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

Costanza Conforti[1,2], **Stephanie Hirmer**[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
cc918@cam.ac.uk

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 applied in this context. In this research paper, we show the high potential of NLP applications to enhance sustainability of projects. 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. In this context, 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 task of *Automatic UPV classification*, which is an extreme multi-class multi-label classification problem. We release *Stories2Insights*, an expert-annotated dataset, provide a detailed corpus analysis, and implement a number of strong neural baselines to address the task. Experimental results show that the problem is challenging, and leave plenty of 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; 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* (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 unfortunately is of considerable lower resolution in de-

[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).

veloping 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 SDat 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 *User-Perceived Value Classification*, which consists in automatic annotation of qualitative interviews using an annotation schema developed in the field of sustainability; (4) we annotate and release *Stories2Insights* (S2I), a corpus for UPV classification; (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

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., 2019; Shi et al., 2020). Here 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).

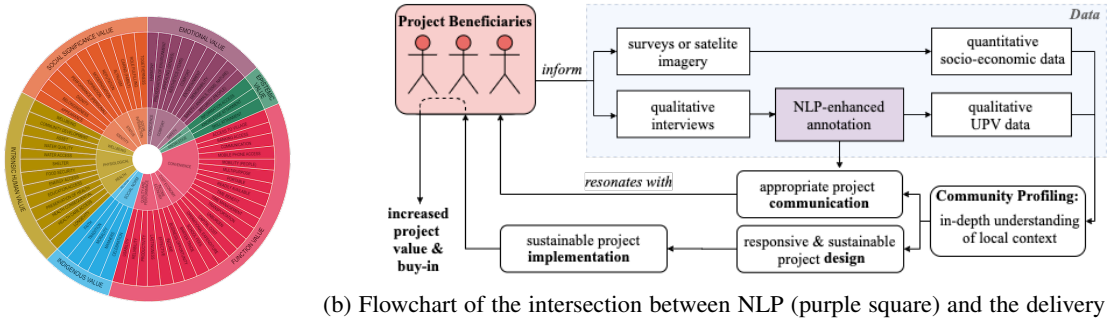
Ethics of AI for Social Good. Despite its positive impact, it is important to recognise that AI can act both as an enhancer and inhibitor of sustainability. As recently shown by Vinuesa et al. (2020), AI might inhibit meeting a considerable number of targets across the SDGs and may result in inequalities 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).

Project biases. Various works highlight the importance of understanding the local context and engaging with local stakeholders, incl. 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 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

The User-Perceived Values (UPV) Framework. As a means to obtain qualitative data with the characteristics mentioned above, we rely on the 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

[3] We have found a single example of the application of NLP, where it is used sporadically to analyse data from a gaming app used in a developing country. (Pulse Lab Jakarta, 2016).



(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).

their inclinations to adopt a new product or service (Nurkka et al., 2009). In the context of developing countries, the UPV framework identifies a set of 64 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. 3 T3 values)
3. *Functional*: contains the T2 values *Convenience*, *Cost Economy*, *Income Economy* and *Quality and Performance* (tot. 24 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. 12 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.

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.

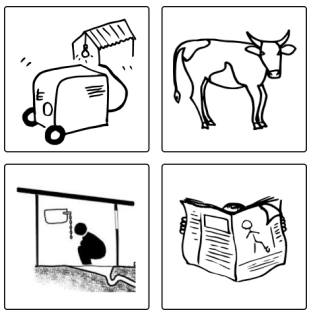
The role of NLP to enhance Sustainable Project Planning. Data conveying the beneficiaries' perspective is seldom considered in practical application, mainly due to the fact that it comes in the form of unstructured qualitative interviews. As introduced above, data needs to be *structured* in order to 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. In this context, we introduce the task of *Automatic UPV classification*, which consists of annotating each sentence of an 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. 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



(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 University of Cambridge); 2b) Women playing the UPV game in village (1)[4]; 2c) Map of case-study villages.

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 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 out of the 46 presented items, based on what is most important to them (*Select stimuli*), (2) to rank them in order of relative importance (*Ranking*); 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);

Data Collection Setting and Guidelines for Interviewers. For each village, 3 interviewers speaking the local language were hired to guide the UPV game. During the interviews, audio recording was used to supplement the note-taking. To ensure consistency and quality of the collected data, a two-day training workshop was held at Makerere University (Kampala, Uganda), and a local research assistant oversaw the entire data collection process including data collection 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 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 the study's integrity, a risk assessment following the U. of Cambridge's *Policy on the Ethics of Research Involving Human Participants and Personal Data* was completed. To protect the participants' identity, names of the villages were omitted.

Data Annotation. The interviews were analysed and annotated by domain experts[6] using the qualitative data analysis software *HyperResearch* (Hesse-Biber et al., 1991). To ensure consistency across

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

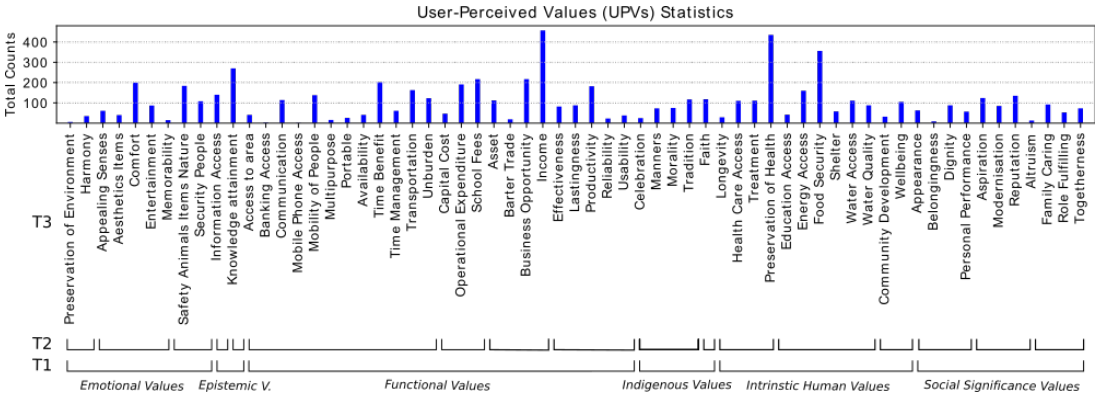


Figure 3: UPV frequencies from the S2I corpus (see Appendix A for UPV definitions).

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 6562 annotated utterances from the interviews. Samples present an average length of 17.2 tokens. The average number of samples per T3 label is 104, with an extremely skewed distribution: the most frequent T3, *Income*, occurs 457 times, while the least common, *Mobile Phone Access*, only twice (Figure 3). 655 samples ($\approx$ 11% of the tot.) are annotated with more than 1 label (see Appendix B for details on label 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 classification problems constitute 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 cant 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.*
- *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 label(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* problemwhere, given a labelled input sample (x, l_2), a system is trained to predict its correct class from a tagset $T = \{l_1, l_2, l_3\}$, for example $x \rightarrow l_2$ (i.e. [0,1,0]). Instead, inspired by previous work in aspect-based sentiment analysis (Wang et al., 2016; Pushp and Srivastava, 2017), we model the task as a *binary classification* problem: given an input sample and a candidate label, the system learns to predict the *relatedness* of the input sample with each one of the possible labels, i.e. $(x, l_1) \rightarrow 0$, $(x, l_2) \rightarrow 1$ and $(x, l_3) \rightarrow 0$.

We consider the true 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:

- $Contentment_{T2}$ = {$Aesthetic_{T3}$, $Comfort_{T3}$, ...}
- $Identity_{T2}$ = {$Appearance_{T3}$, $Dignity_{T3}$...}
- $Status_{T2}$ = {$Aspiration_{T3}$, $Reputation_{T3}$, ...}

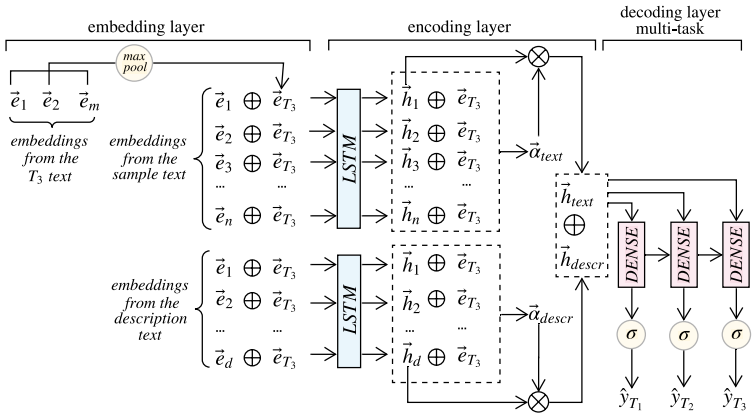


Figure 4: Multi-task neural architecture for UPV classification.

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.

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[7] 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 positive are quick to 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

Baseline Architecture.

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

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 embedding 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., 2015). 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.

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} follow-

Model	*test_set* (T3)			*real_simulation* (T3)		
	P	R	F1	P	R	F1
text	68.87	47.78	56.42	14.36	47.78	22.08
+att	66.08	**58.92**	62.30	16.27	**58.92**	25.50
+descr	**70.05**	58.20	63.24	16.83	58.30	26.11
+att+descr	69.60	58.40	**63.51**	**17.11**	58.40	**26.47**

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

ing 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.

Multi-task Training. A clear hierarchy exists between T3, T2 and T1 values (Section 3). We integrate signal containing such information using multi-task learning (Caruana, 1997; Ruder, 2017). Given an input sample, we predict not only its relatedness w.r.t. a T3 label, but also its relatedness with its corresponding T2 and T1 labels[8]. 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}. We then predict the relatedness of the sample with the given T1 label $\hat{y}_{T1}$ 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 $\sigma(dense_{T3}(h \oplus h_{T2}))$. In this way, 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

Data Preparation. We perform sentence splitting[9] on the 6587 utterances, obtaining 7348 samples. We generate 40 negative samples for each positive one (we found empirically that this was the best performing ratio). Sample weighting was used to account for the different error seriousness (1 for *negative* and *strictly negative* and 0.5 for *mildly negative*). Moreover, to expose the system to more diverse input, we slightly deform each positive sample when generating negative samples.

[8]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].

[9] We use NLTK for tokenization (Loper and Bird, 2002).

		Multi-task train setting					
Label		T3		T2+T3		T1+T2+T3	
Perf.		*ts*	*rs*	*ts*	*rs*	*ts*	*rs*
T3	P	**69.60**	17.11	69.19	19.15	67.83	**19.49**
	R	58.40	58.40	**63.10**	**63.10**	59.89	59.89
	F_1	63.51	26.47	66.01	29.39	63.61	**29.41**
T2	P	–	–	**84.21**	44.43	74.45	**45.11**
	R	–	–	35.02	38.22	**60.94**	**62.31**
	F_1	–	–	49.47	41.47	**67.02**	**52.33**
T1	P	–	–	–	–	85.64	67.31
	R	–	–	–	–	69.03	71.32
	F_1	–	–	–	–	76.45	69.26

Table 2: Results considering all labels granularities (T3, T2 and T1) training the best model, text+att+descr, with the 3 (multi-)task training settings (T3 only, T2+T3, T1+T2+T3). For each setting, *ts* refers to the *test_set* eval, and the *rs* to the *real_simulation* eval.

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).

Hyperparameter Selection and 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 embeddings. To prevent overfitting, the embedding matrix is kept fixed during training. Network hyperparameters are reported in Appendix C for replication. We train the model using binary cross-entropy loss, with early stopping monitoring the development set loss with a patience of 5 epochs. For training, 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 20% proportion from the data as test set.

Evaluation Framework. As the label distribution is highly skewed (1/40 ratio between positive and negative samples), we monitor precision, recall and F_1 score. We consider 2 eval settings: (1) *test_set*, which contains negative samples in the same proportion as in the train set; (2) *real_simulation*, where, for each sample, we generate *all possible* negative samples: this simulates a real scenario where we annotate a new interview with the corresponding UPVs. For multi-task training, we consider 3 layers of performance, corresponding to the labels *T3*, *T2* and *T1*. This is useful to compute because, in the application

context, different levels of granularity (T3/T2/T1 labels) can be monitored.

6.2 Results and Discussion

Models Performance. The results of our experiments are reported in Table 1. Notably, adding attention and integrating signal from descriptions to the base system caused significant improvements in performance. Significantly lower performance is observed in all settings from the *test_set* to the *real_simulation* evaluation setting. This is due to a substantial drop in precision, which proves the extreme difficulty of the task due to the significant imbalance between labels. Note, however, that recall remains stable over changes in evaluation setting. 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.

Multi-task Training. We consider the best performing model and run experiments with the three considered multi-task train settings (Section 5.1). 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 useful information from the label hierarchy which is indeed useful for UPV classification.

Error Analysis. We perform a detailed error analysis of the best performing model's predictions in the *real_simulation* setting, which proved to be more challenging. As reported in Table 3, we observe a correlation between a T3 label's support in the corpus and the system's precision in predicting that label: with almost no exception, all labels where the system obtained a precision lower than 30 had a support similar or lower than 3%. Not surprisingly, particularly good performance is often obtained on T3 labels which often correlate with specific terms (as *School Fees*, or *Faith*). 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.

T1	T3	P	R	F_1	Supp	%
Emotional	Harmony	50.0	66.7	57.1	035	0.53%
	Appealing	33.3	54.5	41.4	062	0.94%
	Aesthetics	37.5	25.0	30.0	043	0.66%
	Comfort	39.1	43.9	41.4	209	3.19%
	Entertainment	68.4	65.0	66.7	085	1.3%
	Safety	49.3	70.8	58.1	217	3.31%
	Sec. People	55.9	67.9	61.3	124	1.89%
Epist	Info. Access	44.4	52.2	48.0	158	2.41%
	Knowl. attain.	57.9	53.2	55.5	319	4.86%
Function	Access to area	16.7	30.0	21.4	049	0.75%
	Communication	01.6	100	03.2	119	1.81%
	Mob. of People	33.3	63.2	43.6	164	2.5%
	Availability	00.8	100	01.5	043	0.66%
	Time Benefit	50.8	66.7	57.7	226	3.44%
	Time Manag.	25.8	66.7	37.2	077	1.17%
	Transportation	39.2	67.4	49.6	246	3.75%
	Unburden	37.0	32.3	34.5	140	2.13%
	Capital Cost	30.0	60.0	40.0	044	0.67%
	Oper. Expend.	56.1	60.4	58.2	213	3.25%
	School Fees	66.0	76.7	71.0	218	3.32%
	Asset	20.0	16.0	17.8	133	2.03%
	Business Opp.	63.0	29.3	40.0	254	3.87%
	Income	48.6	82.6	61.2	496	7.56%
	Effectiveness	20.0	21.4	20.7	090	1.37%
	Lastingness	31.6	31.6	31.6	099	1.51%
	Productivity	43.3	53.1	47.7	193	2.94%
	Usability	20.0	10.0	13.3	039	0.59%
Indigen.	Manners	62.5	41.7	50.0	074	1.13%
	Morality	55.6	27.8	37.0	081	1.23%
	Tradition	95.8	62.2	75.4	123	1.87%
	Faith	68.2	83.3	75.0	165	2.51%
Intrinsic Human	Healthc. Acc.	46.7	45.2	45.9	131	2.0%
	Treatment	39.5	55.6	46.2	141	2.15%
	Pres. of health	41.0	71.6	52.1	482	7.35%
	Educ. Acc.	32.3	100	48.8	042	0.64%
	Energy Acc.	61.0	78.1	68.5	162	2.47%
	Food Security	55.6	67.9	61.1	378	5.76%
	Shelter	33.3	43.8	37.8	070	1.07%
	Water Access	48.8	80.8	60.9	139	2.12%
	Water Quality	38.7	52.2	44.4	124	1.89%
	Comm. Devel.	0.0	0.0	0.0	034	0.52%
	Wellbeing	56.0	60.9	58.3	111	1.69%
Social Significance	Appearance	42.9	54.5	48.0	059	0.9%
	Pers. Perf.	28.6	15.4	20.0	057	0.87%
	Aspiration	25.0	16.0	19.5	143	2.18%
	Modernisation	22.2	11.8	15.4	096	1.46%
	Reputation	66.7	51.6	58.2	141	2.15%
	Fam. Caring	60.0	47.4	52.9	093	1.42%
	Role Fulf.	0.0	0.0	0.0	056	0.85%
	Togetherness	77.8	43.8	56.0	076	1.16%

Table 3: Single label performance and support in the S2I corpus. Results obtained with the best model (T1+T2+T3 training), rounding predictions at 0.5 and evaluating with the *real_simulation* setting.

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 S2I 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 traditional value of 0.5 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 open a constructive dialogue between the fields of NLP and SD, and result in new interesting applications.

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.

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

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 futurecall 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.

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. 2019. Artificial intelligence for social good. *arXiv preprint arXiv:1901.05406*.

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. *arXiv preprint arXiv:1605.05894*.

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.

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.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

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*.

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.

Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train once, test anywhere: Zero-shot learning for text classification. *arXiv preprint arXiv:1712.05972*.

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. *CoRR*, abs/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. *CoRR*, abs/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.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Effective lstms for target-dependent sentiment classification. *arXiv preprint arXiv:1512.01100*.

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.

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 W Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

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.

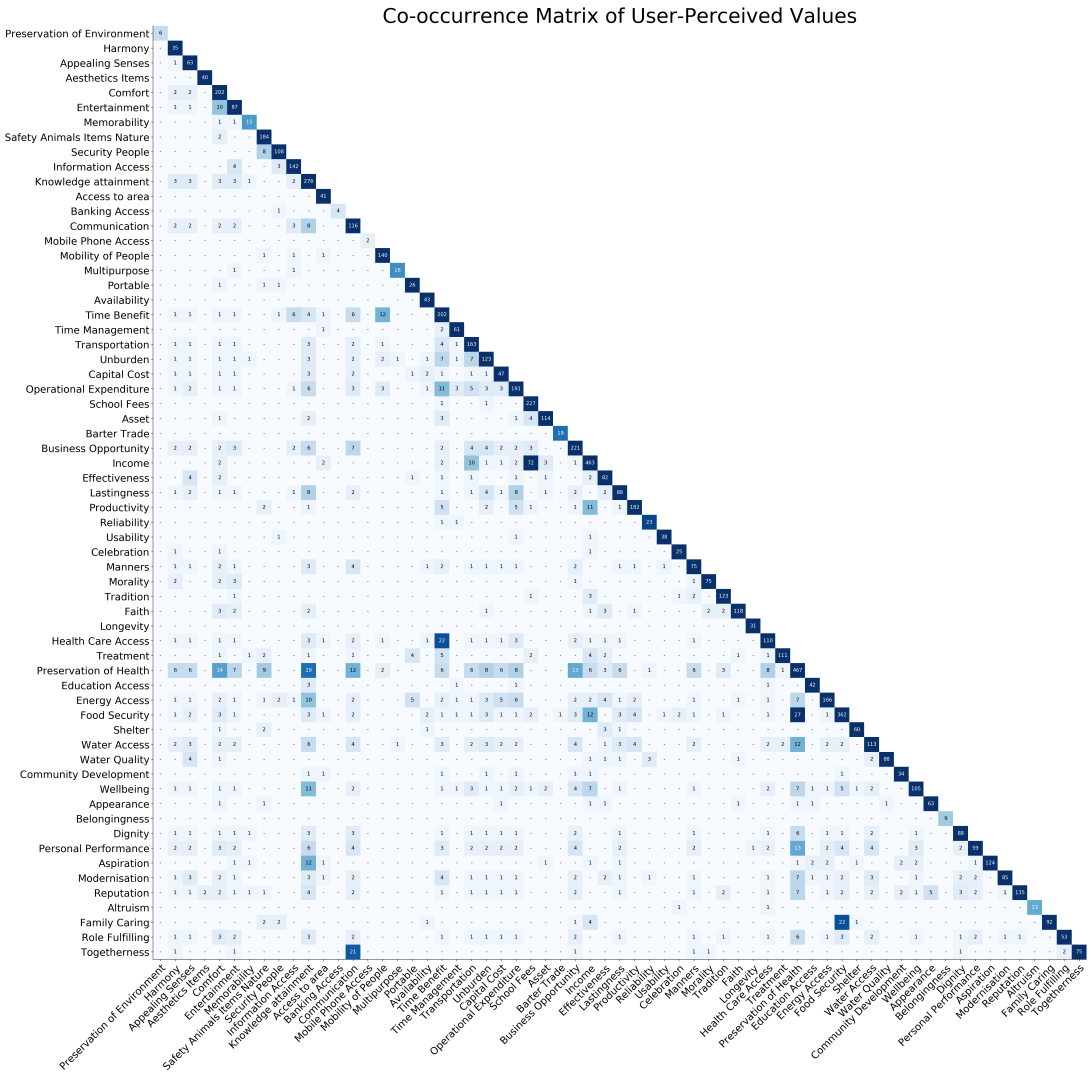
Leiming Yan, Yuhui Zheng, and Jie Cao. 2018. Few-shot learning for short text classification. *Multimedia Tools and Applications*, 77(22):29799–29810.

Appendix A Definitions of User-Perceived Values

Emotional	**Conscience**	
	Preservation of Environment	Being at peace with one another
	Harmony	Preservation of natural resources
	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
	Skill attainment	The ability to learn a new skill
Function	**Convenience**	
	Access to area	Having continuous access to the village or city
	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 of People	Ability to move 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
	Transportation	Conveying and transporting someone or something
	Unburden	Making a task easier by simplifying
	Cost Economy	
	Capital Cost	Fixed one time expenditure through purchase of an item or service
	Operational Expenditure	Cost savings achieved through the operation of an item or service
	School Fees	Ability to pay for school fee
	Income Economy	
	Asset	Something that can be of future benefit
	Barter Trade	Non-monetary trade of goods or services
	Business Opportunity	Sense of entrepreneurship beyond the normal occupation
	Income	Ability to make money through the sale of a good or service
	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	
	Community Development	Improvement of services or infrastructure for benefit of collective group or people
	Wellbeing	A good or satisfying living condition

Social 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	
	Aspiration	Desire or aim to become someone better or more powerful or wise
	Modernisation	Transition to a modern society away from a traditional to the manner of a developed 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

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



Appendix C Adopted (Hyper-)Parameters.

parameter	value	parameter	value
mildly neg s. ratio	2	embedding size	300
neg sample ratio	2	LSTM hid. size	128
strictly neg s. ratio	6	dropout (all l.)	0.2
max sample len	15	batch size	32
max descr len	15	no epochs	70
max UPV code len	4	optimizer	*Adam*

Appendix D Single-Label Performance.

ROC curves for each T3 label, grouped by T1 categories. Reported results are obtained with the best performing model (Base+Attention+Description) trained with the T1+T2+T3 multi-task framework. We evaluate with the *real_simulation* setting (Section 6.1), that is, we consider the associated T3 labels in the gold as positive instances, and we generate all possible negative samples.