I Don't Have That Much Data! Reusing User Behavior Models for Websites from Different Domains

Maxim Bakaev[1](✉), Maximilian Speicher[2], Sebastian Heil[3], and Martin Gaedke[3]

[1] Novosibirsk State Technical University, Novosibirsk, Russia
bakaev@corp.nstu.ru
[2] C&A Europe, Düsseldorf, Germany
maximilian.speicher@canda.com
[3] Technische Universität Chemnitz, Chemnitz, Germany
{sebastian.heil,martin.gaedke}@informatik.tu-chemnitz.de

Abstract. User behavior models see increased usage in automated evaluation and design of user interfaces (UIs). Obtaining training data for the models is costly, since it generally requires the involvement of human subjects. For interaction's subjective quality parameters, like aesthetic impressions, it is even inevitable. In our paper, we study applicability of trained user behavior models between different domains of websites. We collected subjective assessments of Aesthetics, Complexity and Orderliness from 137 human participants for more than 3000 homepages from 7 domains, and used them to train 21 artificial neural network (ANN) models. The input neurons were 32 quantitative metrics obtained via computer vision-based analysis of the homepages screenshots. Then, we tested how well each ANN model can predict subjective assessments for websites from other domains, and correlated the changes in prediction accuracies with the pairwise distances between the domains. We found that the Complexity scale was rather domain-independent, whereas "foreign-domain" models for Aesthetics and Orderliness had on average greater prediction errors for other domains, by 60% and 45%, respectively. The results of our study provide web designers and engineers with a first framework to assess the reusability and difference in prediction accuracy of the models, for more informed decisions.

Keywords: Web design · User experience · Machine learning · Training data

1 Introduction

Even though the thorough evaluation of user interfaces (UIs) became widely popular already in the early 90 s (e.g. [1]), it has not ceased to be a hot topic. User interfaces are becoming increasingly complex and sophisticated, which a visit to the Internet Archive's Wayback Machine easily proves. This, however, also raises the complexity of setting up and analyzing corresponding assessments. Besides, certain methods for evaluation are

M. Bielikova et al. (Eds.): ICWE 2020, LNCS 12128, pp. 146–162, 2020.
https://doi.org/10.1007/978-3-030-50578-3_11

often considered costly and inefficient in the industry. Especially the ability to carry out user tests is limited by the available resources ([2, 3, p. 180]). In many cases, this leads to the application of simpler and faster methods, like A/B testing, which is, however, not perfectly suited for determining qualitative aspects such as the usability or user experience of an interface [4]. One alternative to traditional user testing that has been repeatedly suggested in the literature is to employ models that predict subjective quality parameters – like usability – from (a) static [5] or (b) visual [6] properties of a user interface, or (c) from tracked interactions [7, 8].

Why is (efficient) evaluation of UIs important? With today's plethora of available websites and apps, it is crucial to properly test them in order to gain user acceptance. Users spend most of their time on other websites and disapprove of usability and user experience flaws [9]. Now, the more efficient an evaluation method is, the fewer resources are required, both, time- and money-wise, which leads to easier stakeholder buy-in, particularly in industry settings (yet, effectiveness must not be traded for efficiency). On top, the more user-friendly the interface and the more resource-efficient its creation process, the more sustainable it becomes, which is a consideration becoming increasingly important nowadays [3].

What are the advantages of user behavior models? Leveraging user behavior models to predict subjective interaction quality parameters is a promising approach to effective evaluation that uses fewer resources than traditional methods. First, libraries such as MOA and `scikit-learn` are widely available and make training machine-learning models relatively easy. Second, once such models have been trained, they can be applied as many times as wanted, without lengthy testing sessions and the involvement of real users.

So, what is the problem? Even though user behavior models need to be trained only once, obtaining high-quality training data is often a problem and huge amounts of data might be needed to obtain well-working models (e.g., ~23 GB of raw tracking data in the case of [10]). Therefore, it would be worthwhile to reuse existing models for as many UIs as possible, hence reducing the need for collecting hard-to-obtain training data. Yet, Speicher et al. [8] hypothesized that such models are only applicable within clusters of very similarly structured websites (since user interactions seem to be very sensitive to low-level details of an interface). This is the very question we intend to investigate in this paper.

Based on a set of features that are potentially more robust than user interactions with an interface, we build artificial neural network models (ANNs) for websites from a certain domain and investigate how accurately they can predict subjective assessments for different domains. Overall, this paper makes the following contributions:

1. We train ANN models for 7 different domains of websites, based on subjective quality assessments from 137 users.
2. We show that these ANN models can to a certain degree predict subjective interaction quality parameters of websites from other domains.
3. We show that there is a connection between prediction accuracy and distance between website domains, and we propose the corresponding distance measure.

In Sect. 2, we overview related work, while in Sect. 3 we describe our experimental study. In Sect. 4, we analyze the data and propose the regression model that relates the models' prediction accuracies and the distances between the domains.

2 Related Work

User behavior models are considered effective in representing research results in HCI and a solid basis for software tools that support UI designers [11], particularly in the evaluation of web UI prototypes and designs. Generally, they predict an interaction quality parameter, based on two sets of input: target user characteristics and UI representation. Interaction models are built for particular tasks (more rarely, task specification can be part of the input), whereas user experience models, which are a rather novel research topic, are more inclined towards reflecting cognitive processes and neural structures.

Despite the increasing recognition, their use in practical Web Engineering so far remains limited, for which we see two main reasons. First, building and training a new model for a project context imposes high skill requirements: a software development team rarely includes a computer scientist, a cognitive psychologist, etc. Second, even though more and more models are made available, it remains unclear how granular the input needs to be, i.e., how much re-training is needed for another group of users or a changed UI.

The endeavor undertaken in this paper relates to the topic of transfer learning, which in practical ML sometimes is also called pre-training. According to [12], transfer learning takes place when the knowledge contained in an existing model for a task T1 in a given domain D1 supports the learning of a not yet existing model for a task T2 from a domain D2, whereas $D1 \neq D2$. While also $T1 \neq T2$, the tasks should be related [13], which is the case for predicting quality parameters for websites from different domains. However, our approach is more radical in the sense that we intend to directly apply the model for D1 to D2, rather than to support the learning of a new model. This corresponds to skipping the second step (fine training) in the utilization of pre-trained user behavior models, which means a trade-off: saving on training data, but losing on the evaluation model's accuracy.

2.1 AI in UI Evaluation and Design

Classifiers for predicting quality parameters of UIs used in existing research include Random Forests, Naïve Bayes, ANNs, and non-ML-based models, among others. For instance, [10] collected a number of user interactions (mostly mouse and scrolling behavior) on search engine results pages and trained models that were able to predict the relevance of search results better than a generative state-of-the-art approach. They used Random Forests as the classifier of their choice. However, their solution is restricted to a very specific type of webpage and a single quality parameter. In [8], they employ a similar, but extended approach by tracking a similar (but larger) set of user interactions and learning several models in parallel to predict 7 different usability parameters (according to the INUTT instrument). Their classifier of choice is an incremental version of Naïve Bayes.

Such models in the context of UI evaluation and design have certain advantages and disadvantages. On the one hand, it is very cumbersome for a developer or researcher to manually identify patterns in website structure or user interactions that correlate with certain quality parameters (such as, "Users that change scrolling direction at least twice rate a website as more confusing"). In [8], the authors have tried this, but the correlations they found are mostly rather low ($r < .3$) and derived from the models they learned. Discovering these connections is much easier for machine learning classifiers. On the other hand, the models trained by classifiers are mostly not human-interpretable and the models themselves remain a black box.

The work of Grigera et al. [7] builds on a non-ML approach. They identified patterns of user behavior that hint at certain "usability smells", e.g., "user clicks a link and returns shortly after" → misleading link, and implemented a finder for each smell. This is a robust, easily understandable approach that is applicable to a large range of websites, but limited by existing knowledge about user behavior, not easily adjustable, and might prevent the detection of new patterns beyond the perception of the developers. None of the research described above aims at applying their learned models to user interfaces from a different domain. In [8], they tried but concluded that if it is possible, it is at best possible for interfaces that are structurally very similar. The approach in [7] is applicable to a range of websites from different domains, but not based on machine-learning approaches. Therefore, a comparison with our work is out of scope in this regard.

Indeed, [8] partly inspired the topic of this paper since we hypothesize that with different, more robust input attributes, applying models across domains of websites could be possible. For this, we orient at [6], since global, visual features of websites are potentially not as prone to differences in structure as user interactions. Their work builds on static visual properties of websites – metrics, as obtained through a screenshot-processing visual analyzer – and ANNs to predict subjective quality assessments (e.g., perceived complexity of a website).

2.2 The UI Visual Analysis Tools

The more traditional approach for extracting quantitative metrics of UIs is based on the analysis of UI code or model representation. It boasts high performance and accuracy and is particularly suitable for web UIs whose HTML/CSS code is easily available [14]. Code-based analysis is widely used to check compliance with accessibility guidelines and other standards and recommendations but is less suited for the assessment of such a subjective thing as user experience. On the other hand, the increasingly popular UI vision-based analysis, which is based on image recognition techniques, generally deals with the screenshot of a webpage as rendered in a browser. The main advantage of this UI "visual analysis" approach is that it assesses the UI as the target user witnesses it, so it is naturally good at considering layouts, spatial properties of UI elements, graphical content, etc. For instance, in [15], the authors perform automated data extraction from images and make use of Gestalt principles of human visual perception – this understandably would be highly problematic to do with code analysis. At the same time, the disadvantages of the vision-based approach include computational expensiveness and so far low accuracy for some of the metrics.

In view of the abundance and diversity of metrics proposed by various researchers in the rapidly developing metrics-based UI analysis field, we have previously developed the WUI Measurement Integration Platform[1] [14]. It is capable of collecting web UI metrics from different providers and storing them in the common structured representation for further analysis. The platform sends a web UI screenshot or website URI to a remote service using its supported protocol, waits for the output (WebSocket is mostly used) and saves it in the platform's database. Currently, the platform works with the two main UI visual analysis tools, which we also use for the purposes of the current research:

1. Visual Analyzer (VA), developed by Technical University of Chemnitz (Germany) and Novosibirsk State Technical University (Russia) [14];
2. Aalto Interface Metrics service (AIM), by Aalto University (Finland)[2] [16].

The potential number of UI metrics that can be obtained via the vision-based analysis is understandably boundless (the two analyzers that we exploit for this work are just a small portion of the available tools). It thus seems logical to assume that, just like for the general image recognition techniques, artificial neural networks should be an appropriate modeling method.

2.3 ANNs in User Behavior Modeling

Lately, artificial neural networks are back in fashion, with the advent of deep learning in AI. They have reasonable computational cost but are known to be "hungry" for diverse data, so their practical use in the fields where training data are scarce is limited. User behavior modeling is somehow divided with respect to this since the abundance of data varies due to the exact interaction quality parameter being predicted and the corresponding input. Still, the relatively novel recurrent neural networks are used for modeling sequences of user behaviors and are being introduced to predicting behavior on the web. Particularly, in [17], they consider domain switch – where two successive behaviors belong to different domains, which in that work are understood as "service categories in a large-scale web service".

We can speculate that for predicting user experience (as reported by users in their subjective assessments, making the training data quite costly) there is no guarantee that ANNs would be the most accurate method. Or, at least, quite special architectures and approaches would need to be developed for each of the subjective impressions, which has actually been done, e.g., for aesthetics [18]. However, in our current work, we are going to employ rather unsophisticated ANNs, since our goal is to obtain generalizable patterns of the models' applicability across website domains, not propose the most effective prediction model. So, our choice is further reinforced by the known "universal approximator" capability of ANNs, which theoretically makes them more general than, e.g., linear regression (which is, in a way an ANN with a single layer) or certain other methods.

[1] http://va.wuikb.info.
[2] https://interfacemetrics.aalto.fi/.

Since we are not going to perform neural architecture search and tinker with the ANNs' hyper-parameters, this somehow relaxes the requirements towards the amount of training data we would need. A popular "rule of thumb" for linear models is having 10 cases per predictor, so given the number of quantitative metrics the two chosen analyzers can produce (about 35), we would need to collect training data for about 350 websites per domain.

3 Research Hypotheses and Method

The goal of our experimental study was to check the applicability of models across domains of websites. Particularly, we formulated the following hypotheses:

- **Hypothesis 1:** There are significant differences in the quality of ANN user behavior models due to the website's domain.
- **Hypothesis 2:** The difference is smaller for domains that are more similar.

Material. In our experiment, we used screenshots of homepages of websites belonging to one of the 7 distinct domains described in Table 1. The requirements were:

1. The homepage is in English language (or the homepage of the website's international version).
2. Not representing a famous brand/company.
3. Maximum diversity of designs in the set.
4. The nominal number of websites per domain is 500.

Then we used our dedicated tool to automatically make screenshots of webpages located at the collected URIs. Since there is ongoing exploration of whether or not having above-the-fold screenshots is enough for predicting users' impressions, we settled on a compromise: for the universities (Univer) domain, the full webpage was captured, whereas for the other domains the capture was performed only for 1280×960 or 1280×900 pixels. Afterwards, the set of the automatically collected screenshots was manually inspected. The screenshots having some technical problems (most often, a pop-up covering a significant portion of the screen) or not obviously belonging to the specified domain were removed.

To investigate the influence of domain similarity on the applicability of models, we calculated pairwise domain distances for each combination of the 7 domains. For this calculation, each category was mapped onto the DMOZ hierarchy[3] of categories, as shown in Table 1. The domains of Food, Games, Health, News, and Univer have direct equivalents in DMOZ. For Culture and Gov, we identified sets of DMOZ categories that best match the websites of these domains contained in our dataset. As the resulting categories have the same depth in the hierarchy, all nodes to which a domain is mapped have the same distance to other domains.

Figure 1 shows the relevant section of the DMOZ category hierarchy used for domain distance calculation. To calculate the distance between two domains, we use the length

[3] using http://curlie.org/.

Table 1. Homepage domains and their mappings to DMOZ categories.

Domain name	Number of screenshots	Description	DMOZ Categories
Culture	807	Websites of museums, libraries, exhibition centers, other cultural institutions	Reference/Libraries, Reference/Museums
Food	388	Websites dedicated to food, cooking, healthy eating, etc.	Recreation/Food
Games	455	Websites dedicated to computer games	Games
Gov	370	E-government, non-governmental organizations' and foundations' websites	Society/Government, Society/Organizations, Society/Activism
Health	565	Websites dedicated to health, hospitals, pharmacies, medicaments	Health
News	347	Online and offline news editions' websites, news portals	News
Univer	497	Official websites of universities and colleges	Reference/Education/Colleges and Universities
	3429		

of the shortest path between the nodes corresponding to the two domains as per Table 1. This implies identifying the lowest common ancestor (LCA) and adding vertex distances $dist_v$ between both nodes and their LCA:

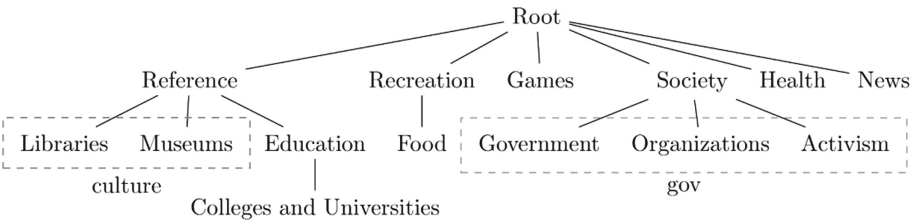

Fig. 1. DMOZ category hierarchy used for domain distance calculation (domains Culture and Gov comprising several DMOZ categories highlighted with boxes).

$$dist(D_1, D_2) = dist_v(dm(D_1), LCA(D_1, D_2)) + dist_v(LCA(D_1, D_2), dm(D_2)) \quad (1)$$

$$dm(D) = \underset{c \in DMOZ(D)}{\arg\min} \; dist_v(c, Root). \tag{2}$$

For distance calculation, domains D are represented by the corresponding DMOZ category that is the highest in the hierarchy, *dm(D)*. Table 2 presents the resulting domain distances for each domain pair.

Table 2. Domain distances based on the proposed measure.

Domain name	Culture	Food	Games	Gov	Health	News	Univer
Culture	0	4	3	4	3	3	3
Food	4	0	3	4	3	3	5
Games	3	3	0	3	2	2	4
Gov	4	4	3	0	3	3	5
Health	3	3	2	3	0	2	4
News	3	3	2	3	2	0	4
Univer	3	5	4	5	4	4	0

Design. The experiment used a within-subject design. The main independent variable was the screenshot domain (*Domain*). Derived independent variables were the pairwise distances between the domains (*Dist*), the 32 metrics for each screenshot (M_i – see the list in Table 3), and the subjects' assessments of each screenshots per the three subjective Likert scales (each ranging from 1, the lowest degree of the characteristic, to 7, the highest degree):

- How visually complex the WUI appears in the screenshot: *Complexity*;
- How aesthetically pleasant the WUI appears: *Aesthetics*;
- How orderly the WUI appears: *Orderliness.*

The dependent variable was the quality of the ANN models in predicting subjective assessments for each domain, as represented by absolute (*MSE*) and relative (MSE_{REL}) mean square errors. MSE_{REL} was calculated as the ratio between the model's *MSE* for the *d*-th domain and the *MSE* for the "native" domain of the model (i.e. the one whose data was used for training the model). Obviously, when *d* was the native domain, $\mathrm{MSE_{REL}} = 100\%$.

Participants and Procedure. In total, there were 137 participants (67 female, 70 male) in the survey, whose ages ranged from 17 to 46 (mean 21.18, SD = 2.68). They were mostly Bachelor's and Master's students of Novosibirsk State Technical University (NSTU), but also students and staff of some other universities, and specialists working in the IT industry. The majority of participants were Russians (89.1%), the rest

Table 3. Derived independent variables (M_i): the metrics for the screenshots.

Group	Metric	Mean	SD
Visual Analyzer (VA)	PNG filesize (in MB)	0.844	0.505
	JPEG 100 filesize (in MB)	0.848	0.453
	No. of UI elements	27.9	22.1
	No. of UI elements' types	4.430	1.279
	Visual complexity index	1248	1220
AIM – Colour Perception	Unique RGB colours	13742	10061
	HSV colours avg Hue	153	152
	HSV colours avg Saturation	0.225	0.140
	HSV colours std Saturation	0.271	0.083
	HSV colours avg Value	0.715	0.170
	HSV colours std Value	0.271	0.070
	HSV spectrum HSV	14157	8927
	HSV spectrum Hue	16396	7975
	HSV spectrum Saturation	16965	3865
	HSV spectrum Value	254.8	5.4
	Hassler Susstrunk dist A	18.0	14.1
	Hassler Susstrunk std A	28.3	14.5
	Hassler Susstrunk dist B	20.2	14.9
	Hassler Susstrunk std B	28.8	13.1
	Hassler Susstrunk dist RGYB	27.7	19.6
	Hassler Susstrunk std RGYB	41.3	17.4
	Hassler Susstrunk colorfulness	49.6	22.3
	Static clusters	3859	2030
	Dynamic CC clusters	693	449
	Dynamic CC avg cluster colors	12.4	1.4
AIM – Perceptual Fluency	Edge congestion	0.252	0.082
	Quadtree Dec balance	0.711	0.246
	Quadtree Dec symmetry	0.564	0.051
	Quadtree Dec equilibrium	1.000	0.002
	Quadtree Dec leaves	2876	2002
	Whitespace	0.340	0.265
	Grid quality (No. of alignment lines)	91.7	61.4

being from Bulgaria, Germany, South Africa, etc. The subjects took part in the experiment voluntary and no random selection was performed. All the participants had normal or corrected to normal vision and reasonable experience with websites.

The participants were provided a link to the online questionnaire that we specially developed for this study. In the survey, the screenshots were selected randomly from the pool of the available ones (with priority given to the ones that had a lower number of evaluations at the moment of selection) and presented to participants successively. The completeness of evaluation, i.e. ranking by all the 3 scales, was mandatory and controlled by the software. The default number of screenshots to be evaluated in each session was set at 100 for most of the participants. The assessment of the screenshots of the Univer domain was performed in a separate session (see in [19]), about 9 months before the other 6 domains, for which the screenshots were mixed into the single pool.

ANN Models. To construct and train ANN models, we used the Colab[4] service freely offered by Google (TensorFlow 1.15.0 environment with Keras, etc.). There was a separate model for each website domain and each subjective impression scale, so there were 21 models in total. In each model, the input values were the 32 metrics for the screenshots of the respective domain, and the single output was the respective subjective assessment.

The most widely used loss function for ANNs that perform a regression task is the mean squared error (MSE), which we will also use to represent the quality of the models. As for the architecture, the goal of our research was not to find the best one but to have comparable models for all domains and quality parameters. Therefore, we adopted the same generic architecture for all the datasets. The main hyper-parameters of the ANNs were set as specified in the following code:

```
def build_model(x):
  model = Sequential()
  model.add(Dense(units = 64, activation = 'relu',\
      input_shape = [len(x.keys())]))
  model.add(Dense(units = 64, activation = 'relu'))
  model.add(Dense(units = 1))
  model.compile(loss = 'mse', \
      optimizer = 'rmsprop', \
      metrics = ['mae', 'mse'])
  return model
```

The normalization of the input data was performed as follows:

```
def norm(x):
  return (x - x.describe().transpose()['mean']) / \
  x.describe().transpose()['std']
```

[4] Our full implementation is available at https://colab.research.google.com/drive/1PFFMkE9vSE7aWBlKdFSLEu0jnSQX7fHw.

For the models' training, the following configuration was specified:

```
TEST_SPLIT = 0.2
VAL_SPLIT = 0.2 # of the remaining 0.8
early_stop = ks.callbacks.EarlyStopping \
    (monitor = 'val_loss', patience = 10, \
    restore_best_weights = True)
```

Since `restore_best_weights` only works if the training was stopped by `EarlyStopping`, the nominal number of training epochs was set to 1000.

4 Results

4.1 Descriptive Statistics

For each of the 3429 screenshots, we attempted to calculate 32 metrics through our WUI Measurement Integration Platform performing in "batch" mode. However, for 345 (10.1%) of the screenshots the VA and AIM services would silently fail to produce some or all metrics (for some reason, *Whitespace* was especially problematic). The screenshots with incomplete metric values had to be excluded from further analysis, even though we do realize that this discard was not random. The metrics' means and standard deviations for the remaining 3084 screenshots are presented in Table 3.

For the 3084 valid screenshots there were 15134 full assessments, so on average 4.9 participants would provide their *Complexity*, *Aesthetics* and *Orderliness* ratings for a screenshot (see in Table 4). For the Univer domain, which was assessed in a separate session, this number was 8.6.

Table 4. Derived independent variables: the subjective impressions scales.

Domain name	Full assessments	Valid screenshots	*Complexity*		*Aesthetics*		*Orderliness*	
			Mean	SD	Mean	SD	Mean	SD
Culture	3280	746 (92.4%)	3.629	0.814	4.243	0.987	4.289	0.895
Food	1585	369 (95.1%)	3.658	0.811	4.699	0.945	4.657	0.865
Games	1570	362 (79.6%)	3.570	0.928	4.244	1.139	4.325	1.000
Gov	1494	346 (93.5%)	3.805	0.820	3.858	0.920	4.140	0.858
Health	2381	541 (95.8%)	3.728	0.789	4.154	0.900	4.399	0.822
News	1445	328 (94.5%)	4.157	0.857	3.795	0.833	4.164	0.817
Univer	3379	392 (78.9%)	3.570	0.636	4.047	0.825	4.417	0.632
	15134	**3084 (89.9%)**	**3.711**	**0.826**	**4.166**	**0.976**	**4.343**	**0.863**

We found significant Kendall's τ_b correlations between *Complexity* and *Aesthetics* ($\tau_{3084} = -0.046$, $p < 0.001$), as well as between *Aesthetics* and *Orderliness* (τ_{3084}

$= 0.520$, $p < 0.001$). *Complexity* and *Orderliness*, however, did not have a significant correlation ($p = 0.359$).

4.2 The ANN Models

Each of the 21 models that we constructed and trained was evaluated with its native testing dataset and 6 foreign ones (i.e., assessments for the screenshots of another domain), thus producing 147 *MSE* values. On average, predictions for the foreign datasets produced greater *MSEs*: +23% for *Complexity*, +60% for *Aesthetics* and +45% for *Orderliness*. T-tests suggested that for *Aesthetics* ($t_{47} = -6.11$, $p < 0.001$) and *Orderliness* ($t_{47} = -2.97$, $p = 0.005$) absolute *MSEs* were significantly different due to the model type (native or foreign). For *Complexity* ($t_{47} = -1.41$, $p = 0.166$), no significant effect was found. Detailed values for the absolute and relative *MSEs* per the subjective evaluation scales are presented in Tables 5, 6 and 7.

Table 5. The results of the *Complexity* models' evaluations (MSE and MSE_{REL}, %).

Testing dataset training dataset	Culture	Food	Games	Gov	Health	News	Univer	Avg. MSE_{REL} for foreign models
Culture	0.820	1.145	1.116	1.308	1.320	1.491	0.937	
		140%	136%	159%	161%	182%	114%	**149%**
Food	1.126	1.231	1.689	1.100	1.260	1.343	1.120	
	91%		137%	89%	102%	109%	91%	**103%**
Games	1.229	1.062	1.346	1.375	1.405	1.549	1.047	
	91%	79%		102%	104%	115%	78%	**95%**
Gov	1.269	0.852	1.356	1.166	1.003	1.612	1.005	
	109%	73%	116%		86%	138%	86%	**102%**
Health	0.945	1.114	1.452	1.068	1.079	1.348	0.889	
	88%	103%	135%	99%		125%	82%	**105%**
News	1.456	1.497	2.778	1.745	1.616	1.497	1.506	
	97%	100%	186%	117%	108%		101%	**118%**
Univer	0.963	0.937	1.510	1.325	1.292	1.536	0.665	
	145%	141%	227%	199%	194%	231%		**190%**
								123%

The correlation between the models' MSEs averaged per domain and the respective domains' dataset sample sizes was not significant ($p = 0.296$). This finding suggests that the models had adequate amounts of training data, which caused no under- or over-fitting. Still, to compare the effects of Domain and of training data, we tried pooling all the domain-specific datasets into a single large one. We trained 3 ANN models (per the

Table 6. The results of the *Aesthetics* models' evaluations (MSE and MSE_{REL}, %).

Testing dataset training dataset	Culture	Food	Games	Gov	Health	News	Univer	Avg. MSE_{REL} for foreign models
Culture	0.958	2.219	1.593	1.261	2.044	1.565	1.146	
		232%	166%	132%	213%	163%	120%	**171%**
Food	1.899	1.009	2.675	1.943	1.943	1.873	2.287	
	188%		265%	193%	193%	186%	227%	**208%**
Games	1.277	2.215	1.401	1.758	2.369	2.963	1.611	
	91%	158%		125%	169%	212%	115%	**145%**
Gov	1.264	2.168	1.605	1.196	0.934	1.099	1.126	
	106%	181%	134%		78%	92%	94%	**114%**
Health	1.680	1.918	1.912	1.525	1.174	1.047	1.395	
	143%	163%	163%	130%		89%	119%	**135%**
News	1.777	2.143	2.299	1.416	1.357	0.866	1.569	
	205%	248%	266%	164%	157%		181%	**203%**
Univer	1.316	2.495	1.493	1.166	1.301	1.084	1.024	
	128%	244%	146%	114%	127%	106%		**144%**
								160%

3 subjective scales) with the same hyper-parameters as we used before and evaluated them with testing sets, in which all the websites were mixed as well. The obtained *MSE* values were −33% for *Complexity*, −18% for *Aesthetics*, and −30% for *Orderliness*, compared to the averaged *MSEs* for the domain-specific models. So, for the two latter scales, the effect of a native training dataset was greater than of more training data.

4.3 Effects of the Domains' Distances

We found significant Pearson correlations between MSE_{REL} and *Dist* for *Aesthetics* ($r_{49} = 0.313$, $p = 0.028$) and *Orderliness* ($r_{49} = 0.343$, $p = 0.016$), but not for *Complexity* ($r_{49} = 0.223$, $p = 0.123$). However, if the native models (distance = 0) were excluded from the consideration, such correlations were no longer significant. But for this set of foreign models, we unexpectedly found significant **negative** correlations between the absolute *MSE* and *Dist*, for *Complexity* ($r_{42} = -0.437$, $p = 0.004$) and *Orderliness* ($r_{42} = -0.347$, $p = 0.024$), though not for *Aesthetics* ($r_{42} = -0.149$, $p = 0.348$). The averaged values for the absolute and relative *MSEs* per *Dist* are presented in Fig. 2 and Fig. 3 respectively.

Further, we performed regression analysis using the backwards selection method (entry 0.05, removal 0.1). We introduced 3 dummy variables with the values {0/1}: $Scale_C$, $Scale_A$ and $Scale_O$ to reflect to which of the subjective impression scales (*Complexity*, *Aesthetics* and *Orderliness*) belongs the model that produced the MSE_{REL}. The

Table 7. The results of the *Orderliness* models' evaluations (MSE and MSE_{REL}, %).

Testing dataset Training dataset	Culture	Food	Games	Gov	Health	News	Univer	Avg. MSE_{REL} for foreign models
Culture	1.048	1.409	1.093	1.452	1.829	1.705	0.848	
		134%	104%	139%	175%	163%	81%	**133%**
Food	1.828	1.311	2.119	2.037	1.937	1.644	1.562	
	139%		162%	155%	148%	125%	119%	**142%**
Games	1.585	2.038	1.506	1.935	2.643	3.174	1.564	
	105%	135%		128%	176%	211%	104%	**143%**
Gov	1.287	1.676	1.210	1.061	1.116	1.385	1.263	
	121%	158%	114%		105%	131%	119%	**125%**
Health	1.557	1.511	1.414	1.465	0.992	1.211	1.312	
	157%	152%	142%	148%		122%	132%	**142%**
News	1.908	1.511	2.306	1.558	1.494	1.073	1.644	
	178%	141%	215%	145%	139%		153%	**162%**
Univer	1.275	1.495	1.456	1.066	1.366	1.862	0.839	
	152%	178%	174%	127%	163%	222%		**169%**
								145%

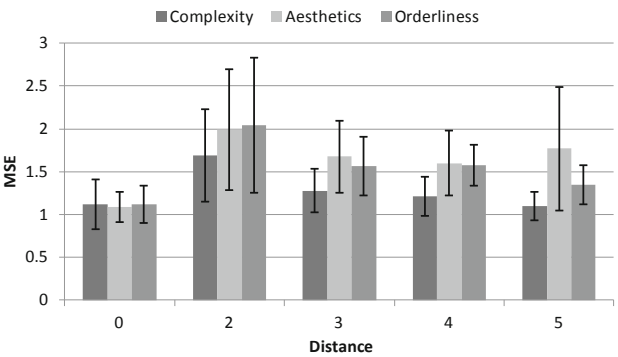

Fig. 2. Averaged absolute MSEs per distances between the domains.

resulting model included the factors of *Dist* ($p < 0.001$, Beta $= 0.274$), $Scale_A$ ($p < 0.001$, Beta $= 0.35$) and $Scale_O$ ($p = 0.02$, Beta $= 0.208$) and was highly significant ($F_{3,143} = 9.62$, $p < 0.001$), although the $R^2 = 0.168$ was rather low:

$$MSE_{REL} = 0.958 + 0.084Dist + 0.318Scale_A + 0.189Scale_O. \quad (3)$$

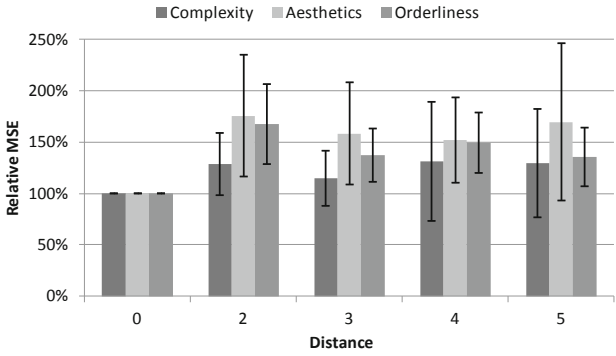

Fig. 3. Averaged relative MSEs per distances between the domains.

5 Discussion

Before we conclude this paper, we intend to have a look at the limitations of the described approach as well as questions that were left open.

First, while Complexity was the only scale without a significantly higher avg. MSE for foreign models, the peculiarity of this scale is further reinforced by its much lower correlation with the other two scales. Hence, we feel the need for more studies in various detailed dimensions of user experience.

Second, there also was an unexpected finding that absolute MSEs had significant negative correlations with the distance between the website domains. We thus believe that the measure of distance that we proposed deserves more exploration, possibly with more domains. Also, rather than relying on topical domains, it would be worthwhile to investigate a more structure-based approach to clustering and distance, e.g., as proposed by Hachenberg and Gottron [20].

Finally, the ANN models that were used for this research are only valid for the specific user groups that provided the subjective assessments. That is, they might not be representative target groups for all of the investigated domains. Hence, for better generalizability of our results, future work should investigate the influence of assessments from different user types on prediction accuracy and the correlations above.

6 Conclusions

In this paper, we sought to apply reuse, which has proven to be rather efficient in software engineering, to machine learning models and training data. For that, we built and trained 21 ANN models for websites from 7 different domains and evaluated how accurately they can predict subjective assessments of *Complexity*, *Aesthetics* and *Orderliness* for other ("foreign") domains. The assessments for the 3 subjective scales were provided by 137 participants of various nationalities, while the input data for the models were 32 metrics obtained through visual-based web UI analysis tools.

Concerning **Hypothesis 1** formulated prior to our experimental study, we found that although all the "foreign" models had on average higher mean square errors (+23%

for *Complexity*, +60% for *Aesthetics* and +45% for *Orderliness*), the difference for *Complexity* was not significant.

Exploring **Hypothesis 2**, we found that the measure of distance between the domains that we proposed in our study significantly affected *Aesthetics* and *Orderliness*. The regression model that we built for all the 3 scales was highly significant (but with a low $R^2 = 0.168$) and suggested that on average an extra point of distance adds 8.4% to the model's MSE, compared to the domain-specific ("native") model.

As for the validity of our study, we need to note the rather modest prediction accuracy of the ANN models, which should probably not be used for practical purposes. Yet, this is understandable since we did not seek to increase the models' MSEs by performing Neural Architecture Search, tweaking the hyper-parameters, etc. As our focus was on studying the effects of website domain similarity, we were reluctant to introduce these extra factors to the models. The amounts of training data per domains that we obtained for the study appear adequate, as we found no significant correlation between the models' MSEs and the sample sizes ($p = 0.296$). On the other hand, the control models that were trained on joined domain datasets had better MSE values, which is in line with the notorious "unreasonable effectiveness of data" in ML.

So, can we trust the predictions of ANN models for other domains than the original one? To answer this, we want to provide the reader with three key takeaways:

1. Our results suggest it is safe to assume user models for *Complexity* do not yield significantly less accurate results for foreign domains.
2. Domain distance indeed correlates with prediction accuracy for *Aesthetics* and *Orderliness*, so if you intend to reuse models, try to do so only for close domains. You can assume roughly 8.4% additional MSE per extra point of distance.
3. More research is required and it is always good, although often costly, to have more subjective assessments, but our study shows, with numbers, the trade-off for using available models and training data from different website domains.

When programmers' time became the prime cost in software, reuse came to be an integral part of SE. We believe that it can become similarly worthy in ML, at least for domains where training data is limited or expensive to get. So, in our work, we made a first step towards calculated trade-offs in the reuse of trained ML user behavior models.

Acknowledgment. The reported study was funded by RFBR and DST according to the research project No. 19-57-45006. We thank Vladimir Khvorostov from NSTU for his technical work on collecting the screenshots, the assessments, and the metrics. We are also grateful to all the colleagues who participated and organized assessments of websites.

References

1. Nielsen, J.: Enhancing the explanatory power of usability heuristics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 152–158 (1994)
2. Nebeling, M. et al.: Crowdstudy: general toolkit for crowdsourced evaluation of web interfaces. In: Proceedings of the 5th ACM SIGCHI Symposium on Engineering Interactive Computing Systems, pp. 255–264 (2013)

3. Frick, T.: Designing for Sustainability: a Guide to Building Greener Digital Products and Services. O'Reilly Media, Inc., Sebastopol (2016)
4. Nielsen, J.: Putting A/B Testing in Its Place. Nielsen Norman Group, 14 August 2005. https://www.nngroup.com/articles/putting-ab-testing-in-its-place/. Accessed 13 Jan 2020
5. Beirekdar, A., Keita, M., Noirhomme, M., Randolet, F., Vanderdonckt, J., Mariage, C.: Flexible reporting for automated usability and accessibility evaluation of web sites. In: Costabile, M.F., Paternò, F. (eds.) INTERACT 2005. LNCS, vol. 3585, pp. 281–294. Springer, Heidelberg (2005). https://doi.org/10.1007/11555261_25
6. Bakaev, M.: Assessing similarity for case-based web user interface design. In: Alexandrov, D.A., Boukhanovsky, A.V., Chugunov, A.V., Kabanov, Y., Koltsova, O. (eds.) DTGS 2018, Part I. CCIS, vol. 858, pp. 353–365. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02843-5_28
7. Grigera, J., et al.: Automatic detection of usability smells in web applications. Int. J. Hum Comput Stud. **97**, 129–148 (2017)
8. Speicher, M., Both, A., Gaedke, M.: Ensuring web interface quality through usability-based split testing. In: Casteleyn, S., Rossi, G., Winckler, M. (eds.) ICWE 2014. LNCS, vol. 8541, pp. 93–110. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08245-5_6
9. Nielsen, J.: The Negativity Bias in User Experience. Nielsen Norman Group, 23 October 2016. https://www.nngroup.com/articles/negativity-bias-ux/. Accessed 14 Jan 2020
10. Speicher, M. et al.: TellMyRelevance! predicting the relevance of web search results from cursor interactions. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, pp. 1281–1290 (2013)
11. Chen, X., et al.: The emergence of interactive behaviour: a model of rational menu search. In: Proceedings of the 33rd ACM Conference on Human Factors in Computing Systems, pp. 4217–4226 (2015)
12. Lin, Y.-P., Jung, T.-P.: Improving EEG-based emotion classification using conditional transfer learning. Front. Hum. Neurosci. **11**, 334 (2017)
13. Torrey, L., Shavlik, J.: Transfer learning. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, pp. 242–264. IGI Global (2010)
14. Bakaev, M., et al.: Auto-extraction and integration of metrics for web user interfaces. J. Web Eng. **17**(6), 561–590 (2018)
15. Estuka, F., Miller, J.: A pure visual approach for automatically extracting and aligning structured web data. ACM Trans. Internet Technol. **19**(4), 1–26 (2019)
16. Oulasvirta, A. et al.: Aalto Interface Metrics (AIM): a service and codebase for computational GUI evaluation. In: 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings, pp. 16–19. ACM (2018)
17. Kim, D. et al: Domain switch-aware holistic recurrent neural network for modeling multi-domain user behavior. In: 12th ACM International Conference on Web Search and Data Mining, pp. 663–671 (2019)
18. Dou, Q., et al.: Webthetics: quantifying webpage aesthetics with deep learning. Int. J. Hum Comput Stud. **124**, 56–66 (2019)
19. Boychuk, E., Bakaev, M.: Entropy and compression based analysis of web user interfaces. In: Bakaev, M., Frasincar, F., Ko, I.-Y. (eds.) ICWE 2019. LNCS, vol. 11496, pp. 253–261. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-19274-7_19
20. Hachenberg, C., Gottron, T.: Locality sensitive hashing for scalable structural classification and clustering of web documents. In: Proceedings of the ACM CIKM (2013)