Automated Detection of User Deception in On-line Questionnaires with Focus on Eye Tracking Use

Metod Rybar
Slovak University of Technology, Faculty of Informatics and Information technologies, Ilkovicova 2, 842 16 Bratislava 4, Slovakia
Email: metod.rybar@stuba.sk

Maria Bielikova
Slovak University of Technology, Faculty of Informatics and Information technologies, Ilkovicova 2, 842 16 Bratislava 4, Slovakia
Email: maria.bielikova@stuba.sk

***Abstract*—On-line questionnaires are today widely used for various tasks, from census data collection to knowledge testing in job interviews. However, there is currently no automated system that can help us decide if the answers from the questionnaires are reliable or estimate how reliable the are. Deception is a part of everyday human behavior and deception is also present when answering on-line questionnaires. People are trying to make themselves look better or are just withholding information for malicious reasons. In our paper we present a method for automatic prediction of honesty for answers in a questionnaire. We demonstrate that by using new technologies like eye-tracking, we can create an automated system which can help us estimate reliability and truthfulness of the answers from on-line questionnaires. In our paper we have proposed and evaluated several metrics that can be used for automated detection of user deception in on-line questionnaires and we have also created and tested our first automated system for deception detection, based on these metrics.**

I. Introduction

Lying is part of everyday life. According to studies, people lie at least once or twice per day. Most common lies are about personal preferences and feelings, but people are also lying about their actions and plans, or achievements and pitfalls. People tend to lie more often if they can get psychological reward from the lie and less often if they are trying to avoid punishment [1].

People are not only lying in interactions with other humans, but also when they are filling out questionnaires. Studies have shown that people are trying to make look themselves better in personality questionnaires, which are used in job interviews [2].

Questionnaires are today widely used for various reasons, from assessing personalities using psychological questionnaires, to collecting consumer preferences or opinions about websites. There is also trend of moving these questionnaires from paper to on-line environment, which makes filling out of the questionnaires easier and also can lead to their quicker evaluation. This also helps the conductors of the questionnaire survey to collect data from more users, as the distribution of on-line questionnaire is lot easier than distribution and collection of printed questionnaires.

However, it is difficult to decide if the answers from the questionnaires are valid and truthful. To help with this problem, questionnaires often employ methods like asking the same question differently several times, which makes the questionnaires more time demanding for users. Moreover, the questionnaires with various alternatives of the same question should be extensively tested, which includes many users who fill the questionnaires before their actual use. This is possible with test inventories used in psychology, but almost impossible in the Web environment full of various questionnaires.

If there was a method which could help us decide if the answers from the questionnaires are truthful, it would help us reduce the size of the questionnaires and also potentially help us achieve better results from data analysis of the answers.

Better results from questionnaires answered by users could also lead to improved personalization for users in cases, where users are not honest while answering our questions, for various reasons.

In our paper we have proposed a method for automatic processing of data collected by eye trackers from users answering on-line questionnaires, which leads to the prediction of honesty for the answers. These predictions can then be used to decide if the questionnaire as a whole is trustworthy and e.g. can be used in some user study or other research as reliable source of answers. We evaluated our method on Big Five personality traits questionnaire implemented in on-line questionnaire system in a user study with 50 participants.

II. State of the art of deception detection

According to meta-analysis which analyzed 116 of different studies with 120 different samples [1], there are at least 158 different metrics that were already tested to decide, if they can be used for deception detection. However, most of these have only weak or none links to the deception detection.

Researchers were trying to develop methods for deception detection for a long time. First breakthrough came with the invention of polygraph. However, there was not much development since. Most changes came to developing better methods of questioning. There are currently three widely used methods: (i) Control Question Technique, (ii) Guilty Knowledge Test and (iii) Concealed Information Test. However, the accuracy of these questioning techniques is varying according to different studies from 37 up to 90 percent, depending on the skill of the investigator and used technique [3]–[5].

When it comes to more modern methods of deception detection or methods currently in development, we can mention

as example relatively new method which uses video analysis of user movement when answering questions. In this method, the metrics used to differentiate between honest and deceiving answers consist mostly of head and hands movement. This method also uses machine learning techniques to calculate a model of deception, which can automatically classify the answers as truthful or deceptive. They achieved accuracy up to 82 percent using model trained with neural network [6].

To sum up, there are existing methods of deception detection and also methods in development which can be quite accurate if used properly. However, none of these methods can be easily used to evaluate answers from on-line questionnaires, because they use an interaction with an investigator, who is asking specifically constructed questions. Also many metrics currently used cannot be easily collected from users who are filling out on-line questionnaires as they require physical presence.

Therefore, we see a need of more research into this topic, to create metrics, which can be used for deception detection in environment of on-line questionnaires, i.e. can be applied to machine learning models, which could help us detect deceptive answers collected via on-line questionnaires automatically.

III. Data collection and metrics proposal

To use our proposed method of automatic deception detection, we need reliable data for training of the model. Because we have found no suitable datasets available, we have proposed our own method for data collection and have created our own dataset of answers labeled as truthful or deceptive, together with metrics linked to them.

Our criteria for these data were that they should be good indicators of cognitive load, as cognitive load is correlated with deception, and that the data should also be usable to create additional metrics that indicate deception. These data should also be easily collectible from users when they are filling out on-line questionnaires e.g. sitting behind computer.

Based on our research we have picked as our main source of data an eye-tracker [7], which can provide us with information about the region of the screen the user is looking at in great detail, his fixations and saccades, response times and also magnitude of pupil dilation. It was demonstrated in [2] that eye-tracking can serve as deception indicator. Pupil dilation is also good indicator of deception and cognitive load according to several studies [8]–[10].

To collect this implicit feedback, we have created system for on-line questionnaires. We decided to make the user interface as simple as possible and to use bigger elements on the screen, to make detection of user fixations more precise [11]. We have also kept contrast and brightness of the screen stable during the experiment to prevent it from affecting pupil dilation. Interface used in our experiment is shown in Figure 1.

To achieve the situation that a person would try to deceive us when answering questions in a questionnaire, we have decided to put participants of the experiment into a position of job interviewee, who should try to make himself look as ideal candidate for hypothetical job. This approach was also used in psychological study presented in [2].

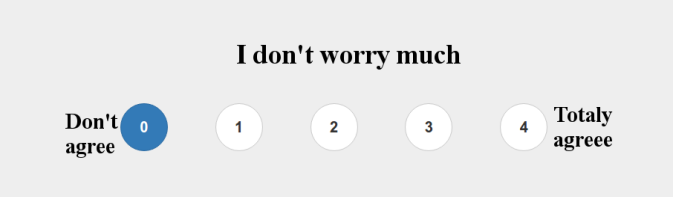

Fig. 1. Interface used for the experiment

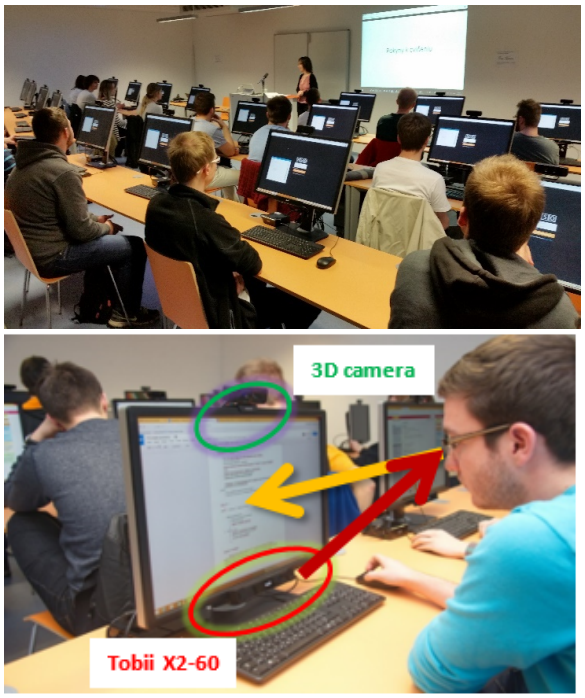

Fig. 2. UX group lab at the Research Centre of User Experience and Interaction

We have divided participants into two separate groups. Each participant was given two tasks, but each group received the instructions in switched order, to avoid the influence of instructions order on the resulting dataset.

To put some time between the questionnaires the participants were answering and to make participants believe, that the answers from the questionnaires were not the main objective of the experiment, participants have been doing movie classification task between the two questionnaires.

The tasks for the questionnaires were defined as follows:

1) Answer the following questions so you look as good as possible.
2) Answer the following questions as truthfully as possible.

These instructions put an incentive on the users to answer deceptively (task number 1) or truthfully (task number 2). We have also asked each user after reading the instruction, to describe verbally how will they answer the following questions. These answers were later used to filter answers in cases, when users were not answering as expected.

To collect the data, we used the infrastructure of the Research center of user experience and interaction (http://uxi.sk)

available at the Faculty of Informatics and Information Technologies of the Slovak University of Technology in Bratislava. It consists of the laboratory with 20 computers equipped with Tobii Pro X2-60 and software for data collection from all the nodes (see Figure 2). These eye-trackers work at 60Hz frequency, so they are capable to collect 60 gaze points per second. Collected data include also information about pupil dilation.

The software infrastructure collects data synchronously from the eye-tracker, the questionnaire system and other peripheries [12]. This was used to tag the eye-tracker data with events of the participants interactions with the questionnaire system. Along with the eye-tracker data, we have also collected video streams from the participants screens and keyboard and mouse interactions of each user.

To calculate the fixations and saccades from eye-tracker data we have used EyeMMV toolbox [13] with parameters set to

- t1=0.08
- t2=0.1
- maximal duration = 0.150
- maxX = 1
- maxY = 1

These values were calculated based on sizes of the elements in the interface of our questionnaire system. Raw data processed into fixations and saccades were then used to create metrics usable for deception detection.

Pupil dilation metric. We have created a metric based on pupil dilation. The metric is created from averages of the pupil dilation each 400 milliseconds. This time was chosen based on previous research mentioned in [9]. From these averages, we have calculated metric that consist from dilation difference of the pupil – 800 milliseconds before answering the question and 1 200 milliseconds after the answering occurred (most values were set according our experimental results).

Number of fixations while answering a question. Another metric based on data from eye-tracker was a number of fixations during the time spent by the user on the question.

Longest fixation while answering a question. Next metric which uses data from eye-tracker was the longest fixation that occurred during the time participant was looking on the screen with the question.

Average duration of fixations. We have calculated also the average duration of fixations. This was calculated from duration of all the fixations that were recorded from users during the time the question was on the screen.

The first fixation for particular answer. More specific metric to our research of deception detection in on-line questionnaires was the first fixation recorded on the answers. In this case, we have recorded different values based on the answer on which the first fixation was detected from the user. The value of 3 was recorded for the most extreme answers (0 and 4 in the questionnaire, see Fig 1); the value of 1 for neutral answer (2), and value of 2 for answers 1 and 3 from the questionnaire.

Response time. As the last metric we have used response time of users. We have calculated the time it took the participant to answer the question from the time the question was shown to him on the screen to the time he has actually answered it.

We have experimented with more metrics and with various variants of the metrics presented above. We present however only those metrics, which showed to contribute deception detection.

IV. Evaluation of proposed metrics

We have performed statistical analysis of our proposed metrics based on data collected in the experiment with 50 participants. The experiment took place during two weeks in several sessions in the UX group lab.

We proposed several hypotheses, but we include only those that were confirmed by the statistical analysis:

1) In deception condition, the participants look more often on extreme answers
2) In deception condition a difference in pupil diameter is bigger
3) In deception condition reaction times are faster
4) In deception condition number of fixations are lower
5) In deception condition longest fixation is longer
6) In deception condition average fixation duration is longer

The statistical analysis of metrics significance for above hypotheses was done using standard Mann-Whitney statistical test for not normal statistical distribution, as all the metrics did not comply with parameters normal distribution.

The results of the statistical analysis are summarized in Table I. All metrics presented in the hypotheses achieved statistical significance $p < 0.05$ in Mann-Whitney statistical test (no metric showed the normal distribution). Also all p-values, except of p-value for average fixation duration (hypothesis 6), achieved strong significance of $p < 0.001$. Thanks to the statistical analysis we could confirm all our hypotheses presented in this paper.

TABLE I. Results of statistical analysis of chosen metrics

Hypothesis number	Mann-Whitney p-value
1.	$56.32E-60$
2.	$567.62E-132$
3.	$2.54E-09$
4.	$9.19E-09$
5.	$217.31E-06$
6.	$14.04E-03$

We have used metrics presented in our hypotheses to create a vector for each answer of each question from the questionnaire from every user. We have created several vectors based on statistical significance of these metrics.

We have used these vectors to train Support vector machine (SVM) with RBF kernel. We have decided to use SVM(RBF) because of its good results in classification problems involving two classes and few metrics. RBF kernel is well suited for our data space, as the metrics are not spread linearly.

Polynomial kernel is also well suited for such data, but has higher complexity and it takes much longer to train model with higher accuracy. We wanted to avoid time necessary for training to allow quick creation of deception models for various questionnaires.

We have also tested several other machine learning methods including SVM with linear and polynomial kernel, decision trees or Naive Bayes, but all of them preformed worse in accuracy, as expected.

We have tested our model with various vectors of metrics, using cross validation to check its accuracy and other parameters. Results of these tests are shown in Table II, Table III and Table IV. We have trained our model for vectors containing all metrics with $p < 0.05$, for vectors containing all metrics with $p < 0.001$ and for each of the metric on its own. For accuracy of the model with one parameter we present only the best one, the model created with the first fixation on answer in Table IV.

TABLE II. SVM MODEL RESULTS FOR VECTORS $p < 0.05$

Accuracy	Precision truth	Precision deception	Recall truth	Recall deception
0.62	0.54	0.57	0.60	0.51

TABLE III. SVM MODEL RESULTS FOR VECTORS $p < 0.001$

Accuracy	Precision truth	Precision deception	Recall truth	Recall deception
0,62	0,67	0,61	0,51	0,75

TABLE IV. SVM MODEL RESULTS FOR THE FIRST FIXATION VECTOR

Accuracy	Precision truth	Precision deception	Recall truth	Recall deception
0,62	0,54	0,56	0,60	0,50

As we can see from our results, the model was the most successful using only strongly statistically significant metrics where $p < 0.001$. Interesting were results for the first fixations on answers on its own, where this model achieved the same accuracy as model with all strongly statistically significant metrics, but it did not achieve the precisions of this model.

These results show that statistical significance of metrics is important for creation of our model, and stronger significance leads to better results. The result of the first fixations metric also shows that some metrics are more successful on their own and are therefore more suitable for the model creation than others. It is clear from the statistical analysis and the model training that higher statistical significance generally leads to better preforming model of deception detection.

The deceptive answers are detected by putting vectors of metrics linked to answer we want to test into SVM model. As we had marked the answers as deceptive or truthful based on the instruction that participants were given, we could test if the SVM prediction was correct.

In metric creation, we had most issues with creating usable metric for pupil dilation, as the pupil has very short reaction times. At the end, the reaction was the strongest just as the users were answering the questions in deceptive condition and this metric has shown good correlation to deception.

Lowest statistical significance was showed for average fixation duration. Averages of the duration of fixations are probably not the best indicators of deception, as users tend to change fixation duration between reading question and answering it. This was indicated by maximal duration of fixations, which has shown much stronger statistical significance.

Strongest correlation was achieved in fixation order, when statistically significantly more users were looking first at extreme questions in deceptive condition. This was also demonstrated on the model precision, as the model created using only this metric was close to precision of our best model. This also confirms that statistical significance of a metric has impact on the model precision.

V. CONCLUSION

In this paper we have demonstrated that it is possible, in the environment of on-line questionnaires, to obtain data and extract metrics (based on gaze) from them, which can be used to detect deception. If these data are correctly processed, they can also be used to train machine learning model, which can detect deceptive answers automatically.

Accuracy of our model is not yet sufficient for reliable deception detection, but with more metrics retrieved and added to our model we expect it to get better. It is important to point out that we can rate every single answer from questionnaire as truthful or deceptive on its own, so if we want to tell if entire questionnaire is answered mostly truthfully, lower accuracy is needed – if lot of the answers are flagged as deceptive, the probability that the whole questionnaire was not answered truthfully is higher. In near future we therefore see more potential in using our or similar models to evaluate questionnaires as a whole instead of flagging concrete answers as dishonest.

Lower accuracy of our model is caused by several factors. As each person is an individual, their physiological signals are slightly different and therefore comparison of different users is difficult. We have used several normalization techniques in the process, but there is still room for improvement.

Other problem with accuracy comes from the fact that we cannot be 100 percent sure if users were indeed answering the questions as instructed. We have taken several steps to insure this, as instructing them in different ways or asking them to describe how they were answering in their own words, but there were surely several answers that were not in the right category for our training and testing sets, and therefore this had definitely impact on the accuracy of our model. We will need to take more precautions and find new ways for data collection to avoid this uncertainty.

We also feel the need to mention moral aspects of deception detection, which could be written down in its own paper. Therefore, we only mention that our or similar methods, if used not correctly or without proper understanding of what they indicate, can lead to unfortunate conclusions.

Psychological studies detect deception by using statistical analysis of questionnaires from thousands of people and creating set of questions that will show if participant is answering honestly or deceptive based on the way in which these thousands of people before had answer the questions.

Our approach would require much less participants to create deception models.

We see potential in future research of this topic and will work on our model to make it more accurate. We plan to also use data from skin conductance or skin temperature. We also plan to add more metrics extracted from eye-tracker.

Additional to collecting data by putting people into artificial conditions, we also work currently on data collection in collaboration with psychologists. They help us to create truthful and deceptive datasets without putting people into artificial conditions, as this can influence data. So we will collect needed data just by letting enough people fill out the questionnaires with psychologically proven deception indicators.

We also work on developing new methods of data collection for training of our model, in which we would not instruct users to answer in a certain way, but we would decide if their answers were truthful or deceptive by external factors. One such method could let users answer simple questions about which we know that the users know the answer, but they would have incentive to answer deceptively e.g. for financial gain. By this we could clearly separate truthful and deceptive answers for machine learning training purposes.

We also consider for our future work on additional metrics using data from galvanic skin response detectors and hearth rate detector sensors, that are also available in our lab. We believe, that using additional metrics can improve our model, because during our experiments and evaluation more statistically significant metrics lead to better accuracy of our model.

We also believe that using new methods of machine learning as deep neural networks can also improve our method, as the data from users are typically very spread out in the mathematical dimensional space, which can be destructive for typical methods as SVM, and neural networks, given enough data, can be used to improve accuracy of the model despite these obstacles.

ACKNOWLEDGMENT

This work is partially supported the grants APVV-15-0508, VG 1/0646/15 and it is the partial result of the Research and Development Operational Programme for the project University Science Park of STU Bratislava, No. ITMS 26240220084, co-funded by the European Regional Development Fund."

REFERENCES

[1] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception." *Psychological bulletin*, vol. 129, no. 1, pp. 74–118, 2003.

[2] E. a. J. van Hooft and M. P. Born, "Intentional response distortion on personality tests: Using eye-tracking to understand response processes when faking." *Journal of Applied Psychology*, vol. 97, no. 2, pp. 301–316, 2012.

[3] G. Ben-Shakhar and E. Elaad, "The validity of psychophysiological detection of information with the Guilty Knowledge Test: a meta-analytic review." *The Journal of applied psychology*, vol. 88, no. 1, pp. 131–151, 2003.

[4] C. J. Patrick and W. G. Iacono, "Validity of the control question polygraph test: The problem of sampling bias." *Journal of Applied Psychology*, vol. 76, no. 2, pp. 229–238, 1991.

[5] B. Verschuere, G. Crombez, E. H. W. Koster, and A. De Clercq, "Antisociality, underarousal and the validity of the Concealed Information Polygraph Test," *Biological Psychology*, vol. 74, no. 3, pp. 309–318, 2007.

[6] T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon, J. F. Nunamaker, D. P. Twitchell, G. Tsechpenakis, and D. N. Metaxas, "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," *IEEE Intelligent Systems*, vol. 20, no. 5, pp. 36–43, 2005.

[7] M. Barla, M. Simek, and M. Bielikova, "Comparing Eye-tracking Data Using Machine Learning," *Journal of Eye Movement Research*, vol. 8, no. 4, p. 192, 2015.

[8] W. Steptoe, A. Steed, A. Rovira, and J. Rae, "Lie tracking," *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, p. 1039, 2010.

[9] J. Wang, M. Spezio, and C. F. Camerer, "Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games," *American Economic Review*, vol. 100, no. 3, pp. 984–1007, 2010.

[10] J. Wang, "Pupil dilation and eye tracking," in *A handbook of process tracing methods for decision*, 2011, pp. 1–33.

[11] R. Moro, J. Daraz, and M. Bielikova, "Defining Areas of Interest for the Dynamic Web Pages," *Journal of Eye Movement Research*, vol. 8, no. 4, p. 190, 2015.

[12] R. Moro, D. Jakub, and M. Bielikova, "Visualization of gaze tracking data for ux testing on the web," in *Porceedings of DataViz 2014: Data Vizualisation Workshop (associated with Hypertext 2014), CEUR. Hypertext Extended Proceedings. Vol. 1210.*, 2014.

[13] V. Krassanakis, V. Filippakopoulou, and B. Nakos, "EyeMMV toolbox: An eye movement post-analysis tool based on a two-step spatial dispersion threshold for fixation identification," *Journal of Eye Movement Research*, vol. 7(1), no. 1, pp. 1–10, 2014.