Lie Detection and Removal in the Parental Reflective Functioning Questionnaire

Francesco Maria Calistroni, Alvise Dei Rossi, Elisa Tremolada

Introduction

Psychological questionnaires have become particularly popular in settings where assessing the presence of psychological disorders in a timely and economic fashion is deemed more important than obtaining a detailed description. Examples may be decisions to force an individual to undergo Compulsory Medical Treatment or, more generally, court cases.

This paper is based on the results of an experiment conducted on a sample of 339 individuals, who were asked to answer the **Parental Reflective Functioning Questionnaire (PRFQ)** twice: once truthfully, once lying. In this second case, they were asked to imagine that they were trying to obtain custody of their children in family court.

We first concern ourselves with the identification of an accurate method for the classification of *honest* vs *faked* profiles. However, the final aim of the paper is that of establishing if, and with what level of accuracy, it is possible to develop a method for the reconstruction of the "true" underlying answer profile, from the "fake" one.

Research problem

As anticipated above, the aim of this paper is twofold: firstly, we aim at developing an efficient method for lie detection (i.e. for classifying honest and dishonest responders) in pyschological questionnaires. Secondly, we attempt to find an efficient procedure for the reconstruction of the honest profile of the faker, once she has been identified.

Lying in psychological questionnaires takes two main forms: *faking bad*, i.e. exaggerating a negative behaviour, or *faking good*, i.e. exaggerating a socially desirable behaviour. In this specific instance, the participants were instructed to lie in the sense of *faking good*: as they were asked to modify their responses in order to win a family custody case, they would clearly have attempted to make them more socially desirable.

The issue of lie detection in psychological questionnaires is not a new one - in fact, specialised answer scales, such as the Validity scale for the MCMI questionnaire (Andrews and Bender, 2020), are usually available and embedded in each questionnaire as part of its development. These scales obtain satisfactory results in detecting *faked* profiles for the questionnaires for which they have been developed. However, accuracy can be improved with machine learning methods such as the ones used in this paper. Moreover, these scales have no utility for the reconstruction of the original honest profile from that of a "faker", a research interest which has recently emerged in the field (Cardaioli et al., 2021).

The Parental Reflective Functioning Questionnaire

As stated by Cardaioli et al., faking in psychological tests must be considered within its specific domain. This work focused on lying within the context of a recently developed psychological questionnaire: the Parental Reflective Functioning Questionnaire (PRFQ). It was defined by its inventors as *"a brief self-report measure designed to assess parental reflective functioning (PRF), i.e. the capacity to treat the infant as a psychological agent"*(Luyten et al., 2017).

Reflective functioning, considered fundamental for the development of the individual, is strongly connected to the ability to perform *mentalization*, a term that refers to the individual's ability to *hold others' minds in mind*. Indeed, several studies have associated the absence of mentalizing capacity with psychopathologies such as depression and BPD. Furthermore, research focused on the origins of this capacity has highlighted the fundamental role of parents in the development of mentalizing and reflective functioning abilities. Based on this evidence, researchers have analyzed the concept of "Parental Reflective Functioning" (PRF), defined as *"the caregiver's capacity to reflect upon his/her own internal mental experiences as well as those of the child"* (Luyten et al.,2017). Moreover, disruptions in PRF can be very harmful to child development, both in the case of deficient and in that of excessive PRF, the latter also being referred to as "parental hypermentalizing" (Luyten et al., 2017).

Therefore, it is realistic to assume that a family court would be interested in knowing the PRF levels of a parent before assigning them custody of their offspring. In this specific setting, which is the one reproduced in the dataset this work was based on, it becomes particularly important to be able to discriminate between fakers' and honest profiles. This is also a good example of a setting in which the reconstruction of the "true" underlying profile from a "fake" one may be of interest, since the reconstructed profile may show psychopathological tendencies of interest to the court.

Therefore, we attempt both at identifying the malingered profiles and the modified responses within those, and at reconstructing honest responses from the faked ones.

Exploratory Data Analysis

Description of dataset

The dataset we worked on is comprised of 678 rows and 18 columns. The columns correspond to the 18 questions posed in the PRFQ. The rows correspond to the 339 participants who were asked to respond to the questionnaire twice: the first time, they were asked to respond honestly; the second time, they were asked to fake good, i.e. to give socially desirable responses. We will refer to the two conditions as "Honest condition" (H) and "Dishonest condition" (D).

In condition D, responders were instructed to imagine that they were going through a Family Court evaluation in order to determine custody of their children. They were asked to answer items in such a way as to look like a good parent.

By design, the questionnaire is divided into 3 sections (PreM, C, IC), made up of 6 questions each. Past studies carried out on PRFQ showed evidence that these 3 *factors* are orthogonal, meaning that they don't present a clear correlation between each other (Luyten et al., 2017, pp.7-8). A brief description of the three factors is reported below.

- **PreM Q1-Q6**: these questions regard the Pre-mentalizing state. A high score is not desirable, as it indicates a parent who is not able to mentalize with their child. *Medium-low scores are socially desirable*.
- **C Q7-Q12**: these questions regard certainty of a parent with respect to the mental state of their child. Here, *desirable scores are the medium ones*, indicating a parent who is neither too certain or too uncertain of their child's mental state.
- **IC Q13-Q18**: these questions try to assess how curious and interested a parent is with respect to their child's internal experiences. Here, *desirable scores are the medium-high ones*, for the same reasons as above.

Preliminary observations

This section is dedicated to the description of the distribution of the answers in the PRFQ questionnaire and to the description of what can be expected from honest and dishonest responders, both in terms of variation in the value of the answers and of correlation between them.

The error bars in *Figure 1* represent approximately a 95% confidence interval for the mean answer of that profile. We can see that dishonest responders tend to exaggerate their answers, thus resulting overly certain and invasive with respect to their children's inner thoughts. Answers with the bigger separation between the average honest and dishonest profiles (e.g. PreM3, C1, C3) are expected to be the most important features for classification.

It has to be noted that, when the participants were instructed to lie, the amount of lying requested wasn't specified. Thus, every participant was free to decide how much to lie and in which direction. A stacked bar plot of the changed answers is shown in *Figure 2*.

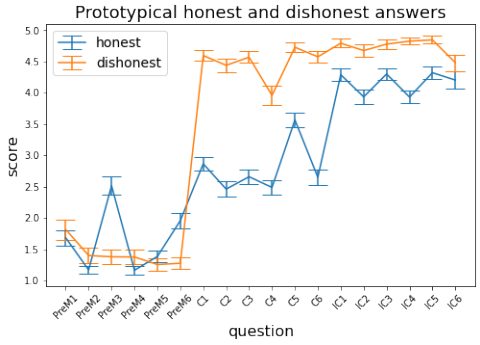

Figure 1. Comparison between average H and D profiles

Moreover, we noticed that most participants opted to lie about half of the time, while very few people decided to lie all the time or

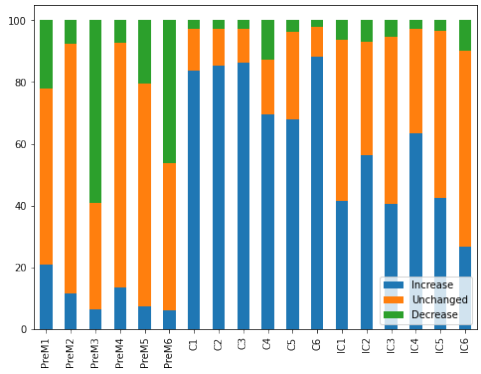

Figure 2. Overview of percentage of changed answers

almost never. Most of the changed answers for the dishonest profiles are changed in the opportune direction (*faking good*). However, some answers are changed in the opposite direction, posing a challenge especially to the reconstruction of the honest profile.

We also reported in *Figure 3* the correlation matrices between the honest and the dishonest responders. For the honest profiles, in accordance to the previous studies mentioned (Luyten et al., 2017 and Mazzeschi et al., 2019), answers belonging to different factors present very weak or null correlation, while answers within the same factor range from a weak to medium correlation. Dishonest profiles, on the other hand, present stronger correlations both between groups and within groups. This is likely due to the strong polarization of the dishonest answers in the perceived optimal direction.

Classification

We begin by framing our first purpose as a binary classification problem, employing answers to PRFQ for the classification of honest and dishonest responders. All the work presented has been carried out using Python programming language and, specifically, the Scikit-learn library.

In particular, we selected six classes of classification methods: Support Vector Machine (SVM), Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbours (KNN) and Neural Networks (NN).

For each algorithm we performed a 10-fold cross-validation and a grid-search. Values of the hyperparameters and the obtained accuracy scores are reported in *Table 1*.

In addition to high accuracy, we attempted to develop explainable classification algorithms. To this aim, we preferred methods which allowed us to clearly identify which features weighted more in the classification process. Due to their particular structure, Decision Trees and Random Forests were identified as the most explainable models. *Decision Trees* are built by splitting the data multiple times according to certain cutoff values in the features, in order to maximize Infogain. *Random Forests* are an ensemble of Decision Trees obtained through a bagging procedure. *Figure 4* shows the structure of the Decision Tree with the decision rules used in the splitting procedure and the feature importance histogram (i.e. a measure indicating the relative importance of a certain feature in the classification procedure) for the Random Forest.

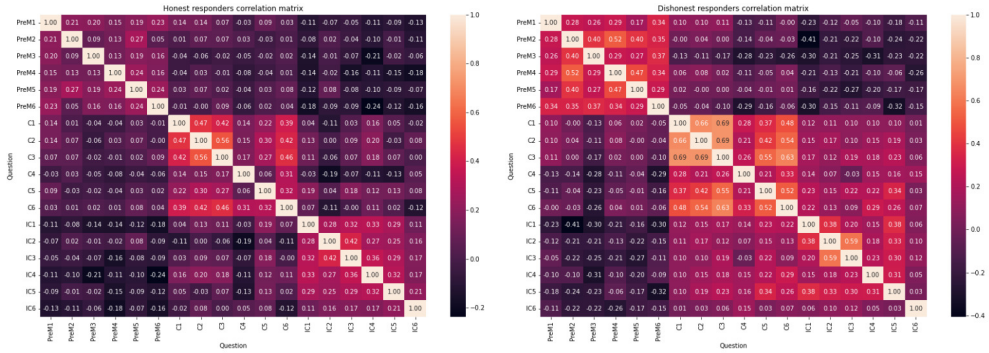

Figure 3. Correlation matrices: correlations between honest responders on the left; between dishonest on the right.

Model	Parameters	Values	Accuracy
Logistic Regression	c	[0.1,**1**,5,10,15,20,25,50,100]	90.6
SVM linear	c	[0.001,0.01,0.03,0.1,1]	---
SVM RBF	c γ	[0.001,0.01,0.03,0.1,1] [0.001,0.01,0.1,1,10]	---
SVM poly	c degree	[0.001,**0.01**,0.03,0.1,1] [2,3,**5**]	91.1
Random Forest	n. of estimators max n. of features max depth	[5, 10, 25, **50**, 100] [**2**,4,6,8] [**2**,4,6]	91.0
Decision Tree	max n. of features max depth min samples leaf	[1,**2**] [1,2,**3**,4] [1,3,**5**]	86.9
KNN	n. of neighbours	[1, 2, ... , **13**, ... , 100]	91.6
Feed-forward NN	n. of hidden layers n. of units (neurons) optimizer dropout values	[1,**2**,3] [512, 1024, 2048] [**SGD (Nesterov)**, ADAM] [0.1,0.2,**0.3**,0.4]	89.2

Table 1. Hyperparameters considered for grid-search and accuracy scores for classification

Based on maximum differences between H and D profiles in *Figure 1*, on the decision rules and on the feature importance plot in *Figure 4*, we concluded that 5 questions are the main contributors to the classification procedure: *PreM3 (inverted),C1,C2,C3,C6*. Dishonest answers to questions belonging to group C (certainty w.r.t. one's child's mental state) present significantly higher scores w.r.t. honest ones. While this behaviour most likely relates to the fact that responders wish to appear more considerate, it ends up simulating an obsessive attention to their child. Conversely, dishonest answers to PreM3 (*"I find it hard to actively participate in make believe play with my child."*) present significantly lower scores w.r.t. honest ones. However, values of PreM3 answers have been inverted in the construction of the classification score (see Algorithm below), due to the nature of the question. Hence, the lying behaviour is in line with that observed in group C: responders in condition D exaggerate their willingness to participate in their child's mental activities.

By summing up the values of the five questions, we obtained a "classification score" that, according to a certain threshold, directly classifies between honest and dishonest responders. The simple arithmetic for this procedure is shown below.

Algorithm: Classification score

if $(C1 + C2 + C3 + C6 + (6 - PreM3)) > threshold$ **then**
  $Subject \leftarrow Dishonest$
else
  $Subject \leftarrow Honest$
end if

The optimal threshold was obtained by applying a cross-validation-like procedure, using as thresholds values in the range [5,25]. We found that the best threshold corresponds to 20 with a reported accuracy of 88.9%. While this result is slightly below the best accuracy obtained in the classification section, it has the important advantage of being easily interpretable, like a sort of *validity scale*, by non-experts.

Lie removal: whole-profile and single-question level

In this section, we move on to the second objective of this paper: reconstructing, starting from a participant's dishonest profile, the "true" honest profile or single answer, i.e. what the participant would have answered if she hadn't lied. We divided attempted approaches between those which use raw answer scores from the questionnaire and those which use scores preprocessed via TF-IDF (for details, refer to the TF-IDF section). All approaches were evaluated via comparison of RMSE and of accuracy of the reconstruction. The benchmark used for comparison is the averaging method, which adds the difference between the typical honest responder and the typical dishonest one to every dishonest profile. It is to be noted that responders to the questionnaire could only give integer responses on the Likert scale (1-5). Hence, results obtained through regressors are rounded to the closest integer in order to allow accuracy calculation.

Raw-scores methods

Three approaches were applied to PRFQ raw answer scores: multi-output regressors, question-specific regressors and question-specific classifiers (indeed, we can consider the single honest response reconstruction task as a 5-class classification problem).

- **Multi-output regressors** (MOR) take as input the complete 18-dimensional questionnaire and give as output its complete honest reconstruction. Seven different MORs were trained: SVM, Random Forest, KNN, Ridge, Lasso, GBM and Autoencoder. Note that the type of regressor and the hyperparameters used are the same for all questions.
- **Question-specific regressors and classifiers**, on the other hand, still take as input the complete questionnaire, but give as output only a single reconstructed question. It is fair to say they are "specialized" in the single question they're dedicated to, thus allowing more flexibility: we can choose 18

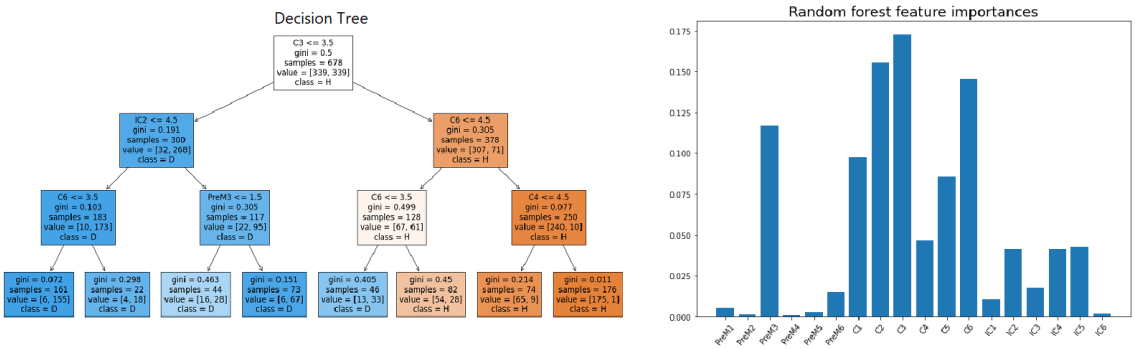

Figure 4. Decision rules for Decision Tree and Relative feature importance for Random Forest

classifiers/regressors among different classes of algorithms and with different hyperparameters (e.g. the reconstruction of PreM1 is done via a Decision Tree, while that of C6 is done via SVM). Of course, this is done at the cost of explainability and computational resources.

Model	Parameters	Values
SVR linear	c	[0.001, 0.01, 0.1, 1, 10]
SVR **RBF**	c γ	[0.001, 0.01, 0.1, **1**, 10] [0.001, 0.01, **0.1**, 1, 10, 100]
SVR poly	c degree	[0.001, 0.01, 0.1, 1, 10] [2, 3, 5]
Random Forest	n. of estimators max n. of features max depth min n. of samples/leaf	[5, 10, 20, **50**] [**2**, 4, 6, 8] [2, 4, **6**] [4, 6, **8**]
KNN	n. of neighbours	[2, 3, 4,..., **48**, 49, 50]
Ridge	α	[100, 150, 200,..., **750**,..., 1500]
Lasso	α	[0.010, 0.012,..., **0.08**,...,0.1]
Gradient Boosting	n. of estimators η (learning rate) max depth min n. of samples/leaf	[**50**, 100, 150, 250] [**0.01**, 0.1, 0.3] [2,4] [4,6,8]
Autoencoder	n. of hidden layers optimizer batch size η (learning rate)	[1,**3**,5] [sgd, **Adam** [8, **16**, 32, 64] [0.01,**0.03**,0.1,0.3]

Table 2. Hyperparameters considered for grid-search

Hyperparameters and metrics for the evaluation of all three approaches are shown in *Table 2* and *Table 3*, respectively.

Note that we included RMSE in *Table 3* as a metric, in addition to accuracy, for both single-question level and the whole profile. In fact, looking at question-specific models, one could obtain high accuracy levels for reconstruction just by using classifiers. However, classifiers' errors - although they happen less frequently - can be very large (e.g. predicting "5" for a question whose answer was "1"), while regressors' errors, albeit more frequent, tend to be much smaller. A consequence of this is that while the accuracy for classifiers is higher, the RMSE of regressors is lower. Thus, both were included for evaluation.

All methods applied do improve performance with respect to the benchmark. In particular, question-specific models show better performance with respect to MORs.

Before moving on to the next section, it must be noted that in all methods applied so far, we always assumed that all answers of a dishonest profile shall be reconstructed. However, as evident from *Figure 2*, for many responders this assumption is not satisfied, as they only change a portion of their answers. In an attempt to solve this issue, we decided to construct other procedures using TF-IDF.

TF-IDF transformed scores methods

Given a dishonest profile, we wanted to be able to understand if the subject is lying for a specific question rather than her profile overall. In order to do so, two different methods based on TF-IDF were tried. TF-IDF, short for Term Frequency – Inverse Document Frequency, is a score often used in information retrieval and NLP to indicate how important a term is in a document, part of a larger corpus. The general idea is to highlight words which are often used within a document but rarely so in the corpus. The score, in practice, is computed by considering the relative frequency of the term within the document (TF) and its inverse frequency within the collection of documents (IDF).

In our case, given a responder's answer to a question, we've computed the TF score by considering the frequency of that answer in the whole profile, while the IDF score was computed as log(N/n), where N was the number of participants (339) and n was the number of times that answer was given by honest responders to the same question.

Identification of changed answers

The first method for identifying faked responses consisted in simply multiplying TF and IDF scores, computed as above. The idea is to identify a threshold TF-IDF value, over which we assume the answer has been faked. We select the faking detection thresholds as the percentile score in the TF-IDF values distribution for each questionnaire item that maximizes the precision for faking. This turned out to be the 89th percentile.

Note that, since each of the questions considered shows a different TF-IDF distribution, the actual TF-IDF threshold individuated was different for every question - even if the percentile is common across answers. An accuracy of about 68% and a precision of 80% in the detection of faked answers was achieved by this method. However, the accuracy obtained per question actually varies widely, as shown in the first row of *Table 4*. Moreover, note that if an honest profile is fed by mistake into the pipeline, only about 12% of the answers are classified as faked, further demonstrating the robustness of the method.

While it makes this first procedure computationally cheaper, we

Model	PreM1	PreM2	PreM3	PreM4	PreM5	PreM6	C1	C2	C3	C4	C5	C6	IC1	IC2	IC3	IC4	IC5	IC6	Mean Accuracy	RMSE	rounded RMSE
No Change	56.90	80.80	34.50	79.10	72.30	47.50	13.30	12.10	11.20	17.70	28.30	9.70	52.20	36.60	54.30	33.90	54.00	63.70	42.12	6.83	6.83
Averaging	56.90	80.80	20.40	79.10	72.30	24.50	33.30	38.90	31.90	18.60	28.30	35.70	27.40	33.30	54.30	42.20	28.90	63.70	42.81	4.81	5.03
SVR	51.30	88.20	17.70	89.70	72.30	31.00	37.40	34.80	31.50	30.10	30.10	40.40	44.00	33.60	44.20	47.80	49.80	50.70	45.81	4.22	4.38
Random Forest	32.70	89.40	19.50	90.90	67.60	25.70	39.80	38.60	34.50	30.10	34.30	38.30	32.20	34.20	30.10	45.10	31.60	18.30	40.72	4.09	4.31
KNN	46.90	89.40	22.40	90.90	68.20	25.40	39.50	35.40	35.70	31.00	34.30	37.70	32.80	34.50	43.30	45.10	35.70	45.10	44.07	4.11	4.32
Ridge	23.30	89.40	22.10	90.90	67.60	26.00	39.80	35.70	35.40	28.30	33.10	37.40	32.20	33.90	30.10	45.10	32.20	13.90	39.80	4.09	4.32
Lasso	45.70	89.40	23.30	90.90	67.80	26.20	39.50	36.90	34.80	29.80	32.20	37.10	32.20	34.50	30.10	45.10	32.20	29.20	42.05	4.09	4.30
Gradient Boosting	15.30	89.10	20.40	90.00	71.10	25.70	39.80	34.20	35.10	29.50	32.80	36.80	32.20	34.50	30.10	44.80	32.20	13.00	39.26	4.10	4.36
Autoencoder	23.60	88.80	23.00	88.50	72.00	26.00	37.40	37.20	36.00	32.40	35.20	37.10	32.80	35.10	33.90	46.00	41.60	14.80	41.19	4.14	4.36
Best single question classifiers	65.80	89.40	34.80	91.10	76.70	48.10	41.60	38.40	38.90	34.50	36.00	42.80	53.70	41.60	54.30	49.90	56.60	65.20	53.30	4.94	4.94
Best single question regressors	45.90	91.80	22.40	94.10	74.10	27.10	45.90	36.50	36.50	23.50	45.90	37.60	35.30	37.60	32.90	47.10	42.40	18.80	44.19	4.03	4.20

Table 3. Accuracy scores and RMSEs for all raw scores lie removal methods; best results are highlighted.

Method	PreM1	PreM2	PreM3	PreM4	PreM5	PreM6	C1	C2	C3	C4	C5	C6	IC1	IC2	IC3	IC4	IC5	IC6
Classic TF-IDF + threshold	69.4	87.1	37.6	95.3	78.8	42.4	74.1	74.1	76.5	63.5	60	80	56.5	43.5	52.9	56.5	60	72.9
Separate TF-IDF + classifiers	77.9	92.1	66.2	93.7	80.3	57	85.5	90.2	88.6	83.5	70.9	90.6	58.2	63.8	59.5	69.3	63.4	73.2

Table 4.TF-IDF lie detection accuracy at single question-level for classical and separate TF-IDF methods.

Method	PreM1	PreM2	PreM3	PreM4	PreM5	PreM6	C1	C2	C3	C4	C5	C6	IC1	IC2	IC3	IC4	IC5	IC6	Mean accuracy	RMSE
Full profile reconstruction	60	**87.1**	**32.9**	**85.9**	**72.9**	38.8	**48.2**	**36.5**	**37.6**	30.6	**30.6**	**34.1**	51.8	**37.6**	**54.1**	45.9	40	60	**49.14**	**4.89**
Partial profile rec.(TF-IDF 1 st method)	56.9	**87.1**	30.3	**85.9**	71.2	38.6	43.2	**36.5**	32.1	**32.6**	**30.6**	26.4	**52.4**	29.1	44.5	45.9	38.5	54.2	46.44	5.21
Partial profile rec.(TF-IDF 2 nd method)	**61.2**	**87.1**	30.6	**85.9**	70.6	**40**	40	**36.5**	**37.6**	22.4	**30.6**	**34.1**	50.6	35.3	**54.1**	**49.4**	**45.9**	**61.2**	48.51	**4.89**

Table 5. Profile reconstruction accuracies and RMSEs for all methods applied. Best results are highlighted in bold.

argue that finding a unique percentile might cause the performance to drop, and that the classic formulation of TF-IDF leads to a further loss of information with respect to single answers. In fact, we know from previous sections that classifiers and regressors are more efficient when they are "specialized" in a single question. Therefore, we propose as a second method that TF and IDF be used as separate features to train a binary classifier (rather than multiplied) for every question, in order to determine if the responder is faking that specific answer. We employed the same methods used in the previous sections and selected the best 18 classifiers, based on the accuracy of the prediction. Results are shown in the second row of *Table 4*. When compared to the results of the previous method we see an improvement over almost all questions.

Lie removal via TF-IDF features

Finally, we apply the TF-IDF methods to the general lie removal problem. In fact, if we were able to correctly predict often enough if a question has been faked, we could reconstruct only the answers deemed faked rather than all of them. Of course, the viability of this procedure largely depends on the performance of the anomaly detectors described in the previous section.

Single-question classifiers were applied for the reconstruction of the answers which the anomaly detectors recognized as faked. Results are shown in *Table 5* (note that here we didn't use 10-fold CV but training/test set, hence results slightly vary from *Table 3*).

Unfortunately, the proposed procedure does not seem to improve the accuracy of the reconstruction of the profile with respect to methods which don't use anomaly detectors in a preliminary phase. However, while classic TF-IDF performs poorly, our second method achieves similar accuracy and same RMSE as those obtained using best-performing methods for whole-profile reconstruction.

We argue that this performance could be greatly improved if future works managed to construct, perhaps via other variations of the TF-IDF score, a more powerful single-question anomaly detector to be used in the preliminary phase.

Conclusion

In this work, we develop a procedure for the successful detection of lying in the PRFQ, also creating an heuristic for non-experts. In addition, significant improvements in honest profile reconstructions are obtained with respect to the averaging benchmark, particularly using specialized single-question models. Notably, TF-IDF based methods along with single-question classifiers can perform on par with complete profile reconstruction, but with fewer questions reconstructed. Possible improvements in the anomaly detector procedure, such as variations of TF-IDF, might in the future make this approach even more robust.

References

- Andrews, Jennifer, and Bender, Sara. "Millon Clinical Multiaxial Inventory (MCMI)." The Wiley Encyclopedia of Personality and Individual Differences: Measurement and Assessment (2020): 287-292.
- Cardaioli, Matteo, et al. "Malingering Scraper: A Novel Framework to Reconstruct Honest Profiles from Malingerer Psychopathological Tests." International Conference on Neural Information Processing. Springer, Cham, 2021.
- Luyten, Patrick, et al. "The parental reflective functioning questionnaire: Development and preliminary validation." PloS one 12.5 (2017): e0176218.