Labeling guide

I. Purpose of labeling

The task of labeling data is essential to ensure the quality of any machine learning model [1]. The primary goal of this labeling task is to read and evaluate a number of sentences based on a selection of labels and assign one (or more) labels to each sentence. By reading and evaluating the content of each sentence, the idea is that we can make a qualified assumption about what the URL in each sentence links to, e.g., a dataset, a software package, or something else [2]. The labeled data will be used to fine-tune a pre-trained SciBERT model [3] and subsequently used to classify a larger set of sentences to identify URLs that link to datasets.

II. Task overview

Based on a list of labels, you will assign labels to 129 sentences that were extracted from forty different NeuroImage articles. The labels were created based on a manual reading and analysis of sentences containing URLs from ten NeuroImage articles. Based on the initial coding, the codes were gathered into groups and defined. All labels available for this labeling task is available in table I). You will be using Taguette [4], which is an open-source qualitative data analysis tool, for the labeling.

A. How to decide on labels?

For each sentence, you will use your understanding of the label's definitions, the examples presented in table I, your prior knowledge and understanding of the sentence to determine the appropriate label. In figure 1 and 2, four sentences are highlighted with yellow and each sentence has been labeled using the schema. I will explain why they were labeled the way they were.

The first sentence in fig. 1 "200 unrelated subjects were selected from the Human Connectome Project (HCP) 1200 Subjects Data Release with available resting (task-free) and task fMRI data from a 3T MRI scanner (https://db.humanconnectome.org/data/projects/HCP_1200)." is given the label **Dataset**, because the sentence mentions that subjects were selected and because it specifies the type of data acquired (resting and task fMRI data). As such, the URL is likely to provide access to a dataset. Furthermore, the HCP dataset is a relatively well-known dataset in the neuroimaging field, and you are free to use your knowledge of things like this to decide on a label.

The second sentence in fig. 1 "This study agreed to the Open Access Data Use Terms (https://www.humanconnectome.org/study/hcp-young-adult/document/wu-minn-hcp-consortium-open-access-data-use-terms) and was exempt from the UCSF IRB because investigators could not readily ascertain the identities of the individuals to whom the data belonged." is given the label

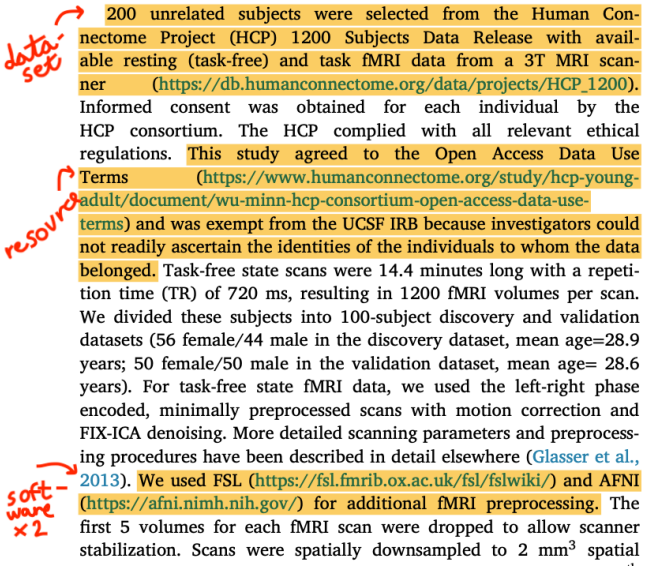

Fig. 1. This is a screenshot of a research paper. Three sentences containing URLs have been highlighted with yellow to illustrate how the sentences were extracted. Next to each sentence, three labels are written with red text. The choice of label is described in section II-A

Resource, because the sentence discusses the conditions and regulations surrounding data use. This makes it more likely that the link is a resource for understanding the legal aspects surrounding using the HCP data and not a link to a dataset.

The third sentence in fig. 1 "We used FSL (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) and AFNI (https://afni.nimh.nih.gov/) for additional fMRI preprocessing." contains two URLs. For this classification task, both URLs would be extracted and the same sentence will be used for classification. In this case, both would be given the label **Software**, because the sentence extracted for both talk about preprocessing. This indicate that the links most likely lead to a software's or tool's website.

In figure 2, another sentence containing two URLs is highlighted. Figure 3 illustrates how these URLs and sentences will look in Taguette. In this case, both sentences will be given the label **Atlas/map**. The first URL, http://www.brainnetome.org/, is classified as such, because the URL follows a mention of the Brainnetome atlas. Similarly, the second URL, http://www.diedrichsenlab.org/imaging/suit.htm, is classified as Atlas/map because the URL follows a mention of the the SUIT atlas.

Regarding sentences that contain multiple links, they cannot not always be given the same label, because the URLs might not both be related at the same category, as was the case in the two previous examples. It is therefore important to be aware of which URL you are working with.

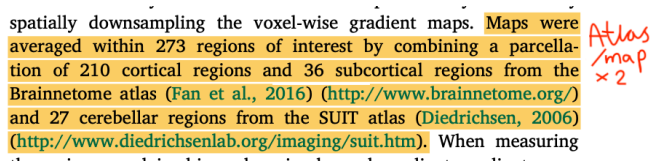

Fig. 2. This is a screenshot of a research paper. One sentence containing multiple URLs has been highlighted with yellow.

http://www.brainnetome.org/ ['Maps were averaged within 273 regions of interest by combining a parcella-tion of 210 cortical regions and 36 subcortical regions from the Brainnetome atlas (Fan et al., 2016) (http://www.brainnetome.org/) and 27 cerebellar regions from the SUIT atlas (Diedrichsen, 2006) (http://www.diedrichsenlab.org/imaging/suit.htm).'] 10.1016/j.neuroimage.2022.119526

http://www.diedrichsenlab.org/imaging/suit.htm ['Maps were averaged within 273 regions of interest by combining a parcella-tion of 210 cortical regions and 36 subcortical regions from the Brainnetome atlas (Fan et al., 2016) (http://www.brainnetome.org/) and 27 cerebellar regions from the SUIT atlas (Diedrichsen, 2006) (http://www.diedrichsenlab.org/imaging/suit.htm).'] 10.1016/j.neuroimage.2022.119526

Fig. 3. This image simulates how the URLs from fig. 2 will be extracted and input into Taguette.

In summary, based on the content of the sentences you determine which label is the most appropriate for each sentence based on its description and examples presented in the label schema. If you feel more than one label is appropriate, use more labels. You are not required to open any of the URLs before you classify the sentence, and you are not required to read the papers.

III. Accessing and using Taguette

You will be performing the labeling using the platform Taguette (https://www.taguette.org/). You need to have a user that I can share the project with. Once you access the project, you will find a number of documents titled 'Anonymous[Animal]' that contain the same list of 129 URLs and associated sentences (as seen on fig. 4). Pick a document that has not been worked in yet. To label the sentences, you have to highlight them and click the button 'new highlight' (fig. 5). Chose whichever label you feel best categorize the sentence (all labels presented in table I are imported into Taguette). If none of the labels feel appropriate, you can chose to 'Create a tag'. If you chose this option, please provide an informative description of the tag.

IV. Tips for annotation

- Use the provided label schema to pick a label. It contains definitions of the labels as well as examples to help you decide.
- Sometimes, there are multiple URLs in one sentence. When you label a sentence, you have to take the URL in front of the sentence into consideration, as explained previously in relation to fig. 2 and 3.
- Some URLs are mentioned in multiple sentences. SciBERT will receive all of the sentences as one input and use them for the classification. As such, when you label these, take all of the sentences into consideration and evaluate them together.

Fig. 4. Screenshot of my project on Taguette. Once a sentence has been labeled, it becomes yellow. I added the red and green boxes outside of Taguette. The red boxes highlight the URLs, and the green boxes highlight the DOI of the article.

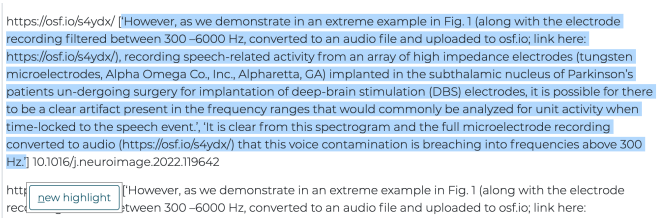

Fig. 5. Screenshot of a highlighted sentence in Taguette. Upon highlighting a sentence, you have to click the button 'new highlight' to give it a label.

- You are free to use more than one label for a sentence if you find them appropriate.
- You are free to use your prior knowledge about a certain URL to label the sentences.
- If you find that none of the provided labels are appropriate, create and define a new label.

If you encounter any challenges or uncertainties during the annotation process, feel free to contact me at carva@itu.dk for assistance.

References

[1] R. S. Geiger *et al.*, ""Garbage In, Garbage Out" Revisited: What Do Machine Learning Application Papers Report About Human-Labeled Training Data?" *Quantitative Science Studies*, vol. 2, no. 3, pp. 795–827, Nov. 2021, arXiv:2107.02278 [cs], ISSN: 2641-3337. DOI: 10.1162/qss_a_00144. [Online]. Available: http://arxiv.org/abs/2107.02278 (visited on 10/25/2023).

[2] L. Salsabil *et al.*, "A Study of Computational Reproducibility using URLs Linking to Open Access Datasets and Software," en, in *Companion Proceedings of the Web Conference 2022*, Virtual Event, Lyon France: ACM, Apr. 2022, pp. 784–788, ISBN: 978-1-4503-9130-6. DOI: 10.1145/3487553.3524658. [Online]. Available: https://dl.acm.org/doi/10.1145/3487553.3524658 (visited on 10/30/2023).

[3] I. Beltagy, K. Lo, and A. Cohan, *SciBERT: A Pretrained Language Model for Scientific Text*, en, arXiv:1903.10676 [cs], Sep. 2019. [Online]. Available: http://arxiv.org/abs/1903.10676 (visited on 10/25/2023).

[4] R. Rampin and V. Rampin, "Taguette: Open-source qualitative data analysis," en, *Journal of Open Source Software*, vol. 6, no. 68, p. 3522, Dec. 2021, ISSN: 2475-9066. DOI: 10.21105/joss.03522. [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.03522 (visited on 10/29/2023).

[5] A. Butterfield, G. E. Ngondi, and A. Kerr, Eds., *A Dictionary of Computer Science*, en, 7th ed. Oxford University Press, 2016, ISBN: 978-0-19-176812-5. [Online]. Available: https://www.oxfordreference.com/display/10.1093/acref/9780199688975.001.0001/acref-9780199688975.

[6] D. A. Dickie *et al.*, "Whole Brain Magnetic Resonance Image Atlases: A Systematic Review of Existing Atlases and Caveats for Use in Population Imaging," *Frontiers in Neuroinformatics*, vol. 11, 2017, ISSN: 1662-5196. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fninf.2017.00001 (visited on 10/28/2023).

[7] A. C. Evans, A. L. Janke, D. L. Collins, and S. Baillet, "Brain templates and atlases," *NeuroImage*, 20 YEARS OF fMRI, vol. 62, no. 2, pp. 911–922, Aug. 2012, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2012.01.024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811912000419 (visited on 10/28/2023).

[8] A. Hess, R. Hinz, G. A. Keliris, and P. Boehm-Sturm, "On the Usage of Brain Atlases in Neuroimaging Research," en, *Molecular Imaging and Biology*, vol. 20, no. 5, pp. 742–749, Oct. 2018, ISSN: 1536-1632, 1860-2002. DOI: 10.1007/s11307-018-1259-y. [Online]. Available: http://link.springer.com/10.1007/s11307-018-1259-y (visited on 10/28/2023).

[9] M. Jenkinson, M. Chappell, M. Jenkinson, and M. Chappell, *Introduction to Neuroimaging Analysis* (Oxford Neuroimaging Primers). Oxford, New York: Oxford University Press, Oct. 2017, ISBN: 978-0-19-881630-0.

[10] R. D. Markello *et al.*, "Neuromaps: Structural and functional interpretation of brain maps," en, *Nature Methods*, vol. 19, no. 11, pp. 1472–1479, Nov. 2022, Number: 11 Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-022-01625-w. [Online]. Available: https://www.nature.com/articles/s41592-022-01625-w (visited on 10/28/2023).

[11] Andriy Burkov, *The Hundred-Page Machine Learning Book*, en. 2019. [Online]. Available: http://themlbook.com/ (visited on 08/23/2023).

[12] X. Han *et al.*, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, Jan. 2021, ISSN: 2666-6510. DOI: 10.1016/j.aiopen.2021.08.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666651021000231 (visited on 10/28/2023).

[13] J. Sterne, *Plug-in*, Dec. 2022. [Online]. Available: https://www.britannica.com/technology/plug-in (visited on 10/28/2023).

TABLE I
SCHEMA WITH EXPLANATIONS FOR EACH (GROUP OF) LABELS.

Group	Labels	Description	Example
The labels in this group are centered around data. Data is defined as something distinguished from other types of information on which computers operate, with the distinguishing characteristics being: "(...) that it is organized in a structured, repetitive, and often compressed way." [5].	Dataset	This label is used for sentences containing URLs that are associated with datasets. A dataset is defined as: "A coherent body of data with well-defined selection criteria and internal structure" [5]. In the context of neuroimaging, the dataset can consist of a variety of different modalities, e.g., fMRI, EEG, CT, and PET.	"EEG datasets used to create the figure in this commentary are freely available at osf.io/gazx2/, osf.io/eucqf/, osf.io/thsqg, and osf.io/bndjg/." "200 unrelated subjects were selected from the Human Connectome Project (HPC) 1200 Subjects Data Release with available resting (task-free) and task fMRI data from a 3T MRI scanner (https://db.humanconnectome.org/data/projects/HCP_1200)."
	Processed dataset	This label is used for sentences containing URLs that are associated with processed datasets. A processed dataset is defined similarly to a dataset [See Dataset], but it will be prefaced by words like "preprocessed" or "averaged".	"Code used to reproduce the plots in Fig. 1, as well as averaged ERP data, is available from osf.io/guwnm/." "Preprocessed fMRI data are available at https://osf.io/b8pfa/?view_only=b6dbb5dd6a044989a7eecdc99facb43c."
The labels in this group are data adjacent. They are considered adjacent because they do not contain individual images or raw data that can be used to, but are instead processed to create something else. Data adjacent forms of information include atlases and models.	Atlas/map	This label is used for sentences containing URLs that are associated with brain atlases or maps. An atlas is a summarized representation of one or more individual scans of brains, and they are used for registering or mapping structural or functional parts of the brain [6–9]. They provide a standardized spatial reference system to help determine specific sites within the brain [6, 9]. The term map has been used synonymously (e.g., [6, 7] refer to MNI152 as an atlas whereas [10] refer to it as a map), and some authors use atlas and template interchangeably [6].	"Maps were averaged within 273 of interest by combining a parcellation of 210 cortical regions and 36 subcortical regions from the Brainnetome atlas (Fan et al.,2016) (http://www.brainnetome.org/) and 27 cerebellar regions from the SUIT atlas (Diedrichsen, 2006) (http://www.diedrichsenlab.org/imaging/suit.htm)." "Users can also download all code used to generate the functional connectivity maps from https://gitlab.com/cfmm/marmoset-connectivity."
	Model	This label is used for sentences containing URLs that are associated with machine learning models. A (machine learning) model is an algorithmically built statistical model which is trained on a dataset to find patterns. The model is assumed to be used to solve problems, e.g. classification or prediction [11]. A specific type of model is a pre-trained model, where the trained model has been shared so that others can use it as is or they can use transfer learning to customize the model to a task [12].	"Another version of pre-trained AlexNet was im-ported from Caffe Model Zoo (https://caffe.berkeleyvision.org/model_zoo.html)."
The labels in this group are all made up of code. Code is any piece of program text written in a programming language.	Software, incl. plugins, toolbox, packages, and functions	This label is used for sentences containing URLs that are associated with software, plugins, toolboxes, packages, functions, or similar. Software is: "(. . .) commonly used to refer to the programs executed by a computer system" [5]. A plugin (also called add-on or extension) is a software component that adds new features or functions to an existing program [13]. A (software) toolbox is a set of tools that can develop, repair, or enhance programs or hardware [5]. Application packages are also called software packages, and they contain a collection of programs that are directed at an application [5]. Functions are defined as a: "(...) program unit that given values for inputparameters computes a value." [5].	"Subsequently the results were loaded in a Matlab Tool Box, Brainstorm (Tadel et al. 2011), an accredited software freely available for download online under the GNU general public license (http://neuroimage.usc.edu/brainstorm)." "All image transformations were done with Clipping Magic (https://clippingmagic.com), ImageMagick, GIMP, Microsoft Paint, the MATLAB SHINE toolbox, and custom MATLAB code." "Task condition block regressors were convolved with a hemo-dynamic response function using the 'spm_get_bf" function in SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/)."
	Analysis	This label is used for sentences containing URLs that are associated with code that perform analysis on data.	"Codes for the fMRI data analysis at https://github.com/Yozafirova/monkey-fMRI-codes and for the CNN data analysis at https://github.com/RajaniRaman/face_body_integration."
The label in this group are centered around any URLs that have to do with the methodology, excluding software.	Resource	This label is used for sentences containing URLs that are associated with resources used in the research methodology. These resources can include images, equipment, or other materials that played a role in the research process, excluding software used for data analysis.	"Both experiments employed static images (modified from Shutterstock, https://www.shutterstock.com)." "MRI data were acquired on a 3T Discovery MR750 scanner (Gen-eral Electric Healthcare, Milwaukee, WI, USA) equipped with a 32-channel head coil (Nova Medical, Wilmington, MA, USA) at the Center for Cognitive and Neurobiological Imaging at Stanford University (www.cni.stanford.edu)." This study agreed to the Open Access Data Use Terms (https://www.humanconnectome.org/study/hcp-young-adult/document/wu-minn-hcp-consortium-open-access-data-use-terms) and was exempt from the UCSF IRB because investigators could not readily ascertain the identities of the individuals to whom the data belonged.'
Other	Not a URL	The extraction of the URLs is not perfect and may have picked up on text that is not actually a URL. If that is the case, use this label.	
	Not enough information	Sometimes, the extraction of the URLs and sentences is not perfect, and sometimes the sentence does not contain enough information to determine what the URL links to.	
	Free text	If there are any sentences containing URLs that cannot be classified using any of the previous labels, write a new label.	