Revival of the Cover Letter? Experimental Evidence on the Performance of AI-driven Personality Assessments

Completed Research Paper

Christoph Kecht
Technical University of Munich
Arcisstr. 21, 80333 Munich, Germany
christoph.kecht@tum.de

Michael Kurschilgen
UniDistance Suisse
Schinerstr. 18, 3900 Brigue, Switzerland
michael.kurschilgen@unidistance.ch

Magnus Strobel
Technical University of Munich
Arcisstr. 21, 80333 Munich, Germany
magnus.strobel@tum.de

Abstract

Organizations have long been trying to assess job applicants' personality using self-reported psychometric tests, such as the Big Five personality test. However, these tests are not robust against incentives to pretend having certain desirable traits, for example, the disposition for being a good team player. We test whether machine learning classifiers trained on written self-descriptions, such as cover letters, predict people's true cooperativeness better than psychometric tests. Based on data from a controlled online experiment with 400 participants, we find that – when people have incentives to fake their personality – linguistic classifiers based on self-descriptions significantly outperform psychometric classifiers based on the Big Five. Moreover, we find that a fine-tuned, pre-trained natural language model can detect incentives to fake in people's self-descriptions. While further research is needed to achieve tamper-proof models, our findings illustrate the potential of automated personality tests based on job applicants' cover letters.

Keywords: Personality Assessment, Cooperativeness, Big Five, Linguistic Inquiry and Word Count, Machine Learning, Natural Language Processing

Introduction

The ability to identify the best available candidate for a job opening is crucial for business success. Besides evaluating applicants' skills and experience, organizations have long been trying to gather robust cues about candidates' personality (Scepura 2020; Varela et al. 2004). A person's disposition for being a good team player, for example, has become an increasingly sought-after trait in many industries (Chen and Gong 2018; Lazear and Shaw 2007). To gauge a candidate's personality, companies typically use self-reported psychometric tests like the 'Big Five' (Goldberg 1990), which attempt to distill people's personality traits from their answers to a number of Likert-scale questions. For instance, the Big Five trait of *Agreeableness* has recurrently been found to be a good predictor of a person's disposition for being a cooperative team player (Kagel

and McGee 2014; Koole et al. 2016; Volk et al. 2011). Yet, psychometric tests like the Big Five have a critical flaw: they are not robust to the presence of incentives to *pretend* having certain personality traits (Morgeson et al. 2007; Tett and Simonet 2021). Put simply, an applicant who anticipates that *looking* like a good team player increases her chances of being hired, will make sure to score high on *Agreeableness*. More often than not, job applicants will have a good idea of what recruiters expect, and thus an incentive to sugarcoat their personality.

Recent advances in artificial intelligence (AI), and in particular, natural language processing (NLP), such as Facebook's BART model (Lewis et al. 2020) and OpenAI's GPT-3 (Brown et al. 2020), have opened an enticing new path for the prediction of personality (Boyd and Pennebaker 2017; Stachl et al. 2020). Most importantly, language – both spoken and written – has the potential to be considerably more robust than psychometric measures against people's temptation to fake their personality (Newman et al. 2003). Even if job applicants – in an attempt to please the recruiter – managed to modify *what* they say in a cover letter, extant research suggests that they may find it substantially more difficult to modify *how* they say it (Bond and Lee 2005; Hancock et al. 2007; Hauch et al. 2015; Newman et al. 2003; Zhou et al. 2004). The present paper studies the question whether cover letters – analyzed by an appropriately trained AI – are better suited to assess applicants' personality traits than easy-to-fake psychometric tests.

To address this question, we combine state-of-the-art machine learning methods with tools from experimental economics. We conducted an online experiment with 400 participants. In the first stage of the experiment, we elicited participants' *true* cooperativeness using a public goods game. In stages two and three, we asked participants to write a 3,000-character long self-description (in the spirit of a cover letter) and to fill out a 10-item Big Five personality questionnaire. In stage four, we elicited participants' cooperativeness a second time as a manipulation check. To induce exogenous variation in participants' incentives to fake their personality, we randomly assigned them to two different treatments: 25% of participants were assigned to a *Salient-Info* group, in which they received the information that the most cooperative participants would receive an additional bonus payment. The information was given *after* stage one (i. e., participants' initial cooperation decision) but *before* stages two and three (i. e., the self-descriptions and the Big Five questionnaire). The remaining 75% of participants were assigned to a *Baseline* group, in which no such information was given.

In three separate steps, we apply machine learning methods to this experimental dataset. First, using data from the *Baseline* group, we train six state-of-the-art classifiers to predict participants' cooperativeness, based on either linguistic scores extracted from the written self-descriptions or responses to the 10-item Big Five personality questionnaire. Second, we test the out-of-context performance of those classifiers (trained with data from the *Baseline* group) in predicting the *true* cooperativeness of participants from the *Salient-Info* group. Third, we examine whether AI can detect the presence of incentives to fake in natural language. For that purpose, we fine-tune a pre-trained German BERT language model to predict based on the raw text of a written self-description whether that text originated from the *Baseline* group or from the *Salient-Info* group.

We find that when people have salient incentives to fake their cooperativeness, linguistic classifiers based on written self-descriptions significantly outperform psychometric classifiers based on the Big Five. In addition, the fine-tuned, pre-trained BERT language model is able to detect incentives to fake in people's self-descriptions. Our findings provide a glimpse at the untapped potential of assessing job applicants' *true* personality on the basis of their cover letters. Finally, our study illustrates the usefulness of experimentally-generated datasets for developing algorithms that not only predict out-of-sample but also out-of-context.

The remainder of this paper is structured as follows: In the next section, we introduce important concepts related to psychometric tests, linguistic scores, and NLP. We then describe our online experiment in detail, followed by an explanation of our machine learning approach. After presenting the results of our analysis, we conclude our work with a discussion of the results and an overview of current limitations and opportunities for future research.

Theoretical Background and Related Work

Psychometric Tests

The use of personality scores obtained from self-reported psychometric tests is based on the idea that personality (e. g., people's habitual patterns of thought, feeling, and behavior) can be subsumed by a set of personality traits (Borghans et al. 2012). The most influential and widely accepted trait theory is the Big Five model of Goldberg (1990), which posits that personality can be captured along the five different traits *Openness*, *Conscientiousness*, *Extroversion*, *Agreeableness*, and *Neuroticism* ("OCEAN"). Each of these traits consists of different facets, which can be measured by different items. For instance, *Openness* includes the facets of 'fantasy', 'aesthetics', 'feelings', 'actions', 'ideas', and 'values', which can be captured by the items "I have a vivid imagination" and "I have difficulty understanding abstract ideas" (Matz et al. 2016). In these self-reported personality tests, test-takers indicate on a Likert scale how much each item applies to them. Based on the responses to the associated items, a score is calculated for each personality trait. In the absence of incentives to fake, personality scores are predictive for subjective well-being (Anglim et al. 2020), resilience (Oshio et al. 2018), as well as many other life outcomes (Ozer and Benet-Martínez 2006; Soto 2019). However, in the presence of incentives to fake, personality scores lose their predictive power (Morgeson et al. 2007; Tett and Simonet 2021) due to the straightforward phrasing of their items. For instance, one of the items of *Agreeableness* – which has recurrently been found to predict cooperativeness (Kagel and McGee 2014; Koole et al. 2016; Volk et al. 2011) – reads "I see myself as critical, quarrelsome". Applicants who want to be perceived as a team player, will simply respond with "Disagree strongly" to this statement.

Linguistic Scores

Language – both spoken and written – has the potential to be considerably more robust against faking. A prominent example reported by Newman et al. (2003) is the case of a mother who drowned her children in a lake. Talking to journalists, the mother unconsciously used the past tense when referring to her children. The mother's use of the past tense eventually led the FBI to the conclusion that she already knew that her children were dead (Adams 1996). A vast number of related studies have investigated the relationship between language and personality (e. g., Mehl et al. 2006; Moreno et al. 2021; Pennebaker and King 1999; Schwartz et al. 2013; Yarkoni 2010). To extract emotional, cognitive, and structural components of language, Linguistic Inquiry and Word Count (LIWC) has evolved as the gold standard (Pennebaker et al. 2015). LIWC compares each word of a given text with an internal library that contains words from different linguistic categories. These categories include social processes (e. g., "Family: dad, daughter, aunt"), cognitive processes (e. g., "Causation: therefore, reason"), and perceptual processes (e. g., "Feel: feel, sleek"). Each time a word matches one of the categories, it increases the count of this category by one (Tausczik and Pennebaker 2010). As an output, LIWC calculates a score for each category, indicating the share of words in this category over all words in the text. In the absence of incentives to fake, linguistic scores are predictive for personality traits elicited by the Big Five personality test. Examples include video interviews (Hickman et al. 2022), Twitter tweets (Qiu et al. 2012), Facebook status updates (Schwartz et al. 2013), and corporate behavior of senior executives (Yang and Lau 2019). Furthermore, there is a connection between language and individual well-being (Schwartz et al. 2016), academic success (Pennebaker et al. 2014; Robinson et al. 2013), relationship satisfaction (Slatcher et al. 2008), cooperation (Rand et al. 2015), and gender (Newman et al. 2008). Using linguistic scores extracted from study participants' opinions on controversial topics, Newman et al. (2003) classified 67% of liars and truth-tellers correctly. These results and further related studies (e. g., Bond and Lee 2005; Hancock et al. 2007; Hauch et al. 2015; Zhou et al. 2004) suggest that in the presence of incentives to fake, linguistic scores retain their predictive power.

Natural Language Processing

NLP research strives to understand how computers understand text written in natural language (Chowdhury 2003). Recent developments in this field include language models that were pre-trained on large corpora of text, for example, Facebook's BART model (Lewis et al. 2020) and OpenAI's GPT-3 (Brown et al. 2020). The latter was trained with 175 billion parameters and, therefore, significantly surpasses most existing language models even without requiring fine-tuning for a specific task (Brown et al. 2020). One advantage of

language models (compared to simple word frequency-based approaches, such as LIWC) is that they are able to recognize in which context a word is being used. To this end, language models internally apply neural networks, for example, recurrent neural networks (Rumelhart et al. 1986) and long short-term memory networks (Hochreiter and Schmidhuber 1997). Due to their suitability for a wide range of NLP tasks, such as text classification, pre-trained language models have been applied in related studies (e. g., Kecht et al. 2021; Parasurama and Sedoc 2021). For instance, Parasurama and Sedoc (2021) use a BERT model to predict the gender of a resume with an accuracy of 71.6%. Another advantage of these models is their easy applicability, for example, using the open-source Python library "transformers" (Wolf et al. 2020), as demonstrated by Kecht et al. (2021).

Experiment

Design

To test the predictive power of AI-driven personality assessments based on written self-descriptions, and benchmark it with the performance of established psychometric measures based on the Big Five, we conduct a controlled online experiment. The goal of the experiment is to construct a dataset with (i) individual-level data on people's self-descriptions, their answers to a Big Five questionnaire, and an incentivized measure of their actual cooperativeness, and (ii) exogenous variation of people's incentives to fake their personality.

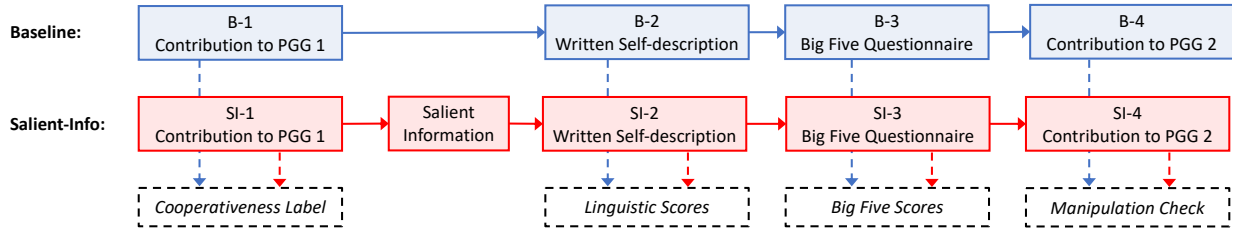

Figure 1. Design of the online experiment.

Figure 1 illustrates the four stages of our experiment that are of particular relevance for the scope of this paper. At the beginning of the experiment, we randomly assigned subjects to a *Baseline* or a *Salient-Info* group with a probability of 75% and 25%, respectively. We choose this distribution to ensure a sufficient number of observations in the *Baseline* group, which will be used to train classifiers, as follows: In stage one, we divided subjects into groups of four players to elicit their *true* cooperativeness in a public goods game (referred to as "PGG" in Figure 1). The public goods game (Isaac and Walker 1988) is the most widely-used measure of cooperative behavior in experimental economics (for a recent overview of the literature, see Engel et al. 2021). In this game, players are endowed with 20 points which they can either retain or invest into a joint project. Each player's payoff function reads: $\pi_i = 20 - g_i + 0.4 \cdot (g_i + \sum_{j=1}^{3} g_j)$, where g_i denotes the player's contribution, g_j are the other players' contributions to the project, and 0.4 is the marginal payoff of contributing to the joint project. The socially optimal outcome $\pi_i = 32$ is achieved when all players contribute all their 20 points to the joint project. However, individually each player is tempted to unilaterally increase their individual payoff to $\pi_i = 44$ by contributing 0 points while the other players continue contributing 20 points. Hence, the setup resembles a teamwork situation in which individual team members face a dilemma between what is best for the team as a whole, and what is best for them individually. We will interpret subjects' contributions g_i as our discrete measure of their true cooperativeness.

After completing stage one, subjects in both the *Baseline* group and the *Salient-Info* group were shown the following information:

> *On the following pages, we ask you to complete 3 additional personality tests.*

To give subjects in the *Salient-Info* a salient monetary incentive to fake being cooperative in the subsequent three stages, they were shown the following information:

> *Based on these 3 personality tests, a committee will decide if you belong to the 40% most cooperative participants. If you belong to the 40% most cooperative participants, you will receive a bonus of €10.*

In stage two, subjects wrote a self-description of approximately 3,000 characters. We set this threshold to resemble the length of a typical one-page cover letter. In this task, we asked subjects to describe themselves by, for instance, elaborating on their skills, hobbies, experiences, dreams, and hopes. In stage three, subjects performed a self-reported 10-item Big Five personality test (Gosling et al. 2003). This personality test measures the items on a 7-point Likert scale and contains two items per personality trait. In stage four, subjects played a second public goods game and made an unconditional contribution decision similar to stage one. This second public goods game served as a manipulation check (see Section "Randomization and Manipulation Checks" for details).

Procedures

After the four main experimental stages described above, we elicited a conditional cooperation test (Fischbacher et al. 2001), subjects' beliefs about their position in the cooperativeness ranking, and a socio-demographics questionnaire (gender, age, citizenship, native tongue, level of education, and the number of siblings). We programmed the experiment in oTree (Chen et al. 2016) and conducted it online with students from the subject pool of experimenTUM, the experimental laboratory at the Technical University of Munich. Between June 30 and July 9, 2020, we conducted 17 online sessions with 400 subjects in total.

For the further data analysis, we excluded ten subjects who did not comply with our requirements for the written self-descriptions (e. g., by using copy & paste). To prevent the results of the linguistic analysis being polluted by a lack of language proficiency, only German native speakers were invited to participate in the experiment. Despite these precautions, the post-experimental questionnaire revealed that 94 subjects were not German native speakers. Excluding these 94 subjects leaves us with 296 subjects for the subsequent analysis (217 subjects in the *Baseline* group and 79 subjects in the *Salient-Info* group). Of these 296 subjects, 143 were males, 152 were females, and one person preferred not to identify their gender. On average, subjects were 24 years old, with 258 subjects being younger than 28 years. 161 subjects were high school graduates, 105 had a Bachelor's degree, 28 had a Master's degree, one held a PhD, and one had a general education certificate. Therefore, the majority of our subjects is likely to soon enter the job market.

Randomization and Manipulation Checks

We check for the correct execution of our random assignment by comparing the contributions to the first public goods game (stage one) across the two treatments. Since we only introduced treatment differences *after* stage one, we should find no difference. Indeed, a Mann-Whitney U test does not show statistical differences ($p=0.695$) between the contributions to the first public goods game in the *Baseline* ($\bar{\mu}=9.06$) and the *Salient-Info* ($\bar{\mu}=8.66$). We also check for the effectiveness of our treatment manipulation, i. e., that the salient information provided between stage one and stage two indeed incentivized subjects to sugarcoat their cooperativeness. If our manipulation worked as intended, contributions to the second public goods game (stage four) should be higher in *Salient-Info* than in *Baseline*. Indeed, a Mann-Whitney U test finds contributions of subjects in the *Salient-Info* ($\bar{\mu}=13.27$) treatment to be significantly higher ($p<0.001$) than in the *Baseline* ($\bar{\mu}=9.00$). For a graphical representation of the mean contributions in both public goods games, see Figure 6 in the Appendix. We conclude that our treatment manipulation was highly effective, i. e., subjects in the *Salient-Info* group tried to fake being cooperative.

Machine Learning Approach

We tackled our research question by comparing the predictive power of written self-descriptions and answers to the Big Five personality test in the absence (i. e., out-of-sample predictions) and presence (i. e., out-of-context predictions) of salient incentives to fake being cooperative. Figure 2 visualizes our machine learning design for the written self-descriptions. For both the *Baseline* and the *Salient-Info* group, we compiled a dataset containing the written self-descriptions from stage two as features. We summarize our discrete cooperation measure from the first public goods game (referred to as "PGG 1" in Figure 2) into a binary cooperativeness label, i. e., two classes denoting whether a subject's contribution is above the median of the *Baseline* group.

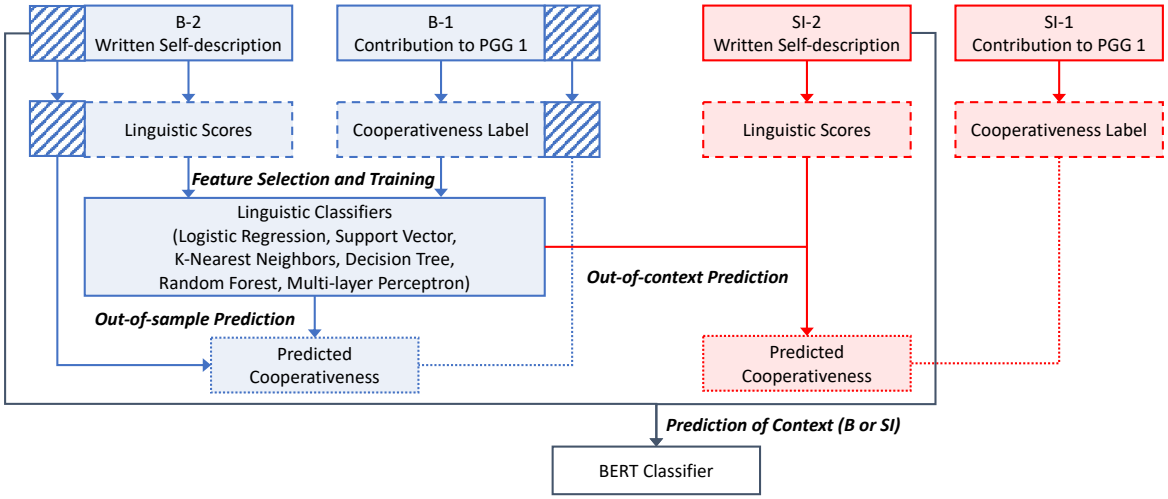

Figure 2. Machine learning design.

In the first step, we used the German LIWC dictionary of 2015 to convert the self-descriptions into 97 linguistic scores (Pennebaker et al. 2015). In the second step, we conducted a correlation analysis to identify the most predictive features among the LIWC scores (see Section "Feature Selection" for details). In the third step, we trained six binary classifiers (see Section "Classifiers" for details) based on the LIWC scores and identified the optimal hyperparameter set for each classifier using a nested cross-validation approach to maximize the out-of-sample performance (see Section "Training and Model Selection" for details). To this end, we held out 20% of the data (visualized in Figure 2 by the shaded rectangles) in each of the five iterations of the nested cross-validation approach and used this data to calculate the out-of-sample performance by comparing the predicted binary cooperativeness with the previously assigned binary cooperativeness labels. In the fourth step, we used the resulting machine learning models to make out-of-context predictions on the LIWC scores we obtained from the *Salient-Info* group to investigate how salient incentives to fake being cooperative affect the predictive power of the linguistic scores.

We followed the same steps for training and evaluating machine learning classifiers on the basis of subjects' answers to the Big Five personality test. Since the responses to the ten items were provided on a 7-point Likert scale, we used these responses as features to train the classifiers, without additional pre-processing.

In the fifth and final step, we predict the context (i. e., whether a subject had salient incentives to fake being cooperative or not) based on the raw text of the self-descriptions using a pre-trained German BERT language model (see Section "Prediction of Incentives to Fake" for details). This allows us to exploit the full potential of raw text data rather than reducing it to the LIWC scores used for the out-of-sample and out-of context predictions. For the Big Five scores, there was no need to transform the answers from the questionnaire to numeric features. Therefore, there is no point in an analogous step based on the Big Five scores.

Feature Selection

Since incorporating features that are not predictive for our label (i. e., whether subjects' true cooperativeness is above the median of the *Baseline* group) would reduce the classifiers' ability to learn from the data, we used a filter method (Chandrashekar and Sahin 2014) to select the most informative scores. Based on data from the *Baseline* group, we conducted a correlation analysis between subjects' true cooperativeness and their linguistic scores and Big Five scores, respectively.

Regarding the linguistic scores, we observe that the LIWC categories *Sadness*, *Future focus*, and *Periods* are negatively correlated with subjects' true cooperativeness at the 10% level, whereas *3rd pers plural*, *Common Adverbs*, *Anxiety*, *Health*, *Drives*, and *Religion* show positive correlations at the the 10% level. Consequently, we selected these nine categories as features for the training of our machine learning models. The Pearson correlation coefficients and respective p-values can be found in Table 4 in the Appendix.

For the Big Five scores, the results show that the traits *Openness*, *Conscientiousness*, and *Agreeableness* are all positively and significantly correlated with subjects' true cooperativeness at the 10% level. Hence, we corroborate the results from the literature on personality traits and cooperation, which find a positive association between *Agreeableness* and cooperation (Kagel and McGee 2014; Koole et al. 2016; Volk et al. 2011). On the item-level, we find that the statements "I see myself as …" (1) "Open to new experiences, complex", (2) "Conventional, uncreative", (3) "Dependable, self-disciplined", and (4) "Sympathetic, warm" are significantly correlated with subjects' true cooperativeness at the 10% level (Grissa et al. 2016). To solely use informative and relevant data for the training of our machine learning models, we selected these four items as features. The Pearson correlation coefficients and respective p-values for all traits and items can be found in Table 3 in the Appendix.

Classifiers

As machine learning models, we used six classifiers for binary classification from the Python machine learning library "scikit-learn" (Pedregosa et al. 2011) that have been applied in related studies. In particular, we used the following classifiers: Logistic Regression (e. g., Hassanein et al. 2021; Tandera et al. 2017), Support Vector Machine (e. g., Mairesse et al. 2007; Pratama and Sarno 2015; Sumner et al. 2012; Tandera et al. 2017), K-Nearest Neighbors (e. g., Mairesse et al. 2007; Pratama and Sarno 2015), Decision Tree (e. g., Mairesse et al. 2007; Sumner et al. 2012), Random Forest (e. g., Hassanein et al. 2021; Sumner et al. 2012), and Multi-layer Perceptron Neural Network (e. g., Tandera et al. 2017). Furthermore, as a benchmark for our predictions, we used four standard dummy classifiers provided by "scikit-learn": (1) the Dummy Minority Classifier, which always predicts that a given player's cooperativeness is above the median, (2) the Dummy Majority classifier, which always predicts that a given player's cooperativeness is below or equal to the median, (3) the Dummy Uniform Classifier, which tosses a fair coin to decide whether a given player's cooperativeness is above the median, and (4) the Dummy Stratified Classifier, which determines each label randomly based on the training set's class distribution and therefore provides a tougher benchmark than the other dummy classifiers. Moreover, from a business perspective, the Dummy Stratified Classifier represents a recruiter who decides whether *current* applicants are above the median in terms of their cooperativeness or not, based on her knowledge about the distribution of cooperativeness among *past* applicants.

We investigated the performance of the initial six classifiers – compared to the Dummy Stratified Classifier – in two distinct situations: first, in the absence of salient incentives to fake being cooperative (out-of-sample predictions on data from the *Baseline* group), and second, in the presence of salient incentives to fake being cooperative (out-of-context predictions on data from the *Salient-Info* group).

Training and Prediction

To train our models and find the best hyperparameters, we instantiated a nested stratified five-fold cross-validation approach. The nested stratified five-fold cross-validation approach consists of an outer loop and an inner loop, each conducting five iterations, respectively. The outer loop serves to evaluate the model's out-of-sample performance, whereas the inner loop strives to improve the model's performance through hyperparameter optimization.

In the outer loop, the data is split into five disjoint stratified folds (i. e., the dataset's distribution is maintained across the individual folds to ensure that each fold is representative of the entire data set). Four folds compose a training dataset, and the fifth fold is a testing dataset for evaluating the model's performance on samples held out from the training process. This prevents having biased results from one of the two classes (i. e., above-median or not above-median cooperators) being under-represented or over-represented in the training dataset or testing dataset. In the five iterations, four folds were used as a training set, while each of the remaining five folds was then used exactly once as a test set to evaluate the model.

In the inner loop, we used the *GridSearchCV* module from the Python library "scikit-learn" (Pedregosa et al. 2011). This module conducts an exhaustive cross-validated grid-search over a specified parameter grid and returns the best hyperparameter set and the respective values. For each classifier, we specified a reasonable parameter grid based on the library's documentation. Furthermore, we pre-processed the data in the inner loop by standardizing the features (subtracting the mean and scaling to unit variance), and by randomly

oversampling the minority class (in our case, the above-median cooperators). Based on a specified grid, *GridSearchCV* iterates over all possible hyperparameter combinations and returns the best model for each classifier and training set. After obtaining the best hyperparameter set for each classifier from the inner loop, we selected the hyperparameter set that best generalizes to out-of-sample data. That is, we evaluated the classifiers' out-of-sample predictions with the test data from the outer loop and chose the hyperparameter set with the lowest generalization error. Finally, we trained each classifier once more using the entire data from the *Baseline* group and the identified optimal hyperparameter set. We then made out-of-context predictions based on the *Salient-Info* group's features with the models trained on the data from the *Baseline* group.

To evaluate and select the models with the smallest generalization error, we used the Matthews correlation coefficient (MCC) as a metric. The advantage of MCC over other measures is that it takes all of the four confusion matrix categories (i. e., true positives, false positives, true negatives, and false negatives) into account (Chicco et al. 2021). This is of particular importance with regard to application processes since a false negative classification could lead to the rejection of an applicant whose true cooperativeness is above the median. Similarly, a false positive classification could lead to hiring an applicant whose true cooperativeness is below the median. Furthermore, according to Chicco and Jurman (2020), the MCC is superior to the F1 score because it does not depend on which class is defined as the positive class, and superior to the accuracy measure, which can be misleading in imbalanced datasets. The MCC ranges from +1 (perfect agreement) to -1 (perfect disagreement), with 0 indicating no relationship between predictions and true labels. For a better comparison with extant work, we will also report the balanced accuracy measure in the Results Section.

Prediction of Incentives to Fake

To detect the presence of incentives to fake in the written self-descriptions, we fine-tune a pre-trained German BERT language model to predict whether a given text originated from the *Baseline* group or from the *Salient-Info* group. We used the Python library "Simple Transformers" that internally calls the "transformers" library (Wolf et al. 2020), and reused deepset's German BERT model (deepset 2019), which was trained on German Wikipedia data, German legal data, and German news articles. Due to the model's universal applicability, it supports a wide range of NLP tasks, such as question answering or text summarization. To achieve suitable predictions based on the raw texts, we fine-tuned the model by training it for this particular task on 80% of the entire data and used the remaining 20% for evaluating the fine-tuned model's performance. We repeated this approach in a stratified cross-validation loop to avoid a bias to a particular training or testing sample. In each iteration, we fine-tuned the model for up to 20 epochs. Since the written-self descriptions of 3,000 characters exceed the model's maximum length of 512 tokens, we applied the sliding window approach (Zaheer et al. 2020) provided by the Python library "Simple Transformers". This approach internally splits the long text into multiple chunks and predicts the probability of representing a particular class for each chunk individually. The absence or presence of salient incentives to fake being cooperative is finally predicted by aggregating the probabilities of the individual chunks.

Results

In a nutshell, we have three main results: First, in the absence of salient incentives to fake being cooperative (i. e., out-of-sample predictions), classifiers based on linguistic scores predict a person's actual cooperativeness at least as good as classifiers based on personality scores. Second, the presence of salient incentives to fake being cooperative (i. e., out-of-context predictions) reduces the predictive power of both the linguistic and the psychometric classifiers but the latter substantially more than the former. Third, the fine-tuned, pre-trained language model based on the raw text of written self-descriptions, predicts the presence of salient incentives to fake being cooperative (i. e., prediction of the context) significantly better than the Dummy Stratified Classifier.

Prediction of Cooperativeness in the Absence of Salient Incentives to Fake

Following the machine learning approach described above, we trained six different machine learning models using the selected personality and linguistic scores from the *Baseline* group as features. Figure 3 visualizes the aggregated MCC and balanced accuracy over all five iterations of the outer loop for each type of classi-

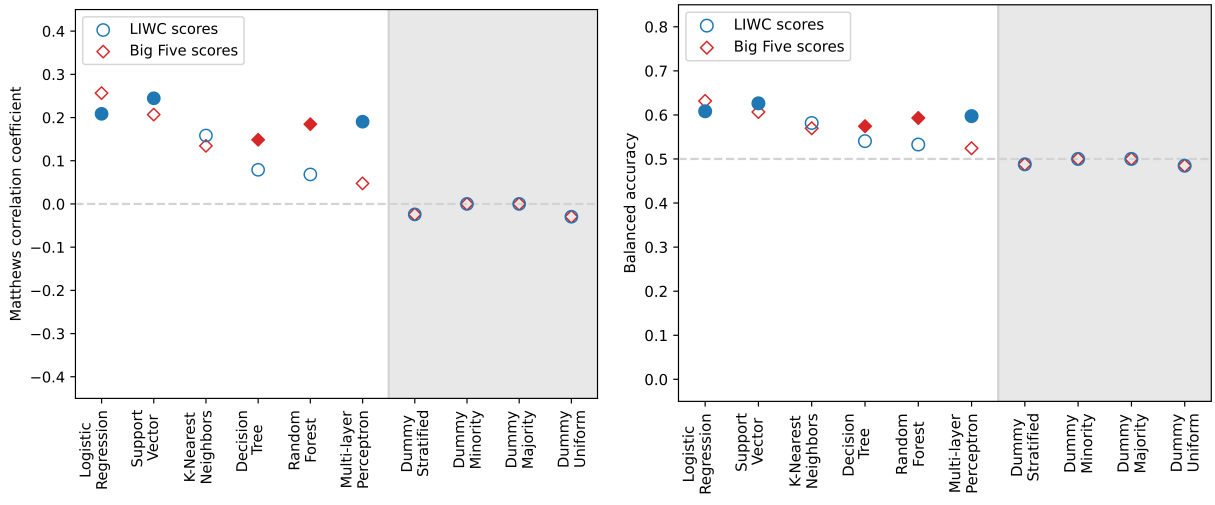

Figure 3. MCC and balanced accuracy for predictions in the absence of salient incentives to fake being cooperative (out-of-sample performance).

Filled markers indicate a significant difference ($p < 0.1$) between a classifier and the corresponding Dummy Stratified Classifier.

Scores	Classifier (CLF)	TN	FP	FN	TP	MCC CLF	MCC Dummy Stratified	p	Bal. Acc. CLF
LIWC	Logistic Regression	81	56	30	50	0.21	-0.02	*0.092	0.61
	Support Vector	86	51	30	50	0.24	-0.02	**0.031	0.63
	K-Nearest Neighbors	84	53	36	44	0.16	-0.02	0.180	0.58
	Decision Tree	83	54	42	38	0.08	-0.02	0.484	0.54
	Random Forest	98	39	52	28	0.07	-0.02	0.246	0.53
	Multi-layer Perceptron	90	47	37	43	0.19	-0.02	*0.065	0.60
Big Five	Logistic Regression	72	65	21	59	0.26	-0.02	0.111	0.63
	Support Vector	72	65	25	55	0.21	-0.02	0.239	0.61
	K-Nearest Neighbors	79	58	35	45	0.13	-0.02	0.345	0.57
	Decision Tree	94	43	43	37	0.15	-0.02	*0.092	0.57
	Random Forest	94	43	40	40	0.18	-0.02	*0.051	0.59
	Multi-layer Perceptron	82	55	44	36	0.05	-0.02	0.693	0.52

Table 1. Performance of classifiers in the absence of salient incentives to fake.

Results of McNemar's tests for pairwise comparisons between the classifiers' predictions on subjects' true cooperativeness (out-of-sample predictions) and those of the Dummy Stratified Classifier (* $p < 0.1$, ** $p < 0.05$).

fier including the four dummy classifiers. The red diamonds represent classifiers based on Big Five scores whereas the blue circles represent classifiers based on LIWC scores. A filled marker indicates that the corresponding classifier performs significantly better than the Dummy Stratified Classifier (for an explanation why we chose the Dummy Stratified Classifier as a benchmark, we refer to Section "Classifiers"). Table 1 complements the visualization of Figure 3 by displaying the confusion matrix for each classifier, as well as the p-values of pairwise McNemar's tests comparing the predictions of each classifier to the predictions of the Dummy Stratified Classifier. As Figure 3 and Table 1 show, all six MCCs are greater than zero. Four out of the six classifiers based on Big Five scores (Logistic Regression, Support Vector, Decision Tree, and Random Forest) and four out of the six classifiers based on LIWC scores (Logistic Regression, Support Vector, K-Nearest Neighbors, and Multi-layer Perceptron) achieve an MCC close to 0.2, indicating a weak positive relationship between subjects' predicted cooperativeness and their true cooperativeness. The balanced accuracy of these eight classifiers ranges between 0.57 and 0.63. By definition, the Dummy Minority Classifier and the Dummy Majority Classifier achieve an MCC of 0 and a balanced accuracy of 0.5, whereas the Dummy Uniform Classifier achieves an MCC of -0.03 and a balanced accuracy of 0.48. An analysis of the individual iterations of the outer loop reveals standard deviations between 0.07 and 0.22 in terms of MCC and standard deviations between 0.04 and 0.12 in terms of balanced accuracy for the classifiers based on Big Five scores.

In contrast, the standard deviation of the classifiers based on LIWC scores within the individual iterations of the outer loop ranges between 0.07 and 0.18 (MCC), and between 0.04 and 0.09 (balanced accuracy). The standard deviation of the Dummy Stratified Classifier equals 0.10 in terms of MCC and 0.05 in terms of balanced accuracy. To test whether our classifiers' predictions are significantly better than the Dummy Stratified Classifier (which reaches an MCC of -0.02 and a balanced accuracy of 0.49), we conducted pair-wise McNemar's tests between the classifiers' predictions on subjects' true cooperativeness based on their Big Five and LIWC scores and those of the Dummy Stratified Classifier (Dietterich 1998; McNemar 1947). Two classifiers based on Big Five scores (Decision Tree and Random Forest) and three classifiers based on LIWC scores (Logistic Regression, Support Vector, and Multi-layer Perceptron) predict significant better than the Dummy Stratified Classifier. For the subsequent analysis in the presence of salient incentives to fake being cooperative, we selected the hyperparameter set for each classifier that achieved the highest MCC in the outer loop and retrained each classifier with the entire data of the *Baseline* group. Thereby, we reduce the impact of the train-test split introduced by the cross-validation approach.

Prediction of Cooperativeness in the Presence of Salient Incentives to Fake

To study the effect of salient incentives to fake being cooperative on the performance of linguistic scores and Big Five scores, we used the scores from the *Salient-Info* group as features and predicted whether subjects' true cooperativeness is above the median or not. Figure 4 visualizes the MCC and balanced accuracy for these predictions, complemented by a confusion matrix and the p-value of a pairwise McNemar's test between the predictions of each classifier and the predictions of the Dummy Stratified Classifier in Table 2.

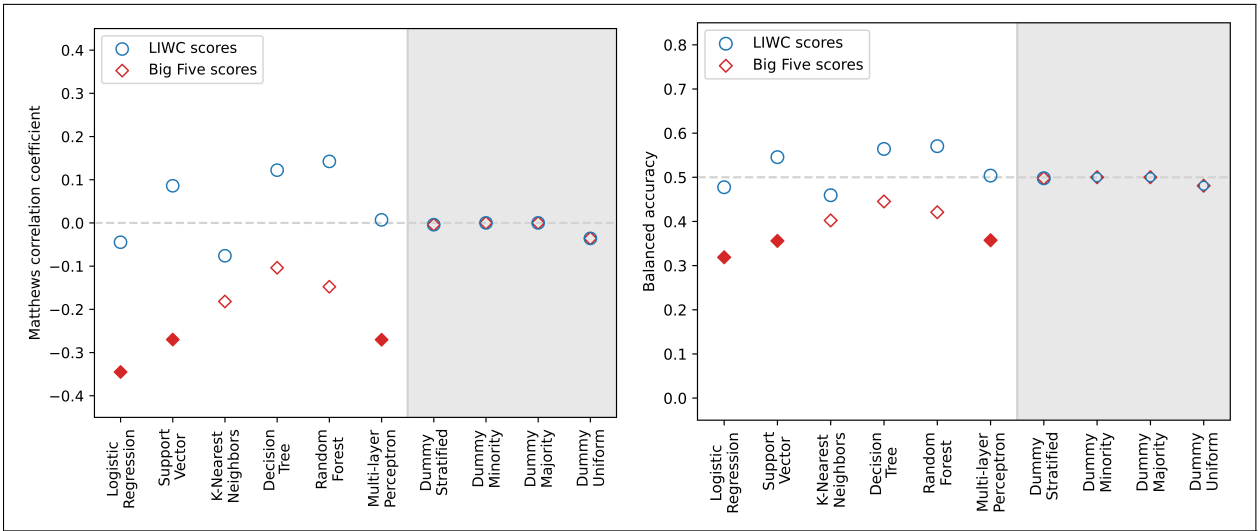

Figure 4. MCC and balanced accuracy for predictions in the presence of salient incentives to fake being cooperative (out-of-context performance).

Filled markers indicate a significant difference ($p < 0.1$) between a classifier and the corresponding Dummy Stratified Classifier.

As Figure 4 shows, the six classifiers based on Big Five scores perform worse than the Dummy Stratified Classifier, whereby this difference is statistically significant in three cases (see Table 2). On the other hand, four of six classifiers based on LIWC scores achieve a higher MCC and a higher balanced accuracy than the Dummy Stratified Classifier (MCC: -0.01, balanced accuracy: 0.5), but do not outperform the Dummy Stratified Classifier significantly. Again, by definition, the performance of the Dummy Minority Classifier and the Dummy Majority Classifier ramains unchanged (MCC: 0, balanced accuracy: 0.5), whereas the Dummy Uniform Classifier achieves an MCC of -0.04 and a balanced accuracy of 0.49. However, three classifiers based on LIWC scores significantly outperform the corresponding classifier based on Big Five scores, namely the Support Vector Classifier (p=0.007), the Decision Tree Classifier (p=0.049), and the Random Forest Classifier (p=0.025).

Scores	Classifier (CLF)	TN	FP	FN	TP	MCC CLF	MCC Dummy Stratified	p	Bal. Acc. CLF
LIWC	Logistic Regression	17	37	9	16	-0.04	-0.01	0.302	0.48
	Support Vector	33	21	13	12	0.09	-0.01	0.635	0.55
	K-Nearest Neighbors	28	26	15	10	-0.08	-0.01	0.749	0.46
	Decision Tree	35	19	13	12	0.12	-0.01	0.451	0.56
	Random Forest	40	14	15	10	0.14	-0.01	0.222	0.57
	Multi-layer Perceptron	22	32	10	15	0.01	-0.01	0.643	0.50
Big Five	Logistic Regression	15	39	16	9	-0.35	-0.01	***0.007	0.31
	Support Vector	19	35	16	9	-0.27	-0.01	**0.043	0.36
	K-Nearest Neighbors	24	30	16	9	-0.18	-0.01	0.201	0.40
	Decision Tree	20	34	12	13	-0.10	-0.01	0.243	0.45
	Random Forest	26	28	16	9	-0.15	-0.01	0.361	0.42
	Multi-layer Perceptron	17	37	15	10	-0.27	-0.01	**0.026	0.36

Table 2. Performance of classifiers in the presence of salient incentives to fake.
Results of McNemar's tests for pairwise comparisons between the classifiers' predictions on subjects' true cooperativeness (out-of-context predictions) and those of the Dummy Stratified Classifier (* p < 0.1, ** p < 0.05, *** p < 0.01).

Prediction of Incentives to Fake

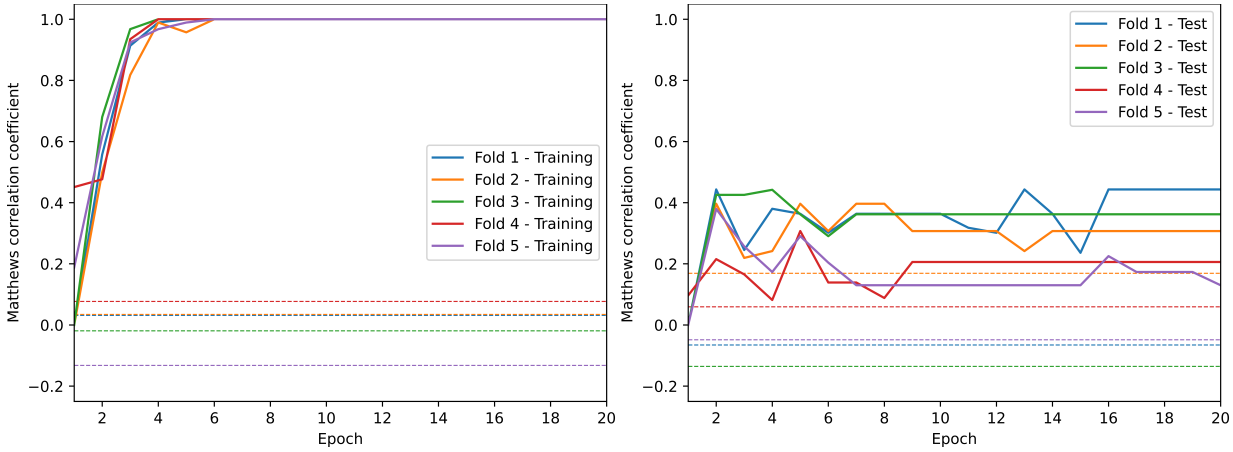

Figure 5. MCCs for predictions of incentives to fake based on the raw text of written self-descriptions (training and test performance).

Figure 5 visualizes the MCC of the pre-trained German BERT model during the training (left) and the test phase (right), respectively, including the MCC of the corresponding Dummy Stratified Classifiers (depicted using the dashed lines). As Figure 5 shows, the performance on the training data converges to an MCC of 1 after four epochs in all five iterations of the cross-validation loop, resulting in over-fitted models after the second epoch. After the second epoch, the MCCs of the five folds of the test data range between 0.22 (weak positive relationship) and 0.44 (moderate positive relationship). A more detailed analysis of the confusion matrices reveals that the BERT models achieve a high specificity (ranging between 0.98 and 1.00) and, therefore, are able to identify the absence of salient incentives to fake reliably. However, the low sensitivity (ranging between 0.06 and 0.31) indicates that the BERT models fail to reliably identify all participants who had incentives to fake their cooperativeness. After aggregating over the five folds, these models significantly outperform corresponding Dummy Stratified Classifiers (MCCs ranging between -0.14 and 0.17), as McNemar's test shows ($p<0.01$ for all epochs).

Discussion

In the absence of salient incentives to fake one's personality, both the classifiers based on Big Five scores and the linguistic scores extracted from written self-descriptions are suitable to assess subject's coopera-

tiveness. However, job applications typically resemble a situation where applicants have an incentive to present themselves as cooperative as possible (Birkeland et al. 2006; Rosse et al. 1998). In the presence of such incentives, we observe that both the Big Five scores and the linguistic scores lose predictive power. The decrease in predictive power is especially high for the Big Five scores, which suggests that written self-descriptions are less vulnerable to faking than self-reported psychometric tests. The fact that we did not find a statistically significant difference neither in the selected personality scores of the *Baseline* and *Salient-Info* group nor in seven of the nine selected LIWC categories, raises the question why we did observe a decrease in the predictive power of these scores. A potential explanation might be that although the difference was not statistically significant, it was still sufficient to mislead the classifiers.

In this study, participants' contributions in a public goods game served as a proxy for their individual cooperativeness. Although the public goods game is, by far, the most common measure of cooperation in experimental economics (Kagel and Roth 2020), it does not constitute the ground truth of a person's cooperativeness. In principle, an uncooperative person could always pretend being more cooperative, not just in the second public goods game of our *Salient-Info* treatment. However, it is important to note that the interpretation of our results does not require one treatment *without* incentives to fake and one *with* incentives to fake. It only requires an exogenous variation of the *degree* of incentives to fake. Our manipulation check showed that this is the case. Beyond cooperativeness, our approach could easily be adapted to examine other dimensions of people's personality, provided there is a well-established proxy measure. Honesty (Abeler et al. 2019), trust, and patience (Falk et al. 2018) would be some straightforward candidates.

We evaluated our classifiers using the MCC, which takes all four dimensions of the confusion matrix equally into account (Chicco et al. 2021). Yet in reality, whether a false-positive prediction or a false negative prediction poses a higher risk, will depend on the organization's hiring strategy. On the one hand, an organization could fear the risk of hiring an uncooperative applicant more than the risk of missing out on a (otherwise equally-qualified) cooperative applicant. Such organizations have a high volume of qualified applications and can thus afford to reject a candidate upon the smallest doubt about the applicant's fit. Such an organization wants a classifier that reliably filters out all uncooperative applicants, i. e., a classifier with a low false-positive rate. On the other hand, an organization may fear the risk of missing out on a cooperative applicant more than the risk of hiring an actually uncooperative applicant. These are, for instance, organizations with a low volume of qualified applications, who are compelled to pursue a more risky hiring strategy in order to prevail in a competitive market. Such organizations want a classifier that does not discard erroneously any cooperative applicant, i. e., a classifier with a low false-negative rate. The confusion matrices of our classifiers (both in the absence and in the presence of salient incentives to fake), reveal a high false-positive rate and a low false-negative rate for all classifiers. Therefore, our classifiers would be particularly helpful for the second type of organizations.

This paper's central finding, i. e., that linguistic classifiers based on written self-descriptions significantly outperform psychometric classifiers based on the Big Five, is a promising first step towards an automated tool for predicting applicants' personality on the basis of their cover letters. Yet, the practical implementation of our approach to classify job applicants' cooperativeness based on their cover letters remains challenging, due to the current lack of suitable corpora of training data to train a classifier and to fine-tune an existing language model, and the potential selection bias inherent to most real-world datasets. Outside of controlled experiments, it is very hard to find data with random assignment to different incentive situations, and thus without potential selection bias. Real-world job openings that vary in terms of the salience of the desired personality traits would typically also vary in other dimensions, such as salary, autonomy, and required skill set.

Besides the technical complexity, it remains questionable whether the automation of personality assessments is socially desirable. On the one hand, developing an automated tool to predict applicant's personality on the basis of their cover letter could unlock considerable efficiency gains by improving the match between company and employee. Moreover, such a tool could contribute to leveling the playing field between large organizations, who have the means to compensate for the current dearth of reliable low-cost predictors by running elaborate assessment centers, and small businesses, who lack those means. On the other hand, automation could lead to discrimination, i. e., the unfair treatment of individuals based on certain protected attributes such as education, gender, or ethnicity (Ferrer et al. 2021). The AI-driven analysis of motiva-

tion letters could, for instance, potentially discriminate against individual applicants with a specific gender, if gender could be predicted from self-descriptions and gender is correlated with the variable of interest. While removing certain attributes prevents the algorithm from using them, this might also result in a loss of accuracy. In addition, discrimination could also arise from an unbalanced training data set. Although our data set was rather balanced in terms of gender, it was not balanced in terms of other important attributes like ethnicity and socio-economic background. Before applying such an algorithm for actual evaluation of job applicants in the field, it would be indispensable to ensure proper training with the specific target population.

Conclusion

This paper has analyzed and compared the performance of the Big Five personality traits and written self-descriptions in predicting cooperativeness when people have salient incentives to fake. The contribution of this paper is threefold. First, to the best of our knowledge, this is the first paper to report empirical evidence on how NLP-based assessments perform relative to psychometric tests, how incentives to fake affect the performance of AI-driven personality assessments, and whether AI can detect the presence of incentives to fake in natural language. Second, we investigate these questions in the context of a real-world problem since job applications resemble a situation in which applicants have incentives to fake being cooperative. Our results suggest that in this situation, self-reported personality tests are not suitable to assess applicants' cooperativeness. In contrast, cover letters offer untapped potential for an automated assessment of cooperativeness of which both small and large organizations could benefit if the assessment is aligned with their hiring strategy. Third, our interdisciplinary study bridges the gap between machine learning and experimental economics. Conducting the experiment in a controlled setting enabled us to exogenously vary the treatment, and thereby eliminate biases that would be present in real-world data. Consequently, we could draw profound conclusions about how salient incentives to fake affect the predictive power of our classifiers.

Nevertheless, our study has several limitations that should be tackled by future work. First, our linguistic scores were extracted from German written self-descriptions. It is unclear whether our results based on the linguistic scores carry over to other languages. Second, the small sample size of 296 participants might have hampered the classifiers' ability to learn from the data, and limited the power of our statistical tests. To appraise the robustness of our findings, future work should conduct further experiments with more participants, other languages, and more diverse socio-demographics. Third, whereas related studies provide evidence for a relationship between cooperativeness and language (e. g., Rand et al. 2015) as well as personality traits (e. g., Kagel and McGee 2014; Koole et al. 2016; Volk et al. 2011), we did not observe a statistically significant difference in our selected personality scores between the *Baseline* and *Salient-Info* group and only in two of nine selected LIWC categories. This observation raises the question whether there are additional influencing factors beyond the LIWC categories that are predictive for subjects' true cooperativeness. The fact that our BERT classifier is able to detect a person's incentive to fake from the raw text of her self-description, suggests that the LIWC scores might not sufficiently capture some more subtle linguistic cues in the presence of incentives to fake. To assess job applicants' cooperativeness more accurately, the development of a two-step classifier could be a promising path. In this case, a first classifier predicts the presence or absence of incentives to fake, and depending on the result, a second classifier is selected to predict the actual cooperativeness. This second classifier could even be extended to predict the cooperativeness as a percentage score relative to top-performing employees, in the spirit of van den Broek et al. (2019). Fourth, we evaluated our models against dummy classifiers provided by the Python library "sklearn". For a more realistic evaluation, the predictions of the classifiers should be compared against assessments of professional human recruiters faced with the same cover letters.

References

Abeler, J., Nosenzo, D., and Raymond, C. 2019. "Preferences for truth-telling". *Econometrica* (87:4), pp. 1115–1153.

Adams, S. H. 1996. "Statement analysis: What do suspects' words really reveal". *FBI Law Enforcement Bulletin* (65:10), pp. 12–20.

Anglim, J., Horwood, S., Smillie, L. D., Marrero, R. J., and Wood, J. K. 2020. "Predicting psychological and subjective well-being from personality: A meta-analysis". *Psychological Bulletin* (146:4), pp. 279–323.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., and Smith, M. A. 2006. "A Meta-Analytic Investigation of Job Applicant Faking on Personality Measures". *International Journal of Selection and Assessment* (14:4), pp. 317–335.

Bond, G. D. and Lee, A. Y. 2005. "Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language". *Applied Cognitive Psychology* (19:3), pp. 313–329.

Borghans, L., Duckworth, A. L., Heckman, J. J., and ter Weel, B. 2012. "The Economics and Psychology of Personality Traits". *Journal of Human Resources* (43:4), pp. 972–1059.

Boyd, R. L. and Pennebaker, J. W. 2017. "Language-based personality: A new approach to personality in a digital world". *Current Opinion in Behavioral Sciences* (18), pp. 63–68.

Brown, T. B. et al. 2020. "Language Models are Few-Shot Learners". In: *Proceedings of the 34th Conference on Neural Information Processing Systems*. Vol. 33. Online, pp. 1877–1901.

Chandrashekar, G. and Sahin, F. 2014. "A survey on feature selection methods". *Computers & Electrical Engineering* (40:1), pp. 16–28.

Chen, D. L., Schonger, M., and Wickens, C. 2016. "oTree – An open-source platform for laboratory, online, and field experiments". *Journal of Behavioral and Experimental Finance* (9), pp. 88–97.

Chen, R. and Gong, J. 2018. "Can self selection create high-performing teams?" *Journal of Economic Behavior & Organization* (148), pp. 20–33.

Chicco, D. and Jurman, G. 2020. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". *BMC Genomics* (21:6).

Chicco, D., Tötsch, N., and Jurman, G. 2021. "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation". *BioData Mining* (14:13).

Chowdhury, G. G. 2003. "Natural language processing". *Annual Review of Information Science and Technology* (37:1), pp. 51–89.

deepset 2019. *German BERT | State of the Art Language Model for German NLP*. URL: https://www.deepset.ai/german-bert (visited on Mar. 5, 2022).

Dietterich, T. 1998. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms". *Neural Computation* (10:7), pp. 1895–1923.

Engel, C., Kube, S., and Kurschilgen, M. 2021. "Managing expectations: How selective information affects cooperation and punishment in social dilemma games". *Journal of Economic Behavior & Organization* (187), pp. 111–136.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. 2018. "Global evidence on economic preferences". *The Quarterly Journal of Economics* (133:4), pp. 1645–1692.

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., and Criado, N. 2021. "Bias and Discrimination in AI: A Cross-Disciplinary Perspective". *IEEE Technology and Society Magazine* (40:2), pp. 72–80.

Fischbacher, U., Gächter, S., and Fehr, E. 2001. "Are people conditionally cooperative? Evidence from a public goods experiment". *Economics Letters* (71:3), pp. 397–404.

Goldberg, L. R. 1990. "An alternative "description of personality": The Big-Five factor structure". *Journal of Personality and Social Psychology* (59:6), pp. 1216–1229.

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. 2003. "A very brief measure of the Big-Five personality domains". *Journal of Research in Personality* (37:6), pp. 504–528.

Grissa, D., Pétéra, M., Brandolini, M., Napoli, A., Comte, B., and Pujos-Guillot, E. 2016. "Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data". *Frontiers in Molecular Biosciences* (3:30).

Hancock, J. T., Curry, L. E., Goorha, S., and Woodworth, M. 2007. "On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication". *Discourse Processes* (45:1), pp. 1–23.

Hassanein, M., Rady, S., Hussein, W., and Gharib, T. F. 2021. "Predicting the Big Five for social network users using their personality characteristics". In: *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pp. 160–164.

Hauch, V., Blandón-Gitlin, I., Masip, J., and Sporer, S. L. 2015. "Are computers effective lie detectors? A meta-analysis of linguistic cues to deception". *Personality and Social Psychology Review* (19:4), pp. 307–342.

Hickman, L., Saef, R., Ng, V., Woo, S. E., Tay, L., and Bosch, N. 2022. "Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews". *Human Resource Management Journal* (n/a).

Hochreiter, S. and Schmidhuber, J. 1997. "Long Short-Term Memory". *Neural Computation* (9:8), pp. 1735–1780.

Isaac, R. M. and Walker, J. M. 1988. "Group size effects in public goods provision: The voluntary contributions mechanism". *The Quarterly Journal of Economics* (103:1), pp. 179–199.

Kagel, J. and McGee, P. 2014. "Personality and cooperation in finitely repeated prisoner's dilemma games". *Economics Letters* (124:2), pp. 274–277.

Kagel, J. H. and Roth, A. E. 2020. *The Handbook of Experimental Economics, Volume 2*. Princeton, NJ, United States: Princeton University Press.

Kecht, C., Egger, A., Kratsch, W., and Röglinger, M. 2021. "Event Log Construction from Customer Service Conversations Using Natural Language Inference". In: *2021 3rd International Conference on Process Mining (ICPM)*. IEEE. Eindhoven, Netherlands, pp. 144–151.

Koole, S. L., Jager, W., van den Berg, A. E., Vlek, C. A. J., and Hofstee, W. K. B. 2016. "On the Social Nature of Personality: Effects of Extraversion, Agreeableness, and Feedback about Collective Resource Use on Cooperation in a Resource Dilemma". *Personality and Social Psychology Bulletin* (27:3), pp. 289–301.

Lazear, E. P. and Shaw, K. L. 2007. "Personnel economics: The economist's view of human resources". *Journal of economic perspectives* (21:4), pp. 91–114.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. 2020. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online, pp. 7871–7880.

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. 2007. "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text". *Journal of Artificial Intelligence Research* (30:1), pp. 457–500.

Matz, S., Chan, Y. W. F., and Kosinski, M. 2016. "Models of Personality". In: *Emotions and Personality in Personalized Services*. Ed. by M. Tkalčič, B. de Carolis, M. de Gemmis, A. Odić, and A. Košir. Human–Computer Interaction Series. Cham: Springer International Publishing, pp. 35–54.

McNemar, Q. 1947. "Note on the sampling error of the difference between correlated proportions or percentages". *Psychometrika* (12:2), pp. 153–157.

Mehl, M. R., Gosling, S. D., and Pennebaker, J. W. 2006. "Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life". *Journal of Personality and Social Psychology* (90:5), pp. 862–877.

Moreno, J. D., Martínez-Huertas, J. Á., Olmos, R., Jorge-Botana, G., and Botella, J. 2021. "Can personality traits be measured analyzing written language? A meta-analytic study on computational methods". *Personality and Individual Differences* (177), p. 110818.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., and Schmitt, N. 2007. "Reconsiderung the Use of Personality Tests in Personnel Selection Contexts". *Personnel Psychology* (60:3), pp. 683–729.

Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W. 2008. "Gender Differences in Language Use: An Analysis of 14,000 Text Samples". *Discourse Processes* (45:3), pp. 211–236.

Newman, M. L., Pennebaker, J. W., Berry, D. S., and Richards, J. M. 2003. "Lying words: Predicting deception from linguistic styles". *Personality and Social Psychology Bulletin* (29:5), pp. 665–675.

Oshio, A., Taku, K., Hirano, M., and Saeed, G. 2018. "Resilience and Big Five personality traits: A meta-analysis". *Personality and Individual Differences* (127), pp. 54–60.

Ozer, D. J. and Benet-Martínez, V. 2006. "Personality and the prediction of consequential outcomes". *Annual Review of Psychology* (57), pp. 401–421.

Parasurama, P. and Sedoc, J. 2021. "Gendered Language in Resumes – An Empirical Analysis of Gender Norm Violation and Hiring Outcome". In: *Proceedings of the 42nd International Conference on Information Systems*. Austin, TX, United States.

Pedregosa, F. et al. 2011. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* (12:85), pp. 2825–2830.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., and Francis, M. E. 2015. *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX, United States.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., and Beaver, D. I. 2014. "When small words foretell academic success: The case of college admissions essays". *PLOS One* (9:12), e115844.

Pennebaker, J. W. and King, L. A. 1999. "Linguistic styles: Language use as an individual difference". *Journal of Personality and Social Psychology* (77:6), pp. 1296–1312.

Pratama, B. Y. and Sarno, R. 2015. "Personality classification based on Twitter text using Naive Bayes, KNN and SVM". In: *2015 International Conference on Data and Software Engineering*, pp. 170–174.

Qiu, L., Lin, H., Ramsay, J., and Yang, F. 2012. "You are what you tweet: Personality expression and perception on Twitter". *Journal of Research in Personality* (46:6), pp. 710–718.

Rand, D. G., Kraft-Todd, G., and Gruber, J. 2015. "The collective benefits of feeling good and letting go: Positive emotion and (dis)inhibition interact to predict cooperative behavior". *PLOS One* (10:1), e0117426.

Robinson, R. L., Navea, R., and Ickes, W. 2013. "Predicting Final Course Performance From Students' Written Self-Introductions". *Journal of Language and Social Psychology* (32:4), pp. 469–479.

Rosse, J. G., Stecher, M. D., Miller, J. L., and Levin, R. A. 1998. "The impact of response distortion on preemployment personality testing and hiring decisions". *Journal of Applied Psychology* (83:4), pp. 634–644.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. "Learning representations by back-propagating errors". *Nature* (323), pp. 533–536.

Scepura, R. C. 2020. "The Challenges With Pre-Employment Testing and Potential Hiring Bias". *Nurse Leader* (18:2), pp. 151–156.

Schwartz, H. A. et al. 2013. "Personality, gender, and age in the language of social media: The open-vocabulary approach". *PLOS One* (8:9), e73791.

Schwartz, H. A. et al. 2016. "Predicting Individual Well-Being Through the Language of Social Media". In: *Proceedings of the Pacific Symposium on Biocomputing 2016*. Ed. by R. B. Altman, A. K. Dunker, L. Hunter, M. D. Ritchie, T. A. Murray, and T. E. Klein. Kohala Coast, Hawaii, United States, pp. 516–527.

Slatcher, R. B., Vazire, S., and Pennebaker, J. W. 2008. "Am "I" more important than "we"? Couples' word use in instant messages". *Personal Relationships* (15:4), pp. 407–424.

Soto, C. J. 2019. "How Replicable Are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project". *Psychological Science* (30:5), pp. 711–727.

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., and Bühner, M. 2020. "Personality Research and Assessment in the Era of Machine Learning". *European Journal of Personality* (34:5), pp. 613–631.

Sumner, C., Byers, A., Boochever, R., and Park, G. J. 2012. "Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets". In: *2012 11th International Conference on Machine Learning and Applications*. Vol. 2, pp. 386–393.

Tandera, T., Hendro, Suhartono, D., Wongso, R., and Prasetio, Y. L. 2017. "Personality Prediction System from Facebook Users". *Procedia Computer Science* (116). 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017), pp. 604–611.

Tausczik, Y. R. and Pennebaker, J. W. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods". *Journal of Language and Social Psychology* (29:1), pp. 24–54.

Tett, R. P. and Simonet, D. V. 2021. "Applicant Faking on Personality Tests: Good or Bad and Why Should We Care?" *Personnel Assessment and Decisions* (7:1), pp. 6–19.

Van den Broek, E., Sergeeva, A., and Huysman, M. 2019. "Hiring Algorithms: An Ethnography of Fairness in Practice". In: *Proceedings of the 40th International Conference on Information Systems*. Munich, Germany.

Varela, J. G., Boccaccini, M. T., Scogin, F., Stump, J., and Caputo, A. 2004. "Personality Testing in Law Enforcement Employment Settings: A Metaanalytic Review". *Criminal Justice and Behavior* (31:6), pp. 649–675.

Volk, S., Thöni, C., and Ruigrok, W. 2011. "Personality, personal values and cooperation preferences in public goods games: A longitudinal study". *Personality and Individual Differences* (50:6), pp. 810–815.

Wolf, T. et al. 2020. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online, pp. 38–45.

Yang, K. and Lau, R. Y. K. 2019. "Detecting Senior Executives' Personalities for Predicting Corporate Behaviors: An Attention-based Deep Learning Approach". In: *Proceedings of the 40th International Conference on Information Systems*. Munich, Germany.

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), pp. 363–373.

Zaheer, M. et al. 2020. "Big Bird: Transformers for Longer Sequences". In: *Advances in Neural Information Processing Systems (NeurIPS 2020)*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33, pp. 17283–17297.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., and Twitchell, D. 2004. "Automating Linguistics-Based Cues for Detecting Deception in Text-Based Asynchronous Computer-Mediated Communications". *Group Decision and Negotiation* (13:1), pp. 81–106.

Appendix

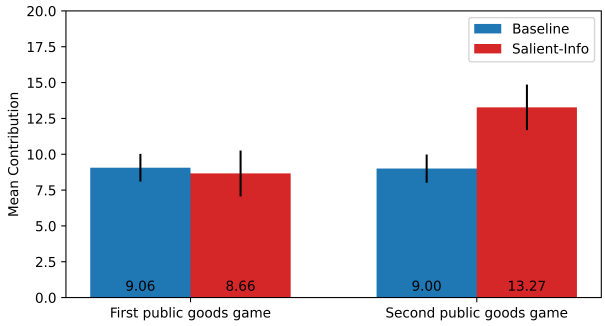

Figure 6. Mean contributions in the first and second public goods game by group.

Personality Trait	r	p	Item	r	p
Openness	0.185	***0.006	Open to new experiences, complex Conventional, uncreative	0.141 -0.149	**0.039 **0.028
Conscientiousness	0.117	*0.085	Dependable, self-disciplined Disorganized, careless	0.160 -0.051	**0.018 0.455
Extraversion	0.038	0.581	Extraverted, enthusiastic Reserved, quiet	0.016 -0.050	0.814 0.460
Agreeableness	0.195	***0.004	Critical, quarrelsome Sympathetic, warm	-0.112 0.218	0.101 ***0.001
Neuroticism	0.019	0.783	Anxious, easily upset Calm, emotionally stable	-0.031 0.002	0.647 0.973

Table 3. Correlations between subjects' true cooperativeness and their personality scores (* p < 0.1, ** p < 0.05, *** p < 0.01).

LIWC category	Label	Examples	r	p
3rd pers plural	they	they, their, they'd	0.126	*0.063
Common Adverbs	adverb	very, really	0.115	*0.090
Anxiety	anx	worried, fearful	0.137	**0.043
Sadness	sad	crying, grief, sad	-0.146	**0.032
Health	health	clinic, flu, pill	0.154	**0.023
Drives	drives	friend, success, bully	0.117	*0.085
Future focus	focusfuture	may, will, soon	-0.127	*0.063
Religion	relig	altar, church	0.113	*0.097
Periods	Period	-	-0.120	*0.078

Table 4. Correlations between subjects' true cooperativeness and their linguistic scores (* p < 0.1, ** p < 0.05).