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[1], *Stephanie Hirmer*[2,4], *Costanza Conforti*[3]

[1] *Department of Information and Computing sciences, Utrecht University*
[2] *Energy and Power Group, University of Oxford*
[3] *Language Technology Lab, University of Cambridge*
[4] *Rural Senses Ltd.*
j.g.vanboven@students.uu.nl

Motivation

Most research on demographic-aware text representation examines only a handful of features which are often modelled separately: while in reality identities are composite, resulting from the mutual influence of different demographic elements [2]. This study addresses this gap by **investigating text classification with a rich set of demographic features**.

Demographic Rich Qualitative UPV-Interviews (DR-QI) corpus

DR-QI (data sheet) contains extracts of qualitative interviews conducted in rural communities in India and Uganda, which are annotated for UPV classification [1]. For each speaker, 10 self-reported categorical demographic features are included.

***Fig 1**: UPV interview in Uganda*

Data Analysis and Models

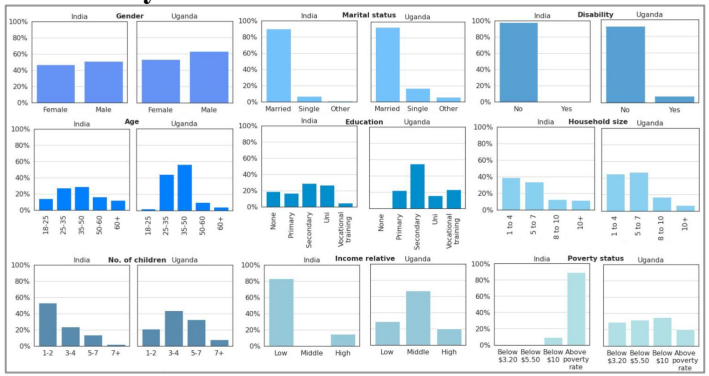

***Fig 2**: Statistical features of demographic features in the DR-QI dataset*

We investigate the effect of adding 10 one-hot encoded demographic features (Fig 2) as model input for UPV classification. We use DistilBERT as the main encoder. In order to protect the privacy of speakers, we train an autoencoder to obscure demographic information.

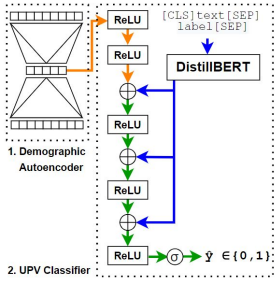

***Fig 3**:Demographic-aware models. Module 2 represents the UPV classifier. We experiment with encoding demographic vectors by a separately trained autoencoder (module 1).*

Experiments

Including demographic information benefits UPV classification (Tab 1) even if this information is encoded, suggesting that autoencoders can be useful for protecting speakers' identity. Further, an ablation study (Tab2) shows a large impact of economic features, while the popular features *age* and *gender* have little impact. This suggests that broadening the range of demographic features can be a promising research direction.

	Precision	Recall	F1
Baseline	65.43	81.84	69.12
Demographic	67.53	**83.41**	70.74
Encoded Dem.	**69.33**	82.52	**72.01**

***Tab 1**: Model performances for the baseline model (no demographic features), demographic model and the encoded demographic model*

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

***Tab 2**: Ablation study. Model performance by removing one of the ten considered features at a time*

Literature

[1] Conforti *et al* (2020). *NLP for achieving sustainable development: the case of neural labelling to enhance community profiling.* EMNLP2020
[2] McCall, L. (2005). The complexity of intersectionality. Signs: *Journal of women in culture and society*, 30(3):1771–1800.

Utrecht University · RURAL senses · UNIVERSITY OF CAMBRIDGE · Language Technology Lab · UNIVERSITY OF OXFORD · Energy and Power Group