



# The generalizability of machine learning models of personality across two text domains

Mathias Berggren<sup>a,\*</sup>, Lisa Kaati<sup>b</sup>, Björn Pelzer<sup>c</sup>, Harald Stiff<sup>c</sup>, Lukas Lundmark<sup>c</sup>, Nazar Akrami<sup>a</sup>

<sup>a</sup> Department of Psychology, Uppsala University, Sweden

<sup>b</sup> Department of Computer and Systems Sciences, Stockholm University, Sweden

<sup>c</sup> Swedish Defence Research Agency, Sweden

## ARTICLE INFO

### Keywords:

Machine learning  
Big Five  
LIWC  
Text analysis

## ABSTRACT

Machine learning of high-dimensional models have received attention for their ability to predict psychological variables, such as personality. However, it has been less examined to what degree such models are capable of generalizing across domains. Across two text domains (Reddit message and personal essays), compared to low-dimensional- and theoretical models, atheoretical high-dimensional models provided superior predictive accuracy within but poor/non-significant predictive accuracy across domains. Thus, complex models depended more on the specifics of the trained domain. Further, when examining predictors of models, few survived across domains. We argue that theory remains important when conducting prediction-focused studies and that research on both high- and low-dimensional models benefit from establishing conditions under which they generalize.

## 1. Introduction

In recent years, there has been a growing interest in assessing personality using machine learning techniques, most often based on people's activities on social media platforms (e.g., Argamon et al., 2009; Arnoux et al., 2017; Azucar et al., 2018; Bai et al., 2013; Kalghatgi et al., 2015; Kosinski et al., 2013; Majumder et al., 2017; Stachl, Au, et al., 2020; Tandra et al., 2017; Tausczik & Pennebaker, 2010; Yarkoni, 2010). The Five-Factor Model (FFM) has attracted the most attention (see Azucar et al., 2018). Some of these models have also been commercialized (e.g., IBM personality insights & Receptiviti).

In particular, there are several machine learning studies that have successfully used peoples' texts to predict their personality within different domains (e.g., Arnoux et al., 2017; Majumder et al., 2017; Pennebaker & King, 1999). For example, a meta-analysis of the power of digital footprints in predicting Big Five personality found average correlations between 0.29 and 0.40 (Azucar et al., 2018, p.150), with no significant moderation due to text usage, suggesting similar ranges for text-based prediction (p.154–155). A meta-analysis of more constrained dictionary-based methods revealed a somewhat lower prediction of self-assessed personality, with an average  $r^2 = 5.1\%$  (corresponding to  $r = 0.23$ ; Koutsoumpis et al., 2022).

Illustrating the interest in the field, more recently, several authors

have discussed the challenges and opportunities of the field (Rauthmann, 2020), such as the role of psychometrics in machine learning studies (e.g., Alexander III et al., 2020; Tay et al., 2020), and of adapting the method to research question (Möttus et al., 2020). Concerning the last point, the general goal of scientific psychology is to *explain* and *predict* human behavior (for applications to machine learning, see, e.g., Bleidorn & Hopwood, 2019; Yarkoni & Westfall, 2017; Möttus et al., 2020). **Explanation** refers to identifying causal underpinnings and critical elements of relationships and constraints under which the causal mechanism holds to inform future theorizing. **Prediction** refers to exploring if/how we can predict future behaviors and ways of improving the accuracy of our predictions. While both these elements are intertwined and important, studies may emphasize one more than the other. Most high-profile machine learning studies have focused on predicting self-assessed personality using high-dimensional datasets with the highest degree of accuracy possible when faced with new random observations from the same population and in the same domain on which the model was initially trained (e.g., Hall & Matz, 2020; Howlader et al., 2018; Stachl, Pargent, et al., 2020; Youyou et al., 2015; for an overview of studies, see Azucar et al., 2018). Such studies provide essential information regarding prediction but perhaps less about how, why, and when the predictors are related to self-assessed personality. The focus is on the performance of the generated model on new observations drawn

\* Corresponding author at: Department of Psychology, Uppsala University, Box 1225, SE-751 42 Uppsala, Sweden.

E-mail address: [Mathias.Berggren@psyk.uu.se](mailto:Mathias.Berggren@psyk.uu.se) (M. Berggren).

<https://doi.org/10.1016/j.paid.2023.112465>

Received 2 June 2023; Received in revised form 4 October 2023; Accepted 27 October 2023

Available online 1 November 2023

0191-8869/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

from the same population rather than the *generalizability* of the model and the stability of predictors. Notably, models can have a good fit to the data and predict the studied concept (e.g., personality) significantly better than chance, but this does not ensure generalizability beyond the domain in which the model was created (e.g., Akrami et al., 2019).

Questions about the generalizability of such big data models remain less explored in the field (but see Bleidorn & Hopwood, 2019; Tay et al., 2020), but, importantly, methods that work well for prediction need not work well for explanation (Möttus et al., 2020). Indeed, atheoretical and high-dimensional maximization of predictions in one domain may limit the generalizability of the model. By high-dimensional, we here mean a model with a larger number of parameters as a result of, for example, including a larger set of atheoretically derived predictors. More generally, high dimensionality can thus also result from other aspects of the employed methods, such as the high number of parameters in neural networks. All else equal, the more dimensions exploratorily assessed, the more chances to pick up predictors whose connection with a trait is entirely specific to the studied population (i.e., in that domain, group, time, etc.) – indeed, one can argue that this is a benefit of high-dimensional models. To understand (explain) the association, it is therefore important to establish whether it is applicable across any domain (e.g., blog, Twitter, Facebook), just a particular domain (e.g., Facebook), or even just for a particular domain at a specific time-point and a certain population (e.g., college-age adults who posted a lot on Facebook in November 2016).

Specifics of the domain and population may further influence how and how well a particular trait can be predicted. For example, people may routinely present themselves as more conscientious in cover letters than when texting with friends. To understand such results in a broader sense (e.g., how well can this trait be predicted), we thus need to study such possible effects of the domain, which requires studying the generalizability of models.

It is important to note that the issue of generalizability goes beyond that of over-fitting within a domain (e.g., Burnham & Anderson, 2002; Goldstein & Gigerenzer, 2009), which machine learning methods guard against with, e.g., cross-validation – training on one part of the dataset and testing on another (e.g., Stone, 1974). Thus, cross-validation tests whether the model will hold for new cases in the same domain, which, while beneficial, means it cannot establish any across-domain generality of the model – unless multiple domains are, in fact, examined.

The issue of generalizability across domains also connects to the issue of *concept drift* within machine learning (e.g., Lu et al., 2018; Stachl, Pargent, et al., 2020, p.622–623). This means that the statistical properties of the domain can change over time (Lu et al., 2018, p.2), which can then result in a deterioration of models that have previously been trained within that domain (Stachl, Pargent, et al., 2020, p.622). For example, internet slang, or the types of apps people use can change over time, so if these variables are used as predictors, the resulting models may decrease in validity. This illustrates that even within a (seemingly) similar domain, models may have trouble generalizing from one-time point to another. With further differences across domains, this issue may be magnified.

Domain-specific results can, of course, still be highly interesting. They may warrant further study as they demonstrate something about that domain and/or population. However, then that domain-specificity needs to be established in the first place, which once again requires generalizability to be examined. Thus, to further our understanding of why people with a certain personality behave as they do and to appropriately apply their prediction models, researchers need to examine not only the performance of the prediction models for the exact conditions under which they were trained but also their generalizability, along with the interpretability and stability of the predictors. For these reasons, this paper focuses on examining the *generalizability* of the computer-generated-personality models.

Regardless of whether we focus on prediction or explanation, one can ask what to expect when it comes to the strength of the association

between self-assessed and computer-generated personality. The answer to this question depends on what we consider the relation to represent – *equivalent measures of personality or prediction of online behavior* where self-assessed personality is aimed to predict online behavior. Different authors emphasize either the first (e.g., Stachl, Pargent, et al., 2020) or second usage (e.g., Renner et al., 2020; for a discussion, see Hinds & Joinson, 2019). Suppose we have the equivalent measures approach in mind. In that case, we need to consider that the correlations between measures (self-assessment) of the same trait (e.g., Agreeableness) tend to vary between 0.50 and 0.90 and are, in most cases, above 0.80 (e.g., Donnellan et al., 2006; Gosling et al., 2003). If we have a choice between measures, we would want computer models of personality to attain the same levels if we were to select them. Suppose we have the predictive-power approach in mind. In that case, we need to consider that the correlations between self-assessed personality and various types of behavior tend to vary around 0.30 (e.g., Roberts et al., 2007). Focusing on these figures provides a more informed expectation regarding the relation between self-assessed and computer-generated personality and online behavior.

### 1.1. The present study

Depending on the data at hand, there are a few methods of generating personality scores using computerized techniques. Computer-generated personality studies have used different sources of data, some of which are available for some domains but not others. In the present study, we use text written by individuals whose self-assessed FFM personality is available. Focusing on text provides the potential to examine features that are available for most people and domains, unlike some social media features such as Facebook likes. When using text, there are two major techniques. One technique is to use high-dimensional machine learning where a large number of features are extracted from the text and then related to self-assessed personality. Another technique uses text analysis software or word counting techniques (also referred to as dictionary methods) to generate less complex low-dimensional models. In the present study, we use high-dimensional machine learning and low-dimensional modeling based on Linguistic Inquiry and Word Count (LIWC; see Pennebaker et al., 2015; see also Tausczik & Pennebaker, 2010). LIWC, based on a series of dictionaries, extracts psychologically meaningful categories from text. Two examples of dictionaries are *anxiety*, which captures the frequency of words like “worried” and “nervous”, and *tentative*, which captures the frequency of words like “maybe”, “perhaps”, and “guess” (see Appendix of Tausczik & Pennebaker, 2010). Such words thus suggest a more anxious and tentative mood, respectively. Thus, although the modeling is still exploratory, the predictors are more constrained, increasingly theoretically meaningful, and, therefore, perhaps, generalizable. We use data from two samples representing diverse domains (Reddit and Essays).

First, we establish the baseline correlations for the different models when trained and tested in the same domain. Next, we examine how that correlation changes when the model is applied to another domain. Finally, we examine the stability of the significant predictors by whether predictors are the same across domains or whether they differ.

The goal of this research is to examine (1) how well different types of models predict within a domain, (2) across domains, and (3) whether there are predictors that survive better across domains. We predict that high-dimensional machine learning models will have superior accuracy within domains compared to the lower-dimensional models but that this shifts in the opposite direction across domains, as we expected the high-dimensional models to be more likely to pick up on more specific predictors within a domain. The study thus contributes to the field by examining how models generalize across domains, compared to just within a domain, and by examining whether some types of models generalize better than others.

## 2. Method

### 2.1. Data

Our personality models are built on two different datasets. The *Essays* data set contains  $N = 2344$  essays and was collected and used in a study by Pennebaker and King (1999). The dataset consists of essays or daily writing submissions from psychology students, who, for example, were asked to write about “*what your thoughts, feelings, and sensations are at this moment*” or to “*Express in your writing what it has been like for you coming to college, and explore your thoughts and feelings of being in college in general*” (both p.1301–1302), with the student's personality scores assessed by answering the 44 items Big Five Inventory (John et al., 1991). The dataset contained some additional participants who had not filled out all their Big Five scores and who were excluded.

The *Reddit* dataset contains posts from different discussion boards on Reddit. First, we searched for all posts of users who had reported their Big Five personality scores on special forums where such scores were meant to be posted. Next, we only included users who had reported continuous scale scores between 0 and 100 by manually examining the entries. Thus, we excluded users reporting their scores as, for example, “high-low” or “yes-no.” Reddit data on personality has previously been used by Gjurković et al. (2020), although we employed a more restrictive inclusion criterion to increase the comparability between the *Essays* and *Reddit* data. The final set of *Reddit* participants resulted in a total sample of  $N = 1200$  users. This dataset is available upon request from the corresponding author.

Both datasets provide personality scores on the factor level but lack item-level data to enable calculating internal consistency. However, the Big Five Inventory is widely used and tends to have good psychometric properties (e.g., John et al., 1991). We further only included the 10,000 most common words in the texts, following recommendations, for example, in Yarkoni (2010). These sample sizes are larger than- or comparable to many other samples used in studies on machine learning and personality (e.g., Arnoux et al., 2017; Bai et al., 2013; Stachl, Au, et al., 2020) and should provide enough power for modeling purposes.

### 2.2. Model training – high-dimensional vocabulary-based

The different datasets were used to train two machine learning models using a Support Vector Regressor (SVR). The training and testing setup was implemented in Python 3.8 using the machine learning library Scikit-Learn (v. 0.22; see <https://scikit-learn.org/stable/index.html>; additional libraries used: pandas v. 1.0.0, numpy 1.19.5, liwc-analysis 1.2.4) SVR implementation using a standard setting across all models with a Radial Basis Function (RBF) kernel, weight decay constant  $C = 1.0$ , and  $\epsilon = 0.1$  (we manually tested some variations of these parameters under cross-validation, but this provided less improvement to model performance than varying the vocabulary limits, see below, so they were set to standard values). Three different feature sets were used. The statistic tf-idf was used on words and bigrams with a vocabulary limit set to 20,000 and on character levels 1, 2, 3, and 4 with a vocabulary limit set to 20,000. These values were found by cross-validation of different vocabulary limits. We also included a set of psychological variables from LIWC-2015. The tf-idf features were adopted specifically for each dataset and can be seen as domain-dependent, while the LIWC-2015 features are domain-independent. Both datasets were split during the training into 80 % train (used for cross-validation) and 20 % test split. The splits were kept consistent between all experiments to ensure no overlap between train and test data for any of the models and any of the datasets.

### 2.3. Model training - low-dimensional dictionary-based

We trained linear regression models using the 73 LIWC categories (Pennebaker et al., 2015) in two ways. One was forward selection of

those predictors, which significantly ( $\alpha = 0.05$ ) added to the model, starting with the most significant predictor and stopping when no additional predictor significantly explained the remaining variance not explained by already included predictors. The other was a Least Absolute Shrinkage and Selection Operator (LASSO) using 10-fold cross-validation to extract the lambda parameter with the lowest mean squared error. Both these methods allow model pruning, although LASSO shrinks regression coefficients and typically allows more variables in the model. Forward selection instead fits the training data as best as it can in each step. However, it puts a threshold on how significantly a variable must explain the remaining variance in the training set to be included in the model. LASSO is thus typically better suited for prediction, at least in the same domain. At the same time, forward selection may be better at finding a nonredundant (in the sense that no two variables explain mostly the same variance) set of predictors, which can aid the goal of explanation. All variables were standardized separately within training and test data. The same 80 % training (for cross-validation) and 20 % testing split was used for the high-dimensional models. The training and testing setup was implemented in R 4.0.1 using the MASS (for forward selection; v.7.3, Ripley et al., 2023) and glmnet packages (for LASSO; v.4.0, Friedman et al., 2023).

## 3. Results

First, we evaluated the LASSO training models within a domain. See Supplementary materials, Table S1 for final lambda parameters. These models had mean-square errors (MSE) ranging between 0.92 and 0.98 (median 0.97) in the *Essays* dataset and 0.93 and 0.98 (median 0.94) in the *Reddit* dataset. The standard deviation in these errors over the 10-folds ranged between 0.021 and 0.034 (median 0.029) for the *Essays* dataset and 0.022 and 0.042 (median 0.030) for the *Reddit* dataset. Thus, as the errors and their variation were quite similar across datasets, the reliability of models appeared similar across them. MSE:s for 10-fold cross-validation of the high-dimensional models exhibited larger variation, reflecting the higher-dimensionality (and therefore higher variation) of the models, although these, too, were similar in magnitude between datasets. Means: *Essays*, 0.87–0.99 (median 0.96); *Reddit*, 0.91–0.99 (median 0.93); standard deviations: *Essays*, 0.064–0.162 (median 0.131); *Reddit*, 0.075–0.181 (median 0.096).

Next, we examined the correlations ( $r$ ) between self-assessed and computer-generated personality scores within each domain, using the three models outlined in the method section (see Tables 1 and 2). The correlation between self-assessed and computer-generated personality averaged across the personality factors varied between 0.14 and 0.38 for the different models. The high-dimensional model based on *Essays* had the highest average correlation (see Table 1). The overall average correlation between self-assessed and computer-generated personality for the *within-domain* analyses was 0.23 for the *Essays* and 0.19 for the *Reddit* data (total 0.21).

Next, we examined the correlations between self-assessed and computer-generated personality using models trained on data from another domain (examining cross-domain generalizability). Here, we found that the averaged correlations across the personality factors varied between 0.03 and 0.16, with the LIWC-LASSO model built on the *Essays* showing the highest average correlation. The average correlation between self-assessed and computer-generated personality for the cross-domain analyses across factors and models was 0.12 for the *Essays* models and 0.05 for the *Reddit* models (total 0.09).

More importantly, we aimed to test the generalizability of the models by comparing the within-domain to the cross-domain model performance. This was tested with the ‘r.test’ function in ‘psych’ R-package (v.2.0.12, Revelle, 2023). As shown in Tables 1 and 2, compared to the within-domain, the cross-domain correlations were lower in 23 of 30 cases, with 12 being significantly lower (see supplemental material, Table S4). The difference between within and cross-domain correlations was more pronounced for the high-dimensional models, with 8 (out of

**Table 1**Correlations [95 % CI] between self-assessed and computer generated (trained on **Essays Data**) personality.

| Domain/Model                                | Openness                  | Conscientiousness        | Extraversion             | Agreeableness            | Neuroticism              | $M_r$ (SD)  |
|---|---------------------------|--------------------------|--------------------------|--------------------------|--------------------------|-------------|
| Within domain (Essay model on Essay data)   |                           |                          |                          |                          |                          |             |
| LIWC-Select                                 | <b>0.22</b> [0.13, 0.31]  | 0.07 [−0.02, 0.16]       | <b>0.13</b> [0.04, 0.21] | <b>0.11</b> [0.02, 0.20] | <b>0.16</b> [0.07, 0.25] | 0.14 (0.06) |
| LIWC-LASSO                                  | <b>0.25</b> [0.16, 0.33]  | <b>0.10</b> [0.01, 0.19] | <b>0.15</b> [0.06, 0.24] | <b>0.13</b> [0.04, 0.22] | <b>0.16</b> [0.07, 0.24] | 0.16 (0.06) |
| High-dimensional model                      | <b>0.43</b> [0.35, 0.50]  | <b>0.39</b> [0.31, 0.46] | <b>0.40</b> [0.32, 0.47] | <b>0.37</b> [0.29, 0.45] | <b>0.31</b> [0.22, 0.39] | 0.38 (0.04) |
| Across domains (Essay model on Reddit data) |                           |                          |                          |                          |                          |             |
| LIWC-Select                                 | 0.12 [−0.01, 0.24]        | <b>0.15</b> [0.03, 0.27] | 0.10 [−0.03, 0.22]       | <b>0.14</b> [0.01, 0.26] | <b>0.19</b> [0.06, 0.31] | 0.14 (0.03) |
| LIWC-LASSO                                  | <b>0.13</b> [0.00, 0.25]  | <b>0.15</b> [0.02, 0.27] | <b>0.16</b> [0.03, 0.28] | <b>0.16</b> [0.03, 0.28] | <b>0.19</b> [0.07, 0.31] | 0.16 (0.02) |
| High-dimensional model                      | <b>0.06</b> [−0.07, 0.19] | −0.01 [−0.14, 0.12]      | <b>0.21</b> [0.08, 0.33] | −0.01 [−0.13, 0.12]      | 0.08 [−0.05, 0.21]       | 0.07 (0.09) |

LIWC-Select = regression model based on LIWC dictionaries with only significant features included in the model, LIWC-LASSO = least absolute shrinkage and selection operator based on LIWC dictionaries. The model was trained on 80 % and tested on 20 % of the dataset. All correlations are based on 20 % of the dataset. **Boldfaced** correlations are significant on level 0.05. Median correlations are underlined.

**Table 2**Correlations [95 % CI] between self-assessed and computer generated (trained on **Reddit Data**) personality.

| Domain/Model                                 | Openness                 | Conscientiousness         | Extraversion              | Agreeableness             | Neuroticism               | $M_r$ (SD)  |
|--|--------------------------|---------------------------|---------------------------|---------------------------|---------------------------|-------------|
| Within domain (Reddit model on Reddit data)  |                          |                           |                           |                           |                           |             |
| LIWC-Select                                  | <b>0.13</b> [0.01, 0.26] | 0.12 [−0.01, 0.24]        | <b>0.19</b> [0.06, 0.31]  | <b>0.19</b> [0.07, 0.31]  | 0.07 [−0.05, 0.20]        | 0.14 (0.05) |
| LIWC-LASSO                                   | <b>0.13</b> [0.01, 0.26] | 0.12 [−0.01, 0.24]        | <b>0.23</b> [0.11, 0.35]  | <b>0.25</b> [0.13, 0.37]  | <b>0.17</b> [0.04, 0.29]  | 0.18 (0.06) |
| High-dimensional model                       | <b>0.30</b> [0.18, 0.41] | <b>0.14</b> [0.01, 0.26]  | <b>0.17</b> [0.05, 0.29]  | <b>0.24</b> [0.12, 0.36]  | <b>0.17</b> [0.05, 0.29]  | 0.20 (0.07) |
| Across domains (Reddit model on Essays data) |                          |                           |                           |                           |                           |             |
| LIWC-Select                                  | 0.06 [−0.03, 0.15]       | <b>0.03</b> [−0.06, 0.12] | −0.00 [−0.09, 0.09]       | −0.03 [−0.12, 0.06]       | 0.06 [−0.04, 0.15]        | 0.02 (0.04) |
| LIWC-LASSO                                   | <b>0.10</b> [0.01, 0.19] | 0.07 [−0.02, 0.16]        | 0.03 [−0.07, 0.12]        | 0.04 [−0.05, 0.13]        | <b>0.06</b> [−0.03, 0.15] | 0.06 (0.03) |
| High-dimensional model                       | <b>0.10</b> [0.01, 0.19] | <b>0.13</b> [0.04, 0.22]  | <b>0.07</b> [−0.02, 0.16] | <b>0.07</b> [−0.02, 0.16] | −0.05 [−0.14, 0.04]       | 0.06 (0.07) |

LIWC-Select = regression model based on LIWC dictionaries with only significant features included in the model, LIWC-LASSO = least absolute shrinkage and selection operator based on LIWC dictionaries. The model was trained on 80 % and tested on 20 % of the dataset. All correlations are based on 20 % of the dataset. **Boldfaced** correlations are significant on level 0.05. Median correlations are underlined.

10) of the cross-domain correlations being significantly lower compared to their corresponding model within-domain. However, it shall be noted that this is in part because they were higher within-domain in the Essays dataset. For the low-dimensional models, 4 (out of 20) dropped significantly. Models trained in the Essays domain generalized better to the Reddit domain than vice versa, 8 (out of 10) of the low-dimensional models predicted significantly when generalized in this direction (and all LASSO models did). This was not true for the high-dimensional models, for which only 1 model (of 5) was generalized in this direction. In the other direction, both low- and high-dimensional models generalized poorly. We return to this point in the discussion.

Treating the respective correlations as individual observations, two-tailed paired samples Wilcoxon's  $U$  tests showed that the within-domain correlation was not significantly higher than that for across domain for the Essays models,  $U = 86$ ,  $p = .1514$ , (see Table 1), but was so for the Reddit models,  $U = 120$ ,  $p < .001$ , (see Table 2). The non-significance in the first case seems to be because the lower-dimensional did not drop notably in this direction, while the high-dimensional models did, while all types of models appeared to drop in the other direction. Thus, the performance was significantly impaired when models were used outside the domain in which they were built. Further, while there were no significant differences in performance between the Essays and Reddit models when used within their domain ( $U = 126$ ,  $p = .5949$ , two-tailed), independent samples Mann-Whitney's  $U$  test showed that the Essays models performed significantly better than the Reddit models when used outside its domain,  $U = 186$ ,  $p = .002$  (two-tailed).

We also examined the stability of the features that significantly contributed to the predictions by looking at the recurring features across datasets. We could not do this for the features in the high-dimensional models, as the features in these models are far too many to examine individually and are often not immediately psychologically meaningful. We will thus focus on LIWC dictionaries. To examine this question, we first trained the data in one domain, kept only those variables included

in the model, and then trained another model with only those variables in the other domain. The aim was to examine which parts of the models remained as predictors when generalized to the other domain. The features that survived to the end in both directions (Essays → Reddit and Reddit → Essays) without switching signs are presented in Table 3. Predictors (with beta-weights) for all models are presented in the SI, Table S2-3. About 40 % of predictors in the LASSO models did not generalize between domains, and about 70 % in the Select models, although about 60 % and 30 % did. The predictors that remained in both directions (see Table 3) were even more constrained, however, especially for the Select models, which only had two similar predictors in the end, both for the same trait.

Finally, to further examine the importance of the different variables within and across domains, we conducted permutation importance tests on the LASSO models. This means randomly shuffling one variable at a time so it will no longer predict the outcome and examine how much the performance of the model drops. One hundred permutations were conducted per variable, and the average score was used. Performance was examined on the test data. Results are presented in Table S5. Although many of the variables that survived in both directions in Table 3 were on

**Table 3**

Recurring LIWC-Dictionaries kept through both domains in the LASSO models. Dictionaries in bold were also recurring in the LIWC-Select models. The + and − indicates the direction of the beta-coefficients.

| Personality Trait | LIWC-Dictionaries  |
|-------------------|--|
| Openness          | +Affect, +Certainty, +Hear, +Insight, +Pronouns, +Religion, −Reward, −Time |
| Conscientiousness | +Home, −Death, −Negate, −Negative Emotions, −Risk                          |
| Extraversion      | +Conjunctions, +Drives, +Religion, −Fillers, −Negate, −Tentative           |
| Agreeableness     | +Affiliation, +Leisure, −Risk, −Swear words                                |
| Neuroticism       | +Anxiety, +Sadness, −Adjectives, −Work, −2nd person (e.g. "you")           |

the upper end of the permutation importance tests, it was fairly rare that they were so for both domains at once. This further indicates that the most important variables for models differed substantially between domains. A few variables remained towards the upper end, though. The predictions of conscientiousness were fairly (positively) dependent on participants' expressions about their home in both domains. The same was true for extraversion and the use of conjunctions in their writing (positive coefficient), and for neuroticism and expressions of anxiety (positive coefficient) and the use of adjectives (negative coefficient).

#### 4. Discussion

Using datasets from two distinct domains (Essays & Reddit), we examined within-domain prediction, across-domain generalizability, and stability of features in computer-generated personality. We found an average (across Big-Five Factors) correlation between self-assessed and computer-generated personality of 0.21 and 0.09 when testing models within and across domains, respectively. Thus, effect sizes were generally small, and models dropped in performance when used outside the domain where they were constructed.

As predicted, high-dimensional, compared to low-dimensional, models had similar (in the Reddit domain) or superior (in the Essays domain) predictive accuracy within a domain. This confirms the usefulness of high-dimensional models for finding novel patterns within a domain. Further confirming our predictions, within domain predictive power did not help high-dimensional models to generalize across domains. The low-dimensional models similarly did not generalize from Reddit to Essays, but contrastingly, they generalized quite well from Essays to Reddit. In particular, the LASSO Essays models were significant for all Big Five factors and did not drop significantly in accuracy.

Recently, there has been increased recognition of the need to examine the generalizability of psychological models (see, e.g., Yarkoni, 2022). As our results suggest, machine-learning models are not exempted from the problem of generalizability. The comparison between high- and low-dimensional models suggests that atheoretical prediction-maximization within one domain may even lessen the generalizability of models (see also Vijayakumar & Cheung, 2018 for simulation results). Future studies may further look into how these results replicate across domains.

For the high-dimensional models, we further employed standard machine learning techniques, while for the low-dimensional models, we used more standard regression models as used within psychology. This meant that there were more variables in the high-dimensional models but also that the regression functions differed, with only linear functions of predictors included in the low-dimensional models. As our results indicate, different models' behavior may differ greatly depending on whether they are tested within- or across domains. Future studies may further examine what aspects of models most affect their generalizability. This also includes the type of machine learning models used. Although we used standard support vector regression and LASSO models, there are many other machine learning models, and it is possible that some other methods would provide predictions that generalize more readily. However, this would need to be demonstrated.

We believe machine learning has an important role in improving psychology's generalizability and thereby both improve the accuracy of psychology's predictions and its explanatory understanding of various phenomena. However, to achieve this, researchers should train their models in different domains and examine how results change between them. So far, the focus of machine learning studies in psychology has been on maximizing prediction within a domain, and good predictions within domains have been forwarded as a successful achievement of these models (e.g., Youyou et al., 2015). However, until the generalizability of models has been examined, there is no way to ascertain how constrained those models are and whether they have picked up central, more general, or specific predictors. Thus, examining models in different domains in future studies on machine learning in psychology appears

crucial. Here, we have examined how our models generalize between two fairly diverse text domains. Future studies could examine generalizability between more diverse and more similar domains. For example, can a Reddit model trained on a certain subset of Reddit (e.g., political topics) be generalized to another subset of it (e.g., computer game topics, movie topics, etc.)? Or can a Reddit model trained on data before (e.g.) 2020 be generalized to data after 2020? One illuminating procedure might be to rank domains by similarity to the training domain and examine how far the model can generalize.

We have further examined how predictors within a model, rather than just models themselves, generalize between domains. Examining specific predictors appears important for furthering our understanding of how personality predicts (and is predictable by) human behaviors. As we have shown, many predictors did generalize across domains for the LASSO models. However, this was not as true for the Select models, nor when examining the final set of predictors that remained in both directions. This might suggest that the drop is particularly affected by the relative importance of different predictors, as the Select models would pick up on the strongest predictor in each case and then continue until no significant predictor can be added, while the LASSO models tend to include more predictors, but shrinks the regression weights.

The above interpretation was supported by the permutation importance analyses, which showed that the importance of predictors varied substantially across domains. Thus, the most important predictor in one domain need not be a very important predictor in another. It should be noted, however, that this can be affected by covariation amongst predictors. If two predictors explain roughly the same share of the variation, then one may come out as a stronger predictor in one domain, while the other becomes excluded, although this may change in another domain, depending on how they covary with other variables, and how those affect the outcome in each domain. Thus, it can be important to further study such covariation amongst predictors and whether different predictors explain very similar parts of the outcome.

Those predictors that did survive in both directions may be of particular importance for predicting personality across diverse domains (while the permutation analyses reveal importance within each domain). Supporting this, several such predictors appeared central to the personality traits. For example, people higher in neuroticism appeared to express more anxiety and sadness, whether they wrote Essays or comments on Reddit. Similarly, people higher in Extraversion appeared more excitable, expressing more words relating to drives and less tentativeness and negations. Examining which predictors generalize more readily and which are domain-specific can provide researchers with a better understanding of the most central and stable behavioral patterns corresponding to different personality traits. Conversely, finding specific predictors may provide insights into important interactions between personality and situation. Not all predictors had that immediate intuitive explanation, but they may provide the basis for further examination. For example, more neurotic people wrote fewer (of the examined) adjectives, possibly reflecting a tendency to describe things less vividly (e.g.) "a baby" instead of "a *cute* baby," corresponding to a more anhedonic style of writing. Such interpretations, however, require future corroborations. When it comes to the use of predictors within machine learning studies, research may benefit from a balance between the use of high-prediction models with multiple predictors and smaller models with more interpretable predictors that may aid theory-building.

We further found that our low-dimensional models generalized better in one direction, Essays to Reddit, than the other, Reddit to Essays. This may be due to sample size; larger samples provide more robust predictions (the Essays sample was roughly twice the size of the Reddit sample). The Reddit data is also likely to be noisier, making it more difficult to find important predictors. We did not find evidence that this affected the models' reliabilities notably, as the variation in MSE:s over cross-validation were similar across datasets. However, it could still have led to lower performance overall. It might also be that some

domains provide results that generalize more readily than others. The Essays, in which participants were instructed to write about their thoughts and feelings (see Pennebaker & King, 1999), may have been more conducive to making participants express their personality, allowing the machine learning models to pick up on more central predictors, that therefore also generalized to the Reddit domain. In the Reddit domain, conversely, participants would generally have been talking about some topic at hand instead of themselves and expressing their personality more indirectly in relation to the topics discussed. This might have made the predictions more specific to this domain. It is also possible that the difference in population between domains (psychology students versus a broader population on Reddit) could have influenced this result. One such influence might be because psychology students are more familiar with personality traits. Thus, perhaps this also affects how they write texts – expressing themselves more in terms of core behaviors/emotions of those traits – which then affects how well models can predict within a domain and how well those predictions generalize. Future studies may examine whether some domains and/or populations reliably provide more generalizable results than others. This could also further our understanding of the conditions when personality most strongly (and reliably) predict human behavior.

Another pattern in our data is that, while the LASSO models tended to predict slightly higher than the Select models, this difference was quite marginal – even though the Select models included fewer predictors than the LASSO ones (see Supplementary Tables S2-S3). This might be because the Select-models picked up on important predictors. However, which predictors were the strongest seemed to depend highly on the domain. Thus, Select-models may be useful for finding a smaller set of important predictors useful for parsimoniously explaining the pattern in the data. LASSO models may do so, too, by varying the lambda parameter to increase the threshold required for including a predictor in the model. However, other authors have found poor performance of (backward) Select models for predicting personality with text when using cross-validation (Martínez-Huertas et al., 2022), illustrating the need to test the robustness of such predictions even within a domain.

Although we have examined more and less theoretical modeling, it would be wrong to call our low-dimensional models confirmatory (for some theory about personality) – our model-building remains exploratory but with predictors with different degrees of psychological meaning. Thus, it remains to be explored to what degree selecting predictors from theoretical accounts of personality helps with generalizing results. The advantage of explanatory theories is that they can help us find generalizable patterns – provided that the theories are valid. Machine learning methods provide a way to establish whether those predictions hold, and examining the generalizability of models helps show whether theories can provide models that generalize better than exploratory modeling.

Throughout this text, we have argued for an increased focus on the generalizability of machine learning models (and their predictors) in psychology. Our results are by no way the final say about the generalizability of machine learning models, nor should our results be taken as an indication that models' generalizability cannot improve. Future studies should further look into how this can be achieved. However, our results do support that better prediction within a domain need not translate to better prediction across domains. Thus, unless generalizability is examined, it is not sure that models and predictors will hold up to new situations. These results may further be connected to how simpler heuristics can function better in uncertain environments (Gigerenzer & Gaissmaier, 2011). Understanding such domain generality or specificity of models should help advance our understanding of psychological phenomena. It has recently been argued that psychology is in a crisis of generalizability, as models and hypotheses are rarely examined across diverse domains (stimuli, tasks, research sites, etc., see Yarkoni, 2022). As we see it, machine learning, if conducted in diverse domains, may help with such examinations of the generalizability of models, predictors, and theories to help achieve a better understanding of their

stability and/or variability. Generalizability has been less talked about in psychology than replicability (within the same domain), perhaps because examples are scarce where there is clear importance in understanding the generalizability of predictions. We hope the examinations we have conducted here help to bring the importance of generalizability to further consideration in future studies on machine learning and personality.

### CRediT authorship contribution statement

All authors developed the study concept and design. H. Stiff and L. Lundmark collected the data. M. Berggren, N. Akrami, H. Stiff, and L. Lundmark performed data analysis. M. Berggren, N. Akrami and L. Kaati, and B. Pelzer contributed to drafting and revising the manuscript. The final version of the manuscript is approved for submission by all authors.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Data availability

Data will be made available on request.

### Acknowledgement

We thank Dr. James W. Pennebaker for providing us the Essay dataset.

### Funding statement

The research was supported by grants from Riksbankens Jubileumsfond to Nazar Akrami (P15-0603:1).

### Open access statement

To maintain anonymity, data is not uploaded anywhere but is available for researchers upon request by contacting the corresponding author. R-scripts of analyses are available on the Open Science Framework: <https://osf.io/9vcak/>.

### Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.paid.2023.112465>.

### References

- Akrami, N., Fernquist, J., Isbister, T., Kaati, L., & Pelzer, B. (2019, December). Automatic extraction of personality from text: challenges and opportunities. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 3156–3164). IEEE.
- Alexander, L., III, Mulfinger, E., & Oswald, F. L. (2020). Using big data and machine learning in personality measurement: Opportunities and challenges. *European Journal of Personality*, 34(5), 632–648.
- Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2), 119–123.
- Arnoux, P. H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R., & Sinha, V. (2017). 25 tweets to know you: A new model to predict personality with social media. In *Eleventh International AAAI Conference on Web and Social Media*.
- Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, 124, 150–159.
- Bai, S., Hao, B., Li, A., Yuan, S., Gao, R., & Zhu, T. (2013, November). Predicting big five personality traits of microblog users. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (Vol. 1, pp. 501–508). IEEE.

- Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, 23(2), 190–203.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference* (2nd ed.). Springer-Verlag.
- Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, 18(2), 192.
- Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., Qian, J., & Yang, J. (2023). Lasso and elastic-net regularized generalized linear models. <https://cran.r-project.org/web/packages/glmnet/glmnet.pdf>.
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62, 451–482.
- Gjurković, M., Karan, M., Vukojević, I., Bošnjak, M., & Šnajder, J. (2020). Pandora talks: Personality and demographics on Reddit. *arXiv preprint arXiv:2004.04460. Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, NAACL 2021*.
- Goldstein, D. G., & Gigerenzer, G. (2009). Fast and frugal forecasting. *International Journal of Forecasting*, 25(4), 760–772.
- Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6), 504–528.
- Hall, A. N., & Matz, S. C. (2020). Targeting item-level nuances leads to small but robust improvements in personality prediction from digital footprints. *European Journal of Personality*, 34(5), 873–884.
- Hinds, J., & Joinson, A. (2019). Human and computer personality prediction from digital footprints. *Current Directions in Psychological Science*, 28(2), 204–211.
- Howlader, P., Pal, K. K., Cuzzocrea, A., & Kumar, S. M. (2018, April). Predicting facebook-users' personality based on status and linguistic features via flexible regression analysis techniques. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (pp. 339–345).
- John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The big five inventory—Versions 4a and 54*. Berkeley, CA.
- Kalghatgi, M. P., Ramannavar, M., & Sidnal, N. S. (2015). A neural network approach to personality prediction based on the big-five model. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, 2(8), 56–63.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15), 5802–5805.
- Koutsoumpis, A., Oostrom, J. K., Holtrop, D., van Breda, W., Ghassemi, S., & de Vries, R. E. (2022). The kernel of truth in text-based personality assessment: A meta-analysis of the relations between the Big Five and the Linguistic Inquiry and Word Count (LIWC). *Psychological Bulletin*, 148(11–12), 843–868.
- Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12), 2346–2363.
- Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2), 74–79.
- Martínez-Huertas, J. Á., Moreno, J. D., Olmos, R., Martínez-Mingo, A., & Jorge-Botana, G. (2022). A failed cross-validation study on the relationship between LIWC linguistic indicators and personality: Exemplifying the lack of generalizability of exploratory studies. *Psych*, 4, 803–815.
- Möttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., ... Zimmermann, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the Big Five traits. *European Journal of Personality*, 34(6), 1175–1201.
- Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic inquiry and word count: LIWC2015*. Austin, TX: Pennebaker Conglomerates ([www.LIWC.net](http://www.LIWC.net)).
- Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6), 1296.
- Rauthmann, J. F. (2020). A (more) behavioural science of personality in the age of multi-modal sensing, big data, machine learning, and artificial intelligence. *European Journal of Personality*, 34(5), 593–598.
- Renner, K. H., Klee, S., & von Oertzen, T. (2020). Bringing back the person into behavioural personality science using big data. *European Journal of Personality*, 34(5), 670–686.
- Revelle, W. (2023). Procedures for psychological, psychometric, and personality research. <https://cran.r-project.org/web/packages/psych/psych.pdf>.
- Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., & Firth, D. (2023). Support functions and datasets for Venables and Ripley's MASS. <https://cran.r-project.org/web/packages/MASS/MASS.pdf>.
- Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, 2(4), 313–345.
- Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., & Hussmann, H. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30), 17680–17687.
- Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., ... Bühner, M. (2020). Personality research and assessment in the era of machine learning. *European Journal of Personality*, 34(5), 613–631.
- Stone, M. (1974). Cross-validated choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 111–133.
- Tandera, T., Suhartono, D., Wongso, R., & Prasetyo, Y. L. (2017). Personality prediction system from facebook users. *Procedia computer science*, 116, 604–611.
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54.
- Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality*, 34(5), 826–844.
- Vijayakumar, R., & Cheung, M. W. L. (2018). Replicability of machine learning models in the social sciences: A case study in variable selection. *Zeitschrift für Psychologie*, 226(4), 259–273.
- Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3), 363–373.
- Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45(e1), 1–78.
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122.
- Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036–1040.