



Data Descriptor

# #PraCegoVer: A Large Dataset for Image Captioning in Portuguese

Gabriel Oliveira dos Santos , Esther Luna Colombini and Sandra Avila \*

Institute of Computing, University of Campinas (Unicamp), Campinas 13083-852, Brazil;  
g197460@dac.unicamp.br (G.O.d.S.); esther@ic.unicamp.br (E.L.C.)

\* Correspondence: sandra@ic.unicamp.br

**Abstract:** Automatically describing images using natural sentences is essential to visually impaired people's inclusion on the Internet. This problem is known as *Image Captioning*. There are many datasets in the literature, but most contain only English captions, whereas datasets with captions described in other languages are scarce. We introduce the #PraCegoVer, a multi-modal dataset with Portuguese captions based on posts from Instagram. It is the first large dataset for image captioning in Portuguese. In contrast to popular datasets, #PraCegoVer has only one reference per image, and both mean and variance of reference sentence length are significantly high, which makes our dataset challenging due to its linguistic aspect. We carry a detailed analysis to find the main classes and topics in our data. We compare #PraCegoVer to MS COCO dataset in terms of sentence length and word frequency. We hope that #PraCegoVer dataset encourages more works addressing the automatic generation of descriptions in Portuguese.

**Dataset:** <https://doi.org/10.5281/zenodo.5710562>



**Citation:** dos Santos, G.O.; Colombini, E.L.; Avila, S. #PraCegoVer: A Large Dataset for Image Captioning in Portuguese. *Data* **2022**, *7*, 13. <https://doi.org/10.3390/data7020013>

Academic Editor: Joaquín Torres-Sospedra

Received: 1 December 2021

Accepted: 15 January 2022

Published: 21 January 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Dataset License:** CC BY-NC-SA 4.0

**Keywords:** #PraCegoVer; image captioning in Portuguese; image captioning; image-to-text

## 1. Summary

The Internet is becoming increasingly accessible, reaching a wide variety of audiences. However, little progress has been made in including people with disabilities. The scenario is even worse for visually impaired people since a significant part of the Internet content is exclusively visual, for instance photos and advertising images. Screen readers can transform textual information into audio but, regarding visual content, they are still mostly dependant on annotations added to the source code of websites, which in turn are not that descriptive.

In light of this situation, in 2012, PraCegoVer [1] arose as a social movement, idealized by Patrícia Braille, that stands for the inclusion of people with visual impairments besides it has an educational propose. The initiative aims to call attention to the accessibility question. It stimulates users to post images tagged with #PraCegoVer and add a short description of their content. This project has inspired many local laws that establish that all posts made by public agencies on social media must refer to #PraCegoVer and contain a short description of the image.

Automatically describing image content using natural sentences is essential for including people with visual impairments on the Internet, making it more inclusive and democratic. However, it is still a big challenge that requires understanding the semantic relation of the objects present in the image, their attributes, and the actions they are involved in to generate descriptions in natural language. Thus, linguistic models are also needed to verbalize the semantic relations in addition to visual interpretation methods. This task is known as *image captioning*.

The image captioning task has been accelerated thanks to the availability of a large amount of annotated data in relevant datasets, for instance, Flickr8k [2], Flickr30k [3], and MS COCO Captions [4]. The images in these datasets are associated with multiple reference descriptions, whose average length is about 10 words and has low variance. Many large-scale datasets have been created [5,6], but in contrast to the previous ones, they employ automated pipelines. Although many datasets were proposed in this literature, most of them contain only English captions. In contrast, datasets with captions described in other languages are scarce.

Simply translating datasets from English to other languages is a cheap way to train models to generate non-English captions. Nevertheless, the literature on Natural Language Processing has already shown that it introduces noise in the data that can harm the performance of models. In particular, the works of Xue et al. [7] and Rosa et al. [8] have shown that the model performance is considered hampered when translated datasets are used in comparison with use datasets originally annotated in the target language.

Hence, inspired by the PraCegoVer project, we introduce the #PraCegoVer dataset, which is a multi-modal dataset with images and Portuguese descriptions. As far as we know, this is the first dataset proposed for the Image Captioning problem with captions<sup>1</sup> in Portuguese. This dataset relies on 533,523 posts collected from 14,000 different profiles on Instagram. Moreover, each instance comprehends an image associated with one reference description. Also, the average reference length is 39.3 words, and the standard deviation is 29.7. These values are considerably high in comparison to other datasets in this literature. These characteristics make #PraCegoVer more challenging than the popular ones because current state-of-art models have difficulty in learning in the context of single reference and long sentences.

Our key contributions are three-fold:

1. We introduced the first dataset for the problem of image captioning with captions in Portuguese. We hope that #PraCegoVer dataset encourages more works addressing the automatic generation of descriptions in Portuguese. We also intend to contribute to the blind Portuguese speaker community.
2. We developed an end-to-end framework for data collection, data preprocessing, and data analysis from a hashtag on Instagram, which is helpful for social media studies (Section 4). In addition, we carried a thorough exploratory analysis to identify the most significant image classes and topics within the captions.
3. We proposed an algorithm to cluster post duplication based on visual and textual information to remove instances with similar content.

Full-reproducible source code is available in our Github repository<sup>2</sup>.

## 2. Related Work

The image captioning task has been accelerated thanks to the availability of a large amount of annotated data in relevant datasets, for instance, Flickr8k [2], Flickr30k [3], and MS COCO Captions [4].

Microsoft Common Objects in COntext (MS COCO) Captions is a dataset created from the images contained in MS COCO [9] and human-generated captions. MS COCO Captions dataset comprises more than 160k images collected from Flickr, distributed over 80 object categories, with five captions per image. Its captions are annotated by human annotators using the crowdsourcing platform Amazon Mechanical Turk (AMT). The annotators were told, among other instructions, not to give people proper names and write sentences with at least eight words. As a result, the descriptions' average sentence length is about 10 words with no proper names.

Many large-scale datasets have been created [6,10–20], but in contrast to the previous ones, they employ automated pipelines. One example of a dataset that follows this approach is the Conceptual Captions dataset [5] which has more than 3.3M pairs of images and English captions. It was created by crawling web pages and extracting images and the

alt-text HTML attribute associated with them. Images and captions are automatically filtered as well as cleaned aiming to select informative and learnable data.

To explore real-world images, Agrawal et al. proposed nocaps [21], a benchmark that consists of validation and test set with 4500 and 10,600 images, respectively, annotated with 11 human-generated captions per image. This dataset is created by filtering images from the Open Images V4 dataset [22] and selecting images based on their object categories. Moreover, nocaps has more objects per image than MS COCO, and it has 600 object categories, whereas MS COCO has only 80. This benchmark evaluates models on how well they generate captions for objects not present in the dataset on which they are trained.

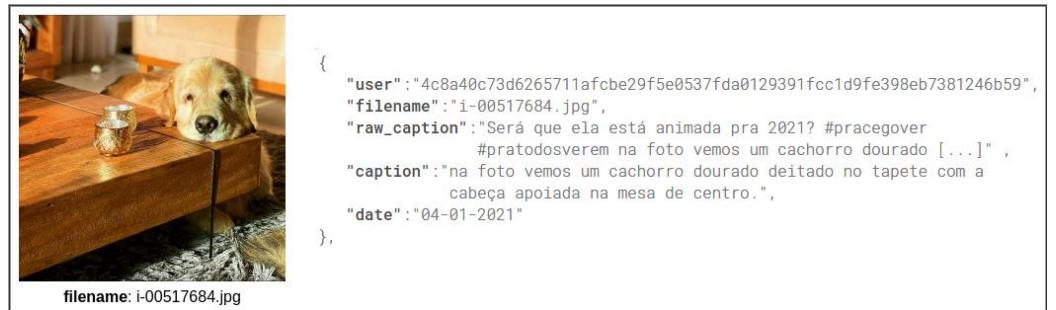
Recently, Gurari et al. proposed VizWiz-Captions dataset [23] focused on the real use case of the models by blind people. It represents a paradigm shift of image captioning towards goal-oriented captions, where captions faithfully describe a scene from everyday life and answer specific needs that blind people might have while executing particular tasks. This dataset consists of 39,181 images taken by people who are blind, each image paired with five captions annotated by using the AMT platform. They also have metadata that indicates whether a text is present on the image and the image quality issues. The overlap between VizWiz-Captions and MS COCO content is about 54%, which means a significant domain shift in the content of pictures taken by blind photographers and what artificially constructed datasets represent.

InstaPIC-1.1M [24] was created by collecting posts from Instagram, comprising 721,176 pairs of image-caption from 4.8k users. Based on the 270 selected hashtags, they crawled the Instagram APIs to filter the posts and collect the images and captions. The major problem in the InstaPIC-1.1M dataset is that the captions may not reflect the image content because they are based on what Instagram users write about their posts, which can be quite vague and do not describe the visual content. For example, “Autumn is the best” and “We take our adventure super seriously. #selfiesunday” are vague captions present in its training set.

Our approach also is based on Instagram posts, but in contrast to InstaPIC-1.1M and Conceptual Captions, we collect only captions where #PraCegoVer is tagged. Moreover, we clean the captions maintaining just the audio description part written by supporters of the PraCegoVer movement. Thus, similarly to VizWiz-Captions, our dataset’s descriptions are addressed to visually impaired people and reflect the visual content. Still, our dataset contains captions with 40 words on average, while those in MS COCO Captions have only ten words, and the variance of sentence length in our dataset is also more significant. On the other hand, since our dataset is freely annotated, the descriptions might contain proper names that can be removed, and we consider this an essential point for future improvements. Finally, #PraCegoVer is the only dataset, as far as we know, that comprises Portuguese descriptions, in contrast to the others that comprehend English captions. Also, the average sentence length the variance of length in terms of the number of words make #PraCegoVer a challenging dataset.

### 3. Data Records

#PraCegoVer dataset is hosted on Zenodo (<http://doi.org/10.5281/zenodo.5710562>, accessed on 14 January 2022) and can be accessed upon request. The dataset is composed of the main file `dataset.json` and a collection of compressed files named `images.tar.gz` containing the images. The file `dataset.json` comprehends a list of json objects with the attributes: `user`, `filename`, `raw_caption`, `caption` and `date`. These attributes represent, respectively, the anonymized user that made the post, the image file name, the raw caption, the clean caption, and the post date. Each instance in `dataset.json` is associated with exactly one image in the `images` directory whose filename is pointed by the attribute `filename`, as shown in Figure 1. Also, we provide a sample with five instances from #PraCegoVer, so the users can download the sample to get an overview of the dataset.



**Figure 1.** Example of instance from #PraCegoVer.

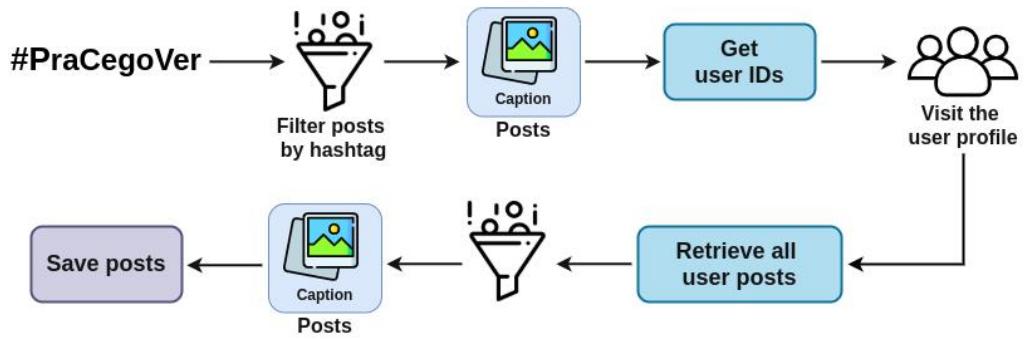
## 4. Method

### 4.1. Data Collection

Several companies and government agencies have joined the campaign #PraCegoVer, thereby posting images, including their audio description, on social networks such as Facebook, Instagram, and Twitter. Although these data are available on the internet, they are not easy to collect because they generally limit their public APIs (Application Programming Interface). Moreover, the restrictions may vary among the platforms. For instance, Facebook provides an API that only allows access to posts from one user, making it challenging to collect data on a large scale. On the other hand, Instagram permits access to public posts from many users, limiting the search to posts published in the last seven days. We tried to collect data from Twitter, but the tweets that use the hashtag #PraCegoVer are only a copy of the same content posted on Instagram. Also, because of the 280-character limit, the users split the text into many tweets. Then, leveraging these data requires more preprocessing for Twitter. Thus, we believe that the effort to collect and preprocess data from Twitter was worthless.

This work has collected data only from Instagram since it focuses on image sharing and filters posts by hashtag. As mentioned before, Instagram limits the filter by hashtag to posts published in the last seven days. Therefore, to overcome this obstacle, we first search for posts related to the hashtag #PraCegoVer and save just the profiles (user id). In the following step, we visit these profiles looking for more posts marking the hashtag we are interested in. Inside the profile pages, there are no limitations regarding date or volume of posts. Thus, we have access to all images published by that user because they are public, and finally, we can collect the posts.

We execute this process daily and incrementally, storing: images, their width and height, their captions, post identifiers, post owners, post date, and the collection date. In this way, we can collect posts published any time ago, instead of up to the past seven days as restricted in the hashtag page. We highlight that we ensure that the crawler never accesses posts from private profiles, which would require an acceptance from each account's owner. However, there may be profiles that became private after we had collected their posts. Figure 2 illustrates our pipeline of data collection.



**Figure 2.** Diagram illustrating the pipeline of data collection. We start filtering the posts by hashtags and save the profile identifiers. From these identifiers, we retrieve the usernames, which are used to visit each profile page. We visit all the profile pages and retrieve the posts. Icons made by Freepik and prosymbols from “[www.flaticon.com](http://www.flaticon.com)” (accessed on 14 January 2022).

#### 4.2. Duplication Detection and Clustering

We collected our data from Instagram, where people usually share similar posts multiple times. The content shared suffers small changes such as adding a new logotype, cropping and rotating the image, changing colour scale, and so forth. We consider these images as duplications because they do not bring new information to the dataset. Thus, removing such items from the dataset is essential because models may get overfitted in those duplicated examples. Also, there are posts in which the images are duplicated but not the caption. We can use these different captions to create a structure similar to MS COCO Captions, where one image is associated with multiple captions. This section describes our proposed algorithm to identify and cluster post duplications. This algorithm uses image features to cluster similar posts and leverages the textual information to eliminate eventual ambiguity.

##### 4.2.1. Duplications

The concept of duplication depends on the application and the data. Then, it is important to define what we consider duplicated posts. Formally, let a post be a tuple of image and caption,  $post = (image, caption)$ , and  $dist(\cdot, \cdot)$  be a distance function, then we define:

**Definition 1** (Image Duplication). *Given two images  $image_1$  and  $image_2$ , they are duplicated if  $dist(image_1, image_2) \leq t_{img}$ , for some predefined threshold  $t_{img}$ . We denote this duplication by  $image_1 \sim image_2$ .*

**Definition 2** (Caption Duplication). *Given two captions  $caption_1$  and  $caption_2$ , they are duplicated if  $dist(caption_1, caption_2) \leq t_{cpt}$ , for some predefined threshold  $t_{cpt}$ . We denote this duplication by  $caption_1 \sim caption_2$ .*

**Definition 3** (Post Duplication). *Given two posts  $post_1 = (image_1, caption_1)$  and  $post_2 = (image_2, caption_2)$ , they are considered as a duplication if, and only if  $image_1 \sim image_2$  and  $caption_1 \sim caption_2$ . We denote this duplication by  $post_1 \sim post_2$ .*

From this definition, we have the transitivity property that is the basis for our algorithm to cluster duplications, Algorithm 1.

**Property 1** (Transitivity). *Given the posts  $post_1$ ,  $post_2$  and  $post_3$ , then  $post_1 \sim post_2$  and  $post_2 \sim post_3 \Rightarrow post_1 \sim post_3$ .*

Figure 3 illustrates an example of two posts considered duplicated. They have similar images and texts.



**(a)** User 1. Caption: “Na imagem, o Zaad Mondo aparece sob uma superfície tomada completamente por juníperos. São esses frutos tão pequenos que compõem o seu acorde tão marcante”.

**(b)** User 2. Caption: “Na imagem, o Zaad Mondo aparece sob uma superfície tomada completamente por juníperos. São esses frutos tão pequenos que compõem o seu acorde tão marcante”.

**Figure 3.** Two similar images posted on Instagram by two different profiles: User 1 and User 2. It can be seen that image (b) is similar to image (a), however, it contains a logo in the lower right corner that is not present in image (a). Moreover, both posts have the same caption, thus we consider them a duplication.

---

#### Algorithm 1: Clustering duplications

---

```

Require: number of posts  $n$ , distance matrices  $D_{cpt}$ ,  $D_{img}$ , and thresholds  $t_{cpt}$ ,  $t_{img}$ 
1:  $graph_{sim}[i][j] \leftarrow 0 \forall i, j \in \{1, \dots, n\}$ 
2: for all  $i \in [1 \dots n]$  do
3:   for all  $j \in [i + 1 \dots n]$  do
4:     if  $D_{img}[i][j] \leq t_{img}$  and  $D_{cpt}[i][j] \leq t_{cpt}$  then
5:        $graph_{sim}[i][j] \leftarrow 1$ 
6:        $graph_{sim}[j][i] \leftarrow 1$ 
7:     end if
8:   end for
9: end for
10:  $visited \leftarrow \emptyset$ 
11:  $clusters \leftarrow \emptyset$ 
12: for all  $v \in [1 \dots n]$  do
13:   if  $v \notin visited$  then
14:      $dups \leftarrow \emptyset$ 
15:      $DFS(graph_{sim}, n, v, visited, dups)$ 
16:      $clusters.append(dups)$ 
17:   end if
18: end for
19: return  $cluster$  {A list with sets of duplications clustered.}

```

---

#### 4.2.2. Duplication Clustering

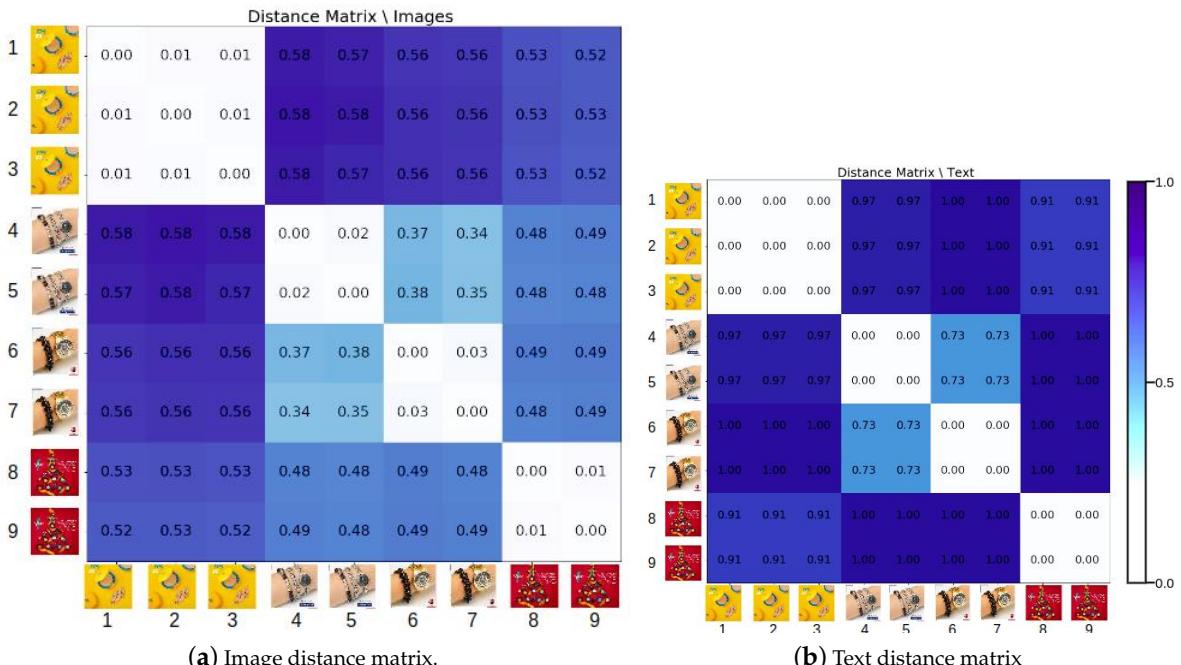
Clustering the duplicated posts is an essential step of the dataset creation process because these duplications may lead to unwanted bias. Also, we must avoid similar instances in the train and test sets to guarantee that the results obtained by algorithms trained on this dataset are representative.

We have designed the clustering Algorithm 1 based on Definition 3 and Transitivity Property 1. On this subject, we create a similarity graph that keeps the duplicated posts in the same connected component. Formally, the similarity graph is an undirected graph

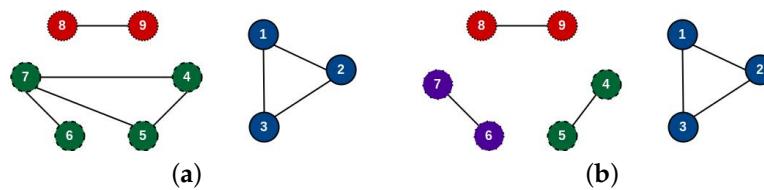
$G(V, E)$  such that for all  $post_i$  we create a vertex  $v_i$ , and for each pair of vertices  $v_i, v_j \in V$ , there exist the edge  $(v_i, v_j) \in E$  if, and only if  $post_i \sim post_j$ . Note that from the Transitivity Property if there exists a path  $(v_i, e_i, v_{i+1}, e_{i+1}, \dots, v_j)$ , then  $post_i \sim post_{i+1} \sim \dots \sim post_j$ . Thus, it is guaranteed that all duplicated posts will be kept in the same connected component of similarity graph. Algorithm 1 is the pseudocode of our clustering method.

Algorithm 1 requires the distance matrices  $D_{cpt}$  and  $D_{img}$  of captions and images, respectively. To compute the distance between two captions, we calculate the distance between the TF-IDF vectors extracted from them (Section 4.3.1). Similarly, we compute the distance between two images using their feature vectors (Section 4.3.2). Also, it requires caption and image thresholds, denoted by  $t_{cpt}$  and  $t_{img}$ . The distance matrices can be constructed using any distance metric. We found empirically that the cosine distance is a good metric for both images and captions. Moreover, we conducted a grid search varying the thresholds  $t_{img}$  and  $t_{cpt}$  in the set  $\{0.02, 0.05, 0.10, 0.20\}$ . We concluded that low thresholds result in many small clusters of duplicated posts, while high thresholds result in few large clusters with some miss clustered posts. In our experiments, we obtained the best clusters when we set the thresholds to  $t_{img} = t_{cpt} = 0.10$ .

To illustrate this algorithm, let us consider nine posts, and suppose the distance matrices with respect to images and texts are shown in Figure 4a and Figure 4b, respectively. Also, consider the thresholds  $t_{img} = 0.35$  and  $t_{cpt} = 0.10$ . It can be seen that if only consider the image distance, then we will have the graph in Figure 5a, where each vertex represents a post and the clusters  $\{1,2,3\}$ ,  $\{4,5,6,7\}$ , and  $\{8,9\}$  represent connected components, because the distances among posts in the same cluster are lower than or equal to  $t_{img} = 0.35$ . However, when we also leverage the textual information, the cluster  $\{4,5,6,7\}$  is split into  $\{4,5\}$  and  $\{6,7\}$ , as illustrated in Figure 5b, because the text distance among the captions of posts in these clusters is lower than or equal to  $t_{cpt} = 0.10$ . Therefore, textual information can improve the identification of duplications by enhancing the context.



**Figure 4.** Distance matrix constructed from the pair-wise cosine distance based on (a) image features, and (b) text features.



**Figure 5.** (a) Similarity graph based only on image distances. (b) Similarity graph based only on both image and text distances. It can be seen that when just image features are taken into account, the algorithm return the clusters {1,2,3}, {4,5,6,7}, and {8,9}, because the distances are lower than or equal to  $t_{img} = 0.35$ . However, when we also consider the textual information, it returns the clusters {1,2,3}, {4,5}, {6,7}, and {8,9}.

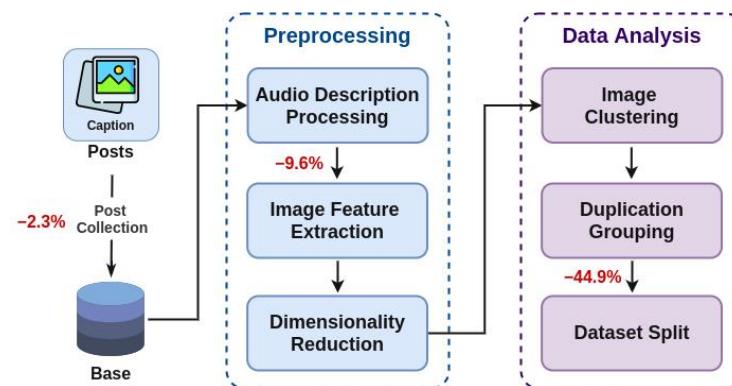
#### 4.3. Preprocessing and Data Analysis

In this section, we describe the text and image preprocessing necessary to execute Algorithm 1 that clusters post duplications and to split the final dataset into training, validation, and test sets.

Figure 6 illustrates our pipeline and highlights the percentage of posts lost in each step concerning the total amount of collected posts. We split our pipeline into two phases: preprocessing and data analysis. The *preprocessing* phase consists of the general processing of texts and images. On the other hand, the steps in the *data analysis* phase are not only used as preprocessing but are also used to explore the data by visualizing the clusters and duplications.

The first step of our pipeline comprehends collecting the data from Instagram. Then, we clean the captions to obtain the descriptions, and we extract images and text features. Next, we reduce the dimensionality of image feature vectors to optimize processing, clustering the images to analyze the data and remove duplicates. Finally, we split the dataset into training, validation, and test sets.

Moreover, we highlight in Figure 6 the loss in each step of the pipeline. About 2.3% of posts are lost during the post-collection process because of profiles that become private during this process. Besides, 9.6% of posts have malformed captions. They do not follow the main pattern, consisting of the “#PraCegoVer” followed by the description. Thus, it is tough to extract the actual caption from the whole text. Therefore we remove them. Finally, 44.9% of the posts have duplicated either caption or image, which may easily overfit the models. We also remove these cases from the dataset. In total, about 56.8% of data is lost or removed. It is worth noting that data loss is inherent to the data source and the process.

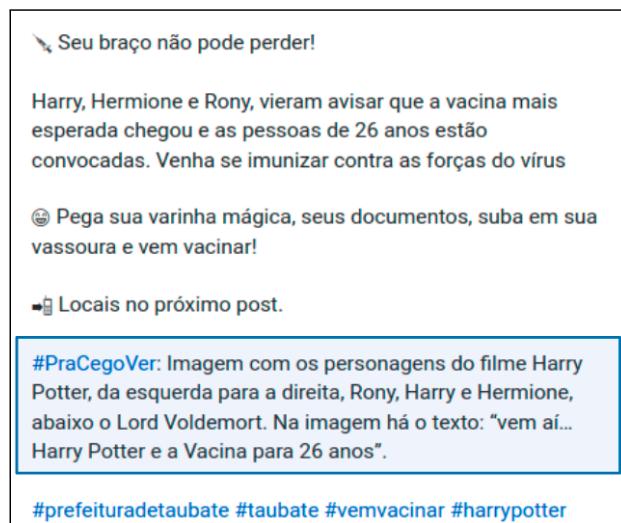


**Figure 6.** Overview of the whole pipeline from the data collection to the dataset split. First, we collect the data, clean the captions to obtain the audio descriptions, and extract the image and text features. Finally, we analyze the data to remove duplications and split the dataset into training, validation, and test sets. We highlight the percentage of posts lost in each step. Icons made by Freepik and Smashicons from “[www.flaticon.com](http://www.flaticon.com)” (accessed on 14 January 2022).

#### 4.3.1. Audio Description Processing

Although those that joined the #PraCegoVer initiative produce a considerable amount of labeled data “for free”, there are some inherent problems in the free labeling process. For instance, misspellings and the addition of emoticons, hashtags, and URL link marks in the captions. Furthermore, the captions often have some texts besides the audio description itself. Therefore, it is needed to preprocess the texts to extract the actual audio description part. We identified patterns in those texts by thoroughly reading many of them, and we used regular expressions to find the audio description within the texts. For instance, the audio description, in general, comes right after the hashtag #PraCegoVer, so we first crop the caption keeping just the text after this hashtag. Then, we use regular expressions to remove emoticons, hashtags, URL links, and profile marks. However, it might lead to wrong punctuations at the end of the texts that we also remove. For example, some posts have captions with a mark of “end of audio descriptions”. Thus, we also use it as an end delimiter. Finally, we convert the texts into lower case, remove stopwords, and transform them into TF-IDF (Term Frequency-Inverse Document Frequency) vectors. These feature vectors will then be used to remove post duplications and in topic analysis during the data visualization phase (see Section 5.2).

Figure 7 shows a real example of the caption. We highlight the part of the text that comprehends an audio description.



**Figure 7.** Example of a real caption in which is tagged the hashtag #PraCegoVer. After the extraction of audio description, we have only the highlighted text. Observe that there are emoticons and hashtags in the original caption, but we remove them.

#### 4.3.2. Image Processing

To better explore our dataset and to remove duplicated posts, we used algorithms that rely on features extracted from images. To do so, we extracted image features using MobileNetV2 [25], a convolutional neural network (CNN), and a popular choice for feature extractors. MobileNetV2 returns image feature vectors with dimension 1280. Thus, we decided to reduce dimensionality to construct a better feature space and optimize memory usage.

We used Principal Component Analysis (PCA) [26] to estimate the number of dimensions needed to keep around 95% of the explained variance, and we found 900 dimensions. Then, we reduce the dimensionality from 1280 to 900 dimensions by using UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) [27], a non-linear dimension reduction method that preserves structures from the high-dimensional space into the lower dimensional embedding. The features are projected to a lower-dimensional space that may shrink the noise. Then it can be a better feature space. We executed the algo-

rithm considering 900 dimensions, a neighborhood of 80, the minimum distance between embedded points equals 0, and correlation as the metric.

#### 4.3.3. Image Clustering

Our algorithm for removing duplicates (Algorithm 1) requires distance matrices for images and texts. The construction of such matrices has a quadratic complexity in time and space, which means that both the time and memory size required to execute this algorithm is proportional to the total number of posts squared. Then, it is not feasible to run the algorithm considering the entire dataset because it would consume a considerable amount of memory and take too long.

We hypothesized that duplicated posts were within the same cluster because they are very similar. Then, we carried a qualitative analysis, and we confirmed our hypothesis. A solution for the memory consumption problem relies on this fact. We clustered the posts taking into account only their image. Then, instead of running the algorithm for removing duplicates considering the entire dataset, we execute it only for the posts that compose a cluster because duplicates might be within the same cluster. This way, we no longer need a memory size proportional to the total number of posts squared. However, we need a memory size proportional to the cluster size squared. Since we created small clusters, as shown in Section 5.1, with at most 60 thousand posts, the memory consumption is reduced in about 98.7%.

To cluster the images, we used HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [28] to cluster the reduced image feature vectors. The hierarchical characteristic of this algorithm helps to find classes and subclasses. However, because of the nature of our data, which is sparse, the algorithm generates many outliers. To overcome this problem, we execute the algorithm iteratively: first, we cluster the data using HDBSCAN, and then we reapply this clustering algorithm to those data points assigned as outliers and repeat this process until it has reached a satisfactory amount of outliers. After we have the clustered images, we compute the distance matrices for captions and images only for those posts whose images belong to the same cluster. Thus, we significantly reduce memory usage.

#### 4.3.4. Dataset Split

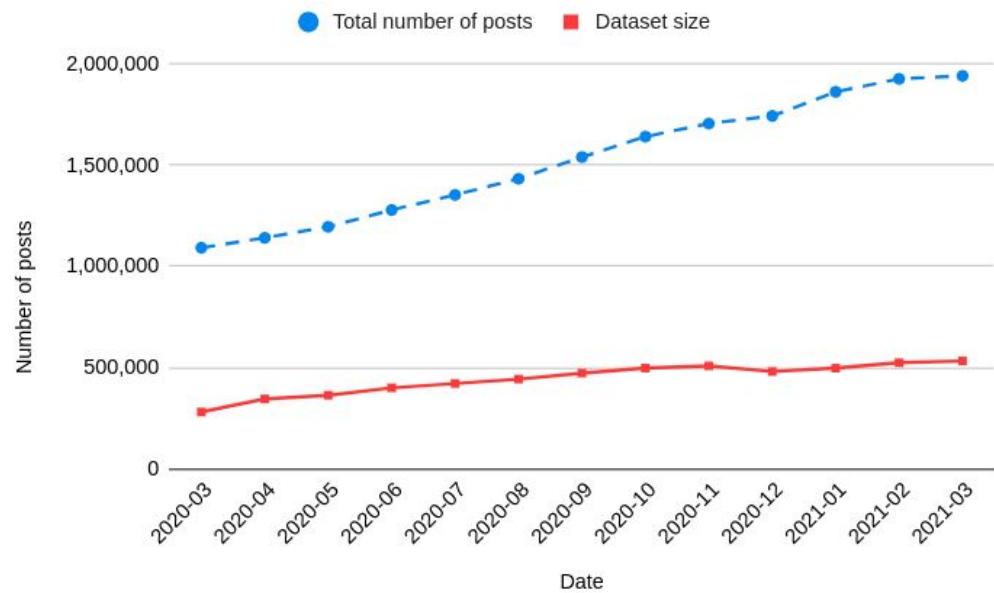
We thoroughly split #PraCegoVer dataset, guaranteeing the quality of sets, and above all, the quality of the test set. We aim to avoid eventual bias related to profiles and duplications. Thus, we start by explicitly removing the detected duplications. Then, we cluster the posts by owners and add all posts from a profile into the same set (either training, validation, or test), such that two different sets do not contain posts from the same profile. This way, we can test and validate the models in cross-profile sets. Finally, we split our dataset considering the proportion 60% for training, 20% for validation, and 20% testing.

### 5. Technical Validation

#### 5.1. Dataset Statistics

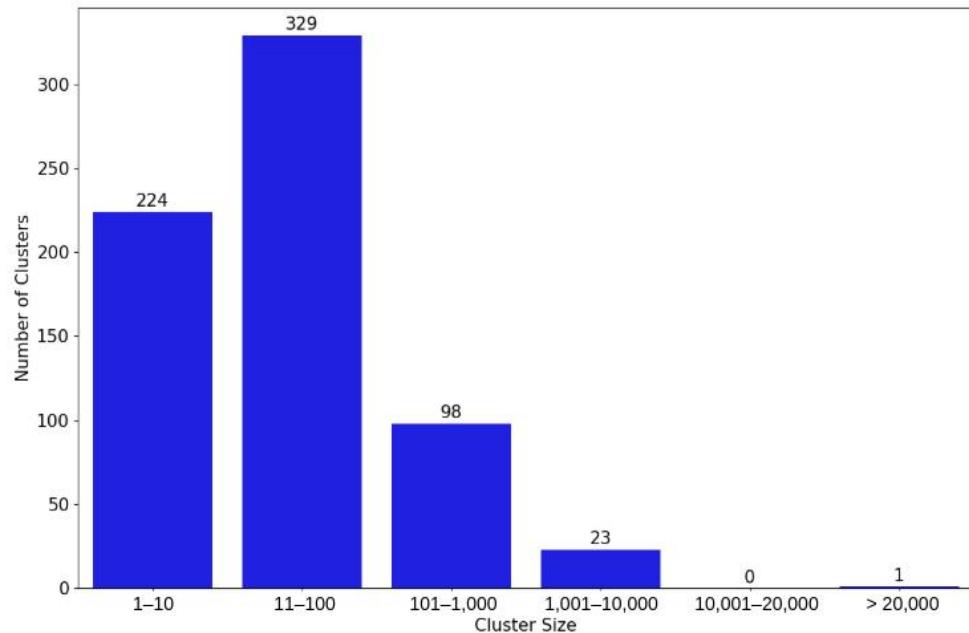
We have collected more than 533,523 posts from 14,666 different profiles on Instagram. Nevertheless, after we clean the data, there will be only 43.2% of the total remaining for training models. The other 56.8% are removed throughout the pipeline as illustrated in Figure 6. Furthermore, our dataset contains a wide variety of images, which is essential to train models that generalize to real-world images on the Internet. Also, we highlight that the #PraCegoVer dataset is growing over time because we are continuously collecting more data. Figure 8 shows the growth of the total number of posts tagging #PraCegoVer (dashed line) and the rise of #PraCegoVer dataset size through time. We can see that the total amount of posts tagging #PraCegoVer is greater than we can collect. It is due to Instagram mechanisms that limit our access to the posts. On average, our dataset has 30% of the total amount of posts. Despite this fact, the #PraCegoVer dataset still has enormous

growth potential because the number of organizations joining the #PraCegoVer movement is increasing.



**Figure 8.** The total number of posts tagging #PraCegoVer (dashed line) and #PraCegoVer dataset size (solid line) throughout the time.

We leverage the cluster structure used to remove duplicate posts to explore overview images and topics in our dataset regarding the data analysis. Figure 9 shows a histogram with the number of clusters of images by size. In all, we have 675 clusters, where one cluster contains outliers. Then, to ease the visualization, we grouped the clusters by size, after removing duplications, considering the size ranges: [1–10], [11–100], [101–1,000], [1001–10,000], [10,001–20,000], and 20,001 or more.



**Figure 9.** Histogram showing the number of clusters of images whose size is within each band. There is only one cluster with more than 20K images, this is a cluster of outliers, and it contains 60K images.

## 5.2. Visualization

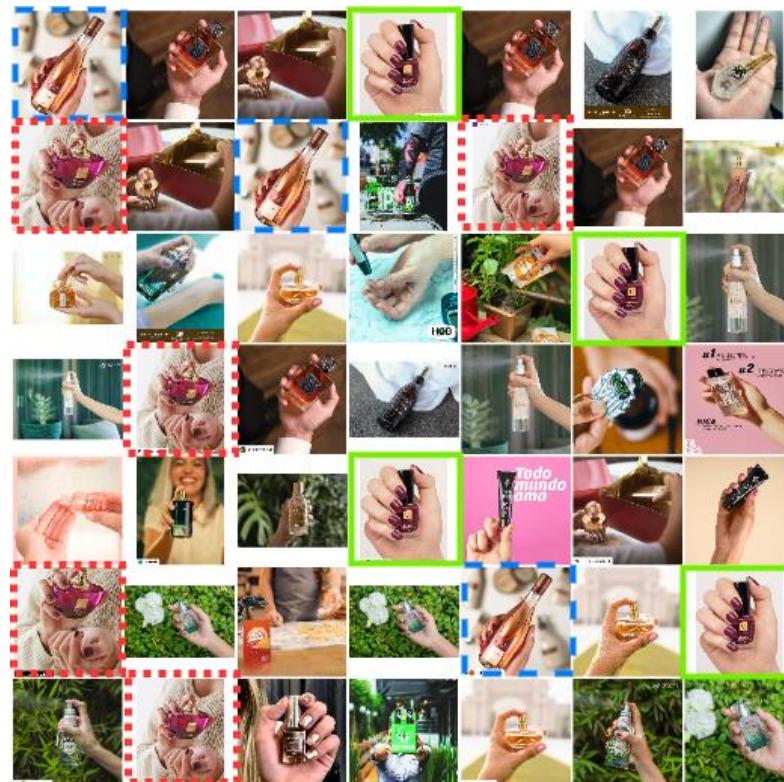
Here, we present image samples from representative and well-formed clusters. Note that we drew these samples before removing duplications. Also, we draw a sample of images from each cluster to visualize and possibly create classes of images afterward. For example, Figure 10 shows a cluster with images of many types of perfume and beauty products. We highlighted a few duplicated images, where the images with borders in the same color and line are considered duplications. Duplications are frequent in the clusters. As shown, one image may be duplicated more than once, and it is worth noting the importance of preprocessing these cases.

The sample of Figure 11 shows images of airplanes in several places, with or without other surrounding elements, flying or not, etc. It is worth noting the variety of aircraft positions, such as showing just the cockpit, wings, and turbine engines. This cluster is well-formed and consists of a representative set of images of airplanes.

Figure 12 shows a cluster that depicts birds from different species, in a diversity of positions, flying or not. Although a few images in this cluster illustrate advertisement campaigns, as can be seen by the texts in them, the images have in common that they show the birds as the primary elements.

As we can see both clusters illustrated by Figures 13 and 14 include images of cartoons. Note that these figures represent samples from different clusters. However, it may be observed that the cartoons of Figure 13 are different from the ones in Figure 14. We can note differences in stroke, colors, and style.

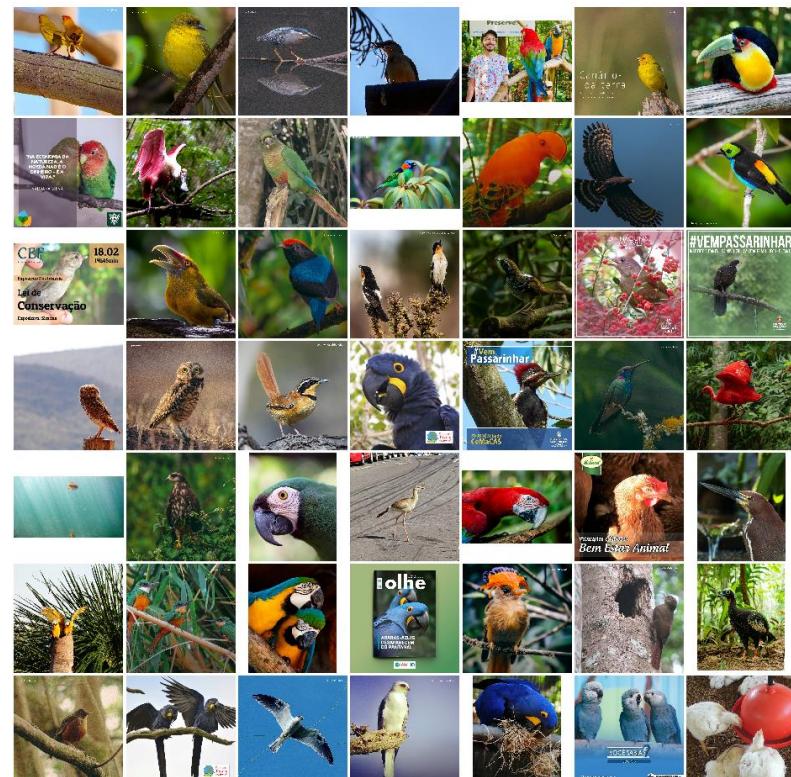
The cluster shown in Figure 15 consists of advertisements that illustrate draft laws presented in the Brazilian Congress. This type of post accounts for a significant part of this dataset because many government agencies and official politicians' profiles frequently publish posts in this format.



**Figure 10.** A sample of images from a cluster whose majority of the images are related to Perfumes. We highlighted the duplicated images, such that the ones with borders in the same color and line are considered duplications. Observe that there are many duplications in this cluster, and it is worth noting the importance of preprocessing these cases.



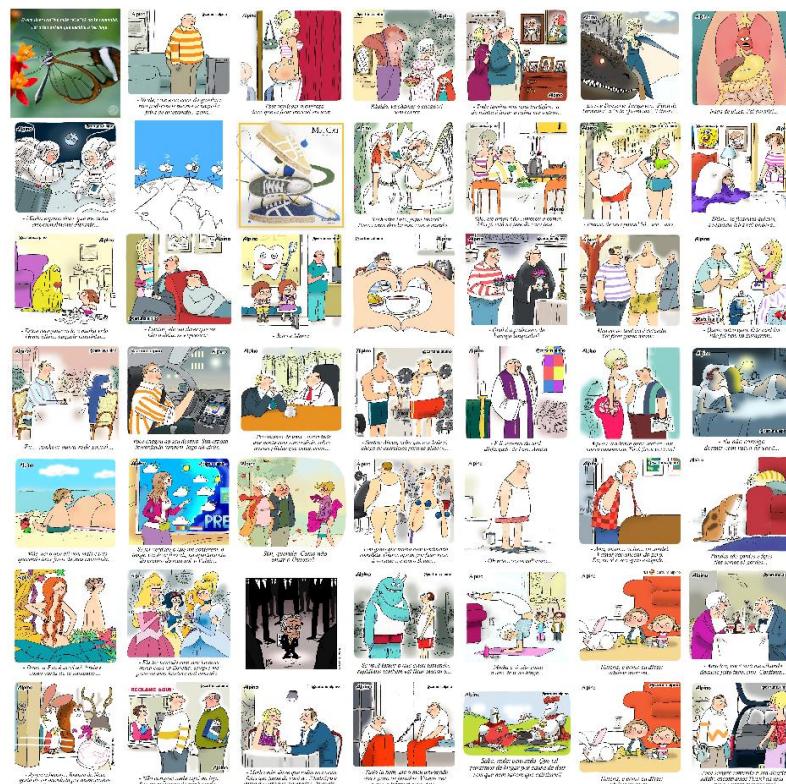
**Figure 11.** Sample from a cluster of airplanes. It is worth noting the variety of positions of the airplanes, some images show just part of them, such as the wings, the turbines, etc.



**Figure 12.** Samples from a cluster of birds. There is a diversity of species of birds as well as a variety of number of animals in the photographs.



**Figure 13.** Samples from a cluster of cartoons. It worth noting that most of the cartoons are made by the same author, and thus they have the same stroke.



**Figure 14.** Samples from a cluster of cartoons. Note that despite the images present in this cluster being cartoons, they have a different style from the ones in Figure 13.



**Figure 15.** Samples from a cluster of informative texts. This cluster illustrates draft laws presented in Brazilian Congress.

Moreover, to visualize the general topics occurring in captions of #PraCegoVer dataset, we carried statistical modeling, Topic Modeling, using Latent Dirichlet Allocation (LDA) algorithm [29]. This kind of analysis allows us to understand our data better. In Figure 16 we present word clouds of the most interpretable topics. Also, in these word clouds, we show the most frequent words for each topic.

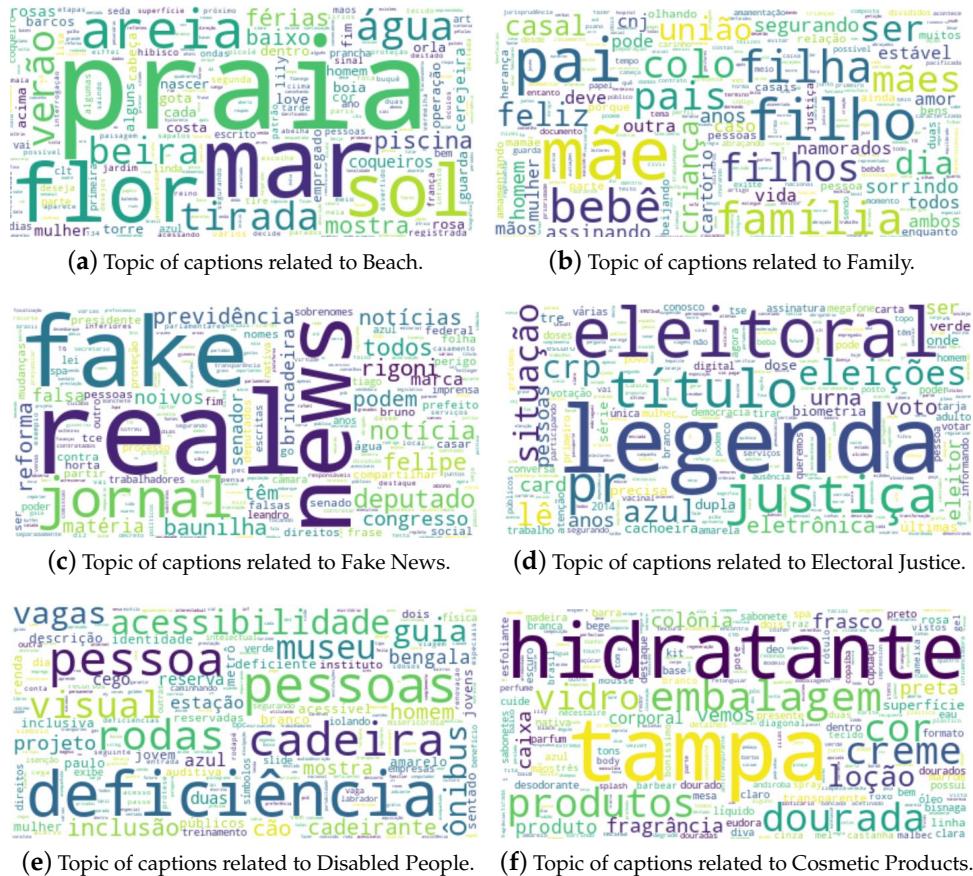
In Figure 16 we illustrate few topics in our dataset. Topic (a) shows elements of beach such as sand, sea, sun, and beach. We can see that Topic (b) is about Family because it comprehends words such as father, mother, baby, children, etc. In Topic (c) the words “fake” and “news” are the most often followed by “real” and “newspaper”, indicating it is related to “Fake News”. Topic (d) illustrates words related to Justice, specifically the Brazilian Electoral Justice, as seen in the terms “Electoral”, “Elections”, “Justice”, and “Allegiances”. Regarding Topic (e), it is based on concepts related to Disability, such as wheelchair, accessibility, and inclusion. Topic (f) is related to Cosmetic Products, having frequent words such as moisturizing cream, packaging, fragrance, perfume, etc.

### 5.3. Comparative Analysis

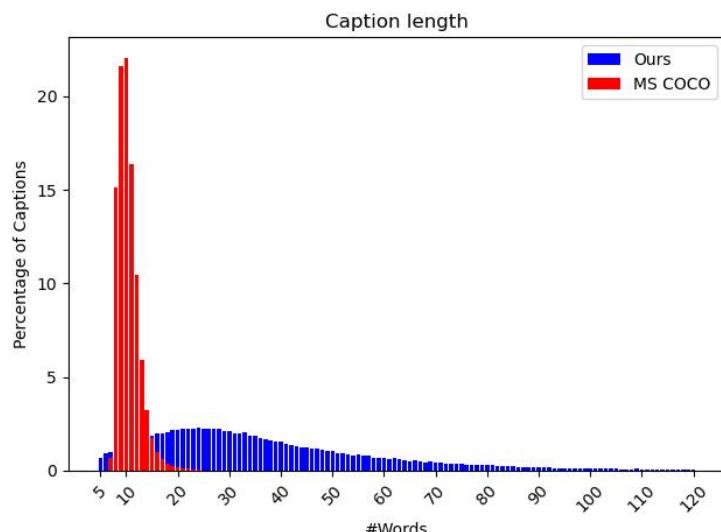
This section describes our dataset’s statistics compared to those of the MS COCO dataset since it is the most used dataset. To do so, we draw a sample of the whole #PraCegoVer and, after we preprocess it, we end up with the set that we named #PraCegoVer-173K.

Figure 17 shows the distribution of descriptions by length, in terms of number of words, in #PraCegoVer-173K and MS COCO datasets. The descriptions in our dataset have, on average, roughly 40 words, whereas MS COCO has only about 10. Also, the standard deviation of caption length in #PraCegoVer-173K is 29.7, while in MS COCO it is 2.2. These two characteristics make our dataset more challenging than MS COCO. Still, considering that the most employed evaluation metric in Image Captioning Literature, CIDEr-D [30], relies on MS COCO to set its hyperparameters, it does not work well in datasets where the caption length differ significantly from MS COCO. Moreover, it is essential to highlight that the majority of state-of-the-art models are trained using Self-Critical Sequence Training [31],

a Reinforcement Learning approach that aims to maximize CIDEr-D. Thus, the mean and variance of caption length in a dataset play an essential role in the final result.

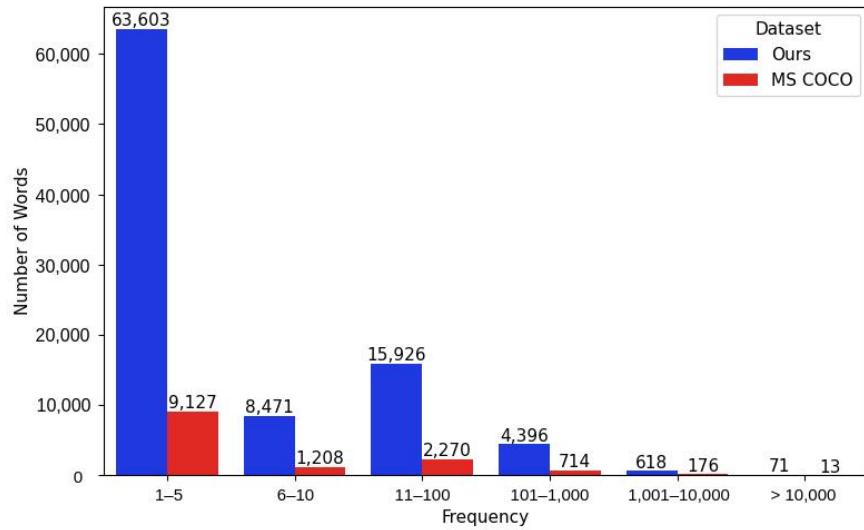


**Figure 16.** Word clouds showing the most frequent words in each topic found in the dataset. The topics were modeled using the LDA algorithm. It can be identifiable topics related to Beach, Family, Fake News, Electoral Justice, Disabled People, and Cosmetic Products.



**Figure 17.** Histogram of the distribution of captions by length in terms of number of words. We plot the caption length distribution for #PraCegoVer-173K (blue) and MS COCO (red) datasets. Figure reproduced from Santos et al. [32].

Moreover, we plot in Figure 18 the distribution of words by frequency, i.e., the number of occurrences of that word, for #PraCegoVer-173K and MS COCO datasets. On the x-axis, we show ranges of word frequency, and on the y-axis, we show the number of words whose frequency is within that band. As we can note, our dataset has by far more words occurring five or fewer times in the captions. Considering words as classes predicted by the models, if we train such models, they will “see” only a few examples of words with low frequency, then the models will not learn those classes. Therefore, this characteristic also makes our dataset more challenging than MS COCO.



**Figure 18.** Histogram of word frequency of #PraCegoVer-173K (blue) and MS COCO (red) datasets. We plot the number of words for each considering frequency range.

#### 5.4. Experiments

To validate our dataset and have a benchmark, we carried out experiments with AoANet [33], one of the state-of-art algorithms for MS COCO Captions. In our experiments, we selected two subsets of #PraCegoVer, initially with 100K and 400K posts. After we clean them, the resulting sets have 63K and 173K examples and we name them as #PraCegoVer-63K and #PraCegoVer-173K, respectively. Table 1 shows the statistics of each dataset used in our experiments. We trained and tested AoANet models on MS COCO Captions, #PraCegoVer-63K and #PraCegoVer-173K, three times for each dataset, considering the same hyperparameters. We trained the models firstly to optimize the Cross-Entropy Loss and then directly maximizing CIDEr-D score [30] using Self-Critical Sequence Training (SCST) [31]. We evaluated the models considering the same metrics used on MS COCO competition: BLEU [34], ROUGE [35], METEOR [36] and CIDEr-D [30].

**Table 1.** Statistics of each dataset used in our experiment. “Avg. Sent. Length” stands for the average sentence length, and “Std. Sent. Length” stands for the standard deviation of the sentence length.

Dataset	Dataset Size	Train Size	Validation Size	Test Size	Vocabulary Size	Avg. Sent. Length	Std. Sent. Length
MS COCO	123,287	113,287	5000	5000	13,508	10.6	2.2
#PraCegoVer-63K	62,935	37,881	12,442	12,612	55,029	37.8	26.8
#PraCegoVer-173K	173,337	104,004	34,452	34,882	93,085	39.3	29.7

#### 5.4.1. Results and Analysis

Table 2 shows that the performance of the model trained on MS COCO Captions drops considerably for all evaluation metrics compared to the one trained on #PraCegoVer-63K. We expected this result because of the difference in mean and variance of caption length between our dataset and MS COCO. Using the SCST approach to maximize the CIDEr-D score makes the models learn sentences with similar size more easily because CIDEr-D takes sentence size directly into account, and predicted sentences that differ in length from the reference are hardly penalized. Moreover, the longer the sentences a model predicts, the more likely they contain words not present in the reference. Thus, the models trained with SCST combined with the CIDEr-D score learn to predict short sentences. However, as our reference descriptions have 40 words on average, the performance is expected to be poor.

**Table 2.** Experimental results obtained by training AoANet model on #PraCegoVer-63K and MS COCO Captions. The performance of the model trained on MS COCO Captions drops considerably for all evaluation metrics compared to the ones trained on #PraCegoVer-63K and #PraCegoVer-173K. This result is expected because of the difference in mean and variance of caption length between our dataset and MS COCO.

Dataset	CIDEr-D	ROUGE-L	METEOR	BLUE-4
MS COCO Captions	$120.5 \pm 0.3$	$57.5 \pm 0.2$	$27.7 \pm 0.0$	$36.5 \pm 0.1$
#PraCegoVer-63K	$4.7 \pm 0.7$	$14.5 \pm 0.4$	$7.1 \pm 0.1$	$1.6 \pm 0.2$
#PraCegoVer-173K	$3.0 \pm 0.2$	$12.9 \pm 0.2$	$5.3 \pm 0.1$	$0.9 \pm 0.0$

#### 5.4.2. Qualitative Analysis

Figure 19 illustrates some images of #PraCegoVer-63K with their reference captions and the descriptions generated by the model trained on our dataset. It is worth noting that the model predicted incomplete sentences most of the time, as illustrated in Figure 19a. Also, it generates sentences where the same word is repeated many times, as can be seen clearly in Figure 19b,d. Finally, regarding advertisements where textual information is often present in the images, the model can not capture the texts on them and predicts meaningless punctuation signs, as shown in Figure 19c.



**(a) Reference:** Na foto, Thalita Gelens e Thaís Silva estão abraçadas com Luana Génot na livraria Travessa. Ao fundo, diversos livros coloridos estão na prateleira. Nas laterais da foto, existem 2 banners: um deles vermelho, com o logo da, e o outro com a divulgação do livro da Luana.

**Generated:** Foto de uma mulher segurando um livro com livros. Ao fundo, uma estante com livros. Texto: "A sua. É sua festa. É sua!"

**(b) Reference:** Em um ambiente externo, uma gata de pelagem branca e caramelo, está deitada de olhos fechados, sua ninhada de filhotinhos esta ao seu redor, um dos gatinhos esta olhando para a câmera, ele tem olhos cor de mel e pelagem branca, preta e caramelo, o restante dos filhotes, estão desfocados. No canto inferior direito, está escrito, "PremieRpet" em letras alaranjadas.

**Generated:** Foto de um gatinho de gato em uma gatinha de pelagem.



**(c) Reference:** Quadrado laranja. Ao centro o texto em cor branca: 13ª Semana pela paz em casa. Em ambos os lados pequenas barras na cor azul com os respectivos dados: 2.333 processos movimentados. 610 sentenças. 552 despachos. 348 medidas protetivas. 239 audiências.

**Generated:** Imagem com fundo amarelo. Texto: "você tem o que você ?". Texto: "você tem direito; você tem direito; você tem direito; você tem direito: R \$ 10%"



**(d) Reference:** Fotografia aérea sobre o pedágio da Terceira Ponte. A foto contém alguns prédios, um pedaço da Terceira Ponte e o fluxo de carros.

**Generated:** foto aérea aérea aérea da cidade de Florianópolis mostrando casas casas, mostrando casas casas. Ao fundo, algumas casas e casas.

**Figure 19.** Examples of images followed by their reference captions and the descriptions generated by the model trained on subset #PraCegoVer-63K optimizing CIDEr-D.

## 6. Usage Notes

Here, we present a detailed description of the dataset, a datasheet for the #PraCegoVer dataset, as proposed by Gebru et al. [37].

### 6.1. Motivation

#### 6.1.1. For What Purpose Was the Dataset Created?

#PraCegoVer dataset has been created to provide images annotated with descriptions in Portuguese for the image captioning task. With this dataset, we aim to alleviate the lack of datasets with Portuguese captions for this task.

### 6.1.2. Who Created the Dataset?

#PraCegoVer dataset was created by G.O.S., E.L.C., and S.A., on behalf of the Institute of Computing at the University of Campinas (Unicamp).

### 6.1.3. Who Funded the Creation of the Dataset?

The creation of #PraCegoVer dataset is partially funded by FAPESP grant (2019/24041-4).

## 6.2. Composition

### 6.2.1. What Do the Instances That Comprise the Dataset Represent?

The instances represent public posts collected from Instagram tagged with #PraCegoVer, comprising images and captions.

### 6.2.2. How Many Instances Are There in Total?

The dataset comprehends 533,523 instances.

### 6.2.3. What Data Does Each Instance Consist of?

Each instance consists of a json object, in the file `dataset.json`, with the attributes: `user`, `filename`, `raw_caption`, `caption`, `date`. These attributes represent, respectively, an identifier for the post owner (anonymized), the image file name, the raw caption (as originally written by the post author), the clean caption, and the post date. Each instance in `dataset.json` is associated with exactly one image in the `images` directory whose filename is pointed by the attribute `filename`, as illustrated in Figure 1.

### 6.2.4. Is There a Label or Target Associated with Each Instance?

The image captioning task consists of generating captions for images, thus the label of each instance is represented by its caption.

### 6.2.5. Are There Recommended Data Splits?

This dataset comes with two specified train/validation/test splits, one for #PraCegoVer-63K (train/validation/test: 37,881/12,442/12,612) and another for #PraCegoVer-173K (train/validation/test: 104,004/34,452/34,882). These splits are subsets of the whole dataset.

### 6.2.6. Are There Any Errors, Sources of Noise, or Redundancies in the Dataset?

#PraCegoVer dataset relies on data labeled in the wild, and captions are automatically cleaned from raw captions using regular expressions. Thus, captions attributes are susceptible to errors that are inherent to the source. Still, many posts are re-posted on Instagram, changing just a few details in the image and text. Then, some instances are very similar, and can be considered as duplication.

### 6.2.7. Does the Dataset Contain Data That Might Be Considered Confidential?

No, we only collect posts marked as public by their owners.

### 6.2.8. Does the Dataset Contain Data That, if Viewed Directly, Might Be Offensive, Insulting, Threatening, or Might Otherwise Cause Anxiety?

An initial analysis shows that although there exist words that can be offensive, as illustrated in Table 3, they are insignificant because they occur rarely. Note that overall such words occupy the position 30,000th or lower in the rank of word frequency. The dataset consists of data collected from public profiles on Instagram that were not thoroughly validated. Thus, there might be more examples of offensive and insulting content.

**Table 3.** This table illustrates words that may be offensive and their respective frequency and rank. Some words are clustered once they are just variations of genre or number (singular or plural), e.g., retardado(s)/retardada(s); in these cases we consider the ranking of the variation most frequent.

Words	#Occurrences	Ranking
idiota	143	28,520
puta	93	30,246
trouxa	62	39,845
viado	61	40,780
caralho	29	56,239
retardado(s)/retardada(s)	40	57,707
imbecil	49	68,800
quenga	14	82,439
escroto(s)/escrota(s)	25	82,865
mulato(s)/mulata(s)	51	85,614
sapatona	11	92,921
xana	10	96,428
vadia	10	107,233

#### 6.2.9. Is It Possible to Identify Individuals Either Directly or Indirectly?

The dataset consists of data collected from public profiles on Instagram. There are many examples of images of people, and, thus, the individuals present in those images can be identified.

#### 6.2.10. Does the Dataset Contain Data That Might Be Considered Sensitive in Any Way?

The dataset consists of data collected from public profiles on Instagram. Therefore, the images and raw captions might contain data revealing racial or ethnic origins, sexual orientations, religious beliefs, political opinions, or union memberships.

### 6.3. Collection Process

#### 6.3.1. How Was the Data Associated with Each Instance Acquired?

We collected the data directly from Instagram, and they were not entirely validated. Also, we extracted the audio description from the raw caption by using regular expressions.

#### 6.3.2. What Mechanisms or Procedures Were Used to Collect the Data?

We have implemented an automated crawler as detailed in Section 4.1.

#### 6.3.3. Who Was Involved in the Data Collection Process and How Were They Compensated?

The users from Instagram spontaneously published the posts, and we collected them using a crawler developed by the authors of this paper. Thus, regarding the collection process, the student, G.O.S., is compensated with the scholarship funded by the São Paulo Research Foundation (FAPESP) (2019/24041-4).

#### 6.3.4. Over What Time-Frame Was the Data Collected? Does This Time-Frame Match the Creation Time-Frame of the Data Associated with the Instances?

We daily collect the posts that were created any time ago. Hence, the creation and collection time-frames may not match.

#### 6.3.5. Were Any Ethical Review Processes Conducted?

No ethical review processes were conducted.

#### 6.3.6. Did You Collect the Data from the Individuals in Question Directly, or Obtain It via Third Parties or Other Sources?

We collected data from Instagram using a crawler, thus via third parties.

### 6.3.7. Were the Individuals in Question Notified about the Data Collection?

Since the data were automatically collected from public profiles, the individuals were not notified.

### 6.3.8. Did the Individuals in Question Consent to the Collection and Use of Their Data?

The Brazilian Law No. 13,709 (Portuguese version: [http://www.planalto.gov.br/ccivil\\_03/\\_ato2015-2018/2018/lei/l13709.htm](http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm), accessed on 14 January 2022, English version: [https://iapp.org/media/pdf/resource\\_center/Brazilian\\_General\\_Data\\_Protection\\_Law.pdf](https://iapp.org/media/pdf/resource_center/Brazilian_General_Data_Protection_Law.pdf), accessed on 14 January 2022), also known as General Data Protection Law, establishes rules for personal data collecting, storing, handling, and sharing. According to Article 11, Item II(c), the processing of sensitive personal data can occur without consent from the data subject when it is indispensable for studies carried out by a research entity, whenever possible, ensuring the anonymization of sensitive personal data.

Moreover, the individuals that have public profiles on Instagram consent to the use of their data once they accept the Data Policy of the platform<sup>3</sup>, thus they consent to have their data accessed and downloaded through third-party services. Therefore, we have not notified the individuals.

### 6.3.9. Has an Analysis of the Potential Impact of the Dataset and Its Use on Data Subjects Been Conducted?

Such analysis was not conducted. However, we conducted an initial analysis of the bias within our dataset. Please, refer to Section 6.5.4.

## 6.4. Preprocessing

### 6.4.1. Was Any Preprocessing of the Data Done?

We preprocess the raw caption by using regular expressions to extract the audio description part within the caption. Also, we use the Duplication Clustering Algorithm (see Algorithm 1) to cluster the instances of duplicated posts.

### 6.4.2. Was the “Raw” Data Saved in Addition to the Preprocessed Data?

In addition to the preprocessed data, we provide all the raw data.

### 6.4.3. Is the Software Used to Preprocess the Instances Available?

All the scripts used to preprocess the data are available on the repository of this project (<https://github.com/larocs/PraCegoVer>, accessed on 14 January 2022).

## 6.5. Uses

### 6.5.1. Has the Dataset Been Used for Any Tasks Already?

This dataset has been used for the image captioning task in Santos et al. [32].

### 6.5.2. Is There a Repository That Links to Any or All Papers or Systems That Use the Dataset?

No.

### 6.5.3. What Other Tasks Could the Dataset Be Used for?

This dataset could be used for image classification, text-to-image Generation, and sentiment analysis of the posts concerning specific periods such as electoral periods.

### 6.5.4. Is There Anything about the Composition of the Dataset or the Way It Was Collected and Preprocessed That Might Impact Future Uses?

We collected the data from public posts on Instagram. Thus the data is susceptible to the bias of its algorithm and stereotypes. We conducted an initial analysis of the bias within our dataset. Figure 20 shows that women are frequently associated with beauty, cosmetic products, and domestic violence. Moreover, black women co-occur more often

with terms such as “racism”, “discrimination”, “prejudice” and “consciousness”, whereas white women appear with “spa”, “hair” and “lipstick”, and indigenous women are mostly associated with beauty products. Similarly, black men frequently appear together with the terms “Zumbi dos Palmares”, “consciousness”, “racism”, “United States” and “justice”, while white men are associated with “theatre”, “wage”, “benefit” and “social security”. In addition, Table 4 shows that women are more frequently associated with physical words (e.g., thin, fat); still, fat people appear more frequently than thin people. Figure 21 illustrates that fat women are also related to swearing words, “mental harassment”, “boss”, while thin women are associated with “vitamin”, fruits, “healthy skin”. To sum up, depending on the usage of this dataset, future users may take these aspects into account.



**Figure 20.** Word clouds showing the most frequent words associated with women and men from different ethnicities.

**Table 4.** This table illustrates how physical characteristics are more related to woman. In addition, “fat people” occur more frequently than “thin people”; these scenarios should be taken into account to avoid biases on models.

Words	#Occurrences	Ranking
gordo	467	18,145
gordos	26	62,002
gorda	806	8579
gordas	176	27,263
magro	286	20,161
magros	20	68,425
magra	410	17,853
magras	96	30,195



(a) Word cloud related to thin women.



(b) Word cloud related to fat women.

**Figure 21.** Word clouds showing the most frequent words associated with fat and thin women. Thin women are associated with vitamins, fruits, and a healthy lifestyle in general. In contrast, fat women are associated with plus-size style, boss, mental harassment, and swear words.

## 6.6. Distribution

### 6.6.1. Will the Dataset Be Distributed to Third Parties Outside of the Entity on Behalf of Which the Dataset Was Created?

We released the dataset under license CC BY-NC-SA 4.0. We request that the ones who use this dataset cite this paper. Commercial use of this dataset is not permitted.

### 6.6.2. How Will the Dataset Be Distributed?

This dataset is available at <http://doi.org/10.5281/zenodo.5710562> (accessed on 14 January 2022) with restricted access because it contains images of people, and it consists of data collected from public profiles on Instagram. Thus, the images and raw captions might contain sensitive data that reveal racial or ethnic origins, sexual orientations, religious beliefs. Hence, under Brazilian Law No. 13,709 (please, refer to Section 6.3.8), to avoid the unintended use of our dataset, we decided to restrict its access, ensuring that the dataset will be used for research purposes only. Still, we will make it available under request explaining the objectives of the research.

### 6.6.3. When Will the Dataset Be Distributed?

#PraCegoVer dataset will initially release upon acceptance of this paper, and new versions will be released from time to time.

### 6.6.4. Will the Dataset Be Distributed under a Copyright or Other Intellectual Property License, and/or under Applicable Terms of Use?

The dataset is released under CC BY-NC-SA 4.0 license. We request that the ones who use this dataset cite this paper. Commercial use of this dataset is not permitted.

### 6.6.5. Do Any Export Controls or Other Regulatory Restrictions Apply to the Dataset or to Individual Instances?

We will share this dataset upon request. The ones interested in #PraCegoVer have to tell us the Institution they are working on, briefly describe their project and how this dataset will be used.

## 6.7. Maintenance

### 6.7.1. Who Is Supporting/Maintaining the Dataset?

#PraCegoVer dataset is maintained by G.O.S.. All comments or requests can be sent to the email address [g194760@dac.unicamp.br](mailto:g194760@dac.unicamp.br).

### 6.7.2. How Can the Curator of the Dataset Be Contacted?

All comments or requests can be sent to G.O.S. through the email address [g194760@dac.unicamp.br](mailto:g194760@dac.unicamp.br).

### 6.7.3. Will the Dataset Be Updated?

#PraCegoVer dataset will be continuously updated, and the authors of this dataset will release new versions with new or deleted instances on the Zenodo repository (<http://doi.org/10.5281/zenodo.5710562>, accessed on 14 January 2022). In addition, we will release notes on the #PraCegoVer repository (<https://github.com/larocs/PraCegoVer>, accessed on 14 January 2022) with the updates.

### 6.7.4. Will Older Versions of the Dataset Continue to Be Maintained?

We will keep track of old dataset versions. Thus, they will be available for download.

### 6.7.5. If Others Want to Contribute to the Dataset, Is There a Mechanism for Them to Do So?

Contributors shall contact the maintainer, G.O.S., by email through the address [g194760@dac.unicamp.br](mailto:g194760@dac.unicamp.br).

**Author Contributions:** G.O.d.S., E.L.C. and S.A. conceived the experiments, G.O.d.S. conducted the experiments, G.O.d.S., E.L.C. and S.A. analyzed the results. All authors reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** G.O.d.S. is funded by the São Paulo Research Foundation (FAPESP) (2019/24041-4)<sup>4</sup>. E.L.C. and S.A. are partially funded by H.IAAC (Artificial Intelligence and Cognitive Architectures Hub). S.A. is also partially funded by FAPESP (2013/08293-7), a CNPq PQ-2 grant (315231/2020-3), and Google LARA 2020.

**Institutional Review Board Statement:** The Brazilian National Health Council's Resolution no. 510/2016<sup>5</sup> states the specific ethical procedures for Social Science and Humanities research (Item II, Article 1): "Research using publicly accessible information will not be registered or evaluated by the National Committee of Ethics in Research (CEP/CONEP) system, according to Law No. 12,527, of 18 November 2011". Our dataset comprises data collected from Instagram's public profiles, with no access restriction. Thus, the data collection does not require CEP/CONEP approval.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** This dataset is available at <http://doi.org/10.5281/zenodo.5710562> (accessed on 14 January 2022).

**Acknowledgments:** We would like to thank Artificial Intelligence (Recod.ai) lab. of the Institute of Computing at University of Campinas (Unicamp) for permitting us to use resources to carry out our research.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Notes

- 1 Hereinafter, we use caption and description interchangeably.
- 2 <https://github.com/larocs/PraCegoVer>, accessed on 14 January 2022.
- 3 [https://www.facebook.com/help/instagram/519522125107875/?helpref=hc\\_fnav&bc\[0\]=Instagram%20Help&bc\[1\]=Policies%20and%20Reporting](https://www.facebook.com/help/instagram/519522125107875/?helpref=hc_fnav&bc[0]=Instagram%20Help&bc[1]=Policies%20and%20Reporting) (accessed on 14 January 2022).
- 4 The opinions expressed in this work do not necessarily reflect those of the funding agencies.
- 5 Resolution no. 510/2016 (in Portuguese): <http://conselho.saude.gov.br/resolucoes/2016/Reso510.pdf>, accessed on 14 January 2022.

## References

1. Web para Todos. Criadora do Projeto #PraCegoVer Incentiva a Descrição de Imagens na Web. 2018. Available online: <http://mwpt.com.br/criadora-do-projeto-pracegover-incentiva-descricao-de-imagens-na-web> (accessed on 14 January 2022).
2. Hodosh, M.; Young, P.; Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.* **2013**, *47*, 853–899. [[CrossRef](#)]
3. Plummer, B.; Wang, L.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vis.* **2017**, *123*, 74–93. [[CrossRef](#)]
4. Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollar, P.; Zitnick, C.L. Microsoft COCO captions: Data collection and evaluation server. *arXiv* **2015**, arXiv:1504.00325.
5. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 2556–2565. [[CrossRef](#)]
6. Sidorov, O.; Hu, R.; Rohrbach, M.; Singh, A. Textcaps: A dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 742–758. [[CrossRef](#)]
7. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 483–498. [[CrossRef](#)]
8. Rosa, G.; Bonifacio, L.; de Souza, L.; Lotufo, R.; Nogueira, R.; Melville, J. A cost-benefit analysis of cross-lingual transfer methods. *arXiv* **2021**, arXiv:2105.06813.
9. Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755. [[CrossRef](#)]
10. Rashtchian, C.; Young, P.; Hodosh, M.; Hockenmaier, J. Collecting image annotations using amazon’s mechanical turk. In *Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*; Association for Computational Linguistics: Los Angeles, CA, USA, 2010; pp. 139–147. [[CrossRef](#)]
11. Farhadi, A.; Hejrati, M.; Sadeghi, M.; Young, P.; Rashtchian, C.; Hockenmaier, J.; Forsyth, D. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 15–29. [[CrossRef](#)]
12. Elliott, D.; Keller, F. Image description using visual dependency representations. In *Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Seattle, WA, USA, 2013; pp. 1292–1302. <https://aclanthology.org/D13-1128>
13. Zitnick, C.; Parikh, D.; Vanderwende, L. Learning the visual interpretation of sentences. In *Proceedings of the IEEE International Conference on Computer Vision*, Sydney, Australia, 1–8 December 2013; pp. 1681–1688. [[CrossRef](#)]
14. Kong, C.; Lin, D.; Bansal, M.; Urtasun, R.; Fidler, S. What are you talking about? Text-to-image coreference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 23–28 June 2014; pp. 3558–3565. [[CrossRef](#)]
15. Harwath, D.; Glass, J. Deep multimodal semantic embeddings for speech and images. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, Scottsdale, AZ, USA, 13–17 December 2015; pp. 237–244. [[CrossRef](#)]
16. Gan, C.; Gan, Z.; He, X.; Gao, J.; Deng, L. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 21–26 July 2017; pp. 3137–3146. [[CrossRef](#)]
17. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [[CrossRef](#)]
18. Levinboim, T.; Thapliyal, A.; Sharma, P.; Soricut, R. Quality Estimation for Image Captions Based on Large-scale Human Evaluations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 3157–3166. [[CrossRef](#)]
19. Hsu, T.; Giles, C.; Huang, T. SciCap: Generating Captions for Scientific Figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021; pp. 3258–3264. [[CrossRef](#)]

20. Lam, Q.; Le, Q.; Nguyen, V.; Nguyen, N. UIT-ViC: A Dataset for the First Evaluation on Vietnamese Image Captioning. In *International Conference on Computational Collective Intelligence*; Springer International Publishing, Cham, Switzerland, 2020; pp. 730–742. [[CrossRef](#)]
21. Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; Anderson, P. nocaps: Novel object captioning at scale. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8948–8957. [[CrossRef](#)]
22. Krasin, I.; Duerig, T.; Alldrin, N.; Ferrari, V.; Abu-El-Haija, S.; Kuznetsova, A.; Rom, H.; Uijlings, J.; Popov, S.; Kamali, S.; et al. OpenImages: A Public Dataset for Large-Scale Multi-Label and Multi-Class Image Classification. 2017. Available online: <https://storage.googleapis.com/openimages/web/index.html> (accessed on 14 January 2022).
23. Gurari, D.; Zhao, Y.; Zhang, M.; Bhattacharya, N. Captioning images taken by people who are blind. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 417–434. [[CrossRef](#)]
24. Park, C.; Kim, B.; Kim, G. Attend to you: Personalized image captioning with context sequence memory networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 895–903.
25. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520. [[CrossRef](#)]
26. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61*, 611–622. [[CrossRef](#)]
27. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426.
28. McInnes, L.; Healy, J.; Astels, S. HDBSCAN: Hierarchical density based clustering. *J. Open Source Softw.* **2017**, *2*, 205. [[CrossRef](#)]
29. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022. [[CrossRef](#)]
30. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575. [[CrossRef](#)]
31. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7008–7024. [[CrossRef](#)]
32. Santos, G.; Colombini, E.; Avila, S. CIDEr-R: Robust Consensus-based Image Description Evaluation. In *Proceedings of the Seventh Workshop on Noisy User-Generated Text (W-NUT 2021)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 351–360. [[CrossRef](#)]
33. Huang, L.; Wang, W.; Chen, J.; Wei, X. Attention on attention for image captioning. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4634–4643. [[CrossRef](#)]
34. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318. [[CrossRef](#)]
35. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
36. Lavie, A.; Agarwal, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June 2007; pp. 228–231. [[CrossRef](#)]
37. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.; Wallach, H.; Daumé, H., III; Crawford, K. Datasheets for datasets. *Commun. ACM* **2021**, *64*, 86–92. [[CrossRef](#)]