

The Peer Data Labelling System (PDLs). A Participatory Approach to Classifying Engagement in the Classroom

Graham Parsonage^{1,2}, Matthew Horton¹, and Janet Read¹

¹ University of Central Lancashire, Preston, UK

² University of the West of Scotland, Paisley, UK

gbparsonage1@uclan.ac.uk, graham.parsonage@uws.ac.uk,

mplhorton@uclan.ac.uk, jcread@uclan.ac.uk

<https://chici.org/>

Abstract. The paper introduces a novel and extensible approach to generating labelled data called the Peer Data Labelling System (PDLs), suitable for training supervised Machine Learning algorithms for use in CCI research and development. The novelty is in classifying one child’s engagement using peer observation by another child, thus reducing the two-stage process of detection and inference common in emotion recognition to a single phase. In doing so, this technique preserves context at the point of inference, reducing the time and cost of labelling data retrospectively and stays true to the CCI principle of keeping child-participation central to the design process. We evaluate the approach using the usability metrics of effectiveness, efficiency, and satisfaction. PDLs is judged to be both efficient and satisfactory. Further work is required to judge its effectiveness, but initial indications are encouraging and indicate that the children were consistent in their perceptions of engagement and disengagement.

Keywords: data labelling · artificial intelligence · engagement.

1 Introduction

Learning is a complex process which relies on many factors, not least the skill of the teacher in maintaining pupils’ attention to their learning activities so that they complete any set tasks. As children use more technology in the classroom, it becomes enticing to consider what an intelligent system might be able to do independently to keep a child engaged on a task. In this study we explore the extent to which pupils can assist in the design of such a system and their acceptance of its judgments.

In Child Computer Interaction (CCI) it is common to engage children in design activities. In our study we “employ” children as labellers of data by using their expertise to decide if a peer is engaged on task or not. We consider this to be a novel approach to assist in training a recogniser. Our contributions include reflections on the approach taken, survey findings indicating pupils’ level

of acceptance of such a method and a data set that others in the CCI community can use and develop. Validation of the children’s judgments is currently ongoing and is not included in this study.

The paper proposes a novel and extensible approach to generating labelled data suitable for training supervised Machine Learning (ML) algorithms for use in CCI research and development called the Peer Data Labelling System (PDLS). The novelty is in classifying one child’s engagement using peer observation by another child. This reduces the two stage process of detection, (the capture of the data) and inference, (the latter coding of the data) common in emotion recognition to a single synchronous phase. In doing so, this technique preserves context at the point of inference, reduces the time and cost of labelling data retrospectively and stays true to the CCI principle of keeping child-participation central to the design process. We evaluate the approach using the usability metrics of effectiveness, efficiency and satisfaction.

1.1 Learning and Engagement

Pupil engagement is widely considered to be a positive factor in, and an important driver of, pupil attainment [3]. Multiple definitions of engagement exist [12] but for the purpose of this study, we consider engagement on task, namely a pupil’s interaction with a computerised learning activity completed within a school classroom. Whilst school age education in the UK has largely returned to the physical classroom, the Covid-19 pandemic fast-forwarded the development and adoption of hybrid and blended learning pedagogical approaches [30]. This created new requirements for tools and techniques that can aid teachers in monitoring and interpreting pupils’ level of engagement with academic tasks both online and in the classroom.

1.2 Approaches to Recognising Children’s Engagement

The study of children’s understanding of emotions based on facial expressions and other stimuli is well researched [13], [29]. Children start to be able to discern emotion from an early age [9] and are also able to differentiate between contexts of expressions, for example they can understand that a parent crying at a TV drama is not the same as one crying following an injury [23]. Hence we argue that context is an important factor on the accuracy of children’s recognition and classification of emotion [28].

A popular and established system for emotion recognition is the Facial Action Coding System (FACS) [11]. One drawback to FACS is the considerable training required which at the time of writing is estimated by the Paul Ekman Group to be between 50 and 100 hours [10]. An alternative approach commonly used both in academia and commercially is to automate the emotion classification process using algorithms such as AFFDEX [20], [1] or FACET [19]. Whilst the algorithmic approach has the potential to save considerable time, there is concern that current emotion recognition systems are less accurate than their human counterparts when employed on children [2].

1.3 Existing Data Sets for Machine Learning that Include Children

Specialised child-centered data sets are relatively scarce. Princeton University Library have curated a directory of databases containing face stimulus sets available for use in behavioural studies of which just four are specific to children [24]. This lack of material restricts the options for CCI researchers looking for data as a starting point on which to train their models.

1.4 Machine Learning and Child Computer Interaction

There is a rich vein of work within the CCI Community enshrining child participation as core to a child-centered design process [15], [27], [8], [25]. Hourcade [14] organises the key principles of CCI research into ten pillars, the second of which, “Deeply engage with stakeholders” enshrines the principle of child participation as the core of a child-centred design process. At a time where a growing number of academic studies are exploring ML based systems and intelligent interfaces both within the CCI community [26], [7], [22] and the wider HCI community [17], [4], [6]. We propose an approach to data labelling that makes child participation intrinsic not only to the development of the system but also core to the system’s outputs.

2 Studies

Two studies were conducted at a single UK secondary school (ages 11 - 16). The aim of the first study was to generate video data that captured the engagement status of children while they completed a computerised task in a classroom. Values for the engagement status of the child completing the task were recorded synchronously by peer observation effectively reducing the two stage operation of detection and inference to a single stage operation while maintaining context during inference and in a time and resource effective manner. The second study assessed the children’s experience of, and confidence in, the data labelling process and a theoretical system based on its output.

2.1 Participants

Forty-five pupils took part in the studies. Twenty-two children, (12 boys and 10 girls) aged between 11 and 15 took part in the first study and a further twenty-three children, (10 boys, 13 girls) aged between 11 and 12 took part in the second study. Prior to the study commencing, written consent was obtained from the school, parents or carers, and the pupils. The pupils were also advised that they could withdraw their data after completing the task regardless of any previous consent given by themselves or third parties. No incentives or rewards were offered to the children who took part in the study.

2.2 Apparatus

Three artefacts were prepared for the studies, the first was a website of material about cryptography. The material was designed to support at least 15 minutes of activity which was the time allocated for each child to interact with the cryptography webpage and was deemed, by the teachers, to be suitable for children within an eleven to fifteen year age range.

The second artefact was an online form with a drop-down list that allowed the (child) observer to log the engagement level of the pupil completing the cryptography task. Using the form, the observer recorded the engagement level as; engaged (interested and working) or disengaged (disinterested or distracted). When the observer felt that the learner had changed engagement category they then logged the updated value.

The final artefact used only in the second study was a short paper based questionnaire. Pupils completed the questionnaire to gauge their feelings about the logging process. Pupils were asked:

1. How accurately they thought their classmate had judged their engagement level whilst completing the task
2. How accurately they thought they had judged their classmate's engagement level whilst completing the task
3. How accepting they would be if a system was utilised in the classroom to monitor their engagement level
4. To what degree would they trust the system to identify disengagement

A Likert scale ranging from 1 - 10 was used to rate the pupils' responses where 1 equated to low and 10 equated to high. For instance for Question 1, a recorded score of 1 would indicate that the pupil thought the accuracy of their classmate's judgement of their engagement level was low whilst a score of 10 would indicate a perceived high accuracy of judgment.

2.3 Procedure

The children worked in pairs each taking turns at being the learner and the observer switching roles half way through the study. The learner completed the online task on their laptop. The observer was positioned so that they could watch the learner completing the task but could not see their laptop screen and logged the learner's engagement status. The importance of the logging process was emphasised to the children as having equal importance to the computerised task.

For the second study, after completing the online task, the children were asked to complete the questions and record any other observations about the study.

3 Results

The first set of studies produced 22 videos of which 17 were usable. 2 videos were discarded as they had audio but no image frames and 3 videos were complete but

had no engagement statuses recorded. The 17 usable videos and engagement logs yielded 2 hours, 33 minutes and 48 seconds of video of which 2 hours, 27 minutes and 32 seconds has labels generated from the pupil logs. This resulted in 221,300 labelled JPEG images. The observers logged 57 instances of an engaged status totalling 2 hours, 12 minutes and 33 seconds yielding 198,825 labelled images. Forty-four instances of a disengaged status were logged totalling 14 minutes and 59 seconds yielding 22,475 images. The average duration of an instance of learner engagement was 2 minutes and 20 seconds and the average duration of learner disengagement was 20 seconds. The frequency of the logged data ranged from a single recording of engaged through to 26 recorded statuses ($M = 3.35$, $SD = 3.6$).

Time spent on the task ranged from 2 minutes and 24 seconds to 20 minutes and 13 seconds ($M = 09:03$, $SD = 05:26$). The logged duration ranged in time from 2 minutes and 8 seconds to 19 minutes and 52 seconds ($M = 08:41$, $SD = 05:28$). Six minutes and 16 seconds of video were discarded as they had no logging status. The majority of the discarded data occurred at the beginning of the videos in the period after the learner had started the video camera generating the starting timestamp and before the observer recorded their first engagement status.

In addition 22 questionnaires were completed from the second study the results of which are presented in Table 1. For a discussion see Section 4.1 Satisfaction.

4 Discussion

4.1 Evaluating the Usability of the Process

The stated aims of this paper were to introduce a novel and extensible approach to generating labelled data suitable for training supervised ML algorithms for use in CCI research and development which were then evaluated using the usability metrics effectiveness, efficiency and satisfaction outlined in ISO 9241-11 [16].

Efficiency We judge PDLS to be both a time and cost efficient system that compares favourably against the options considered. FACS coding by human experts requires both extensive training and a has a considerable time and cost overhead. PDLS labels the data at the point of capture using peer judgments thus avoiding these pitfalls. Algorithmic implementations such as AFFDEX and products that implement them such as iMotions can be configured to perform evaluations in real time but are considerably more costly than PDLS which requires no specialist equipment other than a laptop and a camera both of which are relatively low cost and freely available. PDLS is extensible and suitable for gathering and labelling data concurrently.

Satisfaction Children indicated their satisfaction with both their own and their peers effectiveness in reaching a classification and the potential of a system built

Table 1. Children’s Responses to Survey Questions (scale 1 - 10)

Classmate’s Judgment	Own Judgment	Acceptance of System	Trust in System
9	9	7	10
8	9	7	10
9	10	8	7
10	10	10	10
9	9	5	5
6	9	8	7
3	8	7	3
5	5	3	3
8	9	4	5
5	8	4	4
8	10	5	4
8	6	4	4
8	10	9	9
8	7	4	6
9	9	9	9
10	10	8	6
9	9	9	9
9	10	9	9
6	5	4	4
10	10	6	8
10	-	-	-
10	10	7	8

upon data from the study to make effective judgments. They expressed confidence in their own ability to accurately measure the engagement level of their classmate (R2). When asked to rate the accuracy of their judgements on a Likert scale of 1 to 10 where 1 is not accurate and 10 is very accurate, the average recorded score was 8.667 (SD = 1.623). They were marginally less positive about the ability of their classmate to assess their own engagement levels whilst still expressing confidence (M = 8.045, SD = 1.914) (R1). The children were also asked how accepting they would be if a system were deployed to monitor their level of engagement in the classroom and how trusting they would be in the accuracy of its judgements. The children were neutral to accepting of the proposed system (M = 6.523, SD = 2.159) (R3) and its predictions (M = 6.666, SD = 2.456) (R4) with both scores lower than their confidence in their own and their peers ability.

Effectiveness Evaluating the effectiveness of PDLs is challenging and requires further work, however the initial signs are promising. The children’s judgments appear to be consistent and there are few outliers in the data indicating that the classifications are cohesive and the children are measuring the same phenomena. Whilst we can’t say with certainty that the children’s judgments are correct, a random sample of ten of the 44 videos that were classified as disengaged indicates that in the majority of cases the learner is exhibiting behaviour which may

show disengagement or distraction from the task (Table 2). Certainly their focus often appears to be elsewhere. The exception may be video 212 where although the learner appeared amused by something there is no obvious indication that they were not engaged. Study 212 had the most statuses recorded across both categories, (26 for a logged duration of 11 minutes and 18 seconds), or one every 26 seconds on average with an average duration of ≈ 7 seconds for each logging of disengagement. As such it is feasible that the observer’s judgements were not in line with the other children.

Table 2. Characteristics of Children’s observations of disengagement

Study ID clip	Observation of Behaviour
171.2	The learner appears distracted and looks away from the screen
172.4	The learner is laughing
173.4	The learner is talking and hits out at someone off camera
196.2	The learner is laughing and appears distracted
212.12	The learner is smiling and scratching their head
212.24	The learner is smiling but appears to be working
213.6	The learner is smiling and scratching their ear
213.10	The learner is smiling and looks away from the screen in parts but appears to be working
219.1	The learner is talking and looking away from the screen
237.2	The learner appears to be working but is holding a conversation unrelated to the task

4.2 A Child-Centred Process

Our final stated objective was to stay true to the CCI principle of keeping child-participation central to the design process. In using the children’s own classifications to generate the data set, they become central not just to the design process but also to the operation of a system built using that data set. They are in effect judging themselves. Firstly, they classify each others level of engagement in the classroom using the PDLS method. The labelled data is then used by the system to learn about engagement, this learning process is entirely dependent on the children’s classifications. Once operational the system monitors the children in the classroom and uses what it has learnt from them to classify their engagement level. As such, PDLS not only uses the children’s judgment to label the data but by the very nature of the supervised machine learning process their participation and input will form the basis of future system development and deployment.

4.3 Data Bias, Authenticity and Future Work

Data bias is a recurrent theme in ML literature [21], [18] and beyond. In the UK in 2020 there was uproar that the algorithm designed to predict exam results

was unfair and disadvantaged students from certain demographics resulting in teachers predicting grades [5]. As Intelligent systems become increasingly embedded into society it is an inherent responsibility of designers and developers to ensure that the decisions made by the technology are fair. When making this point we note that the data collected for this study is produced from a single computerised task in one school and the output from any ML model built based on this data will reflect these limitations.

To address these limitations further studies should reflect children's diverse backgrounds increasing the scope of the data set and therefore the quality of the judgments produced by ML models trained upon it. In addition, the scope and circumstance of the observed tasks can be extended to provide new context to the observations. Whilst the work to date has involved a computerised task and webcam it is feasible that judgments could be recorded of children completing more traditional activities which do not involve computers.

5 Conclusion

This paper presents PDLs, a peer observation approach to generating a labelled data set suitable for use in CCI research. The system is evaluated against the usability metrics, effectiveness, efficiency and satisfaction and is judged to be both efficient and satisfactory. Further work is ongoing to judge its effectiveness but initial indications are encouraging and indicate that the children were consistent in their perceptions of engagement and disengagement. The CCI principle of Child Participation is central to the PDLs process which generates labelled data in both a time and cost effective manner. Children were surveyed for their feelings on the accuracy of both their own and their peers' judgment of engagement status after completing the task and expressed their confidence in both these aspects.

6 Acknowledgments

We would like to thank the Head Teacher, staff and pupils of Ribblesdale High School and in particular the Head of Computer Science, Mr Steven Kay for their invaluable assistance and participation in this study.

References

1. Bishay, M., Preston, K., Strafuss, M., Page, G., Turcot, J., Mavadati, M.: Affdex 2.0: A real-time facial expression analysis toolkit. arXiv preprint arXiv:2202.12059 (2022)
2. Bryant, D., Howard, A.: A comparative analysis of emotion-detecting ai systems with respect to algorithm performance and dataset diversity. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 377–382 (2019)
3. Christenson, S., Reschly, A.L., Wylie, C., et al.: Handbook of research on student engagement, vol. 840. Springer (2012)
4. Chromik, M., Butz, A.: Human-xai interaction: A review and design principles for explanation user interfaces. In: Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18. pp. 619–640. Springer (2021)
5. Coughlan, S.: Why did the a-level algorithm say no? (August 2020), <https://www.bbc.co.uk/news/education-53787203>, accessed on 06.01.2023
6. Desolda, G., Esposito, A., Lanzilotti, R., Costabile, M.F.: Detecting emotions through machine learning for automatic ux evaluation. In: Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18. pp. 270–279. Springer (2021)
7. Dietz, G., King Chen, J., Beason, J., Tarrow, M., Hilliard, A., Shapiro, R.B.: Artonomous: Introducing middle school students to reinforcement learning through virtual robotics. In: Interaction Design and Children. p. 430–441. IDC '22, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3501712.3529736>, <https://doi.org/10.1145/3501712.3529736>
8. Druin, A.: The role of children in the design of new technology. *Behaviour and information technology* **21**(1), 1–25 (2002)
9. Durand, K., Gallay, M., Seigneuric, A., Robichon, F., Baudouin, J.Y.: The development of facial emotion recognition: The role of configural information. *Journal of experimental child psychology* **97**(1), 14–27 (2007)
10. Ekman, P.: Facial action coding system (Jan 2020), <https://www.paulekman.com/facial-action-coding-system/>
11. Ekman, P., Friesen, W.V.: Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978)
12. Groccia, J.E.: What is student engagement? New directions for teaching and learning **2018**(154), 11–20 (2018)
13. Gross, A.L., Ballif, B.: Children’s understanding of emotion from facial expressions and situations: A review. *Developmental review* **11**(4), 368–398 (1991)
14. Hourcade, J.P.: Child-computer interaction. Self, Iowa City, Iowa (2015)
15. Inkpen, K.: Three important research agendas for educational multimedia: Learning, children, and gender. In: AACE World Conference on Educational Multimedia and Hypermedia. vol. 97, pp. 521–526. Citeseer (1997)
16. Iso - international organization for standardization. iso 9241-11:2018(en) ergonomics of human-system interaction — part 11: Usability: Definitions and concepts (2018), <https://www.iso.org/obp/ui/>, accessed 0n 18.01.2023
17. Jasim, M., Collins, C., Sarvghad, A., Mahyar, N.: Supporting serendipitous discovery and balanced analysis of online product reviews with interaction-driven metrics and bias-mitigating suggestions. In: Proceedings of the 2022 CHI Conference on

- Human Factors in Computing Systems. CHI '22, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3491102.3517649>, <https://doi.org/10.1145/3491102.3517649>
18. Jiang, H., Nachum, O.: Identifying and correcting label bias in machine learning. In: International Conference on Artificial Intelligence and Statistics. pp. 702–712. PMLR (2020)
 19. Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M.: The computer expression recognition toolbox (cert). In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG). pp. 298–305. IEEE (2011)
 20. McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., Kaliouby, R.e.: Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit. In: Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems. pp. 3723–3726 (2016)
 21. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* **54**(6), 1–35 (2021)
 22. Nguyen, H.: Examining teenagers’ perceptions of conversational agents in learning settings. In: Interaction Design and Children. p. 374–381. IDC '22, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3501712.3529740>, <https://doi.org/10.1145/3501712.3529740>
 23. Pollak, S.D., Messner, M., Kistler, D.J., Cohn, J.F.: Development of perceptual expertise in emotion recognition. *Cognition* **110**(2), 242–247 (2009)
 24. Databases (a-z) - face image databases - research guides at princeton university (Jan 2022), <https://libguides.princeton.edu/facedatabases>, accessed 0n 18.01.2023
 25. Read, J.C., Horton, M., Fitton, D., Sim, G.: Empowered and informed: Participation of children in hci. In: IFIP Conference on Human-Computer Interaction. pp. 431–446. Springer (2017)
 26. Rubegni, E., Malinverni, L., Yip, J.: “don’t let the robots walk our dogs, but it’s ok for them to do our homework”: Children’s perceptions, fears, and hopes in social robots. In: Interaction Design and Children. p. 352–361. IDC '22, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3501712.3529726>, <https://doi.org/10.1145/3501712.3529726>
 27. Scaife, M., Rogers, Y., Aldrich, F., Davies, M.: Designing for or designing with? informant design for interactive learning environments. In: Proceedings of the ACM SIGCHI Conference on Human factors in computing systems. pp. 343–350 (1997)
 28. Theurel, A., Witt, A., Malsert, J., Lejeune, F., Fiorentini, C., Barisnikov, K., Gentaz, E.: The integration of visual context information in facial emotion recognition in 5-to 15-year-olds. *Journal of experimental child psychology* **150**, 252–271 (2016)
 29. Widen, S.C.: Children’s interpretation of facial expressions: The long path from valence-based to specific discrete categories. *Emotion Review* **5**(1), 72–77 (2013)
 30. Zhao, Y., Watterston, J.: The changes we need: Education post covid-19. *Journal of Educational Change* **22**(1), 3–12 (2021)