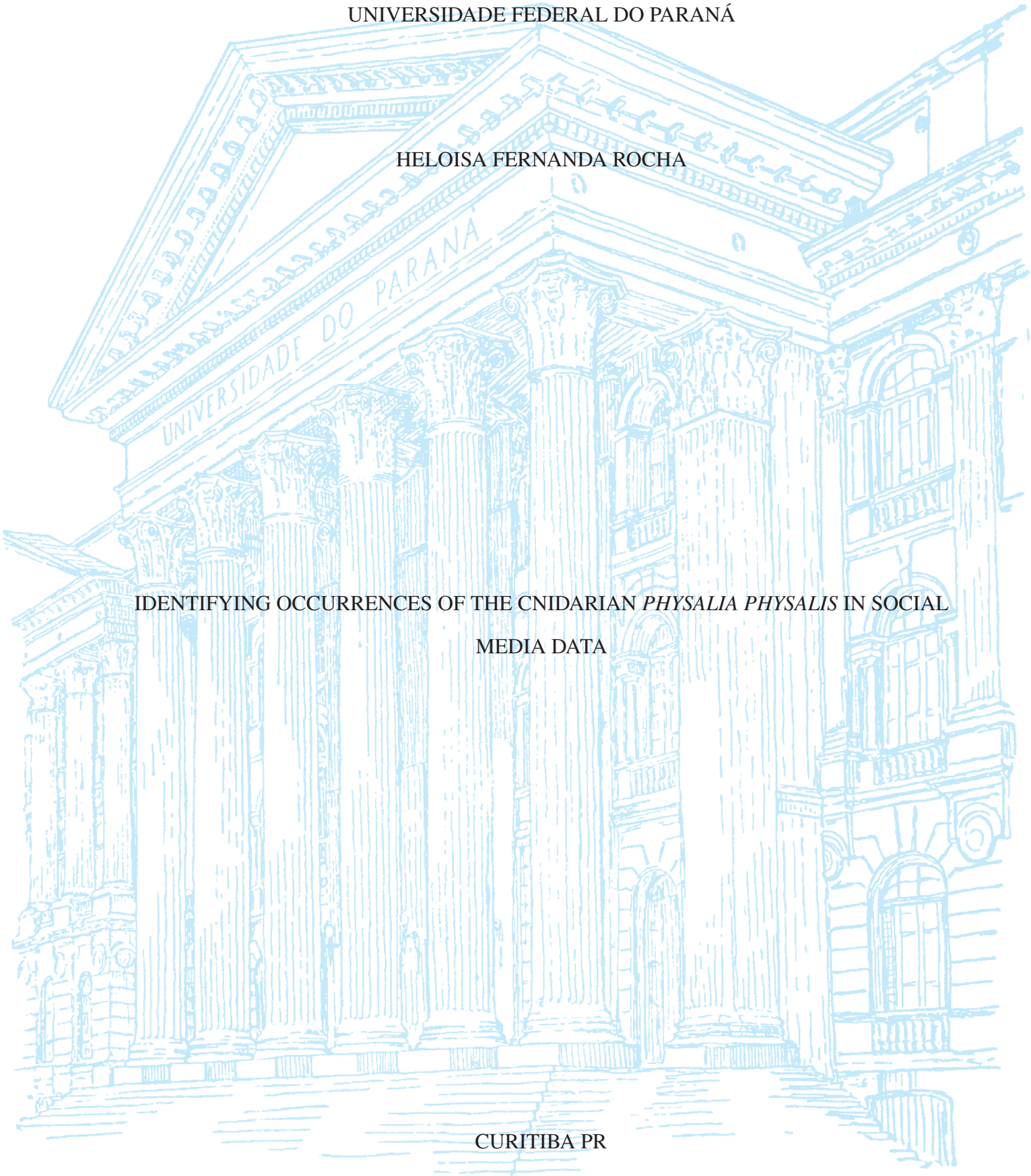
UNIVERSIDADE FEDERAL DO PARANÁ

HELOISA FERNANDA ROCHA



IDENTIFYING OCCURRENCES OF THE CNIDARIAN *PHYSALIA PHYSALIS* IN SOCIAL MEDIA DATA

CURITIBA PR

2023

HELOISA FERNANDA ROCHA

IDENTIFYING OCCURRENCES OF THE CNIDARIAN *PHYSALIA PHYSALIS* IN SOCIAL MEDIA DATA

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Prof. Dra. Carmem Satie Hara.

CURITIBA PR

2023

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIA E TECNOLOGIA

Rocha, Heloisa Fernanda
Identifying occurrences of the cnidarian Physalia physalis in social media data / Heloisa Fernanda Rocha. – Curitiba, 2023.
1 recurso on-line : PDF.

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática.

Orientador: Carmem Satie Hara

1. Processamento de linguagem natural (Computação). 2. Interfaces de usuário multimodal (Sistemas de computação). 3. Mídia social. I. Universidade Federal do Paraná. II. Programa de Pós-Graduação em Informática. III. Hara, Carmem Satie. IV. Título.

Bibliotecário: Elias Barbosa da Silva CRB-9/1894

ATA Nº20

ATA DE SESSÃO PÚBLICA DE DEFESA DE MESTRADO PARA A OBTENÇÃO DO GRAU DE MESTRA EM INFORMÁTICA

No dia vinte e tres de outubro de dois mil e vinte e tres às 10:30 horas, na sala SALA DE VIDEOCONFERÊNCIA, DEPARTAMENTO DE INFORMÁTICA, foram instaladas as atividades pertinentes ao rito de defesa de dissertação da mestranda **HELOISA FERNANDA ROCHA**, intitulada**: Identifying Occurrences of the Cnidarian Physalia Physalis in Social Media Data**, sob orientação da Profa. Dra. CARMEM SATIE HARA. A Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação INFORMÁTICA da Universidade Federal do Paraná, foi constituída pelos seguintes Membros: CARMEM SATIE HARA (UNIVERSIDADE FEDERAL DO PARANÁ), CARLOS AUGUSTO PROLO (UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE), GENOVEVA VARGAS-SOLAR (CNRS, LIRIS, DATABASE GROUP), AURORA TRINIDAD RAMIREZ POZO (UNIVERSIDADE FEDERAL DO PARANÁ). A presidência iniciou os ritos definidos pelo Colegiado do Programa e, após exarados os pareceres dos membros do comitê examinador e da respectiva contra argumentação, ocorreu a leitura do parecer final da banca examinadora, que decidiu pela APROVAÇÃO. Este resultado deverá ser homologado pelo Colegiado do programa, mediante o atendimento de todas as indicações e correções solicitadas pela banca dentro dos prazos regimentais definidos pelo programa. A outorga de título de mestra está condicionada ao atendimento de todos os requisitos e prazos determinados no regimento do Programa de Pós-Graduação. Nada mais havendo a tratar a presidência deu por encerrada a sessão, da qual eu, CARMEM SATIE HARA, lavrei a presente ata, que vai assinada por mim e pelos demais membros da Comissão Examinadora.

CURITIBA, 23 de Outubro de 2023.

Assinatura Eletrônica
24/10/2023 14:33:34.0
CARMEM SATIE HARA
Presidente da Banca Examinadora

Assinatura Eletrônica
24/10/2023 14:11:52.0
CARLOS AUGUSTO PROLO
Avaliador Externo (UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE)

Assinatura Eletrônica
27/10/2023 16:12:16.0
GENOVEVA VARGAS-SOLAR
Avaliador Externo (CNRS, LIRIS, DATABASE GROUP)

Assinatura Eletrônica
24/10/2023 13:10:14.0
AURORA TRINIDAD RAMIREZ POZO
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **HELOISA FERNANDA ROCHA** intitulada: **Identifying Occurrences of the Cnidarian Physalia Physalis in Social Media Data**, sob orientação da Profa. Dra. CARMEM SATIE HARA, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestra está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 23 de Outubro de 2023.

Assinatura Eletrônica
24/10/2023 14:33:34.0
CARMEM SATIE HARA
Presidente da Banca Examinadora

Assinatura Eletrônica
24/10/2023 14:11:52.0
CARLOS AUGUSTO PROLO
Avaliador Externo (UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE)

Assinatura Eletrônica
27/10/2023 16:12:16.0
GENOVEVA VARGAS-SOLAR
Avaliador Externo (CNRS, LIRIS, DATABASE GROUP)

Assinatura Eletrônica
24/10/2023 13:10:14.0
AURORA TRINIDAD RAMIREZ POZO
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

RESUMO

As necessidades de conhecimento da biodiversidade são constantes, enquanto recursos para pesquisa, sejam financeiros, de tempo e humanos são escassos. Por outro lado, a Internet oferece um enorme volume de dados que podem ser explorados em favor da ciência da conservação. As caravelas-portuguesas (*Physalia physalis*) oferecem risco à população, e dados sobre sua ocorrência nem sempre estão disponíveis para estudo da espécie. Este trabalho utiliza processamento de linguagem natural e visão computacional como técnicas para treinamento de modelos de aprendizagem de máquina como ferramentas para classificar dados extraídos de mídias sociais. Tais modelos podem ser utilizados como parte de um processo automatizado de *Extract-Transform-Load* para a criação de uma base de dados de ocorrências de *Physalia physalis* na costa Brasileira a partir de dados extraídos de mídias sociais. Como preparação para o treinamento dos modelos de aprendizagem de máquina foram coletadas e rotuladas postagens extraídas do Instagram como sendo: `aceita` ou `rejeitada` como ocorrência legítima de *Physalia physalis* na costa Brasileira, seguindo critérios de uma oceanógrafa. Entre os modelos treinados com a legenda das postagens estão a Regressão Logística e o BERT multilíngue. O BERT também foi experimentado como extrator de características para alimentar o modelo de Regressão Logística e também retreinado com nossos dados e usado como classificador. TF-IDF também foi usado em conjunto com Regressão Logística. Foram experimentadas diferentes técnicas de normalização de texto, além da otimização dos hiperparâmetros desses modelos. ResNet50 pré-treinada com ImageNet foi escolhida para experimentos com as imagens das postagens. Foram realizados experimentos com a CNN retrainada com os nossos dados e com diferentes abordagens para lidar com dados desbalanceados. Também combinamos os resultados dos modelos individuais usando diferentes regras de fusão e, usando produto, alcançamos a precisão de 94% e F1 Score de 89%. Durante o processo de anotação observamos que postagens rejeitadas pela oceanógrafa por causa de informações espaciais poderiam representar ruído para treinamento de modelos de aprendizado de máquina. Devido a isso, decidimos adaptar os rótulos, considerando como aceitas as postagens rejeitadas por causa de informações espaciais. Com os rótulos adaptados, alcançamos um aumento de 13% no F1 Score com BERT e 9% no F1 Score com ResNet50.

Palavras-chave: processamento de linguagem natural; visão computacional; multimodal; mídias sociais.

ABSTRACT

Biodiversity knowledge needs are constant, while financial, time and human resources for research are scarce. On the other hand, the Internet offers a huge amount of data that can be exploited in conservation science. The Portuguese man-of-war (*Physalia physalis*) is a risk to the population, and data about its occurrences are not always available for researchers. This work use natural language processing and computer vision as techniques for machine learning models' training as a tool for classification of data extracted from social media. 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 from data extracted from social media. In preparation for training machine learning models we collected and labeled posts extracted from Instagram as being: `accepted` or `rejected` as legitimate occurrences of *Physalia physalis* on the Brazilian coast, following the criteria established by an oceanographer. Among the trained models are Logistic Regression and Multilingual BERT. BERT was also used 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. Different text normalization techniques were experimented, in addition to hyperparameters optimization of these models. ResNet50 pre-trained with ImageNet was chosen for image experiments. We experimented different approaches to deal with imbalanced data and tried retrain the CNN with our data. We also combined the results of the individual models using different fusion rules, and using product, we achieved the precision of 94% and F1 Score of 89%. During the annotation process we observed that posts rejected by the oceanographer because of spatial information could represent noise for training machine learning models. Due to this, we decided to adapt the labels, considering as accepted posts rejected because spatial information. With adapted labels, we achieved a increase of 13% of F1 Score with BERT and 9% of F1 Score with ResNet50.

Keywords: natural language processing; computer vision; multimodal; social media.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

BERT	Bidirectional Encoder Representations from Transformers
Bi-GRU	Bidirectional Gated Recurrent Units
Bi-LSTM	Bidirectional Long-Term Short-Memory
BoW	Bag-of-Words
CNN	Convolutional Neural Network
DL	Deep Learning
ETL	Extract-Transform-Load
FFNN	Feedforward Neural Network
FN	False Negative
FP	False Positive
GRU	Gated Recurrent Units
GB	Gradient Boosting
KNN	K-Nearest Neighbors
LR	Logistic Regression
LSTM	Long-Term Short-Memory
mBERT	Multilingual BERT
ML	Machine Learning
MLP	Multilayer Perceptron
MMML	Multimodal Machine Learning
NB	Naive Bayes
NLP	Natural Language Processing
ReLU	Rectified Linear Unit
RF	Random Forest
RNN	Recurrent Neural Network
SOTA	State-of-the-art
SINAN	*Sistema de Informação de Agravos de Notificação*
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machines
TF-IDF	Term Frequency–Inverse Document Frequency
TN	True Negative
TP	True Positive
WE	Word Embeddings

CONTENTS

1 INTRODUCTION

The Portuguese man-of-war (*Physalia physalis*) is a multicellular organism whose tentacles have stinging cells (cnidocytes), which contain filaments that release toxins. Among the reactions caused by these toxins are: very strong pain, burns, dyspnea, malaise, nausea, vomiting, headache, chills, drowsiness, arterial hypotension and cardiac arrhythmias (Cavalcante et al., 2020). This species occurs along the entire Brazilian coast and accidents with these animals have been reported frequently (Cavalcante et al., 2020). Figure 1.1 shows an example of the species.

Figure 1.1: Portuguese man-of-war *Physalia physalis*. Source: https://www.inaturalist.org/photos/180339092

In Brazil, poisonings caused by venomous animals must be registered in the *Sistema de Informação de Agravos de Notificação* (SINAN)[1]. However, SINAN does not have a specific category of envenoming by the Portuguese man-of-war. Accidents involving these animals fall into the "other venomous animals" category. Furthermore, according to Cavalcante et al. (2020), there is evidence of under-reporting of these cases. That is, the limited information about these animals makes it difficult to use official public data in the work of researchers interested in the species.

An alternative to obtain more data about the Portuguese man-of-war is the collection carried out by professionals. However, traditional collection methods consume a lot of time and resources, and often have low coverage (Edwards et al., 2021, 2022; Goodwin et al., 2022).

Another possibility would be the use of citizen science programs, that is, a partnership between scientists and the general public to collect data for scientific research, carried out through campaigns and learning guides. However, organizing public science campaigns requires effort to recruit, train and encourage volunteers (Di Minin et al., 2015; Edwards et al., 2021, 2022).

Finally, an alternative option is the use of involuntary data from social media. This approach is called by some authors passive citizen science or crowdsourcing. It refers to the use of data generated by non-professionals, collected and shared on the Internet, mainly in social media, and which are not connected to any specific citizen science program (Ghermandi and Sinclair, 2019; Jarić et al., 2020; Edwards et al., 2021).

While classical biodiversity research is irreplaceable for understanding the natural world through observations and experiments, the significant amount of data available on the Internet and its potential for use in conservation science has been explored, recognized and encouraged by

[1]http://portalsinan.saude.gov.br/acidente-por-animais-peconhentos

several authors, such as: Di Minin et al. (2015), Daume (2016), Ghermandi and Sinclair (2019), Toivonen et al. (2019), August et al. (2020), Jarić et al. (2020) e Edwards et al. (2021). Just as concluded Daume (2016) in his work **Mining twitter[2] to monitor invasive alien species—an analytical framework and sample information topologies**:

> "In summary, the results confirm that Twitter is a rich source of general biodiversity observations, and particularly opportunistic observations, gathered via descriptive keywords and contributed without knowledge of the species, hold great potential in supplementing biodiversity monitoring, even if this will primarily apply to notable or easily recognizable species".

Among the advantages resulting from the use of social media data are:

- The existence of **abundant and almost always public content**. In addition to images, videos and text, metadata, such as geolocation and timestamp, may be accessible on social networks for free (Ghermandi and Sinclair, 2019; Edwards et al., 2021, 2022). The volume of data available on the Internet is so large that it exceeds the amount that biologists could collect in the traditional way, according to Ghermandi and Sinclair (2019). Also, according to Foglio (2019) the abundance of social media data can lead to better estimates than those obtained by traditional data collection techniques. To give an idea, Instagram has more than 2 billion monthly users and, in every second, 1,074 images are added to the platform[3].

- The approach is **less laborious, less time consuming and has lower cost**. This is especially the case if the extraction is automated, compared to traditional methods or even citizen science campaigns (Ghermandi and Sinclair, 2019; Jarić et al., 2020; Edwards et al., 2021). The human effort and financial impact for other forms of monitoring are high, especially for large-scale collections and for a longer period of time, according to Edwards et al. (2021). According to Di Minin et al. (2015), in contexts where resources for field collection are scarce, the use of social media can reduce the use of resources and direct data collection by professionals to areas that are less known or difficult to access.

- **Data are available almost in real time and continuously**. Using social media as a data source gives the possibility to carry out monitoring and analysis frequently and almost always in real time (Ghermandi and Sinclair, 2019; Edwards et al., 2021, 2022).

- **Allows the acquisition of data with wide geographic scope** (Morais et al., 2021). This is particularly interesting for research involving a large territorial extension such as the Brazilian coast. According to Ghermandi and Sinclair (2019), social media data are relatively easy to apply on a large scale, such as entire populations, ecosystems or biomes.

- **The use of social media as a data source allows new discoveries**. As examples, the work of Santamaria et al. (2020) discovered a new species through a photo shared on Twitter. The work of Nascimento et al. (2022) detected, through posts on Instagram, a predatory behavior of a ghost crab that consumed a Portuguese man-of-war stranded on the beach.

[2]twitter is a social network and microblogging service. In July 2023, Twitter was renamed X, however in this dissertation we will keep the original name: Twitter.

[3]https://www.omnicoreagency.com/instagram-statistics/

The use of social media as a data source does not only have advantages. Some authors point out drawbacks of this approach, such as: quality of the data, reliability, ownership, and future availability of these data (Di Minin et al., 2015; Daume, 2016; Ghermandi and Sinclair, 2019; Edwards et al., 2021, 2022).

Furthermore, traditional approaches to data analysis can become inefficient due to the large volume, diversity and heterogeneity of the data, as argued by Christin et al. (2019), August et al. (2020) and Goodwin et al. (2022) in their works. One way to take advantage of the potential of social media data and transform them into useful information is the use of machine learning techniques, due to its ability to process and analyze large and diverse volumes of data.

Jarić et al. (2020) especially reinforces the use of information technology tools as support for extracting knowledge from Internet data in favor of ecology. He creates a new discipline, called iEcology (i.e., internet ecology), which is defined as a research approach that seeks to quantify patterns and processes in the natural world using data accumulated in digital sources collected for other purposes. Its focus is on the collection, grouping and exploitation of this data.

Other authors, such as Di Minin et al. (2019) and Toivonen et al. (2019), also encourage the use of machine learning for analyzing social media data.

In a project in development at UFPR (Nascimento, 2020), approved by the ethics committee, one of the specific goals is to compile data extracted from social media about occurrences of *Physalia physalis* on the Brazilian coast. In the case of Instagram, searching for content using hashtags is allowed. However, it is necessary to evaluate this content in order to verify whether or not it is a legitimate occurrence of *Physalia physalis* from others, such as tattoos and Portuguese ships. Such verification has been carried out manually.

A previous work related to the project (Carneiro et al., 2022) has already explored the classification of Instagram images as being or not from *Physalia physalis*. However, the caption was not considered in their experiments. During the dataset construction process, it was noted the existence of posts that mention common names and images of the species, but are not direct observations of wildlife. As an example, some posts warn about the risk that *Physalia physalis* poses to bathers, which does not represent a direct observation of the species. This type of situation was also reported by Edwards et al. (2022) in his work. Furthermore, the existence of a *Physalia physalis* image in the post does not necessarily indicate a legitimate occurrence of the species. For this reason, we understand that in addition to the image, it is necessary to analyze the caption of the post.

During the process of annotating the collected data, the oceanographer observed some criteria, which included the requirement, for a post to be accepted as a legitimate occurrence, that the location of the post be on the Brazilian coast. However, for the computer scientist, the lack of 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 machine learning models. In this context, experiments with labels adapted for training machine learning models were carried out resulting in better classifiers.

Text classification using social media data can be challenging due to some characteristics like: short text, presence of misspelling, slang, ambiguity, polysemous words and internet jargon (Dos Santos and Ladeira, 2014; Stiilpen Junior and Merschmann, 2016; de Oliveira and Merschmann, 2021; Edwards et al., 2022). In particular, in Instagram, there are posts that are formed almost exclusively by hashtags and/or emojis. The mixture of languages also appears frequently in texts extracted from this social network. In addition, the use of compound hashtags generates words that do not exist in the lexicon.

Furthermore, Instagram users express themselves using combinations of visual and textual content. To account for this characteristic, it is necessary to build models that can process

and relate information from multiple modalities (i.e., text and image), which is called multimodal machine learning (Baltrušaitis et al., 2019; Toivonen et al., 2019).

The combined use of image and text can improve the performance of data classification tasks from social media, as relevant information can be distributed between visual and textual content (Baltrušaitis et al., 2019; Guo et al., 2019; Toivonen et al., 2019; Edwards et al., 2022). Considering the problem that our work deals with, the caption can contain detailed information about the occurrence of *Physalia physalis*, while the image can be used to recognize it.

Although the use of machine learning in the area of ecology is not new, the combination of (involuntary) social media data with machine learning in this area is relatively scarce (Ghermandi and Sinclair, 2019; Edwards et al., 2022) and has been the subject of some works such as:

- Mazars-Simon (2019) in his master's dissertation he developed a system that identifies sea turtles in Flickr[4] images. To built the system, the author trained a Convolution Neural Network (CNN) to classify images into six categories;
- Kulkarni and Di Minin (2021) used a neural network, Natural Language Processing (NLP) techniques and texts from Google News and Twitter as data sources to capture information about endangered species;
- Edwards et al. (2022) performed experiments using: NLP techniques, some classic machine learning algorithms and also a pre-trained transformer-based model, in order to identify wildlife observations on Twitter;

Therefore, we understand that the use of NLP and computer vision techniques seems to be effective solutions for the task of classifying posts. Currently, transformer-based models have achieved great success in many NLP tasks, including text classification (Kalyan et al., 2021). While CNN-based architectures are ubiquitous in the field of computer vision (Zhang et al., 2021). As a main differential compared to works that propose to identify wildlife in social media data, we train models with the text of the post and also with one of the images of the post. We also combine the results from individual classifiers, and we carry out a comparative analysis of the results obtained.

1.1 MOTIVATION

The need for knowledge of biodiversity is constant, while resources for research, whether financial, time or human, are scarce. On the other hand, the Internet offers an enormous volume of data that can be exploited in favor of science.

The Portuguese man-of-war poses a risk to the population, and data on their occurrence are not always available for research of the species.

Thus, the main motivation for this research are the advantages of using data from social media with the support of machine learning techniques. Such techniques will allow the training of a model that classifies posts identifying legitimate occurrences of *Physalia physalis*. Such a classifier can be inserted into a search, classification and data storage system, thus allowing continuous monitoring of the species.

[4]Flickr is an online photo management and sharing application. https://www.flickr.com/

1.2 OBJECTIVES

The main objective of this dissertation is to obtain one or more models that are able to identify in posts extracted from Instagram occurrences of *Physalia physalis*.

In this way, the following specific objectives are proposed:

- Analysis and preparation of the dataset for training machine learning models;
- Train binary classifiers for identify occurrences of *Physalia physalis* considering only the caption of the posts;
- Train binary classifiers for identify occurrences of *Physalia physalis* considering only the images of the posts. This activity is different from the work developed by Carneiro et al. (2022), which aimed to classify images as being or not from *Physalia physalis*;
- Combine the results of text-only and image-only models to obtain a model that considers the caption and images of the posts for identify occurrences of *Physalia physalis*;
- Perform a comparative analysis of the trained models in order to answer the following questions:
  - Which model performed best for classifying posts as a legitimate occurrence of *Physalia physalis*?
  - By analyzing only the text of the post, it is possible to identify occurrences of *Physalia physalis* with good precision?
  - Which text-only model had the best performance for identifying legitimate occurrences of *Physalia physalis*?
  - Is a text-and-image (multimodal) model better at recognizing legitimate occurrences of *Physalia physalis* than text-only or image-only trained models?

1.3 CONTRIBUTIONS

As the main contribution of this work it is expected to obtain a model that can be used as part of an automated Extract-Transform-Load (ETL) process of a database on occurrences of *Physalia physalis* on the Brazilian coast from data extracted from social media.

It is also expected that the approach developed in this work can be later applied to automate the identification of other species.

Another contribution of this work is a comparison between the use and performance of different modals as a resource for training machine learning models, in the context of conservation science and social media data.

The dataset that was constructed and labeled manually, and which can be extended with new data through the application of the models obtained, is also a contribution of this work.

1.4 RESEARCH ORGANIZATION

This work is organized as follows: Chapter 2 presents the necessary theoretical basis for the dissertation. Concepts related to machine learning, classic algorithms, neural networks, combination of classifiers, natural language processing, image classification and multimodal machine learning are presented.

In Chapter 3, the works related to this dissertation are presented: works that used machine learning models and involuntary data for a wildlife classification task. Moreover, it presents works unrelated to conservation science that performed classification tasks and used natural language processing in Brazilian Portuguese texts with informal characteristics or multimodal machine learning.

Chapter 4, presents how the data used in the development of the dissertation was collected and labeled and what are the challenges of the problem that this dissertation deals with. This chapter also presents an exploratory analysis of the data.

We conducted several experiments with the goal of obtaining a model capable to determine if the post is a legitimate occurrence of *Physalia physalis* in the Chapters 5 (textual analysis), 6 (image analysis) and 7 (multimodal and comparative analysis). Chapter 8 concludes the dissertation and presents some future works.

2 REVIEW OF LITERATURE

In this Chapter, the necessary theoretical basis for the dissertation is presented. Here the concepts related to Machine Learning (Section 2.1), Classic Algorithms (Section 2.2), Neural Networks (Section 2.3), Combining Machine Learning (Section 2.4), Natural Language Processing (Section 2.5), Image Classification (Section 2.6) and Multimodal Machine Learning (Section 2.7) are presented.

2.1 MACHINE LEARNING

The term Machine Learning (ML) was coined by Samuel (1959). ML is a field of computer science and can be understood as the ability of the computer to learn without being explicitly programmed. ML algorithms are exposed to examples and, from them, they build a model capable of finding patterns or making predictions (Goodfellow et al., 2016; Zheng and Casari, 2018; Mohri et al., 2018; Burkov, 2019; Leskovec et al., 2020; Carvalho et al., 2021).

Training a machine learning model should be considered as an approach to solve a problem when it is not possible to write a deterministic algorithm to solve it (Goodfellow et al., 2016).

2.1.1 Tasks

There are several types of tasks that a machine learning algorithm can perform. Some examples of tasks are:

- Classification: it is the task of automatically assigning a label to an unlabeled example (Goodfellow et al., 2016; Burkov, 2019; Mohri et al., 2018). A machine learning model that performs classification tasks can be called a classifier.
- Regression: it is the task of predicting a real value for an unlabeled example (Goodfellow et al., 2016; Burkov, 2019; Mohri et al., 2018).
- Clustering: it is the task of partitioning a dataset into homogeneous subsets (Mohri et al., 2018).

Our work focuses is on the classification task, as it seeks to identify legitimate posts about occurrences of *Physalia physalis* in Instagram posts.

2.1.2 Learning Scenarios

Some of the machine learning scenarios (or types) are:

- Supervised learning: in this scenario the algorithm receives labeled training data and performs predictions for unseen data. This is the scenario most commonly associated with classification and regression tasks (Goodfellow et al., 2016; Burkov, 2019; Mohri et al., 2018).
- Unsupervised learning: In this scenario, the algorithm receives unlabeled training data and performs predictions on unseen data. Clustering is an example of an unsupervised learning task (Goodfellow et al., 2016; Burkov, 2019; Mohri et al., 2018).

Although there are other possible scenarios, our work focuses on supervised learning, since labeled data were used for training machine learning models.

2.1.3 Workflow

A typical supervised machine learning workflow consists of the following steps:

- **Dataset Construction**: in this step, the data that will be used to build the model is collected and labelled.
- **Preprocessing**: The collected data cannot be fed to the machine learning model in its raw state. For making then adequate, some treatments and transformations are necessary, such as (Zafarani et al., 2014):
  - **Cleaning and Preparation**: Cleaning and preparation techniques directly depend on the types of data collected. This dissertation uses text and image. Cleaning and preparation techniques for text are described in Section 2.5, while image-related preprocesses are described in Section 2.6.
  - **Feature Extraction** or **Vectorization**: For a machine learning model to be able to process data, they need to be represented in an appropriate format. More precisely, they have to be represented as a vector, which can also be called representation or feature vector (Zheng and Casari, 2018; Mohri et al., 2018). The text feature extraction methods are described in Subsection 2.5.2. For images, the vectorization is discussed in Section 2.6.
  - **Holdout**: Data is shuffled and partitioned in training, validation and testing (Burkov, 2019; Mohri et al., 2018).
- **Training**: In this step, the algorithm is exposed to the training data, and from them it builds a model capable of predicting labels for data not yet seen. In other words, considering a classification task in which the training set is formed by tuples in the format (x, y), where x is the feature vector and y is the label, it is in this step that, from these data, the machine learning algorithm finds a model that maps x to y, which can be represented by the function $m(.)$ such that $m(x) = y$. Once the model is learned, we can calculate $m(x)$, whose result is the prediction of the label for the unlabeled instance, that is,y (Zafarani et al., 2014; Mohri et al., 2018; Goodwin et al., 2022).
- **Validation**: In this step, validation data is used in two ways. First, to verify the performance of the model, that is, to verify if the learned model is able to assign the labels correctly to data not yet seen. Here, the metrics discussed in Subsection 2.1.6 are applied. Second, to select suitable hyperparameters, the hyperparameters will be discussed in Subsection 2.1.7.
- **Test**: in this step the test data is used to evaluate the performance of the final model. Here the metrics discussed in Subsection 2.1.6 are applied (Mohri et al., 2018; Goodwin et al., 2022).

Figure 2.1 illustrates the machine learning workflow.

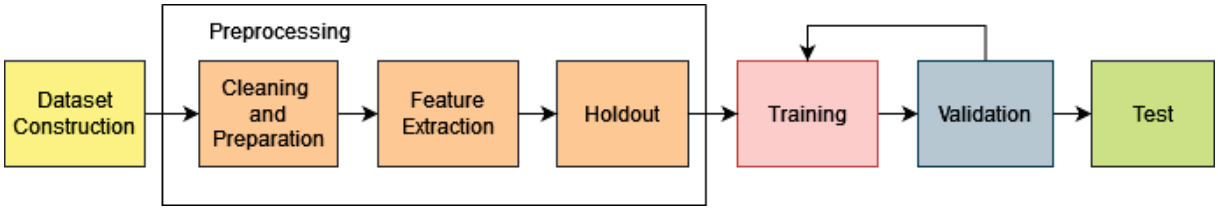


Figure 2.1: Machine Learning Workflow. Source: the author

2.1.4 Generalization

The central challenge of machine learning is getting a model that performs well on unseen data, not just data used to train the model. This is called generalization (Goodfellow et al., 2016; Burkov, 2019).

When a model is not able to predict even the labels of the data it was trained on, this is called underfitting. On the other hand, when a model can predict the labels of the training data very well, but is not able to predict the labels of the data not yet seen, this is called overfitting (Goodfellow et al., 2016; Burkov, 2019). Thus, a model capable of generalizing is one that overcomes underfitting and does not overfit, that is, it performs well both on training data and on data not yet seen (Jurafsky and Martin, 2021).

2.1.5 Imbalanced Data

When a dataset does not have a uniform distribution between classes, it can be said that the dataset is imbalanced. Imbalanced datasets are problematic because the model will spend most of its effort adjusting to the most representative class (Zheng and Casari, 2018).

While some algorithms, like Decision Trees, are capable of dealing with imbalanced data, this can be a problem for many machine learning algorithms. To deal with this situation there are some techniques (Burkov, 2019):

- Oversampling: consists of artificially increasing, in the training set, the number of examples of the underrepresented class. An example of an algorithm that can be used to oversample is the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002);
- Undersampling: consist of randomly removing from the training set some examples from the majority class;
- Class weighting: depending on the machine learning algorithm used for training, greater weight can be assigned to underrepresented class. It is useful to tell the model to "pay more attention" to examples from the minority class.

2.1.6 Model Performance Assessment

Some of the metrics used to evaluate supervised machine learning models in classification tasks are: Confusion Matrix, Accuracy, Precision, Recall and F1Score (Burkov, 2019).

2.1.6.1 Confusion Matrix

The confusion matrix allows the visualization of the classifier's errors and successes. It is represented by a square matrix of size n × n where n is the number of classes. One axis of

the confusion matrix is the label that the model predicted and the other axis is the actual label (Burkov, 2019; Jurafsky and Martin, 2021). Figure 2.2 presents an example of a confusion matrix:

		PREDICTED LABELS	
		NEGATIVE	POSITIVE
TRUE LABELS	NEGATIVE	Number of True Negatives	Number of False Positives
	POSITIVE	Number of False Negative	Number of True Positive

Figure 2.2: Confusion Matrix Example. Source: Author

When we test a binary classifier on an example whose class is known, there are only four different results (Kubat, 2017):

- True Positive (TP): the example is positive and the classifier correctly recognizes it as positive;
- True Negative (TN): the example is negative and the classifier correctly recognizes it as negative;
- False Positive (FP): the example is negative, but the classifier labels it as positive;
- False Negative (FN): the example is positive, but the classifier labels it as negative.

When applying the classifier to an entire set of examples, whose actual classes are known, each of these four results will occur a different number of times. From these numbers it is possible to calculate the Accuracy, Precision, Recall and F1Score (Kubat, 2017).

2.1.6.2 Accuracy

Accuracy represents the frequency of correct classifications generated by the classifier. It is calculated using the number of correctly classified examples divided by the total number of classified examples, according to Formula 2.1 (Kubat, 2017; Burkov, 2019).

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

In imbalanced datasets the accuracy gives a misleading picture of the classifier's performance. The Precision and Recall metrics are best suited to evaluate classifiers in these cases (Kubat, 2017).

2.1.6.3 Precision

Precision reports the frequency of true positives among all examples considered positive by the classifier. It is calculated using the number of true positives divided by the number of positives estimated by the classifier, according to Formula 2.2 (Kubat, 2017; Burkov, 2019).

$$Pre = \frac{TP}{TP + FP} \tag{2.2}$$

In other words, precision is the probability that the classifier is right to label an example as positive (Kubat, 2017; Burkov, 2019). Precision is indicated to be used as a metric when the number of false positives needs to be low (Kubat, 2017).

2.1.6.4 Recall

Recall reports the frequency of true positives among all positive examples in the dataset. It is calculated using the number of true positives divided by the total number of positive examples in the dataset, according to Formula 2.3 (Kubat, 2017; Burkov, 2019).

$$Rec = \frac{TP}{TP + FN} \tag{2.3}$$

Recall is indicated to be used as a metric when the number of false negatives needs to be low (Kubat, 2017).

2.1.6.5 F1 Score

The F1 Score is a combination of precision and recall. It is calculated through the harmonic mean between precision and recall, according to Formula 2.4.

$$F1 = 2 \times \frac{Pre \times Rec}{Pre + Rec} \tag{2.4}$$

Since F1 Score is an average between precision and recall, this means that both have the same weight, so we can interpret its result as follows:

- A classifier will get a high F1 Score if both precision and recall are high.
- A classifier will get a low F1 Score if both precision and recall are low.
- A classifier will get an average F1 Score if one of the measures is low and the other is high.

Similar to precision and recall, F1 Score is indicated for evaluating problems with imbalanced datasets.

2.1.7 Hyperparameters

Hyperparameters are properties of a machine learning algorithm. Their values influence the way the algorithm works. Hyperparameters are not learned by the algorithm from the given data. They need to be set by the developer before training the model (Burkov, 2019).

Each algorithm has its own hyperparameters. Some examples common to neural networks are: number of epochs and learning rate.

The selection of hyperparameters can be done through experimentation. Some of the techniques that can be used are: Grid Search and Random Search (Burkov, 2019; Mohri et al., 2018; Goodwin et al., 2022).

As we saw in the workflow, the search for the best hyperparameters is made in the validation step. The reason behind to use the validation set instead of test set to carry out hyperparameter optimization is to avoid overfitting of the hyperparameters to the test data. Therefore, using the validation set gives confidence that the results on the test data are a true measure of how well our model generalizes (Nielsen, 2015).

2.1.7.1 Grid Search

Grid Search is a technique for hyperparameter optimization. To apply this technique, we first define a set of possible hyperparameter values (parameter grid) that we wish to experiment with. The algorithm will then exhaustively train models with every possible combination of hyperparameters from that grid, calculating a metric using cross-validation (describe in Section 2.1.8). The result will be a ranking of the trained models.

2.1.8 Cross-Validation

In the machine learning workflow, in one of the steps the data is partitioned into training, validation and testing. Although this division avoids overfitting, in problems that have few labeled data, the validation and test sets may not be large enough to be representative (Burkov, 2019; Jurafsky and Martin, 2021).

One solution to this problem is called cross-validation. It works as follows: first, the dataset is partitioned into training and testing. Then the training set is randomly partitioned into k distinct subsets. Then k training iterations are performed, in each iteration one of these subsets is reserved as a validation subset, while the model is trained using the remaining k-1 subsets, at the end of each iteration the model is evaluated on the validation subset. The performance of the model will be represented by the average of the metrics of each iteration (Burkov, 2019; Jurafsky and Martin, 2021).

In other words, instead of dividing the dataset into training, validation, and testing as described in Section 2.1.3, the training and validation sets are dynamically partitioned during training.

Once the best hyperparameter values for the model are found, the entire training set is used to build a model that will be evaluated using the test set (Burkov, 2019).

2.1.9 Nondeterminstic Behavior

ML models are not deterministic by nature. There are two major sources of nondeterminism. The first is algorithmic, it refers to the randomness that developers intentionally introduce to improve model generalization. As examples: shuffle the training dataset and dropout (dropout is describe in Section 2.3). The second is non-algorithmic, it refers to the randomness that developers cannot control. The most significant non-algorithmic nondeterminism is caused by data parallelism (Xu et al., 2022).

Thus, to ensure some reproducibility, is possible to set the random seed during the training. Strategies such as cross-validation and multiple training runs with different random states, can increase confidence in the model results.

2.2 CLASSIC ALGORITHMS

Among the classic machine learning algorithms that have been widely used in supervised learning are: Decision Trees, Logistic Regression, Naive Bayes (NB), Support Vector Machines (SVM) and K-Nearest Neighbors (KNN).

2.2.1 Logistic Regression

Logistic Regression (LR), in particular, has traditionally been used to perform text classification tasks, including social media texts and studies in the field of ecology (Jurafsky and Martin, 2021; Edwards et al., 2022).

The name comes from statistics and is due to the fact that the mathematical formulation of LR is similar to that of linear regression (Burkov, 2019).

LR is a discriminative classifier, it means that this algorithm only tries to learn to distinguish between the classes. In a text classification scenario it attempts to directly compute the probability of class c given document d, it is $P(c|d)$, and learning to assign high weight to document features that directly improve its ability to discriminate between possible classes (Jurafsky and Martin, 2021).

LR as a probabilistic classifier that makes use of supervised machine learning has four components (Jurafsky and Martin, 2021):

- A feature representation of the input. It is an input data formed by tuples in the format (x, y), where x is the feature vector and y is the true label;
- A classification function that computes $\hat{y}$, the estimated class, via $P(y|x)$ (e.g, sigmoid function);
- An objective function for learning, usually involving minimizing error on training examples (e.g., cross-entropy loss function);
- An algorithm for optimizing the objective function (e.g., stochastic gradient descent).

Considering a binary problem and a single input x. To calculate the probability $P(y = 1|x)$, LR learns, from a training set, a vector of `weights` and `bias`. Each weight w_i is a real number, and is associated with one of the input features x_i. The weight w_i represents the importance of the attribute x_i for the classification decision, and can be positive or negative. The bias b is another real number added to the weighted inputs, which determines where to set the classification boundary (Jurafsky and Martin, 2021; Szeliski, 2022).

To make a decision on a new instance, LR first multiplies each x_i by its weight w_i, sums up the weighted features, and adds the bias b. The result is a single number z that expresses the weighted sum of evidence that the predicted class belongs to a given class. This can be expressed by Equation 2.5 (writes in dot product notation) (Jurafsky and Martin, 2021; Szeliski, 2022).

$$z = \mathbf{w} \cdot \mathbf{x} + b \tag{2.5}$$

As the weights are real values, the output might be negative, and z ranges from $-\infty$ to ∞. To create a probability, z must be passed through the `sigmoid function`. The sigmoid takes a real-valued number and maps it into the range [0,1]. The sigmoid has the following Equation (Burkov, 2019; Jurafsky and Martin, 2021):

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.6}$$

Where e is the base of the natural logarithm, also called Euler's number. Thus, when we apply the sigmoid to the sum of the weighted features, $\sigma(z)$, we obtain a number between 0 and 1 (Jurafsky and Martin, 2021).

For make a decision about which class to apply to the example x, it uses a decision boundary, like 0.50. Thus $\hat{y} = 1$ if the probability $P(y = 1|x)$ is greater than 0.5. The decision boundary can be express by the following Equation (Burkov, 2019; Jurafsky and Martin, 2021):

$$decision(x) = \begin{cases} 1 & \text{if } P(y = 1|x) > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

What the LR produces via Equation 2.5 is $\hat{y}$. It makes it learning the parameters (i.e., weights and bias) that make $\hat{y}$ for each training observation as close as possible to the true y. To do this, two components are needed: 1) A metric for how close the estimate $\hat{y}$ is to the true y: the cross-entropy loss function. 2) An optimization algorithm to iteratively update weights so as to minimize the loss function: the gradient descent (Jurafsky and Martin, 2021; Szeliski, 2022).

2.2.1.1 The Cross-Entropy Loss Function

To measure how close the classifier output is to the correct output. It is used a loss function that prefers the correct class labels of the training examples to be more likely. This is called conditional maximum likelihood estimation. Meaning that the parameters weights and bias that maximize the log probability of the true y labels in the training data given the observations x are chosen by the function (Jurafsky and Martin, 2021; Szeliski, 2022).

2.2.1.2 Gradient Descent

The goal of gradient descent is to find the optimal weights, it means minimize the loss function. Gradient descent finds a minimum of a function by figuring out in which direction, in the space of the parameters, the function's slope is rising the most steeply, and than moves itself in the opposite direction. The intuition behind this is that if you are hiking in a canyon and trying to descend faster to the river at the bottom, you can look 360 degrees around you, find the direction where the ground is steepest, and walk downhill in that direction (Jurafsky and Martin, 2021; Szeliski, 2022).

2.2.1.3 Hyperparameters

The LR implementation in the Scikit Learn library (Pedregosa et al., 2011)[1] has some hyperparameters, of which the following were tested in our experiments:

- Solver: the optimization algorithm: lbfgs (default) and liblinear (indicated for small datasets).
- Penalty: regularization terms that is used to penalize large weights, with the goal of reduce overfitting: L2 (default) and L1.
- C: controls the strength of the regularization penalty. A smaller value of C results in stronger regularization, while a larger value of C results in weaker regularization.

2.3 NEURAL NETWORKS

Neural networks are machine learning models that process information in a way inspired by the human brain. They are formed by a network with processing units that mimic the functioning of neurons (Christin et al., 2019). A neural network has three main parts (Christin et al., 2019; Jurafsky and Martin, 2021):

- An input layer that receives the data;
- The processing core, which contains one or more hidden layers;

[1]Library used in some of our experiments.

- An output layer that returns the result of the model.

A neural network with many hidden layers is called a deep neural network, although there is no consensus on the number of hidden layers that differentiates a simple network from a deep network (Christin et al., 2019). Training deep neural networks can also be called Deep Learning (DL). Figure 2.3 illustrates the difference between these networks.

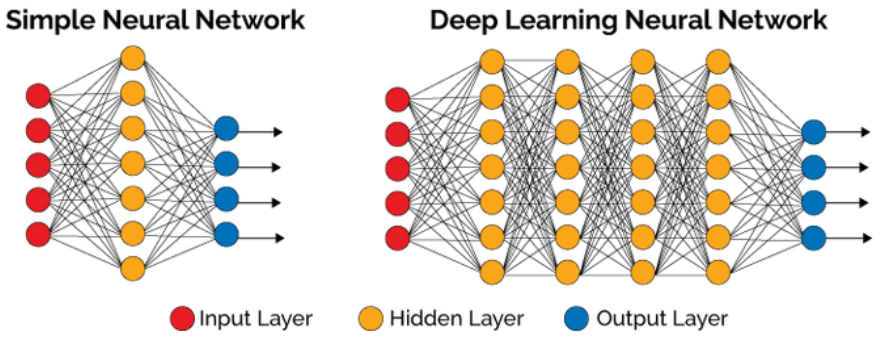


Figure 2.3: Simple Neural Network and Deep Neural Network Illustration. Adapted from https://www.deeplearningbook.com.br

To explain how the neural network learns we will use an artificial neuron model. Consider that this model takes a feature vector as input and produces a single binary output. Figure 2.4 shows such model.

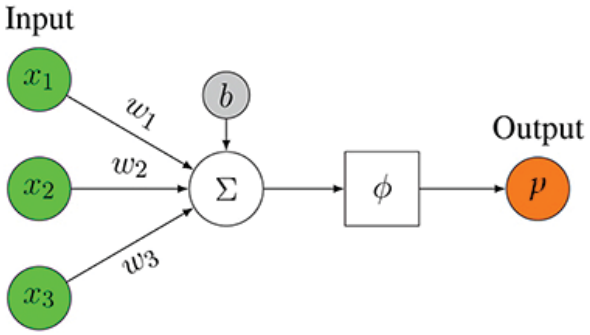


Figure 2.4: Artificial Neuron Model. Source: Adapted from Emmert-Streib et al. (2020).

The elements in the figure are (Nielsen, 2015; Burkov, 2019; Emmert-Streib et al., 2020; Jurafsky and Martin, 2021):

- x_1, x_2, x_3: represent the model inputs. One input for each feature vector position.
- w_1, w_2, w_3: represent the weights. Real numbers that express the importance of the respective inputs to the output. By default, weights are started at random values. It is possible start them at zero or with specific values. But they are really learned during training. They work like a memory that is reinforced as more examples are processed.
- $\sum$: indicates the weighted sum of the multiplication of the input vector by the vector of weights + bias.
- b: the bias. Specifies the threshold for the result produced by the weighted sum ($\sum$) to generate an activation trigger value. The greater the bias, the easier it is to make the neuron emit 1 as an output.

- θ: the activation function. Its goal is to limit the output of the neuron to a range of values. In this model, the function limits the output to 0 or 1.

- p: represent the model output.

Consider a supervised learning scenario in which the training data is formed by tuples in the format (x, y), where x is the feature vector and y is the label. The artificial neuron model (Figure 2.4) takes the feature vector $[x_1, x_2, x_3]$, computes a weighted sum ($\sum$), multiplying each value by a weight ($[w_1, w_2, w_3]$), adds them to a bias b, and then passes the resulting sum through a activation function (θ) to result p as a number 0 or 1. Given the output p, it checks if p is equal to y. If true the weights are maintained and the next input is processed. Otherwise, the weights and bias are updated with the error before processing the next input. The error that is used to update the network weights is calculated by the distance between the output p and the true label y. We call this distance the loss function. The most common loss function used for neural networks is the `cross-entropy loss` (described in 2.2.1) (Nielsen, 2015; Jurafsky and Martin, 2021).

However, if we use this error to update the weights directly, the weights tend to get too large. This can cause the weight vector to get stuck at the local minimum and not be able to find the global minimum of the cost function. The most commonly used algorithm as an optimization method to deal with this problem is the `gradient descent` (described in 2.2.1). Its goal is to identify in which direction there will be a greater increment in the weights and, based on the opposite direction and on a `learning rate`, update the weights (Nielsen, 2015; Goodfellow et al., 2016; Burkov, 2019; Jurafsky and Martin, 2021). In this dissertation we used Adam and AdamW as optimizers, which are algorithms based on gradient descent and are currently the most popular optimizers for DL (Szeliski, 2022).

Each time the entire training set is passed through the network is called an `epoch`. It is possible to define the number of examples that will be used to update the weights, which is called `batch size`. As an example: if the batch size is 1, the weights are updated for each example that passes through the network. If the batch size is 16, the weights are updated for every 16 examples, and so on. Each time a batch passes through the network it is called a `step`.

When the network has more than one layer of neurons (Figure 2.3), the input information is passed through all layers. Each layer receives as input the output of the previous layer. This is called feedforward. This kind of network is called Feedforward Neural Network (FFNN) or Multilayer Perceptron (MLP) (Goodfellow et al., 2016; Emmert-Streib et al., 2020; Jurafsky and Martin, 2021). The error calculated by the optimizer is also propagated to the hidden layers of the network, which is called backpropagation.

The basic hidden layer in FFNN is called `fully-connected layer`. This name comes from the fact that all neurons in this layer receive as input the outputs of each of the neurons in the previous layer (Burkov, 2019; Emmert-Streib et al., 2020). As illustrated in Figure 2.3.

In the artificial neuron model used as an example, we use an activation function that limits the output to either 0 or 1. But there are several kinds of activation functions. In this work we used `sigmoid` and `ReLU`.

The sigmoid function (described in 2.2.1) limits the output to a real number between 0 and 1 (Nielsen, 2015; Jurafsky and Martin, 2021). This type of output is interesting to be used in the last layer of a network in a binary classification problem. This type of output can be considered as the *a posteriori probability* for the input x to belong to the positive class.

Another activation function frequently used in computer vision problems is the Rectified Linear Unit (ReLU). This function returns 0 if it receives a negative value, otherwise it returns the value itself (Nielsen, 2015; Carvalho et al., 2021).

Neural networks are more prone to overfitting, especially when the training dataset is small (Nielsen, 2015). Among the regularization techniques that can be applied to reduce overfitting are: data augmentation, dropout and early stopping.

The best way to better generalize a machine learning model is to train it with more data. However, collecting and labeling more data is not always possible. A way around this problem is to create synthetic examples from the original examples and add it to the training set (Nielsen, 2015; Goodfellow et al., 2016; Burkov, 2019). This is called `data augmentation`. For image, operations such as: zooming, rotating and flipping, can be applied. This type of operation has proven to be effective for increasing the training dataset and increasing the model performance (Nielsen, 2015; Goodfellow et al., 2016; Burkov, 2019; Szeliski, 2022). For tabular data, SMOTE is one of the techniques that can be applied. Although this technique is commonly used to deal with imbalanced datasets, by creating examples for minority class, it can also be used to increase the dataset size without loss of classifier performance (Machado et al., 2022).

The concept of `dropout` is to randomly and temporarily set the output of some neurons to zero during the training (Goodfellow et al., 2016; Burkov, 2019; Leskovec et al., 2020; Jurafsky and Martin, 2021). This action injects noise into the training process and also prevents the network from overly specializing neurons to particular samples or tasks (Szeliski, 2022). The higher the percentage of excluded neurons, the greater is the regularization effect (Burkov, 2019). But this strategy is less effective when the dataset is small (Goodfellow et al., 2016).

When training a neural network, we can measure the loss in the validation set at the end of each epoch. The point at which loss begins to increase rather than decrease is the point where the training process started to learn idiosyncrasies from the training data rather than a generalizable model. A simple approach to avoid this problem is to stop training when the loss in the validation set stops decreasing. This strategy is called `early stopping` (Nielsen, 2015; Goodfellow et al., 2016; Burkov, 2019; Leskovec et al., 2020). In addition, we can compute on the validation data other metrics, such as those discussed in Subsection 2.1.6, at the end of each epoch and save the model whenever the metric improves. This strategy is called `checkpoint`. In this way, we keep the best version of the model to be used to predict new data (Goodfellow et al., 2016).

Neural networks have been widely used for image and natural language processing. Among the neural network architectures that have been widely used in supervised learning are: MLP, CNN and Recurrent Neural Network (RNN).

As we can see, neural networks have many hyperparameters to define: how many layers, how many neurons in each layer, what activation functions to use, the number of epochs, batch size, loss function, optimizer, learning rate, and regularization techniques to use. And it is up to the developer to define these hyperparameters (Jurafsky and Martin, 2021).

2.3.1 Representation Learning

There are entire books devoted to feature engineering. One of them is the book of Zheng and Casari (2018), such is the importance that data representation has for machine learning.

It is important not to confuse "Feature Extraction" with "Feature Engineering" (or Feature Selection). The first deals only with the transformation of raw data into vectors, as discussed in 2.1.3. The second deals with selecting the most appropriate characteristics for a given problem, according to the available data, the model used and the task to be performed (Zheng and Casari, 2018).

When using classic machine learning algorithms, the task of selecting which features should compose the vector is done manually. The developer is responsible for extracting an adequate representation for the problem. On the other hand, in neural networks it is possible to make the network itself learn the characteristics of the data. This is called Representation Learning (Goodfellow et al., 2016).

Representations learned by neural networks generally result in better performance than representations designed by hand. They also allow machine learning systems to be quickly adapted to new tasks, with minimal human intervention (Goodfellow et al., 2016).

Representation Learning is the basis for creating word embeddings, CNNs and pre-trained models, as we will see in the next Sections.

2.3.2 Transfer Learning

One of the problems of DL is the need for large volumes of labeled data to achieve satisfactory performance. The larger the feature vector, the more parameters or the more complex the problem is, the more data is needed. Training neural networks from scratch with few data almost always leads to overfitting. Added to this is the fact that data collection and labeling are tasks that consume time and often financial resources (Toivonen et al., 2019; Kalyan et al., 2021).

Transfer learning is an approach in which a model is pre-trained on a large set of generic data and then used as a starting point for training a new model. This practice is especially interesting in DL, as it allows training models without the need for a large volume of labeled data and with less time and processing resources (Burkov, 2019; Toivonen et al., 2019; Jurafsky and Martin, 2021; Souza et al., 2020; Goodwin et al., 2022).

Its use is indicated when there is a pre-trained model on data and tasks similar to the new problem and when the dataset available for training is too small to train a model from scratch (Goodwin et al., 2022).

In summary, this approach works as follows (Burkov, 2019; Souza et al., 2020):

- Select a pre-trained model that has some affinity with the new problem;
- Remove layers responsible for classification or regression;
- Freeze the remaining layers;
- Add layers adapted to the new problem;
- Train the new layers;
- Optionally, unfreeze the layers and retrain the entire model with a new dataset.

Another form of transfer learning is the use of a pre-trained model as feature extractor, that is, the pre-trained model is used to generate vectors to be used as input into a model for a specific task.

Transfer learning has been widely adopted and achieved state-of-the-art (SOTA) performance in NLP and computer vision problems (Souza et al., 2020; Kalyan et al., 2021).

There are plenty of pre-trained models for computer vision tasks (e.g., ResNet (He et al., 2015), DenseNet121[2], VGG16[3]) and NLP (e.g., BERTimbau (Souza et al., 2020), BERT (Devlin et al., 2019), XML-R (Conneau et al., 2020), Word Embeddings).

Among the advantages of using pre-trained models are (Burkov, 2019; Christin et al., 2019; Jurafsky and Martin, 2021; Kalyan et al., 2021):

[2]https://keras.io/api/applications/densenet/

[3]https://keras.io/api/applications/vgg/

- Pre-trained models can be adapted for specific tasks by just adding news layers, without the need to train a model from scratch.
- They reduce the need for a large amount of labeled data for training, as pre-trained models help the model perform better even with small datasets.
- DL, due to the large number of parameters, tend to overfit when trained from scratch on small datasets. Because pre-training provides a good start, it prevents overfitting.
- Models pre-trained on large datasets learn universal representations. This knowledge can be easily transferred to specific task models reducing training time.
- Pre-trained models can be used as feature extractors.

2.4 COMBINING CLASSIFIERS

A set of classifiers can have their individual decisions combined in some way to classify new examples. This approach is found in the literature as multiple classifiers (or models) or ensemble (Dietterich, 1997; Carvalho et al., 2021).

Classic machine learning models, or even neural networks, can be combined. Basically we can combine models in two ways: we can simply combine the results of trained models using some fusion rule or we can use some specific algorithm for generation and training of models together.

Some examples of ensemble algorithms are: Random Forest (RF), Gradient Boosting (GB) and Extra Trees. These algorithms increase performance by training and combining hundreds of weak Decision Tree models (Burkov, 2019).

One of the motivations for combining multiple models is that when multiple strong uncorrelated models agree, they are more likely to agree on the correct result (Kittler et al., 1996; Burkov, 2019; Carvalho et al., 2021). On the other hand, when combining several weak models (e.g., Random Forest), a robust model can be obtained by decreasing the variance, for example. The main idea is not to rely on a single model (Burkov, 2019).

2.4.1 Fusion Rules

When combining classifier predictions we must observe the type of the model output. If the classifier only provides class labels as output or if the classifier outputs *a posteriori* probability associated with the class label (Dietterich, 1997; Carvalho et al., 2021).

When the classifier outputs only the class, one of the most common rules found in the literature is the Majority Vote Rule. In this rule the majority class among the classes predicted by the models is used as a result (Burkov, 2019; Carvalho et al., 2021).

When each classifier can produce an estimate of the probability for the instance to belong to a class, some of the rules that can be used to fuse the output are (Kittler et al., 1996; Burkov, 2019; Carvalho et al., 2021): Product, Max, Min, Average and Sum.

Consider a binary classification scenario, whose classes are defined by w_0 for negative class and w_1 for positive class. Given a test example x and a set of classifiers m. Each classifier produces the *a posteriori* probability for x to belongs to w_0 ($P(w_0|x)$) and w_1 ($P(w_1|x)$).

Also consider the *a priori* probability of each class, with $P(w_0)$ for the negative class and $P(w_1)$ for the positive class. These probabilities are used in the Product Rule and Sum Rule.

2.4.1.1 Product Rule

In the Product Rule, it is multiplied all the *a posteriori* probabilities estimates of the models for each class and the *a priori* probability of the class. The multiplication that maximizes the probability will be used as the result. This is a very severe rule, as it is enough for one of the classifiers to present a low probability that the final result will be equally low. It is calculated using Equation 2.8:

$$\begin{aligned} product_0 &= P(w_0) \prod^m P(w_0|x) \\ product_1 &= P(w_1) \prod^m P(w_1|x) \\ class &= argmax(product_0, product_1) \end{aligned} \tag{2.8}$$

Where:

- $product_0$: multiplication of all *a posteriori* probabilities for x to belongs to the negative class and the *a priori* probability of the class.
- $product_1$: multiplication of all *a posteriori* probabilities for x to belongs to the positive class and the *a priori* probability of the class.
- class: the multiplication that maximizes the probability.

2.4.1.2 Sum Rule

In the Sum Rule, it is summed all the *a posteriori* probabilities estimates of the models for each class plus the *a priori* probability of the class. The sum that maximizes the probability will be used as the result. It is calculated using Equation 2.9:

$$\begin{aligned} sum_0 &= P(w_0) + \sum^m P(w_0|x) \\ sum_1 &= P(w_1) + \sum^m P(w_1|x) \\ class &= argmax(sum_0, sum_1) \end{aligned} \tag{2.9}$$

Where:

- sum_0: sum of all *a posteriori* probabilities for x to belongs to the negative class + the *a priori* probability of the class.
- sum_1: sum of all *a posteriori* probabilities for x to belongs to the positive class + the *a priori* probability of the class.
- class: the sum that maximizes the probability.

2.4.1.3 Average Rule

In the Average Rule, it is calculate the average of the *a posteriori* probabilities estimates of the models for each class. The average that maximizes the probability will be used as the result. It is calculated using Equation 2.10:

$$\begin{aligned} avg_0 &= \sum^m P(w_0|x)/m \\ avg_1 &= \sum^m P(w_1|x)/m \\ class &= argmax(avg_0, avg_1) \end{aligned} \tag{2.10}$$

Where:

- avg_0: the average of all *a posteriori* probabilities for x to belongs to the negative class.

- avg_1: the average of all *a posteriori* probabilities for x to belongs to the positive class.
- class: the sum that maximizes the probability.

2.4.1.4 Max Rule

In the Max Rule, the maximum *a posteriori* probability estimated by the models for each class is considered. The maximum that maximizes the probability will be used as the result. It is calculated using Equation 2.11:

$$\begin{aligned} max_0 &= \max^m P(w_0|x) \\ max_1 &= \max^m P(w_1|x) \\ class &= argmax(max_0, max_1) \end{aligned} \tag{2.11}$$

Where:

- max_0: the maximum *a posteriori* probability for x to belongs to the negative class.
- max_1: the maximum *a posteriori* probability for x to belongs to the positive class.
- class: the maximum that maximizes the probability.

2.4.1.5 Min Rule

In the Min Rule, the minimum *a posteriori* probability estimated by the models for each class is considered. The minimum that maximizes the probability will be used as the result. It is calculated using Equation 2.12:

$$\begin{aligned} min_0 &= \min^m P(w_0|x) \\ min_1 &= \min^m P(w_1|x) \\ class &= argmax(min_0, min_1) \end{aligned} \tag{2.12}$$

Where:

- min_0: the minimum *a posteriori* probability for x to belongs to the negative class.
- min_1: the minimum *a posteriori* probability for x to belongs to the positive class.
- class: the minimum that maximizes the probability.

2.5 NATURAL LANGUAGE PROCESSING

NLP is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech (Chowdhury, 2003; Caseli and Nunes, 2023).

Transformer-based pre-trained language models are considered SOTA for NLP (Devlin et al., 2019; Kalyan et al., 2021). Language models are those that use statistical and probabilistic techniques to determine the probability of a given word occurring in a sentence (Jurafsky and Martin, 2021). Transformers (Vaswani et al., 2017) are deep neural networks that use an attention mechanism that learns the contextual relationships between words in a sentence. These networks are composed of encoding and/or decoding layers.

A model can be pre-trained using both encoding and decoding layers or just one of them. In general, a encoding-based model consists of an embedding layer followed by encoding layers. The output of the last encoding layer is considered as the contextual representation of the input

sentence (Vaswani et al., 2017; Kalyan et al., 2021). This type of model is generally used in natural language comprehension tasks, such as text classification.

Self-supervised learning enables transformers to learn representations by solving pre-training tasks (Kalyan et al., 2021), such as masking part of the sentence so that the model completes the missing words, which is called Masked Language Modeling.

The goals of self-supervised learning are: a) Learning general language representations, which can be used in different NLP tasks (e.g., text classification) through transfer learning (Subsection 2.3.2); b) Increase the generalizability of the model by learning on large volumes of freely available unlabeled text data.

A pre-trained language model can be used in at least two ways (Kalyan et al., 2021): a) as a feature extractor, that is, the pre-trained model is used to generate context-based vectors that can be used as input to a specific task model (we detail this in Subsection 2.5.2); b) be refined to perform specific tasks; this refinement can range from adding or changing output layers, to re-training the model on a new dataset (i.e., Transfer Learning - Subsection 2.3.2).

Figure 2.5 illustrates the summary of NLP workflow. It shows techniques, approaches and models that can be used in some steps of the workflow.

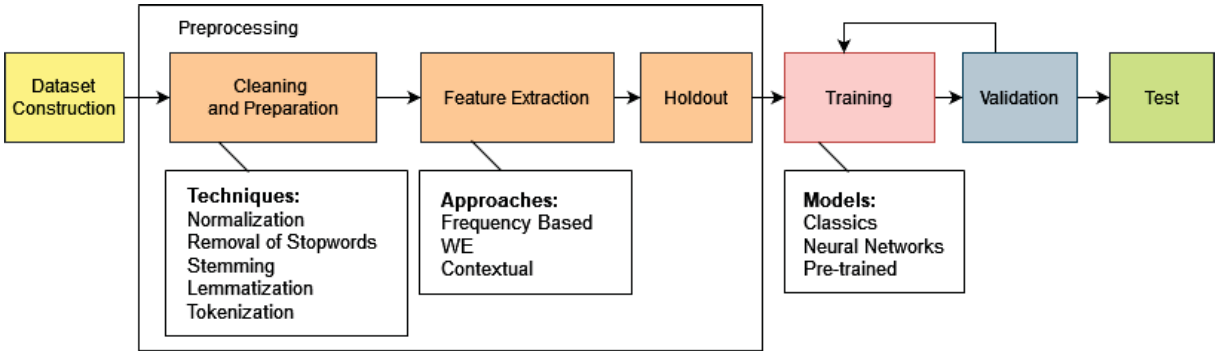


Figure 2.5: Summary of NLP Workflow. Source: the author

2.5.1 Cleaning and Preparation

As discussed in Subsection 2.1.3 the collected data cannot be fed to the machine learning model in its raw state. Before that, data cleaning and preparation is necessary. This section covers cleaning and preparation techniques generally applied in NLP tasks.

2.5.1.1 Normalization

Normalization consists of converting text into a standard format (Jurafsky and Martin, 2021). This process includes treatments such as: converting uppercase letters to lowercase; removal of non-alphabetic characters (e.g., signs, punctuation, line break, emoji, numbers); removal of user mentions; correction of spelling mistakes; accent removal; hashtag removal.

2.5.1.2 Removal of Stopwords

Removal of stopwords, that is, very frequent words, such as: pronouns, articles and prepositions (Zheng and Casari, 2018; Jurafsky and Martin, 2021). This process helps to reduce the size of the vocabulary. However, according to Jurafsky and Martin (2021), in many applications removing these words does not improve performance of the model.

2.5.1.3 Stemming

Stemming is a text normalization technique that switches inflected words to their root (Jurafsky and Martin, 2021). For example: *studies* and *studying*, which are inflected, are transformed into *stud*. The stemming process can reduce a word to another that is grammatically incorrect but still has value for analysis.

2.5.1.4 Lemmatization

Lemmatization is a text normalization technique that consists of the process of deflecting a word to determine its lemma (Jurafsky and Martin, 2021). As an example: *studies* and *studying*, which are inflected, are transformed into *study*. The lemmatization process will always result in a word that actually exists in the grammar.

2.5.1.5 Tokenization

Tokenization is the task of segmenting a document into tokens (Jurafsky and Martin, 2021). A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing (e.g., a word, punctuation symbol, emoji).

Some of these techniques, such as stopwords removal, lemmatization and stemming are also ways to decrease the dimensionality of the feature vector.

2.5.2 Feature Extraction

As discussed in Subsection 2.1.3, for a machine learning model to be able to process data, they need to be represented in the form of a feature vector. According to Vargas et al. (2021) the definition of an appropriate representation is crucial and directly influences the performance of the model.

Among the approaches to represent words that will be presented in this work are: representation based on frequency, Word Embeddings and contextual.

2.5.2.1 Frequency Based Representation

Frequency-based representation encompasses techniques that create vectors based on how often words appear in documents. Among them are: One-Hot Encoding, Bag-of-Words (BoW) and TF-IDF. In this work we will detail the TF-IDF technique.

Term Frequency–Inverse Document Frequency (TF-IDF) is a measure used to determine the relevance of a word for a given set of documents (Zheng and Casari, 2018; Leskovec et al., 2020; Jurafsky and Martin, 2021).

The TF-IDF is calculated by the following formula:

$$TF - IDF(w, d) = TF(w, d) \times log(\frac{N}{DF(w)}) \tag{2.13}$$

Where:

- TF: Term Frequency, quantifies the frequency of the word w in the document d.
- N: Number of documents.
- DF: Document Frequency, number of documents that contain the word w.

The Inverse Document Frequency (IDF) is given by the fraction N / DF. The higher the TF-IDF value, the more important the word is to discriminate a document (Zheng and Casari, 2018; Leskovec et al., 2020; Jurafsky and Martin, 2021).

Using TF-IDF it is possible to convert the text into numerical vectors. As an example, consider 3 documents as a basis for vocabulary and calculation of the TF-IDF:

- Post 1: Linda perigosa #caravelaportuguesa.
- Post 2: Cuidado #caravelaportuguesa.
- Post 3: Linda #caravelaportuguesa.

Calculating how often each word appears in documents (Term Frequency), we obtain the numbers shown in Table 2.1:

Table 2.1: Example - Term Frequency

Word	Post 1	Post 2	Post 3
caravelaportuguesa	1	1	1
cuidado	0	1	0
linda	1	0	1
perigosa	1	0	0

Calculating the TF-IDF we obtain the numbers shown in Table 2.2:

Table 2.2: Example - TF-IDF

Word	Post 1	Post 2	Post 3
caravelaportuguesa	0.00	0.00	0.00
cuidado	0.00	1.09	0.00
linda	0.40	0.00	0.40
perigosa	1.09	0.00	0.00

Therefore, the feature vector of Post 1 is equal to [0.00, 0.00, 0.40, 1.09].

Among the limitations of techniques based on word frequency is that these techniques produce sparse vectors, formed mostly by zeros, and as large as the size of the vocabulary. In addition, words outside the vocabulary, that is, words that do not appear in the training set, are ignored by the model during validation and testing. However, its biggest limitation comes from the lack of context (Jurafsky and Martin, 2021).

The TF-IDF implementation in the Scikit Learn library (Pedregosa et al., 2011)[4] has some hyperparameters, of which the following were tested in our experiments:

- N-gram: An n-gram is a sequence of n tokens. 1 token is a 1-gram or unigram. 2 tokens is a 2-gram or bigram, and 3 tokens is 3-gram or trigram.
- Max features: When building the vocabulary, the method orders the features by the top Term Frequency and ignores features that exceed the given threshold.
- Max DF: When building the vocabulary, the method ignores terms that have a Document Frequency (DF) strictly higher than the given threshold.

[4]Library used in some of our experiments.

- Min DF: When building the vocabulary, the method ignores terms that have a Document Frequency (DF) strictly lower than the given threshold.

2.5.2.2 Word Embeddings

Word Embeddings (WE) encompass techniques that create dense vectors, as opposed to frequency-based techniques.

Representation Learning, discussed in Subsection 2.3.1, is the basis for WE techniques. That is, instead of creating representations manually using feature engineering, a neural network is used to learn the representation directly from the set of documents (Jurafsky and Martin, 2021).

WE represent words as low-dimensional vectors and manage to capture the semantic relationships between words. Words with similar meaning will tend to occur close to each other in the vector space, revealing one of the properties of WE which is that similar words have vectors of similar characteristics (Burkov, 2019; Jurafsky and Martin, 2021).

There are different WE creation processes, such as: Word2Vec, Wang2vec, FastText and GloVE. Furthermore, within these processes there are two variations: Continuous Bag-of-Word and Skip-Gram.

The great advantage of WE is in transfer learning. Currently, it is possible to find several pre-trained word embeddings in Portuguese.

Among the limitations of WE is the fact that the vectors created using this technique are static, that is, they are not able to capture the context in which the words occur. For example, in the sentences: "the woman went to the bank" and "the woman sat on the bank" the word *bank* has the same vector. Also, words outside the vocabulary are ignored by the model during validation and testing (Jurafsky and Martin, 2021).

2.5.2.3 Contextual Representation

Contextual representation encompasses techniques that create context-based vectors, as opposed to other techniques that represent words statically.

In contextual representation each word is represented by a different vector each time it appears in a different context (Devlin et al., 2019; Jurafsky and Martin, 2021). For example, in the sentence: "the woman went to the bank" the word *bank* will have a different vector than in the sentence "the woman sat on the bank", due to the change of context.

To obtain contextual representations, pre-trained language models are used, such as BERT, which is presented in Subsection 2.5.4.

Vectors generated by language models have the following advantages: a) they consider the context; b) they overcome the problem of words outside the vocabulary; c) they are able to encode more information into vectors because of the model's deep layers architecture (Kalyan et al., 2021).

2.5.3 Machine Learning Models for NLP

In this section some models that can be used for text classification are presented.

2.5.3.1 Classics

Among the classic machine learning models that have been widely used in text classification are: SVM, NB, LR and ensembles (e.g., RF and GB) (Gasparetto et al., 2022). Souza et al. (2018) carried out a systematic review on text mining in Portuguese and among the 203 studies surveyed, the SVM and NB models appear among the most used for the text classification task.

2.5.3.2 Neural Networks

Several neural network architectures have been proposed for text classification, as verified by Minaee et al. (2021) and Gasparetto et al. (2022). Among the most used are: MLP, CNN and RNN, such as Long-Term Short-Memory (LSTM), Gated Recurrent Units (GRU), Bidirectional Long-Term Short-Memory (Bi-LSTM) and Bidirectional Gated Recurrent Units (Bi-GRU).

2.5.3.3 Pre-Trained Language Models

As discussed at the beginning of this Section, language models are trained to learn general language representation. However, specific tasks require specific knowledge. For a language model to perform well in specific tasks, its weights must be close to the ideal configuration for the target task (Kalyan et al., 2021).

As discussed in the next Subsection (2.5.4), BERT-type models can be adjusted by adding an additional output layer to be used as a classifier. Furthermore, it can be retrained with new data.

2.5.4 BERT

One of the most outstanding language models today is Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). According to Souza et al. (2020), the main contribution of this model is the use of bidirectional training, that is, it learns the contextual relations between the words of a sentence by reading it all at once, which is different from other models like OpenAI GPT (Radford et al., 2018) that perform unidirectional training.

The pre-trained BERT can be used in two ways: a) as a feature extractor; or b) refined with an additional output layer to be used as a classifier (Devlin et al., 2019; Kalyan et al., 2021).

Among the BERT-based models are: Multilingual BERT (mBERT) (Devlin, 2019), it is a pre-trained model based on the weights of original BERT, available only in the *base* size. It was trained in 104 languages, including Portuguese. BERTimbau (Souza et al., 2020), it is a pre-trained model in Brazilian Portuguese, it is available in *base* and *large* sizes.

The advantage of using multilingual models is that these models are trained with *corpus* from multiple languages, allowing them to be used in problems that use data from more than one language, such as Instagram, in which it is common for users to mix terms of more than one language in the same post (Kalyan et al., 2021).

As part of the preprocessing algorithm, BERT models use WordPiece (Wu et al., 2016) as its tokenization algorithm. This algorithm performs subword-based tokenization. This means that words not used frequently are splitted into subwords. As examples, the word "caravela" is splitted into "cara" and "vela", while the word "água" is not splitted. The main advantage of this technique is that it interchanges between word-based and character-based tokenization, and thus it is able to deal with words outside the vocabulary.

However, using WordPiece implies that the increase of infrequent words in the sentence also increase the number of tokens created during the tokenization process. BERT models have a `sentence length` limit of 512 tokens. Therefore, anything outside the limit is truncated by the model during preprocessing. On the other hand, if the sentence is smaller than the limit, BERT will fill this with a padding token.

BERT preprocessor generates 3 embeddings as model input:

- input-word-ids: The token identifier in the WordPiece vocabulary.

- input-mask: Contains 1 anywhere the input-word-ids is not padding. The mask allows the BERT model to differentiate between the content and the padding (Devlin et al., 2019).

- input-segment: It contains 0 or 1 indicating whether the token belongs to sentence A or sentence B, respectively. This embedding is useful for sentence pair tasks (e.g., Question/Answering, Translation), in the classification problems this vector will be compounded only by zeros (Devlin et al., 2019).

Therefore, for a given token, its input representation is constructed by combining the corresponding: input-word-ids, input-mask and input-segment embeddings (Devlin et al., 2019).

After processing the input representation, BERT will provide 3 outputs:

- sequence-output: contains the contextual embedding for every token from the input sentence.

- encoder-outputs: contains the vectors generated by each transformer layer from the model.

- pooled-output: contains the contextual embedding for each input sentence.

As an example: take the sentence "I love you" and assume that it is compounded by 3 tokens. Passing the sentence through the BERT preprocessor, it will generate 3 embeddings: "input-word-ids": [11,54,25,0,0,...], "input-mask": [1,1,1,0,0,...] and "input-segment": [0,0,0,0,0,...]. This embeddings are used as input for BERT, that generates as output 3 others embeddings: "sequence-output": [[87,58,98,...],[1,99,9,...],[98,33,87,...]], "encoder-outputs": [[57,58,98,...],[1,699,6,...],[98,33,87,...],...] and "pooled-output": [74,57,68,...]. What interests us most is the `pooled-output` embedding, because it represents the embedding of the entire sentence[5].

For pre-training BERT, Devlin et al. (2019) used AdamW optimizer with weight decay of 0.01, learning rate of 1e-4 with warmup over the first 10,000 steps and linear decay, and a dropout probability of 0.1 on all layers. For fine-tuning, most hyperparameters were the same as in pre-training, with the exception of the batch size, learning rate, and number of epochs. They listed the hyperparameters that worked better for fine-tuning: Batch size: 16, 32. Learning rate: 5e-5, 3e-5, 2e-5. Number of epochs: 2, 3, 4.

2.6 IMAGE CLASSIFICATION

Image classification is a computer vision task that tries to understand an image as a whole (Szeliski, 2022). Image classification models take an image as input and output a prediction about which class the image belongs to. According to Szeliski (2022) most modern image recognition techniques are natural applications of DL.

The inputs for an image classification model are the pixel values that comprise the image. Thus the vetorization process is based on converting the image into raw pixels. Images are three-dimensional objects. They have height, width and channels. A RGB image has three colors channels to indicate the amount of red, green and blue, while a grayscale image contains only one channel. Thus each RGB input image is represented by: height x width x 3 (Burkov, 2019; Zhang et al., 2021).

[5]This is a fictitious and simplified example to understand the process.

Image downscaling is the most common preprocessing step in classification tasks. The main reasons for this is: 1) Batch learning through gradient descent requires the same resolution for all images in a batch; 2) Memory limitations; 3) Large image sizes lead to slower training and smaller images can be processed in larger batches; 4) Decrease the dimensionality of the feature vector. As we have seen larger number of features requires more training data (Talebi and Milanfar, 2021).

Another preprocess commonly applied to computer vision tasks is data augmentation. As we saw in Section 2.3, artificially increasing training data can increase model performance. Some of augmentation techniques include: zooming, rotating, flipping, darkening, cropping and adding noise (Nielsen, 2015; Goodfellow et al., 2016; Burkov, 2019; Szeliski, 2022).

Figure 2.6 illustrates the summary of Image Classification Workflow. It shows techniques and models that can be used in some steps of the flow.

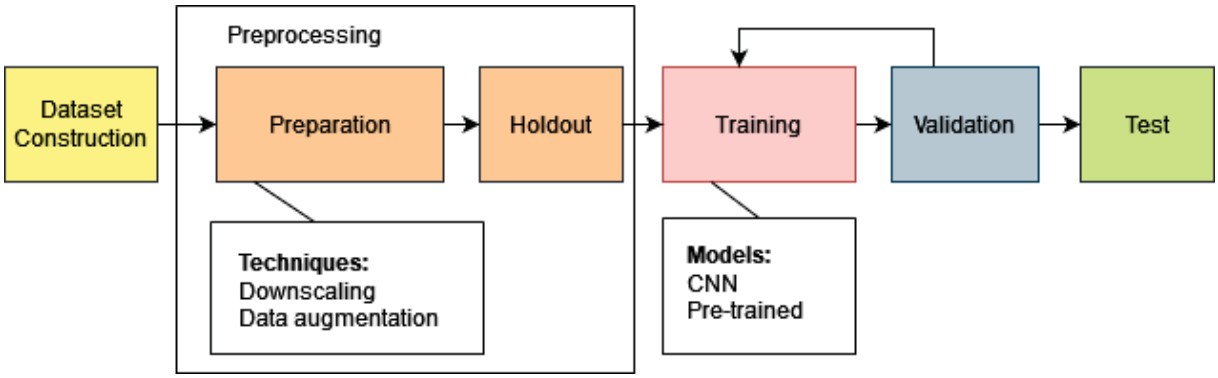


Figure 2.6: Summary of Image Classification Workflow. Source: the author

2.6.1 Convolutional Neural Network

A CNN is a special kind of FFNN, that was designed with image processing in mind (Burkov, 2019; Zhang et al., 2021). CNN-based architectures are ubiquitous in the field of computer vision (Zhang et al., 2021).

It is possible to notice that in images, pixels close to each other usually represent the same type of information: sky, water, leaves, bricks, and so on. The exceptions are the edges: the parts of an image where two different objects "touch" each other. If a neural network is trained to recognize regions with the same information and also the edges, this knowledge would allow the neural network to predict the object represented in the image (Burkov, 2019).

Considering that the most important information in the image is local, we can scan the image using a moving window approach. Then train smaller models, each receiving a square patch as input. The goal of each small model is to learn to detect a specific type of pattern in the input patch. For example, one small model will learn to detect the sky, another one will detect the grass, the third one will detect edges of a building, and so on (Burkov, 2019). This is an overview of how a CNN learns.

Standard CNNs are typically composed of several FFNN layers, including convolution, pooling, and fully connected layers (Section 2.3) (Emmert-Streib et al., 2020).

2.6.1.1 Convolutional Layers

A convolutional layer is an essential part of CNN. This layer has a goal of converting the input into a representation in a more abstract level. However, instead of using the entire input, the

convolutional layer uses a window to slide across the input, performing a convolution operation between each input region and a kernel (or filter). The results are stored in an activation map (or feature map), which can be seen as the output of the convolutional layer. Each kernel can act as a feature extractor and will share its weights with all neurons (Emmert-Streib et al., 2020). The kernel is learned for the network, but the kernel size and the stride (the number of positions the window shifts) are defined by the developer (Burkov, 2019; Emmert-Streib et al., 2020).

Figure 2.7 illustrates a convolution operation with a two-dimensional input. For simplicity the channels are ignored. In the figure the input is a two-dimensional vector with a height of 3 and width of 3 (i.e., shape 3 x 3), while the kernel has a height of 2 and width of 2 (i.e., shape 2 x 2) (Zhang et al., 2021).

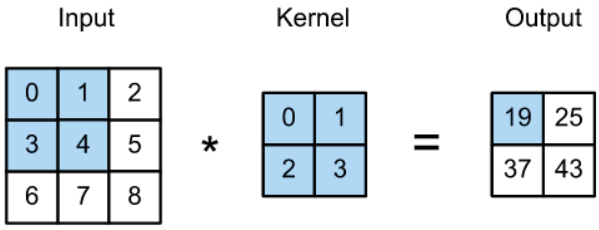


Figure 2.7: Two-dimensional convolution operation. The blue pixels are the first output element as well as the input and kernel elements used for the output computation: 0*0 + 1*1 + 3*2 + 4*3 = 19. Source: (Zhang et al., 2021).

In the two-dimensional convolution operation, one starts with the window positioned in the upper-left of the input vector and slides it through the input vector, always from left to right and from top to bottom. When the window slides to a certain position, the input subvector contained in that window and the kernel vector are multiplied element by element and the resulting vector is summed generating a single scalar value. This result gives the value of the output vector at the corresponding location, forming the activation map (Zhang et al., 2021).

Figure 2.8 illustrates the convolution operation considering 2 channels. So, the resulting output is the sum of the values resulting from the sum of the vector resulting from the multiplication of element by element from the input subvector contained in the window and the kernel vector.

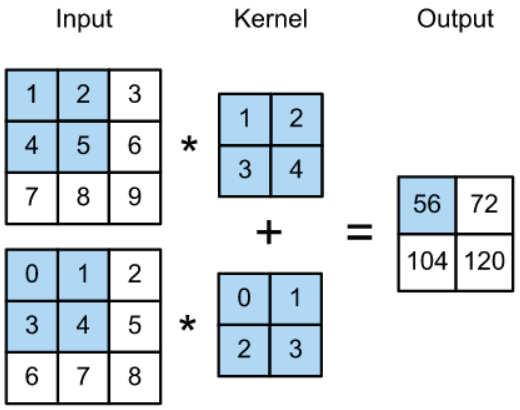


Figure 2.8: Convolution operation with two channels. The output computation of the blue pixels are: (1*1 + 2*2 + 4*3 + 5*4) + (0*0 + 1*1 + 3*2 + 4*3) = 56. Source: adapted from (Zhang et al., 2021).

2.6.1.2 Pooling Layer

A pooling layer is usually inserted between a convolutional layer and the next layer. Pooling layers aim at reducing the dimension of the input with some pre-specified pooling method, resulting

in a smaller input, while keeping as much information as possible (Emmert-Streib et al., 2020; Zhang et al., 2021).

Pooling works in a way very similar to convolution, as a kernel applied using a moving window approach. However, instead of applying a trainable kernel to an input, pooling layer applies a fixed operator (Burkov, 2019; Zhang et al., 2021). There are many types of pooling methods, and among them are: Averaging Pooling and Max-Pooling.

Figure 2.9 illustrates a Max-Pooling operation.

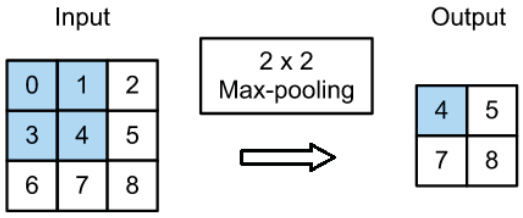


Figure 2.9: Max-pooling with a pooling window shape of 2 x 2. The blue portions are the first output element as well as the input vector elements used for the output computation: max(0,1,3,4) = 4. Source: adapted from (Zhang et al., 2021).

As in the convolutional operations, the pooling operation starts at the upper-left of the input vector and slides from left to right and top to bottom. At each location reached by the pooling window, it calculates the maximum value of the input subvector in the window (Zhang et al., 2021).

2.6.1.3 Flatten Layer

As the output of convolutional and pooling layers are multidimensional, it must to be converted into a single dimension to be fed to the fully connected layer. This operation is called flattening (Shyam, 2021).

When we put all the concepts together we form a complete CNN. Figure 2.10 illustrates a complete CNN. It begins with 28×28 input layer, which are used to encode the pixels for the image. The second layer is a convolutional layer using a kernel 5×5, stride 1 and 20 activation maps. The result is a layer of 20×24×24 hidden feature neurons. The next step is a max-pooling layer, applied to 2×2 regions, across each of the 20 feature maps. The result is a layer of 20×12×12 hidden feature neurons. The penultimate layer is a fully-connected layer. This layer connects every neuron from the max-pooled layer to every one of the 100 neurons. The final layer is a sigmoid (Nielsen, 2015).

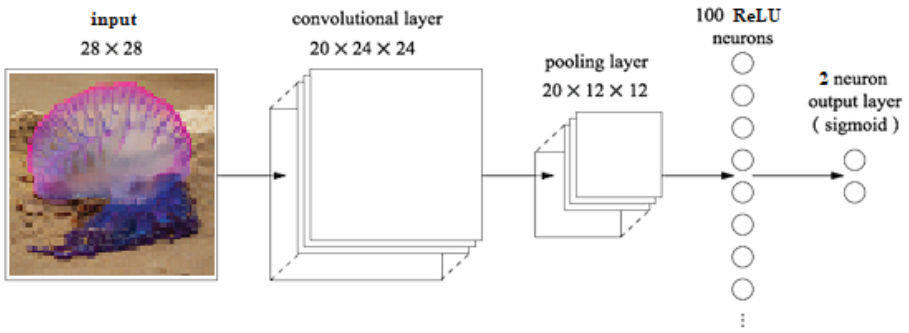


Figure 2.10: Illustration of a complete CNN. Source: adapted from (Nielsen, 2015).

As in any neural network, to train a CNN the developer must define the hyperparameters: how many convolutional layers, how many units in each layer, the kernel parameters, type of pooling and so on, in addition to the hyperparameters already mentioned in Section 2.3. Besides that, training a CNN from scratch demands a big amount of data. Therefore, Transfer Learning (Subsection 2.3.2) is a good strategy to leverage robust networks for downstream tasks (Szeliski, 2022).

A CNN can be used as a feature extractor; or to proceed downstream tasks (e.g., classification or regression).

Among the CNNs pre-trained there is ResNet50 (He et al., 2015), it was pre-trained with Imagenet (Russakovsky et al., 2015). ResNet stands for Residual Network. This network was introduced by He et al. (2015) and won the 1st place on the ILSVRC 2015 classification task with an error of only 3.57%. ResNet50 is a 50-layer convolutional neural network.

2.7 MULTIMODAL MACHINE LEARNING

Multimodal Machine Learning (MMML) aims to build models that can process and relate information from multiple modalities (e.g., text, image, sound) (Baltrušaitis et al., 2019).

Among the approaches found in the literature for the use of multimodalities in machine learning, we can mention: combine results from separately trained models and train multimodal models that take as input multiple modalities (Baltrušaitis et al., 2019; Burkov, 2019).

2.7.1 Combining Results

Called Late Fusion by Baltrušaitis et al. (2019), in this approach models are trained separately, where each one receives as input one of the modalities and then integration of results is done after each modality has been used to make a decision (e.g., classification). This integration can be done using some of the fusion rules discussed in the Section 2.4, such as: average and majority vote (Baltrušaitis et al., 2019).

This approach are not directly dependent on a specific machine learning method. It allows the use of different algorithms for different modalities, allowing more flexibility for training (Baltrušaitis et al., 2019).

Although this is an easy-to-implement approach, the techniques used here were not designed to handle multimodal data (Baltrušaitis et al., 2019).

2.7.2 Training MMML

In general, training machine learning models using data from multiple modalities includes three steps: (Burkov, 2019; Guo et al., 2019):

- Extraction of specific characteristics for each modality;
- Fusion of feature vectors;
- Reasoning stage, such as classification.

2.7.2.1 Feature Extraction

Techniques for extracting features from text were discussed in Subsection 2.5.2. For images, the vectorization was discussed in Section 2.6.

2.7.2.2 Multimodal Fusion

In technical terms, multimodal fusion is the concept of integrating data from multiple modalities in order to predict an outcome (Baltrušaitis et al., 2019).

Among the ways to integrate data from multiple modalities are the average, sum or concatenation of the feature vectors extracted from each modality, which once integrated are used as input for training a machine learning model (Burkov, 2019). For example: considering the vector [i(1), i(2), i(3)] extracted from an image and the vector [t(1), t(2), t(3), t(4)] extracted from a text, the concatenated vector will be [i(1), i(2), i(3), t(1), t(2), t(3), t(4)] (Burkov, 2019).

When using neural networks, it is possible to concatenate the representation layers extracted from sub-networks trained to learn the representation of each of the modalities and use them as input for training a new network. For example, a CNN is trained to learn the representation of images, while an RNN can learn the representation of texts, so it is enough to concatenate the representation layers to use them as input to a network that will do the classification (Burkov, 2019). Some libraries like Keras (Chollet et al., 2015) have methods that allow operations with layers like concatenation and averaging.

This integration approach, in which the feature vectors are integrated before being used by the machine learning model, is called Early Fusion by Baltrušaitis et al. (2019).

2.8 SUMMARY

In this chapter, we presented the theoretical basis for this dissertation.

To carry out the experiments with text, we chose a classic machine learning model available in the Scikit-learn library (Pedregosa et al., 2011): Logistic Regression (presented in Subsection 2.2.1). LR has been chosen because it is a strong baseline for text classification tasks, including social media texts and studies in the field of ecology (Jurafsky and Martin, 2021; Edwards et al., 2022). We also considered a SOTA transformer-based pre-trained language model: mBERT (presented in Subsection 2.5.4), which is available in the Tensorflow Hub (Abadi et al., 2015). We chose mBERT instead BERTimbau, because of one of the characteristics of the text extracted from Instagram: the mix of languages in the same text. As vectorization methods: TF-IDF and mBERT without fine-tuning were used to fed LR.

For perform the experiments with image, we trained CNNs by transferring the learning from a ResNet50 (He et al., 2015) pre-trained with ImageNet (presented in Section 2.6). We chose this network because it had the best performance in the Carneiro et al. (2022) work, more details in Section 3.1.

To proceed with the combination of classifiers, we chose to combine the results using the fusion rules: Product, Max, Min, Average and Sum, described in Section 2.4, since the chosen classifiers have as output the probability associated with the class predicted. As we saw in Section 2.7, there are more sophisticated techniques for using multimodalities in ML. However, we chose to apply a simple technique (i.e., Late Fusion), and left the experimentation of other multimodal learning approaches for future work.

The next Chapter we will present the related works.

3 RELATED WORKS

This Chapter presents the works related to our research. We list here works that used machine learning models and involuntary data for wildlife classification task (Section 3.1). In addition, we list works unrelated to conservation science that performed classification tasks and used NLP in Brazilian Portuguese texts with informal characteristics (Section 3.2) or multimodal machine learning (Section 3.3).

3.1 CONSERVATION SCIENCE AND PASSIVE CITIZEN SCIENCE

In this Section, we list works related to conservation science, which used machine learning for a wildlife classification task using involuntary data (e.g., social media, news sites).

Mazars-Simon (2019), in his master's dissertation, developed a system that searches, classifies, identifies and stores data on sea turtles. He used a dataset of 22,500 images that were shared by conservationists for the project. The dataset was divided into six categories, but the author did not mention whether the data were balanced or not. To build his system, the author trained a CNN to classify images into six categories. With the model trained, he searched for images on Flickr and classified them. If positive for a sea turtle, the system performed individual recognition by comparing the image extracted from Flickr with the images in the database. As a result of training the model for image classification, the author obtained an accuracy of 0.95.

Kulkarni and Di Minin (2021) used texts from Google News and Twitter to train a classifier to identify news about endangered species. The authors used the MurmurHash3[1] as a feature extraction technique and a MLP as a classifier. For training the model, a dataset with 5,464 examples was used, unevenly distributed between the positive and negative classes. As a result, they obtained an F1 Score of 0.96.

Edwards et al. (2022) experimented with different combinations of feature extraction techniques and machine learning algorithms with the aim of identifying wildlife observations in Twitter texts. A dataset with 2,798 tweets was used, equally distributed between positive and negative classes. The authors experimented with BoW, WE and BERT as input for three classifiers: LR, NB and SVM. They also used FastText pipeline[2] and BERT as classifiers. As a result, the model that achieved the best performance was BERT with a F1 Score of 0.96, followed by FastText pipeline and LR trained with BoW, both with 0.94 of F1 Score.

Carneiro et al. (2022) trained neural networks to identify *Physalia physalis* in images extracted from Instagram. A dataset with 12,300 images was used, equally distributed between the positive and negative classes. Part of the training images was obtained from Instagram, part was downloaded through search engines and specialized websites. The authors experimented with different architectures of CNNs with and without pre-training. As a result, they obtained an F1 Score of 0.95 with ResNet50 pre-trained with ImageNet.

Table 3.1 shows the best results of each work listed in this Section.

[1]Technique that converts words into their hashes. https://github.com/aappleby/smhasher/blob/master/src/MurmurHash3.cpp

[2]A neural network that learns to generate WE using the FastText algorithm and has a layer for text classification. https://fasttext.cc/

Table 3.1: Related Works - Conservation Science and Passive Citizen Science. Showing: Work, Problem, Dataset Size (Size), Balanced data (Balanced), Classification type (Type), Architecture and Result. The best results of each work are presented.

Work	Problem	Size	Balanced	Type	Architecture	Result
Mazars-Simon (2019)	Search, classify, identify and store data on sea turtles extracted in images extracted from Flickr	22,500 images	no inf.	multi-class	CNN	acc 0.95
Kulkarni and Di Minin (2021)	Identify news about endangered species	5,464 texts	no	binary	MurmurHash3 + MLP	F1 0.96
Edwards et al. (2022)	Identify wildlife observations in Twitter texts	2,798 text	yes	binary	BERT	F1 0.96
Carneiro et al. (2022)	Identify *Physalia physalis* in images extracted from Instagram	12,300 images	yes	binary	CNN pre-trained	F1 0.95

3.2 NLP WITH SOCIAL MEDIA DATA

In this Section, we list works not related to conservation science, which used NLP for text classification tasks in Brazilian Portuguese. Preference was given to works that used text with informal characteristics (e.g., social media) and small datasets. Our search was limited to works published from 2019 onwards, as it was the year of publication of BERT. The intention with this filter was to find works that used modern NLP techniques, since BERT is considered SOTA for text classification tasks.

Leite et al. (2020), with the aim of detecting the presence of toxic language in Twitter comments, experimented with different refinement approaches for the mBERT and BERTimbau models. The authors used two datasets, one in Portuguese with 19,000 tweets, and the other in English with 14,000 tweets, both unequally distributed between the positive and negative classes. As a baseline, the authors used BoW and an ensemble trained with the Portuguese dataset. The authors performed experiments with mBERT and BERTimbau refined with the Portuguese dataset, mBERT refined with the English dataset, and mBERT refined with the two datasets. As a result, most experiments reached F1 Score between 0.74 and 0.76, including the baseline. The exception was the experiment in which the mBERT was refined only with the English dataset, which obtained an F1 Score of 0.57. The experiments carried out by the authors showed that the results were significantly better (F1 Score between 0.75 and 0.76) when refining the BERT models with a dataset in Portuguese, compared to mBERT refined only with a dataset in English (F1 Score of 0.57). The authors also investigated the effects of the size of the training set and concluded that from 6 thousand examples it was possible to obtain a reliable performance for the task in question. Although the classes in the Portuguese dataset showed a slight imbalance, 44% of the examples were positive versus 56% negative, the authors do not inform whether they carried out any treatment to correct this imbalance.

González-Carvajal and Garrido-Merchán (2020), with the aim of comparing BERT with classic machine learning models, trained models to perform four different tasks. For Task 1, the authors used a dataset of 50,000 movies reviews in English, equally distributed between positive and negative classes. For this task, in addition to BERT, other classic models were trained, including: LR, SVM and NB, and an ensemble. For Task 2, a disaster dataset consisting of 10,000 tweets in English, unequally distributed between true and false classes, was used. For this task, in addition to BERT, an ensemble was trained. For Task 3, the authors used a dataset composed of 12,000 newspaper articles in Brazilian Portuguese, unequally distributed among nine categories (multiclass). For this task, in addition to BERT, an ensemble was trained. For Task 4, a dataset consisting of 6,000 hotel reviews in Chinese, equally distributed between positive and negative classes, was used. For this task, in addition to BERT, an ensemble was trained. In all experiments with classical models and ensemble the TF-IDF was used as a feature extractor.

In all scenarios experimented by the authors, BERT outperformed the classic models, and in the experiment using dataset in Portuguese, BERT obtained 0.90 of accuracy against 0.84 of the classic model. It should be noted that the authors carried out experiments with texts of different formality levels (i.e., newspaper articles, reviews and social media), distributed in different ways between classes, in different amounts and in different languages, offering empirical evidence about the superiority of BERT in different scenarios. The authors do not clarify which version of BERT was used in the experiments with data in Portuguese and Chinese. Therefore, the use of mBERT was considered. Furthermore, although two datasets showed imbalance between classes, the authors did not mention whether they performed any treatment to correct this imbalance.

Diniz et al. (2022) used data collected from Twitter to train machine learning models to detect suicidal thoughts. The authors experimented with mBERT, BERTimbau (large and base), SVM, MLP, RF, GB and Extra Trees. As a feature extractor for the classical and ensemble models, TF-IDF was used. The work involved a dataset containing aroud 3 thousand tweets, imbalanced between the positive and negative classes. The authors applied the SMOTE technique in order to balance the dataset, which resulted in a dataset of 3,400 examples for training. The best performance was obtained by the BERTimbau models (large and base) with F1 Score of 0.95.

Endo et al. (2022) trained and compared different models to detect fake news about Covid-19[3]. As a data source, 11,300 texts extracted from a fake news checking site and a news site linked to a television station were used. As the collected data were imbalanced between the positive and negative classes, the undersampling technique was used to balance the data, leaving the dataset with 2,094 news. The authors experimented with classic machine learning models: SVM, NB, RF and GB, and neural networks: LSTM, GRU, Bi-LSTM, Bi-GRU. For training the classic models they used BoW as a feature extractor and for neural networks they used WE. As a result, the best neural network was Bi-GRU with F1 Score of 0.94, while the best classical models were SVM and RF, both with F1 Score of 0.93.

Feitosa et al. (2022), used data collected from Twitter. Machine learning models were trained and compared in order to infer public opinion about police action in great repercussion security incidents. They experimented with mBERT, BERTimbau and the classic machine learning models: SVM and RF, fed by feature vectors extracted by mBERT and BERTimbau. A dataset containing about 4,400 tweets was used, imbalanced among three classes. As a result, the model that achieved the best performance was BERTimbau with F1 Score of 0.68. Furthermore, classical models trained with vector extracted with BERTimbau performed better compared to classical models trained with vector extracted with mBERT. The authors concluded that the results demonstrate that training models with texts in the domain and with a specific language are fundamental aspects for obtaining good results. In addition to the experiments with the imbalanced dataset, the authors also carried out experiments with the balanced dataset using the SMOTE technique, but only with the classic models. The authors did not present justification for not having performed experiments with balanced data as input for the BERT models. Therefore, due to the lack of experiments with balanced data for all models, we performed a clipping only of the results of the experiments with imbalanced data. In their experiments the authors used different sizes for mBERT (base) and BERTimbau (large). This may explain the difference in performance between them.

Table 3.2 shows a compilation of the architectures used in the works listed in this section. We only present the data for up to three of the best results of each work. Although some authors

[3]Covid-19 is an acute respiratory infection caused by the coronavirus SARS-CoV-2. https://www.gov.br/saude/pt-br/coronavirus

have performed experiments with data in other languages, this table presents the results of models trained only with data in Portuguese.

Table 3.2: Related Works - NLP with Social Media Data. Showing: Work, Problem, Dataset Size (Size), Balanced data (Balanced), Classification type (Type), Feature extraction method (FEM), Classifier and Result.

Work	Problem	Size	Balanced	Type	FEM	Classifier	Result
Leite et al. (2020)	Detecting the presence of toxic language in Twitter	19,000 tweets	no	binary	–	mBERT + pt	F1 0.75
					–	mBERT + pt + en	F1 **0.76**
					–	BERTimbau + pt	F1 **0.76**
González-Carvajal and Garrido-Merchán (2020)	Comparing BERT with classic machine learning models	12,000 news	no	multi-class	TF-IDF	ensemble	acc 0.84
					–	mBERT	acc **0.90**
Diniz et al. (2022)	Detect suicidal thoughts in tweets	3,400 tweets	SMOTE	binary	TF-IDF	RF	F1 0.88
					–	mBERT	F1 0.91
					–	BERTimbau	F1 **0.95**
Endo et al. (2022)	Detect fake news about Covid-19	2,094 news	under-sampling	binary	BoW	SVM and RF	F1 0.93
					WE	Bi-GRU	F1 **0.94**
Feitosa et al. (2022)	Infer public opinion about police action in great repercussion security incidents	4,400 tweets	no	multi-class	BERTimbau	RF	F1 0.61
					BERTimbau	SVM	F1 0.66
					–	BERTimbau	F1 **0.68**

3.3 MULTIMODAL MACHINE LEARNING WITH SOCIAL MEDIA DATA

In this Section, we list works not related to conservation science, which applied multimodal machine learning to classification tasks and used image and text. Preference was given to works that used text with informal characteristics (e.g., social media and memes).

Ofli et al. (2020) trained models to extract information about disaster occurrences from Twitter posts. They compared the results obtained with unimodal models as opposed to a multimodal model in two classification tasks: one binary, with a dataset of 12,700 tweets, and the other multiclass, with a dataset of 8,000 tweets. Both datasets have data imbalanced between classes. For textual data, a CNN was trained from scratch using WE as input. For the images, a VGG16 pre-trained in ImageNet was used, from which the authors adapted the last layer according to each classification task. For training the multimodal model, the representation layers produced by the CNNs were concatenated, which was used as input to a MLP. As a result, the multimodal models achieved the best results: F1 Score of 0.84 and 0.78, for tasks 1 and 2 respectively, followed by models trained only with images, which reached F1 Score of 0.83 and 0.76. In their experiments, the model trained only with the images outperformed the model trained only with the text. This could be due to the characteristics of the problem in question or it could be an inefficiency of the architecture used for text-only training. Note that the authors have not experimented with BERT, which is considered state-of-the-art for NLP tasks. We performed a clipping only of the experiments with binary classification to show in Table 3.3.

Kougia and Pavlopoulos (2021) trained and compared different architectures to detect hateful content in memes. For this, they used a dataset composed of 9,140 memes distributed in an imbalanced way between the positive and negative classes. Some of the architectures experimented by the authors were: 1) BERT refined with meme text and adapted for classification task; 2) BERT refined with meme text and image caption (generated through the Show and Tell [4]) and adapted for task classification; 3) Concatenation of feature vectors extracted using BERT (text) and pre-trained CNN (image), used as input to a neural network; 4) Concatenation of feature vectors extracted using BERT (text) and pre-trained CNN (image), used as input to a

[4]Neural network that generates image captions. https://arxiv.org/abs/1411.4555

classical model; 5) Concatenation of feature vectors extracted using WE (text) and pre-trained CNN (image), used as input to a classical model; 6) Combination, using average, the results of models 1 (text only) and 3 (text and image). As a result, the model resulting from experiment 6 obtained the best result, with F1 Score of 0.76, followed by the BERT model trained only with meme text (experiment 1) with F1 Score of 0.75. The CNN used in the experiments is a DenseNet121 pre-trained in ImageNet and were refined with the images dataset. The meme dataset is somewhat comparable to some Instagram posts that contain short texts. The authors performed three classification tasks, one binary and two multi-label[5]. We performed a clipping only of the experiments with binary classification to show in Table 3.3.

Table 3.3 shows a compilation of the architectures used in the works listed in this section. We only present the data for up to three of the best results of each work.

Table 3.3: Related Works - Multimodal with Social Media Data. Showing: Work, Problem, Dataset Size (Size), Balanced or imbalanced data (Balanced), Modal, Feature extraction and/or integration method (FEIM), Classifier and Result.

Work	Problem	Size	Balanced	Modal	FEIM	Classifier	Result
Ofli et al. (2020)	Extract information about disaster occurrences from Twitter	12,7000 tweets	no	text	WE	CNN	F1 0.80
				image		CNN pre-trained	F1 0.83
				multi	CNN+CNN	MLP	F1 **0.84**
Kougia and Pavlopoulos (2021)	Detect hateful content in memes	9,140 memes	no	meme text + image caption		BERT	F1 0.72
				meme text		BERT	F1 0.75
				multi	average	Combination	F1 **0.76**

3.4 DISCUSSION

Several works highlight the advantages and possibilities of using involuntary data for tasks related to conservation science, such as: Di Minin et al. (2015), Daume (2016), Ghermandi and Sinclair (2019), Toivonen et al. (2019), August et al. (2020), Jarić et al. (2020), Edwards et al. (2021), Morais et al. (2021). There are also a number of studies that apply machine learning to detect aspects of the environment, such as: estimating the number of animals of a given species (Foglio, 2019) and classifying land cover (ElQadi et al., 2020). We can also find studies that used involuntary data for the detection of wildlife, but without the use of automated tools (Dylewski et al., 2017; Taklis et al., 2020; Duailibe et al., 2021).

However, as already noted by Ghermandi and Sinclair (2019) and Edwards et al. (2022), there are few works that applied machine learning to involuntary data for wildlife classification tasks, such as Mazars-Simon (2019), Kulkarni and Di Minin (2021), Edwards et al. (2022) and Carneiro et al. (2022).

Although the works listed in Section 3.1 bear some resemblance to the study problem of this dissertation, only one of the models obtained by the authors can be used in the classification problem presented here: the model obtained by Carneiro et al. (2022). The others have characteristics and objectives different from the problem addressed in this dissertation. In Mazars-Simon (2019) the goal is to classify images of turtles. In Kulkarni and Di Minin (2021) the objective is to classify news as being or not about endangered species, which does not include *Physalia physalis*. In addition to the text used in the training being in English, the characteristics of the text itself, in this case news, are different from the characteristics of social media texts. In Edwards et al. (2022) the goal is to classify tweets as being or not about wildlife observations.

[5]Multi-label classification is a classification problem in which multiple labels can be assigned to the same instance.

Furthermore, the authors used only texts in English to train the models and there were no texts on *Physalia physalis* in their training database. In the case of Carneiro et al. (2022)'s work, the goal is to classify images of *Physalia physalis*. The model obtained by this work could be used in part of our experiments.

In their works, Edwards et al. (2022), Kulkarni and Di Minin (2021) and Toivonen et al. (2019) draw attention to the scarcity of works using text for tasks related to conservation science. The scarcity reported by the authors was also noticed during the searches for works related to this dissertation.

In the search for works that used NLP, it was not possible to find works using texts in Portuguese extracted from Instagram to identify wildlife. Thus, there are no references of which NLP architectures could work better for the problem addressed in this dissertation. Therefore, this work proposes the experimentation of different approaches for text classification.

Even though the classic models have not surpassed the SOTA in the works listed in Section 3.2, they still remain competitive. As can be seen in the results obtained by Endo et al. (2022) and Leite et al. (2020) the difference between the F1 Score obtained by the classical models and the neural network is only 1%. Therefore, this work proposes the training of classic machine learning models.

As a feature extractor, the TF-IDF is more interesting than the simple word count (BoW) because it considers the relevance of the words for generating the feature vector.

Considering the works listed in Section 3.2 and the performance presented by pre-trained BERT models, especially mBERT, this work also proposes the use of this model for training with textual data. It can be used both as a feature extractor and a classifier. The results obtained by the author Diniz et al. (2022) indicate that BERT can perform well even on small datasets.

We also searched for works that trained models using text and image for classification tasks (Section 3.3). As with the search for works using NLP, it was not possible to find works using texts in Portuguese and images extracted from Instagram to identify wildlife. Thus, there are no references of which multimodal machine learning architectures could work best for the problem addressed in this dissertation.

In the tasks tested by Ofli et al. (2020) and Kougia and Pavlopoulos (2021) the multimodal architectures performed better than the unimodal models. The results obtained by the authors reinforce the idea that the combined use of image and text can improve the performance of classification tasks.

Even so, in both works, the difference between the results of the best unimodal model and multimodal is only 1%. It shows that unimodal models are still competitive, depending on the problem, the architecture and size of the dataset used for training.

As we can see, the architectures for training multimodal models vary among the experiments made by the authors, and not always the concatenation of feature vectors as input for training multimodal models obtained the best result, as in the case of Kougia and Pavlopoulos (2021) whose combination of the results of two classifiers outperformed the other architectures.

In summary, as a main differential compared to works that propose to identify wildlife in social media data, we train models with the text of the post and also with one of the images in the post. We also combine the results from individual classifiers. Furthermore, we carried out a comparative analysis of the results obtained with training using only the text, only the image and the combination of classifiers.

4 DATASET CONSTRUCTION

This chapter describes how the data used for training the machine learning models were collected and labeled. This chapter also includes an exploratory analysis of the data. Its important to recall that our research interest is in occurrences of *Physalia physalis* on the Brazilian coast. Therefore, the dataset construction followed this premise.

4.1 DATA COLLECTION

In this work we used data extracted from Instagram. Instagram is a social media developed by Meta. Among its features is the sharing of media (i.e., images 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. Also, users can add hashtags in the caption. A hashtag can be written with text, emojis and numbers. Spaces and special characters will not work (Meta, 2023). Figure 4.1 shows an example of an Instagram post about a Portuguese man-of-war.

Figure 4.1: 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, they are not always added by the user and sometimes the location of the occurrence is informed in the caption or comments of the post;

- Short text. On average the collected posts have 80 tokens and median post size is 55 tokens, while the mode is 8 tokens;
- There are linguistic variations in the text. While some posts show correct use of grammar rules, social media text is informal in nature, with many grammatical errors, slang, abbreviations, neologisms, internet jargon, and language mixing. Furthermore, the use of compound hashtags (e.g., #caravelaportuguesa, #goprobrasil) generates words that do not exist in the lexicon;
- There are captions that do not make sense without the media, such as: "#caravelaportuguesa kkkkk" and "Dando continuidade ao desafio #Praia #Caravela #cnidario #dia2[1]". There are also posts that are formed almost exclusively by hashtags and/or emojis;
- There are posts in which common names of the species are used in contexts that are unrelated to the observation of the species. As an example: when searching using the hashtag #caravelaportuguesa, in addition to posts related to *Physalia physalis*, posts about boats, drawings, tattoos, clothes, jewelry and even restaurants are returned. This type of situation was also reported by Edwards et al. (2022) in his work. The terms "caravela portuguesa" and "água viva" represent a challenge in themselves, since they are polysemous terms. "Caravela portuguesa" can mean both the cnidarian *Physalia physalis* and a type of boat invented by the Portuguese during the Age of Discovery. The term "água viva" can mean either a cnidarian or refer to works such as Clarice Lispector's book[2] or the TV soap opera Água Viva[3] broadcast in 1980. The actual meaning of these terms depends on the context of the post, which can also be a source of ambiguity in very simple sentences like: "Eu vi uma caravela portuguesa"[4];
- There are posts that actually reference wildlife, but are not direct observations of wildlife. As an example, some posts warn about the risk that *Physalia physalis* offers to bathers, which does not represent an accident or a direct observation of the species. This type of situation was also reported by Edwards et al. (2022) in his work.

When a user has a public account and adds hashtags to a post, the post will be visible on the corresponding Hashtag Page (Meta, 2023). A Hashtag Page shows a *Top* section where the most popular posts tagged with the hashtag appear. The page also shows a *Recent* section, where the most recent posts tagged with the hashtag appear. In this section the posts appear in the order in which they were posted (Meta, 2023). Figure 4.2 shows a screenshot of the #caravelaportugesa page.

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 you should not expect to find posts that have the hashtag #cnidarios in the #cnidario page, for example.

The first option considered for obtaining Instagram data was the use of Instagram's API. However, this API does not return location metadata. So we chose to use the Instaloader (2023) library. This library uses scraping to get data from Hashtag Pages and returns data in .json format. The scripts for data collection and download were made in collaboration with Leonardo Camargo, an undergraduate student. More details on the data extraction process can be found in (Camargo et al., 2023).

[1]Translated as: "Continuing the challenge #Praia #Caravela #cnidario #dia2"
[2]https://en.wikipedia.org/wiki/Água_Viva_(novel)
[3]https://pt.wikipedia.org/wiki/Água_Viva_(telenovela)
[4]Translated as: "I saw a portuguese man-of-war"

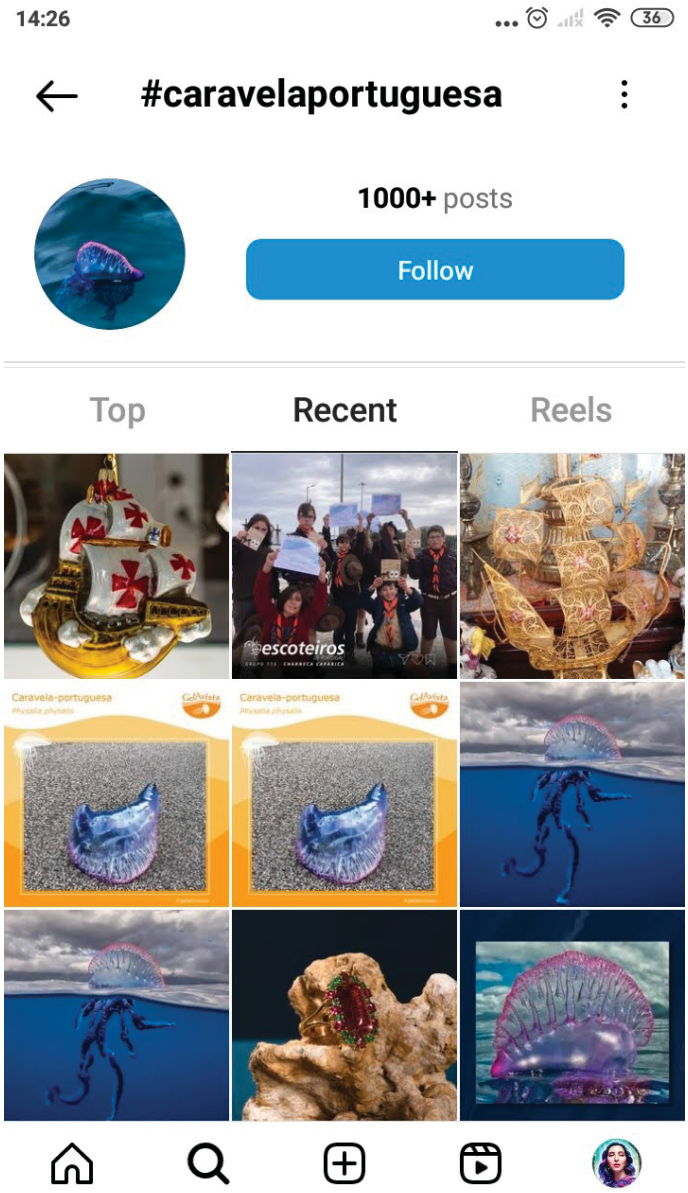


Figure 4.2: Screenshot of the #caravelaportugesa page taken by smart phone. Source: Instagram.

The data of interest for this research are: the caption of the post, media, posting date (timestamp) and location (latitude, longitude, city, location name and country). Figure 4.3 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 on 27/08/2022 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 Portuguese language and thus more assertive than #aguaviva. The result of the search for this hashtag on 27/08/2022 reached around 3,300 posts, which includes, in addition to *Physalia physalis*, boats, handicrafts, tattoos, clothes, among other things.

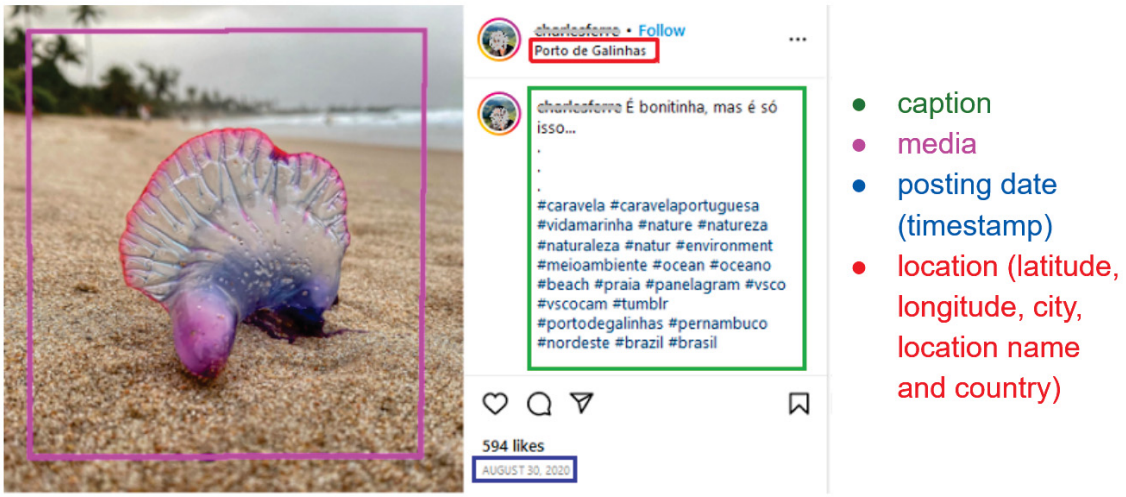


Figure 4.3: Illustration the data of interest for this research. In green the caption of the post, pink for media, blue for posting date and red for location. Source: adapted from Instagram.

The #physaliaphysalis, being the scientific name of the species, was chosen because it has a high potential for obtaining positive posts about *Physalia physalis*. A search for this hashtag on 27/08/2022 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 by being terms with medium potential of return of positive posts about *Physalia physalis*. A search for #cnidarios on 27/08/2022 returned around 3,400 posts which include *Physalia physalis* as well as other cnidarians such as *Velella velella* and medusa.

In all, we collected 6,204 posts dated between 2012 and 2021. Table 4.1 shows the number and period of posts collected per hashtag.

Table 4.1: Number and Period of Posts Collected per Hashtag

Hashtag	No. posts	Period	Interval
#aguaviva	1,680	07/2021 and 12/2021	175 days
#caravelaportuguesa	1,786	04/2012 and 05/2021	3,322 days
#cnidario	528	06/2012 and 03/2021	3,209 days
#cnidário	25	04/2013 and 02/2021	2,841 days
#cnidarios	1,766	12/2014 and 05/2021	2,355 days
#cnidários	230	05/2013 and 02/2021	2,824 days
#physaliaphysalis	189	01/2018 and 12/2021	1,439 days
Total	6,204		

After downloading, the data were saved in spreadsheets to facilitate the annotation process.

For posts that had geolocation (i.e., latitude and longitude) the BigDataCloud (2023) API was used to obtain additional location data (i.e., city, location name, state, country). This data was used as a complement to the existing metadata location.

The data were also enriched by identifying the caption language using the Langdetect (2023) library. 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 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 4.2 shows 3 instances of caption that the library did not detect the Portuguese language.

Table 4.2: Instances of caption that the Langdetect library did not detected 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%
Caravela #caravelaportuguesa #pirata #mergulho #surfer #surfers #cerveja #waves #goprobrasil #surfer #tattoo #style #goprohero7black #aloha #mahalo #freedom #brazilian #pirate #beach #paradise #liberdade #gopro #water #photographer #naturelovers #sea #photography #boatarde	English 85% and Italian 14%

Figure 4.4 shows a screenshot of a spreadsheet created with the downloaded and enriched data.

TIMESTAMP	LAT1	LONG1	COUNTRY1	CITY1	LOCNAME1	COUNTRY2	UF2	CITY2	LOCNAME2	IDIOMA	SHORTCODE	TEXTO	MIDIAS
2016-07-11_1	-14,4	-39,01		Itacaré	Praia de Jeribucaçu	BR	Bahia	Itacaré	Itacaré	pt	BHuL2ILAM2I	Para quem não conhece, apresentamos a vocês as Caravellas! Atenção sempre, no mar existe vida e vc é visita, respeite! #VemParaItacaré ---	
2016-07-01_1	-6,98	-34,83		Cabedelo	praia formosa-cabedelo	BR	Paraíba	Cabedelo	Monte Castelo	pt	BHVGJHCj1F0	⚠⚠⚠ATENÇÃO 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	
2016-06-03_(	39,7	-31,11			Corvo Island	PT	Região Autónoma dos Açores		Costa e Caldeirão - Ilha do Corvo	pt	BGL_sQSDVW	Photo by @lourencoortigao - #nofilterneeded #caravelaportuguesa #dangerous #naofujasnao #Regrann #AzoresWhatElse #Azores #Portugal	

Figure 4.4: Screenshot of a spreadsheet created with the downloaded and enriched data. Source: Author

4.2 DATA ANNOTATION

The data collected were manually annotated by the author and an oceanographer. The first attempt to annotate the data was made by the computer scientist assigning two multiclass labels to the posts, one considering only the caption and the other considering the caption and the media of the posts. Experiments with part of the dataset annotated in this way are published in Rocha and Hara (2022). But this approach failed when we tried to match the labels given by the specialist. The computer scientist evaluated the labels based on interpretation of the text, paying attention to hashtags, context, the presence of the species' common or scientific names, and examining images. The oceanographer, in addition to using a binary label, had three formal criteria to accept a post as a legitimate occurrence of *Physalia physalis* on the Brazilian coast. It must include the following information:

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

Because of this, the labels assigned by the computer scientist were abandoned. However, for the computer scientist, the lack of spatial and/or temporal information may not be important for a model that classifies posts based on their text and/or image. More than that, the dataset annotated by the specialist may introduce noise for training machine learning models.

Figure 4.5 illustrates the problem with noise, 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.

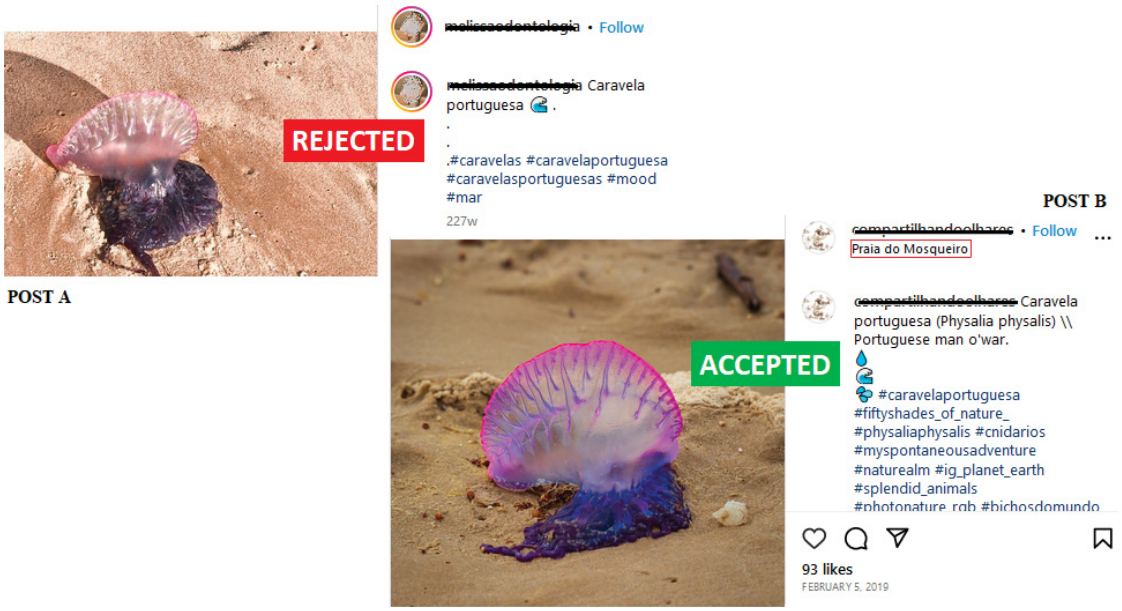


Figure 4.5: Illustration of the problem with using the original label. Both posts have an image of *Physalia physalis*, but post A does not have spatial information, while post B does. Following the expert's criteria, only post B is accepted as a legitimate occurrence of *Physalia physalis* on the Brazilian coast, while post A is rejected. Source: Author.

To overcome this problem, we decided to adapt the labels, considering as accepted posts rejected because spatial information. However, before assigning a binary label to the entire post,

we need to annotated 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 can help us reach a consensus on the binary labels, solve doubts and ambiguities, increase transparency and promote reproducibility of the annotation process.

It was the taxonomic and spatial information classes that allowed adapting the binary label to reduce noise and generate better classification models, as reported in experiments comparing original and adapted labels (Section 5.1 for textual analysis and 6.1 for image analysis). In addition, it makes it possible to carry out other experiments and studies using the enriched data, as an example: training a model to identify accidents with *Physalia physalis* using the taxonomic information.

Another hypothesis would be to exclude posts rejected because of location from the training dataset, however this would reduce the training dataset, which could compromise the results of the models. Therefore, the option of adapting labels is also a strategy that aims to use the entire available training base.

4.2.1 Spatial Criterion

For the spatial criterion, we evaluate whether the post contains spatial information, and in particular if the location is on the Brazilian coast.

Note that a post may have several metadata about location: latitude, longitude, city, location name and country. Besides, we enrich the metadata using latitude, longitude as we describe in Section 4.1, and sometimes the user informs 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 because 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 in a post that talks about a visit to the "Museu Paranense" located in Curitiba. It says: "Navegar é preciso #caravelaportuguesa #portugal #brasil #museuparanaense"[5]. 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 contains 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 contains metadata indicating that the post did not take place on the Brazilian coast, but still the location is in Brazil.

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

- FOREIGNER: The post contains 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 the coastal city, but it is clear that it is not on the beach, as in the examples: “Aquário Marinho do Rio de Janeiro” and “Centro Acadêmico de Vitória - CAV - UFPE”. These posts are classified as COUNTRYSIDE-TEXT when the information is part of the caption, and classified as COUNTRYSIDE-GEO when the information is part of the metadata.

4.2.2 Taxonomic Criterion

For the taxonomic criterion, we evaluate the posts media, especially if the images are photos of *Physalia physalis*, although containing 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 follow.

- 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 of *Physalia physalis*, 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 edits such as: 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 other cnidarians, 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.

4.2.3 Binary Label

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

4.2.3.1 ACCEPTED

To be accepted as a legitimate occurrence of *Physalia physalis* on the Brazilian coast, the post must to meet taxonomic and spatial criteria. It means that: the post must have at least one image classified as: REALISTIC or COLLECTION or EDITED or 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.

4.2.3.2 REJECTED

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

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

4.3 EXPLORATORY DATA ANALYSIS

In this section we make a descriptive analysis of the data annotated, with the goal of finding out the general aspects of the studied problem. To simplify the analysis, we joined 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.

Of the 6,204 posts collected, 151 are repeated one or more times. That is, they appear in the search results for more than one of the searched hashtags with the same identifier.

4.3.1 Distribution by Label

Table 4.3 shows the number of posts, considering the binary label.

Table 4.3: Number of posts per Label

Label	No. Posts	%
ACCEPTED	537	9%
REJECTED	5,667	91%
TOTAL	6,204	

Table 4.4 shows the number of posts per hashtag and label.

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.

Table 4.4: 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%

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 species in posts that have this hashtag.

The search for occurrences using the hashtag #caravelaportuguesa had the best result.

Table 4.5 shows the number of posts per rejection cause, while Table 4.6 shows the number of rejected posts per hashtag and rejection cause.

Table 4.5: Number of posts per rejection cause

Rejection Cause	No. Posts	%
MEDIA	531	9%
LOCATION	768	14%
MEDIA AND LOCATION	4,368	77%
TOTAL	5,667	

Table 4.6: Number of rejected posts per hashtag and rejection cause. The percentage is based on the number of rejected posts by hashtag.

Hashtag	Rejection Cause	No. Posts	%
#aguaviva (1,657)	MEDIA	164	10%
	LOCATION	15	<1%
	MEDIA AND LOCATION	1,478	89%
#caravelaportuguesa (1,364)	MEDIA	101	7%
	LOCATION	566	42%
	MEDIA AND LOCATION	697	51%
#cnidario (2,472)	MEDIA	265	11%
	LOCATION	85	3%
	MEDIA AND LOCATION	2122	86%
#physaliaphysalis (174)	MEDIA	1	1%
	LOCATION	102	59%
	MEDIA AND LOCATION	71	40%

Figure 4.6 shows a chart with the number of accepted posts by period. The chart shows that occurrences of *Physalia physalis* increase in the warmest periods of the year. Some authors such as: Ferreira-Bastos et al. (2017) and Aquino et al. (2019) correlate the increase of the number of accidents with *Physalia physalis* with the increase in tourism on the coast during summer breaks. This indicates that the data collected from Instagram reflects what is found in the literature about occurrences of *Physalia physalis*.

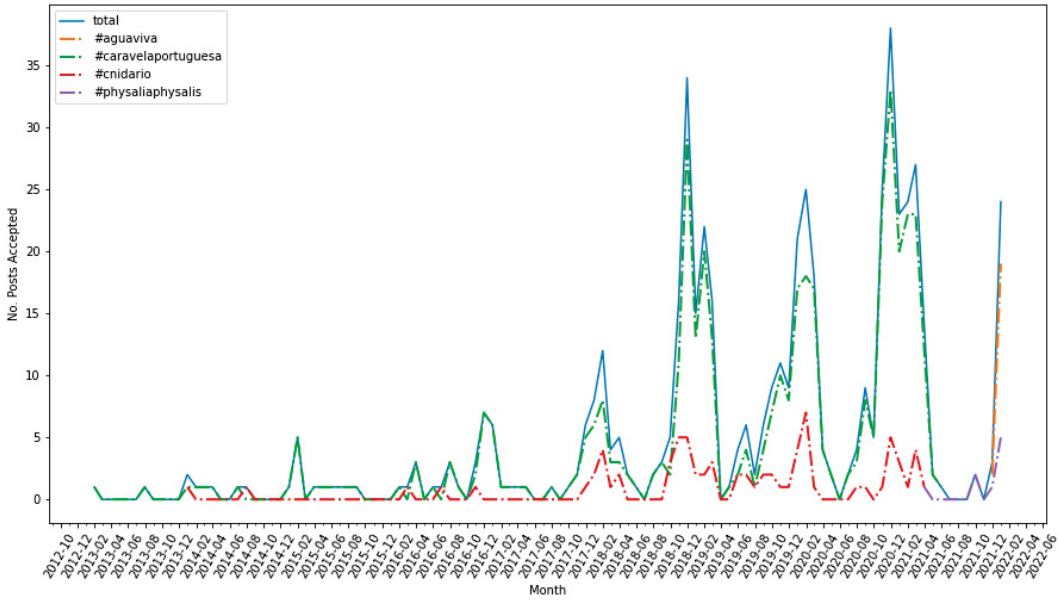


Figure 4.6: Number of Accepted Posts by Period. Source: Author

However, this chart should not be used to draw the conclusion that there was an increase in the number of occurrences over the time, since the chart shows an increase of occurrences from 2017. As shown in Table 4.1 and in the chart itself, the periods collected varied between the hashtags. Furthermore, as these are data from social networks, perhaps the increase is due to increased use of the network and not the occurrence of *Physalia physalis* itself.

4.3.2 Textual Information

In this subsection we explore the textual characteristics of the data collected. 12 (<1%) posts has empty caption.

4.3.2.1 Language

Table 4.7 shows the number of posts in the Portuguese language per hashtag.

Table 4.7: Number of posts in the Portuguese language per Hashtag. Showing: Hashtag, Number of posts collected (No. Collected), Number of posts in Portuguese language (No. Portuguese), Percentage of posts in Portuguese language (%).

Hashtag	No. Collected	No. Portuguese	%
#aguaviva	1,680	811	48%
#caravelaportuguesa	1,786	1,443	81%
#cnidario	2,549	1,561	61%
#physaliaphysalis	189	34	17%
Total	6,204	3,849	62%

Considering only posts identified as being located in Brazil (1,536), we found 43 posts in other languages, most in English (23 posts), followed by Spanish (7 posts). Figure 4.7 shows the distribution of posts by language considering only posts identified as being located in Brazil.

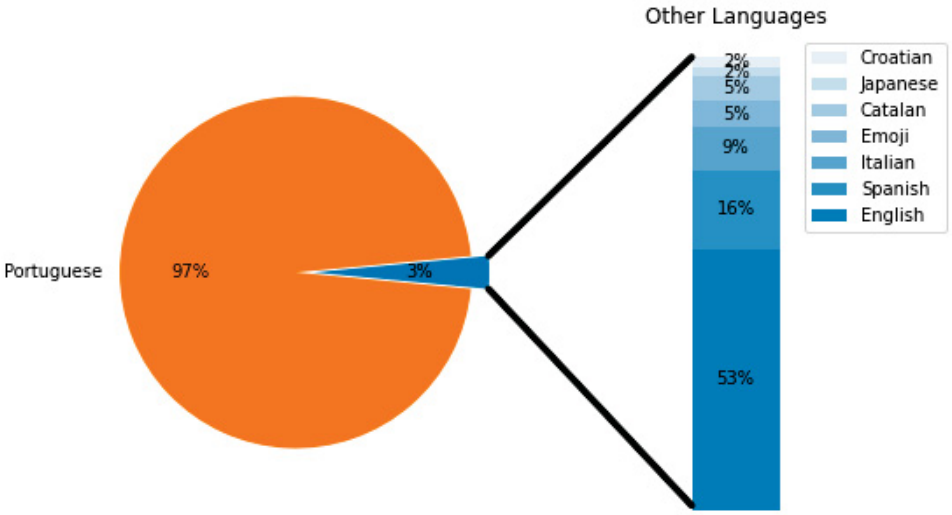


Figure 4.7: Distribution of posts by language. Considering only posts identified as being located in Brazil. Source: Author

4.3.2.2 Hashtags

6,025 (97%) posts have one or more hashtags in the caption. 458 (7%) posts have only hashtags in their caption. Here, we consider all hashtags, not only the ones used for searching and data collection. Table 4.8 shows the number of posts that have hashtags per label.

Table 4.8: Number of Posts that have Hashtags per Label

Label	No. Posts Collected	No. Posts Have Hashtag	%
ACCEPTED	537	517	96%
REJECTED	5,667	5,508	97%
TOTAL	6,204	6,025	97%

4.3.2.3 Emoji

2,705 (43%) posts have one or more emojis in the caption. 7 (<1%) posts have only emojis in the caption. Table 4.9 shows the number of posts that have emoji per label.

Table 4.9: Number of Posts that have Emoji per Label

Label	No. Posts Collected	No. Posts Have Emoji	%
ACCEPTED	537	235	43%
REJECTED	5,667	2,470	43%
TOTAL	6,204	2,705	43%

4.3.2.4 User Mentions

Instagram users can mention other users in the caption or comments. 1,250 (20%) posts have one or more mentions of other users in the caption. Table 4.10 shows the number of posts that have mentions of other users per label.

Table 4.10: Number of Posts that have User Mention per Label

Label	No. Posts Collected	No. Posts Have User Mention	%
ACCEPTED	537	69	12%
REJECTED	5,667	1,181	20%
TOTAL	6,204	1,250	20%

4.3.2.5 URLs

375 (6%) posts have one or more URLs in the caption. Table 4.11 shows the number of posts that have URL per label. Maybe the reason for this low number is because Instagram does not allow the use of links in captions or comments.

Table 4.11: Number of Posts that have URL per Label

Label	No. Posts Collected	No. Posts Have URL	%
ACCEPTED	537	17	6%
REJECTED	5,667	358	3%
TOTAL	6,204	375	6%

4.3.2.6 Numbers

2,486 (40%) posts have one or more numbers in the caption. These numbers can be contact numbers, prices, dates and others. Table 4.12 shows the number of posts that have numbers per label.

Table 4.12: Number of Posts that have Numbers per Label

Label	No. Posts Collected	No. Posts Have number	%
ACCEPTED	537	161	30%
REJECTED	5,667	2,325	41%
TOTAL	6,204	2,486	40%

4.3.2.7 Instagram Fonts

Some Instagram users have been using Instagram Fonts Generators, which are websites that generate Instagram-compatible Unicode glyphs[6]. These websites create "pseudo-alphabets" by taking advantage from Unicode symbols that look like the normal Latin alphabet, but have some differences, such as being bolder or italic, for example (IGfonts, 2023).

Figure 4.8 shows an example of post using Instagram Fonts.

Let's take as an example the word: 𝔭𝔥𝔶𝔰𝔞𝔩𝔦𝔞, which letter 𝔭 has the name "Mathematical Fraktur Small P"[7] which is part of "Mathematical Alphanumeric Symbols" Unicode block. While

[6]A glyph is the specific shape, design, or representation of a character. https://en.wikipedia.org/wiki/Glyph
[7]https://www.compart.com/en/unicode/U+1D52D

Figure 4.8: Example of post using Instagram Fonts. Source: Instagram

"normal" texts are writing using the symbols that belong to Basic Latin and/or Latin-1 Unicode blocks.

We identifid 73 (1%) posts that have Instagram Fonts. It is hard to identify them, because anything that we write uses a Unicode symbols. Thus we tried to identify text that forms words, but are not normalized. For this, we used the Unicodedata module from Python 3.

Table 4.13 shows the number of posts that have Instagram Fonts per label.

Table 4.13: Number of Posts that have Instagram Fonts per Label

Label	No. Posts Collected	No. Posts Have Instagram Fonts	%
ACCEPTED	537	8	1%
REJECTED	5,667	65	1%
TOTAL	6,204	73	1%

4.3.2.8 Similarity

888 (14%) posts have text with over 90% similarity with texts from other posts. In all, there are 337 texts that have a similarity above 90% with other texts. The number of similar texts ranges from 2 to 54. Some of these similar posts are reposts, others have a difference of a word or two or emojis. Sometimes they are identical.

Table 4.14 shows 3 examples of similar texts.

Table 4.14: Three Examples of Similar Texts

Text	No. Similar
#caravelaportuguesa	41
Title : Agua viva Serie: Ilha da Âncora Búzios Technical :photography Fineart prints available Visit my website: www.denisegrecophoto.com Ontem foi dia de mergulhar com a @buziosdivers , impecável. Super recomendo #aguaviva #artdeco #nikonus	12
Ao contrário do que muitas pessoas pensam, este animal não é uma água-viva e sim uma caravela, a qual é formada pela união física de vários animais. Estão tão unidos fisicamente que chegamos a pensar que se trata de um único animal e não uma colônia.	3

To obtain the similarity, we used the Levenshtein Python library (Levenshtein, 2023). The following preprocess was applied to the data before calculating the similarity: replacement of the emoji by its name in Portuguese, using the library Emoji (2023); removal of any non-word characters or non-space; lowercase text convertion and; removing the empty text.

Of the 337 similar texts, 14 do not have the same label among the similar ones. Table 4.15 shows similar texts that were labeled differently.

Table 4.15: Similar texts that was labeled differently. Showing: Number of Similar (S), Number of Accepted (A), Number of Rejected because of the Location (RL), Number of Rejected because of the Media (RM) and Number of Rejected because of the Media and Location (RML). Glyph that cannot be printed here (e.g., Instagram Fonts and Emojis) were replaced by UNK.

Text (up to 60 characters)	S	A	RL	RM	RML
#caravelaportuguesa	41	6	20	0	15
#caravelaportuguesa #mar	40	6	20	0	14
Ops... #caravelaportuguesa	39	5	20	0	14
#caravelaportuguesa #madeira	3	1	1	0	1
Caravela-Portuguesa (Portuguese Man-of-War)	3	1	1	0	1
#caravelaportuguesa#physaliaphysalis	3	1	2	0	0
A caravela-portuguesa (Physalia physalis) é o único organismo em c...	3	1	2	0	0
Dec 11, 2021 IG UNK DAY UNK Photo by @dudulacerdab Locati...	3	1	0	2	0
#portuguese #manowar #caravelaportuguesa	2	1	1	0	0
Esta foto sensacional de Matty Smith foi compartilhada por nossos c...	2	1	1	0	0
Água-viva!!! #caravelaportuguesa	2	1	0	0	1
Esse feed é para alertar banhistas e surfistas da Praia do Futuro. Tem...	2	1	1	0	0
#caravelaportuguesa #praia	2	1	1	0	0
Physalia physalis, popularmente conhecida como caravela... Praia do...	2	0	1	0	1

The problem with similar texts is that they may cause overfitting if they are in a representative number and, those that have been labeled differently may represent noise for training machine learning models.

4.3.3 Spatial Information

3,530 (57%) posts have some information that allows the identification of the location, whether in text or metadata. 2,648 (42%) posts have some location **metadata**. Table 4.16 shows the number of posts per type of spatial information.

Table 4.16: Number of posts per type of spatial information

Has Spatial info.	Spatial info. class	No. Posts	%
YES (3,530)	COAST-TEXT	667	10%
	COAST-GEO	401	6%
	COUNTRYSIDE-TEXT	145	5%
	COUNTRYSIDE-GEO	326	2%
	FOREIGNER	1,991	32%
NOT	NOTHING	2,674	43%

1,068 (17%) posts are identified as on the Brazilian coast. Their spatial information are classified as COAST-TEXT or COAST-GEO. Table 4.17 show the number of posts assigned as location on the Brazilian coast per Hashtag.

Table 4.17: Number of posts assigned as location on the Brazilian coast per Hashtag

Hashtag	No. Collected	No. Brazilian coast	%
#aguaviva	1,680	187	11%
#caravelaportuguesa	1,786	523	29%
#cnidario	2,549	342	13%
#physaliaphysalis	189	16	8%
Total	6,204	1,068	17%

4.3.4 Taxonomic Information

Table 4.18 shows the number of posts per type of taxonomic information. Note that a post can have up to 10 media and these may have been classified differently from each other.

Table 4.18: Number of posts per type of taxonomic information

Taxonomic information	No. Posts	%
REALISTIC	1089	17%
EDITED	164	2%
DISPLACED	27	<1%
CLOSE	9	<1%
COLLECTION	26	<1%
ACCIDENT	13	<1%
ART	139	2%
CNIDARIA	389	6%
VIDEO	817	13%
NOTHING	3944	63%

Table 4.19 shows the number of posts that have positive or negative taxonomic information for *Physalia physalis*.

Table 4.19: Number of posts that have positive or negative taxonomic information for *Physalia physalis*

Taxonomic info.	No. Posts	%
NEGATIVE	4,264	69%
POSITIVE	1,297	21%
ONLY VIDEO	643	10%

643 posts have only videos and as we mentioned before videos were not analyzed for the presence of *Physalia physalis*.

Table 4.20 shows the number of posts that have positive taxonomic information per Hashtag.

4.4 FILTERED DATA

In this section we present the filters applied to train the machine learning models and perform a descriptive analysis of the real data used to develop the models. The proportion of some data presented in the descriptive analysis (Section 4.3) remained the same after applying filters, in this case these data will not be repeated here.

The filters were designed with the aim of using the same dataset for training models regardless of the data modality (text or image), allowing it to be possible to compare the result of

Table 4.20: Number of posts that have positive taxonomic information per Hashtag

Hashtag	No. Posts	No. Positive	%
#aguaviva	1,680	35	2%
#caravelaportuguesa	1,786	981	55%
#cnidario	2,549	164	6%
#physaliaphysalis	189	117	62%
Total	6,204	1,297	20%

the final models. For training, validation and testing of machine learning models, we use the filtered data as follows:

- We deleted posts with empty text: 12 posts;
- We deleted repeated posts: 164 posts;
- We deleted posts with similar text: 488 posts. A threshold of 96% similarity was defined for posts to be deleted. The threshold was found after experiments with different strategies and threshold values. The experiments can be found in Section 5.2;
- We deleted posts with video only: 643 posts. It means posts that don't have image, just videos. This is because we chose to do standard image classification and leave video classification for future work;
- We deleted post with location outside of Brazil: 1,991 posts;
- We kept only posts identified as Portuguese language or emoji only in the caption: 3,856 posts. We kept emoji posts because emojis can be translated to Portuguese.

After applying the filters, **2,610** posts remained. Table 4.21 shows the number of remaining posts per hashtag.

Table 4.21: Number of remaining posts per Hashtag. %/Collected is percentage in relation to total posts collected. %/Remained is percentage in relation to total posts remained.

Hashtag	No. Collected Posts	No. Posts Remained	%/Collected	%/Remained
#aguaviva	1,680	664	10%	25%
#caravelaportuguesa	1,786	809	13%	31%
#cnidario	2,549	1,118	18%	43%
#physaliaphysalis	189	19	<1%	<1%
Total	6,204	2,610	42%	

Table 4.22 shows the number of remaining posts per hashtag and label.

4.4.1 Key Terms

We investigated the presence of the key terms for this research: "caravela portuguesa" and "physalia physalis", in the text of the posts. Table 4.23 show the number of remaining posts per key terms and label.

Table 4.22: Number of remaining posts per Hashtag and Label. The percentage is based on the number of posts by hashtag.

Hashtag	Label	No. Posts	%
#aguaviva	ACCEPTED	22	3%
	REJECTED	642	97%
#caravelaportuguesa	ACCEPTED	394	49%
	REJECTED	415	51%
#cnidario	ACCEPTED	42	4%
	REJECTED	1,076	96%
#physaliaphysalis	ACCEPTED	12	63%
	REJECTED	7	37%

Table 4.23: Number of remaining posts per Key Terms and Label. Showing Key Term, Number of posts that has the key term (No. Posts), Label, Number of posts that has the key term per label (No. Posts/Label) and the percentage of Number of posts that has the key term per label (%/Label)

Key Term	No. Posts	Label	No. Posts/Label	%/Label
caravela portuguesa	846	ACCEPTED	419	49%
		REJECTED	427	51%
physalia physalis	188	ACCEPTED	121	64%
		REJECTED	67	36%

4.4.2 Original Label vs Adapted Labels

As discussed in the beginning of the Section 4.2, posts rejected because of location may represent noise for training machine learning models. Thus, we decided to consider as accepted those rejected because spatial information. We call it adapted labels.

We consider three annotated datasets:

- `original`: original labels assigned by the oceanographer.
- `adapted-lack-location`: consider as ACCEPTED posts rejected only for lack of spatial information (i.e., spatial information equals NOTHING).
- `adapted-not-in-coast`: consider as ACCEPTED posts rejected because of spatial information (i.e., spatial information different from COAST-TEXT and COAST-GEO).

Table 4.24 shows the number of remaining posts per original label and adapted labels.

Table 4.24: 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%
TOTAL	2,610		2,610		2,610	

Table 4.25 shows the number of remaining rejected posts per reason for rejection.

Table 4.26 shows the number of remaining posts rejected because of location per type of spatial information.

Table 4.25: Number of remaining rejected posts rejected per reason

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

Table 4.26: 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

Experiments comparing the original label with the adapted labels can be seen in Sections 5.1 for textual analysis and 6.1 for image analysis. Despite the low number of rejected posts due to location, as we can see in Table 4.26, it already has an effect on the quality of the machine learning model obtained, as we reported in experiments comparing original and adapted labels.

4.4.3 Posts Size

After applying the filters, the posts sizes ranged from 1 to 440 tokens[8]. Figure 4.9 shows the distribution of posts according to the number of tokens. There is a higher concentration of posts with size between 5 and 40 tokens. This concentration is similar regardless of the posts label. On average the filtered posts have 78 tokens and median post size is 55 tokens, while mode is 20 tokens.

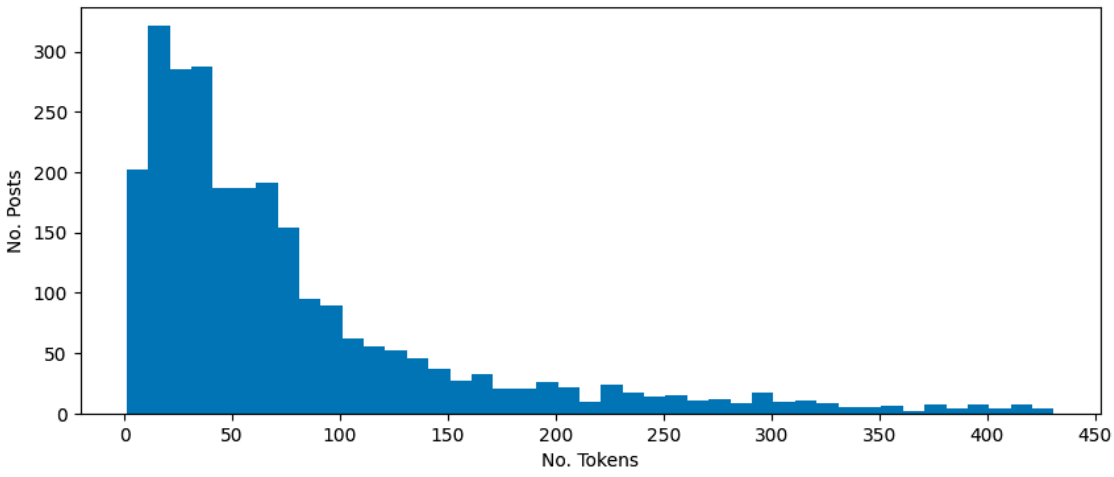


Figure 4.9: Distribution of posts per number of tokens. Source: Author

As mentioned in Subsection 2.5.4, BERT models use a tokenizer algorithm that can create sub-words and due to it the number of tokens can be greater than those created by word-based tokenizers (e.g., NLTK). After applying BERT tokenization, the posts sizes range from 1 to 679 tokens. Also, there are 54 posts that have more than 512 tokens.

[8]NLTK library (Bird et al., 2009) was used for tokenization.

Figure 4.10 shows the distribution of posts according to the number of tokens created by BERT. It is possible to notice that it is different from the chart where the data was tokenized using NLTK. There is a higher concentration of posts with size between 20 and 70 tokens.

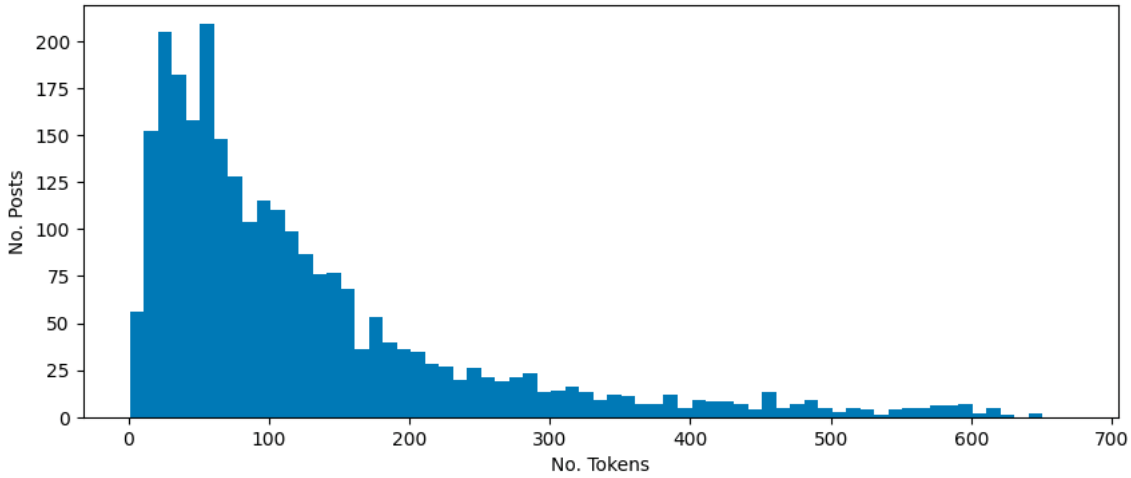


Figure 4.10: Distribution of posts per number of tokens created by BERT. Source: Author

In this Chapter we presented the process of constructing the dataset used in the experiments that will be presented in the next Chapters. We also presented an exploratory data analysis that allowed us to visualize in numbers the characteristics of the collected and annotated data, providing the basis for understanding the experiments applied and results obtained in this work. Part of the annotation process and exploratory analysis described in this Chapter is published in Rocha et al. (2023).

5 TEXTUAL ANALYSIS

In this Chapter we conducted several experiments using the caption of the posts. The main goal is to obtain a model capable to determine if the post is a legitimate occurrence of *Physalia physalis*, using only the caption of the posts, with high precision and F1 Score. Code and models developed in this Chapter are available online[1].

A second goal is to find out what kind of preprocessing techniques work best for our data, since there is no general guideline that works for all NLP problems.

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 for evaluating problems whose data are imbalanced and when there is a need for the number of false positives to be low (Kubat, 2017).

Furthermore, considering that the main contribution of this work is to obtain a model that 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 social media, 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 your project, and it will be considered as the main measure to decide between models.

Most experiments in this Chapter were performed with filtered data as described in Section 4.4, and divided 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, train set with 1279 samples and validation set with 548 samples. There are exceptions to this in some experiments. They are highlighted at the appropriate time.

In the experiments we trained a classic machine learning model available in the Scikit-learn library (Pedregosa et al., 2011): Logistic Regression. We also considered a state-of-the-art deep learning method for NLP tasks: mBERT[2], which is available in the Tensorflow Hub (Abadi et al., 2015).

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 on our data. For the majority of the experiments 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[3] provided by TensorFlow Hub. For other hyperparameters, we used the same ones used by Devlin et al. (2019). Furthermore, we used class weighting to handle data imbalance. There are exceptions to the hyperparameters used in some experiments. They are highlighted at an opportune moment.

For training with LR we kept its default parameters with the exception of the parameter `class_weight` which was set to `'balanced'`. This option automatically adjust 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.

[1] https://github.com/RESMA-PPGINF-UFPR-CAPES-PRINT/CaravelasTextualAnalysis

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

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

For TF-IDF, we kept the default values of the method. By default, this method converts data to lowercase, ignores tokens less than 2 characters long, emojis, punctuation and signs, with the exception of underline. At 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 (Subsection 2.5.4). There are exceptions to the parameters used in some experiments. They are highlighted at an opportune moment.

With the exception of the experiments in Section 5.1, experiments in this Chapter were performed with label adapted-lack-location (defined in Subsection 4.4.2).

In summary, the general workflow of textual analysis, illustrated in figure 5.1, begins with the construction of the dataset, which includes data collection, annotation and filtering. Then, the data is separated into test and development (Holdout). The cleaning and preparation step includes creating datasets for experiments with: similarity, post's size, text normalizations, as well as combining the normalizations. Two approaches are applied in the feature extraction step. As cross-validation is used, training and validation occur in 5 rounds alternating the validation set. In this step, the selection of label (original vs adapted), similarity threshold and normalizations is carried out. Hyperparameter optimization are made and class weighting are applied to deal with imbalance data. In the last step (test), once the best hyperparameters are known, the models are trained using the entire development set and evaluated on the test set.

5.1 ORIGINAL LABEL VS ADAPTED LABEL

As mentioned in Section 4.2, posts rejected only for lack of location can represent noise for training machine learning models.

To find out if there is any impact of considering as accepted posts rejected only for lack of spatial information, we trained two models with the posts' caption, one using `original` labels assigned by the specialist and the other with `adapted-lack-location` label that consider as ACCEPTED posts rejected only because of lack of spatial information.

The experiments in this Section were performed with filtered data with the exception of the filter that eliminates similar posts. The results obtained are presented in Table 5.1.

Table 5.1: Results of Experiments with Adapted Label and 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	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)

Despite the low number of rejected posts because of lack of location, as presented in Table 4.26, the results with the adapted labels are significantly higher than results with the original labels. For this reason, we adopted the adapted-lack-location label for the rest of the experiments in this Chapter. As spatial information is not a feature used for training the models in this work, it is safe to use the adapted label in the experiments.

The reason for only considering one of the adapted labels (adapted-lack-location) in this Chapter is because the idea of using the adapted-not-in-coast label arose during the image experiments. Experiments with text and adapted-not-in-coast label is left for future work.

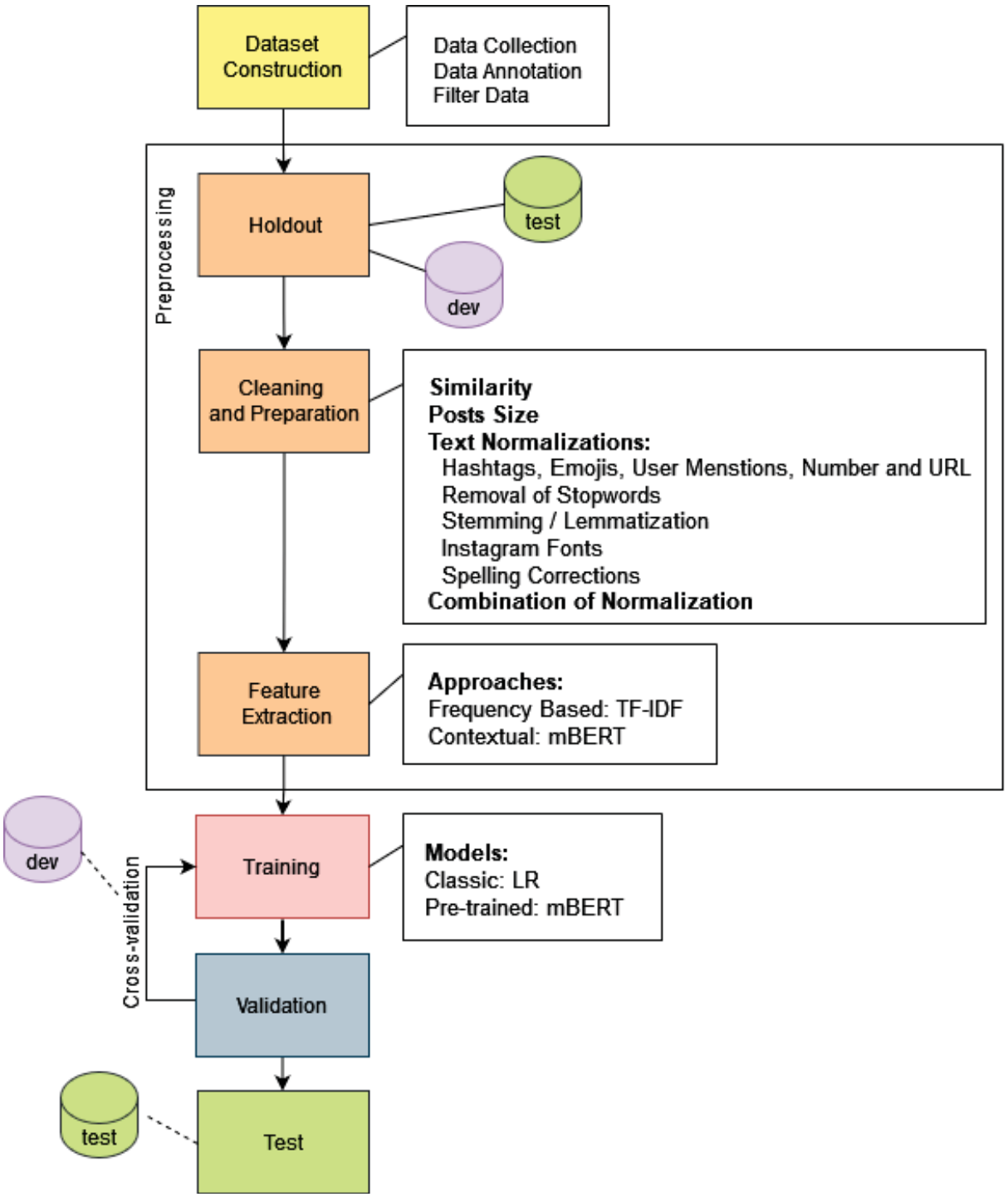


Figure 5.1: Textual Analysis Workflow with an overview of techniques, approaches, and models that we used in textual analysis. Source: the Author

5.2 SIMILARITY

As mentioned in Subsection 4.3.2.8, we have a number of similar posts and, in addition, some of them do not have the same label (label conflict). The problem with similar texts is that they may cause overfitting if they are in a representative number, while those that have been labeled differently may represent noise for training machine learning models. Edwards et al. (2022) in his work removed similar tweets in order to avoid the overfitting.

To find out if there is any impact of removing similar posts from the dataset, we performed the following experiments:

1. Strategy 1: removing of all similar posts except one, randomly selected. In case of label conflict: the one whose class occurs more frequently is selected and in case of a tie all are removed.

2. Strategy 2: removing of all similar posts except one, randomly selected. In case of label conflict: all are removed.

The experiments were performed with filtered data with the exception of the filter that eliminates similar posts. TF-IDF + LR and a similarity threshold ranging from 75% to 100% were adopted. We used the results of the trained model with adapted-lack-location label as the baseline (see Table 5.1). The results obtained are presented in Tables 5.2 and 5.3.

Table 5.2: Results of Similarity Experiments - Strategy 1 - TF-IDF + LR. Showing Precision, Recall and F1 Score with their respective standard deviation

Similarity	Precision	Recall	F1	Training Size
baseline	0.791 (0.009)	0.899 (0.027)	0.842 (0.015)	1,367
≥75%	0.793 (0.018)	0.902 (0.024)	0.843 (0.009)	1,178
≥80%	0.780 (0.033)	0.917 (0.018)	0.842 (0.019)	1,197
≥85%	0.787 (0.049)	0.927 (0.022)	0.850 (0.027)	1,216
≥90%	0.771 (0.020)	0.896 (0.035)	0.829 (0.024)	1,246
≥91%	0.785 (0.027)	0.882 (0.045)	0.829 (0.019)	1,250
≥92%	0.767 (0.029)	0.916 (0.042)	0.834 (0.026)	1,255
≥93%	0.778 (0.017)	0.906 (0.006)	0.837 (0.011)	1,259
≥94%	0.769 (0.024)	0.899 (0.027)	0.828 (0.020)	1,265
≥95%	0.786 (0.024)	0.911 (0.032)	0.843 (0.017)	1,274
≥96%	0.810 (0.027)	0.914 (0.025)	0.859 (0.018)	1,278
≥97%	0.793 (0.031)	0.915 (0.008)	0.849 (0.018)	1,285
≥98%	0.766 (0.011)	0.906 (0.021)	0.830 (0.006)	1,292
≥99%	0.792 (0.034)	0.926 (0.020)	0.853 (0.021)	1,304
≥100%	0.813 (0.018)	0.916 (0.020)	0.861 (0.011)	1,311

The goal of removing similar posts is to reduce overfitting and noise. Although results with 100% similarity are the best, removing only what is 100% similar leaves the database still noisy. Therefore, the results show that 96% presents a good balance between performance and noise elimination.

It can be observed that there was no difference in the results of strategy 1 and 2 from 95% of similarity forward. This likeness is due to the other filters, which when applied result in equal datasets.

To find out if there is any impact of removing similar posts from the dataset in mBERT + LR and mBERT models, we performed experiments by removing posts with 96% and 100% of similarity from the dataset. The results obtained are presented in Tables 5.4 and 5.5 respectively.

Analysing the results we can see that removing posts with 96% of similarity also showed good results for these 2 models.

Table 5.3: Results of Similarity Experiments - Strategy 2 - TF-IDF + LR. Showing Precision, Recall and F1 Score with their respective standard deviation

Similarity	Precision	Recall	F1	Training Size
baseline	0.791 (0.009)	0.899 (0.027)	0.842 (0.015)	1,367
≥75%	0.780 (0.040)	0.921 (0.025)	0.844 (0.026)	1,176
≥80%	0.734 (0.021)	0.893 (0.019)	0.806 (0.019)	1,197
≥85%	0.776 (0.022)	0.904 (0.016)	0.835 (0.011)	1,215
≥90%	0.787 (0.013)	0.919 (0.024)	0.848 (0.013)	1,245
≥91%	0.790 (0.027)	0.913 (0.021)	0.847 (0.018)	1,249
≥92%	0.766 (0.015)	0.905 (0.019)	0.830 (0.009)	1,254
≥93%	0.794 (0.016)	0.902 (0.032)	0.844 (0.016)	1,258
≥94%	0.758 (0.017)	0.929 (0.016)	0.835 (0.013)	1,265
≥95%	0.786 (0.024)	0.911 (0.032)	0.843 (0.017)	1,274
≥96%	0.810 (0.027)	0.914 (0.025)	0.859 (0.018)	1,278
≥97%	0.793 (0.031)	0.915 (0.008)	0.849 (0.018)	1,285
≥98%	0.766 (0.011)	0.906 (0.021)	0.830 (0.006)	1,292
≥99%	0.792 (0.034)	0.926 (0.020)	0.853 (0.021)	1,304
≥100%	0.813 (0.018)	0.916 (0.020)	0.861 (0.011)	1,311

Table 5.4: Results of Similarity Experiments - Strategy 1 - mBERT + LR. Showing Precision, Recall and F1 Score with their respective standard deviation

Similarity	Precision	Recall	F1	Training Size
baseline	0.590 (0.020)	0.800 (0.022)	0.679 (0.019)	1,367
≥ 96%	0.623 (0.016)	0.781 (0.008)	0.693 (0.013)	1,278
≥ 100%	0.657 (0.024)	0.790 (0.056)	0.716 (0.029)	1,311

Table 5.5: Results of Similarity Experiments - Strategy 1 - mBERT. Showing Precision, Recall and F1 Score with their respective standard deviation

Similarity	Precision	Recall	F1	Training Size
baseline	0.823 (0.041)	0.886 (0.046)	0.852 (0.014)	1,367
≥ 96%	0.831 (0.031)	0.900 (0.037)	0.863 (0.018)	1,278
≥ 100%	0.859 (0.034)	0.824 (0.023)	0.841 (0.013)	1,311

5.3 POSTS' SIZE

In the search for related works, we found the works of Forte Martins et al. (2021) and Feitosa et al. (2022) which, in the preprocessing step, excluded texts shorter than 5 and 2 words respectively.

In the Cabral. et al. (2021)'s work, a model trained only with texts with more than 50 words performed significantly better than models trained with texts without this restriction.

To understand whether the posts' size of the training sample could influence the performance of the models, we performed the following experiments only with TF-IDF + LR model:

- Exclusion of texts with less than 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40 and 50 tokens from the training dataset;
- Exclusion of texts with less than 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40 and 50 tokens from the training dataset. However, limiting the dataset size to 650 posts, in order to avoid

biases related to the decrease in the training size as the restriction on the size of the posts increases.

Observe that only the training dataset had posts deleted. The validation dataset remained unchanged.

As baseline, we trained models with raw text. The results obtained are presented in Table 5.6.

Table 5.6: Results of Experiments with Posts' Size. Showing Precision, Recall and F1 Score with their respective standard deviation. Size2+ means that posts smaller than 2 tokens were excluded from the training dataset, size3+ posts smaller than 3 tokens were excluded and so on.

Treatment	Precision	Recall	F1	Training Size
Unlimited				
baseline	0.809 (0.027)	0.913 (0.028)	0.858 (0.019)	1,278
size2+	0.811 (0.026)	0.913 (0.028)	0.859 (0.020)	1,276
size3+	0.811 (0.025)	0.914 (0.025)	0.859 (0.018)	1,274
size4+	0.807 (0.023)	0.916 (0.019)	0.858 (0.014)	1,270
size5+	0.805 (0.021)	0.920 (0.018)	0.859 (0.014)	1,253
size6+	0.805 (0.021)	0.919 (0.017)	0.858 (0.013)	1,249
size7+	0.801 (0.021)	0.916 (0.017)	0.854 (0.010)	1,232
size8+	0.799 (0.020)	0.916 (0.017)	0.853 (0.011)	1,222
size9+	0.803 (0.022)	0.916 (0.017)	0.855 (0.012)	1,206
size10+	0.804 (0.022)	0.917 (0.017)	0.857 (0.011)	1,197
size20+	0.800 (0.021)	0.906 (0.025)	0.849 (0.014)	1,032
size30+	0.819 (0.019)	0.894 (0.027)	0.855 (0.016)	896
size40+	0.831 (0.023)	0.877 (0.023)	0.853 (0.015)	763
size50+	0.843 (0.021)	0.849 (0.023)	0.846 (0.007)	666
Limited to 650 posts				
baseline	0.790 (0.018)	0.909 (0.034)	0.845 (0.016)	650
size2+	0.790 (0.018)	0.909 (0.034)	0.845 (0.016)	650
size3+	0.790 (0.018)	0.909 (0.034)	0.845 (0.017)	650
size4+	0.791 (0.019)	0.910 (0.028)	0.846 (0.015)	650
size5+	0.793 (0.024)	0.916 (0.022)	0.850 (0.017)	650
size6+	0.793 (0.023)	0.916 (0.022)	0.850 (0.017)	650
size7+	0.790 (0.021)	0.914 (0.024)	0.848 (0.017)	650
size8+	0.791 (0.020)	0.916 (0.023)	0.849 (0.017)	650
size9+	0.793 (0.021)	0.913 (0.024)	0.848 (0.018)	650
size10+	0.792 (0.020)	0.914 (0.030)	0.849 (0.021)	650
size20+	0.796 (0.026)	0.906 (0.030)	0.847 (0.020)	650
size30+	0.807 (0.026)	0.890 (0.028)	0.846 (0.015)	650
size40+	0.830 (0.030)	0.874 (0.019)	0.851 (0.009)	650
size50+	0.843 (0.022)	0.849 (0.024)	0.846 (0.010)	650

Considering the F1 Score, the results showed that there was no significant difference between the experiments. However, by excluding posts smaller than 40 tokens (size40+) from the training sample, the precision has improved, while recall has decreased.

5.4 TEXT NORMALIZATION

Text extracted from social media have distinct characteristics, such as: use of hashtags, emojis, neologisms, mix of languages and informal writing. These features increase the challenge of determining which normalization procedures should be applied during the preprocessing step in order to maximize the performance of the classifiers. Furthermore, each treatment consumes time and processing energy. Thus, it is important to determine which ones may in fact affect the classification result. This is especially important when the classification model is continuously used in real applications, on new incoming text.

We looked for references about what types of normalizations could be applied to data in Portuguese extracted from social media. We came across many NLP works applying different treatments, usually in sequence, as in the works of Mota et al. (2021) and Diniz et al. (2022).

However, we found few works that compare the impact of preprocessing applied to machine learning models (Dos Santos and Ladeira (2014); Stiilpen Junior and Merschmann (2016); de Oliveira and Merschmann (2021); Cabral. et al. (2021)). In fact, there are no general guidelines to follow in order to determine which techniques to apply. In the book of Jurafsky and Martin (2021), the authors state that before almost any NLP task, the text must be normalized. But they leave it up to the developer to choose which treatments to apply.

In this Section we conducted experiments to determine the effect of the normalization techniques on the performance of models trained for our problem. We performed experiments with different techniques and evaluated their performance in comparison to models trained with raw data.

Some of the treatments applied in our experiments resulted in some empty samples. In that case, the sample in question was removed from the experiment. As an example, TF-IDF ignores emojis. Thus, in the case of a caption is composed only of emojis, the vectorization process with this method results in a vector composed only of zeros. In other words, in the experiment with raw text this zeroed sample was kept, while in the experiment with emoji removal this sample was excluded.

The transformations considered for text normalization are described in the next subsections. They have been chosen either because they consider the distinct features of our dataset or because they are commonly applied in NLP works.

5.4.1 Hashtags

As mentioned in Subsection 4.3.2.2, 97% of posts have one or more hashtags in the caption, and this percentage is practically the same for the ACCEPTED and REJECTED classes.

To determine whether hashtag-related treatments improve the performance of the models, we considered 3 different ways to normalize the text:

1. **Remove hashtag**: removal of all hashtags from the text;
2. **Tokenize hashtag**: replacement of hashtags by a `hashtag` token. The authors González-Carvajal and Garrido-Merchán (2020) and Edwards et al. (2022) also applied this type of normalization in their works;
3. **Only hashtag**: keeping only hashtags in the text.

The **only hashtag** and **remove hashtag** treatments resulted in empty samples, thus these samples were removed from experiments.

The results obtained are presented in Table 5.7.

Table 5.7: Results of Experiments with Normalizations Related to Hashtags. Showing: Precision, Recall and F1 Score with their respective standard deviation. Vocabulary size (V Size) generated by TF-IDF. Dev database size resulting after apply the normalization (D Size).

Treatment	TF-IDF + LR				mBERT + LR			mBERT			D Size
	Precision	Recall	F1	V Size	Precision	Recall	F1	Precision	Recall	F1	
raw text	0.809 (0.027)	0.913 (0.028)	0.858 (0.019)	17,485	0.649 (0.015)	0.801 (0.020)	0.717 (0.013)	0.835 (0.021)	0.883 (0.029)	0.858 (0.012)	1,827
remove	0.800 (0.050)	0.715 (0.043)	0.754 (0.030)	12,024	0.504 (0.018)	0.699 (0.029)	0.586 (0.017)	0.700 (0.038)	0.734 (0.048)	0.715 (0.012)	1,736
tokenize	0.677 (0.030)	0.830 (0.030)	0.746 (0.027)	12,065	0.518 (0.023)	0.690 (0.039)	0.591 (0.027)	0.676 (0.022)	0.772 (0.055)	0.720 (0.020)	1,827
hashtag only	0.742 (0.048)	0.901 (0.027)	0.813 (0.029)	6,707	0.646 (0.035)	0.798 (0.054)	0.714 (0.039)	0.807 (0.069)	0.841 (0.062)	0.820 (0.036)	1,760

Considering F1 Score, it is noted that none of the models trained with normalized data outperformed the model trained with raw data. Although they did not outperform the model trained with raw data, the models trained only with hashtags performed better than models trained with other treatments. In addition, the absence of hashtags significantly worsened the performance of the models, which demonstrates that hashtags have a good discriminative power for our research problem, and should be maintained.

5.4.2 Emojis

As mentioned in Subsection 4.3.2.3, 43% of posts have one or more emojis in the caption, and this percentage is the same for the ACCEPTED and REJECTED classes.

To understand if emoji-related treatments improve the performance of the models, we considered 3 ways to normalize the text:

1. **Remove emoji**: removal of all emojis from the text;
2. **Tokenize emoji**: replacement of emoji by an `emoji` token;
3. **Translate emoji**: translate the emoji to its meaning in Portuguese, using the Emoji (2023) library.

The **remove emoji** treatment resulted in empty samples, thus these samples were removed from experiments.

The results obtained are presented in Table 5.8.

Table 5.8: Results of Experiments with Normalizations Related to Emojis. Showing: Precision, Recall and F1 Score with their respective standard deviation. Vocabulary size (V Size) generated by TF-IDF. Dev database size resulting after apply the normalization (D Size).

Treatment	TF-IDF + LR				mBERT + LR			mBERT			D Size
	Precision	Recall	F1	V Size	Precision	Recall	F1	Precision	Recall	F1	
raw text	0.809 (0.027)	0.913 (0.028)	0.858 (0.019)	17,485	0.649 (0.015)	0.801 (0.020)	0.717 (0.013)	0.835 (0.021)	0.883 (0.029)	0.858 (0.012)	1,827
remove	0.783 (0.021)	0.924 (0.025)	0.848 (0.022)	17,489	0.637 (0.021)	0.818 (0.034)	0.716 (0.023)	0.845 (0.043)	0.876 (0.043)	0.859 (0.027)	1,823
tokenize	0.806 (0.024)	0.910 (0.021)	0.855 (0.015)	17,485	0.636 (0.023)	0.800 (0.024)	0.708 (0.020)	0.841 (0.032)	0.875 (0.050)	0.856 (0.022)	1,827
translate	0.810 (0.023)	0.906 (0.023)	0.855 (0.015)	17,648	0.631 (0.007)	0.791 (0.018)	0.702 (0.008)	0.825 (0.031)	0.896 (0.030)	0.858 (0.017)	1,827

Considering F1 Score, the majority of models trained with normalized data showed no significant difference in performance compared to the model trained with raw text. While 3 models showed worse performance when trained with normalized texts.

5.4.3 User Mentions, Numbers and URLs

Replacement and removal of user mentions, numbers and URLs are common normalization processes found in the literature. Even if these elements do not have a large number of occurrences in our data, as we presented in Subsections 4.3.2.4, 4.3.2.5 and 4.3.2.6, we wanted to determine if treatments related to them can improve the performance of the model. Thus we considered the following transformations:

1. **Remove mention**: removal of user mentions from the text. The authors Vargas et al. (2021) and de Oliveira and Merschmann (2021) also applied this type of normalization in their works;

2. **Tokenize mention**: replacement of user mentions by an `user` token. The authors Cheema et al. (2021) and Edwards et al. (2022) also applied this type of normalization in their works;

3. **Remove URL**: removal of URLs from the text. The authors Ofli et al. (2020), de Oliveira and Merschmann (2021), Diniz et al. (2022) and Feitosa et al. (2022) also applied this type of normalization in their works;

4. **Tokenize URL**: replacement of URLs by an `URL` token. The authors Cabral. et al. (2021) and Cheema et al. (2021) also applied this type of normalization in their works;

5. **Remove number**: removal of numbers from the text. The authors Dos Santos and Ladeira (2014), Ofli et al. (2020) and Diniz et al. (2022) also applied this type of normalization in their works;

6. **Tokenize number**: replacement of numbers by a `number` token.

The results obtained are presented in Table 5.9.

Table 5.9: Results of Experiments with Normalizations Related to User Mention, Numbers and URLs. Showing: Precision, Recall and F1 Score with their respective standard deviation. Vocabulary size (V Size) generated by TF-IDF. Dev database size resulting after apply the normalization (D Size).

Treatment	TF-IDF + LR				mBERT + LR			mBERT			D Size
	Precision	Recall	F1	V Size	Precision	Recall	F1	Precision	Recall	F1	
raw text	0.809 (0.027)	0.913 (0.028)	0.858 (0.019)	17,485	0.649 (0.015)	0.801 (0.020)	0.717 (0.013)	0.835 (0.021)	0.883 (0.029)	0.858 (0.012)	1,827
User mentions											
remove	0.809 (0.027)	0.913 (0.028)	0.858 (0.019)	17,078	0.634 (0.017)	0.784 (0.017)	0.701 (0.015)	0.842 (0.041)	0.891 (0.022)	0.865 (0.016)	1,827
tokenize	0.810 (0.024)	0.912 (0.026)	0.858 (0.018)	17,080	0.640 (0.012)	0.799 (0.031)	0.710 (0.015)	0.819 (0.046)	0.923 (0.035)	0.867 (0.027)	1,827
Numbers											
remove	0.806 (0.026)	0.917 (0.021)	0.858 (0.016)	17,043	0.644 (0.022)	0.804 (0.020)	0.715 (0.016)	0.852 (0.027)	0.899 (0.025)	0.875 (0.018)	1,827
tokenize	0.820 (0.022)	0.913 (0.025)	0.864 (0.019)	17,139	0.647 (0.014)	0.797 (0.020)	0.714 (0.012)	0.846 (0.040)	0.903 (0.039)	0.873 (0.022)	1,827
URLs											
remove	0.812 (0.027)	0.914 (0.025)	0.860 (0.019)	17,333	0.645 (0.018)	0.799 (0.026)	0.714 (0.017)	0.823 (0.041)	0.906 (0.033)	0.861 (0.018)	1,827
tokenize	0.812 (0.027)	0.914 (0.025)	0.860 (0.019)	17,335	0.644 (0.015)	0.801 (0.022)	0.714 (0.015)	0.852 (0.023)	0.864 (0.033)	0.857 (0.017)	1,827

Considering F1 Score, for LR models, removal or tokenization of users, numbers and urls did not present significant difference in the results compared to the model trained with raw

data, with the exception of user removal which worsened the performance of the mBERT+LR model.

Conversely, mBERT model benefited from some treatments, particularly number-related and user-related normalizations that outperformed the model trained with raw data, with the exception of url-related normalizations, which did not show a significant difference.

An explanation for user-related and number-related treatments to outperform raw text in mBERT models, may be related to the subwords created by mBERT in the data tokenization process (Section 2.5.4). As an example take the user `@helofernanda`. When tokenized it becomes `['@', 'hel', '##of', '##erna', '##nda']`, while the token `user` becomes `['user']`. The same happens with numbers, as an example take the cell phone number: `(41)99946-4575`, when tokenized it becomes `['(', '41', ')', '999', '##46', '-', '457', '##5']`, while the token `number` becomes `['number']`. The same happens with URLs, but the explanation for the not so good performance in the latter case could be the number of occurrences of URLs in our data, only 6% of the text has URLs, against 20% that have mentions of users and 40% that have numbers.

Even if they did not improve the performance in the case of training with LR, these treatments can be interesting to reduce the size of the vocabulary.

5.4.4 Lemmatization and Stemming

Preprocess aimed at vocabulary reduction, like Lemmatization (Subsection 2.5.1.4) and Stemming (Subsection 2.5.1.3), are common in NLP works. To understand the effectiveness of using these treatments to train the models for our problem, 2 experiments were carried out:

1. **Lemmatization**: deflecting the caption words to their lemma. We used spaCy library (Honnibal et al., 2020) for this treatment. The authors Stiilpen Junior and Merschmann (2016), Cabral. et al. (2021), Forte Martins et al. (2021) and Vargas et al. (2021) also applied this type of normalization in their works;

2. **Stemming**: switching the caption words to their root. We used NLTK library (Bird et al., 2009) for this treatment. The author Dos Santos and Ladeira (2014), Stiilpen Junior and Merschmann (2016), de Oliveira and Merschmann (2021) and Diniz et al. (2022) also applied this type of normalization in their works.

The results obtained are presented in Table 5.10.

Table 5.10: Results of Experiments with Normalizations Related to Lemmatization and Stemming. Showing: Precision, Recall and F1 Score with their respective standard deviation. Vocabulary size (V Size) generated by TF-IDF. Dev database size resulting after apply the normalization (D Size).

Treatment	TF-IDF + LR				mBERT + LR			mBERT			D Size
	Precision	Recall	F1	V Size	Precision	Recall	F1	Precision	Recall	F1	
raw text	0.809 (0.027)	0.913 (0.028)	0.858 (0.019)	17,485	0.649 (0.015)	0.801 (0.020)	0.717 (0.013)	0.835 (0.021)	0.883 (0.029)	0.858 (0.012)	1,827
lemma	0.807 (0.022)	0.917 (0.020)	0.858 (0.011)	16,002	0.643 (0.014)	0.809 (0.031)	0.716 (0.006)	0.834 (0.020)	0.886 (0.031)	0.859 (0.011)	1,827
stem	0.817 (0.020)	0.909 (0.029)	0.860 (0.013)	12,828	0.633 (0.022)	0.778 (0.051)	0.697 (0.019)	0.865 (0.019)	0.868 (0.025)	0.866 (0.016)	1,827

In the same way that demonstrated the work of Cabral. et al. (2021), lemmatization applied alone did not improve the TF-IDF + LR model performance. Although lemmatization and stemming did not improve the performance of our LR models, mBERT model trained with stemmed data outperformed the model trained with raw data by 1% of F1 Score.

There is not doubt that lemmatization and stemming contributed to reduce the dimensionality of the vector generated by TF-IDF, although performance did not improve with either treatments.

5.4.5 Stopwords

Another treatment commonly found in the literature is the removal of stopwords. As describe in Subsection 2.5.1.2, the reason behind this treatment is to decrease the vocabulary size by removing the irrelevant words, which are very common words that occur independently of the label and are therefore useless for discriminating the problem.

Although Jurafsky and Martin (2021) already pointed out in their work that in many applications removing stopwords does not improve performance, we would like to determine if the same applies in our context. Thus, this transformation has also been considered. To remove stopwords we used NLTK library (Bird et al., 2009).

The results obtained are presented in Table 5.11.

Table 5.11: Results of Experiments with Normalization Related to Stopwords. Showing: Precision, Recall and F1 Score with their respective standard deviation. Vocabulary size (V Size) generated by TF-IDF. Dev database size resulting after apply the normalization (D Size).

Treatment	TF-IDF + LR				mBERT + LR			mBERT			D Size
	Precision	Recall	F1	V Size	Precision	Recall	F1	Precision	Recall	F1	
raw text	0.809 (0.027)	0.913 (0.028)	0.858 (0.019)	17,485	0.649 (0.015)	0.801 (0.020)	0.717 (0.013)	0.835 (0.021)	0.883 (0.029)	0.858 (0.012)	1,827
remove stopwords	0.803 (0.023)	0.922 (0.017)	0.858 (0.015)	17,384	0.649 (0.030)	0.823 (0.025)	0.726 (0.025)	0.813 (0.028)	0.884 (0.041)	0.846 (0.007)	1,827

Considering F1 Score, there was no difference in the performance of TF-IDF + LR model trained with raw data and the model trained with preprocessed data. In the other hand, the mBERT + LR model trained with preprocessed data outperformed the model trained with raw data. While there was a small drop in the performance of the mBERT model with this treatment.

Considering that BERT takes into account the context of the sentence, the removal of stopwords may precisely be removing words that give context to the sentence, and consequently worsening the performance of this model.

5.4.6 Instagram Fonts

As mentioned in Subsection 4.3.2.7, 73 (2%) posts have Instagram Fonts. Despite the low use, the problem here is that different symbols generate different tokens. As an example, for the words: "physalia" and "𝔭𝔥𝔶𝔰𝔞𝔩𝔦𝔞" will be created two tokens by the TF-IDF vectorizer, both with the same semantics, while the mBERT tokenizer turns the second one into a `UNK` token.

To normalize the caption, we considered 2 methods:

- **ASCII**: conversion to ASCII using Unidecode library (Unidecode, 2023). This library switches everything to ASCII. This normalization also collaborates to normalizes combined characters such as ç, é, à. Some texts may have originally been written with these characters in their decomposed form, which generate broken tokens when vectorized by TF-IDF;

- **Unicodedata**: conversion using Unicodedata. It is a module from Python 3 that can be used to normalize the text written with Instagram Fonts.

The **ASCII** treatment resulted in empty samples, thus these samples were removed from the experiments.

The results obtained are presented in Table 5.12.

Table 5.12: Results of Experiments with Normalizations Related to Instagram Fonts. Showing: Precision, Recall and F1 Score with their respective standard deviation. Vocabulary size (V Size) generated by TF-IDF. Dev database size resulting after apply the normalization (D Size).

Treatment	TF-IDF + LR				mBERT + LR			mBERT			D Size
	Precision	Recall	F1	V Size	Precision	Recall	F1	Precision	Recall	F1	
raw text	0.809 (0.027)	0.913 (0.028)	0.858 (0.019)	17,485	0.649 (0.015)	0.801 (0.020)	0.717 (0.013)	0.835 (0.021)	0.883 (0.029)	0.858 (0.012)	1,827
ascii	0.785 (0.026)	0.911 (0.021)	0.843 (0.022)	16,712	0.625 (0.019)	0.810 (0.034)	0.706 (0.023)	0.872 (0.029)	0.850 (0.053)	0.859 (0.021)	1,824
unicodedata	0.809 (0.023)	0.910 (0.030)	0.856 (0.018)	17,162	0.644 (0.027)	0.801 (0.022)	0.714 (0.023)	0.835 (0.022)	0.897 (0.050)	0.864 (0.016)	1,827

Conversion to ASCII, although tested for the purpose of normalizing Instagram Fonts, changes all the text and did not benefit LR models. The text treated with Unicodedata, did not show a significant difference in F1 Score compared to the raw data.

5.4.7 Spelling Correction

As mentioned in Section 4.1, while some posts correctly follow the grammatical rules, others contain errors, and use slang and/or abbreviations. By applying spelling correction, it is possible to reduce these linguistic variations and consequently reduce the dimensionality of the feature vector.

To find out if there is any impact of applying spelling correction on models training for our problem, we considered the following transformation: normalization with Enelvo (Bertaglia and Nunes, 2016), which is a package developed in Python, for user generated content in Portuguese.

One interesting feature of this package is that one can create dictionaries to force correction or/and ignore terms. As an example, we can always force the correction of term "phisalia" to "physalia".

The results obtained are presented in Table 5.13.

Table 5.13: Results of Experiments with Spelling Correction. Showing: Precision, Recall and F1 Score with their respective standard deviation. Vocabulary size (V Size) generated by TF-IDF. Dev database size resulting after apply the normalization (D Size).

Treatment	TF-IDF + LR				mBERT + LR			mBERT			D Size
	Precision	Recall	F1	V Size	Precision	Recall	F1	Precision	Recall	F1	
raw text	0.809 (0.027)	0.913 (0.028)	0.858 (0.019)	17,485	0.649 (0.015)	0.801 (0.020)	0.717 (0.013)	0.835 (0.021)	0.883 (0.029)	0.858 (0.012)	1,827
enelvo	0.813 (0.025)	0.916 (0.028)	0.861 (0.019)	17,225	0.640 (0.012)	0.796 (0.026)	0.709 (0.009)	0.850 (0.030)	0.872 (0.051)	0.860 (0.029)	1,827

Considering F1 Score, the difference in performance of models trained with data normalized by Enelvo compared to models trained with raw data was insignificant for TF-IDF + LR and mBERT, and there was a small drop in mBERT + LR model. Beside that, among all the nomalizations tried, this one took the longest to run.

In general, considering F1 Score, in experiments with TF-IDF + LR, it is noted that none of the models trained with normalized data outperformed significantly the model trained with raw data. In the experiments with mBERT + LR, only the model trained with text without stopwords outperformed the model trained with raw data by 1%. In the experiments with mBERT,

5 models trained with normalized data (stemming, tokenize and remove numbers, tokenize and remove user mentions) outperformed the model trained with raw data between 1% to 2%.

5.5 COMBINATION OF NORMALIZATIONS

The combination of two or more treatments in the preprocessing step is common in NLP tasks. In this Section we performed experiments with combinations of some of the normalizations experimented in Section 5.4.

Since training mBERT + LR models did not present good results, we did not perform experiments with this architecture in this Section. Only experiments with TF-IDF + LR and mBERT were performed. Considering that the normalizations with the most interesting results were different for these 2 models, we created different combinations of normalization for each one. In total, 191 normalization combinations per model were experimented.

We experimented with combinations of the following normalizations with TF-IDF + LR: Tokenize user mentions, tokenize numbers, remove URLs, translate emojis, conversion with Unicodedata, remove stopwords, conversion to ASCII and stemming. We experimented with combinations of the following normalizations with mBERT: Remove user mentions, remove numbers, tokenize URLs, tokenize emojis, conversion with Unicodedata, remove stopwords, conversion to ASCII and stemming.

The order in which the normalizations were applied was as follows: conversion with Unicodedata; remove or tokenize user mentions; remove or tokenize URLs; remove or tokenize numbers; remove, tokenize or translate emojis; remove stopwords; stemming and conversion to ASCII.

TF-IDF + LR models were trained with cross-validation across 5 folds, while mBERT models were trained with 3 folds and by 15 epochs.

The highest precision and F1 Score obtained by mBERT models trained with normalized data were 0.887 and 0.894 respectively. Table 5.14 shows the top 3 results of mBERT models trained with normalized data ordered by highest F1 Score and ordered by highest precision. To compare the results obtained with models trained with normalized data, we show the results of the model trained with raw text with the same hyperparameters.

Table 5.14: The Top 3 Normalization Combinations - mBERT models. Showing the Rank, Identifier for the combination of normalizations (ID), emoji-related treatment applied (Emoji), user-mention-related treatment applied (Mention), number-related treatment applied (Number), url-related treatment applied (URL), whether stemming was applied or not (Stemm), whether stopwords was removed or not (Stopwords), Instagram-fonts-related treatment applied (IF), Precision, Recall and F1 Score with their respective standard deviation.

Rank	ID	Emoji	Mention	Number	URL	Stemm	Stopwords	IF	Precision	Recall	F1
raw text		none	none	none	none	none	none	none	0.835 (0.021)	0.883 (0.029)	0.858 (0.012)
The Top 3 - F1 Score											
1º	5625	tokenize	none	remove	none	true	none	none	0.869 (0.031)	0.923 (0.048)	0.894 (0.019)
2º	5169	none	none	none	none	true	none	none	0.882 (0.030)	0.906 (0.040)	0.893 (0.012)
3º	8686	tokenize	remove	none	none	none	true	none	0.866 (0.018)	0.923 (0.037)	0.893 (0.021)
The Top 3 - Precision											
1º	9091	tokenize	remove	remove	none	none	true	none	0.887 (0.023)	0.886 (0.011)	0.887 (0.014)
2º	9598	none	remove	remove	tokenize	none	true	none	0.883 (0.029)	0.891 (0.025)	0.886 (0.014)
3º	5169	none	none	none	none	true	none	none	0.882 (0.030)	0.906 (0.040)	0.893 (0.012)

The highest precision and F1 Score obtained by TF-IDF + LR models trained with normalized data were 0.823 and 0.866 respectively. Table 5.15 shows the top 3 results of TF-IDF + LR models trained with normalized data ordered by highest F1 Score and ordered by highest precision. To compare the results obtained with models trained with normalized data, we include the results of the model trained with raw text.

Table 5.15: The Top 3 Normalization Combinations - TF-IDF + LR models. Showing the Rank, Identifier for the combination of normalizations (ID), emoji-related treatment applied (Emoji), user-mention-related treatment applied (Mention), number-related treatment applied (Number), url-related treatment applied (URL), whether stemming was applied or not (Stemm), whether stopwords was removed or not (Stopwords), Instagram-fonts-related treatment applied (IF), Precision, Recall and F1 Score with their respective standard deviation.

Rank	ID	Emoji	Mention	Number	URL	Stemm	Stopwords	IF	Precision	Recall	F1
raw text		none	none	none	none	none	none	none	0.809 (0.027)	0.913 (0.028)	0.858 (0.019)
The Top 3 - F1 Score											
1º	6511	none	none	tokenize	remove	none	none	none	0.822 (0.024)	0.914 (0.026)	0.866 (0.020)
2º	6535	none	none	tokenize	remove	none	none	none	0.821 (0.020)	0.913 (0.025)	0.864 (0.018)
3º	5359	none	none	tokenize	none	none	none	none	0.820 (0.022)	0.913 (0.025)	0.864 (0.019)
The Top 3 - Precision											
1º	7236	translate	tokenize	tokenize	none	true	none	none	0.823 (0.019)	0.907 (0.025)	0.863 (0.017)
2º	6511	none	none	tokenize	remove	none	none	none	0.822 (0.024)	0.914 (0.026)	0.866 (0.020)
3º	8388	translate	tokenize	tokenize	remove	true	none	none	0.822 (0.019)	0.906 (0.026)	0.861 (0.017)

Considering F1 Score, it is possible to see that even models trained with combined normalizations were not able to outperform the model trained with raw text by more than 1% in the case of TF-IDF + LR, while the mBERT trained with normalization combinations outperformed the model trained with raw text by up to 3%.

It is interesting to note that the increase in F1 Score, even if small, in the TF-IDF + LR models trained with normalized data was the result of the increase in precision. While in the mBERT models, there was improvement in both precision and recall.

What is not possible to see in Table 5.15 is that TF-IDF + LR model's F1 Score from the 2nd position to the 143rd position is 86%, which means that there are 142 combinations of normalizations that resulted in an F1 Score of 86%. In the results ordered by highest precision, we notice a similar phenomenon: from the 1st position to the 51st the precision is 82%. Finally, when we observe the results that maximize both the F1 Score (86%) and the precision (82%) we found 51 possible combinations of normalizations. That is, we found 51 combinations of normalizations that achieved very similar results for the TF-IDF + LR model.

Something similar happened with mBERT models. When we look at the results sorted by F1 Score, the first 44 positions achieved an F1 Score of 89%. The difference is in the precision: only 7 combinations were able to reach 88%. When we observe the results that maximize both the F1 Score (89%) and the precision (88%) we found only 6 possible combinations of normalizations.

Considering that both F1 Score and precision are important metrics for our problem, the combination that most favored mBERT models for these metrics was combination with ID 5169, while for TF-IDF + LR models it was combination with ID 6511.

5.6 HYPERPARAMETER OPTIMIZATION

As we saw in Subsection 2.1.7, the model's hyperparameters are not learned during training. In this Section we performed hyperparameter optimization of the TF-IDF + LR and mBERT models, training them with raw and normalized data according to the combinations that most favored the mBERT and LR models (Section 5.5). Here, we trained the models with cross-validation across 5 folds.

5.6.1 Logistic Regression

To search for the best hyperparameters we used the Grid-search technique (Subsection 2.1.7.1). Table 5.16 shows the hyperparameters experimented and chosen, while Table 5.17 shows the results with and without hyperparameter optimization.

Table 5.16: Grid-Search parameter range and best parameters for TF-IDF + LR

	Parameter	Default Value	Values Tried	Best Parameter	
				raw text	norm text
LR	C	1.0	1 to 100 | 0.01 to 1.0	11	10
LR	solver	lbfgs	lbfgs | liblinear	lbfgs	lbfgs
LR	penalty	L2	L1, L2	L2	L2
TF-IDF	N-gram	1g	1g | 2g | 3g | 1,2g | 1,2,3g | 2,3g	1g	1g
TF-IDF	Max features	None	2,000 to 17,000	None	None
TF-IDF	Max DF	1.0	0.05 to 1.0	1.0	1.0
TF-IDF	Min DF	1	0.01 to 1.0	1	1

Table 5.17: Results of Hyperparameter Optimization for TF-IDF + LR. Showing Precision, Recall and F1 Score with their respective standard deviation.

	With Optimization			Without Optimization		
	Precision	Recall	F1	Precision	Recall	F1
raw text	0.859 (0.013)	0.910 (0.025)	0.884 (0.013)	0.809 (0.027)	0.913 (0.028)	0.858 (0.019)
norm text	0.858 (0.013)	0.917 (0.023)	0.886 (0.014)	0.822 (0.024)	0.914 (0.026)	0.866 (0.020)

Observing the results of hyperparameter optimization for TF-IDF + LR, we can note that models trained with raw and normalized data have practically the same results. This indicates that, for the data used in this dissertation, model optimization does not require prior text normalization.

5.6.2 mBERT

Table 5.18 shows the hyperparameters experimented and chosen, while Table 5.19 shows the results with and without hyperparameter optimization. Here we are treating the sentence length as a hyperparameter, but it is not a BERT hyperparameter. We are experimenting the impact of different sizes of sentence length combined with hyperparameters.

As in the results of hyperparameter optimization for TF-IDF + LR, the results of hyperparameter optimization for mBERT, show that models trained with raw and normalized data have practically the same results, indicating that, for the data used in this dissertation, model optimization does not require prior text normalization.

Table 5.18: Hyperparameters tried range and best parameters for mBERT

Parameter	Values Listed*	Values Tried	Best Parameter		
			raw text	norm text F1**	norm text Pre***
Epoch	2, 3, 4	2, 3, 4, 5, 10, 15, 20, 50	15	20	50
Batch size	16, 32	8, 16, 32, 64	16	8	8
Learning rate	5e-5, 3e-5, 2e-5	6e-5, 5e-5, 3e-5, 2e-5	3e-5	3e-5	3e-5
Learning rate warmup + linear decay	0.01	None, 0.01, 0.1, 0.2	0.1	0.1	0.1
Sentence length		between 64 and 512	344	164	512

*Values listed by Devlin et al. (2019) that worked better for fine-tuning in their experiments.
**Model's parameters with best F1 Score.
***Model's parameters with best precision.

Table 5.19: Results of Hyperparameter Optimization for mBERT. Showing Precision, Recall and F1 Score with their respective standard deviation.

	With Optimization			Without Optimization*		
	Precision	Recall	F1	Precision	Recall	F1
raw text - best F1 and Precision	0.873 (0.025)	0.907 (0.030)	0.889 (0.016)	0.823 (0.041)	0.886 (0.046)	0.852 (0.014)
norm text - best F1	0.871 (0.018)	0.912 (0.026)	0.890 (0.011)	0.861 (0.011)	0.907 (0.036)	0.883 (0.016)
norm text - best Precision	0.878 (0.014)	0.893 (0.045)	0.885 (0.025)			

*Hyperparameters used as default in Chapter: batch size 16 + learning rate 3e-5 + sentence length 128 + 50 epochs + warmup 0.10

5.7 FINAL MODEL

Now that we know which normalizations and hyperparameters could increase the results of the models, it's time to obtain a final model capable to determine whether the post is a legitimate occurrence of *Physalia physalis*, using only the caption of the posts.

To reach a final model, we trained models using the entire development set and evaluated them on the test set.

We trained two TF-IDF + LR models using the best hyperparameters found after optimization (Table 5.16): one with raw data and other with data normalized, since the results after optimization were very close between the models trained with raw and normalized data.

We also trained mBERT models using the best hyperparameters reached with raw data, hyperparameters with best F1 Score achieved with normalized data, and hyperparameters with best precision achieved with normalized data (see Table 5.18). For each configuration, five training rounds were performed with different random states. Also, we used 2 checkpoints, one for the best precision and other for best F1 Score.

The results obtained are presented in Table 5.20. For mBERT models, we show the mean performance over 5 runs with the standard deviation.

Considering F1 Score and precision, the results show that the difference in performance of TF-IDF + LR model trained with normalized data compared to model trained with raw data

Table 5.20: Final Results - Text Models. For mBERT models, we show the mean performance over 5 runs with the standard deviation

	Precision	Recall	F1
TF-IDF + LR			
raw text	0.812	0.889	0.849
norm text	0.810	0.900	0.853
mBERT			
raw text - best F1	0.826 (0.019)	0.903 (0.014)	0.863 (0.006)
raw text - best Precision	0.868 (0.012)	0.809 (0.051)	0.837 (0.024)
norm text - best F1	0.837 (0.014)	0.889 (0.012)	0.862 (0.003)
norm text - best Precision	0.869 (0.009)	0.804 (0.046)	0.834 (0.022)

was insignificant. Similar results had already been observed in the hyperparameter optimization results (Table 5.17).

In the same way, for mBERT, the results show that there was no major difference between models trained with raw data and normalized data. This was also observed in the hyperparameter optimization results (Table 5.19).

5.7.1 Error Analysis

We compared the performance of the best performing classifiers. As we performed 5 runs for each mBERT configuration, we present here the results of the models that achieved the best performance among the 5 runs.

Figure 5.2 presents the confusion matrices achieved by the models TF-IDF + LR.

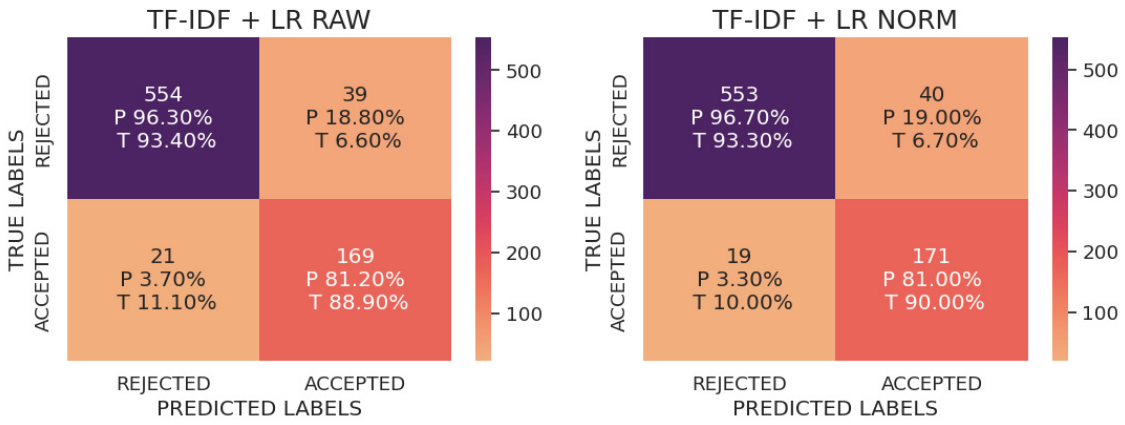


Figure 5.2: Confusion Matrices achieved by the models TF-IDF + LR trained with raw data (TF-IDF + LR RAW) and normalized data (TF-IDF + LR NORM). Beyond the classifier's errors and successes in absolute numbers, they show normalized numbers over the predicted label (P), and normalized numbers over the true label (T).

In Figure 5.3 we present the confusion matrices achieved by the BERT models trained with raw data.

In Figure 5.4 we present the confusion matrices achieved by the BERT models trained with normalized data.

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.

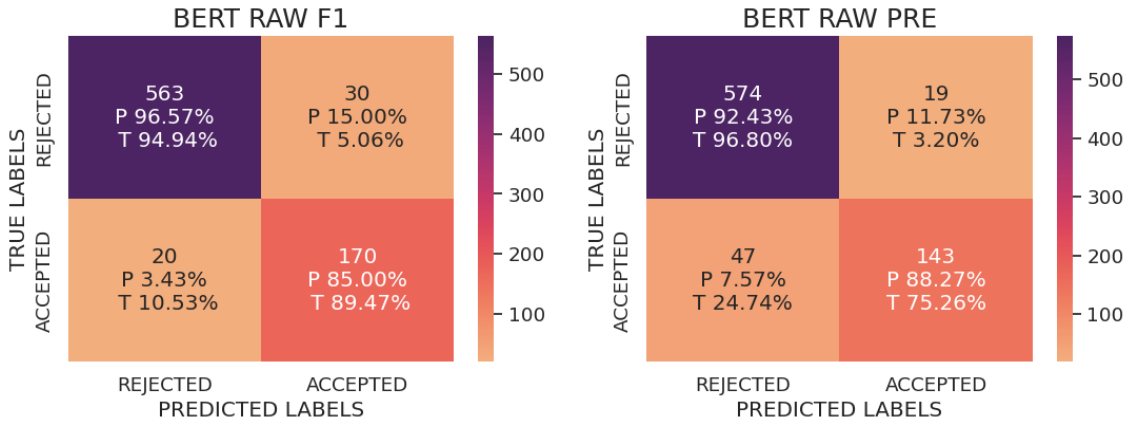


Figure 5.3: Confusion Matrices of mBERT models trained with raw data. BERT RAW F1 shows the confusion matrix of model that achieved the best F1 Score among the 5 runs, while BERT RAW PRE shows the confusion matrix of model that achieved the best precision among the 5 runs. Beyond the classifier's errors and successes in absolute numbers, they show normalized numbers over the predicted label (P), and normalized numbers over the true label (T).

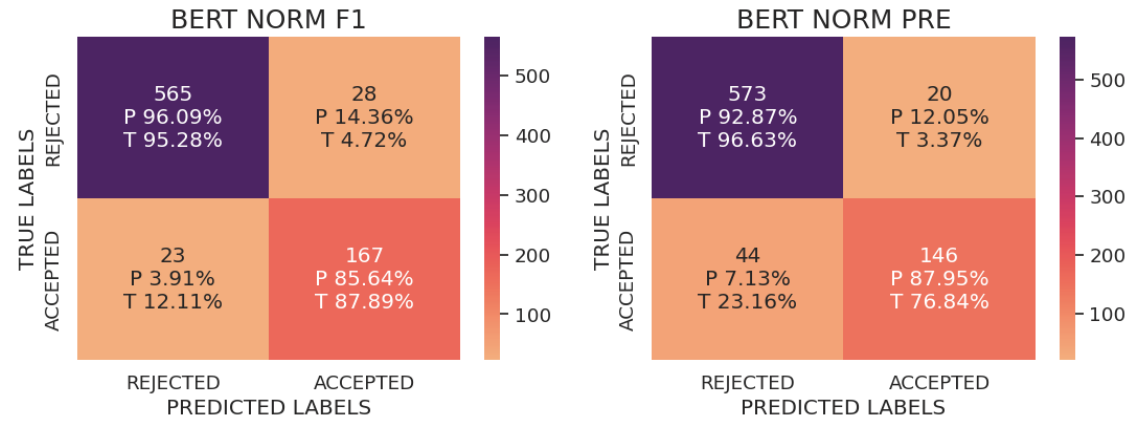


Figure 5.4: Confusion Matrices of mBERT models trained with normalized data. BERT NORM F1 shows the confusion matrix of model that achieved the best F1 Score among the 5 runs, while BERT NORM PRE shows the confusion matrix of model that achieved the best precision among the 5 runs. Beyond the classifier's errors and successes in absolute numbers, they show normalized numbers over the predicted label (P), and normalized numbers over the true label (T).

As we mentioned in the beginning of the Chapter, the ideal for our application is to minimize the number of false positives.

In all, the same 16 posts were misclassified by all models, 8 of which are false positives. Table 5.21 shows three samples of false positive by all text models. The complete list can be found in Appendix A (Table A.1).

5.7.2 The Best Text Model

Considering the results of precision, since a high precision means a lower number of false positives, the best model to classify Instagram posts as accepted or rejected as legitimate occurrence of *Physalia physalis*, using only the caption of the posts, was BERT RAW PRE: mBERT model trained with raw data and the one that achieved the highest precision among the five training rounds. This model achieved a precision of 0.883, recall of 0.753 and F1 Score of 0.813.

Table 5.21: Examples of False Positive by All Text Models. Showing Caption, Spatial Information (Spatial Info.), Taxonomic Information (Tax. Info.), Reason for Rejection (R. Rejection) and mean Probability with standard deviation (Prob). Emojis were replaced by EMOJI

I	Caption (up to 250 characters)	Spatial Info.	Tax. Info.	R. Rejection	Prob
1	Um pequeno lembrete: "Com Caravela não se brinca!" #portuguesemanowar #burned #caravelaportuguesa #marinelife #naturelovers #sandandsea	COAST-GEO	ACCIDENT	MEDIA	0.95 (0.063)
2	Bebezíneo de Caravela Portuguesa (foi o que disseram pra gente). #pará #salinopolis #praiadascorvinas #regiaonorte #turismobrasil #caravelaportuguesa #jambutour	COAST-TEXT	CNIDARIA	MEDIA	0.91 (0.138)
3	Milhares delas, pequeninas e venenosas... #poison #águasvivas #caravelaportuguesa #beach #swimming #ocean #atlantic #ericeira #portugal #portuguesemanofwar #wild#surf	NOTHING	CNIDARIA	MEDIA AND LOCATION	0.89 (0.171)

It is worth remembering that the model was trained with adapted-lack-location label and the filtered dataset, so when used in production this must be taken into account.

In all, the BERT RAW PRE misclassified 66 posts, where 19 are false positives. Table 5.22 shows three samples of false positive by this model. The complete list can be found in Appendix A (Table A.2). Among the 19 misclassified as positive, 8 are rejected because of media, 9 because of the media and location and 2 because the location is not on the coast. 8 images are classified as NOTHING, 5 as CNIDARIA, 3 as ART, 2 as REALISTIC and 1 as ACCIDENT.

Table 5.22: Examples of False Positive by BERT RAW PRE. Showing Index (I), Caption, Spatial Information (Spatial Info.), Taxonomic Information (Tax. Info.), Reason for Rejection (R. Rejection) and Probability (Prob). Emojis were replaced by EMOJI.

I	Caption (up to 300 characters)	Spatial Info.	Tax. Info.	R. Rejection	Prob
1	A caravela-portuguesa é o único organismo em colônia heteromorfa, no grupo dos cnidários. Ou seja, são seres que vivem em colônia, isto é, estão conectados anatomicamente e não sendo um único ser...	COUNTRYSIDE-GEO	REALISTIC	LOCATION	1.0
2	Um pequeno lembrete: "Com Caravela não se brinca!" #portuguesemanowar #burned #caravelaportuguesa #marinelife #naturelovers #sandandsea	COAST-GEO	ACCIDENT	MEDIA	0.97
3	Bebezíneo de Caravela Portuguesa (foi o que disseram pra gente). #pará #salinopolis #praiadascorvinas #regiaonorte #turismobrasil #caravelaportuguesa #jambutour	COAST-TEXT	CNIDARIA	MEDIA	1.0

Table 5.23 shows samples of false negative by BERT RAW PRE model. The complete list can be found in Appendix A (Table A.3). Of the 47 false negatives, 33 were really accepted by the oceanographer, while 14 were rejected because of lack of location (and were considered as accepted when we adapted the label). 38 images are classified as REALISTIC and 9 as EDITED.

Table 5.23: Examples of False Negative by BERT RAW PRE. Showing Index (I), Caption, Spatial Information (Spatial Info.), Taxonomic Information (Tax. Info.), when REJECTED the Reason for Rejection (R. Rejection) and Probability (Prob). Emojis were replaced by EMOJI.

I	Caption (up to 300 characters)	Spatial Info.	Tax. Info.	R. Rejection	Prob
1	Hoje foi registrado novamente o aparecimento de caravelas pelas praias de Porto belo EMOJI Fiquem atentos! Registro da seguidora @mariana_vycente #praia #aguaviva #ft #litoral #portobelo #bombinhas #sc #brasil #caixadaço...	COAST-TEXT	REALISTIC	ACCEPTED	0.0
2	#pernambuco #caravelaportuguesa	COAST-GEO	REALISTIC	ACCEPTED	0.002
3	EMOJI Os cnidários, também conhecidos como celenterados, são animais que se reúnem no filo Cnidaria.... EMOJIA maioria dos cnidários é encontrada em ambientes marinhos, e os principais representantes desse grupo são as águas-vivas, as anêmonas-do-mar, os corais e as caravelas.... EMOJI Existem...	NOTHING	REALISTIC	LOCATION	0.0
4	Somos tudo aquilo que dizemos que somos... EMOJI@flarlesonpedrosa #nova coleção verão 2016/17 @hurley aqui na #FRAN6STORE . Enviamos para todo EMOJI Obrigado Senhor! #Camping. A sua segunda #house é aqui! Por tudo somos grato! Diárias à partir de R$ 19.90 EMOJI #HOSTEL FRAN6...	COAST-TEXT	REALISTIC	ACCEPTED	0.0

The truth is that sometimes it is very difficult, even for a human, to identify a legitimate occurrence using only the caption. The reader can try to perform this task using Table 5.22, of course ignoring the fact that all those posts listed are rejected. Certainly, reading just the text, the reader would accept post no. 3[4] as a legitimate occurrence, since the text suggests that the image is from a *Physalia physalis* puppy and also uses hashtags indicating location on the Brazilian coast (i.e., #salinopolis and #praiadascorvinas). But, the image is not from a *Physalia physalis* (i.e., Taxonomic Information equals CNIDARIA), and therefore rejected by the oceanographer.

Other example is the post no. 2[5], the text suggests that a poisoning by *Physalia physalis* occurred and the image was classified as ACCIDENT. The reader maybe believes that if there was an accident soon there was an occurrence. But, the post was rejected by the oceanographer, since the post does not meet the taxonomic criterion: "a media that clearly shows a *Physalia physalis*" (see Section 4.2).

Also the model trained with LR + TFIDF misclassified the posts no. 2 and 3 (see Table 5.21).

Even though some texts are talking about Portuguese man-of-war and have spatial information on the Brazilian coast (see Table A.2 for the complete list of false positives). Note that the majority of false positives (i.e., 17) were rejected because of the media or media and location (as in the case of posts no. 2 and 3, already commented). Only 2 posts were rejected because the location is not on the Brazilian coast.

It is important to say that the simple presence of terms such as "caravela portuguesa" or "physalia physalis" does not indicate a legitimate occurrence of *Physalia physalis*. As we can see in Tables 5.22 and 5.23, the hashtag #caravelaportuguesa appears both among those rejected and among those accepted.

33% (846) of the posts present in the dataset used for training, validation and testing of the models have the key term "caravela portuguesa" in the text. Among them 49% are accepted and 51% are rejected, that is, "caravela portuguesa" alone is not enough to distinguish between accepted or rejected (see Table 4.23).

We can also observe the false negatives (see Table 5.23), although there are posts that clearly talk about an occurrence of *Physalia physalis*, as the post no. 1[6]. When we read the post no. 2 is impossible to say if they are talking about an occurrence or not.

In this work we adapted the labels based on spatial information. But there could be a refinement of the adapted labels based on the text of the post. As an example, the post no. 3[7] in Table 5.23 was considered accepted because it was rejected because of the lack of location. However, it is an educational text about cnidarians and not a description of the occurrence of *Physalia physalis*.

Perhaps the post no. 4 (Table 5.23) causes some doubts when it appears among those accepted, since the text is talking about some product and hostel[8]. However, the post meets the taxonomic and spatial information necessary to be accepted as a legitimate occurrence of *Physalia physalis*.

---

[4]Translate as: Little Portuguese man-of-war Baby (that's what they told us)

[5]Translate as: A little reminder: "You can't play with Caravela!"

[6]Translate as: "Today the appearance of caravels was recorded again on the beaches of Porto belo EMOJI Stay careful! Registration of follower @mariana_vycente..."

[7]Translate as: "EMOJI Cnidarians, also known as coelenterates, are animals that come together in the phylum Cnidaria.... EMOJI Most cnidarians are found in marine environments, and the main representatives of this group are jellyfish, sea anemones, the corals and caravels.... EMOJI There are.."

[8]Translate as "We are everything we say we are... EMOJI@flarlesonpedrosa #new summer 2016/17 collection @hurley here at #FRAN6STORE. We send to everyone EMOJI Thank you Lord! #Camping. Your second #house is here! For everything we are grateful! Daily rates from R$ 19.90".

To find out if the model is really applicable to generate a base on occurrences of *Physalia physalis* in a real scenario, we evaluated the model by simulating 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, the post 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 posts `accepted` and `rejected because of media`, totaling 257 posts. Table 5.24 shows the results of the best text model evaluated on the full test set and partial test set.

Table 5.24: Results of the best text model evaluated on the partial and full test sets. Showing: Dataset, Precision, Recall and F1 Score.

Dataset	Precision	Recall	F1
full test set	0.883	0.753	0.813
partial test set	0.931	0.766	0.840

From the results presented in Table 5.24, it is possible to observe an increase in all metrics when we evaluated the model by simulating the flow in production, indicating that the BERT RAW PRE model can be used to generate a base on occurrences of *Physalia physalis* in a real scenario.

One last experiment carried out was to train a mBERT model with adapted-not-in-coast label, and the same hyperparameters and configuration used in BERT RAW PRE. This will allow us to combine the results of this model with the results of the model trained only with image in Chapter 7. Five training rounds with different random states were performed. We also evaluated the models with the partial test set. The results are presented in Table 5.25.

Table 5.25: Results of the Best Text Model trained with adapted-lack-location and adapted-not-in-coast labels, 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.

Label	Full Test Set			Partial Test Set		
	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)

Finally, we present a compilation of the results of the best model trained with text. Table 5.26 show the results of the models that achieved the highest precision among all runs by label, evaluated with the partial and full test sets.

Table 5.26: The Best Text Model. Showing: Label, Precision, Recall and F1 Score.

Label	Full Test Set			Partial Test Set		
	Precision	Recall	F1	Precision	Recall	F1
adapted-lack-location	0.883	0.753	0.813	0.931	0.766	0.840
adapted-not-in-coast	0.873	0.846	0.859	0.947	0.879	0.912

As final analysis, when we observe the results presented in Table 5.26, it is possible to notice a significant increase in recall with the model trained with label adapted-not-in-coast, reflecting an increase in F1 Score of around 7% when we evaluated the model by simulating the flow in production. This increase of recall is also observed in the average of 5 rounds (Table 5.25).

When we observe the precision, an increase of 2% is observed only when the model is evaluated on the partial test set. These results are a good motivation to carry out more experiments with text and the adapted-not-in-coast label in future work.

5.8 DISCUSSION

As we discussed in the last Subsection, the task of classifying posts as accepted/rejected as a legitimate occurrence of *Physalia physalis* using only the text is very difficult even for a human. Especially considering that the expert's criteria do not include the interpretation of the post's text. However, one of the questions that this research seeks to answer is precisely "By analyzing only the text of the post, it is possible to identify occurrences of *Physalia physalis* with good precision?". The answer is yes, especially considering the characteristics of the data and the difficulty of the task.

In the end, one of the preprocessing techniques that presented good results for our problem was the exclusion of posts based on similarity, which was used as a filter for carrying out the other experiments in this work (Section 4.4). Throughout the experiments with text normalization (Section 5.4 and Section 5.5), in general, the models trained with normalized data presented superior results than the models trained with raw data. However, observing the results obtained with hyperparameter optimization (Section 5.6) and the Final models it is possible to note that models trained with raw and normalized data have practically the same results, indicating that, for the data used in this dissertation, model optimization does not require prior text normalization.

What draws attention in the final results is the small difference in performance between TF-IDF + LR and mBERT. Considering the best F1 Score result of TF-IDF + LR (0.853) and mBERT (0.863) the difference is only 1% (see Table 5.20). This demonstrates that the use of the classic model remains competitive, especially if the training time of BERT models is considered.

Obviously BERT models training time depends on the amount of data and the hyperparameters such as the number of epochs and batch size. During the training of the final models the two LR models were executed in 23 seconds each one, while on average each mBERT run took 14 minutes.

This Chapter presented the experiments carried out with text from posts collected from Instagram, with the aim of obtaining a model capable of identifying legitimate occurrences of *Physalia physalis*. In the next Chapter we will present experiments with the posts' image.

6 IMAGE ANALYSIS

In this Chapter we conducted experiments using the images of the posts. The main goal was to obtain a model capable to determine if the post is a legitimate occurrence of *Physalia physalis*, using only the images of the post, with a high precision and F1 Score. Code and models developed in this Chapter are available online[1].

To do this, we trained CNNs with one of the images from each post. As the posts can have more than one image, we randomly selected one image per post to perform the training. In particular, for the positive posts, we selected one image that met the taxonomic criterion (Subsection 4.2.2).

The experiments were performed using Python and Keras (Chollet et al., 2015). We trained the CNNs by transferring the learning from a ResNet50 (He et al., 2015) pre-trained with Imagenet (Russakovsky et al., 2015) and adapted to our problem. We chose the same architecture described as the best model in (Carneiro et al., 2022). The authors trained several CNNs in order to classify images as being or not from *Physalia physalis*. In their experiments, a Precision of 0.94 and F1 Score of 0.95 was obtained for the positive class. Our goal in this Chapter is different from the work developed by Carneiro et al. (2022) which aimed to classify images as being or not from *Physalia physalis*, while we are interested in determining whether the `post` is a legitimate occurrence of *Physalia physalis*.

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 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 Carneiro et al. (2022).

As in the textual analysis, most experiments in this Chapter were carried out with filtered and divided data as described at the beginning of the Chapter 5. We have also applied cross-validation across 5 folds. The same class weighting calculation as LR and BERT were applied to train the CNNs.

To evaluate the machine learning models trained in this Chapter, we used the same the metrics used in textual analysis, it is precision, recall and F1 Score.

In summary, the general workflow of image analysis, illustrated in Figure 6.1, begins with the construction of the dataset, a process common to text analysis, which includes data collection, annotation and filtering. Then, the data is separated into test and development (Holdout) and prepared with downscaling and data augmentation. As cross-validation is used, training and validation occur in 5 rounds alternating the validation set. In this step, the label selection (original vs adapted) is carried out, the hyperparameters are optimized and 2 techniques are experimented to deal with data imbalance: undersampling and class weighting. In the last step (test), once the best hyperparameters are known, the models are trained using the entire development set and evaluated on the test set.

[1]https://github.com/RESMA-PPGINF-UFPR-CAPES-PRINT/CaravelasImageAnalysis

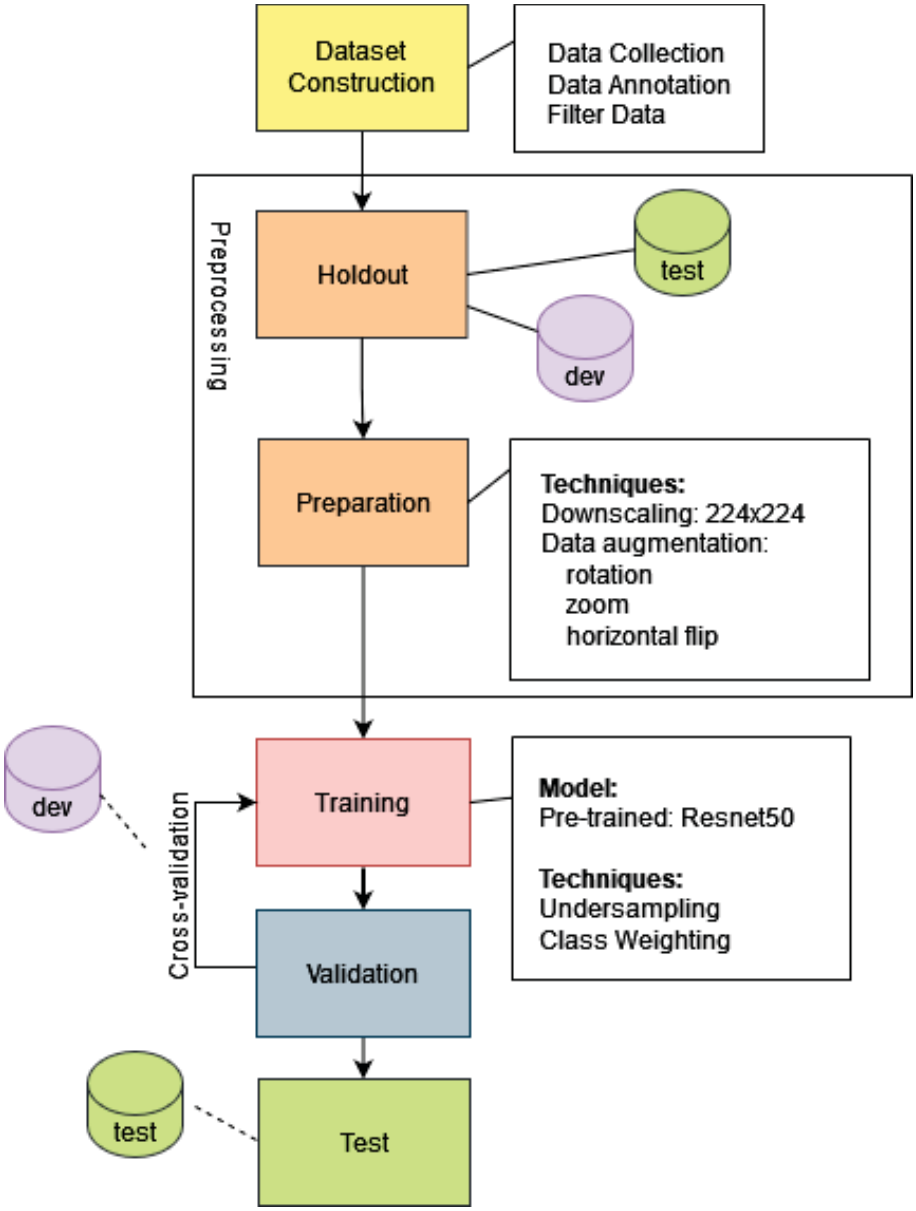


Figure 6.1: Image Analysis Workflow with an overview of techniques and models that we used in image analysis. Source: the Author

6.1 ORIGINAL LABEL VS ADAPTED LABELS

As we experimented with posts' caption, we wanted to find out if there is any impact of considering as accepted posts rejected because of location in models trained with posts' images. Thus, we performed experiments using the following datasets:

- `original`: with original labels assigned by the oceanographer.
- `adapted-lack-location`: consider as ACCEPTED posts rejected only for lack of spatial information (i.e., spatial information equals NOTHING).
- `adapted-not-in-coast`: consider as ACCEPTED posts rejected because of spatial information (i.e., spatial information different from COAST-TEXT and COAST-GEO).

The experiments were performed with filtered data with the exception of the filter that eliminates similar posts. We used as hyperparameters: 50 epochs with early stopping based on

the loss calculated in the validation set, batch size of 32 and Adam optimizer with learning rate of 0.001.

The results obtained are presented in the Table 6.1.

Table 6.1: Results of Experiments with Adapted Labels and 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-not-in-coast	0.967 (0.049)	0.986 (0.020)	0.976 (0.035)
adapted-lack-location	0.965 (0.024)	0.957 (0.059)	0.961 (0.042)

As in the experiment with caption, the experiments with images show that the adapted labels produced better classification models than the original label, with the best F1 Score with the `adapted-not-in-coast` label. It represents an increase of 9% of F1 Score compared with the original label. For this reason, we adopted this label for the rest of the experiments in this Chapter. As spatial information is not a feature used for training the models in this work, it is safe to use the adapted label in the experiments.

6.2 HYPERPARAMETER OPTIMIZATION

In this Section we performed hyperparameter optimization of the CNN, training it using two methods to deal with imbalanced data: Undersampling and Class Weighting. We chose these methods because the first one was used in the Carneiro et al. (2022) experiments and the last one we used in all experiments in the Textual Analysis. After applying undersampling we got a train dataset with 648 samples.

In addition, we carried out experiments with and without retraining the ResNet50 with our data. In all, we created four main training scenarios: undersampling + retraining, undersampling without retraining, class weighting + retraining and class weighting without retraining.

We use the following hyperparameters as default: 50 epochs with early stopping based on the loss calculated in the validation set and Adam optimizer.

Table 6.2 shows the hyperparameters experimented and chosen, while Table 6.3 shows the results obteined with hyperparameter optimization.

Table 6.2: Hyperparameters tried range and best parameters for ResNet50

Parameter	Values Tried	Best Parameter			
For imbalanced data	Undersampling | Class Weighting	Undersampling		Class Weighting	
Retrained	True | False	True	False	True	False
Batch size	8, 16, 32, 64	64	16	32	64
Learning rate	1e-3, 1e-4, 1e-5	1e-4	1e-3	1e-5	1e-3
No. neurons with ReLu	2, 64, 256, 512, 1024	256	256	512	64
Pooling window	2x2, 7x7	2x2	2x2	2x2	7x7

6.3 FINAL MODEL

To reach a final model to classify Instagram posts as accepted or rejected as legitimate occurrence of *Physalia physalis*, using only the images of the posts, we trained models using the best

Table 6.3: Results of Hyperparameter Optimization for ResNet50. Showing the method used to deal with imbalanced data (Method), whether ResNet50 was retrained with our data (Retrained), Precision, Recall and F1 Score with their respective standard deviations.

Method	Retrained	Precision	Recall	F1
Undersampling	True	0.989 (0.008)	0.948 (0.018)	0.968 (0.009)
	False	0.980 (0.007)	0.959 (0.015)	0.969 (0.007)
Class Weighting	True	0.988 (0.017)	0.984 (0.021)	0.986 (0.019)
	False	0.980 (0.024)	0.986 (0.022)	0.983 (0.023)

hyperparameters found after optimization (Table 6.2) and used the same main scenarios used in the hyperparameters optimization: undersampling + retraining (US RETRAINED), undersampling without retraining (US WITHOUT), class weighting + retraining (CW RETRAINED) and class weighting without retraining (CW WITHOUT). For each scenario, five training rounds were performed with different random states.

We trained models using the entire development set and evaluated them on the test set. After applying undersampling we got a training set with 926 samples.

We used as default hyperparameters: 50 epochs with early stopping based on the loss calculated in the test set and Adam optimizer. Also, we configure 2 checkpoints, one for the best Precision and other for the best F1 Score.

The results obtained are presented in the Table 6.4.

Table 6.4: Final Results - Image Models. Showing the method used to deal with imbalanced data (Method), whether ResNet50 was retrained with our data (Retrained), Precision, Recall and F1 Score. We show the mean performance over 5 runs with the standard deviation.

Method	Retrained	Precision	Recall	F1
Undersampling	True	0.933 (0.026)	0.895 (0.052)	0.913 (0.019)
	False	0.918 (0.013)	0.889 (0.030)	0.903 (0.013)
Class Weighting	True	0.935 (0.012)	0.912 (0.017)	0.923 (0.008)
	False	0.904 (0.006)	0.923 (0.011)	0.913 (0.007)

6.3.1 Error Analysis

We compared the results of the models that achieved the best performance among the 5 runs for each scenario. Figure 6.2 shows the confusion matrices achieved by these models. It is possible to observe that, as textual models, the models trained with images also had 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.

Although the CW RETRAINED model seems better on average due to F1 Score 1% higher than the others (see Table 6.4), looking at the confusion matrices of US RETRAINED and CW RETRAINED, it is possible to notice that US RETRAINED has 2% higher precision than CW RETRAINED, this is 96.9% versus 95.10%.

In all, the same 12 posts were misclassified by all models, 3 of which are false positives. Figure 6.3 shows false positive posts by all models.

The presence of figures that show a *Physalia physalis* among those misclassified may cause doubts at first, but we must remember that the models were trained with labels assigned to the post and not with labels assigned to the images, besides, as pointed out in the Subsection 4.2.2: a picture containing a *Physalia physalis* does not always mean the post will be accepted as

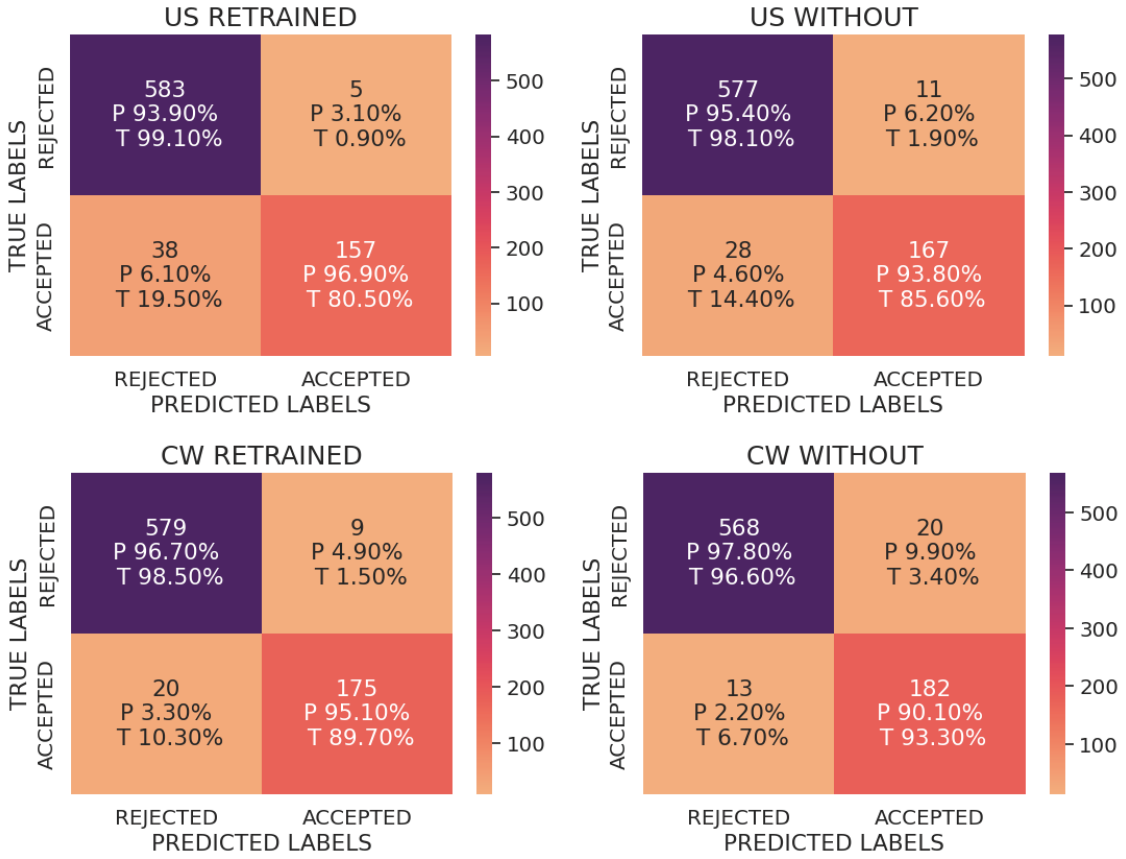


Figure 6.2: Confusion Matrices of CNN models trained with undersampling + retraining (US RETRAINED), undersampling without retraining (US WITHOUT), class weighting + retraining (CW RETRAINED) and class weighting without retraining (CW WITHOUT). The values displayed are from the models that achieved the best Precision among the 5 runs. 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).

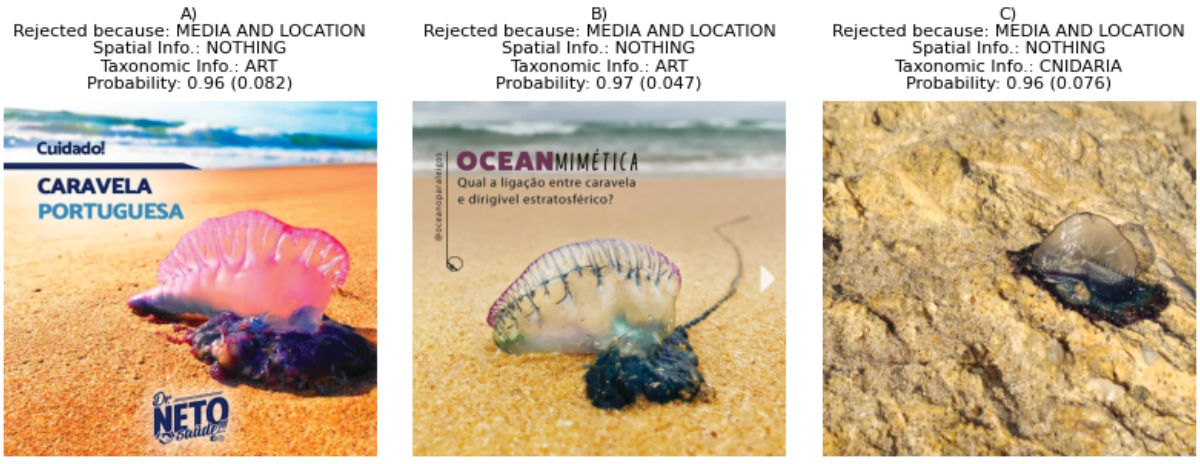


Figure 6.3: False Positive by All Image Models. At the top of each figure is informed the Reason for Rejection, Spatial Information, Taxonomic Information and mean Probability from all models with standard deviation.

a legitimate occurrence. As this is a real application, in some cases it was not possible to capture the nuances of what the oceanographer saw in the image, which made her rejects the post as a legitimate occurrence. Still, what is written in the images can give us an idea of the reason for the rejection even if there is a *Physalia physalis* in the image. For example, in figure A it says:

"Cuidado. Caravela Portuguesa. Dr Neto saúde!!!"[2]. In figure B it is written: "Oceanmimética. Qual a ligação entre caravela e dirigível estratosférico"[3]. The first post may be a warning about the risks of *Physalia physalis*, and reading the caption of the post that's right. The second clearly speaks of some curiosity related to this cnidarian, and reading the caption of the post that's right.

Figure 6.4 shows false negative posts by all models. Note that among the figures are posts that were originally accepted by the oceanographer (i.e., A, B, C, D and E), and also some whose label was adapted to ACCEPTED because they were rejected because of location by the oceanographer (i.e., F, G, H and I). Note also that with the exception of B, E and F, the others show *Physalia physalis* very small or almost indistinguishable.

6.3.2 The Best Image Model

Considering the Precision of the models, since a high Precision means a lower number of false positives, the best model to classify Instagram posts as accepted or rejected as legitimate occurrence of *Physalia physalis*, using only the images of the posts, was US RETRAINED: model retrained with our data, using undersampling as the method to deal with imbalanced data and achieved the highest precision among the five training rounds. This model achieved a Precision of 0.969, Recall of 0.805 and F1 Score of 0.879.

It is worth remembering that the model was trained with an adapted label and the filtered dataset, so when used in production this must be taken into account.

In all, the best model misclassified 43 posts, 5 of which are false positives. Figure 6.5 shows false positive posts by this model.

Figure 6.6 shows samples of false negative posts by the best model. See Appendix A for a complete list of false negative by this model (Figures A.1, A.2, A.3 and A.4). Among the 38 misclassified as negative, 10 show people or parts of the body (e.g. C and D), 22 images are REALISTIC and 15 images are EDITED. 21 of these posts were originally accepted by the oceanographer and 17 were originally rejected by her due to location (16 due to lack of location and 1 due to location in the countryside).

To find out if the model is really applicable to generate a base on occurrences of *Physalia physalis* in a real scenario, we evaluated the model by simulating 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, the post 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 posts `accepted` and `rejected because of media`, totaling 257 posts. Table 6.5 shows the results of the best image model evaluated on the full test set and partial test set.

Table 6.5: Results of the best image model evaluated on the partial and full test sets. Showing: Dataset, Precision, Recall and F1 Score.

Dataset	Precision	Recall	F1
full test set	0.969	0.805	0.879
partial test set	0.992	0.851	0.916

From the results presented in Table 6.5, it is possible to observe an increase in all metrics when we evaluated the model by simulating the flow in production, indicating that the US RETRAINED model can be used to generate a base on occurrences of *Physalia physalis* in a real scenario.

[2]Translate as: "Careful. Portuguese Caravel. Dr Neto health!!!"

[3]"Oceanmimetics. What is the connection between caravel and stratospheric airship"

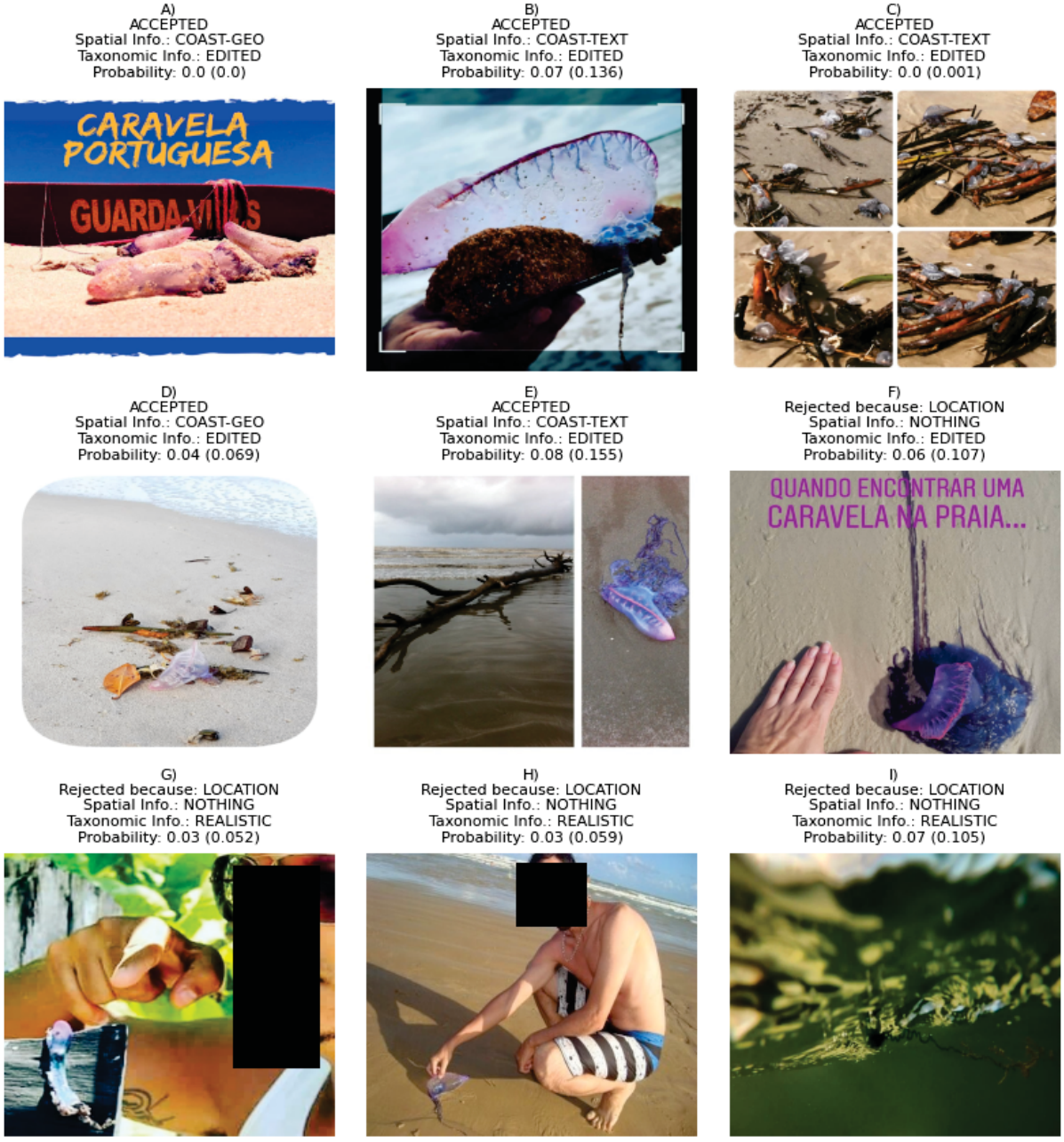


Figure 6.4: False Negative by All Image Models. At the top of each figure is informed the Accepted or Reason for Rejection if rejected, Spatial Information, Taxonomic Information and mean Probability from all models with standard deviation.

One last experiment carried out was to train the ResNet50 with adapted-lack-location label, and the same hyperparameters and configuration used in US RETRAINED. This will allow us to combine the results of this model with the results of the model trained only with text in the next Chapter. Five training rounds with different random states were performed. We also evaluated the models with the partial test set. The results are presented in the Table 6.6.

Finally, we present a compilation of the results of the best model trained with images. Table 6.7 show the results of the models that achieved the highest Precision among all runs by label, evaluated with the partial and full test sets.

As final analysis, when we observe the results presented in Table 6.7, it is possible to notice an increase in recall with the model trained with label adapted-lack-location, reflecting

Figure 6.5: False Positive by Best Image Model. At the top of each figure is informed the Reason for Rejection, Spatial Information, Taxonomic Information and Probability.

Table 6.6: Results of the Best Image Model trained with adapted-lack-location and adapted-not-in-coast labels, 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.

Label	Full Test Set			Partial Test Set		
	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)

Table 6.7: Results of the Best Image Model. Showing: Label, Precision, Recall and F1.

Label	Full Test Set			Partial Test Set		
	Precision	Recall	F1	Precision	Recall	F1
adapted-not-in-coast	0.969	0.805	0.879	0.992	0.851	0.916
adapted-lack-location	0.941	0.842	0.889	0.992	0.894	0.940

an increase in F1 Score of around 2% when we evaluated the model by simulating the flow in production. This increase is not observed in the average of 5 rounds (Table 6.6).

The small difference in results between models trained with the two adapted labels was also observed in the initial experiments (Section 6.1). At that moment we decided to adopt the label with the highest F1 Score to carry out the following experiments with images (i.e., adapted-not-in-coast). However, the 94% precision achieved on the partial test set is a good motivation to carry out more experiments with the adapted-lack-location label in future work.

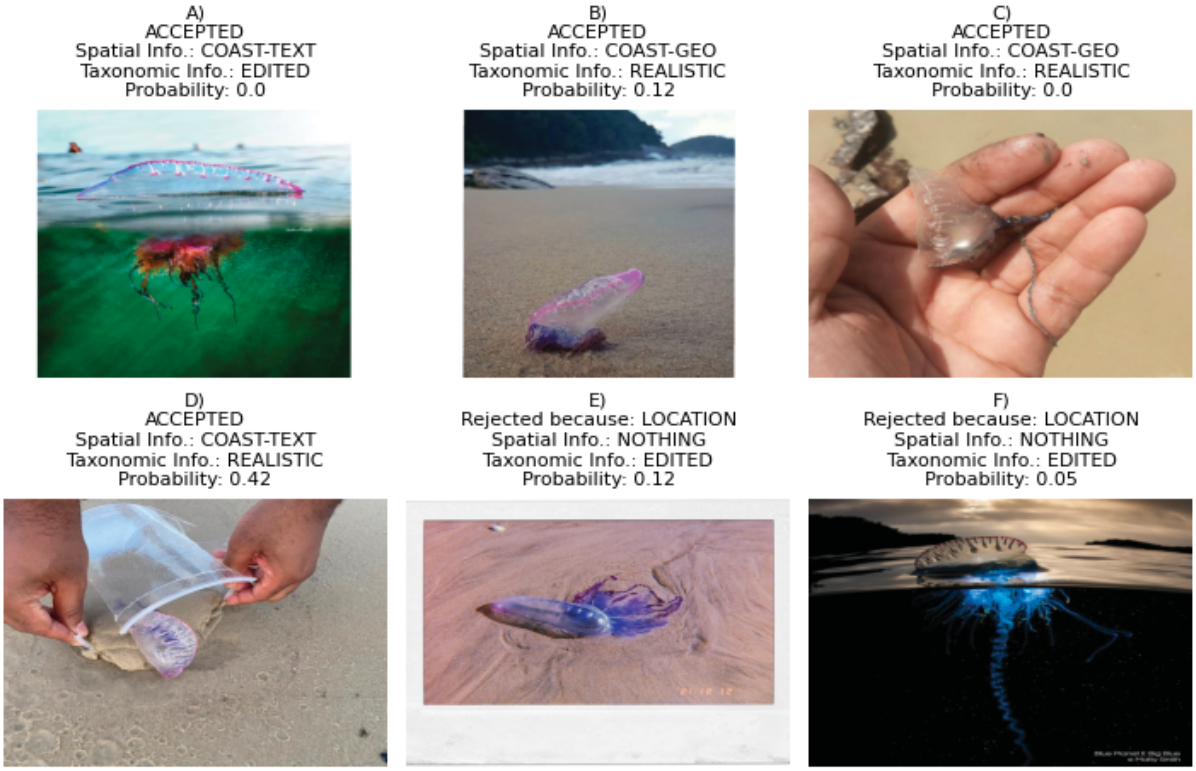


Figure 6.6: samples of False Negative by Best Image Model. At the top of each figure is informed the Accepted or Reason for Rejection if rejected, Spatial Information, Taxonomic Information and Probability.

In this Chapter we presented the experiments carried out with the aim of obtaining a model capable of identifying legitimate occurrences of *Physalia physalis*, using only the images of the posts collected from Instagram. In the next Chapter we will present experiments combining the results of the best models obtained with text and image.

7 MULTIMODAL AND COMPARATIVE ANALYSIS

In this Chapter we conducted experiments combining the results of the best models obtained with text (BERT RAW PRE) and image (US RETRAINED) (Section 7.1). We also, performed an analysis comparing the results of the best unimodal and multimodal models (Section 7.2).

7.1 MULTIMODAL ANALYSIS

In this Section we conducted experiments combining the results of the best models obtained with text (BERT RAW PRE) and image (US RETRAINED). The main goal was to obtain a model capable of determining if the post is a legitimate occurrence of *Physalia physalis* with high precision and F1 Score. For this we adopted a multimodal approach called Late Fusion. The rules used to integrate the results are those described in Subsection 2.4.1, that is: Product, Sum, Average, Max and Min.

To allow comparisons between unimodal and combined models, we show a compilation of the results of the unimodal models that achieved the best precision in Table 7.1.

Table 7.1: Compilation of the Results of the best unimodal models. Showing: Model, Label, Precision, Recall and F1 Score, resulting from the evaluation on the partial and full test sets.

Model	Label	Full Test Set			Partial Test Set		
		Precision	Recall	F1	Precision	Recall	F1
BERT RAW PRE	adapted-not-in-coast	0.873	0.846	0.859	0.947	0.879	0.912
	adapted-lack-location	0.883	0.753	0.813	0.931	0.766	0.840
US RETRAINED	adapted-not-in-coast	0.969	0.805	0.879	0.992	0.851	0.916
	adapted-lack-location	0.941	0.842	0.889	0.992	0.894	0.940

7.1.1 Combining Models Trained with ADAPTED-NOT-IN-COAST Label

In this Subsection we combined the results of models trained with adapted-not-in-coast label, that is, label that considers as ACCEPTED posts rejected because of spatial information (i.e., spatial information different from COAST-TEXT and COAST-GEO). The results are presented in the Table 7.2.

Table 7.2: Results of combining models trained with adapted-not-in-coast label. Showing: Fusion Rule, Precision, Recall and F1 Score

Fusion Rule	Precision	Recall	F1
Average	0.943	0.851	0.895
Product	0.943	0.841	0.889
Max	0.943	0.851	0.895
Min	0.943	0.851	0.895
Sum	0.986	0.723	0.834

Considering results obtained with the combination (Table 7.2) and individual results of the models trained with the same label and evaluated with the same set (Table 7.1), it is observed that only the Sum rule outperforms the precision of both individual models, although it has

the worst recall. The Average, Product, Max and Min rules outperform the F1 Score of both individual models, and also surpass the precision of the model trained only with text.

In Table 7.3 we present the results of the same combination now evaluated on the partial test set.

Table 7.3: Results of combining models trained with adapted-not-in-coast label and evaluated with partial test set. Showing: Fusion Rule, Precision, Recall and F1 Score

Fusion Rule	Precision	Recall	F1
Average	0.984	0.887	0.933
Product	0.984	0.894	0.937
Max	0.984	0.887	0.933
Min	0.984	0.887	0.933
Sum	0.950	0.950	0.950

Considering results obtained with the combination (Table 7.3) and individual results of the models trained with the same label and evaluated with the same set (Table 7.1), it is observed that the combination only surpasses the precision and F1 Score of the model trained only with text. The Average, Product, Max and Min rules outperform the F1 Score of the model trained only with image.

7.1.2 Combining Models Trained with ADAPTED-LACK-LOCATION Label

In this Subsection we combined the results of models trained with adapted-lack-location label, that is, label that considers as ACCEPTED posts rejected only for lack of spatial information (i.e., spatial information equals NOTHING). The results obtained are presented in the Table 7.4.

Table 7.4: Results of combining models trained with adapted-lack-location label. Showing: Fusion Rule, Precision, Recall and F1 Score

Fusion Rule	Precision	Recall	F1
Average	0.933	0.800	0.861
Product	0.948	0.774	0.852
Max	0.933	0.800	0.861
Min	0.933	0.800	0.861
Sum	0.977	0.663	0.790

Considering results obtained with the combination (Table 7.4) and individual results of the models trained with the same label and evaluated with the same set (Table 7.1), as observed in the later combination, the Sum rule outperforms the precision of both individual models, although it has the worst recall. The Average, Product, Max and Min rules outperform the precision and F1 Score of the model trained only with text. The Product rule surpasses the precision of the model trained only with image.

Table 7.5 presents the results of the same combination now evaluated on the partial test set.

Considering results obtained with the combination (Table 7.5) and individual results of the models trained with the same label and evaluated with the same set (Table 7.1), it is observed that the combination only surpasses the precision and F1 Score of the model trained only with text.

Table 7.5: Results of combining models trained with adapted-lack-location label and evaluated with partial test set. Showing: Fusion Rule, Precision, Recall and F1 Score

Fusion Rule	Precision	Recall	F1
Average	0.967	0.823	0.889
Product	0.967	0.844	0.902
Max	0.967	0.823	0.889
Min	0.967	0.823	0.889
Sum	0.943	0.936	0.940

7.1.3 Error Analysis

Here we compare the results of the combined models. Figure 7.1 shows the confusion matrices achieved by combined models using the Product rule and evaluated with the full test set. We could choose the Sum rule since it has the best precision, but this rule has a very low recall. That is why we chose to use the results of the Product rule.

It is possible to observe that, similar to the individual models, the combined ones 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.

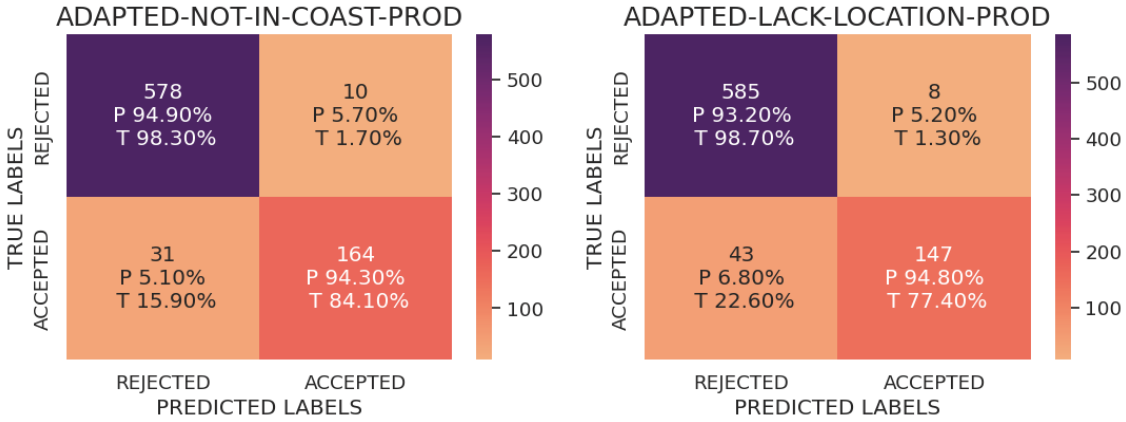


Figure 7.1: Confusion Matrices of combined models using the Product rule and evaluated with the full test base. Beyond the classifier's errors and successes in absolute numbers, they show normalized numbers over the predicted label (P), and normalized numbers over the true label (T).

Figure 7.2 shows the confusion matrices achieved by combined models using the Product rule and evaluated with the partial test set. It is possible to observe that, the difficulty in recognizing positive examples is also noted on the partial test set.

Figure 7.3 shows the confusion matrices achieved by combined models using the Sum rule and evaluated with the partial test set. In this case, it is observed that the models have the same difficulty in recognizing both positive and negative examples.

7.1.4 Discussion

In general, the combined models outperformed the models trained only with text, but did not outperform the models trained only with image.

When we compare the two combinations, considering the results of the models evaluated on the full test set (Tables 7.2 and 7.4), it is noted that combinations using adapted-not-in-coast label are better in terms of precision and F1 Score. Although the difference of precision is 1% in all the rules, the difference of F1 Score is 3% to 4%.

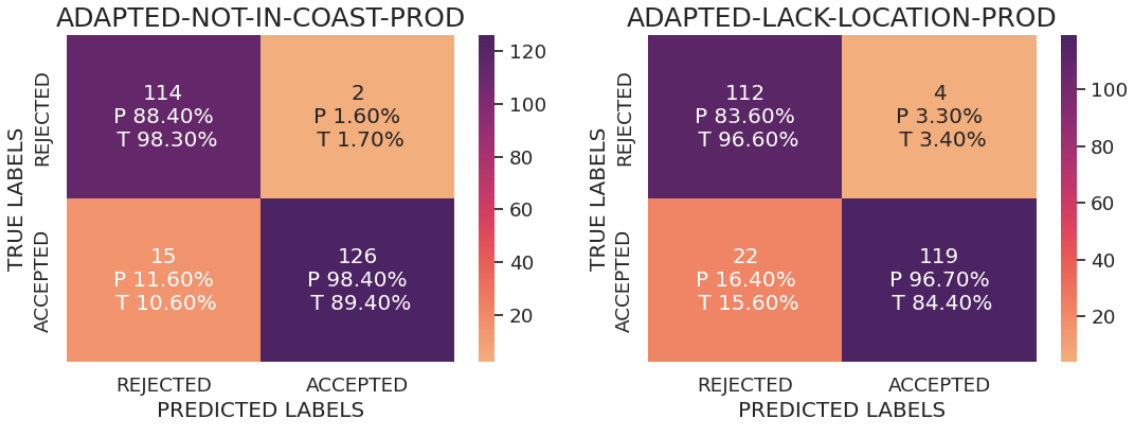


Figure 7.2: Confusion Matrices of combined models using the Product rule and evaluated with the partial test base. Beyond the classifier's errors and successes in absolute numbers, they show normalized numbers over the predicted label (P), and normalized numbers over the true label (T).

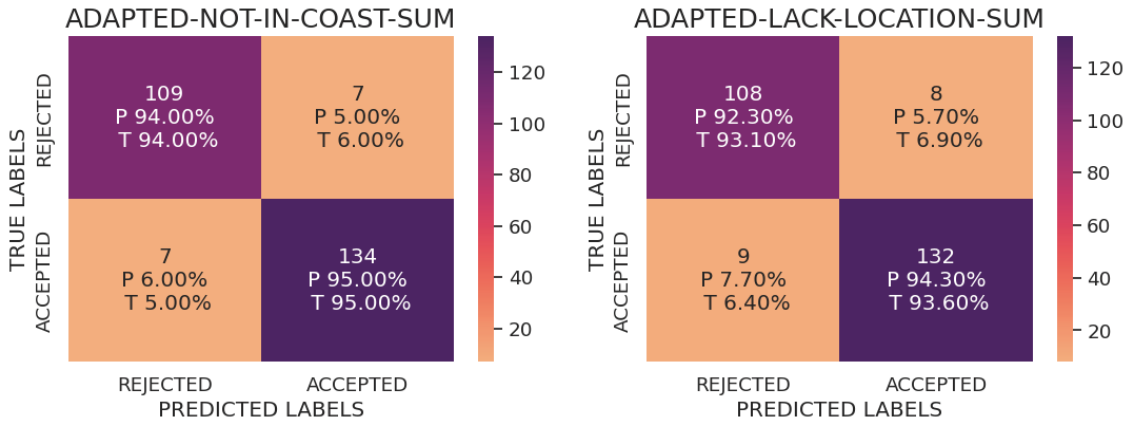


Figure 7.3: Confusion Matrices of combined models using the Sum rule and evaluated with the partial test base. Beyond the classifier's errors and successes in absolute numbers, they show normalized numbers over the predicted label (P), and normalized numbers over the true label (T).

Something similar happens when we observe the results of the models evaluated on the partial test set (Tables 7.3 and 7.5), the precision and F1 Score of the combinations trained adapted-not-in-coast label outperform the combinations trained with adapted-lack-location.

Considering the Sum rule and results of the models evaluated on the partial test set, a significant increase in recall is observed, compared to the results of the other rules. Furthermore, something that can be observed by the confusion matrices 7.3, is that, in this case, the models start to have the same difficulty in recognizing both positive and negative examples. It contrasts with the models analyzed so far, which had more difficulty in recognizing positive examples.

7.1.5 The Best Combined Model

Considering the precision of the models, since a high precision means a lower number of false positives, the best model to classify Instagram posts as accepted or rejected as legitimate occurrence of *Physalia physalis*, combining results of the best textual and image models, is ADAPTED-NOT-IN-COAST-PROD model that combines, using the Product rule, the results of BERT RAW PRE and US RETRAINED trained with adapted-not-in-coast label. This model maximizes precision and F1 Score when evaluated on the partial test set.

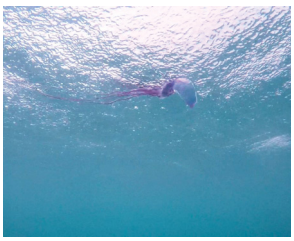
In all, the best model misclassified 40 posts out of 783, 8 of which are false positives. Table 7.6 shows three samples of false positive posts by this model. See Appendix A for a complete list of false positives by this model (Table A.4). In all, 5 posts have images classified as ART, 2 as CNIDARIA and 1 as NOTHING. All these posts misclassified as positive were originally rejected by the oceanographer, 1 being because of media and 7 because of media and location. In fact only the first post displayed in the table would actually be exposed to the classifier in the production scenario.

Table 7.6: Examples of False Positive by Best Combined Model. Showing Image, Caption and Details: Spatial Information (Spatial Info.), Taxonomic Information (Tax. Info.), Original Label and Probabilities

I	Image	Caption (up to 700 characters)	Details
1		EMOJI Mar pequeno tem brincadeira sim pow ! EMOJI - #sc #santacatarina #florianopolis #floripa #ilhadamagia #peace #aloha #surf #corpo #mente #bodysurf #treino #surfdepeito #jacarezinho #swimming #natação #cuidado #caravelaportuguesa #gopro #praialimpa #summer #lixonolixo #verão #island #protetorsolar #h2o #lifestyle #waterman	Spatial Info.: COAST-TEXT Tax. Info.: NOTHING Rejected because MEDIA BERT Probability: 0.999 ResNet50 Probability: 0.025
2		Tenho recebido mensagens e visto relatos das aparições das Caravelas Portuguesas (Physalia physalis) nas nossas praias. Sendo assim venho fazer um alerta sobre os riscos do contato com esses animais. O aparecimento delas nessa época do ano é comum e ocorre por conta da correnteza marítima, só que o que elas tem de exótico, também tem de perigo. Apesar de parecer inofensiva, o problema está no contato com seus tentáculos, eles liberam substâncias extremamente urticantes que podem causar queimaduras de terceiro grau e em alguns casos, pode até mesmo ser fatal. Qualquer contato com a caravela-portuguesa deve ser evitado, mesmo que o animal esteja morto. Caso o banhista seja queimado, não deve t...	Spatial Info.: NOTHING Tax. Info.: ART Rejected because MEDIA AND LOCATION BERT Probability: 0.996 ResNet50 Probability: 0.837
3		Atenção a Caravela Portuguesa A Caravela Portuguesa (Physalia Physalia) é um invertebrado que pertence ao grupo dos Cnidários, o mesmo grupo da água viva, anêmona, entre outros animais. Às caravelas Portuguesas possuem cnidocitos com substância urticante, que são disparados ao contato. O IMPA anunciou e alertou a presença de Caravelas Portuguesas em todo litoral brasileiro, isso é comum nessa época do ano por conta das correntes marítimas, caso vejam uma Caravela Portuguesa na água ou na areia, mantenham distância e chame o salva vidas mais próximo para efetuar a retirada, não entre em contato com o animal, mesmo na areia, seus tentáculos ainda podem liberar as substâncias urticantes. Caso...	Spatial Info.: NOTHING Tax. Info.: ART Rejected because MEDIA AND LOCATION BERT Probability: 0.999 ResNet50 Probability: 0.01

Table 7.7 shows three samples of false negative posts by this model. See Appendix A for a complete list of false negatives by this model (Table A.5). From 32 posts misclassified as negative, 14 were originally rejected by the oceanographer, being all because of location, which means they would not be exposed to the classifier in the production scenario. 18 posts have images classified as REALISTIC and 14 as EDITED.

Table 7.7: Examples of False Negative by Best Combined Model. Showing Index (I), Image, Caption and Details: Spatial Information (Spatial Info.), Taxonomic Information (Tax. Info.), Original Label and Probabilities. Emojis were replaced by EMOJI.

I	Image	Caption (up to 700 characters)	Details
1		Hoje ela deu o ar de sua graça, desfilando em por cima dos mergulhadores ! EMOJI#planetsubfilmes #mar #aguaviva #caravelaportuguesa #portodegalinhas #brasil #sea #ocean #scubadiving #scubaworld #scuba #dive #padi #work #myplace #fish #biodiversidade #living #brasil #nordeste #viverepreservar	Spatial Info.: COAST-TEXT Tax. Info.: REALISTIC ACCEPTED BERT Probability: 0.04 ResNet50 Probability: 0.009
2		Somos tudo aquilo que dizemos que somos... EMOJIEMOJIEMOJIEMOJI EMOJI@flarlesonpedrosa #nova coleção verão 2016/17 @hurley aqui na #FRAN6STORE . Enviamos para todo EMOJI EMOJIEMOJIEMOJIEMOJIEMOJIEMOJIEMOJIEMOJI Obrigado Senhor! #Camping. A sua segunda #house é aqui! Por tudo somos grato! Diárias à partir de R$ 19.90 EMOJIEMOJIAlohaEMOJIEMOJI EMOJI #HOSTEL FRAN6 EMOJI#PASSEIOS EMOJI@FLYBOARDMACEIO EMOJI #Aulasdesurf ReservasEMOJI EMOJI55 82 993925252 EMOJIATENÇÃOEMOJI NÃO ESQUEÇA DE PASSAR O SEU PROTETOR FACIAL...	Spatial Info.: COAST-TEXT Tax. Info.: REALISTIC ACCEPTED BERT Probability: 0.001 ResNet50 Probability: 0.976
3		Alem das aguas da Litoranea em Sao Luis estarem com o status de “Improprias ao banho” hoje foi possivel encontrar milhares de Caravelas Portuguesas Bebes (Physalia physalis) na praia, o que normalmente seria verificado no seu maior periodo de reproducao, em outubro. #aguaviva #aguavivacaravela #aguavivacaravelaportuguesa #caravelaportuguesa #physaliaphysalis #physalis #litoranea #litoraneasaoluis #litoraneasaoluisma #jellyfish #praiasdomaranhao	Spatial Info.: COAST-TEXT Tax. Info.: EDITED ACCEPTED BERT Probability: 0.999 ResNet50 Probability: 0.0

7.2 COMPARATIVE ANALYSIS

In this Section we perform an analysis comparing the results of the best models, first we compare the unimodal models: BERT RAW PRE and US RETRAINED, and than we compare the multimodal model with unimodal ones.

7.2.1 Unimodal Models

We perform an analysis comparing the results of the best model trained only with text, BERT RAW PRE (mBERT model trained with raw data and one that achieved the highest precision among all models trained with text), and the best model trained only with images, US RETRAINED (ResNet50 pre-trained with Imagenet and retrained with our data, the one that we used undersampling as method to deal with imbalanced data and achieved the highest precision among all models trained with image).

Table 7.8 shows the results of the best text and image models. It is possible to notice that the best model trained with image outperforms the best model trained with text by 9% of precision and 7% of F1 Score.

Table 7.8: Comparing Results of the Best Text and Image models. Showing: Model, Precision, Recall and F1 Score.

Model	Precision	Recall	F1
BERT RAW PRE	0.883	0.753	0.813
US RETRAINED	0.969	0.805	0.879

7.2.2 All Models

We perform an analysis comparing the results of the best model trained only with text: BERT RAW PRE, the best model trained only with images: US RETRAINED, and the best combined model: ADAPTED-NOT-IN-COAST-PROD.

Table 7.9 show the results of the best models, evaluated with full and partial test sets.

When evaluated in the full test set as well as in the database that simulates production (partial), the combined model does not surpass the precision of the model trained only with images. Considering F1 Score the combined model outperformed both individual models.

Table 7.9: Comparing Results of the Best Models, evaluated with full and partial test sets. Showing: Model, Precision, Recall and F1 Score

	Full Test Set			Partial Test Set		
Model	Precision	Recall	F1	Precision	Recall	F1
BERT RAW PRE	0.883	0.753	0.813	0.931	0.766	0.840
US RETRAINED	0.969	0.805	0.879	0.992	0.851	0.916
ADAPTED-NOT-IN-COAST-PROD	0.943	0.841	0.889	0.984	0.894	0.937

7.2.3 Discussion

Among the specific objectives proposed by this work is to perform a comparative analysis of the trained models in order to answer the following questions:

- Question: Which model performed best for classifying posts as a legitimate occurrence of *Physalia physalis*? Answer: US RETRAINED, a ResNet50 pre-trained with Imagenet and retrained with our data, the one that we used undersampling as method to deal with imbalanced data and achieved the precision of 97% and F1 Score of 88%.

- Question: Which text-only model had the best performance for identifying legitimate occurrences of *Physalia physalis*? Answer: BERT RAW PRE, multilingual BERT model trained with raw data, which achieved the highest precision among all models trained with text, with the precision of 88% and F1 Score of 81%.

- Question: By analyzing only the text of the post, it is possible to identify occurrences of *Physalia physalis* with good precision? Answer: Even though the results of text-only models did not outperform the results of models trained only with image, it is possible to state that the result obtained by model BERT RAW PRE is a good result, mainly considering the characteristics of the texts analyzed and the difficulty of the task. As argued in Section 5.8, the task is difficult to execute even for a human.

- Question: Is a text-and-image (multimodal) model better at recognizing legitimate occurrences of *Physalia physalis* than text-only or image-only trained models? Answer: Although the combined models did not outperformed the precision of the model trained

only with images, they presented excellent results, with the model combined using the product rule reaching an F1 Score of 94%, that is, 2% higher than the F1 Score of the best model trained with images only.

In this Chapter we performed experiments combining the results of the best unimodal models and presented a comparative analysis of the results obtained with models trained with text and image from posts collected from Instagram. In the next Chapter we conclude the dissertation and present some future works.

8 CONCLUSION

In this work, we explored the problem of identifying legitimate occurrences of *Physalia physalis* in posts extracted from Instagram using classification task. Classifiers were trained using only the text of the post and only one of the images of the post. We also combined the results of the best individual models and performed a comparative analysis of the results obtained.

In preparation for training machine learning models we collected and labelled posts extracted from Instagram as ACCEPTED or REJECTED for legitimate occurrence of *Physalia physalis* on the Brazilian coast.

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. Different text normalization techniques were tried to maximize the performance of the classifiers. In addition, we optimized the model's hyperparameters. In the end, the model trained with text that achieved the best result was BERT RAW PRE, a multilingual BERT model trained with raw data, which achieved the precision of 88% and F1 Score of 81% (see Table 7.9).

Observing the results obtained with normalized text and hyperparameter optimization, it is possible to note that models trained with raw and normalized data have practically the same results, indicating that, for the data used in this dissertation, model optimization does not require prior text normalization.

Considering the performance of the final text models, the difference between the best F1 Score of TF-IDF + LR and mBERT was only 1% (see Table 5.20). This demonstrates that the use of the classic model remains competitive.

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. In the end, the model trained with image that obtained the best result was US RETRAINED, a model retrained with our data. The one that we used undersampling to deal with imbalanced data and achieved the precision of 97% and F1 Score of 88% (see Table 7.9).

Finally, we combined the results of the unimodal models (BERT RAW PRE and US RETRAINED) using different fusion rules, and applying the product rule, we obtained a model (ADAPTED-NOT-IN-COAST-PROD) with precision of 94% and F1 Score of 89% (see Table 7.9).

These 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, indicating that these models can be used to generate a dataset on occurrences of *Physalia physalis*.

The US RETRAINED model was considered the best model for classifying posts as a legitimate occurrence of *Physalia physalis*. This model achieved precision of 99% and F1 Score of 92% when evaluated with the dataset that simulates the production (see Table 7.9). This results demonstrates that this model 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 a resource 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 Edwards et al. (2022). We also did not find other work that combined the results of models trained with different modals, in the context of conservation science and social media data. In fact the use of images and video are more common in this context. In our experiments, models trained with images outperformed models trained with text in 9% of precision and 7% of F1 Score (see Table 7.8), demonstrating that the image was more relevant than the text for recognizing the species researched. The contribution of the image becomes even clearer when we observe the results of the combined model, where it is noted that the results of the combinations always surpassed the results of models trained only with text, but not always surpassed the results obtained by models trained only with image (see Section 7).

During the annotation process we observed that posts rejected because of location could represent noise for training machine learning models. Due to this, we performed experiments with original (i.e., labels assigned by the oceanographer) and adapted labels (i.e., labels considering as accepted posts rejected because of location). Despite the low number of rejected posts because location, the results obtained by the models trained with the adapted labels were significantly higher than results obtained by the models trained with original label. For example, mBERT trained with adapted-lack-location label outperformed the model trained with original label by 13% of F1 Score (see Table 5.1), while Resnet50 trained with adapted-not-in-coast label outperformed the model trained with original label by 9% of F1 Score (see Table 6.1). This results show that the design of appropriate labels can affect the quality of the machine learning model and the unquestionable use of the label assigned by the specialist can be a source of noise. Something that became clear is that spatial information had an impact on the results, being actually a source of noise for the problem researched. It became even more evident when we evaluated the models on the dataset that simulates the production, which does not include posts rejected because of location.

Although we had good results with adapted labels. There could be a refinement of the adapted labels based on the text of the post. As an example, the post that was considered accepted because it was rejected because of the lack of location (see post no. 3 of Table 5.23). But, it is an educational text about cnidarians and not a description of the occurrence of *Physalia physalis*. It should have been kept as rejected.

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 model. 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 differentiate sightings from accidents using taxonomic information.

We make a suggestion that the specialist consider the text of the posts in hers criteria. For example, there is a post that talks about an accident with *Physalia physalis*, but because the image was of part of the body burned by *Physalia physalis*, it was not accepted as a legitimate occurrence (see post no. 2 of Table 5.22). On the other hand, there is a post that talks about a new clothing collection and hostel, which was accepted as a legitimate occurrence because it has a *Physalia physalis* image and the location on the Brazilian coast (see post no. 4 of Table 5.23).

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

8.1 LIMITATIONS

Among the limitations of this work we can mention:

- We could perform experiments with other multimodal approach, classical models, neural network architectures, pre-trained models, methods for dealing with imbalanced data, and so on. However, time constraints forced us to keep the scope of experiments small.
- We also did not explore the model explainability.
- There could have been more iterations to discuss the labeled data. However, as the last labeled spreadsheet was delivered in April/2023, there was no time for further discussions and refinements.

8.2 FUTURE WORKS

This dissertation 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:

Extensions to textual analysis include:

- Perform more experiments with adapted-not-in-coast label.
- Use other techniques to deal with imbalanced data, such as: oversampling and undersampling; even data augmentation can also be used to augment the training dataset as a whole.
- In addition to the captions of the posts, it would be interesting to also use the comments. There are evidences that there is spatial information in the comments of the posts.
- Experiment with BERTimbau (Souza et al., 2020). The argument for experiment with this model is that multilanguage models may not represent all languages equally due to possible underrepresentation during the pre-training step. Therefore, the use of monolanguage models can alleviate this problem (Kalyan et al., 2021).
- Experiment with other multilingual BERTs, such as: XML-RoBERTa[1] and DistilBERT[2].
- Carry out experiments with other neural network architectures (e.g., LSTM, GRU) and classical models (e.g., SVM, RF).
- Use mBERT retrained with our data before using it as a feature extractor may provide better performance than using it without refinement as we did.
- As was demonstrated in the experiments in Section 5.4.1, hashtags have a good discriminative power for our research problem. Therefore, experiments with the application of other normalizations related to hashtags, such as breaking down compound hashtags, could be interesting.

Possible extensions on the media analysis include:

- Perform more experiments with adapted-lack-location label.

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

[2] https://huggingface.co/distilbert-base-multilingual-cased

- Carry out experiments with object detection, in addition to detecting *Physalia physalis*. It could make it possible to count the animals in the images.
- Conduct experiments with all images of the post. Models can be trained with one or more images as input, or transform posts with multiple images into multiple rows in the training dataset.
- Analyse videos of the posts.
- Experiment with new techniques such Vision Transformer (ViT) (Dosovitskiy et al., 2021) to image classification.

Other future lines of investigation can be cited:

- Apply different filters; in this work the filters applied to the data were designed with the aim of using the same dataset for training machine learning models regardless of the data modality (text or image), allowing it to be possible to compare the result of the final models. However, it may be interesting to reconsider these filters, since some filters are only related to text (i.e., exclusion of empty texts, exclusion of similar texts, restriction to posts in Portuguese) and others with media only (i.e., exclusion of only-video posts). In particular, the results of experiments with similarity (Section 5.2) may actually be statistical accidents. Because by randomly removing some posts we may end up removing better or worse texts for model training, and as the percentage of similarity decreases, the chance of this occurring increases. The exception is experiments with 100% of similarity. Therefore, experiments without this filter may be interesting.
- Conduct experiments with multimodal machine learning by training a model fed with features extracted from both text and images (i.e., Early Fusion).
- Perform more experiments combining the results of the models, such as: other fusion rules, and include the model trained with LR.
- Analyze the reels and stories of Instagram.
- Collect and analyse data from other social medias such as: Twitter, Facebook and TikTok.
- Analyse caption and comments for identifying the location and occurrence date.
- Collect and label more data, particularly to increase the number of positive samples, since this was the class that the classifiers had more difficult to predict.
- Train a model to identify accidents with *Physalia physalis* using the taxonomic information.
- Interpret the decision made by the models by applying methods such as: Feature Visualization and Pixel Attribution (Molnar, 2022).

8.3 LIST OF PUBLICATIONS

During the development of this research, the following papers were published:

- ROCHA, Heloisa F.; HARA, Carmem S.. Identificação de Ocorrências do Cnidário Physalia physalis em Dados Extraídos de Mídias Sociais. In: WORKSHOP DE TESES E DISSERTAÇÕES (WTDBD) - SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD), 37. , 2022, Búzios. Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados. Porto Alegre: Sociedade Brasileira de Computação, 2022 . p. 119-125. DOI: https://doi.org/10.5753/sbbd_estendido.2022.21853

- CAMARGO, Leonardo da S.; ROCHA, Heloisa; NASCIMENTO, Lorena S. do; HARA, Carmem. Coleta de Dados do Instagram sobre Ocorrências de Caravelas-Portuguesas na Costa Brasileira. In: ESCOLA REGIONAL DE BANCO DE DADOS (ERBD), 18. , 2023, Palmas/PR. Anais da XVIII Escola Regional de Banco de Dados, 2023. Porto Alegre: Sociedade Brasileira de Computação, 2023 . p. 51-59. ISSN 2595-413X. DOI: https://doi.org/10.5753/erbd.2023.229499.

- Rocha, H.F., Nascimento, L.S., Camargo, L., Noernberg, M., Hara, C.S. (2023). Labeling Portuguese Man-of-War Posts Collected from Instagram. In: Abelló, A., et al. New Trends in Database and Information Systems. ADBIS 2023. Communications in Computer and Information Science, vol 1850. Springer, Cham. https://doi.org/10.1007/978-3-031-42941-5_32

REFERENCES

Abadi, M. et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/.

Aquino, G. G. E. S., Haddad Jr., V., and Pires, V. A. (2019). Avaliação dos acidentes ocorridos por cnidários no município de Salinópolis/Pará (Brasil). *Biota Amazônia (Biote Amazonie, Biota Amazonia, Amazonian Biota)*, 9(4):37–40.

August, T. A., Pescott, O. L., Joly, A., and Bonnet, P. (2020). Ai naturalists might hold the key to unlocking biodiversity data in social media imagery. *Patterns*, 1(7):100116. https://www.sciencedirect.com/science/article/pii/S2666389920301574.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443. https://doi.org/10.1109/TPAMI.2018.2798607.

Bertaglia, T. F. C. and Nunes, M. d. G. V. (2016). Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 112–120.

BigDataCloud (2023). BigDataCloud. https://www.bigdatacloud.com/. Acessado em 15/02/2023.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc. https://www.nltk.org/.

Burkov, A. (2019). *The hundred-page machine learning book*, volume 1. Andriy Burkov Quebec City, QC, Canada. http://themlbook.com/wiki/doku.php.

Cabral., L., Monteiro., J. M., Franco da Silva., J. W., Mattos., C. L., and Mourão., P. J. C. (2021). Fakewhastapp.br: Nlp and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages. In *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS,*, pages 63–74. INSTICC, SciTePress. https://doi.org/10.5220/0010446800630074.

Camargo, L., Rocha, H., Nascimento, L., and Hara, C. (2023). Coleta de dados do instagram sobre ocorrências de caravelas-portuguesas na costa brasileira. In *Anais da XVIII Escola Regional de Banco de Dados*, pages 51–59, Porto Alegre, RS, Brasil. SBC. https://sol.sbc.org.br/index.php/erbd/article/view/24346.

Carneiro, A., Nascimento, L., Noernberg, M., Hara, C., and Pozo, A. (2022). Portuguese man-of-war image classification with convolutional neural networks. *arXiv preprint arXiv:2207.01171*. https://doi.org/10.48550/ARXIV.2207.01171.

Carvalho, A., Faceli, K., Lorena, A., Gama, J., and Almeita, T. (2021). *Inteligência Artificial - uma abordagem de aprendizado de máquina*. Rio de Janeiro: LTC, 2nd edition.

Caseli, H. and Nunes, M., editors (2023). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN. https://brasileiraspln.com/livro-pln.

Cavalcante, M. M. E., Rodrigues, Z. M. R., Hauser-Davis, R. A., Siciliano, S., Haddad Júnior, V., and Nunes, J. L. S. (2020). Health-risk assessment of portuguese man-of-war (physalia physalis) envenomations on urban beaches in são luís city, in the state of maranhão, brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, 53. https://doi.org/10.1590/0037-8682-0216-2020.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357. https://doi.org/10.1613/jair.953.

Cheema, G. S., Hakimov, S., Müller-Budack, E., and Ewerth, R. (2021). On the role of images for analyzing claims in social media. *arXiv preprint arXiv:2103.09602*. https://doi.org/10.48550/arXiv.2103.09602.

Chollet, F. et al. (2015). Keras. https://keras.io.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, 37(1):51–89. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440370103.

Christin, S., Hervet, , and Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644. https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13256.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.747.

Daume, S. (2016). Mining twitter to monitor invasive alien species — an analytical framework and sample information topologies. *Ecological Informatics*, 31:70–82. https://www.sciencedirect.com/science/article/pii/S157495411500196X.

de Oliveira, D. N. and Merschmann, L. H. d. C. (2021). Joint evaluation of preprocessing tasks with classifiers for sentiment analysis in brazilian portuguese language. *Multimedia Tools Appl.*, 80(10):15391–15412. https://doi.org/10.1007/s11042-020-10323-8.

Devlin, J. (2019). Multilingual bert. https://github.com/google-research/bert/blob/master/multilingual.md. Acessado em 10/03/2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Di Minin, E., Fink, C., Hiippala, T., and Tenkanen, H. (2019). A framework for investigating illegal wildlife trade on social media with machine learning. *Conservation Biology*, 33(1):210–213. https://doi.org/10.1111/cobi.13104.

Di Minin, E., Tenkanen, H., and Toivonen, T. (2015). Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, 3. https://www.frontiersin.org/articles/10.3389/fenvs.2015.00063.

Dietterich, T. G. (1997). Machine-learning research. *AI Magazine*, 18(4):97. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1324.

Diniz, E. J., Fontenele, J. E., de Oliveira, A. C., Bastos, V. H., Teixeira, S., Rabêlo, R. L., Calçada, D. B., Dos Santos, R. M., de Oliveira, A. K., and Teles, A. S. (2022). Boamente: A Natural Language Processing-Based Digital Phenotyping Tool for Smart Monitoring of Suicidal Ideation. *Healthcare*, 10(4). https://doi.org/10.3390/healthcare10040698.

Dos Santos, F. L. and Ladeira, M. (2014). The role of text pre-processing in opinion mining on a social media language dataset. In *2014 Brazilian Conference on Intelligent Systems*, pages 50–54. https://doi.org/10.1109/BRACIS.2014.20.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Duailibe, I. C. F. d. S., Coelho, K. K. F., Filgueira, C. H. M. d. S., Nunes, A. R. O. P., Saraiva, A. C. S., and Nunes, J. L. S. (2021). Uso de mÍdias digitais aplicado À estudos de conservaÇÃo do mero epinephelus itajara no litoral amazÔnico brasileiro = use of digital media applied to conservation studies of the atlantic goliath grouper epinephelus itajara (lichtenstein, 1822) in the brazilian amazon coast. *Boletim do Laboratório de Hidrobiologia*, 31(1). https://periodicoseletronicos.ufma.br/index.php/blabohidro/article/view/14230.

Dylewski, Ł., Mikula, P., Tryjanowski, P., Morelli, F., and Yosef, R. (2017). Social media and scientific research are complementary—youtube and shrikes as a case study. *The Science of Nature*, 104(5):1–7. https://doi.org/10.1007/s00114-017-1470-8.

Edwards, T., Jones, C. B., and Corcoran, P. (2022). Identifying wildlife observations on twitter. *Ecological Informatics*, 67:101500. https://doi.org/10.1016/j.ecoinf.2021.101500.

Edwards, T., Jones, C. B., Perkins, S. E., and Corcoran, P. (2021). Passive citizen science: The role of social media in wildlife observations. *PLOS ONE*, 16(8):e0255416. https://doi.org/10.1371/journal.pone.0255416.

ElQadi, M. M., Lesiv, M., Dyer, A. G., and Dorin, A. (2020). Computer vision-enhanced selection of geo-tagged photos on social network sites for land cover classification. *Environmental Modelling Software*, 128:104696. https://www.sciencedirect.com/science/article/pii/S1364815219308059.

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., and Dehmer, M. (2020). An introductory review of deep learning for prediction models with big data. *Frontiers in Artificial Intelligence*, 3. https://www.frontiersin.org/articles/10.3389/frai.2020.00004.

Emoji (2023). Emoji. https://pypi.org/project/emoji/. Acessado em 10/03/2023.

Endo, P. T., Santos, G. L., Xavier, M. E. d. L., Campos, G. R. N., de Lima, L. C., Silva, I., Egli, A., and Lynn, T. (2022). Illusion of Truth: Analysing and Classifying COVID-19 Fake News in Brazilian Portuguese Language. *Big Data and Cognitive Computing*, 6(2). https://doi.org/10.3390/bdcc6020036.

Feitosa, M., Ferreira, C., Gonçalves, G., and Almeida, J. (2022). Análise da percepção das pessoas no twitter sobre ações policiais. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 73–84, Porto Alegre, RS, Brasil. SBC. https://sol.sbc.org.br/index.php/brasnam/article/view/20518.

Ferreira-Bastos, D. M. R., Haddad Jr., V., and Nunes, J. L. S. (2017). Human envenomations caused by Portuguese man-of-war (Physalia physalis) in urban beaches of São Luis City, Maranhão State, Northeast Coast of Brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, 50(1):130–134. https://doi.org/10.1590/0037-8682-0257-2016.

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

Forte Martins, A. D., Cabral, L., Chaves Mourão, P. J., Monteiro, J. M., and Machado, J. (2021). Detection of misinformation about covid-19 in brazilian portuguese whatsapp messages. In *Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, Proceedings*, page 199–206, Berlin, Heidelberg. Springer-Verlag. https://doi.org/10.1007/978-3-030-80599-9_18.

Gasparetto, A., Marcuzzo, M., Zangari, A., and Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. *Information*, 13(2). https://www.mdpi.com/2078-2489/13/2/83.

Ghermandi, A. and Sinclair, M. (2019). Passive crowdsourcing of social media in environmental research: A systematic map. *Global Environmental Change*, 55:36–47. https://www.sciencedirect.com/science/article/pii/S0959378018309920.

González-Carvajal, S. and Garrido-Merchán, E. C. (2020). Comparing bert against traditional machine learning text classification. *arXiv preprint*. https://doi.org/10.48550/arXiv.2005.13012.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Goodwin, M., Halvorsen, K. T., Jiao, L., Knausgård, K. M., Martin, A. H., Moyano, M., Oomen, R. A., Rasmussen, J. H., Sørdalen, T. K., and Thorbjørnsen, S. H. (2022). Unlocking the potential of deep learning for marine ecology: overview, applications, and outlook. *ICES Journal of Marine Science*, 79(2):319–336. https://doi.org/10.1093/icesjms/fsab255.

Guo, W., Wang, J., and Wang, S. (2019). Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394. https://doi.org/10.1109/ACCESS.2019.2916887.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*. https://doi.org/10.48550/arXiv.1512.03385.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://spacy.io/.

IGfonts (2023). Instagram fonts generator. https://igfonts.io/. Acessado em 19/03/2023.

Instaloader (2023). Instaloader. https://github.com/instaloader/instaloader. Acessado em 15/02/2023.

Jarić, I., Correia, R. A., Brook, B. W., Buettel, J. C., Courchamp, F., Di Minin, E., Firth, J. A., Gaston, K. J., Jepson, P., Kalinkat, G., Ladle, R., Soriano-Redondo, A., Souza, A. T., and Roll, U. (2020). iecology: Harnessing large online resources to generate ecological insights. *Trends in Ecology Evolution*, 35(7):630–639. https://www.sciencedirect.com/science/article/pii/S016953472030077X.

Jurafsky, D. and Martin, J. H. (2021). *Speech and language processing*. 3rd draft edition.

Kalyan, K. S., Rajasekharan, A., and Sangeetha, S. (2021). Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*. https://doi.org/10.48550/arXiv.2108.05542.

Kittler, J., Hater, M., and Duin, R. (1996). Combining classifiers. In *Proceedings of 13th International Conference on Pattern Recognition*, volume 2, pages 897–901 vol.2. https://doi.org/10.1109/ICPR.1996.547205.

Kougia, V. and Pavlopoulos, J. (2021). Multimodal or text? retrieval or BERT? benchmarking classifiers for the shared task on hateful memes. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 220–225, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.woah-1.24.

Kubat, M. (2017). *An introduction to machine learning*, volume 2. Springer. https://doi.org/10.1007/978-3-030-81935-4.

Kulkarni, R. and Di Minin, E. (2021). Automated retrieval of information on threatened species from online sources using machine learning. *Methods in Ecology and Evolution*, 12(7):1226–1239. https://doi.org/10.1111/2041-210X.13608.

Langdetect (2023). Langdetect. https://github.com/Mimino666/langdetect. Acessado em 15/02/2023.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics. https://aclanthology.org/2020.aacl-main.91.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2020). *Mining of massive data sets*. Cambridge university press, 3 edition. https://doi.org/10.1017/9781108684163.

Levenshtein (2023). Levenshtein pypi. https://pypi.org/project/python-Levenshtein/. Acessado em 05/04/2023.

Machado, P., Fernandes, B., and Novais, P. (2022). Benchmarking data augmentation techniques fornbsp;tabular data. In *Intelligent Data Engineering and Automated Learning – IDEAL*

2022: 23rd International Conference, IDEAL 2022, Manchester, UK, November 24–26, 2022, Proceedings, page 104–112, Berlin, Heidelberg. Springer-Verlag. https://doi.org/10.1007/978-3-031-21753-1_11.

Mazars-Simon, A. E. (2019). The Wild in Live Project: A Human/Algorithm learning network to help citizen science in wildlife conservation. Master's thesis, Universidade de Coimbra. http://hdl.handle.net/10316/88052.

Meta (2023). Instagram help center. https://help.instagram.com. Acessado em 15/02/2023.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3). https://doi.org/10.1145/3439726.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press, 2nd edition. https://mitpress.ublish.com/ebook/foundations-of-machine-learning--2-preview/7093.

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition. https://christophm.github.io/interpretable-ml-book.

Morais, P., Afonso, L., and Dias, E. (2021). Harnessing the Power of Social Media to Obtain Biodiversity Data About Cetaceans in a Poorly Monitored Area. *Frontiers in Marine Science*, 8. http://hdl.handle.net/10400.1/17403.

Mota, A., Franco, W., and Mattos, C. (2021). Detecção de desinformação sobre covid-19 no twitter. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 172–181, Porto Alegre, RS, Brasil. SBC. https://sol.sbc.org.br/index.php/stil/article/view/17796.

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

Nascimento, L. S., Noernberg, M. A., Bleninger, T. B., Hausen, V., Pozo, A., Camargo, L. S., Hara, C., and Nogueira Júnior, M. (2022). Social media in service of marine ecology: new observations of the ghost crab ocypode quadrata (fabricius, 1787) scavenging on portuguese man-of-war physalia physalis (linnaeus, 1758). *Aquatic Ecology*, 56(3):859–864. https://doi.org/10.1007/s10452-022-09947-9.

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. Determination Press. http://neuralnetworksanddeeplearning.com.

Ofli, F., Alam, F., and Imran, M. (2020). Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response. *arXiv preprint arXiv:2004.11838*. https://doi.org/10.48550/arXiv.2004.11838.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. https://scikit-learn.org/stable/.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding with unsupervised learning.

Rocha, H. and Hara, C. (2022). Identificação de ocorrências do cnidário physalia physalis em dados extraídos de mídias sociais. In *Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 119–125, Porto Alegre, RS, Brasil. SBC. https://sol.sbc.org.br/index.php/sbbd_estendido/article/view/21853.

Rocha, H. F., Nascimento, L. S., Camargo, L., Noernberg, M., and Hara, C. S. (2023). 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., and Zumpano, E., editors, *New Trends in Database and Information Systems*, pages 369–381, Cham. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42941-5_32.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252. https://doi.org/10.1007/s11263-015-0816-y.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

Santamaria, S., Enghoff, H., and Reboleira, A. S. (2020). The first laboulbeniales (ascomycota, laboulbeniomycetes) from an american millipede, discovered through social media. *MycoKeys*, 67:45–53. https://doi.org/10.3897/mycokeys.67.51811.

Shyam, R. (2021). Convolutional Neural Network and its Architectures. *Journal of Computer Technology & Applications*, 12(2):6–14p.

Souza, E., Costa, D., Castro, D. W., Vitório, D., Teles, I., Almeida, R., Alves, T., Oliveira, A. L., and Gusmão, C. (2018). Characterising text mining: a systematic mapping review of the portuguese language. *IET Software*, 12(2):49–75. https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-sen.2016.0226.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-61377-8_28.

Stiilpen Junior, M. and Merschmann, L. H. C. (2016). A methodology to handle social media posts in brazilian portuguese for text mining applications. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, Webmedia '16, page 239–246, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/2976796.2976845.

Szeliski, R. (2022). *Computer Vision : Algorithms and Applications*. Springer Cham, 2nd edition. http://szeliski.org/Book/.

Taklis, C., Giovos, I., and Karamanlidis, A. A. (2020). Social media: a valuable tool to inform shark conservation in Greece. *Mediterranean Marine Science*, 21(3):493–498. https://doi.org/10.12681/mms.22165.

Talebi, H. and Milanfar, P. (2021). Learning to resize images for computer vision tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 497–506. https://doi.org/10.48550/arXiv.2103.09950.

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H., and Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation*, 233:298–315. https://www.sciencedirect.com/science/article/pii/S0006320718317609.

Unidecode (2023). Unidecode pypi. https://pypi.org/project/Unidecode/. Acessado em 19/03/2023.

Vargas, F. A., de Góes, F. R., Carvalho, I., Benevenuto, F., and Pardo, T. A. S. (2021). Contextual lexicon-based approach for hate speech and offensive language detection. *arXiv preprint arXiv:2104.12265*. https://doi.org/10.48550/arXiv.2104.12265.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*. https://doi.org/10.48550/arXiv.1609.08144.

Xu, X., Liu, H., Tao, G., Xuan, Z., and Zhang, X. (2022). Checkpointing and deterministic training for deep learning. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, CAIN '22, page 65–76, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3522664.3528605.

Zafarani, R., Abbasi, M. A., and Liu, H. (2014). *Social Media Mining: An Introduction*. Cambridge University Press, New York, NY, USA. http://www.socialmediamining.info.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*. https://d2l.ai/index.html.

Zheng, A. and Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.".

APPENDIX A – EXTENSION OF RESULTS

This appendix presents the extent of some results and errors of the classifiers.

Table A.1 shows the complete list of false positive by final text model.

Table A.1: Complete List of False Positive by Final Text Models. Showing Index (I), Caption, Spatial Information (Spatial Info.), Taxonomic Information (Tax. Info.), Reason for Rejection (R. Rejection) and mean Probability with standard deviation (Prob). Emojis were replaced by EMOJI

I	Caption (up to 250 characters)	Spatial Info.	Tax. Info.	R. Rejection	Prob
1	Um pequeno lembrete: "Com Caravela não se brinca!" #portuguesemanowar #burned #caravelaportuguesa #marinelife #naturelovers #sandandsea	COAST-GEO	ACCIDENT	MEDIA	0.95 (0.063)
2	Bebezíneo de Caravela Portuguesa (foi o que disseram pra gente). #pará #salinopolis #praiadascorvinas #regiaonorte #turismobrasil #caravelaportuguesa #jambutour	COAST-TEXT	CNIDARIA	MEDIA	0.91 (0.138)
3	Milhares delas, pequeninas e venenosas... #poison #águasvivas #caravelaportuguesa #beach #swimming #ocean #atlantic #ericeira #portugal #portuguesemanofwar #wild#surf	NOTHING	CNIDARIA	MEDIA AND LOCATION	0.89 (0.171)
4	EMOJI #dialindo #muitoamor #caravelaportuguesa #eupraiana	COAST-GEO	CNIDARIA	MEDIA	0.91 (0.136)
5	Atenção a Caravela Portuguesa A Caravela Portuguesa (Physalia Physalia) é um invertebrado que pertence ao grupo dos Cnidários, o mesmo grupo da água viva, anêmona, entre outros animais. Às caravelas Portuguesas possuem cnidocitos com substância urticante,...	NOTHING	ART	MEDIA AND LOCATION	0.99 (0.022)
6	Caravela-portuguesa Flutua rosa e azul EMOJI Se move com o vento . . . . . #ruananegri #litora #aguaviva #caravelaportuguesa #vidamarinha #encontros #inspiracaododia #marsp #vidalinda #naturezas	COUNTRYSIDE-GEO	REALISTIC	LOCATION	0.99 (0.012)
7	Beira Mar EMOJI em Cananéia ,,, tarde ótima ! #cananeiasp #caravelaportuguesa	COAST-TEXT	NOTHING	MEDIA	0.91 (0.116)
8	EMOJIEMOJI Beleza que deve ser admirada de longe. Assim como várias outras espécies de #cnidários, a caravela-portuguesa, Physalia physalis, possui ao longo dos tentáculos células do tipo cnidócitos, as quais podem causar injúrias ao entrar em contato com nossa pele...	NOTHING	ART	MEDIA AND LOCATION	0.87 (0.202)

Table A.2 shows the complete list of false positive by BERT RAW PRE model.

Table A.2: Complete List of False Positive by BERT RAW PRE. Showing Index (I), Caption, Spatial Information (Spatial Info.), Taxonomic Information (Tax. Info.), Reason for Rejection (R. Rejection) and Probability (Prob). Emojis were replaced by EMOJI.

I	Caption (up to 250 characters)	Spatial Info.	Tax. Info.	R. Rejection	Prob
1	A caravela-portuguesa é o único organismo em colônia heteromorfa, no grupo dos cnidários. Ou seja, são seres que vivem em colônia, isto é, estão conectados anatomicamente e não sendo um único ser...	COUNTRYSIDE-GEO	REALISTIC	LOCATION	1.0
2	Um pequeno lembrete: "Com Caravela não se brinca!" #portuguesemanowar #burned #caravelaportuguesa #marinelife #naturelovers #sandandsea	COAST-GEO	ACCIDENT	MEDIA	0.97
3	Bebezíneo de Caravela Portuguesa (foi o que disseram pra gente). #pará #salinopolis #praiadascorvinas #regiaonorte #turismobrasil #caravelaportuguesa #jambutour	COAST-TEXT	CNIDARIA	MEDIA	1.0
4	Caravela-portuguesa Flutua rosa e azul EMOJI Se move com o vento . . . . . #ruananegri #litora #aguaviva #caravelaportuguesa #vidamarinha #encontros #inspiracaododia...	COUNTRYSIDE-GEO	REALISTIC	LOCATION	1.0
5	No verão é comum o aparecimento de caravelas e águas-vivas no litoral brasileiro , você sabe como diferencia-las? Vem que o CEMP te explica!EMOJI #agua #aguaviva #caravelas #verão #vivaaciência	NOTHING	NOTHING	MEDIA AND LOCATION	0.86
6	Praia de sábado! . . . . . #aguaviva #praia #floripa #suldailha #omelhor #lugar #do #mundo EMOJI	COAST-TEXT	CNIDARIA	MEDIA	0.6
7	EMOJI #dialindo #muitoamor #caravelaportuguesa #eupraiana	COAST-GEO	CNIDARIA	MEDIA	1.0
8	EMOJIAlgumas dicas para você aproveitar essa época linda do ano: VERÃO! . . . . . EMOJI Feliz Ano Novo! EMOJI #celulasincriveis #biology #verao #felizanonovo #2020 #eudefendoauniversidade...	NOTHING	NOTHING	MEDIA AND LOCATION	1.0
9	Atenção a Caravela Portuguesa A Caravela Portuguesa (Physalia Physalia) é um invertebrado que pertence ao grupo dos Cnidários, o mesmo grupo da água viva, anêmona, entre outros animais. Às...	NOTHING	ART	MEDIA AND LOCATION	1.0
10	#caravelaportuguesa #portuguesemanofwar #alforreca #jellyfish #arte #art #engraving #gravura #calcografia #etching #aquatint #nature #natureza A caravela Portuguesa vai estar disponível na...	NOTHING	NOTHING	MEDIA AND LOCATION	1.0
11	VOCÊ SABIA? OS PERIGOS DAS ÁGUAS VIVAS E DAS CARAVELAS Espécie muito semelhante às temidas águas-vivas, a Caravela (Physalia physalis) possui os mesmos tentáculos munidos...	NOTHING	ART	MEDIA AND LOCATION	1.0
12	EMOJI Mar pequeno tem brincadeira sim pow ! EMOJI - #sc #santacatarina #florianopolis #floripa #ilhadamagia #peace #aloha #surf #corpo #mente #bodysurf #treino #surfdepeito #jacarezinho...	COAST-TEXT	NOTHING	MEDIA	1.0
13	Milhares delas, pequeninas e venenosas... #poison #águasvivas #caravelaportuguesa #beach #swimming #ocean #atlantic #ericeira #portugal #portuguesemanofwar #wild#surf	NOTHING	CNIDARIA	MEDIA AND LOCATION	1.0
14	#caravelaportuguesa #cnidarios	NOTHING	NOTHING	MEDIA AND LOCATION	0.91
15	Cuidado! água-viva. Técnicas lendárias do homem sereia e mexilhãozinho. . . . . #maceioalagoas #maceio #maceió #beach #praia #praiadofrancês...	COAST-TEXT	NOTHING	MEDIA	1.0
16	Beira Mar EMOJI em Cananéia „, tarde ótima ! #cananeiasp #caravelaportuguesa	COAST-TEXT	NOTHING	MEDIA	1.0
17	“ Fim de tarde foi a forma que Deus encontrou de nos lembrar de agradecer por mais um dia”. #fimdetarde #maisumdia #convida20 #aracaju #praia #beach #caravelaportuguesa #sunset	COAST-GEO	NOTHING	MEDIA	1.0
18	#praia #aguaviva #cnidarios #celenterados #cnidoblasto	COUNTRYSIDE-GEO	CNIDARIA	MEDIA AND LOCATION	1.0
19	EMOJIEMOJI Beleza que deve ser admirada de longe. Assim como várias outras espécies de #cnidários, a caravela-portuguesa, Physalia physalis, possui ao longo dos tentáculos células do tipo cnidócitos...	NOTHING	ART	MEDIA AND LOCATION	1.0

Table A.3 shows the complete list of false negative by BERT RAW PRE model.

Table A.3: Complete List of False Negative by BERT RAW PRE. Showing Index (I), Caption, Spatial Information (Spatial Info.), Taxonomic Information (Tax. Info.), when REJECTED the Reason for Rejection (R. Rejection) and Probability (Prob). Emojis were replaced by EMOJI.

I	Caption (up to 400 characters)	Spatial Info.	Tax. Info.	R. Rejection	Prob
1	Hoje foi registrado novamente o aparecimento de caravelas pelas praias de Porto belo EMOJI Fiquem atentos! Registro da seguidora @mariana_vycente #praia #aguaviva #ft #litoral #portobelo #bombinhas #sc #brasil #caixadaço #jetski #banhistas #alerta	COAST-TEXT	REALISTIC	ACCEPTED	0.0
2	#pernambuco #caravelaportuguesa	COAST-GEO	REALISTIC	ACCEPTED	0.002
3	EMOJI Os cnidários, também conhecidos como celenterados, são animais que se reúnem no filo Cnidaria. . . . EMOJIA maioria dos cnidários é encontrada em ambientes marinhos, e os principais representantes desse grupo são as águas-vivas, as anêmonas-do-mar, os corais e as caravelas. . . . EMOJI Existem duas formas de cnidários: os pólipos e as medusas. . .	NOTHING	REALISTIC	LOCATION	0.0
4	Somos tudo aquilo que dizemos que somos... EMOJIEMOJIEMOJIEMOJI EMOJI@flarlesonpedrosa #nova coleção verão 2016/17 @hurley aqui na #FRAN6STORE . Enviamos para todo EMOJI EMOJIEMOJIEMOJIEMOJIEMOJIEMOJIEMOJIEMOJI Obrigado Senhor! #Camping. A sua segunda #house é aqui! Por tudo somos grato! Diárias à partir de R$ 19.90 EMOJIEMOJIAlohaEMOJIEMOJI EMOJI #HOSTEL FRAN6...	COAST-TEXT	REALISTIC	ACCEPTED	0.0
5	Que linda né? Tinham muitas dessas hoje na praia, Benício e o papai foram se divertir por lá e acabaram apreciando essas belezas. #aguaviva #praiadeperuibe #belezadapraia #belezadanatureza	COAST-TEXT	REALISTIC	ACCEPTED	0.002
6	Não toque nestes animais. Fique por perto garantindo que ninguém se machuque e peça para outra pessoa chamar o guarda-vidas para retirá-la da praia. Verifique o mar antes de entrar se não há outras. Lembre que seus tentáculos causam queimaduras e podem medir até 40 metros de comprimento! #atenção #CaravelaPortuguesa #SaveThePlanet #marineanimals	NOTHING	EDITED	LOCATION	0.0
7	Na floresta de algas, um monstro marinho se aproximava. Ao avistar uma caravela-portuguesa no mar, afaste-se, e se encontrar na areia, também evite tocar ou pisar, você pode sofrer sérias lesões por queimaduras. . #rambo #ramborambora #praia #caravela #caravelaportuguesa #monstro #musculo #muscle #boneco #brinquedo #stallone #summer #sdv #follow4followback	NOTHING	REALISTIC	LOCATION	0.001
8	ATENÇÃO! EMOJI . Período de ventos fortes nas nossas praias, e com eles, aumento da incidência de caravelas na zona de banho EMOJI . Já fizemos aqui alguns textos dando orientações como proceder em caso de acidente com estas.EMOJI . Fiquem atentos principalmente com as crianças, que podem achar na areia e pegar para brincar.EMOJI . Foto: SD Gama...	COAST-GEO	EDITED	ACCEPTED	0.0
9	Oi pessoas, tudo bem? Então vou explicar pra vocês um pouco sobre esse organismo que muitos confundem com águas-vivas. A CARAVELA. Caravela, organismo colonial. Vivem em alto-mar e possui longos tentáculos de até 20 metros ou mais. As substâncias urticantes que fabrica podem causar sérias queimaduras em seres humanos. Sua fisgada pode ser muito dolorosa...	COAST-GEO	REALISTIC	ACCEPTED	0.032
10	Caravela-portuguesa Portuguese man-of-war EMOJIEMOJI . . . . . #portuguesemanofwar #seaanimals #poison #caravelaportuguesa #aguaviva #usa #france #italy #england #mexico #london #riodejaneiro #saoluisma #southamerica #brazil #minhafotocomzenfone	COAST-TEXT	REALISTIC	ACCEPTED	0.0
11	Elas "andem" aí... #caravelaportuguesa	NOTHING	REALISTIC	LOCATION	0.122
12	#portuguese #manowar #caravelaportuguesa	NOTHING	REALISTIC	LOCATION	0.161
13	Pra completar a trilogia do dia no feed aqui vai a água viva ou medusa que me queimou hoje mais cedo com seus lindos tentáculos. #aguaviva #mar #praia	NOTHING	EDITED	LOCATION	0.001
14	Caravela-Portuquesa EMOJIEMOJI #caraveladomar #caravelaportuguesa #mar #animais #animaismarinhos #verao #ilheus #mar #sea #vidamarinha #amoios #joiadoatlantico #foconafoto #photo #photografy #bahia	COAST-TEXT	REALISTIC	ACCEPTED	0.044
15	Exotica, perigosa, fascinante. Da série: tesouros do mar. Salve-se quem puder, rs #ferias #caravelaportuguesa #danger #perigo #pericolo	NOTHING	REALISTIC	LOCATION	0.0
16	Caravela Portuguesa EMOJI #caravelaportuguesa	NOTHING	REALISTIC	LOCATION	0.077
17	Hoje ela deu o ar de sua graça, desfilando em por cima dos mergulhadores ! EMOJI#planetsubfilmes #mar #aguaviva #caravelaportuguesa #portodegalinhas #brasil #sea #ocean #scubadiving #scubaworld #scuba #dive #padi #work #myplace #fish #biodiversidade #living #brasil #nordeste #viverepreservar	COAST-TEXT	REALISTIC	ACCEPTED	0.0
18	Oi, humana! Nem chegue perto de mimEMOJIEMOJI (olha essa coooooor, impactada EMOJI)	COAST-GEO	REALISTIC	ACCEPTED	0.0
19	Se encostar = Tomar na Jabiraca EMOJI 3 linda porém perigosa #caravela-portuguesa #purple #blue #green #maceio #paripueira	COAST-TEXT	REALISTIC	ACCEPTED	0.0

Continuation of Table A.3

I	Caption (up to 400 characters)	Spatial Info.	Tax. Info.	R. Rejection	Prob
20	Sim, isso é uma água-viva! E eu estou muito encantada por vê-la tão perfeitamente e tão de perto. . . #amazing #jellyfish #marinelife #aguaviva #salvador #tonabahia #viajarfazbem #traveling #caravelaportuguesa #wanderlust #pic #photography #motozplay #motorola . *Edit. Não é uma água-viva EMOJI, é uma CARAVELA-PORTUGUESA! Hahah!	COAST-TEXT	REALISTIC	ACCEPTED	0.0
21	#caravelaportuguesa EMOJIEMOJIEMOJI	NOTHING	REALISTIC	LOCATION	0.099
22	#caravelaportuguesa #garrafaazul	NOTHING	REALISTIC	LOCATION	0.001
23	Caravela catalisadora de fortes emoções	COAST-GEO	REALISTIC	ACCEPTED	0.011
24	Por @mayy.castroo • • • • • #caravelaportuguesa #itapoa	COAST-TEXT	REALISTIC	ACCEPTED	0.322
25	Monstro na praia, seria um pokèmon ? #aguaviva #purple #sealife #aracaju #ig_sergipe #pokemon #portuguesemanofwar #vidamarinha #igerssergipe #caravelaportuguesa	COAST-TEXT	REALISTIC	ACCEPTED	0.002
26	EMOJIEMOJIEMOJI #caravelaportuguesa	NOTHING	REALISTIC	LOCATION	0.084
27	O mar, suas belezas e suas surpresas. #caravelaportuguesa #maeeujuroquetavalonge Amamos tanto o mar que se desse pra vir direto do aeroporto com as malas a gente vinha. Deu tempo nem de colocar a sunga #meugaroto #minifériаs #1de5	COAST-GEO	REALISTIC	ACCEPTED	0.001
28	A caravela-portuguesa (Physalia physalis) é o único organismo em colônia heteromorfa, no grupo dos cnidários. Ou seja, são seres que vivem em colônia, isto é, estão conectados anatomicamente e não sendo um único ser. E são divididos em duas partes: região subnatural e região natural, as duas estando opostas. Vivem nas águas de todas as regiões tropicais dos oceanos...	COAST-GEO	REALISTIC	ACCEPTED	0.002
29	#aracaju #aracajusergipe #caravelaportuguesa	COAST-TEXT	REALISTIC	ACCEPTED	0.001
30	As Caravelas Portuguesas são mais perigosas do que as água vivas, porque seus tentáculos são longos e cheios de células urticantes que em contato com a pele, podem provocar queimaduras de até terceiro grau ou até mesmo uma parada cardíaca. . Em Florianópolis, elas já estão sendo encontradas nas praias. . Durante a temporada, nos locais em que for constatada a presença...	COAST-TEXT	EDITED	ACCEPTED	0.0
31	As caravelas portuguesas estão marcando presença neste novembro em Capão. Tome cuidado, a queimadura desta espécie de água-viva é muito mais danosa do que as nossas tradicionais mães d'água. EMOJI @rcardosofotos	COAST-GEO	REALISTIC	ACCEPTED	0.002
32	Um achado exótico mais lindo EMOJI EMOJIEMOJIEMOJIEMOJIEMOJI #caravelaportuguesa	COAST-GEO	REALISTIC	ACCEPTED	0.001
33	#caravelaportuguesa #running #corridaderua #foconoobjetivo #treinofinalizado	COAST-GEO	EDITED	ACCEPTED	0.0
34	. agua UNK viva . EMOJI #caravelaportuguesa P.S: Achei ela a cara do Guyodai do Changeman! EMOJI	COAST-GEO	REALISTIC	ACCEPTED	0.0
35	Se não bastassem o corona e os tubarões... #caravelaportuguesa	COAST-GEO	REALISTIC	ACCEPTED	0.001
36	Alem das aguas da Litoranea em Sao Luis estarem com o status de "Improprias ao banho" hoje foi possivel encontrar milhares de Caravelas Portuguesas Bebes (Physalia physalis) na praia, o que normalmente seria verificado no seu maior periodo de reproducao, em outubro. #aguaviva #aguavivacaravela #aguavivacaravelaportuguesa #caravelaportuguesa #physaliaphysalis #physalis #litoranea	COAST-TEXT	EDITED	ACCEPTED	0.038
37	Nome científico: Physalia physalis Classificação superior: Physalia Classificação: Espécie Ordem: Siphonophora Filo: Cnidaria Classe: Hydrozoa Todas essas belezinhas estavam na praia de cotovelo hoje de manhã #caravelaportuguesa #cotovelobeach	COAST-TEXT	REALISTIC	ACCEPTED	0.0
38	Pra quem quis brincar no parquinho, o ingresso. #surf #surfing #natural #caravelaportuguesa #goodvibes #sealife #gopro #bahia #gloriaaDeus #gratidao	NOTHING	REALISTIC	LOCATION	0.0
39	Mais lembranças de Aracajú EMOJIEMOJIEMOJIEMOJIEMOJIEMOJI #Tbt #caravelaportuguesa #riovazabarris #aracaju #praiaatalaia #ilhadosnamorados #croadogore #orladoatalaia #aracajusergipe #sergipe #nordeste #brasil #lagoadostambaquis #largodagentesergipana #museudagentesergipana #brasilEMOJI	COAST-TEXT	REALISTIC	ACCEPTED	0.001
41	Cnidários á deriva EMOJI EMOJI #zoo #animals #biologist #science #scientist #biologo #zoologia #cnidario #animaismarinhos #invertebrado	COAST-GEO	REALISTIC	ACCEPTED	0.0
42	Você conhece os moradores de Icarai de Amontada? EMOJIEMOJIEMOJI Essa é a Caravela Portuguesa que possui nome científico de Physalia physalis, essa colônia de animais pertence ao mesmo grupo dos corais, anêmonas e águas-vivas, o filo dos cnidários. São compostos por um agrupamento de vários organismos do mesmo filo, cada um cumprindo um papel para a...	COAST-TEXT	REALISTIC	ACCEPTED	0.001
43	EMOJI CUIDADO EMOJI Calor, praia e água-viva: em um único dia, litoral tem 36 pessoas com lesões causadas por contato com animal Número corresponde a 22% do número de casos de outubro até 19 de dezembro 22/12/2021 - 12h24min #litoral #aguaviva #mar #calor #verao #veraneio #noticia	NOTHING	EDITED	LOCATION	0.002

Continuation of Table A.3

I	Caption (up to 400 characters)	Spatial Info.	Tax. Info.	R. Rejection	Prob
44	Filos do reino animal: CNIDARIOS Na sequência dos filos do reino animal, o próximo a se falar são os Cnidarios. Fazem parte dos cnidarios: medusas ou águas-vivas; caravelas; anemonas; hidras e corais. Esses animais são encontrados em oceanos de quase todo o mundo e são bem conhecidos, afinal, quem nunca ficou com medo de ir na praia e ser queimado por uma água-viva?...	NOTHING	REALISTIC	LOCATION	0.0
45	Um registro especial dessa interação entre animais, pelo fotógrafo @derluh-dantas Falamos um pouco sobre o Grauçá (Ocypode quadrata (Fabricius, 1787)) e iremos conhecer um pouco mais sobre seus hábitos alimentares. São classificados como carnívoros, essencialmente predadores, porém são carniceiros ocasionais. Sua dieta principal depende muito do ambiente,...	COAST-GEO	EDITED	ACCEPTED	0.0
46	EMOJIEMOJI EMOJILembre ! LAVAR COM ÁGUA SALGADA DO MAR E VINAGRE ! NÃO LAVAR COM ÁGUA DOCE ! EMOJIEMOJIEMOJI —- EMOJIEMOJI Um banho no mar ou uma caminhada na areia pode ser interrompido de maneira bastante desagradável por uma “queimadura” de cnidário, como as águas-vivas e as caravelas. Os primeiros sintomas são vermelhidão, ardência e dor intensa no local afetado, que podem durar de...	COAST-GEO	REALISTIC	ACCEPTED	0.0
47	Passeio pela praia... #praia #peruibecity #chuva #nublado #calor #quarta #bike #aguaviva #cnidários	COAST-TEXT	EDITED	ACCEPTED	0.104

End of Table

Figures A.1, A.2, A.3 and A.4 show the complete list of false negative by best image model (US RETRAINED).

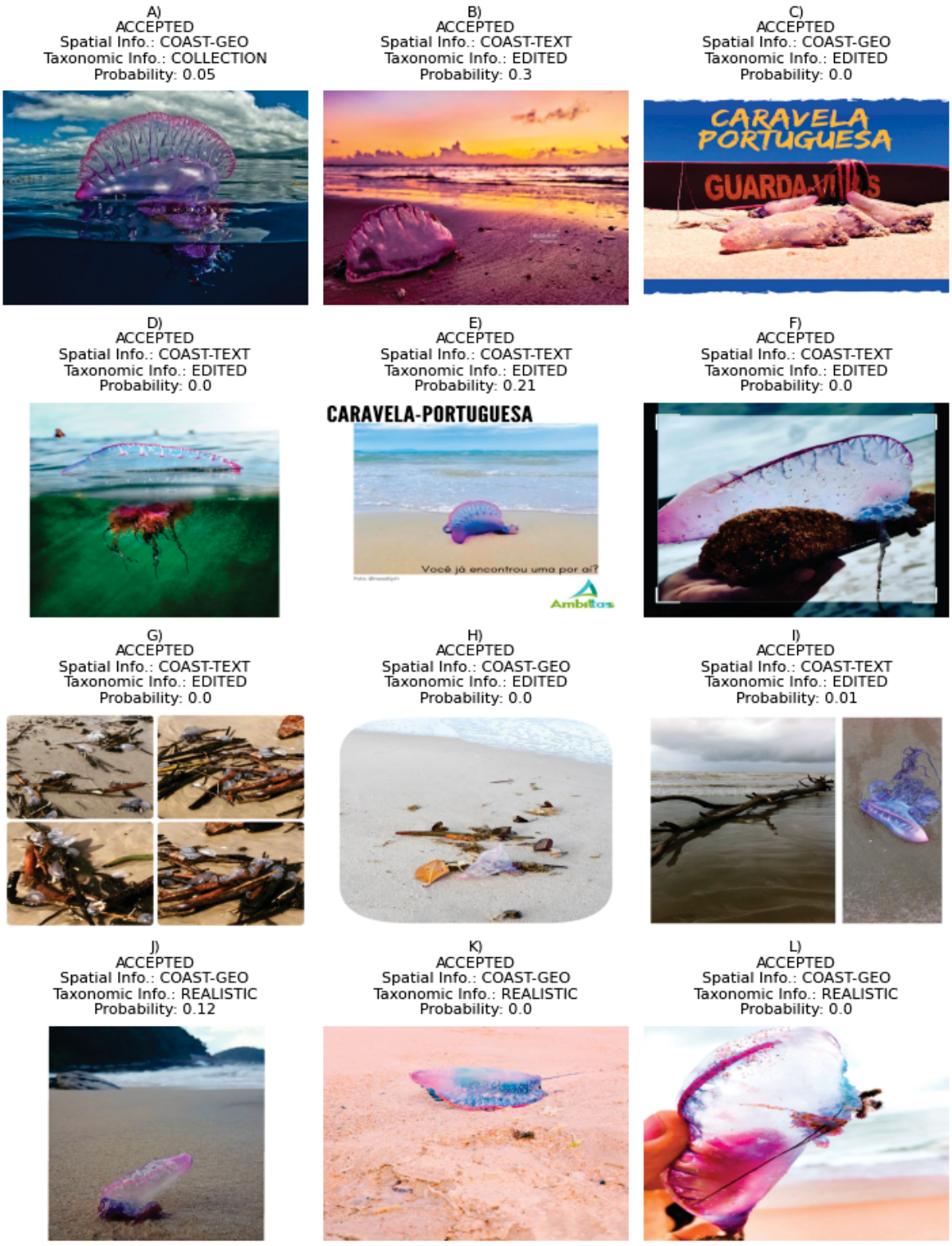


Figure A.1: Complete List of False Negative by Best Image Model. At the top of each figure is informed Accepted or Reason for Rejection if rejected, Spatial Information, Taxonomic Information and Probability.

Figure A.2: Continuation of Complete List of False Negative by Best Image Model. At the top of each figure is informed Accepted or Reason for Rejection if rejected, Spatial Information, Taxonomic Information and Probability.

Figure A.3: Continuation of Complete List of False Negative by Best Image Model. At the top of each figure is informed Accepted or Reason for Rejection if rejected, Spatial Information, Taxonomic Information and Probability.

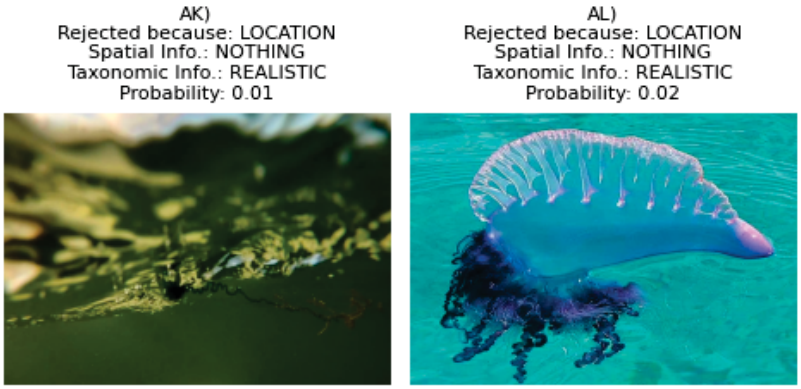


Figure A.4: End of Complete List of False Negative by Best Image Model. At the top of each figure is informed Accepted or Reason for Rejection if rejected, Spatial Information, Taxonomic Information and Probability. Note: The AC and AL posts actually have the same image.

Table A.4 shows the complete list of false positive by best combined model (ADAPTED-NOT-IN-COAST-PROD).

Table A.4: Complete List of False Positive by Best Combined Model. Showing Index (I), Image, Caption and Details: Spatial Information (Spatial Info.), Taxonomic Information (Tax. Info.) and Reason for Rejection (R. Rejection). Emojis were replaced by EMOJI.

I	Image	Caption (up to 700 characters)	Details
1		EMOJI Mar pequeno tem brincadeira sim pow ! EMOJI - #sc #santacatarina #florianopolis #floripa #ilhadamagia #peace #aloha #surf #corpo #mente #bodysurf #treino #surfdepeito #jacarezinho #swimming #natação #cuidado #caravelaportuguesa #gopro #praialimpa #summer #lixonolixo #verão #island #protetorsolar #h2o #lifestyle #waterman	Spatial Info.: COAST-TEXT Tax. Info.: NOTHING R. Rejection: MEDIA BERT Probability: 0.999 CNN Probability: 0.025
2		Tenho recebido mensagens e visto relatos das aparições das Caravelas Portuguesas (Physalia physalis) nas nossas praias. Sendo assim venho fazer um alerta sobre os riscos do contato com esses animais. O aparecimento delas nessa época do ano é comum e ocorre por conta da correnteza marítima, só que o que elas tem de exótico, também tem de perigo. Apesar de parecer inofensiva, o problema está no contato com seus tentáculos, eles liberam substâncias extremamente urticantes que podem causar queimaduras de terceiro grau e em alguns casos, pode até mesmo ser fatal. Qualquer contato com a caravela-portuguesa deve ser evitado, mesmo que o animal esteja morto...	Spatial Info.: NOTHING Tax. Info.: ART R. Rejection: MEDIA AND LOCATION BERT Probability: 0.996 CNN Probability: 0.837
3		Atenção a Caravela Portuguesa A Caravela Portuguesa (Physalia Physalia) é um invertebrado que pertence ao grupo dos Cnidários, o mesmo grupo da água viva, anêmona, entre outros animais. Às caravelas Portuguesas possuem cnidocitos com substância urticante, que são disparados ao contato. O IMPA anunciou e alertou a presença de Caravelas Portuguesas em todo litoral brasileiro, isso é comum nessa época do ano por conta das correntes marítimas, caso vejam uma Caravela Portuguesa na água ou na areia, mantenham distância e chame o salva vidas mais próximo para efetuar a retirada, não entre em contato com o animal, mesmo na areia, seus tentáculos ainda podem liberar as substâncias urticantes. Caso...	Spatial Info.: NOTHING Tax. Info.: ART R. Rejection: MEDIA AND LOCATION BERT Probability: 0.999 CNN Probability: 0.01

Continuation of Table A.4

I	Image	Caption (up to 700 characters)	Details
4		VOCÊ SABIA? OS PERIGOS DAS ÁGUAS VIVAS E DAS CARAVELAS Espécie muito semelhante às temidas águas-vivas, a Caravela (Physalia physalis) possui os mesmos tentáculos munidos de cnidoblastos - células capazes de nos machucar. Por isso, nunca devem ser tocados. Se você sofreu alguma queimadura ao interagir com estes animais, procure um bombeiro ou guarda-vidas mais próximo. Para atenuar a dor em um primeiro momento, uma dica válida é despejar água salgada e/ou gelada no local da lesão. Água doce? Nem pensar! Vale a pena lembrar também que tais espécies representam risco mesmo sem vida e fora da água, portanto, todo cuidado é pouco.... #caravelaportuguesa #aguaviva #mar #verao #praia...	Spatial Info.: NOTHING Tax. Info.: ART R. Rejection: MEDIA AND LOCATION BERT Probability: 0.999 CNN Probability: 0.023
5		#Carcavelos #caravelaPortuguesa	Spatial Info.: NOTHING Tax. Info.: CNIDARIA R. Rejection: MEDIA AND LOCATION BERT Probability: 0.019 CNN Probability: 1.0
6		Para quem não viu. Vivemos um surto de caravelas nas praias de SP e se as praias estivessem realmente cheias, seria uma catástrofe.. Mas o surgimento de caravelas dura aproximadamente 2 a 5 dias O contato com os tentáculos da caravela causa um envenenamento muito doloroso e a dor lembra a de uma queimadura. e o local fica com linhas avermelhadas que correspondem aos tentáculos do animal. Entretanto, quem tiver contato com uma caravela deve tomar certos cuidados: - NÃO coloque água doce em hipótese alguma (piora o envenenamento). - NÃO toque a mão sem luvas no tentáculo, se tiver parte do tentáculo na pele , retire com um graveto ou proteja as mãos com luvas . Em seguida faça compressa...	Spatial Info.: COUNTRYSIDE-GEO Tax. Info.: ART R. Rejection: MEDIA AND LOCATION BERT Probability: 0.993 CNN Probability: 0.999
7		"Evitar o perigo não é, a longo prazo, tão seguro quanto expor-se ao perigo. A vida é uma aventura ousada ou, então, não é nada." . . . #dangerous #adventure #seaanimals #jellyfish #caravelaportuguesa #alforreca #caparica #danger #perigo #youandme #love #passion #followyourdreams #followyourheart #dontgiveup #youcandoit #destiny #tu #arriscar #always #allofmylife	Spatial Info.: NOTHING Tax. Info.: CNIDARIA R. Rejection: MEDIA AND LOCATION BERT Probability: 0.999 CNN Probability: 0.029

Continuation of Table A.4			
I	Image	Caption (up to 700 characters)	Details
8		EMOJIEMOJI Beleza que deve ser admirada de longe. Assim como várias outras espécies de #cnidários, a caravela-portuguesa, Physalia physalis, possui ao longo dos tentáculos células do tipo cnidócitos, as quais podem causar injúrias ao entrar em contato com nossa pele. Por isso, mesmo que seja linda, admire-a à distância e não toque nela! A caravela-portuguesa faz parte da Ordem Siphonophora, e consiste em uma colônia de hidrozoários composta por indivíduos polipóides e medusóides, com até mil zoóides em uma única colônia. O maior tentáculo da colônia, o alimentar (gastrozoóide), pode chegar a 13m nesta espécie. Além do tentáculo alimentar, temos ainda os tentáculos não alimentares de...	Spatial Info.: NOTHING Tax. Info.: ART R. Rejection: MEDIA AND LOCATION BERT Probability: 0.999 CNN Probability: 0.03
End of Table			

Table A.5 shows the complete list of false negative by best combined model (ADAPTED-NOT-IN-COAST-PROD).

Table A.5: Complete List of False Negative by Best Combined Model. Showing Index (I), Image, Caption and Details: Spatial Information (Spatial Info.), Taxonomic Information (Tax. Info.) and Original Label. Emojis were replaced by EMOJI.

I	Image	Caption (up to 700 characters)	Details
1		Embora pareça uma água-viva típica, a caravela-portuguesa na verdade é uma colônia de pequenos cnidários. Nesta relação ecológica, todos são beneficiados, sendo considerada, portanto, uma relação harmônica. #cnidarios #caravelaportuguesa #biologia #praianova #praias #beach #caravela	Spatial Info.: COAST-GEO Tax. Info.: REALISTIC ACCEPTED BERT Probability: 0.999 CNN Probability: 0.003
2		Um registro especial dessa interação entre animais, pelo fotógrafo @derluhdantas Falamos um pouco sobre o Graucá (Ocypode quadrata (Fabricius, 1787)) e iremos conhecer um pouco mais sobre seus hábitos alimentares. São classificados como carnívoros, essencialmente predadores, porém são carniceiros ocasionais. Sua dieta principal depende muito do ambiente, a maioria das populações se alimenta basicamente de macroinvertebrados filtradores enterrados em águas rasas (como tatuíras e mariscos), o que explica sua anatomia, com as garras orientadas para o solo. Além disso, sua alimentação pode ser composta de insetos da vegetação costeira e de animais mortos trazidos pela maré. Como exemplo dessa f	Spatial Info.: COAST-GEO Tax. Info.: EDITED ACCEPTED BERT Probability: 0.0 CNN Probability: 0.0

Continuation of Table A.5

I	Image	Caption (up to 700 characters)	Details
3		Cnidários á deriva EMOJI EMOJI #zoo #animals #biologist #science #scientist #biologo #zoologia #cnidario #animaismarinhos #invertebrado	Spatial Info.: COAST-GEO Tax. Info.: REALISTIC ACCEPTED BERT Probability: 0.0 CNN Probability: 0.79
4		Nome científico: Physalia physalis Classificação superior: Physalia Classificação: Espécie Ordem: Siphonophora Filo: Cnidaria Classe: Hydrozoa Todas essas belezinhas estavam na praia de cotovelo hoje de manhã #caravelaportuguesa #cotovelobeach	Spatial Info.: COAST-TEXT Tax. Info.: REALISTIC ACCEPTED BERT Probability: 0.004 CNN Probability: 0.75
5		Alem das aguas da Litoranea em Sao Luis estarem com o status de “Improprias ao banho” hoje foi possivel encontrar milhares de Caravelas Portuguesas Bebes (Physalia physalis) na praia, o que normalmente seria verificado no seu maior periodo de reproducao, em outubro. #aguaviva #aguavivacaravela #aguavivacaravelaportuguesa #caravelaportuguesa #physaliaphysalis #physalis #litoranea #litoraneasaoluis #litoraneasaoluisma #jellyfish #praiasdomaranhao	Spatial Info.: COAST-TEXT Tax. Info.: EDITED ACCEPTED BERT Probability: 0.999 CNN Probability: 0.0

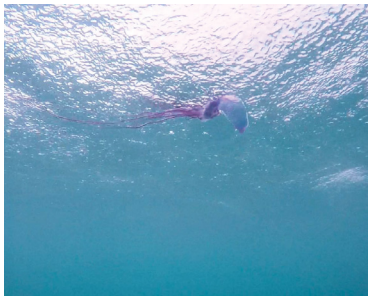
Continuation of Table A.5

I	Image	Caption (up to 700 characters)	Details
6		Andando pela areia da praia hoje na Barra do Sahy (Aracruz-ES) encontramos essa beleza da natureza, ao postar a foto nos stories recebi mensagens do nome dessa espécie rara, pesquisei um pouquinho sobre o que era no famoso Google trata-se da caravela-portuguesa (Physalia physalis) que maravilha-nos com as suas cores e tons. Mas devemos evitar chegar perto demais: este curioso animal é perigoso, a picada extremamente dolorosa e pode evoluir para algo mais grave (a que encontramos parecia estar morta) Ela pertence ao grupo dos cnidários e na verdade não é sequer “um” animal, é um organismo pluricelular, composto de quatro pólipos, ou zooides, distintos – portanto, uma colônia. Estes quatro pó	Spatial Info.: COAST-TEXT Tax. Info.: EDITED ACCEPTED BERT Probability: 0.994 CNN Probability: 0.0
7		#caravelaportuguesa #running #corridaderua #foconoobjetivo #treinofinalizado	Spatial Info.: COAST-GEO Tax. Info.: EDITED ACCEPTED BERT Probability: 0.0 CNN Probability: 0.9
8		Um achado exótico mais lindo EMOJI EMOJIEMOJIEMOJIEMOJIEMOJI #caravelaportuguesa	Spatial Info.: COAST-GEO Tax. Info.: REALISTIC ACCEPTED BERT Probability: 0.0 CNN Probability: 0.938
9		As caravelas portuguesas estão marcando presença neste novembro em Capão. Tome cuidado, a queimadura desta espécie de água-viva é muito mais danosa do que as nossas tradicionais mães d’água. EMOJI @rcardosofotos	Spatial Info.: COAST-GEO Tax. Info.: REALISTIC ACCEPTED BERT Probability: 0.989 CNN Probability: 0.006

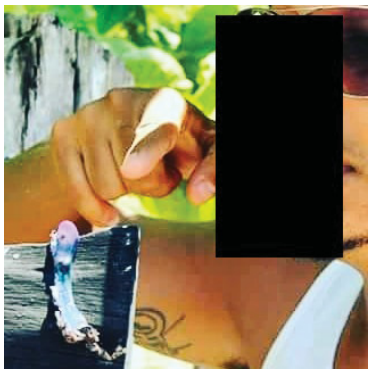
Continuation of Table A.5

I	Image	Caption (up to 700 characters)	Details
10		Caravela Portuguesa, uma das mais perigosas. Encontrada na praia Massaguaçu às 15:00 - 05/12/2020 EMOJIEMOJI EMOJI Erick Rodrigues #caravelaportuguesa #animaismarinhos #vidaaquatica #vidamarinha #animais #praiasbrasileiras	Spatial Info.: COAST-GEO Tax. Info.: REALISTIC ACCEPTED BERT Probability: 1.0 CNN Probability: 0.005
11		Tão linda,mais tão perigosa!caravela portuguesaEMOJI "Foto que fiz ontem tardeEMOJI" @outexbrasil @realoutex #caravela #caravelaportuguesa #aguaviva #águaviva #underwater #sea #animais #mergulholivre #mergulho #oceans #oceano #oceanlife #underwaterphotography #underwaterworld #photographer #biologia #bio #biologiamarinha #outex #outexbrasil #nikon #nikonzeiros #guarujabeach #guarujá #guarujapitangueiras #magic #physaliaphysalis #marinho #underwaterlife #underwateranimals	Spatial Info.: COAST-TEXT Tax. Info.: EDITED ACCEPTED BERT Probability: 1.0 CNN Probability: 0.0
12		Somos tudo aquilo que dizemos que somos... EMOJIEMOJIEMOJIEMOJI EMOJI@flarlesonpedrosa #nova coleção verão 2016/17 @hurley aqui na #FRAN6STORE . Enviamos para todo EMOJI EMOJIEMOJIEMOJIEMOJIEMOJIEMOJIEMOJIEMOJI Obrigado Senhor! #Camping. A sua segunda #house é aqui! Por tudo somos grato! Diárias à partir de R$ 19.90 EMOJIEMOJIAlohaEMOJIEMOJI EMOJI #HOSTEL FRAN6 EMOJI#PASSEIOS EMOJI@FLYBOARDMACEIO EMOJI #Aulasdesurf ReservasEMOJI EMOJI55 82 993925252 EMOJIATENÇÃOEMOJI NÃO ESQUEÇA DE PASSAR O SEU PROTETOR FACIAL EMOJI EMOJI@protetorbrazinco EMOJI EMOJI EMOJI#GOODVIBE EMOJI EMOJI MANTENHAM A PRAIA LIMPA! EMOJIPATROCINADORES #CTFRAN6EMOJI - FuncionalEMOJI @espaco_funcional - Fisioterapia e Quiropraxia EMOJI @rodrigo.costaleite - Parafinas EMOJI @magnet.wax - Comunicação VisualE	Spatial Info.: COAST-TEXT Tax. Info.: REALISTIC ACCEPTED BERT Probability: 0.001 CNN Probability: 0.976

Continuation of Table A.5

I	Image	Caption (up to 700 characters)	Details
13		Passeio pela praia... #praia #peruibecity #chuva #nublado #calor #quarta #bike #aguaviva #cnidários	Spatial Info.: COAST-TEXT Tax. Info.: EDITED ACCEPTED BERT Probability: 0.976 CNN Probability: 0.008
14		#caravela #aguaviva #caravelaportuguesa	Spatial Info.: COAST-GEO Tax. Info.: REALISTIC ACCEPTED BERT Probability: 0.999 CNN Probability: 0.005
15		Hoje ela deu o ar de sua graça, desfilando em por cima dos mergulhadores ! EMOJI#planetsubfilmes #mar #aguaviva #caravelaportuguesa #portodegalinhas #brasil #sea #ocean #scubadiving #scubaworld #scuba #dive #padi #work #myplace #fish #biodiversidade #living #brasil #nordeste #viverepreservar	Spatial Info.: COAST-TEXT Tax. Info.: REALISTIC ACCEPTED BERT Probability: 0.04 CNN Probability: 0.009

Continuation of Table A.5

I	Image	Caption (up to 700 characters)	Details
16		Oi pessoas, tudo bem? Então vou explicar pra vocês um pouco sobre esse organismo que muitos confundem com águas-vivas. A CARAVELA. Caravela, organismo colonial. Vivem em alto-mar e possui longos tentáculos de até 20 metros ou mais. As substâncias urticantes que fabrica podem causar sérias queimaduras em seres humanos. Sua fisgada pode ser muito dolorosa e tóxica, pode apresentar sintomas como calafrios, febre, náusea, vômito e choque. Em alguns casos, as fisgadas são fatais. São colônias formadas principalmente por vários pólipos transparentes que como um todo, ficam flutuando sobre a água dos oceanos. Na colônia, grupos diferentes de pólipos desempenham funções diferentes. Uns promovem a digestão dos alimentos, alguns a reprodução, outros a proteção de toda a colônia, por exemplo.	Spatial Info.: COAST-GEO Tax. Info.: REALISTIC ACCEPTED BERT Probability: 0.998 CNN Probability: 0.002
17		ATENÇÃO! EMOJI . Período de ventos fortes nas nossas praias, e com eles, aumento da incidência de caravelas na zona de banho EMOJI . Já fizemos aqui alguns textos dando orientações como proceder em caso de acidente com estas.EMOJI . Fiquem atentos principalmente com as crianças, que podem achar na areia e pegar para brincar.EMOJI . Foto: SD Gama (@tiagogamanutricionista) Edição: SD Sancho (@nicosanchoef) . #guardavidasceara #preveniresalvar #caravelaportuguesa #guardavidas	Spatial Info.: COAST-GEO Tax. Info.: EDITED ACCEPTED BERT Probability: 0.001 CNN Probability: 0.0
18		Caravela EMOJI #caravelaportuguesa #pirata #mergulho # #surfer #surfers #cerveja #waves #goprobrasil #surfer #tattoo #style #goprohero7black #aloha #mahalo #freedom #brazilian #pirate #beach #paradise #liberdade #gopro #water #photographer #naturelovers #sea #photography #boatarde	Spatial Info.: COAST-GEO Tax. Info.: REALISTIC ACCEPTED BERT Probability: 0.002 CNN Probability: 0.962
19		Agua viva, Morta EMOJI! Verão chegou com essas Pequeninas porém potentes #caravela #caravelaportuguesa #aguavivacaravela dói dói dói muito EMOJI.	Spatial Info.: NOTHING Tax. Info.: REALISTIC Rejected because LOCATION BERT Probability: 0.999 CNN Probability: 0.0

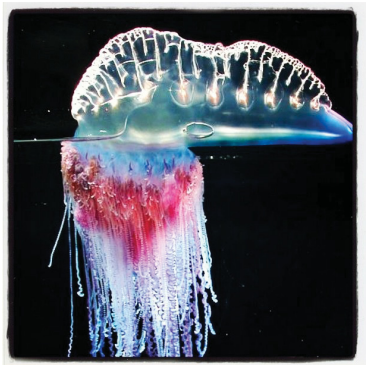
Continuation of Table A.5

I	Image	Caption (up to 700 characters)	Details
20		Pra completar a trilogia do dia no feed aqui vai a água viva ou medusa que me queimou hoje mais cedo com seus lindos tentáculos. #aguaviva #mar #praia	Spatial Info.: NOTHING Tax. Info.: EDITED Rejected because LOCATION BERT Probability: 0.001 CNN Probability: 0.123
21		Filos do reino animal: CNIDARIOS Na sequência dos filos do reino animal, o próximo a se falar são os Cnidarios. Fazem parte dos cnidarios: medusas ou águas-vivas; caravelas; anemonas; hidras e corais. Esses animais são encontrados em oceanos de quase todo o mundo e são bem conhecidos, afinal, quem nunca ficou com medo de ir na praia e ser queimado por uma água-viva? Na escala evolutiva eles são muito importantes, pois foram os primeiros a apresentarem tecidos verdadeiros, porem ainda não formam órgãos. Outra coisa muito notada no seu surgimento foi a cavidade digestiva desses animais, o que permitiu que pudessem se alimentar de uma forma mais variada e maior! Eles podem ser sésseis (não se movimentam) ou podem se locomover, podendo formar também colônias como os corais, que são colonias	Spatial Info.: NOTHING Tax. Info.: REALISTIC Rejected because LOCATION BERT Probability: 0.0 CNN Probability: 0.007
22		EMOJI Você já viu uma caravela? Muitas pessoas pensam que é uma água viva. Mas não, ela não é uma água viva, nem mesmo é um único animal. EMOJI A caravela é um cnidário, formado por QUATRO zooides. Portanto, na realidade as caravelas são uma colônia. Os zooides da caravela não vivem separados, apenas juntos. Dessa forma então, para formar a caravela temos os seguintes zooides: EMOJI Pneumatóforo – Que é a vesícula flutuadora que faz com que os outros membros da colônia boiem. Em alguns casos, como para a sua proteção, a colônia pode liberar parte do gás da vesícula, fazendo com que a vesícula murche e afunde rapidamente. EMOJI Dactilozooides – São os que formam os tentáculos. EMOJI Gastrozooides – Que são respo	Spatial Info.: NOTHING Tax. Info.: EDITED Rejected because LOCATION BERT Probability: 0.997 CNN Probability: 0.0
23		Pra quem quis brincar no parquinho, o ingresso. #surf #surfing #natural #caravelaportuguesa #good-vibes #sealife #gopro #bahia #gloriaaDeus #gratidao	Spatial Info.: NOTHING Tax. Info.: REALISTIC Rejected because LOCATION BERT Probability: 0.0 CNN Probability: 0.007

Continuation of Table A.5

I	Image	Caption (up to 700 characters)	Details
24		Não toque nestes animais. Fique por perto garantindo que ninguém se machuque e peça para outra pessoa chamar o guarda-vidas para retirá-la da praia. Verifique o mar antes de entrar se não há outras. Lembre que seus tentáculos causam queimaduras e podem medir até 40 metros de comprimento! #atenção #Caravela-Portuguesa #SaveThePlanet #marineanimals	Spatial Info.: NOTHING Tax. Info.: EDITED Rejected because LOCATION BERT Probability: 0.004 CNN Probability: 0.01
25		Filo: Cnidaria Classe: Hydrozoa Ordem: Siphonophora Família: Physaliidae Género: Physalia Espécie: P.physalis As caravelas portuguesas são organismos compostos por 4 pólipos da família Physaliidae, ou seja, são uma colônia. Cada um dos pólipos desempenham um papel no organismo, tais pólipos são: gonozooides,gastrozooides,domonoctozooides e pneumatoforo, e desempenham respectivamente as funções: reprodução, estômagos, tentáculos e vesícula cheia de ar. Elas exibem cores muito chamativas, e por incrível que pareça os seus tentáculos podem chegar a até 20 metros, eles possuem cnidócitos venenosos, o seu veneno chega a ser comparado com o da viúva negra, por esse motivo é aconselhável que não se aproximem tanto delas. A função desses tentáculos é capturar presas que possivelmente entre	Spatial Info.: NOTHING Tax. Info.: EDITED Rejected because LOCATION BERT Probability: 0.996 CNN Probability: 0.0
26		:.Caravela-portuguesa:. #aguavivas #caravelaportuguesa #mar #fotografiaamadora #photography #natureza	Spatial Info.: NOTHING Tax. Info.: REALISTIC Rejected because LOCATION BERT Probability: 1.0 CNN Probability: 0.0

Continuation of Table A.5

I	Image	Caption (up to 700 characters)	Details
27		Caravela Portuguesa EMOJI #caravelaportuguesa	Spatial Info.: NOTHING Tax. Info.: REALISTIC Rejected because LOCATION BERT Probability: 0.053 CNN Probability: 0.927
28		#caravelaportuguesa #garrafaazul	Spatial Info.: NOTHING Tax. Info.: REALISTIC Rejected because LOCATION BERT Probability: 0.001 CNN Probability: 0.565
29		#cnidários #caravela #águaviva #sea #seresmarinhos #mundoestranho #oceano #amazing #surpreendente #gorgeous #mimo	Spatial Info.: NOTHING Tax. Info.: EDITED Rejected because LOCATION BERT Probability: 0.999 CNN Probability: 0.002

Continuation of Table A.5

I	Image	Caption (up to 700 characters)	Details
30		EMOJIEMOJIEMOJI #caravelaportuguesa	Spatial Info.: NOTHING Tax. Info.: REALISTIC Rejected because LOCATION BERT Probability: 0.024 CNN Probability: 0.993
31		#CaravelaPortuguesa O nome caravela-portuguesa deve-se à semelhança dos cnidários e as caravelas utilizadas como navios de guerra. É uma colónia de organismos geneticamente idênticos e altamente especializados que aparentam ser uma única criatura e a sua principal toxina é a Physaliatoxina (glicoproteína de 240 kDa) com citotoxicidade e toxicidade hemolítica.	Spatial Info.: NOTHING Tax. Info.: REALISTIC Rejected because LOCATION BERT Probability: 0.999 CNN Probability: 0.0
32		EMOJI CUIDADO EMOJI Calor, praia e água-viva: em um único dia, litoral tem 36 pessoas com lesões causadas por contato com animal Número corresponde a 22% do número de casos de outubro até 19 de dezembro 22/12/2021 - 12h24min #litoral #aguaviva #mar #calor #verao #veraneio #noticia	Spatial Info.: NOTHING Tax. Info.: EDITED Rejected because LOCATION BERT Probability: 0.0 CNN Probability: 0.978

End of Table