

# At the Intersection of NLP and Sustainable Development: Exploring the Impact of Demographic-Aware Text Representations in Modeling Value on a Corpus of Interviews.

Goya van Boven<sup>1</sup>, Stephanie Hirmer<sup>2,4</sup>, Costanza Conforti<sup>3</sup>

<sup>1</sup> Department of Information and Computing sciences, Utrecht University

<sup>2</sup> Energy and Power Group, University of Oxford

<sup>3</sup> Language Technology Lab, University of Cambridge

<sup>4</sup> Rural Senses Ltd.

j.g.vanboven@students.uu.nl

## Abstract

This research explores automated text classification using data from Low- and Middle-Income Countries (LMICs). In particular, we explore enhancing text representations with demographic information of speakers in a privacy-preserving manner. We introduce the **Demographic-Rich Qualitative UPV-Interviews Dataset (DR-QI)**, a rich dataset of qualitative interviews from rural communities in India and Uganda. The interviews were conducted following the latest standards for respectful interactions with illiterate speakers (Hirmer et al., 2021a). The interviews were later sentence-annotated for Automated User-Perceived Value (UPV) Classification (Conforti et al., 2020), a schema that classifies values expressed by speakers, resulting in a dataset of 5,333 sentences. We perform the UPV classification task, which consists of predicting which values are expressed in a given sentence, on the new DR-QI dataset. We implement a classification model using DistilBERT (Sanh et al., 2019), which we extend with demographic information. In order to preserve the privacy of speakers, we investigate encoding demographic information using autoencoders. We find that adding demographic information improves performance, even if such information is encoded. In addition, we find that the performance per UPV is linked to the number of occurrences of that value in our data.

**Keywords:** Demographic-Aware Text Representation, Multi-Class Multi-Label Text Classification, Privacy-Preserving NLP, Computational Social Science and Cultural Analytics

## 1. Introduction

The interplay between language use and demographic factors has been widely studied in sociolinguistics (Labov, 1972; Extra and Gorter, 2001), and has more recently found application in NLP (Hovy, 2015; Vijayaraghavan et al., 2017). While this is an important step in NLP, most research on demographic-aware text representation examines only a handful of features. These are often modelled separately: however, in reality, identities are composite and stratified, and they result from the complex interplay and mutual influence of different demographic elements (McCall, 2005). Studying each element in isolation is therefore too simplistic and ultimately artificial (Herbelot et al., 2012). Moreover, almost all research in this area has focused on High-Income Countries, with a particular strong focus on the US (Blodgett et al., 2016). However, culturally-specific elements play an important role in defining someone’s identity and certain demographic aspects can be even more determinant in shaping a person’s language use, particularly in traditional societies where social roles can be more rigid (Jayachandran, 2020). This paper addresses some of the gaps mentioned above and makes the following contributions: (1) We focus on text classification with a rich set of demographic features, including less often considered attributes such as income and number of children. To our knowledge, it

is the first time that such a comprehensive set of demographic features is investigated in NLP. (2) We collect and release the **Demographic-Rich Qualitative UPV-Interviews Dataset (DR-QI)**, a new dataset for multi-class multi-label text classification collected as part of a sustainable development project in two Low- and Middle-Income countries (LMICs). Samples in DR-QI are labelled with the interviewees’ values they contain (User-Perceived Values or UPVs). (3) We implement a model which integrates textual and demographic features. To further protect speakers’ identities, we propose a simple recipe to protect such information. (4) Our experimental results show that a more comprehensive demographic representation of the speaker is beneficial for UPV classification, and that less studied demographic features contribute to the overall performance more than traditionally considered aspects. This holds true even when encoded in a privacy preserving manner using an Autoencoder. This suggests new research directions for demographic-aware text classification, privacy-preserving representations, and more generally at the intersection of NLP and sustainable development.

## 2. Background

**User-Perceived Values (UPVs) for Sustainable Projects Planning.** We use the UPV framework to better understand what is important to rural communities in India and Uganda. UPV, as defined by

(Hirmer, 2018), refers to “the benefits, concerns, feelings and underlying drivers that vary in importance and act as the main motivators in the lives of the people — as perceived and defined by the [people] themselves at a given time”. The UPV classification schema (Appendix B) divides values into six broad categories which are split up into 17 tiers, which in turn group the 58 UPVs. Accounting for values in project design and implementation in LMICs is important to achieving sustainable development through improved project or intervention buy-in (Howells et al., 2010; Hirmer, 2018).

**NLP for Sustainable Development: UPV Classification.** In recent years, Machine Learning techniques have been applied to a wide range of sustainability-related problems (Gomes et al., 2019; Hager et al., 2017; Shi et al., 2020; Vinuesa et al., 2020); however, research on NLP for sustainable development still constitutes a relatively new field (Conforti et al., 2020; Schröder et al., 2021). In this context, Conforti et al. (2020) proposed to use NLP to partially automate UPV labelling of qualitative interviews. If we consider data from qualitative interviews as representing a “snapshot” of a community at a given point (Caselli et al., 2021), reducing the annotation bottleneck between data collection and project development is very important to ensure that development interventions are timely and accurate (Figure 1).

**Decolonising AI.** AI practices, even when striving for positive effects, often perpetuate colonial structures (Mohamed et al., 2020), to which language technologies are no exception (Bird, 2020). Likewise, AI for sustainable development initiatives risks perpetuating technological solutionism and paternalistic attitudes (Mohamed et al., 2020). Alternatively, a decolonial approach involves vulnerable communities having self-determination and being active participants in the design process, as for example in (Nekoto et al., 2020).

### 3. The DR-QI Dataset

The sustainable uptake of energy services is crucial for achieving the Sustainable Development Goals (SDG), in particular SDG 7 on energy access (EnDev, 2020). In line with this, the DR-QI dataset was collected as part of a wider project investigating the social, environmental and economic impact of off- and weak-grid energy appliances in Low- and Middle-Income Countries undertaken by the Efficiency for Access Coalition together with Rural Senses and SVT. As part of this, 214 interviews were conducted in two LMICs: India and Uganda. Data collection consisted of three parts, (i) the UPV game, (ii) a socio-economic survey, and (iii) the UPV game specific to four appliance types (namely: fan, fridge, TV and water pump). The aim was to shed light into perceived everyday challenges and needs of local communities, as well as to better understand the value placed on said appliances. This is important to ensure that potential users of energy initiatives are at

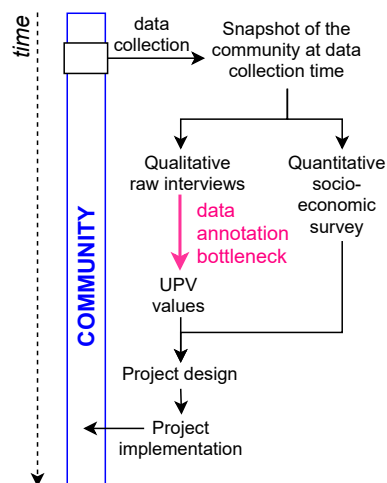


Figure 1: Communities are constantly evolving. Data collected during fieldwork represent a snapshot of the community at a specific moment in time. NLP techniques can be used to reduce the data annotation bottleneck (pink arrow), thus allowing project design to be based on a reliable portrait of the community.

the centre of decision-making. In this work we focus on the data collected as part of (i) and (ii).

#### 3.1. Data Collection

**Location of the communities and interviewee selection.** We collected data from 214 interviewees across 138 rural communities: 56 in India and 68 in Uganda. The following interviewee sampling criteria were applied. Interviewees had to (a) cover a variety of ages; (b) have equal gender distribution; (c) have diverse occupations and (d) income ranges; and (e) cover both, appliance ownership or non-ownership.

**Data collection setting.** In total, for both Uganda and India, 12 local data collectors, with good knowledge of the local languages, were hired to carry out the interviews. In order to minimise interviewer bias (Frey, 2018; Matarazzo et al., 1963), data collectors took part in a training workshop where they practised mock interviews.

**The UPV Game.** Following the methodology proposed by Hirmer (2018) and recently adopted in other NLP works (Conforti et al., 2020), we used the UPV game as a means to obtain responses from illiterate speakers. This method has successfully been used for data collection in LMICs (Hirmer and Guthrie, 2016), and biases, such as priming effect, are minimised (Hirmer et al., 2021a). The UPV game is framed as a semi-structured interview, in which interviewees are asked to: **(1)** select 5 out of 45 presented items<sup>1</sup>; **(2)** rank the items in order of importance and, for each item, **(3)** explain why

<sup>1</sup>Cards depicted everyday items found in rural areas in LMICs and were adapted to the local context with the help of local experts. Indian and Ugandan interviewees were therefore exposed to slightly different, culturally specific cards.



Figure 2: UPV interview taking place in Uganda<sup>2</sup>.

the item was selected. At this last stage *why-probing* is used to capture multiple layers of reasoning (Reynolds and Gutman, 1988; Zaltman and Coulter, 1995; Coulter and Zaltman, 1994).

**Socio-economic survey.** To contextualise the data collected from the UPV game, a socio-economic survey with the same participants was carried out.

**Ethical considerations.** Interviews followed the country’s COVID-guidance: data collectors and participants were provided with safety equipment. We followed the recommendations of Hirmer et al. (2021a) for respectfully contacting and interacting with community members. Each interview lasted between 1–2 hours. Interviewees were compensated with the equivalent of 1 day labour.

### 3.2. Data pre-processing and annotation

Following data collection, interviews were translated into English. This step was necessary not only because English constitutes the de-facto standard language of sustainable development (Coleman, 2010), but also to assure the protection of participants’ identity: as illiterate and isolated rural communities tend to display a high degree of language fragmentation (Hazen, 2002), precise geographic information could implicitly be released by directly sharing the original transcript.

**Annotation process and Ethical Considerations.** We hired 8 annotators based in LMICs, which were familiar with the local context in which the interviews were collected. Annotators received training and were asked to submit a short quiz to validate their understanding before they could start annotation work. We employed the original 58 UPV label schema proposed by Hirmer (2018) (the UPV label schema including definitions can be found in Appendix B). To ease the complex process

<sup>2</sup>While permission to photograph participants was granted, photos were pixelated in order to protect their identity.

of UPV annotation, we implemented a user-friendly platform. Each sentence was separately labelled by between 3 to 5 annotators. We only kept labels which were selected at least 3 times. The entire annotation process took 4 weeks. Annotators were paid a competitive salary appropriate to the local context.

We acknowledge that the choice of annotators might reinforce the emergence of bias (Waseem, 2016; Sap et al., 2019; Geva et al., 2019). To mitigate the presence of annotation bias in the DR-QI corpus, we largely followed the processes laid out by Hirmer et al. (2021a).

### 3.3. Data sharing

We release the DR-QI corpus under a Academic Free License agreement<sup>3</sup>. We are aware of the many ethical issues surrounding data collected from marginalised communities in locations with low democracy (Hovy and Spruit, 2016). Virtually all models trained on language data are dual-use (Benton et al., 2017): in order to avoid potential misuse, we will share our dataset only upon signing a data sharing agreement restricting the data usage to research only.

### 3.4. Data Analysis

In order to help future researchers’ assessment and effective usage of the data, we compiled a data statement (Bender and Friedman, 2018) for the DR-QI dataset (Appendix A).

**Textual Features.** The final corpus collects 5,333 annotated sentences. The sentences, which may present constructions typical of spoken language, are on average 12.7 tokens, with Ugandan speakers being more talkative than Indian speakers (15.5 vs 9.1 tokens per sentence). On average, each sample received 1.21 gold UPV annotations: an example of an annotated sentence is reported in Figure 3. This resulted in 77.7 average occurrences per UPV label, but with a very skewed UPV label distribution (Figure 4): the most common label (*Income*) occurs 562 times, and the least common (*Morality*) just once. This was expected, and was noticed in other works on UPV classification (Conforti et al., 2020).

**Demographic Features.** For each speaker, we considered 10 demographic features, each of which has been categorised in order to decrease the samples granularity. Features were selected based on the information collected in the socio-economic survey. While they do not aim in any way to be exhaustive in describing a speaker, they nevertheless constitute a considerably broader set than typical in NLP works. Due to the exploratory nature of this study, we leave a deeper analysis about feature correlation, and their implications in modeling, for future work. Features are self-reported by the speakers and fall into the following categories:

1. *Gender*<sup>4</sup>: woman or man.

<sup>3</sup>The DR-QI corpus can be found at the link: <https://github.com/gvanboven/DR-QI>

<sup>4</sup>In this paper, we refer to gender as a social construc-

Item	Television
Text	We get all kinds of news like Covid related news which helps us to stay alert.
Country	India
UPVs	[Safety] [Preventative health] [Information access]

Figure 3: A labeled sentence from the DR-QI dataset.

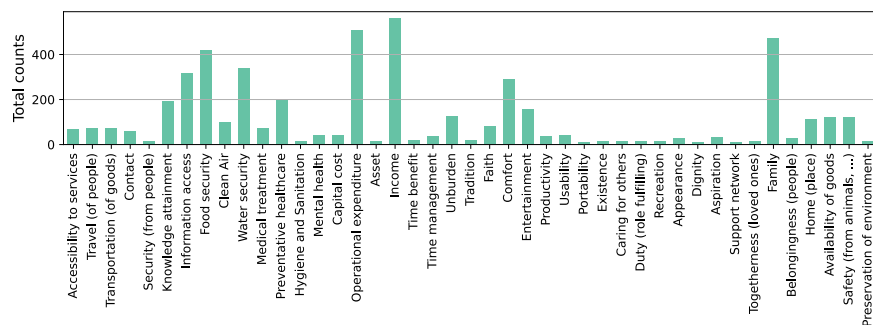


Figure 4: UPV label frequency in the DR-QI dataset.

2. *Marital Status*: married, single, or other (widowed, divorced, ...).
3. *Disability*: yes or no.
4. *Age ranges*: 18-25, 25-35, 35-50, 50-60, 60+.
5. *Education*: None, Primary, Secondary, University, Vocational training.
6. *Household Size*: 1-4, 5-7, 8-10, 10+ people.
7. *Number of children*: 1-2, 3-4, 5-7, 7+.
8. *Income (relative)*: low, middle, high<sup>5</sup>.
9. *Poverty Status*: <\$3.20, <\$5.50, <\$10 a day, or above the poverty rate<sup>6</sup> (Roser and Ortiz-Ospina, 2013).
10. *Occupation*, distinguishing 16 job types.

The features’ distribution across Indian and Ugandan speakers (Figure 5) are largely similar. A first divergence can be observed in *number of children*, with Ugandan interviewees reporting a higher number of children (4.7 vs. 2.2 on average). This reflects culturally specific aspects of Ugandan and Indian societies. The number of children in rural Uganda remains high. This may be attributed to cultural norms and a lack of contraception (Nalwadda et al., 2010; Bbaale and Mpuga, 2011). In contrast, India has seen a significant slow-down. This is partially due to its greater investment in family planning but also due to programmes of mass sterilisation (from 1976 to 1992) (Oliveira et al., 2014). Notably, reported education levels are considerably lower than in high-income societies (in line with Van Hiel et al. (2018)); moreover, the age distribution is centred around young and middle-aged people, with relatively few interviewees older than 60.

**Discussion.** We note that the DR-QI dataset contains a remarkably rich set of demographic features, many of which have never been considered in NLP research so far. Therefore, our dataset will be an invaluable source for future work exploring the impact and the combination of various demographic features on language pro-

tion (Buchanan, 2018; Scott, 1986). All interviewed in this study self-identified as woman or man: therefore, cognizant of the multitude of different genders and sexual identities, this study refers to men and women as cisgender/cissexual.

<sup>5</sup>Relative to the speakers average income for each country.

<sup>6</sup>Income in the local currency was converted to dollars.

cessing. However, NLP research which includes any demographic information should not be disjointed from discussions about privacy preserving measures. As a first measure, to protect the speakers’ identity, we categorised any potentially granular information – such as age, household size, or income – into ranges. Secondly, to remove location-based information, we are only releasing the country where the interview was taken; more granular geographical features, such as the exact village name or its coordinates, are not released. Therefore, the study participants are not identifiable from the rest of the country population (resp. 45 million and 1.38 billion people in Uganda and India). Additionally, we manually went through all sentences in order to remove any identifiable information such as names of people, locations and companies.

Researchers, however, should be aware that *all* data can potentially be identifiable and therefore sensitive (Salganik, 2019), and could be misused (Tockar, 2014). Although our dataset is anonymised, we acknowledge that de-identification of (some of) our participants might potentially be possible, for example in the case that malicious users combined several other datasets with DR-QI (Sweeney, 2002; Narayanan and Shmatikov, 2008). For this reason, as a further privacy-preserving countermeasure, in this paper we explore models which can take advantage of demographic information about the speakers by leveraging opaque encoded representations of such features (further discussion in Section 4.3).

#### 4. Models for UPV Classification

We consider the task of UPV annotation where, given an interview input sentence  $x$ , a model has to predict the UPV label(s)  $L$  that are mentioned in the text. Given the very large tagset considered (58 possible UPVs) and its extremely skewed distribution in our data, we adopt the same framework as in Conforti et al. (2020) and approach the task as a sequence of *binary classification* problems: that is, instead of directly predicting  $x \rightarrow \{l_2, l_4\}$  (i.e.  $x \rightarrow [0, 1, 0, 1]$  for a tagset  $|T| = 4$ ), we predict  $(x, l_1) \rightarrow 0$ ,  $(x, l_2) \rightarrow 1$ , etc.



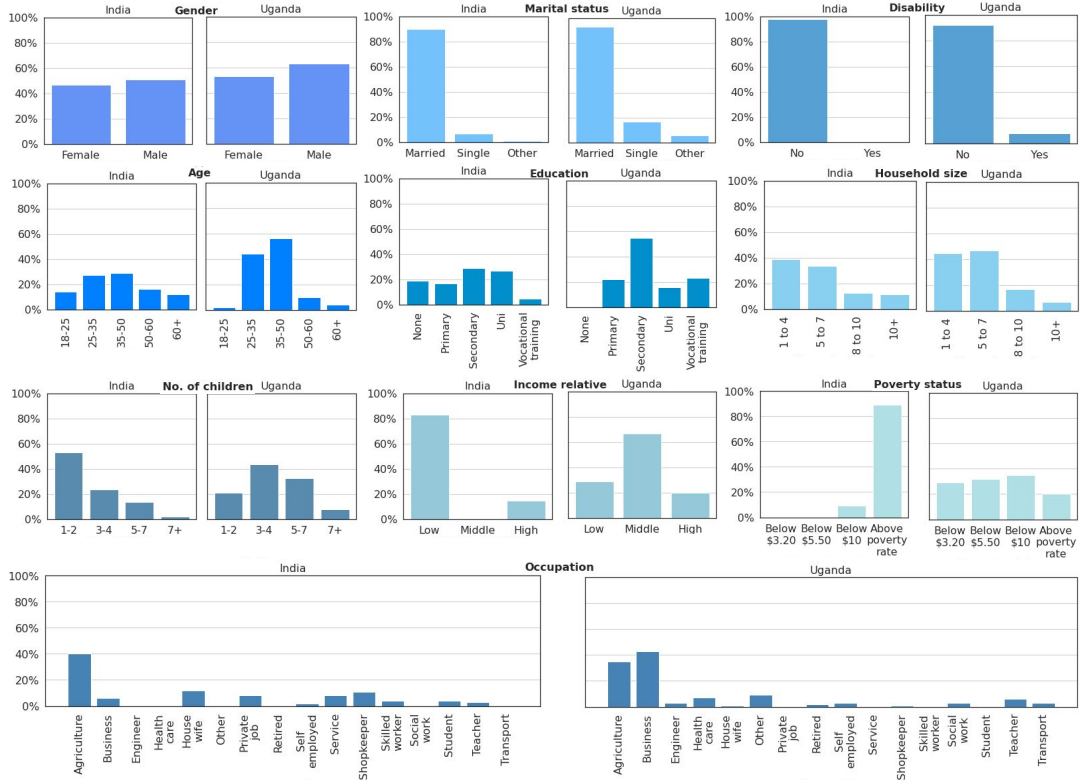


Figure 5: Statistical features of demographic features in the DR-QI dataset

#### 4.1. Transformers

As a strong baseline, we employ a transformer-based classifier which employs DistilBERT (Sanh et al., 2019) as the main encoder. The model receives the input sentences as: [CLS] text [SEP] label [SEP], where text is the input sentence, and label is the textual representation of the UPV label. The relatedness of the given UPV label with the input text is predicted with a dense layer followed by a sigmoid operation on the hidden state  $h_{[CLS]}$  of the special classification embedding [CLS].

#### 4.2. Blending Demographic Attributes

We implement a simple architecture which integrates textual features with the speaker’s demographic information (module 2 in Figure 6):

*Text Encoder.* We employ the same DistilBERT architecture introduced in Subsection 4.1.

*Demographic Encoder.* We one-hot encode the demographic features in 3.4 and concatenate them into a vector  $d$ . We obtain an encoded representation  $h_d = \text{relu}(\text{relu}(d))$ , where  $\text{relu}$  is a dense layer with ReLU activations (Glorot et al., 2011).

*Blending Signals.* We combine the text and the demographic signals into  $h_b$  as<sup>7</sup>:

$$h_b = \text{relu}(h_d \oplus h_{[CLS]}) \quad (1)$$

<sup>7</sup>In our preliminary experiments, this simple approach performed better than more complex blending methods, such as bilinear transformations (Sawhney et al., 2020).

We want to obtain a hidden representation of the input sentence which captures the hidden correlations with its speaker’s demographic attributes. To this goal, we further process  $h_b$  by passing it through a sequence of concatenation operations with the original text representation  $h_{[CLS]}$  and dense layers with ReLU activations. Formally, we obtain the final hidden representation  $h$  as:

$$h = \text{relu}(\text{relu}(h_b \oplus h_{[CLS]}) \oplus h_{[CLS]}) \quad (2)$$

*Output Layer.* We finally predict the sample text’s relatedness with respect to the given label with a sigmoid operation on  $h$ .

#### 4.3. Privacy-Preserving Demographic-Aware UPV Classification

A potentially problematic point of the model described in section 4.2 is that the demographic vector could potentially disclose sensitive information about participants. As discussed in Section 3.4, it might be debatable whether this is a real possibility in the case of our dataset. Nevertheless, as a further step to protect our participants’ identity, we experiment with a UPV classifier which takes as input an encoded, opaque demographic vector.

**Methods.** Research in privacy-preserving NLP has been widely addressed in recent literature (Feyisetan et al., 2020; Feyisetan et al., 2021). This topic assumes a particular relevance when considering that data from vulnerable communities can be particularly susceptible to exploitation (Christiaensen and Subbarao,

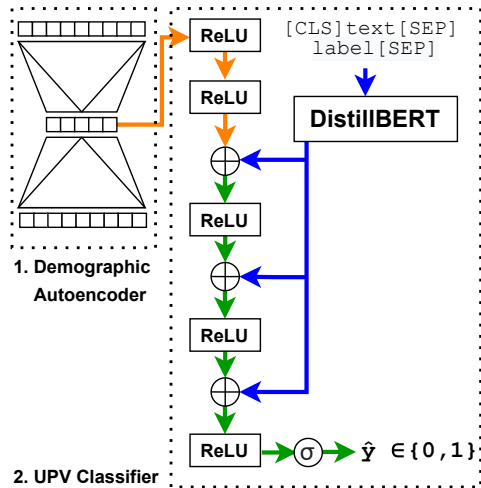


Figure 6: Overview of the demographic-aware models. Module 2 represents the UPV classifier (Sec. 4.2): blue, orange, and green arrows are textual, demographic, and blended signals. We also consider models which take as input encoded demographic vectors (Sec. 4.3) produced by a separately trained autoencoder (module 1).

2005; De Cénival, 2008). Adversarial training (Goodfellow et al., 2014; Ganin et al., 2016) is a popular method used to protect a number of potentially sensitive attributes, and it has been explored for various text classification tasks (Wang et al., 2019; Li et al., 2018; Barrett et al., 2019). However, such methods are practical when a relatively small number of features are considered, such as gender, age, and race (Elazar and Goldberg, 2018). In this paper, we use autoencoders (Ranzato et al., 2007) to produce an encoded representation of our one-hot encoded demographic vectors. Autoencoders have been successfully employed to obscure time-series sensitive sensor data (Malekzadeh et al., 2017), images (Krishna et al., 2021), and videos (D’Souza et al., 2020), with applications ranging from privacy-preserving facial expression recognition (Zeng et al., 2018) to various types of image classification tasks in the medical domain (Phan et al., 2016; D’Souza et al., 2020).

**Models.** Our model is composed of two separately trained modules (Figure 6): (1) a demographic autoencoder, and (2) the main UPV classifier, which takes the encoded demographic vector produced in (1) and combines it with textual features.

*Demographic Autoencoder.* First, we train an autoencoder to reconstruct the original one-hot encoded demographic vectors from their latent representations. Specifically, we jointly train an encoder  $f_\theta$  which maps an input demographic vector  $d$  to its latent representation  $e = f_\theta(d)$ . A decoder  $g_{\theta'}$  then aims to reconstruct  $d$  from  $e$ , as  $d' = g_{\theta'}(e)$ .  $f_\theta$  (resp.  $g_{\theta'}$ ) takes the form

	Hyperparameter	Value
<i>UPV classification</i>	Learning rate	$5e^{-5}$
	Optimizer	Adam
	Epochs	25
	Batch size	64
	Dropout probability	0.2
	Max. sequence length	54
	Early stopping patience	7
<i>Autoencoder</i>	Learning rate	0.005
	Optimizer	Adam
	Epochs	25
	Batch size	32
	Encoder dimensionality	32

Table 1: Hyperparameters used for training our models.

Model	Trainable parameters	Avg runtime per epoch	$F_1$ on dev
<i>Baseline</i>	66,363,649	675 s	72.99
<i>Demographic</i>	67,584,257	587 s	74.40
<i>Encoded demographic</i>	67,713,665	681 s	74.06
<i>Autoencoder</i>	16,517	72 s	-

Table 2: Overview of the number of trainable parameters, the average runtime per epoch and the development set F1-score for each model.

of a sequence of four dense layers with ReLU activations of a decreasing (resp. increasing) number of units. Therefore,  $(d, d') \in \mathbb{R}^n$  and  $e \in \mathbb{R}^m$ , with  $n > m$ .

*UPV Classifier.* We use the same classifier architecture described in Section 4.2, with the difference that it receives the compressed encoded latent representation  $e$  instead of  $d$  (Figure 6).

## 5. Experimental Setting

**Data Splitting and Samples generation.** We use a 60-15-25 train-dev-test split. We exclude UPVs that occur less than 10 times in DR-QI, leaving a total of 43 labels. We obtain positive samples by pairing up sentences from DR-QI with each gold label and generate synthetic negative samples (20 for training, 43 for evaluation) by pairing them with random labels which were not selected by the annotators. Following (Conforti et al., 2020), we encourage the model’s generalisation by deforming input sentences with random word substitution, deletion, and insertion using EDA (Wei and Zou, 2019). We fine-tune the entire DistilBERT network during training.

**Evaluation Framework.** Following Conforti et al. (2020) we report precision, recall and F1 score. In the test set, we include all negative samples for each sentence: this is comparable to the real-world situation in which human annotators select from 43 UPV labels.

**Computing Infrastructure.** Models were implemented in TensorFlow (Abadi et al., 2015) and

	Precision	Recall	F1
<i>Baseline</i>	65.43	81.84	69.12
<i>Demographic</i>	67.53	<b>83.41</b>	70.74
<i>Encoded Dem.</i>	<b>69.33</b>	82.52	<b>72.01</b>

Table 3: Model performances for the baseline model (Section 4.1), demographic model (Section 4.2) and the encoded demographic model (Section 4.3).

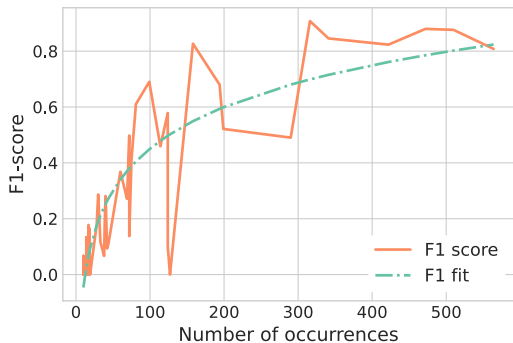


Figure 7: Per-UPV  $F_1$  score (encoded demographic model), against its support in DR-QI.

Keras (Chollet and others, 2015), using Huggingface (Wolf et al., 2020) to load DistilBert.

In Table 1 we provide an overview of the hyperparameters used to train our models. We further report the number of trainable parameters, the training time per epoch and the F1-score on the development set for each model in Table 2.

**Ethical Considerations.** Training transformers requires large amounts of energy and, depending on the energy source, can emit large quantities of  $CO_2$  (Strubell et al., 2019; Henderson et al., 2020). To minimise the environmental impact of this research, we used a pre-trained model, more specifically DistilBERT (Sanh et al., 2019). This considerably limited the training time: models took  $\sim 2$  hours to train on a Microsoft Azure GPU. Moreover, the DistilBERT model might encode biases due to the data it was trained on (Bender et al., 2021). We note, however, that this element is beyond the scope of this research.

## 6. Experiments and Discussion

**Model performance.** Results in Table 3 show that including demographic information about the speaker is beneficial for UPV classification, with consistent gains in performance over all considered metrics. Interestingly, we do not observe a strong performance drop when encoding the demographic representations: a marginal decline in recall (-0.9) is compensated by an increase in precision (+1.8), resulting in an improved  $F_1$  score (+1.3). This seems to indicate that using autoencoders to produce encoded demographic representations constitutes a valid recipe to protect speakers’ identity, without sacrificing on classification per-

Excluded feature	Precision	Recall	F1	$\Delta F1$
<i>All demographics</i>	67.53	83.41	70.74	–
– <i>Age</i>	67.74	84.55	71.78	1.04
– <i>Gender</i>	64.58	84.94	69.81	-0.93
– <i>Marital status</i>	66.38	81.24	69.40	-1.34
– <i>Disability</i>	64.34	83.90	68.36	-2.38
– <i>Education</i>	65.23	82.83	69.70	-1.04
– <i>Occupation</i>	64.12	83.28	68.19	-2.55
– <i>Household size</i>	65.27	84.63	69.64	-1.10
– <i>Children</i>	66.57	79.43	68.21	-2.53
– <i>Income</i>	63.08	81.91	66.93	-3.81
– <i>Poverty status</i>	63.44	81.79	67.11	-3.63

Table 4: Ablation study. Model performance by removing one of the ten considered features at a time.

formance. Moving to single-label classification, we report a correlation between a label’s support in the training data and the model’s performance (Figure 7). Particularly satisfactory results are obtained on *faith*, *clean air*, and *entertainment*. The upward spikes may be attributed to the presence of lexical clues such as *God*, *air* and *recreation* respectively. In turn, the large drop in performance for *unburden* (downward spike) could possibly be explained by the fact that this UPV is usually expressed very implicitly.

**Ablation Experiments and Discussion.** The models described above are provided with the entire range of demographic attributes discussed in Section 3.4. Separating the influence of different demographic attributes is complicated, as social identities intersect and overlap in complex and unique ways (Solomon et al., 2018; Jiang and Fellbaum, 2020). As a proxy for their relative importance, we investigate the impact of each demographic feature with a set of ablation experiments (Table 4). Results show that excluding the features related to economic status seems to have the highest impact on overall performance (*income level* -3.81, *poverty status* -3.63, *occupation* -2.55 in  $F_1$  score). The importance of such features for UPV classification on our data is unsurprising: in both India and Uganda, wealth status often correlates with people’s living conditions and thus access to basic services. This influences the value perception of dwellers, which are in turn, expressed in the UPVs (Hirmer and Guthrie, 2016).

The other two features that seem to contribute substantially to UPV classification are *number of children* and *disability* (resp. -2.53 and -2.38  $F_1$ ). Both elements can play a determinant role in shaping a person’s worldview and defining their priorities, and their influence might be even more pronounced in LMICs. In both case countries, disability is likely to hold social stigma and can result in an inability to actively partake in society (Abimanyi-Ochom and Mannan, 2014), partially due to a lack of access to services but also due to social segregation (Kumar et al., 2012; Mac-Seing et al., 2020). In contrast, parenthood commonly correlates with social status and community respect: in Uganda,

for example, having children is seen as societal and cultural obligation (Nattabi et al., 2012). Moreover, having a high number of children (as is often the case for those interviewed for DR-QI) can strongly affect poverty status, especially in countries where education is expensive. For example, primary and secondary universal education in Uganda is available, but it is not entirely free and the standards are often very poor (Tumwesige, 2020).

**Impact of Gender and Age for UPV Classification.** Ablation experiments for age and gender deserve a separate discussion as they represent ‘classic’ demographic attributes commonly considered in NLP research (Hovy, 2015; Hovy and Søgaard, 2015; Elazar and Goldberg, 2018; Garimella et al., 2019). Beyond NLP, age and gender are consistently included in almost all demographic surveys (Hughes et al., 2016) and are often collected just because they constitute an “easy question to ask” (Larson, 2017). Our results, however, show that while *gender* has a relatively minor impact on overall performance (-0.93), removing *age* has a *positive* impact on performance (+1.04).

**Implications Beyond UPV Classification.** The results reported above highlight that including often neglected demographic features into modelling can play a considerably important role for text classification. Despite the longstanding interest of sociolinguists in the correlation between socio-economic class and language usage (Labov, 1972), such characteristics are rarely considered in NLP tasks. Our results show that these features can be promising to consider.

Intuitively, different demographic features describe different aspects of a speaker’s identity, which needs to be considered in its interdependent, composite nature (Lepori, 2020; Kumar B et al., 2021). This represents a very interesting outcome, which suggests that broadening the range of considered demographic features, motivated by an attentive analysis of the community under study, might constitute a very promising research direction for other NLP tasks and datasets.

**Implications for Sustainable Project Implementation.** Understanding which demographic features – beyond the ‘classic’ features such as *gender* and *age* – also shows which community groupings might be most beneficial when it comes to practical project implementation, which is important for achieving sustainable project development in LMICs. Project design and communication strategies can be better aligned with the needs of alternative community clusters and more significant links between values portrayed/communicated and the predominant demographic (Hirmer et al., 2021b). This in turn may result in improved project buy-in (Figure 1).

## 7. Conclusions

This paper studied the impact of a wide range of demographic features on automatic UPV classification whereby sentences are annotated with the values they

contain. We collect and release DR-QI, an annotated dataset of interviews collected in two LMICs, which constitutes an invaluable resource for studying the interaction of textual and demographic features. Our experiments show that performance increases when demographic attributes are included, even when they are obscured using autoencoders; the relative simplicity of our model, which takes as inputs an interview’s sentence and one-hot encoded demographic features, leaves a lot of room for future research. Moreover, our results show that the relative importance of popular demographic features, such as *gender* and *age*, is negligible when compared to features such as *income* or *occupation*, thus opening up new interesting research directions for text classification.

## 8. Acknowledgements

We thank the anonymous reviewers for their effort in reviewing this paper, their constructive feedback and suggestions. We are particularly grateful to the Rural Senses team for helping to define this research and for their continuous guidance. Finally, we would like to thank the Efficiency for Access Coalition, in particular Makena Ileri and Richa Goyal, for access to the data collected as part of the project to map the impact of off- and weak-grid energy appliances.

## 9. Bibliographical References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
- Abimanyi-Ochom, J. and Mannan, H. (2014). Uganda’s disability journey: progress and challenges: community paper. *African Journal of Disability*, 3(1):1–6.
- Barrett, M., Kementchedjhieva, Y., Elazar, Y., Elliott, D., and Søgaard, A. (2019). Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6330–6335.
- Bbaale, E. and Mpuga, P. (2011). Female education, contraceptive use, and fertility: evidence from Uganda. *Consilience*, (6):20–47.
- Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science.



- Transactions of the Association for Computational Linguistics*, 6:587–604.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Madeleine Clare Elish, et al., editors, *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.
- Benton, A., Coppersmith, G., and Dredze, M. (2017). Ethical Research Protocols for Social Media Health Research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, EthNLP@EACL, Valencia, Spain, April 4, 2017*, pages 94–102. Association for Computational Linguistics.
- Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.
- Blodgett, S. L., Green, L., and O’Connor, B. (2016). Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas, November. Association for Computational Linguistics.
- Buchanan, I. (2018). *A dictionary of critical theory*. Oxford University Press.
- Caselli, T., Cibin, R., Conforti, C., Ecinas, E., and Teli, M. (2021). Guiding Principles for Participatory Design-inspired Natural Language Processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35. Association for Computational Linguistics.
- Chollet, F. et al. (2015). Keras. <https://keras.io>.
- Christiaensen, L. J. and Subbarao, K. (2005). Towards an understanding of household vulnerability in rural Kenya. *Journal of african economies*, 14(4):520–558.
- Coleman, H. (2010). The English language in development. *British council*, pages 1–24.
- Conforti, C., Hirmer, S., Morgan, D., Basaldella, M., and Ben Or, Y. (2020). Natural language processing for achieving sustainable development: the case of neural labelling to enhance community profiling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8427–8444, Online, November. Association for Computational Linguistics.
- Coulter, R. H. and Zaltman, G. (1994). Using the zaltman metaphor elicitation technique to understand brand images. *ACR North American Advances*.
- De Cenival, M. (2008). Ethics of research: the freedom to withdraw. *Bulletin de la Societe de pathologie exotique (1990)*, 101(2):98–101.
- D’Souza, M., Van Munster, C. E., Dorn, J. F., Dorier, A., Kamm, C. P., Steinheimer, S., Dahlke, F., Uitdenhaag, B. M., Kappos, L., Johnson, M., et al. (2020). Autoencoder as a new method for maintaining data privacy while analyzing videos of patients with motor dysfunction: Proof-of-concept study. *Journal of medical internet research*, 22(5):e16669.
- Elazar, Y. and Goldberg, Y. (2018). Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*.
- EnDev. (2020). Productive Use of Energy : Moving to scalable business cases. Technical report, Deutsche Gesellschaft fuer Internationale Zusammenarbeit (GIZ) GmbH Registered, Eschborn, Germany.
- Extra, G. and Gorter, D. (2001). *The other languages of Europe: Demographic, sociolinguistic, and educational perspectives*, volume 118. Multilingual Matters.
- Oluwaseyi Feyisetan, et al., editors. (2020). *Proceedings of the Second Workshop on Privacy in NLP*, Online, November. Association for Computational Linguistics.
- Oluwaseyi Feyisetan, et al., editors. (2021). *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, Online, June. Association for Computational Linguistics.
- Frey, B. B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. Sage Publications.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.
- Garimella, A., Banea, C., Hovy, D., and Mihalcea, R. (2019). Women’s syntactic resilience and men’s grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy, July. Association for Computational Linguistics.
- Geva, M., Goldberg, Y., and Berant, J. (2019). Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In Kentaro Inui, et al., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1161–1166. Association for Computational Linguistics.
- Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial*

- intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings.
- Gomes, C., Dietterich, T., Barrett, C., Conrad, J., Dilkina, B., Ermon, S., Fang, F., Farnsworth, A., Fern, A., Fern, X., et al. (2019). Computational sustainability: Computing for a better world and a sustainable future. *Communications of the ACM*, 62(9):56–65.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Gregory D Hager, et al., editors. (2017). *AAAI Symposium on AI for Social Good*, Stanford University, CA, United States, March.
- Hazen, K. (2002). Identity and language variation in a rural community. *Language*, pages 240–257.
- Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020). Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. *CoRR*, abs/2002.05651.
- Herbelot, A., Von Redecker, E., and Müller, J. (2012). Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 45–54.
- Hirmer, S. and Guthrie, P. (2016). Identifying the needs of communities in rural Uganda: A method for determining the ‘User-Perceived Value’ of rural electrification initiatives. *Renewable and Sustainable Energy Reviews*, 66:476–486.
- Hirmer, S., Leonard, A., Tumwesige, J., and Conforti, C. (2021a). Building representative corpora from illiterate communities: A review of challenges and mitigation strategies for developing countries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2176–2189, Online, April. Association for Computational Linguistics.
- Hirmer, S., Mazzone, A., Leonard, A., and Conforti, C. (2021b). The Power of Language: Exploring values, empowerment dynamics and communication strategies for gender-inclusive energy service design in rural Uganda. *Energy Research and Social Sciences*, 101(2):98–101.
- Hirmer, S. (2018). *Improving the Sustainability of Rural Electrification Schemes: Capturing Value for Rural Communities in Uganda*. Ph.D. thesis, University of Cambridge.
- Hovy, D. and Søgaard, A. (2015). Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China, July. Association for Computational Linguistics.
- Hovy, D. and Spruit, S. L. (2016). The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7–12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.
- Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China, July. Association for Computational Linguistics.
- Howells, M. I., Jonsson, S., Käck, E., Lloyd, P., Bennett, K., Leiman, T., and Conradie, B. (2010). Calabashes for kilowatt-hours: Rural energy and market failure. *Energy Policy*, 38(6):2729–2738.
- Hughes, J. L., Camden, A. A., and Yangchen, T. (2016). Rethinking and updating demographic questions: Guidance to improve descriptions of research samples. *Psi Chi Journal of Psychological Research*, 21(3):138–151.
- Jayachandran, S. (2020). Social norms as a barrier to women’s employment in developing countries. Technical report, National Bureau of Economic Research.
- Jiang, M. and Fellbaum, C. (2020). Interdependencies of gender and race in contextualized word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 17–25, Barcelona, Spain (Online), December. Association for Computational Linguistics.
- Krishna, S., Gupta, R., and Dupuy, C. (2021). ADePT: Auto-encoder based Differentially Private Text Transformation. *arXiv preprint arXiv:2102.01502*.
- Kumar B, S., Chandrabose, A., and Chakravarthi, B. R. (2021). An overview of fairness in data – illuminating the bias in data pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 34–45, Kyiv, April. Association for Computational Linguistics.
- Kumar, S. G., Roy, G., and Kar, S. S. (2012). Disability and rehabilitation services in India: Issues and challenges. *Journal of Family Medicine and Primary Care*, 1(1):69.
- Labov, W. (1972). *Sociolinguistic patterns*. Number 4. University of Pennsylvania press.
- Larson, B. (2017). Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain, April. Association for Computational Linguistics.
- Lepori, M. (2020). Unequal representations: Analyzing intersectional biases in word embeddings using representational similarity analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1720–1728, Barcelona,

- Spain (Online), December. International Committee on Computational Linguistics.
- Li, Y., Baldwin, T., and Cohn, T. (2018). Towards robust and privacy-preserving text representations. *arXiv preprint arXiv:1805.06093*.
- Mac-Seing, M., Zinszer, K., Eryong, B., Ajok, E., Ferlatte, O., and Zarowsky, C. (2020). The intersectional jeopardy of disability, gender and sexual and reproductive health: experiences and recommendations of women and men with disabilities in Northern Uganda. *Sexual and reproductive health matters*, 28(2):1772654.
- Malekzadeh, M., Clegg, R. G., and Haddadi, H. (2017). Replacement autoencoder: A privacy-preserving algorithm for sensory data analysis. *arXiv preprint arXiv:1710.06564*.
- Matarazzo, J. D., Weitman, M., Saslow, G., and Wiens, A. N. (1963). Interviewer influence on durations of interviewee speech. *Journal of Verbal Learning and Verbal Behavior*, 1(6):451–458.
- McCall, L. (2005). The complexity of intersectionality. *Signs: Journal of women in culture and society*, 30(3):1771–1800.
- Mohamed, S., Png, M.-T., and Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4):659–684.
- Nalwadda, G., Mirembe, F., Byamugisha, J., and Faxelid, E. (2010). Persistent high fertility in Uganda: young people recount obstacles and enabling factors to use of contraceptives. *BMC public health*, 10(1):1–13.
- Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE.
- Nattabi, B., Li, J., Thompson, S. C., Orach, C. G., and Earnest, J. (2012). Between a rock and a hard place: stigma and the desire to have children among people living with HIV in northern Uganda. *Journal of the International AIDS Society*, 15(2):17421.
- Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohunge, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Seling, M., Okegbemi, L., Martinus, L., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Danganana, I. A., Kamper, H., Elshahar, H., Duru, G., Kioko, G., Espoir, M., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C. C., Dossou, B. F. P., Sibanda, B., Bassey, B., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November. Association for Computational Linguistics.
- Oliveira, I. T. d., Dias, J. G., and Padmadas, S. S. (2014). Dominance of sterilization and alternative choices of contraception in India: an appraisal of the socioeconomic impact. *PLoS One*, 9(1):e86654.
- Phan, N., Wang, Y., Wu, X., and Dou, D. (2016). Differential privacy preservation for deep autoencoders: an application of human behavior prediction. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- Ranzato, M., Huang, F. J., Boureau, Y.-L., and LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.
- Reynolds, T. J. and Gutman, J. (1988). Laddering theory, method, analysis, and interpretation. *Journal of advertising research*, 28(1):11–31.
- Roser, M. and Ortiz-Ospina, E. (2013). Global Extreme Poverty. *Our World in Data*. <https://ourworldindata.org/extreme-poverty>.
- Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.
- Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. In Anna Korhonen, et al., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1668–1678. Association for Computational Linguistics.
- Sawhney, R., Agarwal, S., Wadhwa, A., and Shah, R. R. (2020). Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426, Online, November. Association for Computational Linguistics.
- Schröder, C., Bürgl, K., Annanias, Y., Niekler, A., Müller, L., Wiegrefe, D., Bender, C., Mengs, C., Scheuermann, G., and Heyer, G. (2021). Supporting Land Reuse of Former Open Pit Mining Sites using Text Classification and Active Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4141–4152.
- Scott, J. W. (1986). Gender: a useful category of historical analysis. *The American historical review*, 91(5):1053–1075.
- Shi, Z. R., Wang, C., and Fang, F. (2020). Artificial

- intelligence for social good: A survey. *Computing Research Repository*, arXiv:2001.01818.
- Solomon, A., Moon, D., Roberts, A. L., and Gilbert, J. E. (2018). Not Just Black and Not Just a Woman: Black Women Belonging in Computing. In *2018 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*, pages 1–5.
- Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Anna Korhonen, et al., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3645–3650. Association for Computational Linguistics.
- Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.
- Tockar, A. (2014). Riding with the stars: Passenger privacy in the nyc taxicab dataset. *Neustar Research*, September, 15.
- Tumwesige, J. (2020). COVID-19 Educational Disruption and Response: Rethinking e-Learning in Uganda. *University of Cambridge*.
- Van Hiel, A., Van Assche, J., De Cremer, D., Onraet, E., Bostyn, D., Haesevoets, T., and Roets, A. (2018). Can education change the world? education amplifies differences in liberalization values and innovation between developed and developing countries. *PloS one*, 13(6):e0199560.
- Vijayaraghavan, P., Vosoughi, S., and Roy, D. (2017). Twitter demographic classification using deep multimodal multi-task learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 478–483, Vancouver, Canada, July. Association for Computational Linguistics.
- Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., and Nerini, F. F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1):1–10.
- Wang, T., Zhao, J., Yatskar, M., Chang, K., and Ordonez, V. (2019). Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5309–5318. IEEE.
- Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In David Bamman, et al., editors, *Proceedings of the First Workshop on NLP and Computational Social Science, NLP+CSS@EMNLP 2016, Austin, TX, USA, November 5, 2016*, pages 138–142. Association for Computational Linguistics.
- Wei, J. and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November. Association for Computational Linguistics.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.
- Zaltman, G. and Coulter, R. H. (1995). Seeing the voice of the customer: Metaphor-based advertising research. *Journal of advertising research*, 35(4):35–51.
- Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., and Dobaie, A. M. (2018). Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, 273:643–649.

## A. Datasheet for the DR-QI dataset

**Data set name:** Demographic-Rich Qualitative UPV-Interviews Dataset (DR-QI)

**Data set developers:** Josephine Tumwesige, Yau Ben-Or, Yair Perry, Vishal Narula, Stephanie Hirmer, Costanza Conforti, Goya van Boven

**Data statement authors:** Stephanie Hirmer, Costanza Conforti, Goya van Boven

**Contacts:** info@ruralsenses.com

### A. CURATION RATIONALE

The DR-QI dataset contains 5,333 sentences. The sentences constitute excerpts from 214 interviews, conducted in 56 rural villages in India and 68 villages in Uganda. The exact names of the villages in which interviews were collected are not released in order to protect the identity of participants. The interviews were part of a larger project to develop an impact framework for off-grid energy appliances in Low- and Middle-Income countries (LMICs).

The sentences in the DR-QI dataset were annotated with the User-Perceived Value (UPV) approach. The dataset contains additional demographic information for each speaker. The dataset releases the 5,333 translated sentences, along with their UPV annotations. For each speaker, ten self-reported categorical demographic features are included.



## B. LANGUAGE VARIETY AND TRANSLATOR DEMOGRAPHIC

The sentences in the DR-QI dataset are in English. Sentences were translated from original transcriptions in 7 local languages.

To ensure translators could take area and tribe specific linguistic nuances into account, all translators were from the regions in which interviews were carried out.

## C. SPEAKER DEMOGRAPHIC

Details on the speakers' demographics are reported in Table 5. For each speaker the following self-reported demographic features are included: age, gender, marital status, disability (yes/no), education level, occupation, household size, number of children, income relative to the participants group from the same country, and absolute poverty status. All demographic features are categorical.

Total speakers	214
Gender	Equal split women/men
Nationality	Mostly Ugandan and Indian
Age	Between 18 and 77.
Socio Economic Status	Variable: political and religious leaders excluded, as well as close family members.

Table 5: Speakers Demographic in the DR-QI dataset.

## D. ANNOTATOR DEMOGRAPHIC

The 8 annotators all lived in LMICs and were familiar with the local context of the interviews. Prior to starting the annotation process, annotators received a training workshop and took part in a short quiz. Annotations were carried out on a user-friendly platform. Annotators received a competitive salary that matched the local context.

## E. SPEECH SITUATION

All interviews were collected between 2020 and 2021. The interviews were conducted in locations familiar to the interviewees, mostly open air. The interviews were conducted both individually and in groups of 6 people following standard focus group methods. To avoid direct inquiry, the interviews were conducted by means of the *UPV game*, which is described in detail in (Hirmer and Guthrie, 2016), resulting in semi-structured interactions. Even if the speech situation can be characterized as a dialogue, the interviewers were instructed to talk as little as possible, in order to avoid external influence, and to elicit answers to simple interactions such as *why probing*. As a consequence, many of the interviews may be better characterized as monologues.

## F. TEXT CHARACTERISTICS

All translators were bilingual, but English wasn't the native language of any of them. As a consequence,

the DR-QI dataset contains some grammatical errors. Moreover, the DR-QI dataset contains many examples of oral constructions. Proper nouns (people, companies, locations, ...) in the DR-QI dataset are anonymized with a special tag.

## G. RECORDING QUALITY

Original recordings aren't released.

## B. UPV Values Definitions

Figure 8 shows the UPV wheel and Table 6 reports the complete lists of UPV labels with their definitions, as they were provided to annotators.

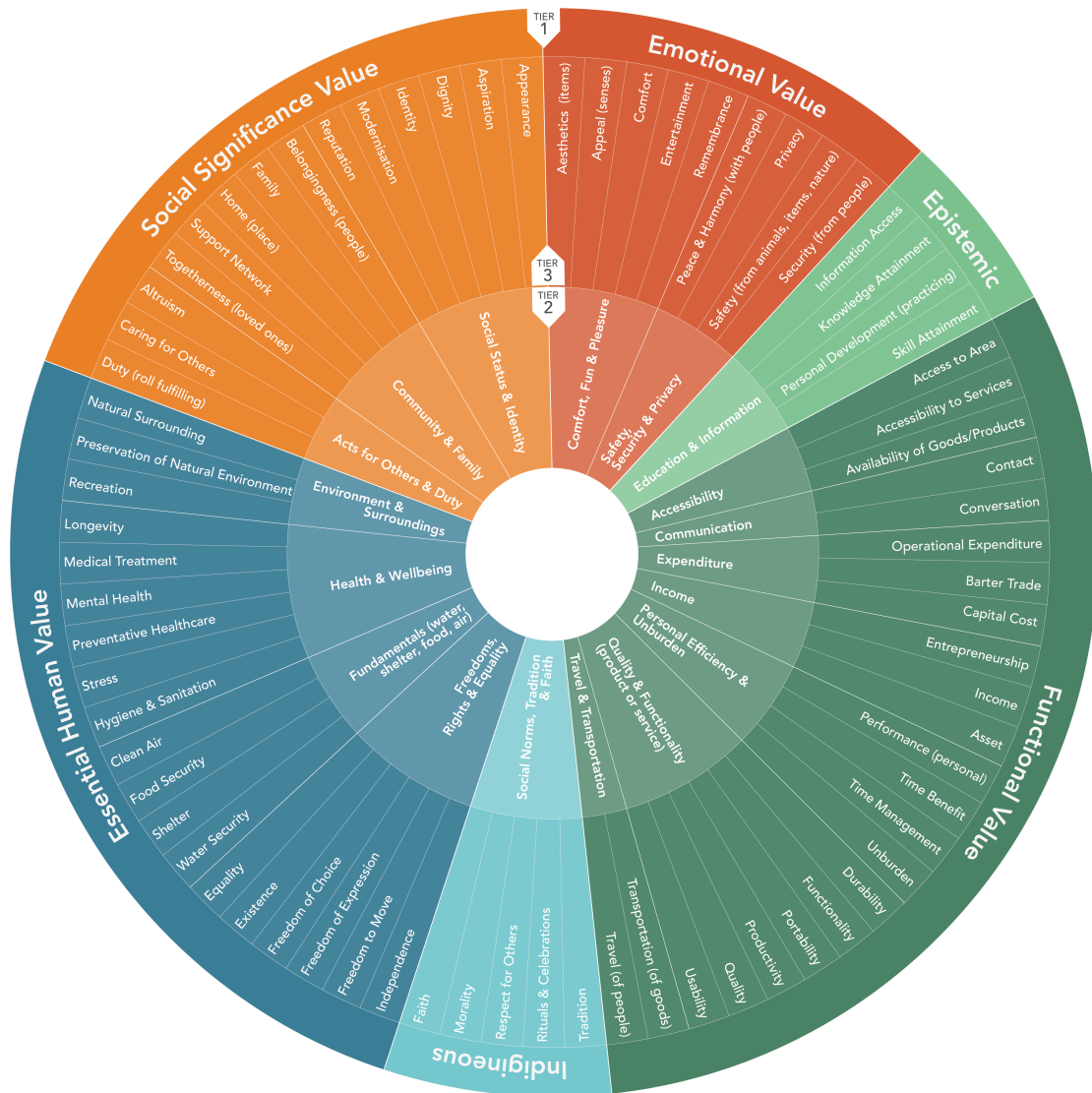


Figure 8: User-Perceived Value (UPV) wheel showing three levels of value categorisation. UPVs are in the middle ring, while tier 2 (inner ring) and tier 1 (outer ring) are higher levels of categorisation.

Tier 2 label	Tier 3 label	Definition
Accessibility	Access to area Availability of goods/products Accessibility to services	Having uninterrupted access between local areas. Being able to get, find, or buy something in the area (good or service). Having continuous access to services.
Travel & Transportation Communication	Travel (of people) Transportation (of goods) Contact Conversation	Being able to move oneself and others from one place to another. Being able to transport something between locations. The ability to interact or communicate with others near or far. The ability to be understood when conversing with others.
Safety, Security & Privacy	Safety (from animals, items, nature) Privacy Peace and Harmony Security (from people)	Being able to prevent harm from accidents resulting from environmental factors. Being able to remain unobserved or disturbed by other people. Being on good terms with others (e.g. community members, persons) Being free from danger, violence and threat posed by other people (e.g. theft, attacks)
Education & Information	Knowledge attainment Information access Skill attainment Personal development (practicing)	Learning or being taught new knowledge. Being informed of current events (locally and globally). Learning or being taught a new skill. Activities that improve living conditions and personal wellbeing.
Fundamentals (Water, Shelter, Food, Air)	Food security Shelter Clean Air Water security	A reliable and continuous supply of a diverse variety of foods. A place giving protection from weather or danger. Ability to breath clean air. The ability to have continuous access to clean water.
Health & Wellbeing	Longevity Medical treatment Preventative healthcare Hygiene and Sanitation Stress Mental health	The ability to extend one's lifespan. The ability to receive medical attention as a consequence of illness or injury. Something that prevents illness or injury. Health-related cleanliness and treatment of waste Something that is mentally or emotionally straining. Ability to achieve mental, social, and emotional steadiness/wellbeing.
Expenditure	Capital cost Operational expenditure	Fixed one-time expenditure incurred through the purchase of an item or service. The cost incurred during the operation of an item or service.
Income	Asset Barter trade Entrepreneurship  Income	Ownership of a commodity with potential economic future returns. The ability to exchange goods or services. The ability or desire to start a business that is beyond the normal occupation in the particular area. The ability to make money from the sale of a good or service.
Personal Efficiency & Unburden	Time benefit Time management Performance (personal) Unburden	The ability to accomplish a task with the minimum expenditure of time. The ability to work or plan towards a schedule. Productivity which someone achieves in his or her work. The ability to simplify tasks, to make something easier.
Social norms, Tradition and Faith	Celebration // Rite Respect for others Morality Tradition Faith	The activities associated with important events (ceremonial practice). Showing or feeling esteem or honor for someone or something. Following accepted rules and conduct. Expected customs embedded into the specific community culture. Belief in a god, or in the doctrines or teachings of religion or spiritualism.
Comfort, Fun & Pleasure	Appeal (senses) Aesthetics (items) Comfort Entertainment Remembrance	Something that is pleasing to the senses (taste and smell). Something or someone that is pleasing to look at. Being comfortable, leading to a positive feeling. Something that affords pleasure, diversion, or amusement. An association with a past event of emotional significance (positive or negative).
Quality & Functionality (product or service)	Durability Productivity Usability Functionality  Portability Quality	An item or service that continues reliably or endures for a long time. An item or service that improves the rate of output or lead to increased yield. An item or service that is easy to operate, handle, or maintain (physical interaction). Features of an item or service that serves a purpose well, fulfils a variety of uses and functions. An item or service that is easily carried, transported, or conveyed by hand. An item or service that exceeds others (i.e. excellent or good).
Freedoms, Rights & Equality	Freedom of choice Freedom of expression Freedom to move Independence Existence Equality	The ability to make choices without interference from individuals or states. The ability to express oneself without interference from individuals or states. Freedom to move from one area to another. The ability to be and act independent from others. To be allowed to live and be recognised as a human being. Having the same opportunities and rights as others.
Acts for others & Duty	Caring for others Duty (roll fulfilling) Altruism	Displaying kindness and concern for people close to the individual. The need to fulfill the tasks or responsibilities associated with a certain role. The principle and practice of unselfish concern for the wellbeing of others.
The environment & Surrounding	Preservation of natural environment Recreation Natural surrounding	The preservation of natural resources and animals. Having access to a recreational space or activities. Being close to and having access to nature.

Table 6: Complete lists of UPV labels with their definitions, as they were provided to annotators for the Demographic-Rich Qualitative UPV-Interviews Dataset.