Readme

Touché24-ValueEval

Version: 2024-08-09 [doi] [task]

Note: New versions will be announced on our mailing list.

Dataset for ValueEval'24 @ Touché: Human Value Detection. The dataset is organized in the following directories:

- training/validation/test (in valueeval24.zip). Contains the sentences and labels of the respective dataset split (60%/20%/20%).
- training-english/validation-english/test-english (in valueeval24.zip). Contains the sentences and labels of the respective dataset split, translated to English (if necessary) using the DeepL API or (for Hebrew) Google Translate API.
- valueeval23 (in valueeval23.zip). Only use for comparison with previous year. Not part of the ValueEval24 competition. It contains the 1576 arguments of the ValueEval'23 test dataset. The "sentences" are the original dataset's premises (often more than a sentence). Somewhat arbitrarily, arguments in favor of are marked as (partially) attaining the respective values, whereas arguments against are marked as (partially) constraining the respective values. This assignment to attain and constrain allows to kickstart classifier development, but should not be used for anything further.

For each directory listed above, the dataset contains the following files:

- sentences.tsv. Contains one sentence per line:
  - Text-ID identifies the text that contains the sentence
  - Sentence-ID gives the index of the sentence in the text
  - Text is the sentence text itself
- labels.tsv. Contains one sentence per line:
  - Text-ID same as for sentences.tsv
  - Sentence-ID same as for sentences.tsv
  - For each of the 19 values two columns:
    - One column <value> attained with a 1 meaning that the sentences refers to this value and (partially) attains it
    - One column <value> constrained with a 1 meaning that the sentences refers to this value and (partially) constrains it If both are 0 the sentence does not refer to that value at all. If both are 0.5 the sentence refers to the value but it is unclear whether it (even partially) attains or constrains it.

Sentences were split first using paragraph information from the sources and then using the Trankit sentence splitter (version 1.1.1; https://github.com/nlp-uoregon/trankit).

Value Taxonomy

The value-categories.json describes the 19 value categories of this task. Format:

```
{
  "<value tag>": {
    "name": "<value name>",
    "goal": "<brief description of the goal associated with the value>",
    "personal-motivation": "<personal motivations for working towards the goalY"
  }, ...
}
```

Reading the dataset in Python

Both labels.tsv and sentences.tsv can be read with Pandas:

```
import pandas
data_frame = pandas.read_csv(file_path, encoding="utf-8", sep="\t", header=0)
```

For use with Transformers, one can use this method:

```
import datasets
import numpy
import os
import pandas
import transformers

values = [ "Self-direction: thought", "Self-direction: action", "Stimulation",  "Hedonism",
"Achievement", "Power: dominance", "Power: resources", "Face", "Security: personal", "Security:
societal", "Tradition", "Conformity: rules", "Conformity: interpersonal", "Humility", "Benevolence:
caring", "Benevolence: dependability", "Universalism: concern", "Universalism: nature",
"Universalism: tolerance" ]
labels = sum([[value + " attained", value + " constrained"] for value in values], [])

pretrained_model = "bert-base-uncased" # example
tokenizer = transformers.AutoTokenizer.from_pretrained(pretrained_model)
```

```
def load_dataset(directory, tokenizer, load_labels=True):
    sentences_file_path = os.path.join(directory, "sentences.tsv")
    labels_file_path = os.path.join(directory, "labels.tsv")

    data_frame = pandas.read_csv(sentences_file_path, encoding="utf-8", sep="\t", header=0)
    encoded_sentences = tokenizer(data_frame["Text"].to_list(), truncation=True)

    if load_labels and os.path.isfile(labels_file_path):
        labels_frame = pandas.read_csv(labels_file_path, encoding="utf-8", sep="\t", header=0)
        labels_frame = pandas.merge(data_frame, labels_frame, on=["Text-ID", "Sentence-ID"])
        labels_matrix = numpy.zeros((labels_frame.shape[0], len(labels)))
        for idx, label in enumerate(labels):
            if label in labels_frame.columns:
                labels_matrix[:, idx] = (labels_frame[label] >= 0.5).astype(int)
        encoded_sentences["labels"] = labels_matrix.tolist()

    encoded_sentences = datasets.Dataset.from_dict(encoded_sentences)
    return encoded_sentences, data_frame["Text-ID"].to_list(), data_frame["Sentence-ID"].to_list()
```

Authors

The ValuesML Team

Project Coordinators

- Bertrand De Longueville, Joint Research Centre (JRC)
- Johannes Kiesel, Bauhaus-Universität Weimar
- Theresa Reitis-Münstermann, Joint Research Centre (JRC)
- Mario Scharbillig, Joint Research Centre (JRC)
- Paula Schulze Brock, Joint Research Centre (JRC)
- Nicolas Stefanovitch, Joint Research Centre (JRC)

Language Leads

- Murat Ardag, Bremen International Graduate School of Social Sciences
- Sharon Arieli, The Hebrew University Business School
- Ella Daniel, Tel Aviv University
- Henrik Dobewall, Finnish Institute for Health and Welfare
- Anna Krasteva, New Bulgarian University
- Thomas Peter Oeschger, University of Basel
- Luana Russo, Maastricht University
- Antonella Seddone, University of Turin

- Joanne Sneddon, The University of Western Australia
- Aurelia Tamo-Larrieux, Maastricht University
- Hester van Herk, Vrije Universiteit Amsterdam
- Johannes Karl, Dublin City University
- Georgios Petasis, NCSR "Demokritos"

Annotators and Curators

- Sandrine Astor, Pacte research centre, School of Political Studies Univ. Grenoble Alpes
- Petra Auer, Free University of Bozen-Bolzano
- Nazan Avcı, Middle East Technical University Northern Cyprus Campus
- Anat Bardi, Royal Holloway University of London
- Fiorella Battaglia, University of Salento, Lecce & Ludwig-Maximilians-Universität, München
- Constanze Beierlein, Hamm-Lippstadt University of Applied Sciences
- Maya Benish-Weisman, The Hebrew University of Jerusalem
- Giuliano Bobba, UNITO
- Christina Christodoulou, NCSR "Demokritos"
- Patricia Collins, Edith Cowan University
- Irene Coppola, ReCEPL-Università degli Studi di Napoli Federico II-
- Ahmet Çoymak, Abdullah Gul University
- Maria Dagioglou, National Centre for Scientific Research "Demokritos"
- Meike Morren, University Amsterdam
- Einat Elizarov, The University of Haifa
- Naama Erlich, Bgu
- Uwana Evers, The University of Western Australia
- Peculiar Tochukwu Ezeigwe-Ephraim, ISCTE - Instituto Universitário de Lisboa
- Maria Cristina Gaeta, Research Centre of European Private Law (ReCEPL), Suor Orsola Benincasa University of Naples
- Lucilla Gatt, Università degli STudi Suor Orsola Benincasa
- Sjoukje Goldman, Amsterdam University of Applied Sciences
- Frederic Gonthier, Sciences Po Grenoble - School of Political Studies, Grenoble Alpes University, France
- Stefanie Habermann, Royal Holloway University
- Mina Hristova, Bulgarian Academy of Sciences
- Demet Islambay Yapali, Independent Researcher
- Ömer Topuz, Abdullah Gul University
- Luigi Izzo, Suor Orsola Benincasa
- Panos Kapetanakis, National Centre for Scientific Research 'Demokritos'
- Agathi Karadima, National Centre for Scientific Research 'Demokritos'
- Dora Katsamori, National Centre for Scientific Research 'Demokritos'
- Reşit Kışlıoğlu, Middle East Technical University Northern Cyprus Campus
- Roberta Koleva, New Bulgarian University
- Joshua Lake, University of Western Australia
- Ingmar Leijen, Vrije Universiteit Amsterdam
- Adva Liberman, The Hebrew University of Jerusalem
- Vanina Ninova, Policy and Citizens' Observatory
- Elif Sandal Önal, Bielefeld University
- Berna Öney, Carl von Ossietzky Universität Oldenburg
- Shani Oppenheim Weller, Hadassah Academic College
- Duygu Ozturk, Istanbul Medipol University
- Ioannis Elissaios Paparrigopoulos, NCRS Demokritos
- Vladimir Ponizovskiy, Ruhr-Universität Bochum
- Tim Reeskens, Tilburg University
- Maria Francesca Romano, Scuola Superiore Sant'Anna
- Torven Schalk, Te Herenga Waka - Victoria University of Wellington

- Ricarda Scholz-Kuhn, University of Basel
- Oscar Smallenbroek, Joint Research Centre
- Evelina Staykova, New Bulgarian University
- Maite Taboada, Simon Fraser University
- Christin-Melanie Vauclair, Iscte- University Institute of Lisbon
- Adam Wyner, Swansea University
- Sheng Ye, East China University of Science and Technology
- Emilia Zankina, Temple University
- Chaya Koleva, Policy and Citizens' Observatory

Data Cleaning

- Nicolas Handke, Leipzig University

Version History

- 2024-08-09
  - Added test labels
- 2024-04-15
  - Last data and Hebrew translations (using Google Translate)
- 2024-04-03
  - Fixed TSV escaping for English translations, loads correctly in pandas now
- 2024-04-02
  - More data
  - Added DeepL translations to English (Hebrew not supported)
- 2024-02-15
  - Fixed MAC OS line endings
- 2024-02-13
  - First version of ValuesML data
- 2023-12-16
  - Initial, ValueEval'23 only

Data Usage Agreement

The dataset may include content which is protected by copyright of third parties. It may only be used for scientific research purposes in the context of human value detection. The dataset may not be redistributed or shared in part or full with any third party. You may not share you access with others or give access to the dataset to unauthorised persons. Any other use is explicitly prohibited.