TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Bai, N., Luo, R., Nourian, P., & Pereira Roders, A. (2021). WHOSe Heritage: Classification of UNESCO World Heritage Statements of "outstanding Universal Value" with Soft Labels. In M.-F. Moens, X. Huang, L. Specia, & S. W.-T. Yih (Eds.), *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021* (pp. 366-384). Article 34 (Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2021.findings-emnlp.34

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

WHOSe Heritage: Classification of UNESCO World Heritage Statements of "Outstanding Universal Value" with Soft Labels

Nan Bai[1], **Renqian Luo**[2], **Pirouz Nourian**[3], **Ana Pereira Roders**[1]
[1]UNESCO Chair in Heritage and the Reshaping of Urban Conservation for Sustainability, Chair of Heritage and Values, Delft University of Technology, Delft, the Netherlands
[2]University of Science and Technology of China, Hefei, China
[3]Chair of Design Informatics, Delft University of Technology, Delft, the Netherlands
{n.bai, p.nourian, a.r.pereira-roders}@tudelft.nl
lrq@mail.ustc.edu.cn

Abstract

The UNESCO World Heritage List (WHL) includes the exceptionally valuable cultural and natural heritage to be preserved for mankind. Evaluating and justifying the Outstanding Universal Value (OUV) is essential for each site inscribed in the WHL, and yet a complex task, even for experts, since the selection criteria of OUV are not mutually exclusive. Furthermore, manual annotation of heritage values and attributes from multi-source textual data, which is currently dominant in heritage studies, is knowledge-demanding and time-consuming, impeding systematic analysis of such authoritative documents in terms of their implications on heritage management. This study applies state-of-the-art NLP models to build a classifier on a new dataset containing Statements of OUV, seeking an explainable and scalable automation tool to facilitate the nomination, evaluation, research, and monitoring processes of World Heritage sites. Label smoothing is innovatively adapted to improve the model performance by adding prior inter-class relationship knowledge to generate soft labels. The study shows that the best models fine-tuned from BERT and ULMFiT can reach 94.3% top-3 accuracy. A human study with expert evaluation on the model prediction shows that the models are sufficiently generalizable. The study is promising to be further developed and applied in heritage research and practice.[1]

1 Introduction

Since the World Heritage Convention was adopted in 1972, 1121 sites has been inscribed worldwide in the World Heritage List (WHL) up to 2019, aiming at a collective protection of the cultural and natural heritage of Outstanding Universal Value (OUV) for mankind as a whole (UNESCO, 1972; von Droste, 2011; Pereira Roders and van Oers, 2011). First proposed in 1976, OUV, meaning the "*cultural and/or natural significance which is so exceptional as to transcend national boundaries and to be of common importance for present and future generations of all humanity*", has been operationalized and formalized into an administrative requirement for new inscriptions on the WHL since 2005. (UNESCO, 2008; Jokilehto, 2006, 2008). All nominations must meet one or more of the ten selection criteria (6 for culture and 4 for nature), focusing on different cultural and natural values.

Since 2007, complete **Statements of OUV** (SOUV) need to be submitted and approved for new World Heritage (WH) nominations, which should include, among others, a section of "*justification for criteria*", giving a short paragraph to explain why a site (also known as property) satisfies each of the criteria it is inscribed under. These statements are to be drafted by the State Parties after scientific research for any tentative nominations, further reviewed and revised by the Advisory Bodies from ICOMOS and/or IUCN, and eventually approved and adopted by the World Heritage Committee for inscription. Similarly, Retrospective SOUV have been required for sites inscribed before 2006 to revise or refill the section *justification of criteria* (IUCN et al., 2010). However, the evaluation of SOUV can be ambiguous in the sense that: 1) the selection criteria are not mutually exclusive and contain common information about historical and aesthetic/artistic values as an integral part (Jokilehto, 2008); 2) the key stakeholders to evaluate the SOUV for a nomination occasionally disagree with each other at early stages, leading to recursive reviews and revisions, though all are considered to be domain experts (Jokilehto, 2008; Tarrafa Silva and Pereira Roders, 2010; von Droste, 2011). A tool to check the accuracy, objectivity, consistency, and coherence of such statements can significantly benefit the inscription process involving thousands of experts worldwide each year.

Not only for new nominations, the SOUV are

[1]Code and data for this project are available at https://github.com/zzbn12345/WHOSe_Heritage

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 366–384
November 7–11, 2021. ©2021 Association for Computational Linguistics

also essential reference points for monitoring and interpreting inscribed heritage sites (IUCN et al., 2010). Researchers and practitioners actively and regularly check if the justified criteria are still relevant for the sites, as to decide on further planning and managerial actions. Moreover, these same statements are also used in support of legal court cases, should WH sites be endangered by human development (Pereira Roders, 2010; von Droste, 2011). Under the support of the *Recommendation of Historic Urban Landscape* and the recent *Our World Heritage* campaign, multiple data sources (e.g., news articles, policy documents, social media posts) are encouraged in such analyses of identifying and mapping OUV (UNESCO, 2011; Bandarin and van Oers, 2012; Ginzarly et al., 2019). The traditional method of manually annotating heritage values and attributes by experts can be time-consuming and knowledge-demanding for analysing massive social media posts by people in cities with WH sites to find OUV-related statements, albeit dominantly applied in practice (Tarrafa Silva and Pereira Roders, 2012; Abdel Tawab, 2019; Tarrafa Silva and Pereira Roders, 2010).

To approximate both ultimate goals of this study: 1) aiding the inscription process by checking the coherence and consistency of SOUV, and 2) identifying heritage values from multiple data sources (e.g., social media posts), a computational solution rooted on SOUV is desired. By training NLP models with the officially written and approved SOUV, a machine replica of the collective authoritarian view could be obtained. This machine replica will not be employed at this stage to justify OUV for new nominations from scratch. Rather, it will assess the written SOUV of WH sites (either existing or new) and classify OUV-related texts with the learned collective authoritarian view. Furthermore, it can investigate the existing SOUV from bottom up and capture the subtle intrinsic associations within the statements and among the corresponding selection criteria (Bai et al., 2021a). This yields a new perspective on interpreting the WHL, which would give insights for furthering amending the concept of OUV and selection criteria to be better discernible.

Therefore, this study aims at training an explainable and scalable classifier that can reveal the intrinsic associations of World Heritage OUV selection criteria, which can be feasible to apply in real-world analyses by researchers and practitioners. As outcome, this paper presents the classifier of UNESCO **W**orld **H**eritage Statements of **O**UV with **S**oft Lab**e**ls (**WHOSe** Heritage).

The contributions of this Paper can be summarized as follows: 1) A novel text classification dataset is presented, concerning a domain-specific task about Outstanding Universal Value for UNESCO World Heritage sites; 2) Innovative variants of label smoothing are applied to introduce the prior knowledge of label association into training as soft labels, which turned out effective to improve performance in most investigated popular models as baselines in this task; 3) Several classifiers are trained and compared on the Statements of OUV classification task as initial benchmarks, supplemented with explorations on their explainability and generalizability using expert evaluation.

2 Related Work

Text classification In the past decades, numerous models have been proposed from shallow to deep learning models for text classification tasks. In shallow learning models, the raw input text is pre-processed to extract features of the text, which are then fed into machine learning classifiers, e.g., Naive Bayes (Maron, 1961) and support vector machine (Joachims, 1998) for prediction. In deep learning models, deep neural networks are leveraged to extract information from the input data, such as convolutional neural networks (CNN) (Kim, 2014; Johnson and Zhang, 2017), recurrent neural networks (RNN) (Tai et al., 2015; Cho et al., 2014), attention networks (Yang et al., 2016) and Transformers (Devlin et al., 2019). Multi-class and multi-label tasks are two extensions of the simplest binary classification, where every sample can belong to one or more classes within a class list (Aly, 2005; Tsoumakas and Katakis, 2007), where the labels may also be correlated (Pal et al., 2020). This work explores the combined application of some popular shallow and deep learning models for a multi-class classification task.

Label Smoothing Label smoothing (LS) is originally proposed as a regularization technique to alleviate overfitting in training deep neural networks (Szegedy et al., 2016; Müller et al., 2019). It assigns a noise distribution on all the labels to prevent the model from predicting too confidently on 'ground-truth' labels. It is widely used in computer vision (Szegedy et al., 2016), speech (Chorowski and Jaitly, 2017) and natural language processing

(Vaswani et al., 2017) tasks. Originally the distribution is uniform across the labels, which is data independent. Recently, other variants of LS are also proposed that are able to incorporate the interrelation information from the data into the distribution (Zhong et al., 2016; Zhang et al., 2021; Krothapalli and Abbott, 2020). In this work, the technique is applied to generate soft labels with a distribution derived from domain knowledge since the classes in this task are clearly interrelated with each other.

Transfer Learning in NLP In many real-world applications, labelled data are limited and expensive to collect. Training models with limited data from scratch affects the performance. Transfer learning (Pan and Yang, 2010) is widely used to solve this by using word embeddings that are pre-trained on massive corpus and fine-tuning them on target task. Earlier works (Mikolov et al., 2013; Pennington et al., 2014) provide static word embeddings that ignore the contextual information in the sentences. More recent works, e.g., ULMFiT (Howard and Ruder, 2018) and BERT (Devlin et al., 2019), take the context into account and generate dynamic contextualized word vectors, showing excellent performance, which also prove to be sufficiently generalizable across many tasks. This task, with a relatively small data size, employs the idea of transfer learning and applies both embedding methods.

3 Data and Problem Statement

3.1 Data Collection and Pre-processing

UNESCO World Heritage Centre openly releases a syndication dataset of the sites in XLS format[2], which includes information of the inscribed World Heritage sites such as *ID, name, short description*, *justification of criteria* et. al.. Among them, the field of *justification* provides a paragraph for each selection criterion the site fulfills[3], contributing as the input data for this task. In total, 1052 out of 1121 WH sites contain the *justification* data[4], while the remaining 69 await the Retrospective SOUV to be approved as introduced in Section 1. As an example, in Venice and Its Lagoon, the paragraph on **criterion (i)** shows:

> *...The lagoon of Venice also has one of the highest concentrations of masterpieces in the world: from Torcello's Cathedral to the church of Santa Maria della Salute.The years of the Republic's extraordinary Golden Age are represented by monuments of incomparable beauty...*[5]

For any inscribed WH site $p_i \in P$, where P is the set of all the sites, it may fulfill one or more of the ten selection criteria. By checking if each criterion is justified for the site p_i, a non-negative vector $\boldsymbol{\gamma}_i := [\gamma_{i,k}]_{\kappa\times 1},\ k \in [1, \kappa], \kappa = 10$ can be formed as the "parental" label for the site:

$$\gamma_{i,k} = \begin{cases} 1, & \text{if } p_i \text{ meets the } k_{\text{th}} \text{ criterion,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Meanwhile, the paragraphs $\boldsymbol{X}_i$ in the *justification* field of p_i, describing all criteria that p_i has, are split into sentences. For the j_{th} sentence $\boldsymbol{x}_{i,j,k}$ describing the criterion k possessed by the site p_i, a non-negative one-hot vector $\boldsymbol{y}_{i,j,k}$ can be formed as the "ground-truth" label for this single sentence:

$$\boldsymbol{y}_{i,j,k} = \boldsymbol{e}_k \in \{0, 1\}^{\kappa}. \quad (2)$$

Each sentence $\boldsymbol{x}_{i,j,k} \in \boldsymbol{X}_i$ is treated as a sample, with two labels: a one-hot "ground-truth label" $\boldsymbol{y}_{i,j,k}$ for the particular sentence, and a multi-class "parental label" $\boldsymbol{\gamma}_i$ for all sentences that belong to the site p_i. The sentence-level setup is desirable here since paragraphs may contain overwhelming information of multiple OUV criteria, as will be shown in Section 3.2. As such, a more specific indication of OUV tendencies in each part of the texts could be differentiated. Complementarily, the fine-grained sentence-level prediction vectors could still be aggregated into paragraph/text levels without losing lower-level details, which will be demonstrated in Figure 2. As the sentences were written, revised, and approved by various domain experts at local and global levels during the inscription process, the labels can be considered as having a good "inter-annotator agreement" (Jokilehto, 2008; Nowak and Rüger, 2010).

The following data pre-processing techniques are applied to construct the final dataset used for training: 1) all letters are turned into lower-case; 2) the umlauts and accents are normalized; 3) numbers are replaced with a special $<$ NUM $>$ token; 4) only sentences with a length between 8 and 64 words are kept, based on the dataset distribution; 5) the sentences are randomly split

[2] http://whc.unesco.org/en/syndication. Copyright © 1992 - 2021 UNESCO/World Heritage Centre. All rights reserved.

[3] This field is not complete in the original XLS dataset. The WHC website is walked through to fill in the missing values.

[4] The statistics are up to the 44th session of the World Heritage Committee held in Fuzhou, China in July 2021, after which the total number of WH sites grew to 1154.

[5] https://whc.unesco.org/en/list/394

Split	C1	C2	C3	C4	C5	C6	N7	N8	N9	N10	Sum
train	333	631	651	774	209	327	386	261	370	572	4514
valid	40	71	83	89	28	49	43	42	42	76	563
test	41	79	72	92	35	47	45	32	50	71	564
test in SD	815	1563	1647	2049	554	876	510	334	465	548	9361
seen w LS	1077	1747	1832	2131	609	1063	1130	630	1047	1251	12517

Table 1: The number of samples in sentence level that contain each criterion as a label, annotated with C1 to C6 for cultural values and N7 to N10 for natural values. The first three rows show the data split using the field *justification*; the fourth row shows a new dataset only for testing using the field *short description* (SD); the last row shows the potential samples the models can see for each criterion after introducing label smoothing (LS).

into train/validation/test sets with a proportion of 8:1:1. Additionally, the official definition sentences of selection criteria[6] as given in Table 4 of Appendix A are respectively appended into the train split with the same one-hot sentence and parental labels for each criterion. Stop-words are not removed since BERT and ULMFiT to be applied generally prefer natural texts with context information. Furthermore, an additional 11_{th} class "Others" is introduced by appending an arbitrary noise of $\gamma_{i,\kappa+1} = 0.2$ to all parental labels $\boldsymbol{\gamma}_i$, and a 0 to all "ground-truth" labels $\boldsymbol{y}_{i,j,k}$, so that the models are not forced to give predictions only to the ten criteria even when the relevance to all of them is weak. For each sentence, the 11_{th} "Others" class and the complement sets of its parental labels could be regarded as the negative classes for classification since the site this sentence describes is not justified with those values. An exemplary pre-processed data sample is shown in Table 6 in Appendix A.

On average, 27.97 ± 11.04 words appear in each sentence. A summary of the number of samples in sentence level in each split for each criterion is presented in the first three rows of Table 1.

Similarly, the paragraphs $\boldsymbol{S}_i$ in the field *short description* of WH site p_i, giving a general introduction of the site, which are not originally written to describe any specific OUV selection criterion, are pre-processed into an additional independent test dataset SD to evaluate the generalizability of the classifiers on unseen data that comes from a slightly different distribution. For those sentences $\boldsymbol{s}_{i,o} \in \boldsymbol{S}_i$, both ground-truth and parental labels are the same as $\boldsymbol{\gamma_i}$ for the site they describe. The total number of samples that contain each criterion in SD dataset is shown in the fourth row of Table 1.

3.2 Association between Classes

Jokilehto (2008) summarized the selection criteria with their main focuses by inspecting the official definitions and the justification texts of WH sites. Details about the definitions of the criteria could be found in Appendix A. However, as stated in Section 1, the criteria are not mutually exclusive. The **criterion (i)** justification of Venice in Section 3.1 will be again used as an example. Judging as a domain expert, it clearly describes criterion (i) as labelled, since it explicitly uses the term "*masterpieces*" and "*monuments of incomparable beauty*". However, traces can still be found on other values: 1) as it describes the "*Cathedral*", "*church*", and "*monuments*", it also concerns the criterion (iv) about architectural *typology*; 2) as it talks about the "*Golden Age*", it also points to criterion (ii) about *influence* and criterion (iii) about *testimony*. In fact, Venice is also justified with criteria (ii), (iii), and (iv). Pragmatically speaking, for sites fulfilling more than one OUV selection criteria, it is hard to avoid talking about the other criteria while isolating one criterion alone (Pereira Roders, 2010).

Furthermore, the association between each pair of criteria can be different. The distinction between criteria is generally larger when the pair comes from a different category (cultural v.s. natural). For a pair of criteria from the same category, the association level can also vary. For example, Jokilehto (2008) pointed out that "*criteria (i) and (ii) can reinforce each other while (iv) is often used as an alternative*". This complex association pattern can also be seen in the co-occurrence matrix $\boldsymbol{A}_{\boldsymbol{\kappa}\times\boldsymbol{\kappa}} := [a_{k,l}]_{\kappa\times\kappa},\ k,l \in [1,\kappa]$ of the criteria in all the inscribed sites P, where the diagonal entries record the number of cases when each criterion is used alone (shown in Figure 4 of Appendix A):

$$a_{k,l} = \begin{cases} \sum_i \left(\gamma_{i,k}\gamma_{i,l}\right), & \text{if } k \neq l, \\ \sum_i \lfloor \frac{\gamma_{i,k}}{\sum_{j\in[1,\kappa]}\gamma_{i,j}} \rfloor, & \text{otherwise.} \end{cases} \tag{3}$$

This intrinsic association is to be used as the prior knowledge for the classification task.

[6] http://whc.unesco.org/en/criteria/

4 Models and Experiments

4.1 Soft Labels Generation

Section 3.2 argues that the selection criteria are not mutually exclusive, and that co-justified criteria of a WH site that have a stronger association may be reflected in the sentences describing a specific criterion. In other words, classifying such sentences is not purely a single-label multi-class classification task. Rather, it also has a multi-label characteristic considering the "parental labels" of the sites.

To leverage the problem between the two sorts of tasks and to prevent the models from being overconfident at the only "ground-truth" labels, this paper proposes to apply the label smoothing (LS) technique with two novel variants to combine the "ground-truth" sentence label $\boldsymbol{y}_{i,j,k}$ and the parental document label $\boldsymbol{\gamma}_i$ into a single vector $\widetilde{\boldsymbol{y}}_{i,j,k}$ as soft labels for training process. This is similar to the hierarchical LS approach proposed by (Zhong et al., 2016) to reflect the prior label similarity distribution. We propose three variants: **vanilla** that assigns identical "noises" to all classes, which will be proved equivalent to the original LS in Appendix B; **uniform** that treats all co-justified associated criteria in the parental label equally; and **prior** that weights the co-justified criteria based on the frequency that the pair co-occurs in matrix $\boldsymbol{A}_{\kappa\times\kappa}$:

$$\widetilde{\boldsymbol{y}}_{i,j,k} = \begin{cases} \mathbf{f}(\boldsymbol{y}_{i,j,k} + \alpha\mathbf{1}), & \text{if } \textbf{vanilla}, \\ \mathbf{f}(\boldsymbol{y}_{i,j,k} + \alpha\boldsymbol{\gamma}_i), & \text{if } \textbf{uniform}, \\ \mathbf{f}(\boldsymbol{y}_{i,j,k} + \alpha\boldsymbol{\mu}_k \odot \boldsymbol{\gamma}_i), & \text{if } \textbf{prior}. \end{cases} \tag{4}$$

Here $\mathbf{f} : \mathbb{R}_+^d \to [0,1]^d$ is a variant of the original softmax function so that it maps a $d-$dimensional vector of non-negative real numbers to a distribution that sums up to 1:

$$\mathbf{f}(\boldsymbol{z})_t = \frac{e^{z_t} - 1}{\sum_{l=0}^{d} e^{z_l} - d}, \text{or } \mathbf{f}(\boldsymbol{z}) = \frac{e^{\boldsymbol{z}} - \mathbf{1}}{e^{\boldsymbol{z}^T}\mathbf{1} - d}, \tag{5}$$

for $t \in [0, d), \mathbf{1} := [1]_{d\times 1}$ and $\boldsymbol{z} := [z_t]_{d\times 1} \in \mathbb{R}_+^d$; α is a scalar that leverages the effect of LS; $\boldsymbol{\mu}_k := [\mu_{l,k}]_{(\kappa+1)\times 1}$ is a criterion-specific non-negative vector showing the inter-criteria associations:

$$\mu_{l,k} = \frac{a_{l,k}}{\sum_i a_{i,k}}, \; l \in [1, \kappa+1], \tag{6}$$

and $\odot$ represents the element-wise Hadamard-Schur product of vectors. This variant of the softmax function introduced in Equation 5 is preferable since it transforms the combined non-negative labels-vectors in Equation 4 to a "probability" distribution while keeping non-related labels still as 0. For example, a combined vector $[2, 0, 1, 0]^T$ becomes $[.62, .08, .22, .08]^T$ with normal softmax, and $[.79, 0, .21, 0]^T$ with this variant.

All three variants are considered as options during training, and tuned as hyperparameters together with the scalar $\alpha \in \{0, 0.01, 0.05, 0.1, 0.2, 0.5, 1\}$. For all variants, the problem is purely multi-class when $\alpha = 0$, and approaches multi-label when α gets larger, giving parental labels larger weights.

The following benefits can be achieved with the use of proposed LS variants: 1) The knowledge of the actual association of classes (selection criteria) are introduced into the training in both uniform and prior variants, giving the model chances to learn these intrinsic associations with soft labels; 2) The freedom on the design decision of whether the problem should be multi-class or multi-label is provided for the model training process; 3) The models can potentially see more instances for each class during training with LS variants, as shown in the last row of Table 1; 4) The computed soft label vector $\widetilde{\boldsymbol{y}}_{i,j,k}$ is mathematically more similar to the prediction vector $\widehat{\boldsymbol{y}}_{i,j,k}$ than one-hot vectors, both of which are discrete "probability" distributions, pushing the use of Cross-entropy Loss closer to its original definition (Rubinstein and Kroese, 2013).

4.2 Baselines

Five models are selected as baselines: 1) N-gram (Cavnar et al., 1994) embedding followed by multi-layer perceptron (MLP); 2) Bag-of-Embeddings (BoE) using GloVe (Pennington et al., 2014); 3) Gated Recurrent Unit (GRU) (Cho et al., 2014) with Attention (Bahdanau et al., 2015; Yang et al., 2016) (denoted as GRU+Attn); 4) Pretrained ULMFiT language model (Howard and Ruder, 2018) further fine-tuned on the full WHL domain dataset; and 5) uncased base BERT model (Devlin et al., 2019). The former three models are trained mostly from scratch (where BoE and GRU+Attn used the GloVe-6B-300d vectors as initial embeddings), while the latter two are extensively pretrained and fine-tuned on this specific classification task. The model implementation details and the hyperparameter configurations are shown in Appendix C.

4.3 Metrics

For the training process, **Cross-Entropy** is used as the loss-function for two soft label vectors, while three metrics are used to evaluate the model per-

formance as a multi-class classification task: 1) **Top-1 Accuracy** which counts the instances when the predicted class with the highest output value matches the ground-truth sentence label; 2) **Top-k Accuracy** which counts the instances when the ground-truth sentence label is among the top k predicted classes with the highest output values; 3) **Macro-averaged F1** which calculates the overall cross-label performance. **Per-class Metrics** (i.e., top-1 precision, recall, and F1) for each selection criteria are also calculated for evaluation purpose.

For the independent SD test set, two metrics are defined here to evaluate the model performance as a multi-label classification task: 1) **Top-1 Match** which counts the instances when at least one of the parental labels matches the predicted class; 2) **Top-k Match** which counts the instances when at least one parental label is among the top k predicted classes. Arguably, the top-1 and top-k matches are more tolerant extensions of top-1 and top-k accuracy into multi-label classification scenarios.

For all evaluation metrics, k is chosen to be 3 following the rationale introduced in Appendix A.

4.4 Experiment Setup

The experiment consists of three successive steps for each baseline (details given in Appendix C):

1. Grid search within a small range is performed to tune the hyperparameters with a single random seed, and the best configuration is selected according to the top-k accuracy on the validation split;

2. LS with different α values under all three conditions (vanilla, uniform, and prior) is tested using the configuration from step 1, repeated with 10 different random seeds, treated as another round of hyperparameter tuning, saving the best LS configuration according to the performance mean and variance over the seeds;

3. The best LS configuration in step 2 is applied to save a model with the same random seed used in step 1 and evaluated together with the baseline model without LS, both on validation/test splits and on SD test set;

Early-stopping is applied during all training processes based on the top-k accuracy on the validation split. The models are implemented in PyTorch (Rao and McMahan, 2019) and experiments are performed on NVIDIA Tesla P100 GPU and Intel Core i7-8850H CPU, respectively. The inference is performed entirely on a CPU to test the models' feasibility in more general application scenarios when GPU can be unavailable for end-users. More details of training resource utilization, model size, and inference time is shown in Appendix D.

5 Results and Analyses

5.1 Experiment Results

The averaged top-k accuracies of experiments conducted with 10 random seeds are shown in Figure 1. In most cases (except for BoE), the models with proposed LS variants (uniform or prior) either strictly or weakly out-perform the baselines (without LS or with vanilla LS) based on multiple experiments. Furthermore, the proposed LS variants seem to make the models more robust to over-fitting and catastrophic forgetting problems, especially with the cases of BERT and ULMFiT. The uniform variant of LS with different α values appears in most models. A possible explanation is that uniform LS introduces the prior knowledge from the parental labels as "noise" in a simple way during the training, balancing yet not challenging the "ground-truth" sentence labels (Müller et al., 2019). Yet, the complex effect of LS on different baselines invites further investigation.

Table 2 shows the performance of the models with and without LS on the validation split, test split, and SD test set. Except for BoE, introducing LS increased the performance of most baselines in most metrics. Generally speaking, the pretrained models dominate the performance, and the highest score for all the metrics occurs in either ULMFiT or BERT, mostly with LS. Still, top-1 accuracy only reaches 71% in the best models, while top-k accuracy manages to reach 94%, suggesting that it would be more reliable to look at the top 3 predictions during application in this task. The models perform remarkably well in the SD test set, though given a relatively simpler task than in training, indicating the generalizability of the classifiers.

The per-class top-1 metrics of the best models in each baseline on the validation and test split (Table 3) make it evident that the difficulty for classifying each selection criterion varies. T-test shows that F1 score is significantly different between the cultural and natural criteria ($t = 8.20, p < .001$), suggesting that natural criteria are probably more clearly defined, while cultural ones might be closely intertwined. The poor performance on

Model	LS Config	val 1	val k	val F1	test 1	test k	test F1	SD 1	SD k
N-gram	w/o LS	67.38	90.82	63.11	59.96	88.87	58.87	70.49	95.13
	uniform 0.1	67.19	91.21	62.11	59.57	89.65	58.24	71.12	95.26
BoE	w/o LS	64.84	91.99	63.11	62.11	91.60	61.93	68.80	94.53
	prior 0.01	64.26	91.60	62.48	62.70	91.41	62.14	66.15	94.14
GRU+Attn	w/o LS	64.26	91.60	60.83	60.55	91.41	59.28	64.27	92.71
	uniform 0.2	64.26	91.80	61.36	61.52	90.23	61.06	66.35	94.06
ULMFiT	w/o LS	69.34	93.95	68.40	66.41	92.38	66.09	70.21	96.15
	prior 0.1	70.12	**94.34**	68.83	67.19	93.16	66.97	70.65	**96.22**
BERT	w/o LS	70.31	**94.34**	69.60	**67.58**	93.55	67.15	**71.56**	95.96
	uniform 0.2	**71.68**	93.95	**70.42**	66.99	**94.53**	**67.34**	71.51	96.15

Table 2: The performance of models with and without LS on validation split, test split (top-1 accuracy, top-k accuracy, and averaged macro F1), and independent SD test set (top-1 match and top-k match), where k=3. The best score for each metric is highlighted in bold, and underlined if the best score occurs in models with LS. The effect of adding LS to each baseline is marked with background colors: blue indicates a rise in performance, red indicates a drop, while grey indicates a tie. The darker background color indicates a larger variation in performance.

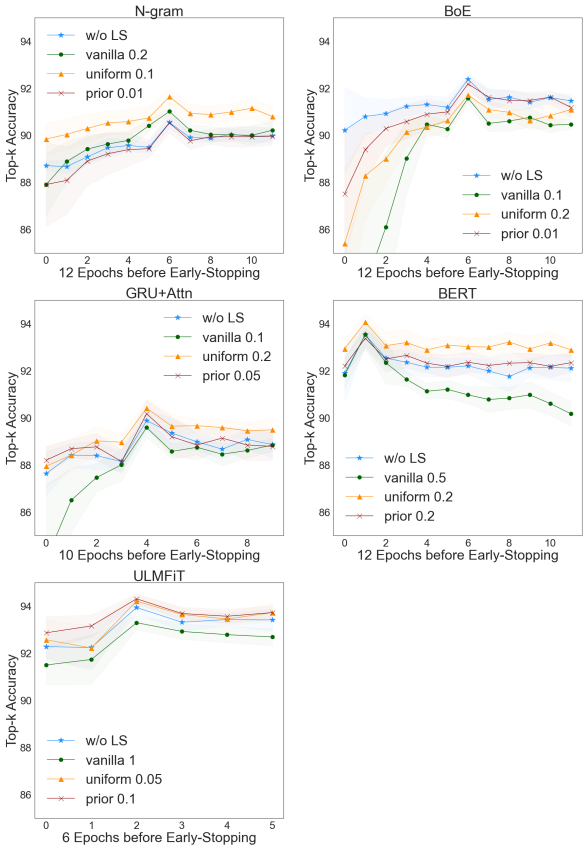


Figure 1: The average training curve of best-performing models in experiments under 10 random seeds for each baseline on validation split. The x-axes show several epochs before the early-stopping happened. The numbers of epochs are different for each baseline as described in Appendix C. Orange curves with triangles show the top-k (k=3) accuracy with uniform LS, red curves with crosses the performance of prior LS, green curves with circles for vanilla LS, and blue curves with stars show the performance without LS. 95 % confidence intervals of the performance based on the 10 random seeds are shown in shades.

OUV	Focus	Prec	Recall	F1
C1	Masterpiece	46.68	71.52	56.18
C2	Values/Influences	69.19	66.34	67.56
C3	Testimony	63.96	58.60	61.01
C4	Typology	61.10	54.23	57.24
C5	Land-Use	40.98	52.30	45.01
C6	Associations	58.28	67.89	61.27
N7	Natural Beauty	78.94	70.89	74.35
N8	Geological Process	66.92	80.42	72.39
N9	Ecological Process	60.16	67.23	63.45
N10	Bio-diversity	86.89	78.54	82.48

Table 3: The average per-class metrics over all models on validation and test splits with LS, and the main focus of each criteria adapted from Jokilehto (2008).

criterion (v) is consistent with its smallest sample size (as shown in Table 1); meanwhile, the models perform reasonably well for criterion (viii) with the second smallest sample size. This suggests that except for sample size, the strong associations between the classes can also influence the difficulty for NLP models (and probably also for human experts) to distinguish the nuance of criteria. Criterion (i) has a far poorer precision than recall, suggesting that samples from other criteria, especially from criterion (iv) based on the confusion matrices shown in Figure 5 of Appendix D, are easily mistaken as this one. This is also comprehensible since criterion (i), emphasizing that a site is a *masterpiece*, can be easily mentioned "unintentionally" in the description of criterion (iv) that regards the value of some specific *architectural typology*.

5.2 Error Analysis and Explainability

Although sometimes challenged (Serrano and Smith, 2020), attention mechanisms are believed to be effective for visualizing NLP model perfor-

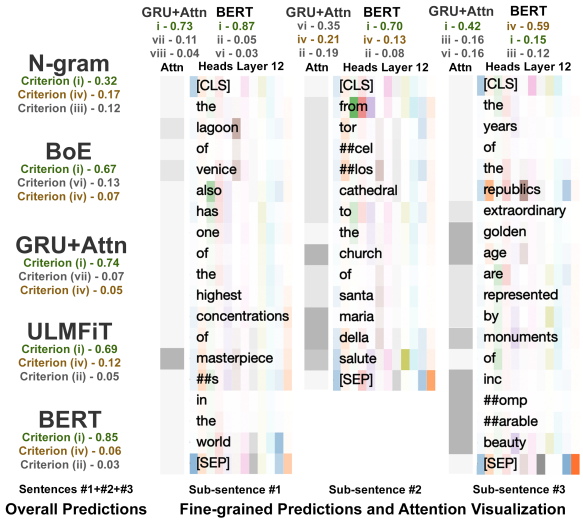


Figure 2: The overall and fine-grained top-3 predictions of models, and attention weights of GRU+Attn and BERT models on the exemplary sub-sentences concerning criterion (i) in Venice. The left part of the image reports the top-3 predictions of all 5 models when the models take the aggregated paragraph as input. The top part reports the fine-grained top-3 predictions of two models on each sub-sentence. The rest of the image visualizes the attention weights. Attention weights of GRU+Attn is visualized in grey-scale, and that of BERT is illustrated using BertViz as coloured bars.

mance in an explainable manner (Yang et al., 2016; Vaswani et al., 2017; Tang et al., 2019; Sun and Lu, 2020). The same example on OUV selection criterion (i) in Venice as in Section 3.1 and 3.2 will be demonstrated here using the trained models from the attention-enabled GRU+Attn and BERT, as shown in Figure 2, with the help of BertViz library (Vig, 2019; Vaswani et al., 2018). GRU+Attn employs a single universal attention mechanism to all inputs, while BERT has 12 attention heads for the [CLS] token on its last layer, both of which manage to capture the meaningful keywords and phrases such as *masterpiece*, *church*, *golden age*, *monuments*, and *incomparable beauty* in the sentences. As a note, Clark et al. (2019) used probing to find out that some BERT attention heads correspond to certain linguistic phenomena. In this study, the attention heads from the last layer also seem to focus on different semantic information of OUV. This observation invites further studies.

Figure 2 also shows the top-3 predictions of the models on the exemplary sentences. In the overall predictions taking the sentences as a paragraph for input, all models manage to give the ground-truth label criterion (i) the highest predicted value

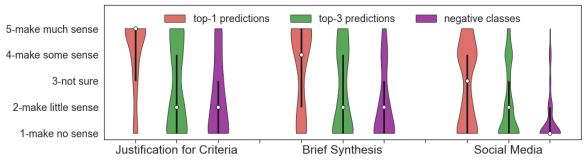


Figure 3: The distribution as violin plots of expert evaluations given to the relevance of selection criteria and sample sentences about Venice from three sources. The scores for top-1 and top-3 classes and the negative class predicted by the models are plotted separately. The 25%, 75% percentiles, and the medians are also shown.

(from 0.32 in N-gram to 0.85 in BERT). Remarkably, all models also include criterion (iv) in the top-3 predictions (from 0.05 in GRU+Attn to 0.17 in N-gram), suggesting that the sentences might also be related to criterion (iv). The fine-grained predictions taking each sub-sentence as input, however, show a different pattern. Although criterion (i) is almost always present in the top-3 predictions, criterion (iv) shows to take a higher place in the second sentence by GRU+Attn, and in the third sentence by BERT. This behaviour is not necessarily an error per se in prediction. Rather, considering the arguments in Section 3.2, those sub-sentences could be indeed relevant to other criteria (in this case, criterion iv) based on the association pattern, q.v. Bai et al. (2021a), indicating why criterion (iv) is always included in the overall predictions.

5.3 Expert Evaluation

Eight heritage researchers with rich experience in identifying heritage values and attributes were invited for a human study adapted from He et al. (2021), Nguyen (2018) and Schuff (2020), to test the models' reliability and generalizability. They were presented with 56 sentences about Venice harvested from "*Justification*" (14) and "*Brief Synthesis*" (13) in SOUV and Social Media platforms (29). Each sentence was given three positive classes as top-1 and top-3 criteria predictions from BERT and ULMFiT models, and one negative class as another random cultural criterion. Not knowing that the criteria are predictions by computer models, the experts were asked to rate the relevance of the sentences and each criterion on a 5-point Likert scale.

The distributions of all the ratings are shown in Figure 3. For all data sources, the expert ratings for top-1 and top-3 predictions are significantly higher than those for negative classes based on Mann-Whitney U tests (See Table 8 in Appendix E). The average ratings of experts for

each sentence-criterion pair show a strong correlation with the average confidence scores of models ($r = 0.618, p < 0.001$). Some heritage experts seem to be rather cautious and reserved to assess informal texts as "culturally significant" without further historical contexts and comparative studies. For example, the third sentence in Table 9 of Appendix E from social media, "*In 1952, the station was finalized on a design by the architect Paul Perilli*" with a predicted label of criterion (i) got extremely divergent expert scores. For some experts, it is clearly related to criterion (i) about *masterpiece* based on the semantic content. However, for the experts who rated a low score, merely declaring that some building is designed by a certain architect does not automatically entail that it is a masterpiece. Further investigations have to be made to fully convince them. Although such an example shows disagreement amongst the experts and between the experts and the computer models, it does not limit the machine's ability to differentiate positive and negative classes. Full details of the human study are presented in Appendix E. The expert evaluation proves that the models are sufficiently reliable and capable of identifying OUV-related statements even from the less formal social media data, useful for the ultimate motivations of this study discussed in Section 1.

6 Discussion and Conclusions

This paper presents a new text classification benchmark from a real-world problem about UNESCO World Heritage Statements of Outstanding Universal Value (OUV). The problem is essentially a multi-class single label classification task, while the classes are not necessarily mutually exclusive. The prior knowledge of the class association is added to the training process as soft labels through novel variants of label smoothing (LS). The study shows that introducing LS improved the performance on most baselines, reaching a top-3 accuracy of 94.3%. The models also performed reasonably well in an independent test dataset and received positive outcomes in a human study with domain experts, suggesting that the classifiers have the potential to be further developed and applied in the World Heritage research and practice.

LS was not tuned together with other hyperparameters during the training. Yet, it still showed an improvement in most baselines. However, the complex effect of LS on different baselines needs more investigation. The top-1 accuracy is limited even on the best models, which is not uncommon in the literature for non-binary multi-class classification when the labels are not sufficiently distinct (Sun et al., 2019). Applying data augmentation and training supplemental binary classifiers may improve the performance on difficult classes. The choice of replacing all numbers into $<$ NUM $>$ tokens might introduce both advantages and drawbacks in terms of semantic context and generalizability when historical dates might be crucial information, which invites more investigations. Moreover, more studies on the generalizability and reliability of the models on data from different distributions (e.g., from policy documents or news article) are needed before further application. This work would support a series of follow-up studies respectively exploring the intrinsic associations of OUV based on the models' behaviour (Bai et al., 2021a), application of the proposed methods in social media mining in Venice (Bai et al., 2021b), and generalizability in case studies worldwide.

This work is intended to aid, but not replace the workload of human stakeholders: for State Parties to identify OUV-related statements through documentation, for Advisory Bodies and WHC to review and revise the yearly nomination proposals, for researchers to investigate massive official discourse and user-generated content, and for the public to visually understand the values of *Their World Heritage* around them. Therefore, this work **WHOSe Heritage** can be another milestone for the digital transformation of World Heritage studies, aiming at a more socially inclusive future practice.

Acknowledgements

The presented study is within the framework of the Heriland-Consortium. HERILAND is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 813883. The authors are grateful for all the constructive comments from the anonymous reviewers.

Broader Impact Statement

This work focuses on exploring and applying NLP techniques to a real-world application of cultural and natural World Heritage (WH) preservation for the sake of social good. The research is to aid the identification and justification of heritage values across the world for various stakeholders, including

both heritage experts and lay-persons, through text classification, as is pointed out in Section 1 and 6. It can lead to better understandings of the OUV criteria and the association among them.

The dataset used in this work is collected by the author(s) from the public website of UNESCO World Heritage Centre via XLS syndication respecting the terms of use and copy rights. The description of the dataset is sufficiently revealed in section 3.1 and Appendix A. All labels used are based on the official OUV justification given by local and global heritage experts and involve no crowd workers or other new annotators. The dataset and the methods used in the paper do not contain demographic/identity characteristics. Once deployed, the model does not learn from user inputs, and it generates no harmful output to users. The expert evaluation involving human study was totally voluntary, did not collect any personal information, and the privacy of the experts was fully protected. Though initially unaware of the true purpose of the evaluation to reduce bias, the experts were explained with the study afterwards.

BERT and ULMFiT with LS proved to perform best in all investigated metrics. However, there is a trade-off to consider for real-world application. As claimed in Appendix D and Section 5.2, ULMFiT has a relatively shorter inference time compared to BERT, while BERT is potentially more explainable due to the attention mechanism. Both models might work optimally for different application scenarios.

Nevertheless, the interpretation of the classification result needs to be carefully conducted by researchers and practitioners, especially during policy decision-making on World Heritage for the social benefit of the entire human species. WH inscription and OUV justification are far more complicated than only reading written texts and identifying the described values. Rather, it is a systematic thematic study based on scientific research and always rooted in a COMPARATIVE study across the globe (Jokilehto, 2008). The actual decisions of including new nominations into the WHL have to be made by human with heritage investigations. This is also evident in the results of expert evaluation and during the open discussion about the exercise with invited experts. As stated in the example shown in Section 5.2, thorough heritage investigations are always needed to determine if a site truly justifies certain OUV selection criteria. Such investigations, however, would be out of the scope of our NLP application study investigating the semantic and syntactic content of written official documents. Therefore, a human has to be involved in the loop during application.

This study and the obtained NLP models are inherently less biased than manual annotation by a single expert in the sense that they avoid adding too much implicit personal experience into the written texts, and that the trained models represent the collective views of many human experts in the past. This can also be seen in some divergent evaluation outcomes by the eight invited experts, as demonstrated in Appendix E: though one specific expert may be more cautious and critical at a certain sample, the overall trend of all experts can consistently differentiate the positive and negative classes. However, the computational models trained on SOUV can also be a double-edged sword in the sense that they are highly dependent on the existing descriptions, which may contain historical unfairness.

Researchers and practitioners, especially those outside of the Computer Science field, need to be explicitly informed and even warned before usage on the limitations of such models, to avoid automation bias, which shows that people favour the results automatically generated from systems for decision-making (Parasuraman and Manzey, 2010). Wrongly under-judging the value of a WH nomination merely based on text classification results and consequently deferring or even refusing the inscription can cause a great loss to human culture in the worst scenario, as it can hamper its access to the available heritage management and conservation programs. Therefore, this work functions as a supplemental tool and reference for the understanding/evaluating of World Heritage OUV implied in text descriptions, which will and shall not replace the human effort and/or deviate the expert knowledge in WH decision-making process. Instead, it has two ultimate goals as use-cases: 1) aiding inscription processes by checking the coherence and/or consistency of OUV statements; 2) mining heritage-values-related texts from multiple data sources (e.g., social media).

References

Abdel Tawab. 2019. The Assessment of Historic Towns' Outstanding Universal Value Based on the Interchange of Human Values They Exhibit. *Heritage*, 2(3):1874–1891.

Mohamed Aly. 2005. Survey on multiclass classification methods. *Neural Netw*, 19:1–9.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

N. Bai, P. Nourian, R. Luo, and A. Pereira Roders. 2021a. "What is OUV" revisited: A computational interpretation on the statements of Outstanding Universal Value. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, VIII-M-1-2021:25–32.

N. Bai, P. Nourian, and A. Pereira Roders. 2021b. Global citizens and world heritage: Social inclusion of online communities in heritage planning. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVI-M-1-2021:23–30.

Francesco Bandarin and Ron van Oers. 2012. *The Historic Urban Landscape*. WILEY-BLACKWELL, Chinchester.

William B Cavnar, John M Trenkle, et al. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, pages 161–175. Citeseer.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111. Association for Computational Linguistics.

Jan Chorowski and Navdeep Jaitly. 2017. Towards better decoding and language model integration in sequence to sequence models. In *Proc. Interspeech 2017*, pages 523–527.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Manal Ginzarly, Ana Pereira Roders, and Jacques Teller. 2019. Mapping historic urban landscape values through social media. *Journal of Cultural Heritage*, 36:1–11.

Zeya He, Ning Deng, Xiang (Robert) Li, and Huimin Gu. 2021. How to "read" a destination from images? machine learning and network methods for dmos' image projection and photo evaluation. *Journal of Travel Research*, 0(0):1–23.

Jeremy Howard and Sylvain Gugger. 2020. Fastai: A layered api for deep learning. *Information (Switzerland)*, 11(2):1–26.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics.

IUCN, ICOMOS, ICROM, and WHC. 2010. Guidance on the preparation of retrospective Statements of Outstanding Universal Value for World Heritage Properties. Technical report.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Proceedings*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570, Vancouver, Canada. Association for Computational Linguistics.

Jukka Jokilehto. 2006. World Heritage: Defining the Outstanding Universal Value. *City & Time*, 2(2):1–10.

Jukka Jokilehto. 2008. What is OUV? Defining the Outstanding Universal Value of Cultural World Heritage Properties. Technical report, ICOMOS, ICOMOS Berlin.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ujwal Krothapalli and A. Lynn Abbott. 2020. Adaptive label smoothing. *CoRR*, abs/2009.06432.

M. E. Maron. 1961. Automatic indexing: An experimental inquiry. *J. ACM*, 8(3):404–417.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, NeurIPS, pages 4694–4703.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1069–1078. Association for Computational Linguistics.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566.

Ankit Pal, Muru Selvakumar, and Malaikannan Sankarasubbu. 2020. MAGNET: multi-label text classification using attention-based graph neural network. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 2, Valletta, Malta, February 22-24, 2020*, pages 494–505. SCITEPRESS.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Ana Pereira Roders. 2010. Revealing the World Heritage cities and their varied natures. In *Heritage 2010: Heritage and Sustainable Development, Vols 1 and 2*, chapter 1 Heritage and Governance for Development, pages 245–253. Green Lines Institute.

Ana Pereira Roders and Ron van Oers. 2011. World Heritage cities management. *Facilities*, 29(7):276–285.

Delip Rao and Brian McMahan. 2019. *Natural Language Processing with PyTorch - Build Intelligent Language Applications Using Deep Learning*. O'Reilly Media, Inc.

Reuven Y Rubinstein and Dirk P Kroese. 2013. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media.

Hendrik Schuff. 2020. Explainable question answering beyond f1: metrics, models and human evaluation. Master's thesis, Universitaet Stuttgart.

Sofia Serrano and Noah A. Smith. 2020. Is attention interpretable? *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 2931–2951.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to Fine-Tune BERT for Text Classification? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11856 LNAI(2):194–206.

Xiaobing Sun and Wei Lu. 2020. Understanding Attention for Text Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1999, pages 3418–3428.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1556–1566. The Association for Computer Linguistics.

Matthew Tang, Priyanka Gandhi, Md. Ahsanul Kabir, Christopher Zou, Jordyn Blakey, and Xiao Luo. 2019. Progress notes classification and keyword extraction using attention-based deep learning models with BERT. *CoRR*, abs/1910.05786.

Ana Tarrafa Silva and Ana Pereira Roders. 2010. The cultural significance of World Heritage cities : Portugal as case study. In *Heritage and Sustainable Development*, June, Évora, Portugal.

Ana Tarrafa Silva and Ana Pereira Roders. 2012. Cultural Heritage Management and Heritage (Impact)

Assessments. In *Joint CIB W070, W092 & TG72 International Conferenceon Facilities Management, Procurement Systems and Public Private Partnership*, January 2012.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *Int. J. Data Warehous. Min.*, 3(3):1–13.

UNESCO. 1972. Convention Concerning the Protection of the World Cultural and Natural Heritage. Technical report, UNESCO.

UNESCO. 2008. Operational guidelines for the implementation of the world heritage convention. Technical report, UNESCO World Heritage Centre.

UNESCO. 2011. Recommendation on the Historic Urban Landscape. Technical report, UNESCO World Heritage Centre.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, François Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Lukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018, Boston, MA, USA, March 17-21, 2018 - Volume 1: Research Papers*, pages 193–199. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 37–42. Association for Computational Linguistics.

Bernd von Droste. 2011. The concept of outstanding universal value and its application: "From the seven wonders of the ancient world to the 1,000 world heritage places today". *Journal of Cultural Heritage Management and Sustainable Development*, 1(1):26–41.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1480–1489. The Association for Computational Linguistics.

Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. 2021. Delving deep into label smoothing. *IEEE Trans. Image Process.*, 30:5984–5996.

Qiaoyong Zhong, Chao Li, Yingying Zhang, H Sun, S Yang, D Xie, and S Pu. 2016. Towards good practices for recognition & detection. In *CVPR workshops IRSVLC2016*.

Appendix

A Selection Criteria and Dataset

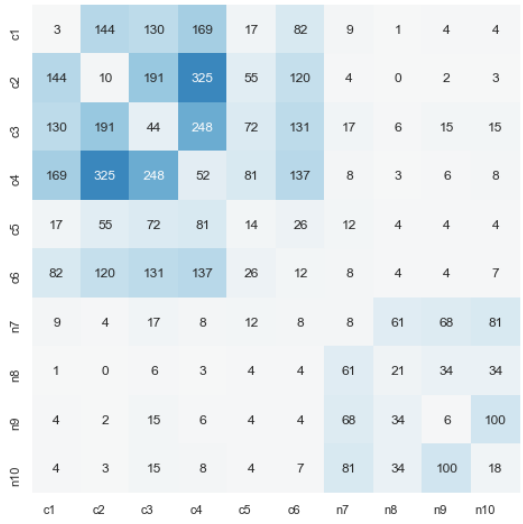


Figure 4: The co-occurrence matrix $\boldsymbol{A}_{\kappa\times\kappa}$ of OUV selection criteria in all UNESCO WH sites.

Selection Criteria Definitions For any site to be inscribed in the World Heritage List, it must satisfy at least one of the ten Outstanding Universal Value (OUV) selection criteria and meet the conditions of integrity and/or authenticity.

However, it is to be stressed that the definition of the selection criteria shown in Table 4 is regularly revised by the World Heritage Committee to reflect the evolution of World Heritage (WH)

OUV	Focus	Definition	Total
C1	Masterpiece	*To represent a masterpiece of human creative genius*;	254
C2	Values /Influences	*To exhibit an important interchange of human values, over a span of time or within a cultural area of the world, on developments in architecture or technology, monumental arts, town-planning or landscape design*;	449
C3	Testimony	*To bear a unique or at least exceptional testimony to a cultural tradition or to a civilization which is living or which has disappeared*;	466
C4	Typology	*To be an outstanding example of a type of building, architectural or technological ensemble or landscape which illustrates (a) significant stage(s) in human history*;	597
C5	Land-Use	*To be an outstanding example of a traditional human settlement, land-use, or sea-use which is representative of a culture (or cultures), or human interaction with the environment especially when it has become vulnerable under the impact of irreversible change*;	157
C6	Associations	*To be directly or tangibly associated with events or living traditions, with ideas, or with beliefs, with artistic and literary works of outstanding universal significance*;	246
N7	Natural Beauty	*To contain superlative natural phenomena or areas of exceptional natural beauty and aesthetic importance*;	146
N8	Geological Process	*To be outstanding examples representing major stages of earth's history, including the record of life, significant on-going geological processes in the development of landforms, or significant geomorphic or physiographic features*;	93
N9	Ecological Process	*To be outstanding examples representing significant on-going ecological and biological processes in the evolution and development of terrestrial, fresh water, coastal and marine ecosystems and communities of plants and animals*;	128
N10	Bio-diversity	*To contain the most important and significant natural habitats for in-situ conservation of biological diversity, including those containing threatened species of outstanding universal value from the point of view of science or conservation.*	156

Table 4: The definition for each UNESCO World Heritage OUV selection criterion and its main topic according to UNESCO (2008), Jokilehto (2008), and Bai et al. (2021a). The last column shows the total number a criterion is justified with a WH site either uniquely or together with other criteria until 2019.

N	Count	Proportion	Example
1	188	16.75%	Sydney Opera House
2	468	41.71%	Babylon
3	304	27,09%	City of Bath
4	103	9.18%	Yellowstone National Park
5	34	3.0%	Acropolis, Athens
6	4	0.36%	Venice and its Lagoon
7	2	0.18%	Mount Taishan

Table 5: The distribution of the total number of selection criteria $\sum_{k=1}^{\kappa} \gamma_{i,k}$ a site is justified with.

Attribute	Symbol	Data
data	$\boldsymbol{x}_{i,j,k}$	*the counter reformation of the late* < NUM > *th century led to a flowering in the creation of calvaries in europe*
single label	k	Criterion (iv)
sentence label	$\boldsymbol{y}_{i,j,k}$	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]
parental label	$\boldsymbol{\gamma}_i$	[0, 1, 0, 1, 0, 0, 0, 0, 0, 0, .2]
length	$\lvert\boldsymbol{x}_{i,j,k}\rvert$	18 (tokens)
site ID	i	905
data split		train

Table 6: An example of data sample.

itself[7]. For example, cultural (criteria i-vi, also denoted as C1-C6) and natural (criteria vii-x, also denoted as N7-N10) OUV used to be justified apart as two sets. Since 2004, the two sets are combined. Although WH sites are usually justified with OUV from one category (cultural or natural), within the domain of mix heritage and cultural landscape, OUV from both categories can co-occur in one site (e.g., Mount Tai has all first seven criteria).

Association between Criteria Among all the 1121 sites inscribed in the World Heritage List up to 2019, only 188 are justified with only one criterion. The distribution of the total number of criteria justified for each site (i.e., $\sum_{k=1}^{\kappa} \gamma_{i,k}$) is shown in Table 5. This is an indication on the extend of how the problem characterizes a multi-label classification nature. It is also the rationale behind the choice of k = 3 for the evaluation metrics Top-k Accuracy and Top-k Match, as 85.5% of sites are justified with no more than 3 criteria. Regardless of the number of co-justified criteria for each site, the co-occurrence matrix $\boldsymbol{A}_{\kappa\times\kappa}$ of all selection criteria is shown in Figure 4. The row-normalized $\boldsymbol{A}_{\kappa\times\kappa}$ becomes the source of the criterion-specific non-negative vectors $\boldsymbol{\mu}_k$ of the **prior** variant of Label Smoothing (LS), as is discussed in Section 4.1. The criteria from the same category are co-justified more often, while criteria (ii-iv), (iii-iv), and (ii-iii) are the most frequently co-occurred pairs.

Dataset Example A data point concerning the WH site "*Kalwaria Zebrzydowska: the Mannerist*

[7] http://whc.unesco.org/en/criteria/

Architectural and Park Landscape Complex and Pilgrimage Park" in Poland justified with Criteria (ii) and (iv) is shown in Table 6, with the attributes of text data $\boldsymbol{x}_{i,j,k}$, sentence label as discrete index k, sentence label as one-hot vector $\boldsymbol{y}_{i,j,k}$ (appended with 0 in the end for the class "Others"), parental label as vector $\boldsymbol{\gamma}_i$ (appended with 0.2), sample length $|\boldsymbol{x}_{i,j,k}|$ in terms of the number of tokens, index of parental WH site i, and the data split.

B Proof of the Equivalence

Here we will show that the Vanilla Label Smoothing (LS) defined in Equations 4 and 5 is equivalent to the original LS assigning noise to all classes.

Proof. The LS defined in Szegedy et al. (2016):

$$q'(k) = (1-\epsilon)\boldsymbol{\delta}_{k,y} + \frac{\epsilon}{K} \tag{7}$$

could be rewritten as following to fit the context of mathematical notations in this paper:

$$\boldsymbol{y}^O_{i,j,k} = (1-\epsilon)\boldsymbol{y}_{i,j,k} + \frac{\epsilon}{K}\mathbf{1}, \tag{8}$$

where $\boldsymbol{y}_{i,j,k}$ is a one-hot vector of "ground-truth" label, K is the total number of classes (instead of $\kappa+1$ in the paper for brevity and generality), ϵ is smoothing parameter as scalar, and $\mathbf{1}$ is a vector of 1s of size $K \times 1$.

On the other hand, the Vanilla LS proposed in this paper could be written as:

$$\boldsymbol{y}^V_{i,j,k} = \mathbf{f}(\boldsymbol{y}_{i,j,k} + \alpha\mathbf{1}) = \frac{e^{\boldsymbol{y}_{i,j,k}+\alpha\mathbf{1}} - \mathbf{1}}{e^{(\boldsymbol{y}_{i,j,k}+\alpha\mathbf{1})^T}\mathbf{1} - K}. \tag{9}$$

We will show that when

$$\epsilon = \frac{(e^\alpha - 1)K}{e^{1+\alpha} + (K-1)e^\alpha - K}, \tag{10}$$

the vectors in Equations 8 and 9 are the same.

First, it is trivial that both the vectors are with the same shape of $\boldsymbol{y}_{i,j,k}$, i.e., $K \times 1$, and that the sums of all entries in both vectors are 1; e.g., observe that the denominator of the right-hand side of Equation 9 is equal to the vectorised summation of the values of the nominator.

Second, we assume, without loss of generality, that the "ground-truth" of the one-hot vector $\boldsymbol{y}_{i,j,k}$ is at its first entry, which means that $\boldsymbol{y}_{i,j,k} = [1, 0, ..., 0]_{K\times 1}$. Then both vectors could be rewritten as:

$$\boldsymbol{y}^O_{i,j,k} = \left[1 - \epsilon + \frac{\epsilon}{K}, \frac{\epsilon}{K}, ..., \frac{\epsilon}{K}\right]_{K\times 1}, \tag{11}$$

$$\boldsymbol{y}^V_{i,j,k} = \left[\frac{e^{1+\alpha}-1}{S}, \frac{e^\alpha - 1}{S}, ..., \frac{e^\alpha - 1}{S}\right]_{K\times 1}, \tag{12}$$

where $S := e^{1+\alpha} + (K-1)e^\alpha - K$.

Substituting Equation 10 into the entries in Equation 11, the first entry could be rewritten as $1 - \epsilon + \frac{\epsilon}{K} = 1 - \frac{(e^\alpha-1)K}{S} + \frac{e^\alpha-1}{S} = \frac{S-(e^\alpha-1)K+e^\alpha-1}{S} = \frac{e^{1+\alpha}+(K-1)e^\alpha-K-Ke^\alpha+K+e^\alpha-1}{S} = \frac{e^{1+\alpha}-1}{S}$. And the other entries could be rewritten as $\frac{\epsilon}{K} = \frac{e^\alpha-1}{S}$. Both types of entries are exactly the same as the ones shown in Equation 12.

Last, we will show that ϵ has a one-to-one relation with α based on Equation 10 when $\alpha \geq 0$. The partial derivative of ϵ with respect to α:

$$\frac{\partial\epsilon}{\partial\alpha} = \frac{Ke^\alpha(e-1)}{(e^{1+\alpha} + (K-1)e^\alpha - K)^2} > 0 \tag{13}$$

is non-negative, suggesting that the function is monotonic. Furthermore, $\epsilon = 0$ when $\alpha = 0$, and $\lim_{\alpha\to+\infty}\epsilon = \lim_{\alpha\to+\infty}\frac{K}{\frac{e^\alpha(e-1)}{e^\alpha-1}+K} = \frac{K}{e-1+K} > 0$ when $\alpha \to +\infty$, suggesting that it is incremental. This means that a unique $\epsilon \in \left[0, \frac{K}{e-1+K}\right)$ always exists for any non-negative α and *vice versa*. □

C Model Implementation Detail

For all baselines, Adam (Kingma and Ba, 2015) is used as the optimizer with L2 regularization. Hyperparameter tuning is conducted as grid-search within a small range for each one being searched (and/or selected according to common experience if not mentioned), based on the top-k accuracy on validation split with an early-stopping criterion of 5 epochs, if not explicitly mentioned below. The models are implemented in PyTorch (Rao and McMahan, 2019) and experiments are performed on NVIDIA Tesla P100 GPU (N-gram, GRU+Attn, BERT) and Intel Core i7-8850H CPU (BoE, ULMFiT), respectively.

N-gram The N-gram model used the TfidfVectorizer from Scikit-learn Python library to get an embedding vector of all 1-grams and 2-grams in the sample that appeared at least twice in the vocabulary. The embedding vectors are then fed in a 2-layer Multi-layer Perceptron (MLP) to get the model prediction.

Hyperparameter tuning is performed on the size of the MLP hidden layer in {50, 100, 150, 200}, batch size in {64, 128, 256}, L2 in {0, 1e-5. 1e-4}, and dropout rate in {0.1, 0.2, 0.5} with 108

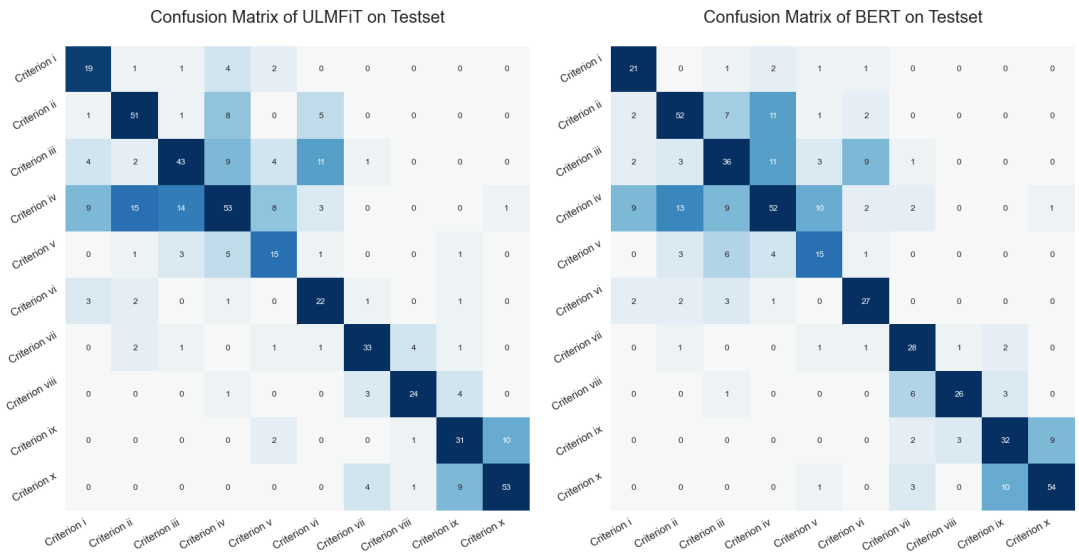


Figure 5: The confusion matrices of ULMFiT and BERT on test split.

configurations. The best configuration applied in later experiments of Label Smoothing (LS) has a hidden dimension of 200, batch size of 128, L2 of 1e-5, learning rate of 2e-4, and dropout rate of 0.5.

BoE The Bag-of-Embedding (BoE) model used the GloVe-6B-300d vectors[8] as initial embeddings, which are set to be tunable during training. Only words that have a higher frequency than a threshold in the full dataset will be kept, while the others will be transformed to a special $<$ UNK $>$ token. The word embeddings of all words in the sentence is averaged before being fed to a 2-layer MLP.

Hyperparameter tuning is performed on the size of the MLP hidden layer in {50, 100, 150, 200}, batch size in {64, 128, 256}, and frequency threshold in {1, 3, 5} with 36 configurations. The best model has a hidden dimension of 200, batch size of 64, cut-off frequency of 1, L2 of 1e-5, learning rate of 5e-4, and dropout rate of 0.1.

GRU+Attn The GRU+Attn model also used the GloVe-6B-300d as embeddings, which are frozen during the training. The embedding sequence is then fed into a GRU network. Word-level attention (Yang et al., 2016) is applied to compute the sentence vector by a learned word context vector and the last hidden state of the GRU. The sentence vector is fed to a 1-layer feed-forward network for the output of the model.

Hyperparameter tuning is performed on the size of the hidden layer in GRU in {64, 128, 256}, whether or not to use bi-directional GRU, batch size in {64, 128, 256}, L2 in {0, 1e-5, 1e-4}, learning rate in {1e-3. 5e-4. 2e-4}, and dropout rate in {0, 0.1, 0.2, 0.5} with 648 configurations. The best model is a uni-dimensional GRU with hidden dimension of 128, batch size of 256, L2 of 1e-5, learning rate of 1e-3, and dropout rate of 0.1.

ULMFiT The ULMFiT model employs the idea of Universal Language Model Fine-tuning from a general-domain pretrained language model on Wikitext-103 with AWD-LSTM architecture (Howard and Ruder, 2018). A domain-specific language model is then fine-tuned with the full UNESCO WHL dataset including SD using fastai API (Howard and Gugger, 2020). One epoch is trained with a learning rate of 1e-2, with only the last layer unfrozen, reaching a perplexity of 46.71. Then the entire model is unfrozen and further trained for 10 epochs, with a learning rate of 1e-3, obtaining a fine-tuned WH domain-specific language model reaching a 30.78 perplexity. Some examples of the language model at this step are shown here, starting with the given phrases marked in bold:

> ***This site is unique because*** *it is the only example of a complex of karst complexes that is clearly recognised as being of outstanding universal value. The island of zanzibar has been inscribed as a world heritage site in <num>. The inscriptions, which bear witness to the civilisation of...*
>
> ***This architecture has a special layout****, especially in the form of the body of the building. The planet's primary feature is the addition of the ideal island, which lies at an elevation of <num>m above the sea floor, and is home to some <num>...*

[8] https://nlp.stanford.edu/projects/glove/

Performance	N-gram	BoE	GRU+Attn	ULMFiT	BERT
Infrastructure	GPU	CPU	GPU	CPU	GPU×4
Training Time per Item (s)	0.34	0.18	0.03	2.53	0.54
Training Time per Epoch (s)	12.69	3.18	1.97	213.61*	46.20
Early-Stopping Criteria	5	5	5	3	10
Training Epochs	32	20	15	7**	10
Trainable Parameters (M)	3.82	1.88	0.18	24.55	109.49
Inference Time per Item (s)	0.0031	0.0007	0.2245	0.0589	0.5542
Inference Time for SD (s)	6.92	1.44	4.44	151.75	1598.06

*1180.20 during language model fine-tuning.
**11 during language model fine-tuning.

Table 7: The model performance in terms of resource occupancy and inference time. The inference is conducted on Intel Core i7-8850H CPU. *Inference time per Item* shows the average time the model uses to make a prediction on one sentence. And *Inference time for SD* shows the total time the model needs to fully process and predict the independent Short Description (SD) test set.

The encoder of the fine-tuned language model is loaded in PyTorch followed by a Pooling Linear Classifier[9] for classifier fine-tuning. Gradual unfreezing is applied in a simplified manner to prevent catastrophic forgetting: 1) for the 1st epoch, only the decoder is unfrozen and trained with a learning rate of 2e-2; 2) for the 2nd to 4th epoch, one more layer is unfrozen each time and trained with a learning rate of 1e-2, 1e-3, and 1e-4, respectively; 3) from the 5th epoch onward, the full model is unfrozen and trained with a learning rate of 2e-5. An early-stopping criterion of 3 is applied.

No extensive hyperparameter tuning is performed since: 1) tuning ULMFiT is expensive on CPU; 2) the hyperparameter configuration from experience suggested by Howard and Gugger (2020) and Howard and Ruder (2018) already performs reasonably well; 3) the purpose of this study is not necessarily finding the best hyperparameter. The final model uses batch size of 64, L2 of 1e-5, and the default dropout rate for the decoder.

BERT The BERT model uses the uncased base model using The Transformers library (Wolf et al., 2020). The pooler output processed from the last hidden-state of the [CLS] token during pretraining is fed into a 1-layer feed-forward network to fine-tune the classifier (Sun et al., 2019). An early-stopping criterion of 10 is applied.

Hyperparameter tuning is performed on the batch size in {16, 24, 48, 64}, L2 in {0, 1e-5, 1e-4}, and dropout rate in {0, 0.1, 0.2} with 36 configurations. The best model uses batch size of 64, L2 of 1e-4, learning rate of 2e-5, and dropout rate of 0.2.

LS Configuration Tuning A single random seed 1337 is used for hyperparameter tuning. Afterwards, ten random seeds in {0, 1, 2, 42, 100, 233, 1024, 1337, 2333, 4399} are used to tune the LS configuration with $\alpha \in \{0, 0.01, 0.05, 0.1, 0.2, 0.5, 1\}$ for all three variants. The best LS configuration is selected based on the sum of the lower bound of 95% confidence interval on both top-1 and top-k accuracy. The best LS configuration is then used to evaluate the model performance on single seed 1337. The total runs on each baseline are, therefore, the sum of the number of hyperparameter configurations and random seeds experiments (which is 210).

[9] https://fastai1.fast.ai/text.models.html

D Extended Model Performance

Resource and Time Table 7 shows some further information on the model performance in terms of training resource utilization, model size, and inference time. Training processes are conducted on CPU or GPU, respectively, while inference is fully conducted with CPU.

It can be noted that the best-performing models ULMFiT and BERT also consume the most resources, in terms of training time and infrastructure usage, and have the largest model sizes. Though most time-consuming during training, ULMFiT takes a remarkably short time for inference on CPU compared to BERT. This suggests that ULMFiT might be an optimal choice for further development and application when time is a critical matter.

Confusion Matrices The confusion matrices of the best-performing ULMFiT and BERT models on the test split are shown in Figure 5. It can be seen that certain criteria are easily confused as the others, such as sentences with a "ground-truth" label of criterion (iv) can be confused as criteria (i), (ii), and (iii), and *vice versa*; while criterion (iii) might be confused easily as criterion (vi), but **NOT** *vice*

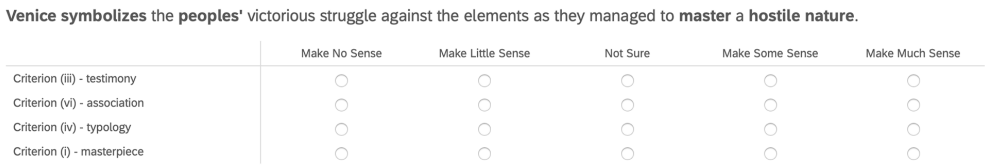


Figure 6: The evaluation interface on Qualtrics.

versa. This complex association relationship is extensively discussed in Bai et al. (2021a).

E Expert Evaluation Details

Materials The materials about the WH site "*Venice and Its Lagoon*" for expert evaluation were harvested from three data sources: 1) all 14 sentences from **Justification for Criteria** section of Statements of OUV (SOUV), where each sentence has one "ground-truth" sentence label and a parental site label of Venice, which is also within the data $\boldsymbol{X}_i$ used during model training and testing; 2) all 13 sentences from **Brief Synthesis** section of SOUV, where sentences only have the same multi-label parental label of Venice, which is similar with the SD test data $\boldsymbol{S}_i$ used for generalization test; 3) **Social Media** data sampled from a total of 1687 social media posts where a textual description is written, collected from Flickr in the region of Venice with a resolution of 5km using Flickr API[10]. Among the 1687 social media posts, there are 820 unique textual descriptions in English. By splitting the unique posts into sentences, removing HTML symbols, and filtering out the texts about camera parameters, image formats, and advertisements, 1132 sentences were obtained. The 1132 sentences were fed into the trained BERT and ULMFiT models. The sentences were further filtered based on the predictions: 1) the total confidence scores of top-3 predictions need to be larger than 0.8 by both models; 2) the Intersection over Union of top-3 predictions by two models needs to be larger than 0.5 (i.e., maximum one different predicted class). As a result, 388 Social Media sentences that potentially convey OUV-related information were obtained. Furthermore, 29 sentences were randomly sampled from those 388 for the expert evaluation.

Survey Design Each of the 56 sentences was fed into BERT and ULMFiT models to obtain the predictions and confidence scores. The predicted selection criteria with the highest confidence scores by both models were considered as the **top-1 predictions**. Two other criteria within the top-3 classes predicted by both models with relatively high confidence scores were considered as the **top-3 predictions** for the survey. Another random cultural criterion that was not predicted by any model to be top-3 classes was considered as the **negative class** for each sentence. Criteria for natural heritage were not sampled as negative classes as they are not easily confused with the positive cultural ones. As a result, each sentence got **four** criteria to be evaluated. All four criteria were presented in a random order for each sentence, asking for an evaluation about the relevance of the sentence conveying the criterion on a 5-point Likert scale (from "*5: make much sense*", to "*1: make no sense*"). The "important" words with higher attention weights in the GRU+Attn model were highlighted in bold. An example of such evaluation on Qualtrics platform is shown in Figure 6. The sentences from the three data sources were grouped in four separate sessions, while the social media data were split into two sessions. The session of "*justification for criteria*" were always presented first during evaluation, also as a practice for the experts. The other three sessions were presented in a randomized order to prevent systematic errors caused by impatience or tiredness. Additional questions about the familiarity for heritage value identification, familiarity about Venice, confidence of evaluation, usefulness of highlighted words, and overall enjoyment and difficulty of the exercise were respectively raised before and after the evaluation, also with 5-point Likert scale. Note the number of samples involved in the in-depth expert evaluation is relatively small, which is not uncommon in qualitative validation. Moreover, we plan to conduct online non-expert human evaluation in follow-up studies, which could involve more participants with larger sample sentences. It would, however, serve a different purpose than the expert evaluation presented.

General Analyses The evaluations took 55.10 ± 20.74 minutes to finish. The eight experts are all

[10] https://pypi.org/project/flickrapi/

Data Source	Type-1	Type-2	M_1	M_2	n_1	n_2	U value	p value
Justification of Criteria	top-1 prediction	top-3 prediction	5	2	120	240	8157.0***	<0.001
	top-1 prediction	negative class	5	2	120	120	3161.0***	<0.001
	top-3 prediction	negative class	2	2	240	120	12638.0*	0.026
Brief Synthesis	top-1 prediction	top-3 prediction	4	2	96	192	6256.0***	<0.001
	top-1 prediction	negative class	4	2	96	96	2401.5***	<0.001
	top-3 prediction	negative class	2	2	192	96	7603.5**	0.006
Social Media	top-1 prediction	top-3 prediction	3	2	232	464	40629.0***	<0.001
	top-1 prediction	negative class	2	1	232	232	13784.5***	<0.001
	top-3 prediction	negative class	2	1	464	232	39284.5***	<0.001

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Table 8: The results of post-hoc Mann-Whitney U tests for the three types of labels within each data source. The medians (M) and counts (n) of each type are given together with the statistics from U tests.

Text	Criteria	Source	Type	BERT	ULMFiT	Expert Ratings
With the unusualness of an archaeological site which still breathes life, Venice bears testimony unto itself.	iii	justification	top-1	0.744	0.825	5,5,5,3,5,5,4,5
Human interventions show high technical and creative skills in the realization of the hydraulic and architectural works in the lagoon area.	i	synthesis	top-1	0.607	0.590	4,5,5,1,4,4,2,5
In 1952, the station was finalized on a design by the architect Paul Perilli.	i	social media	top-1	0.757	0.529	5,4,1,1,1,3,1,1

Table 9: Some example ratings on sentence-criterion relevance by human experts. The confidence scores by the computer models BERT and ULMFiT are also given.

very familiar with the concept of OUV (4.38 ± 0.70) and the heritage values and attributes identification (4.75 ± 0.43), while not all are familiar with OUV justification (3.00 ± 1.50), nor with the cultural heritage in Venice (3.00 ± 1.41). The experts agree that the exercise in the evaluation was very hard (4.13 ± 0.93) and not so enjoyable (2.63 ± 1.32). They are more confident with identifying irrelevant sentence-criterion pairs (3.88 ± 0.78) than evaluating the relevant ones (3.00 ± 1.12). These show that the results of the expert evaluation are sufficiently reliable, that the heritage experts are cautious and critical of the process, that OUV justification is a difficult task even for experts as it is time-consuming and knowledge-demanding, and that a computational model is urgently needed to automate the classification if to be applied with massive social media data. The experts are not fully convinced that the highlighted words helped them with the justification process (2.88 ± 1.05), since the words provide both relevant information (3.13 ± 1.27) and irrelevant information (4.38 ± 0.70). This suggests that the explainability of the model using GRU+Attn attention mechanism needs further development.

Evaluation Results Since the expert evaluations are in ordinal scales, non-parametric statistical tests, including Kruskal-Wallis H tests (analogous to ANOVA) and Mann-Whitney U tests (analogous to $t-$ test), are conducted. The statistic analyses are performed with Scipy[11] and Statsmodels[12] libraries. Kruskal-Wallis H tests show significant differences among the three types of criteria labels for all data sources, including for "*justification of criteria*" [$H(2) = 68.412, p < 0.001$], for "*brief synthesis*" [$H(2) = 40.351, p < 0.001$], and for "*social media*" [$H(2) = 102.321, p < 0.001$]. Post-hoc Mann-Whitney tests were used to compare all pairs of groups, as is shown in Table 8. The all-significant results of U tests show that the human experts gave significantly higher ratings to top-1 predictions than top-3 predictions, and to top-3 predictions than negative classes. In other words, the human experts and computer models are consistently similar in differentiating the positive and negative criteria for the sentences concerning their relevance. Some exemplary ratings of the experts and model predictions are given in Table 9. It shows that the opinion of experts easily diverge, that some experts seem to be rather cautious during evaluation and rate lower for the social media data, and that it is difficult even for human experts to reach an agreement without further discussion.

[11] https://docs.scipy.org/doc/scipy/reference/stats.html
[12] https://github.com/statsmodels/statsmodels