# Natural Language Processing for Achieving Sustainable Development: the Case of Neural Labelling to Enhance Community Profiling

**Costanza Conforti**[1,2], **Stephanie Hirmer**[1,3], **David Morgan**[4], **Marco Basaldella**[2], **Yau Ben Or**[1]
[1]Rural Senses Ltd.
[2]Language Technology Lab, University of Cambridge
[3]Energy and Power Group, University of Oxford
[4]Centre for Sustainable Development, University of Cambridge
info@ruralsenses.com

## Abstract

In recent years, there has been an increasing interest in the application of Artificial Intelligence – and especially Machine Learning – to the field of Sustainable Development (SD). However, until now, NLP has not been systematically applied in this context. In this paper, we show the high potential of NLP to enhance project sustainability. In particular, we focus on the case of community profiling in developing countries, where, in contrast to the developed world, a notable data gap exists. Here, NLP could help to address the cost and time barrier of structuring qualitative data that prohibits its widespread use and associated benefits. We propose the new extreme multi-class multi-label *Automatic User-Perceived Value classification* task. We release *Stories2Insights* (S2I), an expert-annotated dataset of interviews carried out in Uganda, we provide a detailed corpus analysis, and we implement a number of strong neural baselines to address the task. Experimental results show that the problem is challenging, and leaves considerable room for future research at the intersection of NLP and SD.

## 1 Introduction

Sustainable Development (SD) is an interdisciplinary field which studies the integration and balancing of economic, environmental and social concerns to tackle the broad goal of achieving inclusive and sustainable growth (Brundtland, 1987; Keeble, 1988; Sachs, 2015). As a collective, trans-national effort toward sustainability, in 2015 the United Nations approved the *2030 Agenda* (United Nations, 2015), which identifies 17 Sustainable Development Goals (SDGs) to be reached by 2030 (Lee et al., 2016). In recent years, there has been increasing recognition of the fundamental role played by data in achieving the objectives set out in the SDGs (Griggs et al., 2013; Nilsson et al., 2016; Vinuesa et al., 2020).

In this paper, we focus on data-driven planning and delivery of projects[1] which address one or more of the SDGs in a developing country context. When dealing with developing countries, a deep understanding of project beneficiaries' needs and values (hereafter referred to as *User-Perceived Values* or UPVs, Hirmer and Guthrie (2016)) is of particular importance. This is because beneficiaries with limited financial means are especially good at assessing needs and values (Hirji, 2015). When a project fails to create value to a benefiting community, the community is less likely to care about its continued operation (Watkins et al., 2012; Chandler et al., 2013; Hirmer, 2018) and as a consequence, the chances of the project's long-term success is jeopardised (Bishop et al., 2010). Therefore, comprehensive community profiling[2] plays a key role in understanding what is important for a community and act upon it, thus ensuring a project's sustainability (van der Waldt, 2019).

Obtaining data with such characteristics requires knowledge extraction from qualitative interviews which come in the form of unstructured free text (Saggion et al., 2010; Parmar et al., 2018). This step is usually done manually by domain experts (Lundegård and Wickman, 2007), which further raises the costs. Thus, structured qualitative data is often unaffordable for project developers. As a consequence, project planning heavily relies upon sub-optimal aggregated statistical data, like household surveys (WHO, 2016) or remotely-sensed satellite imagery (Bello and Aina, 2014; Jean et al., 2016), which unfortu-

[1] Examples of projects for SD include *physical infrastructures* (as the installation of a solar mini-grid to provide light (Bhattacharyya, 2012)) or of *programmes* to change a population's behaviour (as the awareness raising campaigns against HIV transmission implemented by Avert (2019)).

[2] *Community profiling* is the detailed and holistic description of a community's needs and resources (Blackshaw, 2010).

*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8427–8444,
November 16–20, 2020. ©2020 Association for Computational Linguistics

nately is of considerable lower resolution in developing countries. Whilst these quantitative data sets are important and necessary, they are insufficient to ensure successful project design, lacking insights on UPVs that are crucial to success. In this context, the application of NLP techniques can help to make qualitative data more accessible to project developers by dramatically reducing time and costs to structure data. However, despite having been successfully applied to many other domains – ranging from biomedicine (Simpson and Demner-Fushman, 2012), to law (Kanapala et al., 2019) and finance (Loughran and McDonald, 2016) – to our knowledge, NLP has not yet been applied to the field of SD in a systematic and academically rigorous format[3].

In this paper, we make the following contributions: (1) we articulate the potential of NLP to enhance SD—at the time of writing this is the first time NLP is systematically applied to this field; (2) as a case-study at the intersection between NLP and SD, we focus on enhancing project planning in the context of a developing country, namely Uganda; (3) we propose the new task of *UPV Classification*, which consists in labeling qualitative interviews using an annotation schema developed in the field of SD; (4) we annotate and release *Stories2Insights*, a corpus of UPV-annotated interviews in English; (5) we provide a set of strong neural baselines for future reference; and (6) we show – through a detailed error analysis – that the task is challenging and important, and we hope it will raise interest from the NLP community.

## 2 Background

### 2.1 Artificial Intelligence for Sustainable Development

While NLP has not yet been applied to the field of SD, in recent years there have been notable applications of Artificial Intelligence (AI) in this area. This is testified by the rise of young research fields that seek to help meet the SDGs, as *Computational Sustainability* (Gomes et al., 2019) and *AI for Social Good* (Hager et al., 2017; Shi et al., 2020).

In this context, Machine Learning, in particular in the field of Computer Vision (De-Arteaga et al., 2018), has been applied to contexts ranging from conservation biology (Kwok, 2019), to poverty (Blumenstock et al., 2015) and slavery mapping (Foody et al., 2019), to deforestation and water quality monitoring (Holloway and Mengersen, 2018).

### 2.2 Ethics of AI for Social Good

Despite its positive impact, it is important to recognise that some AI techniques can act both as an enhancer and inhibitor of sustainability. As recently shown by Vinuesa et al. (2020), AI might inhibit meeting a considerable number of targets across the SDGs and may result in inequalities within and across countries due to application biases. Understanding the implications of AI and its related fields on SD, or Social Good more generally, is particularly important for countries where action on SDGs is being focused and where issues are most acute (UNESCO, 2019a,b).

### 2.3 Project biases

Various works highlight the importance of understanding the local context and engaging with local stakeholders, including beneficiaries, to achieve project sustainability. Where such information is not available, projects are designed and delivered based on the judgment of other actors (e.g. project funders, developers or domain experts, (Risal, 2014; Axinn, 1988; Harman and Williams, 2014)). Their judgment, in turn, is subject to biases (Kahneman, 2011) that are shaped by past experiences, beliefs, preferences and worldviews: such biases can include, for example, preferences towards a specific sector (e.g. energy or water), technology (e.g. solar, hydro) or gender-group (e.g. solutions which benefit a gender disproportionately), which are pushed without considering the local needs.

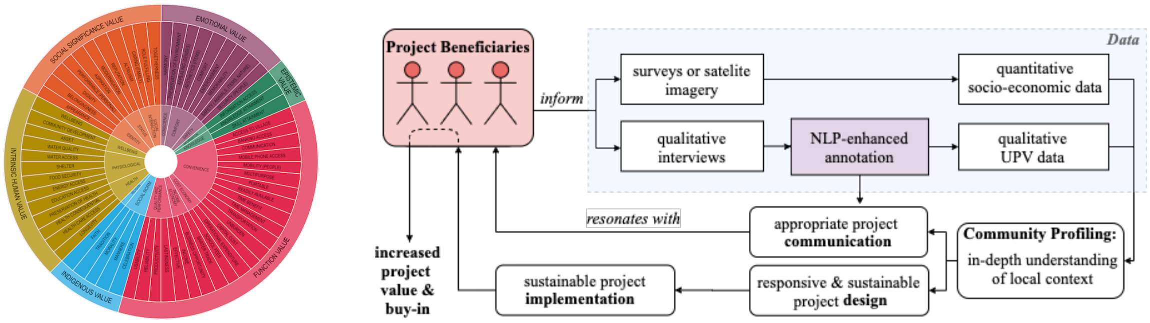
NLP has the potential to increase the availability of community-specific data to key decision makers and ensure project design is properly informed and appropriately targeted. However, careful attention needs to be paid to the potential for bias in data collection resulting from the interviewers (Bryman, 2016), as well as the potential to introduce new bias through NLP.

## 3 User-Perceived Values (UPVs) for Data-driven Sustainable Projects

### 3.1 The User-Perceived Values (UPV) Framework.

As a means to obtain qualitative data with the characteristics mentioned above, we adapt the

[3] We have found sporadic examples of the application of NLP, e.g. for analysing data from a gaming app used in a developing country (Pulse Lab Jakarta, 2016).


SOCIAL SIGNIFICANCE VALUE
EMOTIONAL VALUE
EPISTEMIC VALUE
FUNCTION VALUE
INDIGENOUS VALUE
INTRINSIC HUMAN VALUE
Project Beneficiaries
inform
Data
surveys or satelite imagery
quantitative socio-economic data
qualitative interviews
NLP-enhanced annotation
qualitative UPV data
resonates with
appropriate project communication
Community Profiling:
in-depth understanding of local context
increased project value & buy-in
sustainable project implementation
responsive & sustainable project design

(a) User-Perceived Value wheel.

(b) Flowchart of the intersection between NLP (purple square) and the delivery of SD projects.

Figure 1: Using UPVs (1a) to build sustainable projects: note the role of NLP (purple square in 1b).

User-Perceived Values (UPV) framework (Hirmer, 2018). The UPV framework builds on value theory, which is widely used in marketing and product design in the developed world (Sheth et al., 1991; Woo, 1992; Solomon, 2002; Boztepe, 2007). Value theory assumes that a deep connection exists between what consumers perceive as important and their inclinations to adopt a new product or service (Nurkka et al., 2009).

In the context of developing countries, our UPV framework identifies a set of 58 UPVs which can be used to frame the wide range of perspectives on what is of greatest concern to project beneficiaries (Hirmer and Guthrie, 2016). UPVs (or *tier 3* (T3) values) can be clustered into 17 *tier 2* (T2) value groups, each one embracing a set of similar T3 values; in turn, T2 values can be categorized into 6 *tier 1* (T1) high-level value pillars, as follows: (Hirmer and Guthrie, 2014):

1. *Emotional*: contains the T2 values *Conscience*, *Contentment*, *Human Welfare* (tot. 9 T3 values)
2. *Epistemic*: contains the T2 values *Information* and *Knowledge* (tot. 2 T3 values)
3. *Functional*: contains the T2 values *Convenience*, *Cost Economy*, *Income Economy* and *Quality and Performance* (tot. 21 T3 values)
4. *Indigenous*: containing the T2 values *Social Norm* and *Religion* (tot. 5 T3 values)
5. *Intrinsic Human*: *Health*, *Physiological* and *Quality of Life* (tot. 11 T3 values)
6. *Social significance*: contains the T2 *Identity*, *Status* and *Social Interaction* (tot. 11 T3 values)

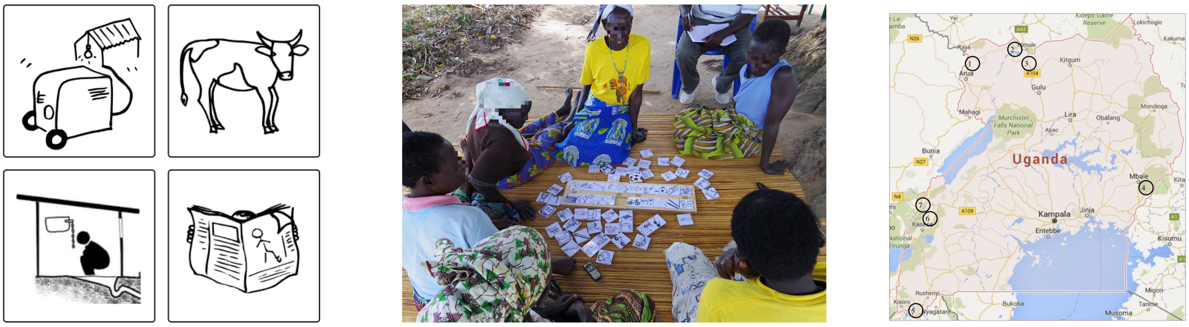
The interplay between T1, T2 and T3 values is graphically depicted in the *UPV Wheel* (Figure 1a). See Appendix A for the full set of UPV definitions.

### 3.2 Integrating UPVs into Sustainable Project Planning.

The UPV approach offers a theoretical framework to place communities at the centre of project design (Figure 1b). Notably, it allows to (a) facilitate more responsible and beneficial project planning (Gallarza and Saura, 2006); and (b) enable effective communication with rural dwellers. The latter allows the use of messaging of project benefits in a way that resonates with the beneficiaries' own understanding of benefits, as discussed by Hirji (2015). This results in a higher end-user acceptance, because the initiative is perceived to have personal value to the beneficiaries: as a consequence, community commitment will be increased, eventually enhancing the project success rate and leading to more sustainable results (Hirmer, 2018).

### 3.3 The role of NLP to enhance Sustainable Project Planning.

Data conveying the beneficiaries' perspective is seldom considered in practical application, mainly due to the fact that it comes in the form of unstructured qualitative interviews. As introduced above, data needs to be *structured* in order to be useful (OECD, 2017; UN Agenda for Sustainable Development, 2018). This makes the entire process very long and costly, thus making it almost prohibitive to afford in practice for most small-scale projects. In this context, the role of AI, and more specifically NLP, can have a yet unexplored opportunity. Implementing successful NLP systems to automatically perform the annotation process on interviews (Figure 1b, purple square), which constitutes the major bottleneck in the project planning pipeline (Section 4.1), would dramatically speed up the entire project life-cycle and drastically reduce its costs.



(a)

(b)

(c)

Figure 2: Playing the UPV game in Uganda. From left to right: 2a) Cards for the items *generator*, *cow*, *flush toilet* and *newspapers* (adapted to the Ugandan context with the support of international experts and academics from the U. of Cambridge; 2b) Women playing the UPV game in village (1)[4]; 2c) Map of case-study villages.

In this context, we introduce the task of *Automatic UPV classification*, which consists of annotating each sentence of a given input interview with the appropriate UPV labels which are (implicitly) conveyed by the interviewee.

## 4 The *Stories2Insights* Corpus: a Corpus Annotated for User-Perceived Values

To enable research in UPV classification, we release S2I, a corpus of labelled reports from 7 rural villages in Uganda (Figure 2c). In this Section, we report on the corpus collection and annotation procedures and outline the challenges this poses for NLP.

### 4.1 Building a Corpus with the UPV game

**The UPV game.** As widely recognised in marketing practice (Van Kleef et al., 2005), consumers are usually unable to articulate their own values and needs (Ulwick, 2002). This requires the use of methods that elicit what is important, such as laddering (Reynolds and Gutman, 2001) or Zaltman Metaphor Elicitation Technique (ZMET) (Coulter et al., 2001). To avoid direct inquiry (Pinegar, 2006), Hirmer and Guthrie (2016) developed an approach to identify perceived values in low-income settings by means of a game (hereafter referred to as *UPV game*). Expanding on the items proposed by Peace Child International (2005), the UPV game makes reference to 46 everyday-use items in rural areas[5], which are graphically depicted (Figure 2a). The decision to represent items graphically stems from the high level of illiteracy across developing countries (UNESCO, 2013).

Building on the techniques proposed by Coulter *et al.* (2001) and Reynolds *et al.* (2001), the UPV game is framed in the form of semi-structured interviews:
(1) participants are asked to select 20 items, based on what is most important to them (*Select stimuli*), (2) to rank them in order of importance; and finally, (3) they have to give reasons as to why an item was important to them. *Why-probing* was used to encourage discussion (*Storytelling*).

**Case-Study Villages.** 7 rural villages were studied: 3 in the West Nile Region (Northern Uganda); 1 in Mount Elgon (Eastern Uganda); 2 in the Ruwenzori Mountains (Western Uganda); and 1 in South Western Uganda. All villages are located in remote areas far from the main roads (Figure 2c). A total of 7 languages are spoken across the villages[6].

**Data Collection Setting and Guidelines for Interviewers.** For each village, 3 native speaker interviewers guided the UPV game. To ensure consistency and data quality, a two-day training workshop was held at Makerere University (Kampala, Uganda), and a local research assistant oversaw the entire data collection process in the field.

**Data Collection.** 12 people per village were interviewed, consisting of an equal split between men and women with varying backgrounds and ages. In order to gather complete insight into the underlying decision-making process – which might be influenced by the context (Barry et al., 2008) – interviews were conducted both individually and in groups of 6 people following standard focus group

[5] Such items included livestock (*cow, chicken*), basic electronic gadgets (*mobile phone, radio*), household goods (*dishes, blanket*), and horticultural items (*plough, hoe*) (Hirmer, 2018).

[5] While permission of photographing was granted from the participants, photos were pixelised to protect their identity.

[6] Rukonjo, Rukiga, Lugwere and Swahili (Bantu family); Sebei/Sabaot, Kupsabiny, Lugbara (Nilo-Saharan family).

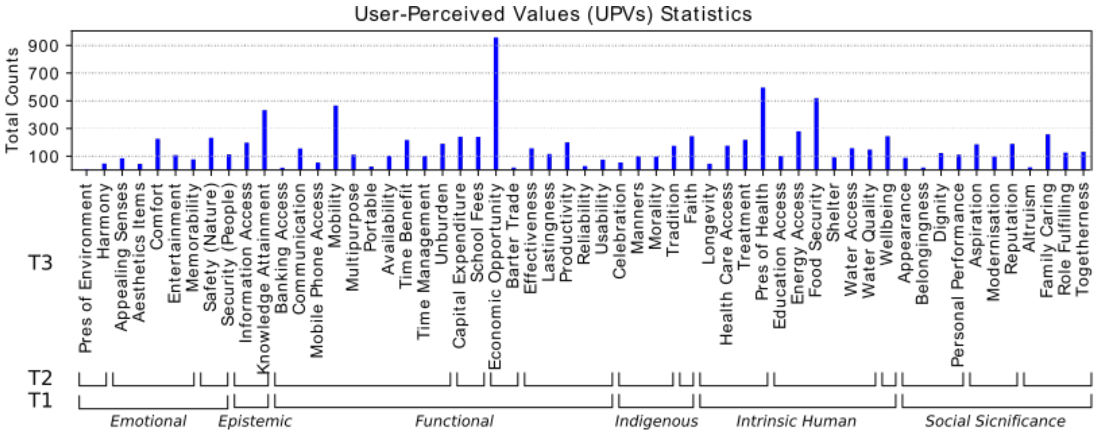


Figure 3: UPV frequencies from the S2I corpus (see Appendix A for UPV definitions).

methods (Silverman, 2013; Bryman, 2016). Each interview lasted around 90 minutes. The data collection process took place over a period of 3 months and resulted in a total of 119 interviews.

**Ethical Considerations.** Participants received compensation in the amount of 1 day of labour. An informed consent form was read out loud by the interviewer prior to the UPV game, to cater for the high-level of illiteracy amongst participants. To ensure integrity, a risk assessment following the University of Cambridge's *Policy on the Ethics of Research Involving Human Participants and Personal Data* was completed. To protect the participants' identity, locations and proper names were anonymized.

**Data Annotation.** The interviews were translated[7] into English, analysed and annotated by domain experts[8] using the computer-assisted qualitative data analysis software *HyperResearch* (Hesse-Biber et al., 1991). To ensure consistency across interviews, they were annotated following Bryman (2012), using cross-sectional indexing (Mason, 2002). Due to the considerable size of collected data, the annotation process took around 6 months.

### 4.2 Corpus Statistics and NLP Challenges

We obtain a final corpus of 5102 annotated utterances from the interviews. Samples present an average length of 20 tokens. The average number of samples per T3 label is 169.1, with an extremely skewed distribution: the most frequent T3, *Economic Opportunity*, occurs 957 times, while the least common, *Preservation of the Environment*, only 7 (Figure 3).

58.8% of the samples are associated with more than 1 UPV, and 22.3% with more than 2 UPVs (refer to Appendix B for further details on UPV correlation). Such characteristics make UPV classification highly challenging to model: the task is an extreme multi-class multi-label problem, with high class imbalancy. Imbalanced label distributions pose a challenge for many NLP applications – as sentiment analysis (Li et al., 2011), sarcasm detection (Liu et al., 2014), and NER (Tomanek and Hahn, 2009) – but are not uncommon in user-generated data (Imran et al., 2016). The following interview excerpt illustrates the multi-class multi-label characteristics of the problem:

1. *If I have a flush toilet in my house I can be a king of all kings because I can't go out on those squatting latrines* [Reputation][Aspiration]
2. *And recently I was almost rapped* (sic.) *when I escorted my son to the latrine* [Security]
3. *That [...] we have so many cases in our village of kids that fall into pit latrine* [Safety][Caring]

Further challenges for NLP are introduced by the frequent use of non-standard grammar and poor sentence structuring, which often occur in oral production (Cole et al., 1995). Moreover, manual transcription of interviews may lead to spelling errors, thus increasing OOVs. This is illustrated in the below excerpts (spelling errors are underlined):

- *Also men like phone there are so jealous for their women for example like in the morning my husband called me and asked that are you in church; so that's why they picked a phone.*

[7] Note that translating into English (or other languages commonly spoken in international workplaces, https://www.un.org/en/sections/about-un/official-languages/) is often a crucial step when applying knowledge to practical application in SD, in this case project decision-making (Bergström et al., 2012).

[8] A team of researchers from the Department of Engineering for Sustainable Development, supported by researchers in Development Studies and Linguistics, all at the University of Cambridge.

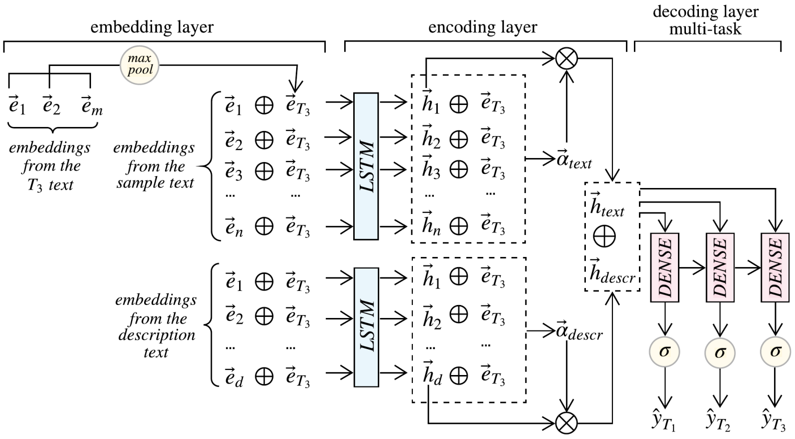

embedding layer
encoding layer
decoding layer
multi-task
max pool
$\vec{e}_1$ $\vec{e}_2$ $\vec{e}_m$
embeddings from the $T_3$ text
embeddings from the sample text
$\vec{e}_1 \oplus \vec{e}_{T_3}$
$\vec{e}_2 \oplus \vec{e}_{T_3}$
$\vec{e}_3 \oplus \vec{e}_{T_3}$
$\vec{e}_n \oplus \vec{e}_{T_3}$
LSTM
$\vec{h}_1 \oplus \vec{e}_{T_3}$
$\vec{h}_2 \oplus \vec{e}_{T_3}$
$\vec{h}_3 \oplus \vec{e}_{T_3}$
$\vec{h}_n \oplus \vec{e}_{T_3}$
$\vec{\alpha}_{text}$
embeddings from the description text
$\vec{e}_d \oplus \vec{e}_{T_3}$
$\vec{h}_d \oplus \vec{e}_{T_3}$
$\vec{\alpha}_{descr}$
$\vec{h}_{text}$
$\oplus$
$\vec{h}_{descr}$
DENSE
DENSE
DENSE
$\sigma$
$\hat{y}_{T_1}$
$\hat{y}_{T_2}$
$\hat{y}_{T_3}$

Figure 4: Multi-task neural architecture for UPV classification.

| | |
|---|---|
| *Original* | Pot keeps water safe and cold |
| *Deletion* | Pot keeps water safe and |
| *Synonym Replacement* | Pot keeps water safe and **freezing** |
| *Insertion* | Pot keeps water **hold** safe and cold |
| *Token Swap* | Pot keeps water **and safe** cold |
| *Char Swap* | Pot **kepes** water safe and cold |

Figure 5: Examples of negative samples generated through data augmentation.

- *A house keeps secrecy for example [...] I can be bitten by a snake if I had sex outside [...] you see, me I cannot because may child is looking for mangoes in the bush and finds me there, how do I explain, can you imagine!!*

## 5 User-Perceived Values Classification

As outlined above, given an input interview, the task consists in annotating each sentence with the appropriate UPV(s). The extreme multi-class multi-label quality of the task (Section 4.2) makes it impractical to tackle as a standard *multi-class classification* problem—where, given an input sample $x$, a system is trained to predict its label from a tagset $T = \{l_1, l_2, l_3\}$ as $x \rightarrow l_2$ (i.e. [0,1,0]). Instead, we model the task as a *binary classification* problem: given $x$, the system learns to predict its *relatedness* with each one of the possible labels, i.e. $(x, l_1) \rightarrow 0$, $(x, l_2) \rightarrow 1$ and $(x, l_3) \rightarrow 0$ [9].

We consider the samples from the S2I corpus as *positive instances*. Then, we generate three kinds of *negative instances* by pairing the sample text with random labels. To illustrate, consider the three T2 classes *Convenience*, *Identity* and *Status*, which contain the following T3 values:

- *Contentment*$_{T2}$ = {*Aesthetic*$_{T3}$, *Comfort*$_{T3}$, ...}
- *Identity*$_{T2}$ = {*Appearance*$_{T3}$, *Dignity*$_{T3}$...}
- *Status*$_{T2}$ = {*Aspiration*$_{T3}$, *Reputation*$_{T3}$, ...}

Moreover, *Contentment*$_{T2}$ $\in$ *Emotional*$_{T1}$ and {*Identity*$_{T2}$, *Status*$_{T2}$} $\in$ *SocialSignificance*$_{T1}$. Given a sample $x$ and its gold label *Aspiration*$_{T3}$, we can generate the following training samples:

- $(x,$ *Aspiration*$_{T3})$ is a *positive sample*;
- $(x,$ *Reputation*$_{T3})$ is a *mildly negative sample*, as $x$ is linked with a wrong T3 with the same T2;
- $(x,$ *Dignity*$_{T3})$ is *negative sample*, as $x$ is a associated with a wrong T3 from a different T2 class, but both T2 classes belong to the same T1; and
- $(x,$ *Aesthetic*$_{T3})$ is a *strictly negative sample*, as $x$ is associated with a wrong label from the another T2 class in a different T1.

In this way, during training the system is exposed to positive (real) samples and negative (randomly generated) samples.

A UPV classification system should satisfy the following desiderata: (1) it should be relatively light, given that it will be used in the context of developing countries, which may suffer from access bias[10] and (2) the goal of such a system isn't to completely replace the work of human SD experts, but rather to reduce the time needed for interview annotation. In this context, false positives are quick to notice and delete, while false negatives are more difficult to spot and correct. Moreover, when assessing a community's needs and values, missing a relevant UPV is worse than including one which wasn't originally present. For these reasons, recall is particularly important for a UPV classifier.

In the next Section, we provide a set of strong baselines for future reference.

### 5.1 Neural Models for UPV Classification

#### 5.1.1 Baseline Architecture

*Embedding Layer.* The system receives an input sample $(x, T3)$, where $x$ is the sample text $(e_1, ..., e_n)$, $T3$ is the T3 label as the sequence of its tokens $(e_1, ..., e_m)$, and $e_i$ is the word embed-

[9] Note that this is different to the classic *binary relevance* method, where a *separated* binary classifier is learned for each considered label (Read et al., 2011).

[10] With *access bias* we refer to contexts with limited computational capacity and cloud services accessibility.

ding representation of a token at position $i$. We obtain a T3 embedding $e_{T3}$ for each T3 label using a max pool operation over its word embeddings: given the short length of T3 codes, this proved to work well and it is similar to findings in relation extraction and targeted sentiment analysis (Tang et al., 2016). We replicate $e_{T3}$ $n$ times and concatenate it to the text's word embeddings $x$ (Figure 4).

*Encoding Layer.* We obtain a hidden representation $\vec{h}_{text}$ with a forward LSTM (Gers et al., 1999) over the concatenated input. We then apply attention to capture the key parts of the input text w.r.t. the given T3. In detail, given the output matrix of the LSTM layer $H = [h_1, ..., h_n]$, we produce a hidden representation $h_{text}$ as follows:

$$M = tanh(\begin{bmatrix} W_h H \\ W_v e_{upv} \otimes e_N \end{bmatrix})$$

$$\alpha_{text} = softmax(w^T M)$$

$$h_{text} = H\alpha^T$$

This is similar in principle to the attention-based LSTM by Wang et al. (2016), and proved to work better than classic attention over $H$ on our data.

*Decoding Layer.* We predict $\hat{y} \in [0, 1]$ with a dense layer followed by a sigmoidal activation.

#### 5.1.2 Including Description Information

Each T3 comes with a short description, which was written by domain experts and used during manual labelling (the complete list is in the Appendix A). We integrate information from such descriptions into our model as follows: given the ordered word embeddings from the UPV description $(e_1, ..., e_d)$, we obtain a description representation $h_{descr}$ following the same steps as for the sample text.

In line with previous studies on siamese networks (Yan et al., 2018), we observe better results when sharing the weights between the two LSTMs. We keep two separated attention layers for sample texts and descriptions. We concatenate $h_{text}$ and $h_{descr}$ and feed the obtained vector to the output layer.

#### 5.1.3 Multi-task Training

A clear hierarchy exists between T3, T2 and T1 values (Section 3). We integrate such information using multi-task learning (Caruana, 1997; Ruder, 2017). Given an input sample, we predict its relatedness not only w.r.t. a T3 label, but also with its corresponding T2 and T1 labels[11]. In practice, given the hidden representation $h = h_{text} \oplus h_{descr}$, we first feed it into a dense layer $dense_{T1}$ to obtain $h_{T1}$, and predict $\hat{y}_{T1}$ with a sigmoidal function. We then concatenate $h_{T1}$ with the previously obtained $h$, and we predict $\hat{y}_{T2}$ with a T2-specific dense layer $\sigma(dense_{T2}(h \oplus h_{T1}))$. Finally, $\hat{y}_{T3}$ is predicted as $\sigma(dense_{T3}(h \oplus h_{T2}))$.

|  | text | +att | +descr | +att+descr |
|---|---|---|---|---|
| P | 77.5 | 78.1 | **80.4** | 78.9 |
| R | 65.5 | **71.0** | 66.5 | 70.6 |
| $F_1$ | 71.0 | 74.2 | 72.8 | **74.4** |

Table 1: Results of ablation study (single-task).

In this way, the prediction $\hat{y}_i$ is based on both the original $h$ and the hidden representation computed in the previous stage of the hierarchy, $h_{i-1}$ (Figure 4).

## 6 Experiments and Discussion

### 6.1 Experimental Setting

#### 6.1.1 Data Preparation

For each positive sample, we generate 40 negative samples (we found empirically that this was the best performing ratio, see Appendix C).

Moreover, to expose the system to more diverse input, we slightly deform the sample's text when generating negative samples. Following Wei and Zou (2019), we implement 4 operations: random deletion, swap, insertion, and semantically-motivated substitution. We also implement character swapping to increase the system's robustness to spelling errors (Figure 5).

We consider only samples belonging to UPV labels with a support higher than 30 in the S2I corpus, thus rejecting 12 very rare UPVs. We select a random 80% proportion from the data as training set; out of the remaining 980 samples, we randomly select 450 as dev and use the rest as test set.

#### 6.1.2 Training Setting

In order to allow for robust handling of OOVs, typos and spelling errors in the data, we use FastText subword-informed pretrained vectors (Bojanowski et al., 2017) to initialise the word embedding matrix. We train using binary cross-entropy loss, with early stopping monitoring the development set loss with a patience of 5. Sample weighting was used to account for the different error seriousness (1 for *negative* and *strictly neg* and 0.5 for *mildly neg*).

[11] The mapping between sample and correct labels [T3, T2, T1] is as follows: *positive*: [1, 1, 1]; *slightly negative*: [0, 1, 1]; *negative*: [0, 0, 1]; *strictly negative*: [0, 0, 0].

| Label | | Multi-task train setting | | |
|---|---|---|---|---|
| | | T3 | T2+T3 | T1+T2+T3 |
| T3 | P | 78.9 | **83.5** | 79.5 |
| | R | 70.6 | 67.0 | **72.0** |
| | $F_1$ | 74.4 | 74.4 | **75.4** |
| T2 | P | – | **92.0** | 84.9 |
| | R | – | 40.5 | **62.3** |
| | $F_1$ | – | 56.2 | **71.9** |
| T1 | P | – | – | 89.8 |
| | R | – | – | 70.1 |
| | $F_1$ | – | – | 78.7 |

Table 2: Results considering all granularities and all (multi-)task training settings (T3, T2+T3, T1+T2+T3).

| T1 | T3 | P | R | $F_1$ | Support | (%) |
|---|---|---|---|---|---|---|
| *Emotional* | Harmony | 16.7 | 50.0 | 25.0 | 47 | 0.9 |
| | Appealing | 30.0 | 75.0 | 42.9 | 85 | 1.7 |
| | Aesthetics | 08.8 | 60.0 | 15.4 | 45 | 0.9 |
| | Comfort | 52.0 | 52.0 | 52.0 | 226 | 4.4 |
| | Entertainment | 40.0 | 54.5 | 46.2 | 108 | 2.1 |
| | Memorability | 16.7 | 12.5 | 14.3 | 77 | 1.5 |
| | Safety | 59.4 | 76.0 | 66.7 | 233 | 4.6 |
| | Sec. People | 46.2 | 75.0 | 57.1 | 113 | 2.2 |
| *Epist* | Info. Access | 84.6 | 55.0 | 66.7 | 198 | 3.9 |
| | Knowl. attain. | 06.2 | 09.8 | 07.5 | 433 | 8.5 |
| *Function* | Communication | 05.4 | 58.8 | 10.0 | 156 | 3.1 |
| | Mobile Acc. | 81.8 | 81.8 | 81.8 | 54 | 1.1 |
| | Mobility | 79.4 | 81.8 | 80.6 | 466 | 9.1 |
| | Multipurpose | 57.1 | 33.3 | 42.1 | 111 | 2.2 |
| | Availability | 01.4 | 33.3 | 02.6 | 104 | 2.0 |
| | Time Benefit | 51.9 | 66.7 | 58.3 | 217 | 4.3 |
| | Time Manag. | 76.9 | 83.3 | 80.0 | 102 | 2.0 |
| | Unburden | 41.9 | 72.0 | 52.9 | 190 | 3.7 |
| | Cap. Expend. | 85.0 | 53.1 | 65.4 | 241 | 4.7 |
| | School Fees | 94.4 | 73.9 | 82.9 | 240 | 4.7 |
| | Econ. Oppor. | 80.4 | 86.3 | 83.2 | 957 | 18.8 |
| | Effectiveness | 17.1 | 24.0 | 20.0 | 157 | 3.1 |
| | Lastingness | 83.3 | 38.5 | 52.6 | 116 | 2.3 |
| | Productivity | 52.4 | 66.7 | 58.7 | 200 | 3.9 |
| | Usability | 25.0 | 33.3 | 28.6 | 75 | 1.5 |
| *Indigen.* | Celebration | 100 | 50.0 | 66.7 | 55 | 1.1 |
| | Manners | 83.3 | 45.5 | 58.8 | 100 | 2.0 |
| | Morality | 20.0 | 22.2 | 21.1 | 98 | 1.9 |
| | Tradition | 85.7 | 70.6 | 77.4 | 175 | 3.4 |
| | Faith | 96.7 | 96.7 | 96.7 | 245 | 4.8 |
| *Intrinsic Human* | Longevity | 09.1 | 60.0 | 15.8 | 46 | 0.9 |
| | Healthc. Acc. | 72.2 | 76.5 | 74.3 | 176 | 3.4 |
| | Treatment | 78.3 | 85.7 | 81.8 | 218 | 4.3 |
| | Educ. Acc. | 80.0 | 54.5 | 64.9 | 103 | 2.0 |
| | Energy Acc. | 82.1 | 84.2 | 83.1 | 280 | 5.5 |
| | Food Security | 64.9 | 87.7 | 74.6 | 519 | 10.2 |
| | Shelter | 42.9 | 54.5 | 48.0 | 92 | 1.8 |
| | Water Access | 68.2 | 78.9 | 73.2 | 158 | 3.1 |
| | Water Quality | 37.0 | 90.9 | 52.6 | 148 | 2.9 |
| | Wellbeing | 09.8 | 59.1 | 16.9 | 245 | 4.8 |
| *Social Significance* | Appearance | 62.5 | 71.4 | 66.7 | 88 | 1.7 |
| | Dignity | 85.7 | 60.0 | 70.6 | 123 | 2.4 |
| | Pers. Perf. | 33.3 | 11.1 | 16.7 | 111 | 2.2 |
| | Aspiration | 56.2 | 56.2 | 56.2 | 186 | 3.6 |
| | Modernisation | 57.1 | 40.0 | 47.1 | 98 | 1.9 |
| | Reputation | 52.9 | 69.2 | 60.0 | 189 | 3.7 |
| | Fam. Caring | 63.6 | 58.3 | 60.9 | 258 | 5.1 |
| | Role Fulf. | 37.5 | 50.0 | 42.9 | 126 | 2.5 |
| | Togetherness | 53.3 | 57.1 | 55.2 | 132 | 2.6 |
| | ***Total*** | *44.9* | *70.3* | *50.5* | | |

Table 3: Single label results in the *Real-World Simulation* setting, with label support in S2I corpus.

Network hyperparameters are reported in Appendix C for replication.

## 6.2 Results and Discussion

### 6.2.1 Models Performance

During experiments, we monitor precision, recall and $F_1$ score. For evaluation, we consider a test set where negative samples appear *in the same proportion* as in the train set (1/40 positive/negative ratio). The results of our experiments are reported in Table 1. Notably, adding attention and integrating signal from descriptions to the base system lead to significant improvements in performance.

### 6.2.2 Multi-task Training

We consider the best performing model and run experiments with the three considered multi-task train settings (Section 5.1.3). We consider 3 layers of performance, corresponding to T3, T2 and T1 labels. This is useful because, in the application context, different levels of granularity can be monitored. As shown in Table 2, we observe relevant improvements in F1 scores when jointly learning more than one training objective. This holds true not only for T3 classification, but also for T2 classification when training with the T3+T2+T1 setting. This seems to indicate that the signal encoded in the additional training objectives indirectly conveys information about the label hierarchy which is indeed useful for classification.

### 6.2.3 Real-World Simulation and Error Analysis

To simulate a real scenario where we annotate a new interview with the corresponding UPVs, we perform further experiments on the test set by generating, for each sample, *all possible* negative samples. We annotate using the T1+T2+T3 model, finetuning the threshold for each UPV on the development set, and perform a detailed error analysis of the results on the test set.

As reported in Table 3, we observe a significant drop in precision, which confirms the extreme difficulty of the task in a real-world setting due to the extreme data imbalancy. Note, however, that recall remains relatively stable over changes in evaluation settings. This is particularly important for a system which is meant to enhance the annotators' speed, rather than to completely replace human experts: in this context, missing labels are more time consuming to recover than correcting false positives.

Not surprisingly, particularly good performance is often obtained on T3 labels which tend to correlate with specific terms (as *School Fees*, or *Faith*). In particular, we observe a correlation between a T3 label's support in the corpus and the system's precision in predicting that label: with very few exceptions, all labels where the system obtained a precision lower than 30 had a support similar or lower than 3%.

The analysis of the ROC curves shows that, overall, satisfactory results are obtained for all T1 labels considered (Appendix D), leaving, however, considerable room for future research.

## 7 Conclusions and Future Work

In this study, we provided a first stepping stone towards future research at the intersection of NLP and Sustainable Development (SD). As a case study, we investigated the opportunity of NLP to enhancing project sustainability through improved community profiling by providing a cost effective way towards structuring qualitative data.

This research is in line with a general call for AI towards social good, where the potential positive impact of NLP is notably missing. In this context, we proposed the new challenging task of *Automatic User-Perceived Values Classification*: we provided the task definition, an annotated dataset (the *Stories2Insights* corpus) and a set of light (in terms of overall number of parameters) neural baselines for future reference.

Future work will investigate ways to improve performance (and especially precision scores) on our data, in particular on low-support labels. Possible research direction could include more sophisticated thresholding selection techniques (Fan and Lin, 2007; Read et al., 2011) to replace the simple threshold finetuning which is currently used for simplicity. While deeper and computationally heavier models as Devlin et al. (2019) could possibly obtain notable gains in performance on our data, it is the responsibility of the NLP community – especially with regards to social good applications – to provide solutions which don't penalise countries suffering from access biases (as contexts with low access to computational power), as it is the case of many developing countries.

We hope our work will spark interest and open a constructive dialogue between the fields of NLP and SD, and result in new interesting applications.

## Acknowledgments

We thank the anonymous reviewers for their effort in reviewing this paper, their constructive feedback and suggestions. We are grateful to Dr. Nigel Collier (University of Cambridge) for providing valuable feedback on early versions of this paper. We also thank  Simon Anthony Patterson for helping to define the UPV hierarchy. Data collection was funded by QUALCOMM European Research Studentships in Technology (Grant No. 1068) and the Engineering and Physical Sciences Research Council (EPSRC) (Grant No. EP/K503009/1). Finally, we are particularly grateful to Rural Senses for helping to define this research and guide us along.

## References

Avert. 2019. Hiv prevention programming. Technical report, Avert HIV and AIDS organisation.

George H Axinn. 1988. International technical interventions in agriculture and rural development: Some basic trends, issues, and questions. *Agriculture and Human Values*, 5(1-2):6–15.

Marie-Louise Barry, Herman Steyn, and Alan Brent. 2008. Determining the most important factors for sustainable energy technology selection in africa: Application of the focus group technique. In *PICMET'08-2008 Portland International Conference on Management of Engineering & Technology*, pages 181–187. IEEE.

Olalekan Mumin Bello and Yusuf Adedoyin Aina. 2014. Satellite remote sensing as a tool in disaster management and sustainable development: towards a synergistic approach. *Procedia-Social and Behavioral Sciences*, 120:365–373.

Anna Bergström, Stefan Peterson, Sarah Namusoko, Peter Waiswa, and Lars Wallin. 2012. Knowledge translation in uganda: a qualitative study of ugandan midwives' and managers' perceived relevance of the sub-elements of the context cornerstone in the parihs framework. *Implementation Science*, 7(1):117.

Subhes C. Bhattacharyya. 2012. Energy access programmes and sustainable development: A critical review and analysis. *Energy for Sustainable Development*, 16(3):260 – 271.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

S Bishop, J Blum, Pursnani Pradeep, Bhavnani Anuradha, et al. 2010. Marketing lessons from the room to breathe campaign. *Boiling Point*, (58):2–17.

Tony Blackshaw. 2010. *Key concepts in community studies*. Sage.

Joshua Blumenstock, Gabriel Cadamuro, and Robert On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Suzan Boztepe. 2007. User value: Competing theories and models. *International journal of design*, 1(2).

Gro Harlem Brundtland. 1987. Our common future—call for action. *Environmental Conservation*, 14(4):291–294.

A Bryman. 2012. Mixed methods research; combining qualitative and quantitative research. *Social Research Methods*, pages 627–651.

Alan Bryman. 2016. *Social research methods*, 4 edition. Oxford university press.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Clare IR Chandler, James Kizito, Lilian Taaka, Christine Nabirye, Miriam Kayendeke, Deborah DiLiberto, and Sarah G Staedke. 2013. Aspirations for quality health care in uganda: How do we get there? *Human resources for health*, 11(1):13.

Ron Cole, Lynette Hirschman, Les Atlas, Mary Beckman, Alan Biermann, Marcia Bush, Mark Clements, L Cohen, Oscar Garcia, Brian Hanson, et al. 1995. The challenge of spoken language systems: Research directions for the nineties. *IEEE transactions on Speech and Audio processing*, 3(1):1–21.

Robin A Coulter, Gerald Zaltman, and Keith S Coulter. 2001. Interpreting consumer perceptions of advertising: An application of the zaltman metaphor elicitation technique. *Journal of advertising*, 30(4):1–21.

Maria De-Arteaga, William Herlands, Daniel B Neill, and Artur Dubrawski. 2018. Machine learning for the developing world. *ACM Transactions on Management Information Systems (TMIS)*, 9(2):1–14.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Rong-En Fan and Chih-Jen Lin. 2007. A study on threshold selection for multi-label classification. *Department of Computer Science, National Taiwan University*, pages 1–23.

Christiane Fellbaum. 2012. Wordnet. *The encyclopedia of applied linguistics*.

Giles M. Foody, Feng Ling, Doreen S. Boyd, Xiaodong Li, and Jessica Wardlaw. 2019. Earth observation and machine learning to meet sustainable development goal 8.7: Mapping sites associated with slavery from space. *Remote Sensing*, 11(3):266.

Martina G Gallarza and Irene Gil Saura. 2006. Value dimensions, perceived value, satisfaction and loyalty: an investigation of university students' travel behaviour. *Tourism management*, 27(3):437–452.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with lstm.

Carla Gomes, Thomas Dietterich, Christopher Barrett, Jon Conrad, Bistra Dilkina, Stefano Ermon, Fei Fang, Andrew Farnsworth, Alan Fern, Xiaoli Fern, et al. 2019. Computational sustainability: Computing for a better world and a sustainable future. *Communications of the ACM*, 62(9):56–65.

David Griggs, Mark Stafford-Smith, Owen Gaffney, Johan Rockström, Marcus C Öhman, Priya Shyamsundar, Will Steffen, Gisbert Glaser, Norichika Kanie, and Ian Noble. 2013. Policy: Sustainable development goals for people and planet. *Nature*, 495(7441):305.

Gregory D Hager, Ann Drobnis, Fei Fang, Rayid Ghani, Amy Greenwald, Terah Lyons, David C Parkes, Jason Schultz, Suchi Saria, Stephen F Smith, et al., editors. 2017. *AAAI Symposium on AI for Social Good*. Stanford University, CA, United States.

Sophie Harman and David Williams. 2014. International development in transition. *International Affairs*, 90(4):925–941.

Sharlene Hesse-Biber, Paul Dupuis, and T Scott Kinder. 1991. Hyperresearch: A computer program for the analysis of qualitative data with an emphasis on hypothesis testing and multimedia analysis. *Qualitative Sociology*, 14(4):289–306.

K Hirji. 2015. Accelerating access to energy: lessons learnt from efforts to build inclusive energy markets in developing countries. *Boil Point*, pages 2–6.

Stephanie Hirmer. 2018. *Improving the Sustainability of Rural Electrification Schemes: Capturing Value for Rural Communities in Uganda*. Ph.D. thesis, University of Cambridge, Department of Engineering.

Stephanie Hirmer and Peter Guthrie. 2014. The user-value of rural electrification: An analysis and adoption of existing models and theories. *Renewable and Sustainable Energy Reviews*, 34:145 – 154.

Stephanie Hirmer and Peter Guthrie. 2016. Identifying the needs of communities in rural uganda: A method for determining the "user-perceived values" of rural electrification initiatives. *Renewable and Sustainable Energy Reviews*, 66:476 – 486.

Jacinta Holloway and Kerrie L. Mengersen. 2018. Statistical machine learning methods and remote sensing for sustainable development goals: A review. *Remote Sensing*, 10(9):1365.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51(3):371–402.

Brian R Keeble. 1988. The brundtland report:'our common future'. *Medicine and War*, 4(1):17–25.

Roberta Kwok. 2019. AI empowers conservation biology. *Nature*, 567(7746):133–134.

Bandy X Lee, Finn Kjaerulf, Shannon Turner, Larry Cohen, Peter D Donnelly, Robert Muggah, Rachel Davis, Anna Realini, Berit Kieselbach, Lori Snyder MacGregor, et al. 2016. Transforming our world: implementing the 2030 agenda through sustainable development goal indicators. *Journal of public health policy*, 37(1):13–31.

Shoushan Li, Guodong Zhou, Zhongqing Wang, Sophia Yat Mei Lee, and Rangyang Wang. 2011. Imbalanced sentiment classification. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2469–2472.

Peng Liu, Wei Chen, Gaoyan Ou, Tengjiao Wang, Dongqing Yang, and Kai Lei. 2014. Sarcasm detection in social media based on imbalanced classification. In *Web-Age Information Management*, pages 459–471, Cham. Springer International Publishing.

Tim Loughran and Bill McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Iann Lundegård and Per-Olof Wickman. 2007. Conflicts of interest: An indispensable element of education for sustainable development. *Environmental Education Research*, 13(1):1–15.

Jennifer Mason. 2002. Organizing and indexing qualitative data. *Qualitative Researching*, 2:147–72.

Måns Nilsson, Dave Griggs, and Martin Visbeck. 2016. Policy: map the interactions between sustainable development goals. *Nature*, 534(7607):320–322.

Piia Nurkka, Sari Kujala, and Kirsi Kemppainen. 2009. Capturing users' perceptions of valuable experience and meaning. *Journal of Engineering Design*, 20(5):449–465.

OECD. 2017. *Development Co-operation Report 2017*. Organisation for Economic Co-operation and Development.

Manojkumar Parmar, Bhanurekha Maturi, Jhuma Mallik Dutt, and Hrushikesh Phate. 2018. Sentiment analysis on interview transcripts: An application of NLP for quantitative analysis. In *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018, Bangalore, India, September 19-22, 2018*, pages 1063–1068. IEEE.

Peace Child International. 2005. Needs and wants game. *vol3*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey S Pinegar. 2006. What customers want: using outcome-driven innovation to create breakthrough products and services by anthony w. ulwick. *Journal of Product Innovation Management*, 23(5):464–466.

Pulse Lab Jakarta. 2016. The 1st research dive on natural language processing for sustainable development. Technical report, Pulse Lab Jakarta Technical Report.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Mach. Learn.*, 85(3):333–359.

Thomas J Reynolds and Jonathan Gutman. 2001. Laddering theory, method, analysis, and interpretation. In *Understanding consumer decision making*, pages 40–79. Psychology Press.

Subas Risal. 2014. Mismatch between ngo services and beneficiaries' priorities: examining contextual realities. *Development in Practice*, 24(7):883–896.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *Computing Research Repository*, arXiv:1706.05098.

Jeffrey D Sachs. 2015. *The age of sustainable development*. Columbia University Press.

Horacio Saggion, Elena Stein-Sparvieri, David Maldavsky, and Sandra Szasz. 2010. NLP resources for the analysis of patient/therapist interviews. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Jagdish N Sheth, Bruce I Newman, and Barbara L Gross. 1991. Why we buy what we buy: A theory of consumption values. *Journal of business research*, 22(2):159–170.

Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial intelligence for social good: A survey. *Computing Research Repository*, arXiv:2001.01818.

David Silverman. 2013. *Doing qualitative research: A practical handbook*. SAGE publications limited.

Matthew S Simpson and Dina Demner-Fushman. 2012. Biomedical text mining: a survey of recent progress. In *Mining text data*, pages 465–517. Springer.

Michael R Solomon. 2002. The value of status and the status of value. In *Consumer value*, pages 77–98. Routledge.

Latanya Sweeney. 2000. Simple demographics often identify people uniquely.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan. The COLING 2016 Organizing Committee.

Katrin Tomanek and Udo Hahn. 2009. Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the fifth international conference on Knowledge capture*, pages 105–112.

Anthony W Ulwick. 2002. Turn customer input into innovation. *Harvard business review*, 80(1):91–98.

UN Agenda for Sustainable Development. 2018. Overview of standards for data disaggregation. Technical report, United Nations Working Paper.

UNESCO. 2013. *Adult and youth literacy: National, regional and global trends, 1985–2015*. UNESCO Institute for Statistics Montreal.

UNESCO. 2019a. Artificial intelligence for sustainable development: challenges and opportunities for unesco's science and engineering programmes. Technical report, UNESCO Working Paper.

UNESCO. 2019b. Artificial intelligence for sustainable development: synthesis report, mobile learning week 2019. Technical report, UNESCO Working Paper.

United Nations. 2015. Transforming our world: The 2030 agenda for sustainable development. *General Assembley 70 session*.

Ellen Van Kleef, Hans CM Van Trijp, and Pieternel Luning. 2005. Consumer research in the early stages of new product development: a critical review of methods and techniques. *Food quality and preference*, 16(3):181–201.

Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1):1–10.

Gerrit van der Waldt. 2019. Community profiling as instrument to enhance project planning in local government. *African Journal of Public Affairs*, 11(3):1–21.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Ryan Watkins, Maurya West Meiers, and Yusra Visser. 2012. *A guide to assessing needs: Essential tools for collecting information, making decisions, and achieving development results*. The World Bank.

Jason Wei and Kai Zou. 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. Association for Computational Linguistics.

WHO. 2016. *World health statistics 2016: monitoring health for the SDGs sustainable development goals*. World Health Organization.

Henry KH Woo. 1992. *Cognition, value, and price: a general theory of value*. Univ of Michigan Pr.

Leiming Yan, Yuhui Zheng, and Jie Cao. 2018. Few-shot learning for short text classification. *Multimedia Tools and Applications*, 77(22):29799–29810.

## Appendix A – Definitions of User-Perceived Values

| | | |
|---|---|---|
| *Emotional* | **Conscience** | |
| | Preservation of Environment | Preservation of natural resources |
| | Harmony | Being at peace with one another |
| | **Contentment** | |
| | Appealing Senses | Being pleasing to the senses taste and smell |
| | Aesthetics Items | Physical appearance of item or person which is pleasing to look at |
| | Comfort | State of being content, having a positive feeling |
| | Entertainment | Something affording pleasure, diversion or amusement |
| | Memorability | Association to a past event with emotional significance |
| | **Human Welfare** | |
| | Safety (Animals Items Nature) | Being protected from or prevent injuries or accidents by animals or nature |
| | Security People | Being free from danger and threat posed by people |
| *Epistemic* | **Information** | |
| | Information Access | Ability to stay informed |
| | **Knowledge** | |
| | Knowledge attainment | The ability to learn or being taught new knowledge |
| *Function* | **Convenience** | |
| | Banking Access | Having continuous access to banking services |
| | Communication | Ability to interact with someone who is far |
| | Mobile Phone Access | Having continuous access to mobile telecommunication services |
| | Mobility | Being able to transport goods, or to carry people from one place to another |
| | Multipurpose | Able to be used for a multitude of purposes |
| | Portable | An item that can easily be carried, transported or conveyed by hand |
| | Availability | Possible to get, buy or find in the area |
| | Time Benefit | Accomplish something with the least waste of time or minimum expenditure of time |
| | Time Management | Being able to work or plan towards a schedule |
| | Unburden | Making a task easier by simplifying |
| | **Cost Economy** | |
| | Capital Expenditure | Cost savings achieved |
| | School Fees | Ability to pay for school fee |
| | **Income Economy** | |
| | Economic Opportunity | Obtaining cash, assets, income through one-off sales or ongoing business opportunities |
| | Barter Trade | Non-monetary trade of goods or services |
| | **Quality and Performance** | |
| | Effectiveness | Adequate to accomplish a purpose or producing the result |
| | Lastingness | Continuing or enduring a long time |
| | Productivity | Rate of output and means that lead to increased productivity |
| | Reliability | The ability to rely or depend on operation or function of an item or service |
| | Usability | Refers to physical interaction with item being easy to operate handle or look after |
| *Indigenous* | **Social Norm** | |
| | Celebration | Association chosen as they play important part during celebration |
| | Manners | Ways of behaving with reference to polite standards and social components |
| | Morality | Following rules and the conduct |
| | Tradition | Expected form of behaviour embedded into the specific culture of city or village |
| | **Religion** | |
| | Faith | Belief in god or in the doctrines or teachings of religion |
| *Intrinsic Human* | **Health** | |
| | Longevity | Means that lead to an extended life span |
| | Health Care Access | Being able to access medical services or medicine |
| | Treatment | To require a hospital or medical attention as a consequence of illness or injury |
| | Preserv. of Health | Practices performed for the preservation of health |
| | **Physiological** | |
| | Education Access | Being able to access educational services |
| | Energy Access | Being able to obtain energy services or resources |
| | Food Security | The ability to have a reliable and continuous supply of food |
| | Shelter | A place giving protection from bad weather or danger |
| | Water Access | Continuous access or availability of water |
| | Water Quality | To have clean water as sickness, colour and taste |
| | **Quality of Life** | |
| | Wellbeing | Obtaining good or satisfying living condition (for people or for the community) |
| *Significance* | **Identity** | |
| | Appearance | Act or fact of appearing as to the eye or mind of the public |
| | Belongingness | Association with a certain group, their values and interests |
| | Dignity | The State or quality of being worthy of honour or respect |
| | Personal Performance | The productivity to which someone executes or accomplishes work |
| | **Status** | |

| | | |
|---|---|---|
| *Social* | Aspiration | Desire or aim to become someone better or more powerful or wise |
| | Modernisation | Transition to a modern society away from a traditional society |
| | Reputation | Commonly held opinion about ones character |
| | **Social Interaction** | |
| | Altruism | The principle and practice of unselfish concern |
| | Family Caring | Displaying kindness and concern for family members |
| | Role Fulfilling | Duty to fulfilling tasks or responsibilities associated with a certain role |
| | Togetherness | Warm fellowship, as among friends or members of a family |

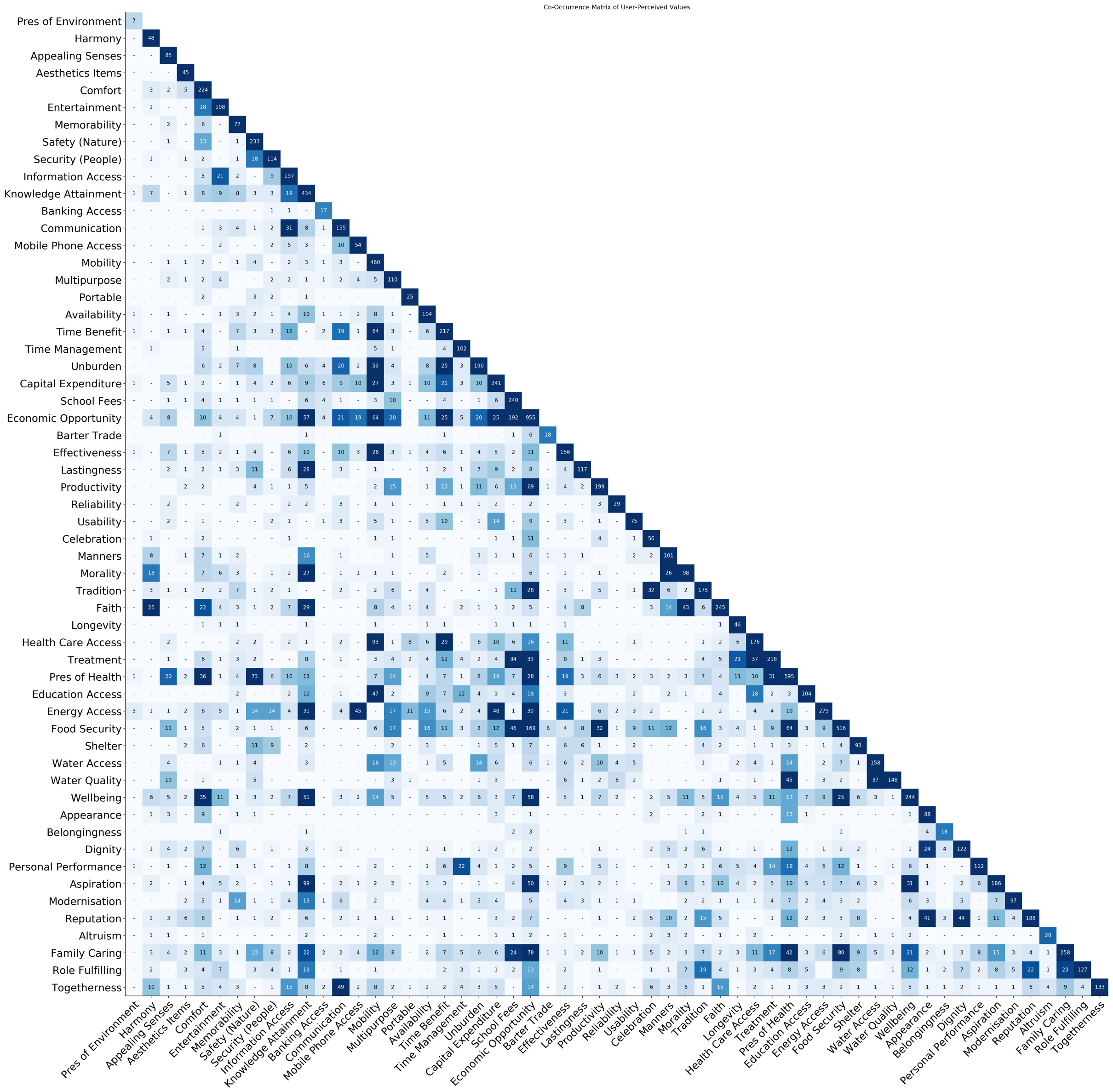


Figure 6: Co-occurrence matrix of T3 labels in the S2I corpus.

## Appendix B – Co-occurrence matrix of User-Perceived Values in the S2I corpus.

The co-occurrence matrix in Figure 6 depicts the inter-relatedness between different T3 labels. The intensity of colour corresponds to the number of samples in the S2I corpus where the given T3 labels co-occur.

The analysis of labels co-occurrence can offer valuable insights on commonly associated User-Perceived Values (UPVs, (Hirmer and Guthrie, 2014)): this can be useful to highlight challenges and problems in the considered community, which might not be known to the dwellers themselves. While some correlations are typical and expected, others are related to the specific Ugandan context, and might be surprising to those external to the location.

For example, *Economic Opportunity*, *Food Security* and *Preservation of Health* appear to frequently co-occur with other T3 labels. Note that the lack of employment opportunity, the availability of food and the quality of healthcare services represent endemic problems in the rural context studied in this paper. As they constitute primary concerns for most interviewees, it is therefore unsurprising that they were mentioned frequently in relation to many of the items selected as part of the UPV game (Section 4). A further illustrative example of the cultural context - in this case rural Uganda - is the high concurrence of *Unburden* and *Mobility*. This can be explained by the fact that rural roads are often of poor quality and villages or areas are inaccessible by motorised vehicles. Henceforth, people are required to find alternatives moves of transport for moving themselves to hospital or crops to the nearest market for sell, for example. As a final example, the frequent mentioning of *Faith*, *Harmony* and *Morality*, which also tend to co-occur in similar contexts, testifies the fundamental role played by religion in the rural villages considered in this study.

The information on the (co-)occurrence of UPVs in a community is also particularly valuable for designing appropriate project communication (Figure 1b), which can increase project buy-in through focused messaging (Section 3).

## Appendix C – Experimental Specifications.

In this Appendix, we report on the exact experimental setting used for experiments to aid experiment reproducibility.

### C.1 Data Specifications

**Data Selection and Splitting.** We select all sentences from the 119 interviews which were at least 3 tokens long and which were annotated with at least one UPV. We then randomly select an 80% proportion of the data as training set, and take the rest as heldout data (with a dev/test split of respectively 450 and 530 samples). Figure 7 shows that the obtained label distribution is similar.

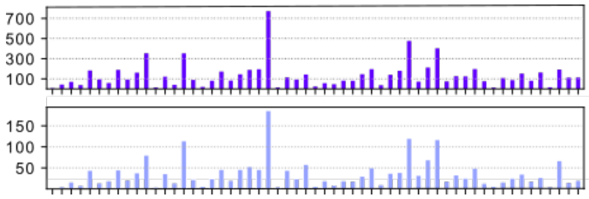


Figure 7: From top to bottom: distribution of UPVs in the training and heldout (dev+test) sets. Labels in the $x$ axis follow the same order as in Figure 3 of the main document.

**Data Anonymization.** In order to prevent the identification of the interviewees (Sweeney, 2000), data was manually anonymized. We anonymized all occurrences of: proper names, names of villages, cities or other geographical elements, and other names that might be sensitive (as names of tribes, languages, ...).

**Data Sample.** We are providing a sample of the data in the supplementary material. Each data sample is associated with the following fields:

- *id*: a unique identifier of the sample;
- *text*: a sentence to be classified;
- *t3_labels*: a list of the gold T3 labels associated with the sample.

For privacy reasons, we are not releasing metadata information associated with the samples (as the interviewee's name, gender, age, or the exact village name).

**Data Preprocessing.** For sentence splitting and word tokenization, we used NLTK's `sent_tokenize` and `word_tokenize` tokenizers (Bird and Loper, 2004)[12]. We use a set of regex to find interviewer comments and questions. Given that *Why-Probing* (Section 4.1, Reynolds and Gutman (2001)) was used, interviewers' comments are very limited and standard.

**Negative Samples Generation.** To generate negative samples (Section 6.1), we slightly modify Wei and Zou (2019)'s implementation[13] EDA (Easy Data Augmentation techniques) by adding a new function for character swapping and by adapting the stopword list. For semantic-based replacement, we rely on NLTK's interface[14] to WordNet (Fellbaum, 2012). Random shuffling and choice are controlled by a seed.

### C.2 Further Specifications

**(Hyper)-Parameters Selection.** All parameters used for experiments are reported in Table 4.

| parameter | value | parameter | value |
|---|---|---|---|
| *mildly neg* s. ratio | 5 | embedding size | 300 |
| *neg* sample ratio | 11 | LSTM hid. size | 128 |
| *strictly neg* s. ratio | 24 | dropout (all l.) | 0.2 |
| max sample len | 25 | batch size | 32 |
| max descr len | 15 | no epochs | 70 |
| max UPV code len | 4 | optimizer | *Adam* |

Table 4: Adopted (hyper-)parameters.

We use 300-dimensional FastText subword-informed pretrained vectors (Bojanowski et al.,

[12] `https://www.nltk.org/api/nltk.tokenize.html`

[13] `https://github.com/jasonwei20/eda_nlp/blob/d75e8bd4631f4d93260cb291aa47852d8eacd51d/code/eda.py#L65`

[14] `https://www.nltk.org/_modules/nltk/corpus/reader/wordnet.html`

2017)[15] to get the word embedding representations for each input sample.

Note that the goal of this paper is to present a new interesting NLP application, namely NLP for Sustainable Development: therefore, our goal here is to provide a set of robust baselines on our new S2I dataset, which can be referenced for future research. For this reason, we don't perform extensive hyper-parameter tuning on the selected models.

The only parameters we optimize are the number of generated negative samples of each type (*mildly negative, negative* and *strictly negative*). The best ratios were found empirically through experiments. The ratio used for optimization are reported in Table 5.

| total | mildly negative | negative | strictly negative |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 5 | 1 | 2 | 2 |
| 10 | 2 | 2 | 6 |
| 15 | 3 | 4 | 8 |
| 20 | 4 | 7 | 9 |
| 25 | 5 | 8 | 12 |
| 30 | 5 | 11 | 14 |
| 35 | 5 | 11 | 19 |
| **40** | **5** | **11** | **24** |
| 45 | 5 | 10 | 30 |
| 50 | 5 | 12 | 33 |
| 55 | 5 | 13 | 37 |
| 60 | 5 | 14 | 41 |

Table 5: Details of the relative number of *mildly negative, negative* and *strictly negative* samples used for experiments. Best ratio (used in all reported experiments) is in bold.

The analysis of the performance progression over training (Figure 8) shows that, in line with Wei and Zou (2019), adding negative examples is useful to improve performance: in our case, the plateau is reached around 40 augmented samples. In particular, we observe gains in all considered output levels (T1, T2 and T3 labels).

**Number of Parameters and Runtime Specifications.** Table 6 reports on the total number of (trainable) parameters and the average runtime/step for each considered model. Embeddings are kept fixed over training to avoid overfitting.

**Computing Infrastructure.** We run experiments on an NVIDIA GeForce GTX 1080 GPU.

**Evaluation Specifications.** For computing the evaluation metrics, we use the sklearn's (Pedregosa et al., 2011) implementation of precision, recall and $F_1$ score[16].

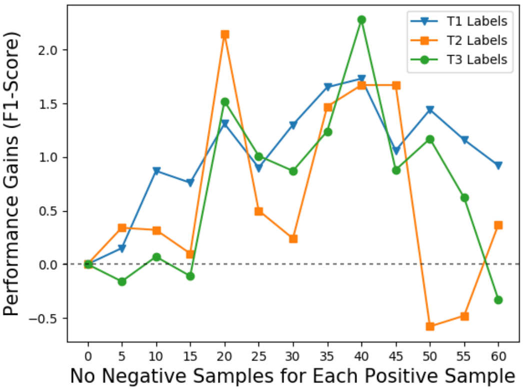


Figure 8: Progression of performance gains in F1-Score, considering the three labels T1, T2 and T3.

| (Multi-)task Setting | Model | #pars | avg runtime/ step |
|---|---|---|---|
| T3 | text | 373,377 | 55s |
| | +att | 590,013 | 56s |
| | +descr | 373,505 | 74s |
| | +att+descr | 806,777 | 75s |
| T2T3 | +att+descr | 844,154 | 78s |
| T1T2T3 | +att+descr | 865,019 | 85s |

Table 6: Number of trainable parameters and average runtime/step for all considered models and (multitask) training settings.

[15] We chose the `wiki.en.zip` model pretrained on the English Wikipedia https://fasttext.cc/docs/en/pretrained-vectors.html

[16] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

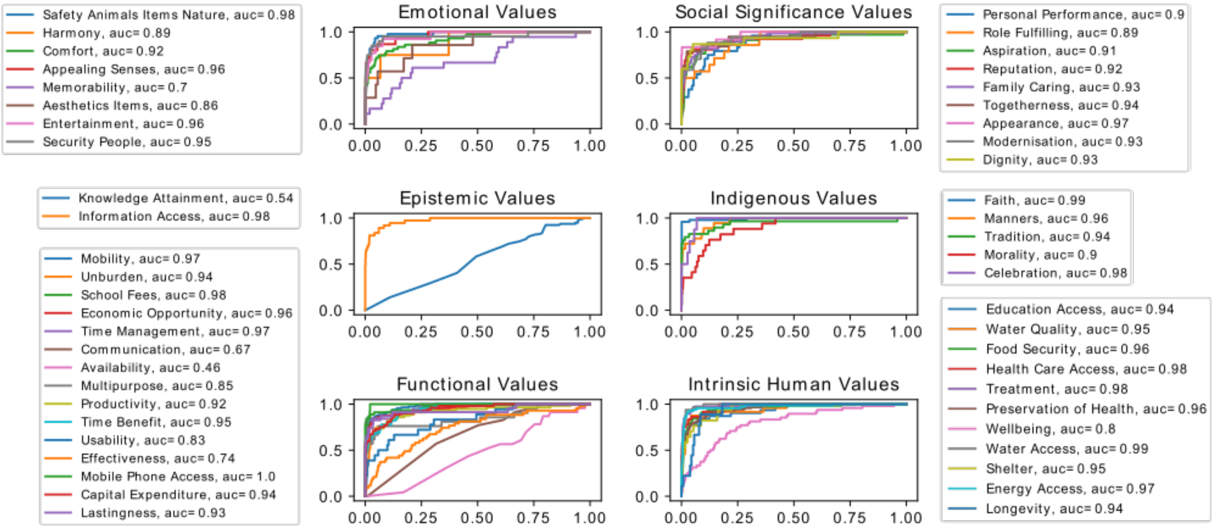


Figure 9: ROC curves for each T3 label, grouped by T1 categories.

## Appendix D – Single-Label Performance.

In this Appendix, we report the ROC curves for each T3 label, grouped by T1 categories. Figure 9 reports results obtained with the best performing model (Base+Attention+Description) trained with the T1+T2+T3 multi-task framework.

We evaluate with the "real-world" evaluation setting (Section 6.1), that is, we generate *all* positive and negative instances for each training sample. In practice, for a test sample $x$ associated with the T3 labels $T3_2$ and $T3_{45}$, we would generate 50 test instances $\{(x, T3_1) \rightarrow 0, (x, T3_1) \rightarrow 0, (x, T3_2) \rightarrow 1, ..., (x, T3_{50}) \rightarrow 0\}$, one for each of the T3 considered during training. All generated test samples would be negative, with the exception of $(x, T3_2)$ and $(x, T3_{45})$.

The single T3 labels' AUC show that satisfactory results are obtained overall for all T1 macro-labels: in particular, we obtain an AUC $>= 70$ for 47 out of 50 labels. Despite these promising results, our best model still struggles with some T3 labels, notably *Knowledge Attainment*, *Availability* and *Communication*. While the paper leaves ample room for future research, preliminary results are encouraging.

Refer to Section 6.2 for further details.