.leonardo, aurora.pozo, carmemhara}@ufpr.br
[2] Universidade Federal do Paraná, Centro de Estudos do Mar,
Pontal do Paraná-PR, Brazil
{lorena.sn, m.noernberg}@ufpr.br

Abstract. The Portuguese man-of-war (*Physalia physalis*), though beautiful, poses a risk to the population due to its potential to cause severe burns. Tracking their occurrences can prevent accidents through alerts to the population and predictive simulation models. However, traditional sources do not always provide records of their sightings. On the other hand, previous studies indicate that social media can be an effective source of information for conservation science. This work uses natural language processing and computer vision to obtain machine learning models to classify data extracted from Instagram. Such models can be used as part of an automated Extract-Transform-Load process to build a database on occurrences of *Physalia physalis* on the Brazilian coast. In preparation for training the models, we collected and manually labeled Instagram posts in order to distinguish the ones about the animal from other subjects, such as ships and tattoos. Given the nature of the problem, the spatial and temporal information associated with the sightings are essential for biologists. Thus, the absence or nonvalidity of such data is often used as a rationale to reject the post. However, the same criteria may not be suitable for training machine learning models to classify new posts automatically. The main goal of this article is to highlight the importance of choosing appropriate labels to train both text and image models, as well as to take into consideration the rejection criteria of the biologist before using a classification model. An experimental study is presented to show the effect of unquestioning adoption of labels given by a specialist, compared to labels adapted for machine learning training.

Keywords: Physalia physalis, dataset construction, labeling, machine learning.

1. Introduction

Portuguese man-of-war (cnidarian *Physalia physalis*) are multicellular organisms whose tentacles have stinging cells that can release harmful toxins. They often appear along the Brazilian coast, and due to their beautiful colors, people tend to touch them. As a result, there are many cases of envenomation, which may lead to severe injuries [6].

* This work is an extended version of the conference paper [26]

Although it is mandatory to be reported, SINAN (the Brazilian injury reporting system) lacks specific indications for *Physalia physalis* poisoning. Moreover, evidence suggests significant underreporting of cases [6]. The limited information about these animals in official public databases lead researchers to look for alternative sources of information.

One such alternative is to use involuntary data from social media. This approach, called passive citizen science, consists of using data generated by non-professionals, collected and shared on the Internet, mainly on social media [12].

Among the advantages resulting from using social media data are [12]: abundant and almost always public content; less labor intensive, less time consuming, and with lower cost, especially if automated, compared to traditional methods; data available almost in real-time and continuously; and data acquisition with wide geographic scope [22]. In addition, the use of information technology tools, such as machine learning, makes it possible to explore the potential of social media data and turn it into useful information.

In the *Caravela* project [23], approved by the Federal University of Paraná ethics committee, one of the specific goals is to compile data extracted from social media about occurrences of *Physalia physalis* on the Brazilian coast. This process was performed manually [24] and the results showed that the posts collected from Instagram represented 60% of the entire dataset, which included other sources such as newspapers (7%), iNaturalist[3] (19%), reports in the literature (8%), and personal observations (6%). The goal of the study was to determine the effectiveness of Instagram as a data source for spatial and temporal information of their sightings, as well as obtain the number, size and types of interactions with other species and human beings.

In this article, we report on the annotation process to show the importance of properly designing the labels that will be associated with each entry, in order to prepare the dataset for training machine learning models. The labeling process is required to distinguish between posts that are legitimate sightings of the species from others, such as tattoos and Portuguese boats.

It has been a rich real-world application experience. It contrasts with other machine learning works that consider benchmark datasets for which labels were already given. The design of the resulting labels involved a close interaction between a computer scientist and an oceanographer, with discussions of the requirements for the end user and for the training process to automate the classification task. Such a classifier can later be inserted into a search, classification and data storage system, which can continuously monitor the species.

During the annotation process, the oceanographer adopted some criteria for accepting a post as a legitimate occurrence, which included the requirement that its location is on the Brazilian coast. However, for the computer scientist, the spatial information may not be important for a model that classifies posts based only on their text and/or image. More than that, the dataset annotated by the specialist may introduce noise for training these models. To this end, the experiments presented in this article show that models trained with the adapted labels perform better than models trained with the labels originally given by the oceanographer. This work is an extended version of a conference paper [26]. The extension includes experiments with two adapted labels instead of only one, as presented in the original work. In the textual analyses, we added experiments with multilingual BERT, as a classifier and also as a feature extractor to feed a Logistic Regression model. In all

[3] https://www.inaturalist.org/

experiments, we used class weighting to deal with imbalanced data and cross-validation. In the original paper, we used these techniques only in textual analyses.

Moreover, considering that in the future the trained model could be used to develop a system to continuously extract posts from Instagram and automatically classify them as legitimate occurrences of *Physalia physalis*, we considered two different flows of this process. In the first one, all the collected posts are given as input to the trained model. In the second, posts that do not meet the spatial requirement are filtered out before being evaluated by the classifier. The latter closely resembles what is expected to be adopted in the real production chain. Our experiments show that it improves all metrics of the classifier compared to the former approach.

In summary, the contributions of the article are threefold:

- Show the importance of the knowledge on the part of the computer scientist of which criteria were used to label the data: our experiments show that adapting labels can produce better classification models;
- Compare the effect of the adapted labels on different modals (text and image), in the context of conservation science and social media data;
- Determine the performance of the classifier adopting two different flows of tasks: with and without prior filtering of posts based on the spatial criteria.

The rest of the article is organized as follows. Section 2 presents some related works that apply machine learning on data extracted from social media. The *Physalia physalis* dataset construction is described in Section 3, followed by details on the annotation process in Section 4. Experiments with adapted labels are presented in Section 5, while experiments simulating the production flow are presented in Section 6. Section 7 concludes the article by presenting some future works.

2. Related Work

Several works highlight the advantages and possibilities of using involuntary data, such as those available on social media, for tasks related to conservation science [12, 22]. In addition, there are some studies that apply machine learning to detect aspects of the environment, such as estimating the number of animals of a given species [14] and classifying land cover [13]. One can also find studies using Instagram to monitor other marine species [29, 19], but without the use of automated tools.

However, as already noted in other studies [15, 11], there are few works that applied machine learning to involuntary data for wildlife classification tasks. In particular, when searching for works that applied Natural Language Processing (NLP), it was not possible to find works that used texts in Portuguese extracted from Instagram to identify wildlife. In this Section, we report 6 works related to conservation science that used machine learning for wildlife classification tasks using involuntary data.

Mazars-Simon [20] developed a system that searches, classifies, identifies, and stores data on sea turtles. He used a dataset with 22,500 images, divided into six categories, which were shared by conservationists for the project. The author did not explain how the categories were defined.

In [18], the authors used 5,464 texts extracted from Google News and Twitter[4] to train a classifier to identify news about endangered species. They created the dataset automatically considering all data extracted from Twitter as positive and the data extracted from Google News as negative.

Another work [11] experimented with different combinations of representation techniques and machine learning algorithms to identify wildlife observations in Twitter texts. A dataset with 2,798 tweets was used. In their work, they described the collection and labeling process.

Convolutional neural networks to identify *Physalia physalis* in images extracted from Instagram were reported in [5]. A dataset with 12,300 images was used. Part of the training images were obtained from Instagram and labeled by an expert, and part was downloaded through search engines and specialized websites.

Cardoso et al. [4] investigated whether pre-trained and freely available deep learning models could support the identification of potential wildlife trade on online platforms. To train the models they collected 2.634 images of pangolins from iNaturalist, Flickr[5] and Google images. In their work, they described the collection and labeling process.

In [17], the authors implemented and evaluated a hierarchical text classification pipeline aiming to investigate the utility of machine learning in automating online analyses of wildlife exploitation. To evaluate the proposed pipeline they collected data on bats (Chiroptera) from Facebook[6], Twitter, Google, and Bing search engines. After filtering and deduplication they had a dataset of 413,937 unique social media posts and 104,279 web articles. The authors described the collection and labeling process.

Although the works listed in this section bear some resemblance to the problem studied in this article, only one of the models obtained by the authors can be used in the classification problem presented here [5]. The others have characteristics and objectives different from the problem addressed in this article. In [20] the goal is to classify images of turtles. In [18] the objective is to classify news as news about endangered species or not, which does not include the *Physalia physalis*. In addition to the text used in the training being in English, the characteristics of the text itself, in this case, news (usually long and formal texts), are different from the characteristics of social media texts (usually short and informal texts). In [11] the goal is to classify Twitter texts as being or not about wildlife observations. Furthermore, the authors used only texts in English to train the models, and there were no texts on our species of interest in their training database. In the work of [4] the goal was to investigate the use of deep learning to identify animal trade using pangolin images as a test case. Although the species of interest is different from ours, this work reinforces that the use of deep learning to identify species in data extracted from social media can present good results. In [17], the author proposed a pipeline for text classification using text on bats as a test case. In addition, the text used in the experiments was in English. In summary, as the main differentiator compared to works that propose to identify wildlife in social media data, we trained models with the caption of the posts and also with one of the images of the post. Moreover, none of these works focus on the impact of choosing appropriate labels for training a machine learning model

[4] Twitter is a social network and microblogging service. In July 2023, Twitter was renamed X, however, in this article we will use the original name: Twitter.

[5] https://www.flickr.com/

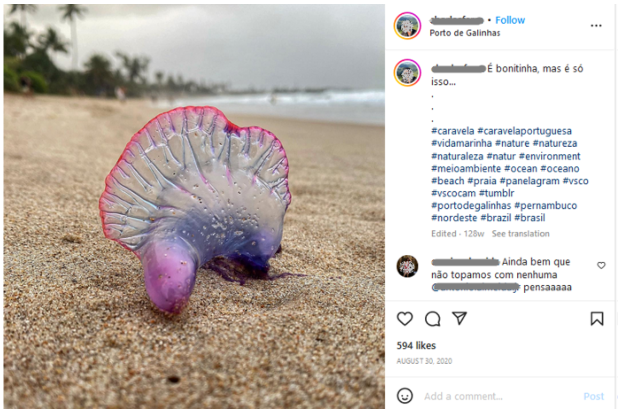
[6] https://www.facebook.com/

or compare the direct use of the dataset with adaptations that could increase the model performance. In the case of [5], the goal is to classify images of the species. In fact, we used the best architecture reported by the authors in our experiments (Sections 5 and 6).

3. Dataset Construction

This section describes how the data used to train the machine learning models proposed in this work were collected and labeled. It is important to recall that our research interest is in occurrences of *Physalia physalis* on the Brazilian coast. Therefore, the construction of the dataset followed this premise.

In this work, we used data extracted from Instagram. Instagram is a social network developed by Meta. Among its features is the sharing of media (i.e., photos and videos). Users can upload up to 10 media in a single post. They can also write a caption for the post and add its location. In addition, users can add hashtags to the caption. A hashtag can be written with text, emojis, and numbers. Spaces and special characters will not work. Figure 1 shows an example of an Instagram post about a Portuguese man-of-war.



[redacted] • Follow
Porto de Galinhas

[redacted] É bonitinha, mas é só isso...
.
.
.
#caravela #caravelaportuguesa #vidamarinha #nature #natureza #naturaleza #natur #environment #meioambiente #ocean #oceano #beach #praia #panelagram #vsco #vscocam #tumblr #portodegalinhas #pernambuco #nordeste #brazil #brasil

Edited · 128w See translation

[redacted] Ainda bem que não topamos com nenhuma [redacted] pensaaaaa

594 likes
AUGUST 30, 2020

Fig. 1. Example of an Instagram post about a Portuguese man-of-war. Source: Instagram

Instagram posts have some characteristics that should be considered:

- A post can have one or more images or videos. It is common to find posts with different media, for example: a picture of a *Physalia physalis*, a video showing the waves, and a picture of trash, all in the same post;
- As Instagram is an image-focused platform, posts do not always have caption;
- Although it is possible for the user to add location metadata to the post, it is not mandatory and sometimes the location of the occurrence is informed in the caption or comments of the post;
- There are linguistic variations in the caption. While some posts show the correct use of grammar rules, social media text is informal in nature, with many grammatical errors, slang, abbreviations, neologisms, internet jargon, and language mixture. Also,

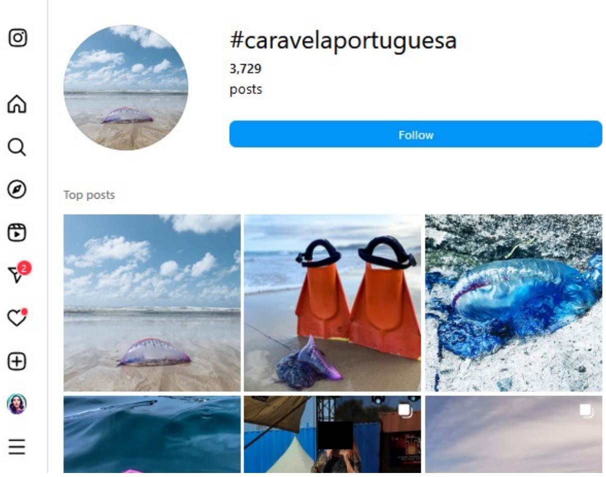


Fig. 2. Screenshot of the #caravelaportugesa page. Source: Instagram

some posts are formed almost exclusively by hashtags and emojis. Furthermore, the use of compound hashtags (e.g., #caravelaportuguesa, #goprobrasil) generates words that do not exist in the lexicon.

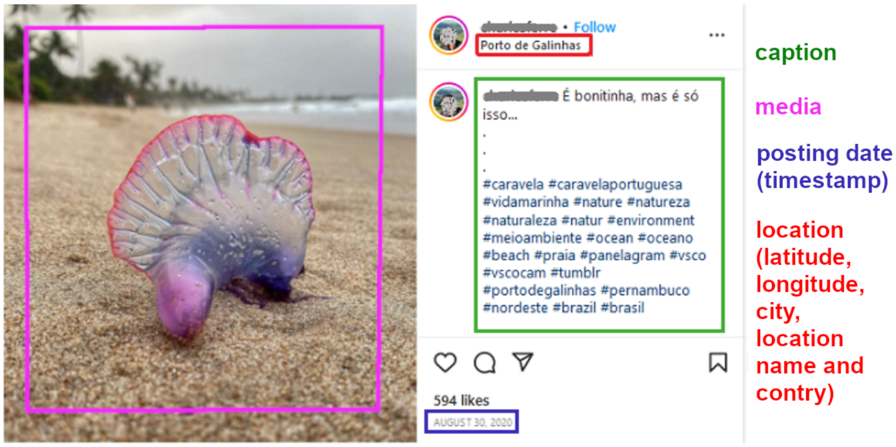
When a user has a public account and adds hashtags to a post, the post will be visible on the corresponding Hashtag Page. A Hashtag Page shows a *Top* section where the most popular posts tagged with the hashtag appear. The page also shows a *Recent* section, containing the most recent posts tagged with the hashtag. In this section, the posts appear in the order in which they were posted. Figure 2 shows a screenshot of the page #caravelaportugesa. It is important to know that each hashtag is unique: #cnidario is different from #cnidarios, #aguaviva is different from #águaviva, and so on. This means that one should not expect to find posts that have the hashtag #cnidarios in the #cnidario page, for example.

The first option considered to obtain Instagram data was the use of Instagram's API. However, this API does not return location metadata. So we chose to use the Instaloader library[7]. This library uses scraping to get data from Hashtag Pages and returns data in .json format. More details on the data extraction process were published in [3]. The data of interest for our research are the caption of the post, media, posting date (timestamp), and location (latitude, longitude, city, location name, and country). We did not collect data on users' profiles. Figure 3 illustrates the data of interest for this research.

We chose to use the data obtained through the following hashtags: #aguaviva, #caravelaportuguesa, #cnidarios, #cnidários, #cnidario, #cnidário and #physaliaphysalis.

The #aguaviva was chosen because it is a generic term sometimes used in posts about *Physalia physalis*. The search result for this hashtag returned about 151,000 posts, which include, in addition to *Physalia physalis*, tattoos, people, clothes, other cnidarians, among other things. The #caravelaportuguesa was chosen because it is the popular name of *Physalia physalis* in the Portuguese language and thus more assertive than

[7] https://github.com/instaloader/instaloader



[redacted] • Follow
Porto de Galinhas

[redacted] É bonitinha, mas é só isso...
.
.
.
#caravela #caravelaportuguesa #vidamarinha #nature #natureza #naturaleza #natur #environment #meioambiente #ocean #oceano #beach #praia #panelagram #vsco #vscocam #tumblr #portodegalinhas #pernambuco #nordeste #brazil #brasil

594 likes
AUGUST 30, 2020

caption
media
posting date (timestamp)
location (latitude, longitude, city, location name and contry)

Fig. 3. Illustration the data of interest for this research. In green is the caption of the post, pink is for media, blue is for posting date, and red is for location. Source: adapted from Instagram

#aguaviva. The result of the search for this hashtag returned around 3,300 posts, which included, in addition to *Physalia physalis*, boats, handicrafts, tattoos, clothes, among other things. The #physaliaphysalis, which is the scientific name of the species, was chosen because it has a high potential to obtain positive posts about *Physalia physalis*. A search for this hashtag returned around 2,500 posts. Cnidaria is the name of the phylum to which *Physalia physalis* belongs. Thus the hashtag #cnidarios, and some variations like #cnidário, #cnidario and #cnidários, were chosen because they are common names of this phylum in the Portuguese language and for being terms with medium potential to return positive posts about *Physalia physalis*. A search for #cnidarios returned around 3,400 posts which include *Physalia physalis* as well as other cnidarians such as *Velella velella* and medusa. All searches were performed on August 27, 2022. In total, we collected 6,204 posts dated between 2012 and 2021.

After downloading, the data were saved in spreadsheets to facilitate the annotation process. All data collected were stored in a password-protected file system. For posts that had geolocation (that is, latitude and longitude), the BigDataCloud API[8] was used to obtain additional location data (that is, city, location name, state, country). This data was used as a complement to existing location metadata.

Data were also enriched by identifying the caption language using the Langdetect library[9]. This library has a function that returns a vector indicating which languages were detected in the text along with the probability of the text being written in the language in question. The Portuguese language was assigned to the post whenever it appeared with some probability, even if small. As previously mentioned, among the characteristics of the data extracted from Instagram are the existence of posts using words from more than one language, in addition to the use of compound hashtags that generate words that do not exist in the lexicon. These characteristics make the task of assigning the language for the

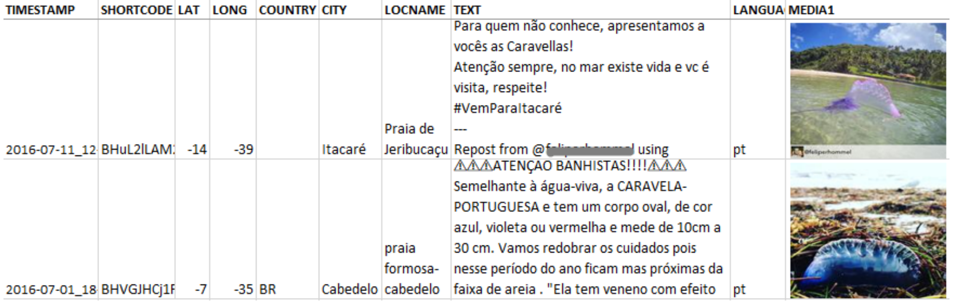
[8] https://www.bigdatacloud.com/
[9] https://github.com/Mimino666/langdetect

Table 1. Caption instances that the library did not detect the Portuguese language

Caption	Languages and Probabilities
Sou imortal!	Catalan 99%
#aguaviva Por USER #nemdói TELEFONE para orçamentos #tattoo #inked #inkwork #inkworld #inklife #inkedgirls #tatuagens #art #blackwork #sketch #brunotattoo #omegainkstudio	English 42%, Swedish 28% and Afrikaans 28%

caption difficult. For this reason, the language assigned by the library was reviewed by the computer scientist, with the goal of identifying the presence of terms used in Portuguese that could characterize the caption as written by a speaker of that language. After the review, the number of posts considered in Portuguese increased from 3,440 (according to the library) to 3,849 (after the review).

Table 1 shows 2 caption instances in which the library did not detect the Portuguese language. Figure 4 shows a screenshot of the spreadsheet created with the downloaded and enriched data.



TIMESTAMP	SHORTCODE	LAT	LONG	COUNTRY	CITY	LOCNAME	TEXT	LANGUA	MEDIA1
2016-07-11_12	BHuL2lLAM:	-14	-39		Itacaré	Praia de Jeribucaçu	Para quem não conhece, apresentamos a vocês as Caravellas! Atenção sempre, no mar existe vida e vc é visita, respeite! #VemParaItacaré --- Repost from @[redacted] using	pt	
2016-07-01_18	BHVGJHCj1F	-7	-35	BR	Cabedelo	praia formosa-cabedelo	⚠⚠⚠ATENÇAO BANHISTAS!!!!⚠⚠⚠ Semelhante à água-viva, a CARAVELA-PORTUGUESA e tem um corpo oval, de cor azul, violeta ou vermelha e mede de 10cm a 30 cm. Vamos redobrar os cuidados pois nesse período do ano ficam mas próximas da faixa de areia . "Ela tem veneno com efeito	pt	

Fig. 4. Screenshot of a spreadsheet created with the downloaded and enriched data

4. Data Annotation

The data collected were manually annotated by a computer scientist and an oceanographer. However, the purpose of the annotation differed for each person. The oceanographer was interested in determining the effectiveness of Instagram in providing temporal and spatial information of sightings on the Brazilian coast. On the other hand, the computer scientist was interested in generating an annotated dataset for training a machine learning model in order to automatically classify new Instagram posts, and thus continually monitor the species.

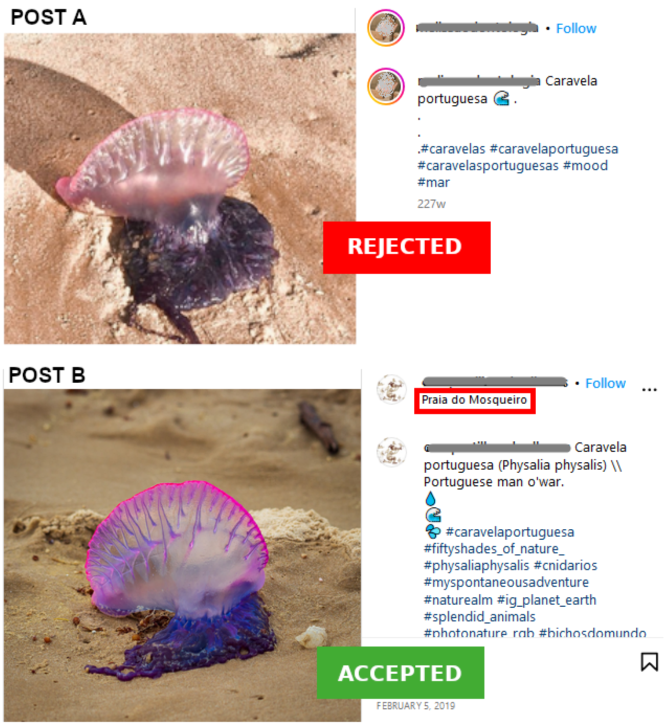
This distinction resulted in a large number of posts classified as legitimate occurrences of the species by one annotator and negative by the other.

The oceanographer used the following criteria to accept a post as a legitimate occurrence of *Physalia physalis*. It had to satisfy three conditions:

- Taxonomic identification: a media that clearly shows a *Physalia physalis*;
- Spatial information: a location on the Brazilian coast;
- Temporal information: include a timestamp associated with the post.

For the computer scientist, on the other hand, the lack of spatial and/or temporal information may not be important for a model that classifies posts based on their text and/or media. Thus, the dataset annotated by the specialist can include noise for training machine learning models.

Figure 5 illustrates the problem. Two posts extracted from Instagram are displayed. Both have an image of *Physalia physalis*, but post A does not have spatial information, while post B does. Following the expert's criteria, only post B is accepted as a legitimate occurrence of *Physalia physalis* on the Brazilian coast, while post A is rejected.



POST A

[redacted] • Follow

[redacted] Caravela portuguesa 🌊 .
.
.
.#caravelas #caravelaportuguesa #caravelasportuguesas #mood #mar

227w

REJECTED

POST B

[redacted] • Follow ...
Praia do Mosqueiro

[redacted] Caravela portuguesa (Physalia physalis) \\ Portuguese man o'war.
💧
🌊
💦 #caravelaportuguesa #fiftyshades_of_nature_ #physaliaphysalis #cnidarios #myspontaneousadventure #naturealm #ig_planet_earth #splendid_animals #photonature_rgb #bichosdomundo

ACCEPTED

FEBRUARY 5, 2019

Fig. 5. Illustration of the problem with using the original label. Both posts have an image of *Physalia physalis*, but post A does not have spatial information, while post B does. Following the expert's criteria, only post B is accepted as a legitimate occurrence of *Physalia physalis* on the Brazilian coast, while post A is rejected. Source: adapted from Instagram

To overcome this problem, we decided to adapt the labels, accepting posts that were rejected because of spatial information. However, before assigning a binary label to the entire post, we needed to annotate their taxonomic and spatial information. These new labels weren't given by the oceanographer but were given by the computer scientist with machine learning in mind. This annotation helped us reach a consensus on the binary labels, solve doubts and ambiguities, increase transparency, and promote the reproducibility of the annotation process.

These taxonomic and spatial classes allowed the adaptation of the binary label in order to reduce noise and generate better classification models, as reported in experiments comparing original and adapted labels (Section 5). In addition, they make it possible to carry out other experiments and studies using the enriched data. As an example, they can be used to train a model to identify accidents with *Physalia physalis* using the taxonomic information.

Spatial Criterion For the spatial criterion, we evaluated whether the post contained spatial information, and in particular, if the location was on the Brazilian coast. Note that a post may have several metadata about location: latitude, longitude, city, location name, and country. In addition, we enriched the metadata using latitude and longitude as we described in Section 3, and sometimes the user informed the location in the caption. There may be a difference between the location obtained by latitude and longitude and the location informed by the user on the metadata (city, location name, and country) or in the caption. The reason for this is that the user can post while being in one location, but referring to something that happened in another location.

Therefore, for the evaluation of spatial information, the following order of precedence was considered: first, the data contained in the caption, second the data from the metadata: city, location name, and country, and lastly location obtained by latitude and longitude. It should be noted that the word or hashtag `portugal` alone was not used as location information, as it often appears as a reference to Portuguese culture, such as the post that talks about a visit to the Museu Paranaense[10]. It says: *"Navegar é preciso #caravelaportuguesa #portugal #brasil #museuparanaense"*[11]. Also, just the state (e.g., #bahia, Ceará), #beach or names of very common beaches (e.g., *Praia Grande*) were not used alone to determine the location.

The spatial information was classified as follows:

- COAST-TEXT: The caption contains some information that allows identifying the location as being on the Brazilian coast (even if there is metadata indicating otherwise).
- COAST-GEO: The post has metadata that allows identifying the location as being on the Brazilian coast.
- COUNTRYSIDE-TEXT: The caption contains some information that indicates that the post did not take place on the Brazilian coast, but still the location is in Brazil.
- COUNTRYSIDE-GEO: The post has metadata indicating that the post did not take place on the Brazilian coast, but still the location is in Brazil.
- FOREIGNER: The post has information in the caption or metadata that indicates that the post did not take place in Brazil.
- NOTHING: There is no information that allows the post location to be identified.

In some cases the location has been identified as being in a coastal city, but it is clear that it is not on the beach, as in the examples: *"Aquário Marinho do Rio de Janeiro"* and

[10] Museu Paranaense is a museum located in the Brazilian city of Curitiba

[11] Translated as: "Sailing is necessary #caravelaportuguesa #portugal #brasil #museuparanaense"

"Centro Acadêmico de Vitória - CAV - UFPE"[12]. These posts are classified as COUNTRYSIDE-TEXT when the information is part of the caption and as COUNTRYSIDE-GEO when the information is part of the metadata.

Taxonomic Criterion For the taxonomic criterion, we evaluated the media of the posts, especially if the images are photos of *Physalia physalis*. However, the existence of a *Physalia physalis* picture does not always mean the post will be accepted as a legitimate occurrence. It is important to recall that a single post can have up to 10 media, so each media can have a different classification. The media of the posts were classified as follows.

- REALISTIC: The media is a realistic picture of *Physalia physalis*, either on the beach sand or in the sea. In addition, when a post contains several images, they appear to be from the same occurrence.
- CLOSE: The media is a realistic picture of *Physalia physalis*, but it is a close-up image, showing only parts of *Physalia physalis*.
- DISPLACED: The media is a realistic picture of *Physalia physalis*, but displaced from its habitat. For example, it is in an aquarium or inside a plastic bag.
- EDITED: The media is a realistic picture of *Physalia physalis* but with minor edits such as the inclusion of authorship, text, or frame, but that do not disturb the visualization of *Physalia physalis*.
- COLLECTION: The images in the post are a collection of images from *Physalia physalis*, but it is possible to notice that they are not from the same occurrence and are possibly images taken from the internet.
- ART: The image is a realistic representation of *Physalia physalis*, such as: compositions, drawings, tattoos. It could be a realistic image of *Physalia physalis* with edits that reveal this is not a sighting of the species.
- ACCIDENT: The media is a picture of a body part "burned" by a *Physalia physalis*.
- CNIDARIA: The media is a picture of another cnidarian, for example: *Velella velella*.
- VIDEO: The media is a video and therefore was not evaluated for this criterion.
- NOTHING: The media cannot be sorted in any other way described above.

Binary Label Based on the spatial and taxonomic annotations, the posts were labeled as ACCEPTED or REJECTED as legitimate occurrences of *Physalia physalis* on the Brazilian coast. For Instagram posts, the temporal criterion is always satisfied, because every post has an associated timestamp. The given label is also associated with a justification as described next.

To be *accepted* as a legitimate occurrence of *Physalia physalis* on the Brazilian coast, the post must meet taxonomic and spatial criteria. It means that: the post must have at least one image classified as REALISTIC, COLLECTION, EDITED, DISPLACED, or CLOSE (even if the post has other media that do not fit these five classes), and have the spatial information classified as COAST-GEO or COAST-TEXT.

One post that does not meet one or more criteria is considered *rejected* as a legitimate occurrence of *Physalia physalis* on the Brazilian coast. The rejection can be justified as:

[12] Translated as: "Marine Aquarium of Rio de Janeiro" and "Vitoria Academic Center - CAV - Federal University of Pernambuco"

- BECAUSE OF THE MEDIA: The post has spatial information classified as COAST-GEO or COAST-TEXT, but no media were classified as REALISTIC, COLLECTION, EDITED, DISPLACED, or CLOSE.
- BECAUSE OF THE LOCATION: The post has at least one media classified as REALISTIC, COLLECTION, EDITED, DISPLACED, or CLOSE, but the spatial information is classified as COUNTRYSIDE-TEXT, COUNTRYSIDE-GEO, FOREIGNER or NOTHING.
- BECAUSE OF THE MEDIA AND LOCATION: The post does not meet both criteria.

4.1. Distribution by Label

Table 2 shows the number of accepted and rejected posts per hashtag. To simplify the analysis, we combined all the data from the hashtags: #cnidario, #cnidarios, #cnidario and #cnidarios. So from now on, when we write #cnidario, we are considering all posts with these hashtags.

In general, the number of positive posts for the occurrence of *Physalia physalis* on the Brazilian coast was small, mainly for the hashtags: #aguaviva, #cnidarios, and #physaliaphysalis. We expected to find more positive posts in the search for #physaliaphysalis, but only 8% of the collected posts were accepted as legitimate occurrences of *Physalia physalis* for this hashtag. Despite the small number of positive occurrences found with #aguaviva (1%), the period surveyed was also the shortest (175 days), which makes room to invest more effort in the search for the species in posts that have this hashtag. The search for occurrences using the hashtag #caravelaportuguesa had the best result, with 422 posts accepted by the specialist.

The table also shows that the number of rejected posts is much larger than the number of accepted ones.

Table 2. Number of posts per Hashtag and Label. The percentage is based on the number of posts by hashtag

Hashtag	Label	No. Posts	%
#aguaviva	ACCEPTED	23	1%
	REJECTED	1,657	99%
#caravelaportuguesa	ACCEPTED	422	24%
	REJECTED	1,364	76%
#cnidario	ACCEPTED	77	3%
	REJECTED	2,472	97%
#physaliaphysalis	ACCEPTED	15	8%
	REJECTED	174	92%

5. Experiments with Adapted Labels

As discussed earlier, posts rejected because of location may represent noise for training machine learning models. In this section, we investigate whether there is any impact on the trained model if we consider as accepted posts rejected because of spatial information. We performed experiments with the caption and one of the images of the posts and built three annotated datasets:

- `original`: the one annotated with the oceanographer criteria.
- `adapted-lack-location`: considers as ACCEPTED posts rejected only for lack of spatial information (i.e., spatial information equals NOTHING).
- `adapted-not-in-coast`: considers as ACCEPTED posts rejected because of spatial information (i.e., spatial information equals NOTHING, COUNTRYSIDE-TEXT, COUNTRYSIDE-GEO OR FOREIGNER).

For the training, validation, and testing of the machine learning models we first filtered the dataset as follows. We deleted posts that have the following characteristics: empty text (12 posts); duplicate posts identified by having the same ID (164 posts); location outside Brazil (1,991 posts); and with video only (643 posts). The latter refers to posts that do not have images, but just videos. We removed them because we chose to do standard image classification. Moreover, we kept posts identified as Portuguese language and those with only emojis in the caption because emojis can be translated into Portuguese. After applying the filters, 2,610 posts remained in the dataset, which were used in the experiments described next. Table 3 shows the number of remaining posts in the original and adapted datasets.

Table 3. Number of remaining posts per Label

Label	original		adapted-lack-location		adapted-not-in-coast	
	No. Posts	%	No. Posts	%	No. Posts	%
ACCEPTED	470	18%	648	25%	658	25%
REJECTED	2,140	82%	1,962	75%	1,952	75%

Table 4 shows the number of rejected posts per reason for rejection. There were 188 posts rejected because of location, that were considered accepted in the adapted datasets.

Table 4. Number of remaining rejected posts per reason for rejection

Reason for Rejection	No. Posts
LOCATION	188
MEDIA AND LOCATION	1,549
MEDIA	403
TOTAL	2,140

Table 5 details these 188 posts per type of spatial information. There are 178 posts rejected because of lack of location (NOTHING). They are considered accepted in the adapted-lack-location dataset, resulting in 648 (470+178) accepted posts in this dataset. There are additional 10 posts rejected because they have location information but not in the coast (COUNTRYSIDE-GEO and COUNTRYSIDE-TEXT). These posts are considered accepted in the adapted-not-in-coast dataset, in addition to those labeled as NOTHING. Thus, the number of accepted posts in the adapted-not-in-coast dataset is 658.

Table 5. Number of remaining posts rejected because of location per type of spatial information

Spatial info. class	No. Posts
NOTHING	178
COUNTRYSIDE-GEO	9
COUNTRYSIDE-TEXT	1
TOTAL	188

Comparing the original and adapted labels, the number of accepted posts in both adapted datasets is not much bigger than in the original dataset (see Table 3). Although the difference is small, it already affects the quality of the machine learning models obtained, as we report in Sections 5.1 and 5.2.

To evaluate the machine learning models trained in this work, the metrics precision, recall and F1 Score were used. They were chosen because they are indicated to evaluate the resulting models in problems whose data are imbalanced and when there is a need to keep the number of false positives low.

Furthermore, considering that in the future the model can be integrated into an automated ETL process of a database on occurrences of *Physalia physalis* on the Brazilian coast, the ideal is to minimize the number of false positives. This prevents false posts from being registered in the database, even with the risk that legitimate posts are discarded, thus guaranteeing credibility and trust in the data included in the database. This means that precision is an important metric for our study, and it will be considered as the main measure to compare models.

We divided the dataset into 30% (test) and 70% (development). Then, we applied cross-validation across 5 folds, where the development set was divided into 70% (train) and 30% (validation). The result was a test set with 783 samples, a train set with 1279 samples, and a validation set with 548 samples. Furthermore, we used class weighting to handle data imbalance.

The experiments were performed using Python and libraries for NLP and image classification, such as NLTK [2], Scikit-learn [25], TensorFlow [1] and Keras [7].

5.1. Textual Analysis

We trained models with the posts' captions, using original and adapted labels. In the experiments, we trained a classic machine learning model available in the Scikit-learn library [25]: Logistic Regression (LR). We also considered a state-of-the-art deep learning method for NLP tasks: Multilingual BERT (mBERT) [8], which is available in the Tensorflow Hub[13]. It is a pre-trained model based on the weights of the original BERT [9], available only in the *base* size. It was trained in 104 languages, including Portuguese.

[13] https://tfhub.dev/tensorflow/bert_multi_cased_L-12_H-768_A-12/4

mBERT was refined with an output layer adapted to our problem (with a sigmoid function), and a dropout layer (with 10% of probability). In addition, we unfroze the model and retrained it with our data. We used as hyperparameters: 50 epochs with early stopping based on the loss calculated in the validation set, batch size of 16, AdamW optimizer, learning rate of 3e-5 with warmup over the first 10% steps and linear decay, and a sentence length of 128 tokens, since this sentence length is the default value from the mBERT preprocessing model[14] provided by TensorFlow Hub. For other hyperparameters, we used the same ones as used by [9].

For training with LR we kept its default parameters except the parameter `class_weight` which was set to `'balanced'`. This option automatically adjusts the weights inversely proportional to the class frequencies in the input data. It is a strategy to deal with imbalanced data. As vectorization methods, TF-IDF and mBERT without fine-tuning were used. For TF-IDF, we kept the default values of the method. By default, this method converts data to lowercase, and ignores tokens less than 2 characters long, emojis, punctuation, and signs, except underline. In the end, the method normalizes the resulting TF-IDF vectors by the Euclidean norm. For mBERT, we used the pooled-output, the contextual embedding that represents the sentence.

The results obtained for classifying posts according to their captions are presented in Table 6.

Table 6. Results of Experiments with Posts' Caption. Showing: Label, Precision, Recall and F1 Score with their respective standard deviation

Label	TF-IDF + LR			mBERT + LR			mBERT		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
original	0.667 (0.022)	0.831 (0.040)	0.740 (0.021)	0.478 (0.033)	0.760 (0.044)	0.585 (0.015)	0.708 (0.040)	0.740 (0.080)	0.722 (0.044)
adapted-lack-location	0.791 (0.009)	0.899 (0.027)	0.842 (0.015)	0.590 (0.020)	0.800 (0.022)	0.679 (0.019)	0.823 (0.041)	0.886 (0.046)	0.852 (0.014)
adapted-not-in-coast	0.805 (0.024)	0.893 (0.032)	0.846 (0.026)	0.618 (0.018)	0.787 (0.036)	0.692 (0.019)	0.841 (0.052)	0.876 (0.048)	0.856 (0.022)

It is possible to observe that there was an increase in all metrics when we compare the results of the models trained with adapted labels with the results of the models trained with original labels.

Another point is that models trained with the label adapted-lack-location showed a small advantage in the recall, around 1% greater than models trained with the label adapted-not-in-coast. However, models trained with adapted-not-in-coast labels showed a small advantage in precision compared to models trained with adapted-lack-location.

5.2. Image Analysis

We trained CNNs with the posts' images, using original and adapted labels. As the posts can have more than one image, we randomly selected one image per post to perform the

[14] https://tfhub.dev/tensorflow/bert_multi_cased_preprocess/3

training. In particular, for positive posts, we selected one image that met the taxonomic criterion.

We trained the CNNs by transferring the learning from a ResNet50 [16] pre-trained with Imagenet [27] and adapted the last layer to our problem. We chose the same architecture described as the best model in [5].

ResNet50 was adapted by adding: an average pooling layer, a flattening layer, a fully connected layer with ReLu, a dropout regularization layer with a percentage of 30%, and an output layer with one neuron with sigmoid. To increase the amount of training images, we used data augmentation. We have applied random rotation, random zoom, and random horizontal flip of the images. Note that data augmentation is only applied to the development set. All images have been resized to a 224x224 size. This data preparation is actually done by the neural network, through the addition of a preprocessing layer. These preprocesses and hyperparameters were chosen based on the work of [5]. We also used the following hyperparameters: 50 epochs with early stopping based on the loss calculated in the validation set, batch size of 32, and Adam optimizer with a learning rate of 0.001.

The results for classifying posts based on their images are presented in Table 7.

Table 7. Results of Experiments with Posts' Image. Showing: Label, Precision, Recall, and F1 Score with their respective standard deviation

Label	Precision	Recall	F1
original	0.844 (0.049)	0.938 (0.066)	0.889 (0.056)
adapted-lack-location	0.965 (0.024)	0.957 (0.059)	0.961 (0.042)
adapted-not-in-coast	0.967 (0.049)	0.986 (0.020)	0.976 (0.035)

As in the experiments with captions, the experiments with images showed that the adapted labels produced better classification models than the original label, with the adapted-not-in-coast label achieving a 9% F1 Score increase compared to the original label.

5.3. Discussion

The experiments for both text and image show that the adapted labels produced better classification models than the original labels since there was an increase of all metrics when we compare the results of the models trained with adapted labels with the results of the models trained with original labels.

Despite the small increase in the number of ACCEPTED posts when we adapted the labels (18% to 25%), as presented in Table 3, the results with the adapted labels are significantly higher than results with the original labels.

6. Simulating the Production Flow

To find out if the models are really applicable to generate a base on occurrences of *Physalia physalis* in a real scenario. In this section, we trained models using the entire development set and evaluated them on two different test sets: the `full test set` (as described at the beginning of Section 5) and a `partial test set`. Using this partial test set, we intended to simulate the flow in production, in which, before being evaluated by the model, the post is evaluated for location data. When the post meets the spatial criteria, it is evaluated by the model. If not, it is rejected before being evaluated by the model. In practice, we create a partial test set with the posts `accepted` and `rejected because of media`, resulting in 257 posts. Figure 6 illustrates the production scenario.

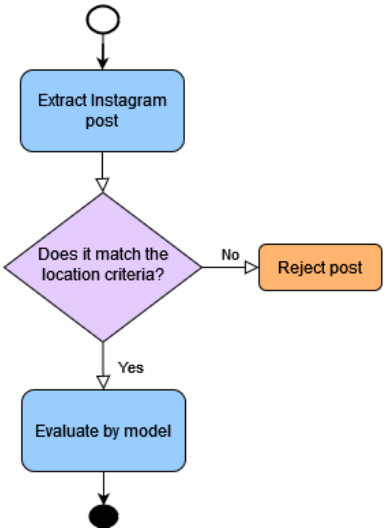


Fig. 6. Illustration of the production scenario. Before being evaluated by the model, the post is evaluated for location data. When the post meets the spatial criteria, the post is evaluated by the model. If not, it is rejected before being evaluated by the model. Source: the Author.

6.1. Textual Analysis

In this part of the experiments, we chose to train only mBERT, since it presented the best precision and F1 Score when trained with adapted labels. Five training rounds with different random states were performed. We also applied hyperparameter optimization, where we obtained the best results with 15 epochs and a sentence length of 344. The results are presented in Table 8.

When we observe the results presented in Table 8, it is possible to notice an increase in all metrics when we evaluate the models by simulating the flow in production.

Error Analysis We compared the performance of the best-performing classifiers. As we executed 5 runs for each mBERT configuration, we present here the results of the models that achieved the best precision among the 5 runs. As we mentioned at the beginning of Section 5, the ideal for our application is to minimize the number of false positives.

Table 8. Results of mBERT trained with entire development set and evaluated with the partial and full test sets. They report the mean performance over 5 runs with the standard deviation

	Full Test Set			Partial Test Set		
Label	Precision	Recall	F1	Precision	Recall	F1
adapted-lack-location	0.868 (0.012)	0.809 (0.051)	0.837 (0.024)	0.925 (0.008)	0.840 (0.050)	0.880 (0.027)
adapted-not-in-coast	0.860 (0.018)	0.858 (0.025)	0.859 (0.004)	0.926 (0.013)	0.879 (0.027)	0.902 (0.015)

In Figure 7 we present the confusion matrices achieved by the mBERT models evaluated with the full test set, while in Figure 8 we present the confusion matrices achieved by the mBERT models evaluated with the partial test set.

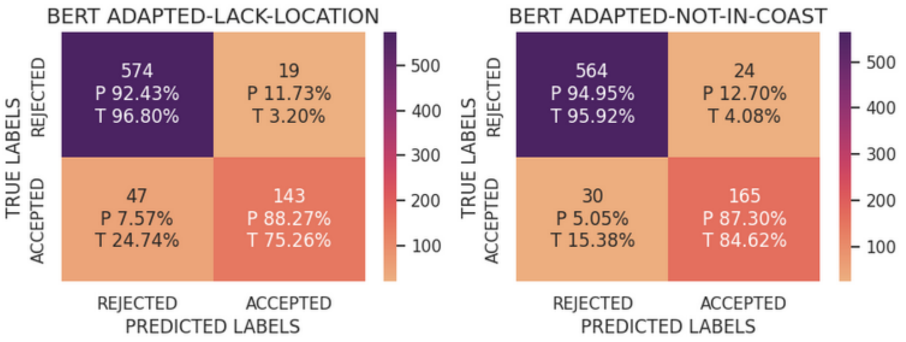


Fig. 7. Confusion Matrices of mBERT models evaluated with the full test set. BERT ADAPTED-LACK-LOCATION shows the confusion matrix of the model trained with adapted-lack-location and achieved the best precision among the 5 runs, this model has an F1 Score of 81%, while BERT ADAPTED-NOT-IN-COAST shows the confusion matrix of the model trained with adapted-not-in-coast and achieved the best precision among the 5 runs, this model has an F1 Score of 86%. Beyond the classifier's errors and successes in absolute numbers, they show normalized numbers over the predicted label (P), and normalized numbers over the true label (T)

It is possible to observe that all models have more difficulty in recognizing positive examples. Note the normalized numbers over the true label (T), the percentage of false negatives is greater than the percentage of false positives.

There is a decrease in false positives in BERT ADAPTED-LACK-LOCATION when evaluated with the full test set. On the other hand, when the models are evaluated with the partial test set, beyond the general increase of all metrics, the model trained with ADAPTED-NOT-IN-COAST presented a decrease of false positive and false negative, and also an increase of precision, resulting in an F1 Score of 91%.

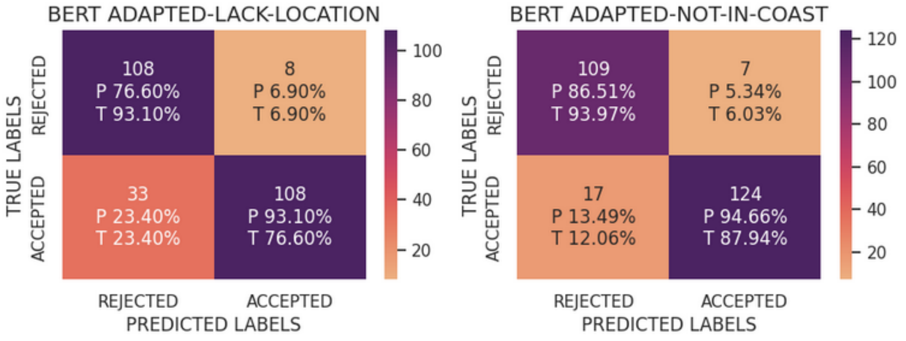


Fig. 8. Confusion Matrices of mBERT models evaluated with the partial test set. BERT ADAPTED-LACK-LOCATION shows the confusion matrix of the model trained with adapted-lack-location and achieved the best precision among the 5 runs, this model has an F1 Score of 84%, while BERT ADAPTED-NOT-IN-COAST shows the confusion matrix of the model trained with adapted-not-in-coast and achieved the best precision among the 5 runs, this model has an F1 Score of 91%. Beyond the classifier's errors and successes in absolute numbers, they show normalized numbers over the predicted label (P), and normalized numbers over the true label (T)

6.2. Image Analysis

As in the experiments with captions, we also trained models using the entire development set and evaluated them on two different test sets: the `full test set` and the `partial test set`. Five training rounds with different random states were performed. The hyperparameters were optimized and 2 techniques were experimented to deal with data imbalance: undersampling and class weighting. We obtained the best results with undersampling, a learning rate of 1e-5, and 512 neurons with ReLu. The results are presented in Table 9.

Table 9. Results of ResNet50 trained with entire development set and evaluated with the partial and full test sets. Showing Label, Precision, Recall and F1 Score. We show the mean performance over 5 runs with the standard deviation

	Full Test Set			Partial Test Set		
Label	Precision	Recall	F1	Precision	Recall	F1
adapted-not-in-coast	0.933 (0.026)	0.895 (0.052)	0.913 (0.019)	0.978 (0.015)	0.926 (0.045)	0.951 (0.020)
adapted-lack-location	0.922 (0.023)	0.866 (0.058)	0.892 (0.021)	0.985 (0.008)	0.896 (0.050)	0.938 (0.025)

Observing the results presented in Table 9, it is possible to notice an increase in all metrics when we evaluate the models by simulating the flow in production.

Error Analysis We compared the performance of the best-performing classifiers. As we performed 5 runs for each configuration, we present here the results of the models that achieved the best precision among the 5 runs. As we mentioned at the beginning of Section 5, the ideal for our application is to minimize the number of false positives.

In Figure 9 we present the confusion matrices achieved by the ResNet50 models evaluated with the full test set, while in Figure 10 we present the confusion matrices achieved by the ResNet50 models evaluated with the partial test set.

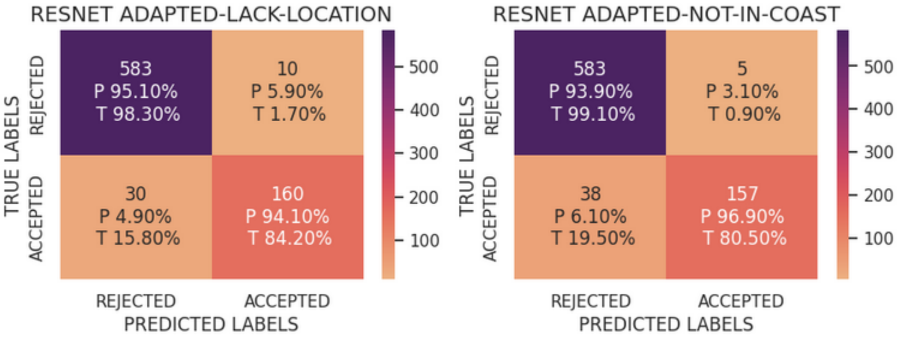


Fig. 9. Confusion Matrices of ResNet50 models evaluated with the full test set. RESNET ADAPTED-LACK-LOCATION shows the confusion matrix of the model trained with adapted-lack-location and achieved the best precision among the 5 runs, this model has an F1 Score of 89%, while RESNET ADAPTED-NOT-IN-COAST shows the confusion matrix of the model trained with adapted-not-in-coast and achieved the best precision among the 5 runs, this model has an F1 Score of 88%. Beyond the classifier's errors and successes in absolute numbers, they show normalized numbers over the predicted label (P), and normalized numbers over the true label (T)

Like in the models trained with captions. It is possible to observe that all models have more difficulty in recognizing positive examples. Note the normalized numbers over the true label (T), the percentage of false negatives is greater than the percentage of false positives.

There is a decrease in false positives in RESNET ADAPTED-NOT-IN-COAST when evaluated with the full test set. On the other hand, when the models are evaluated with the partial test set, beyond the general increase of all metrics, both models presented the same number of false positives and precision. The model trained with ADAPTED-LACK-LOCATION presented a decrease of false negatives and also an increase in recall, resulting in an F1 Score of 94%.

6.3. Discussion

The model trained with text that achieved the best result was a multilingual BERT model trained with the adapted-not-in-coast label, which achieved a precision of 95% and an F1 Score of 91%, when evaluated with the partial test set (see Figure 8). The model trained with images that obtained the best result was the model trained with adapted-lack-location, which achieved a precision of 99% and an F1-Score of 94% when evaluated with the partial test set (see Figure 10).

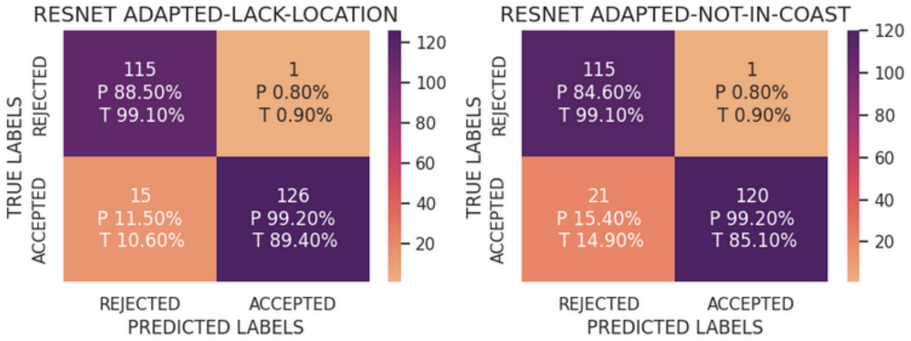


Fig. 10. Confusion Matrices of ResNet50 models evaluated with the partial test set. RESNET ADAPTED-LACK-LOCATION shows the confusion matrix of the model trained with adapted-lack-location and achieved the best precision among the 5 runs, this model has an F1 Score of 94%, while RESNET ADAPTED-NOT-IN-COAST shows the confusion matrix of the model trained with adapted-not-in-coast and achieved the best precision among the 5 runs, this model has an F1 Score of 92%. Beyond the classifier's errors and successes in absolute numbers, they show normalized numbers over the predicted label (P), and normalized numbers over the true label (T)

These experiments also show that between the two best models, the one trained with images outperformed the model trained with text in 4% of precision and 3% of F1 Score. Given the nature of Instagram, which is based primarily on images and video, this is an expected result. However, the good results obtained with text, show that it can also play an important role in the classification process.

Moreover, the results demonstrate that the spatial information had indeed an impact on the results, being actually a source of noise for the problem investigated. It reinforces that adapting the labels was a good choice. Moreover, results presented in this Section show that filtering out posts that do not meet the spatial criteria before evaluating them by the classifier can further improve all metrics.

7. Conclusion

In this article, we explored the problem of identifying legitimate occurrences of *Physalia physalis* in posts extracted from Instagram using a classification task. Classifiers were trained using only the text of the post and only one of the images of the post. We also presented an experience of constructing a dataset on the cnidarian *Physalia physalis* with posts extracted from Instagram. In particular, we showed that even for a binary classification problem, to distinguish between legitimate and false sightings of the species, the design of appropriate labels can affect the quality of the machine learning model trained with the dataset.

Among the models experimented with the text are Logistic Regression and Multilingual BERT. The latter was also experimented as a feature extractor to feed the Logistic Regression model and also retrained with our data and used as a classifier. TF-IDF was also used in conjunction with Logistic Regression.

ResNet50 pre-trained with ImageNet was chosen for image experiments. They were carried out with different approaches to deal with imbalanced data, in addition to retraining the network with our data and optimizing the hyperparameters.

The best models were evaluated on a test set that simulates the flow in production, in which before being evaluated by the model the post is evaluated for location data. It was possible to observe an increase in all metrics, achieving a precision of 95% for text and 99% for images, indicating that these models can be used as part of an automated ETL process of a database on occurrences of *Physalia physalis* on the Brazilian coast from data extracted from Instagram.

One of the contributions of this work is a comparison between the use and performance of different modals (i.e., text and image) as resources for training machine learning models, in the context of conservation science and social media data. To the best of our knowledge, there is no other work using text and image for species recognition in data extracted from social media. Furthermore, there is a small number of works that explore the text for tasks related to conservation science, as already mentioned by [11]. In fact, the use of images and video are more common in this context. Our experiments reinforce the importance of images in these applications, showing that for our target species models trained with images outperformed models trained with text by 4% of precision.

Another contribution of this article was to show the importance of knowledge on the part of the computer scientist of which criteria were used to label the data, since the unquestionable use of the label assigned by the specialist can be a source of noise. This is especially true for works involving species occurrence, where spatial and temporal data are essential but may not be present in all samples. Knowing how the data was labeled can allow the labels to be adapted and therefore produce better results, as demonstrated in this work.

The dataset that was constructed and labeled manually by the oceanographer and the computer scientist, can be extended with new data through the application of the final models. Furthermore, taxonomic and spatial information labels can be used to refine the model itself or used for training new models. Examples of refinements include: recognizing the spatial information contained in the caption and differentiating sightings from accidents using taxonomic information.

As a final contribution, as the best models are neural networks, they can be adapted to new problems, taking advantage of the transfer learning paradigm, which makes it possible to automate the identification of other species.

This work is part of a project that has the goal of monitoring *Physalia physalis* on the Brazilian coast. There are many interesting directions that this work can be extended to contribute to this objective. Among them, we are planning to train multimodal machine learning in order to produce a classifier that takes into consideration both text and images in the same model. We also plan to explore preprocessing techniques for the text and train classification models to automatically extract data from other sources such as newspapers, and other social media platforms.

Other future lines of investigation include: experiments with other language models, such as BERTimbau [28] and XML-RoBERTa[15]; interpretation of decisions made by the models by applying methods such as Feature Visualization and Pixel Attribution [21];

[15] https://huggingface.co/xlm-roberta-large

experiments with new techniques such Vision Transformer (ViT) [10] to image classification.

Acknowledgments. This work was partially supported by CNPq (Process 407644/2021-0), and by CAPES/PrInt-UFPR.

References

1. Abadi, M., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/
2. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media Inc. (2009), https://www.nltk.org/
3. Camargo, L., Rocha, H., Nascimento, L., Hara, C.: Coleta de dados do instagram sobre ocorrências de caravelas-portuguesas na costa brasileira. In: Anais da XVIII Escola Regional de Banco de Dados. pp. 51–59. SBC, Porto Alegre, RS, Brasil (2023)
4. Cardoso, A.S., Bryukhova, S., Renna, F., Reino, L., Xu, C., Xiao, Z., Correia, R., Di Minin, E., Ribeiro, J., Vaz, A.S.: Detecting wildlife trafficking in images from online platforms: A test case using deep learning with pangolin images. Biological Conservation 279, 109905 (2023)
5. Carneiro, A., Nascimento, L.S., Noernberg, M.A., Hara, C.S., Pozo, A.T.R.: Social media image classification for jellyfish monitoring. Aquatic Ecology 58, 3–15 (2024)
6. Cavalcante, M.M.E., Rodrigues, Z.M.R., Hauser-Davis, R.A., Siciliano, S., Haddad Júnior, V., Nunes, J.L.S.: 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 (2020)
7. Chollet, F., et al.: Keras. https://keras.io (2015)
8. Devlin, J.: Multilingual bert. https://github.com/google-research/bert/blob/master/multilingual.md (2019), acessado em 10/03/2022
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: "BERT: Pre-training of deep bidirectional transformers for language understanding". In: Burstein, J., Doran, C., Solorio, T. (eds.) 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://aclanthology.org/N19-1423
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
11. Edwards, T., Jones, C.B., Corcoran, P.: Identifying wildlife observations on twitter. Ecological Informatics 67, 101500 (2022)
12. Edwards, T., Jones, C.B., Perkins, S.E., Corcoran, P.: Passive citizen science: The role of social media in wildlife observations. PLOS ONE 16(8), e0255416 (08 2021)
13. ElQadi, M.M., Lesiv, M., Dyer, A.G., Dorin, A.: Computer vision-enhanced selection of geotagged photos on social network sites for land cover classification. Environmental Modelling & Software 128, 104696 (2020)
14. Foglio, M.: Animal Wildlife Population Estimation Using Social Media Images Collections. Master's thesis, University of Illinois, Chicago, Illinois, USA (2019)
15. Ghermandi, A., Sinclair, M.: Passive crowdsourcing of social media in environmental research: A systematic map. Global Environmental Change 55, 36–47 (2019)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (June 2016)

17. Hunter, S.B., Mathews, F., Weeds, J.: Using hierarchical text classification to investigate the utility of machine learning in automating online analyses of wildlife exploitation. Ecological Informatics 75, 102076 (2023)
18. Kulkarni, R., Di Minin, E.: Automated retrieval of information on threatened species from online sources using machine learning. Methods in Ecology and Evolution 12(7), 1226–1239 (2021)
19. Leitão, A.T.T.S., de O Alves, M.D., dos Santos, J.C.P., Bezerra, B.: Instagram as a data source for sea turtle surveys in shipwrecks in brazil. Animal Conservation 25(6), 736–747 (2022)
20. Mazars-Simon, A.E.: The Wild in Live Project: A Human/Algorithm learning network to help citizen science in wildlife conservation. Master's thesis, Universidade de Coimbra (2019)
21. Molnar, C.: Interpretable Machine Learning. Independently published, 2 edn. (2022), https://christophm.github.io/interpretable-ml-book
22. Morais, P., Afonso, L., Dias, E.: Harnessing the Power of Social Media to Obtain Biodiversity Data About Cetaceans in a Poorly Monitored Area. Frontiers in Marine Science 8 (2021)
23. do Nascimento, L.S.: Monitoring jellyfish population by social media. Tech. rep., Universidade Federal do Paraná (2020), technical report, Pós-Graduação em Sistemas Costeiros e Oceânicos
24. Nascimento, L.S., Hara, C.S., Jr., M.N., Noernberg, M.: Instagram como fonte de dados alternativa no monitoramento da #caravelaportuguesa (physalia phisalis, cnidaria). In: Livro de Memórias do IV SUSTENTARE e VII WIPIS: Workshop internancional de Sustentabilidade, Indicadores e Gestão de Recursos Hídricos (2022)
25. Pedregosa, F., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
26. Rocha, H.F., Nascimento, L.S., Camargo, L., Noernberg, M., Hara, C.S.: Labeling portuguese man-of-war posts collected from instagram. In: Abelló, A., Vassiliadis, P., Romero, O., Wrembel, R., Bugiotti, F., Gamper, J., Vargas Solar, G., Zumpano, E. (eds.) New Trends in Database and Information Systems. pp. 369–381. Springer Nature Switzerland, Cham (2023)
27. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision 115, 211–252 (2015)
28. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Cerri, R., Prati, R.C. (eds.) Intelligent Systems. pp. 403–417. Springer International Publishing, Cham (2020)
29. Sullivan, M., Robinson, S., Littnan, C.: Social media as a data resource for #monkseal conservation. PLoS ONE 14(10) (2019)

Heloisa Fernanda Rocha is a Software Engineer with over 15 years of experience. She holds a master's degree in Informatics from the Federal University of Paraná in 2023. She received a postgraduate degree in Database Administration from Ponta Grossa State University in 2004 and Software Engineering from the Pontifical Catholic University of Paraná in 2015.

Leonardo S. Camargo is a Major in Science Computer graduation student at Universidade Federal do Parana (UFPR) in Brazil and an intern at Telefónica Brasil S.A. He has researched social media data mining and presented the article ”Coleta de Dados do Instagram sobre Ocorrências de Caravelas-Portuguesas na Costa Brasileira” at Encontro Regional de Banco de Dados (ERBD,2022). His interests reside in Data Visualization, Data Mining and Data Science.

Maurício Almeida Noernberg recived the PhD degree in environmental geology from the Federal University of Paraná in 2001. He is currently a Full Professor at Center for

Marine Studies and the Head of Coastal Oceanography and Remote Sensing Lab. at the Federal University of Paraná. His research interests include land-sea interactions, coastal circulation, environmental monitoring, numerical modeling, remote sensing and image processing.

Aurora Pozo is currently a full Professor of Computer Science at Universidade Federal do Paraná, where she heads the interdisciplinary Evolutionary Computation Research Group. Her scientific contributions are in the area of computational intelligence, mainly focused on machine learning, and particularly in evolutionary computation and learning (using genetic programming, particle swarm optimization, and learning classifier systems), multi-objective optimization, and evolutionary deep learning. In addition to technical contributions in the area of computational intelligence, Aurora works to consolidate the area in Brazil, actively participating in events related to the area (BRACIS, ENIAC, CTDIAC). The bidder has coordinated CNPq Universal and Fundação Araucária projects in the area of bioinspired computing and applications. The projects have the collaboration of professors from different National Universities. Internationally, she coordinated projects which allowed fruitful cooperation with the University of the Vasco Country/Spain. In addition to these activities, during this period, she has been dedicated to the qualification of human resources, providing ongoing supervision to both master's and doctoral candidates. Furthermore, she has contributed to numerous high-quality articles published in esteemed journals and conferences, cementing her presence in the academic and research spheres.

Carmem S. Hara is a Full Professor at Universidade Federal do Parana (UFPR) in Brazil, head of the Information and Data Management Group (GDAI), and member of the Wireless and Advanced Networks Group (NR2). Her main research interests include management of complex and graph data models. She holds a PhD degree in Computer and Information Sciences from the University of Pennsylvania, USA (2014). From 2015 to 2019, she served as a member of the Database Special Council of the Brazilian Computer Society (CEBD-SBC), and is currently serving the Council as its senior member (2022-2024). In 2022, she received a "Distinguished Researcher Award" from the Brazilian Database Symposium. In the graduate program in Informatics at UFPR, she has supervised 4 PhD students, and 20 master's thesis. Her current projects involves Data Management for Environmental Monitoring and Modelling (REsMA-PrInt - Funding: CAPES) and Intelligent Systems for Marine Species Monitoring (Funding: CNPq).

Lorena Silva Nascimento is an oceanographer with a PhD in Biology and Ecology of Coastal and Oceanic Systems (Federal University of Paraná – UFPR, Brazil). She has expertise in the taxonomy and ecology of marine zooplankton, with an emphasis on copepods and gelatinous organisms. Her main interest involves the application of citizen science and iEcology in the marine monitoring and analysis of species distribution. She is also enthusiastic about scientific communication, working especially with social media use in environmental education. From February 2022 to February 2024, she served as the journal Aquatic Ecology editor. Currently, Lorena is a postdoctoral fellow in Zoology (Federal University of Paraíba – UFPB, Brazil) and works as associate editor of the journal Ocean and Coastal Research.

Received: March 01, 2024; Accepted: August 06, 2024.