Creating an openly accessible dataset of learning dialogue

Rebecca Ferguson
Institute of Educational Technology, The Open University
rebecca.ferguson@open.ac.uk

Kirsty Kitto
Connected Intelligence Centre, University of Technology Sydney
kirsty.kitto@uts.edu.au

Catherine A. Manly
City University of New York Graduate Center
cmanly@gc.cuny.edu

Oleksandra Poquet
Centre for Research and Interdisciplinarity, University of Paris & INSERM U1284
sasha.poquet@cri-paris.org

ABSTRACT: Content analysis (CA) is a widely used method in the learning sciences, and so has become a well-accepted practice in the domain of learning analytics (LA). Increasingly, we see datasets coded with CA used as labelled datasets to drive machine learning. However, the scarcity of widely shareable datasets means that research groups around the world work independently to code text using CA, with few attempts made to compare results across groups. A risk is emerging that different groups using the same constructs are coding them in different ways, leading to results that will not prove replicable. In this poster, we report on the development of an openly accessible database containing the discussion associated with an international online course on learning analytics, which ran for four weeks on the Slack platform. Participants were aware that their postings would form part of the database, and that any personally identifiable information would be pseudonymised. The database will be shared via GitHub and the SoLAR website to support the development of replicable work on content analysis of learning and teaching dialogue.

Keywords: content analysis, dataset, open access, replication

1 CHALLENGES TO REPLICATION IN LEARNING ANALYTICS

How do we know that the approaches emerging in learning analytics (LA) are valid? Replication is key to validating the theories and models developed using quantitative approaches. Many fields have long established procedures that support reproducibility. For example, data science research communities organize competitions to provide the best solution for core challenges. Specified baseline datasets are analysed using different approaches with the results compared. The released datasets help ensure reproducible research results, and a better understanding of state-of-the-art solutions. These datasets also support new entrants to the field, who can access data, examine how it has been labelled and then work to develop their own sophisticated analytical methods.

Although LA has attempted to develop similar approaches, shareable datasets are uncommon. Some exceptions include a *Journal of Learning Analytics* special section (Dietze et al, 2016) with information about four open datasets; the LAK data challenge (Drachsler et al, 2014); the Pittsburgh datashop[1]; and the release of MOOC data from Stanford[2]. Yet, beyond this, the public release of data is uncommon in LA, and what does get released fails to cover the broad range of learning activities that LA strives to model. Nonetheless, groups continue collecting, cleaning, and exploring learning data, making implicit decisions about how to process and analyse it (Buckingham Shum & Luckin, 2019). Since many such decisions that influence the results are not well documented, different research groups may plausibly make different decisions when cleaning and labelling similar datasets. This lack of openness is a problem, as theoretical constructs adopted in quantitative analysis in LA, particularly those developed in educational research, were derived from rigorous qualitative approaches. They are highly contextual and grounded by rich descriptions of the situation in which they were developed. Our ongoing inability to document the contextual data, in addition to the lack of open and shareable datasets, creates further ambiguity around details relevant to supervised and unsupervised approaches applied towards content analysis to capture particular theoretical lenses.

Reasons as to why shareable datasets are rare and limited in scope are important. Sets of clickstream, discourse, and engagement-pattern data are difficult for research groups to access and cannot be shared without breaching the privacy of individual learners. Yet, the question of replication persists - the ontological and epistemological differences between research methods are frequently overlooked.

2 CREATING AN OPENLY ACCESSIBLE LA DATASET

This poster reports on an effort to create a shareable open dataset that can be used as a baseline in our research community. We seek to generate a discussion about how this initial effort might be scaled. To generate our dataset, we created a short online course about the past, present, and future of learning analytics. We sought ethics approval from the University of South Australia to collect data that will become publicly available, generated through the interactions of the learners. According to the course objectives, its participants were 1) to identify ways in which learning analytics has developed over the past decade; 2) to identify significant challenges for learning analytics in the next five years; and 3) to discuss how work at their institution aligns with the challenges for learning analytics. The curriculum included short open-ended tasks released weekly, that required the participants to engage with short videos by experts in Learning Analytics and provide reflective responses. Each week built on the content from the previous week. The course was developed by four researchers, two – taking on explicit instructional roles in the course, and another two – taking on roles of participants to support emerging discussions.

The invitation to the course with explicit consent information was distributed to the participants of the Learning Analytics Summer Institute 2021. Participants were informed that if they register, they would be able to engage with the themes around learning analytics for four weeks on a closed, specially dedicated Slack channel. They were also informed that any text they share with each other

[1] https://pslcdatashop.web.cmu.edu/

[2] https://datastage.stanford.edu/StanfordMoocPosts/

will form the public dataset, and only their personal names will be replaced by pseudonyms. Fifty-four participants signed up for the course and were added to the private Slack channel. The Slack channel was changed to the public discussion within the LASI'21 community five days into the course, for technical reasons. Once the course was completed, the team manually collected discussion data from the private channel (Figure 1) and downloaded Slack channel data from the public discussion (19 json files of participant activity representative of 19 individual days). We have also recorded screenshots of the discussions to capture the interface of the course.

Such a pilot activity demonstrates how LA researchers can join forces to facilitate data collection to build a dataset that can facilitate replication efforts, particularly though not limited to, automated content analysis. We report statistics to describe the scope of the collected data.

Event_ID	User	TS	Date	Message	Thread	Msg_ID	Likes	Replies	Pinned	Type
1	Winnie The Pooh	2:28AM	3.09.2021	Hello there!You have signed up for the course (	1	1	0	0	1	Post
2	Cinderella	6:08PM	3.09.2021	I'll start the Activity 1 introductions. I'm a profess	2	1	2	0	0	Post
3	Little Red Riding Hood	NA	NA	NA	2	1	0	0	0	Like
4	Ariel	NA	NA	NA	2	1	0	0	0	Like
5	Little Red Riding Hood	2:44AM	6.09.2021	I'm a doctoral candidate at a public research ur	3	1	1	0	0	Post
6	Winnie The Pooh	NA			3	1	0	0	0	Like
7	Ariel	4:35AM	6.09.2021	Hi, I am a doctoral candidate in a research univ	4	1	0	2	0	Post
8	Winnie The Pooh	NA		Hi Ariel, am curious about your area of focus wi	4	2	0	0	0	Reply
9	Ariel	NA		Hello Winnie, right now we are designing dashb	4	3	0	0	0	Reply
10	Richard the Lion Heart	5:44AM	6.09.2021	Activity 1 - Introduction of myself: I'm a universit	5	1	0	0	0	Post
11	Richard the Lion Heart	6:15AM	6.09.2021	Activity 2 - LAK12: There were much fewer subr	6	1	0	0	0	Post
12	Phoenix bird	10:36AM	6.09.2021	Hello, I am Phoenix bird, Doctoral Student in In	7	1	0	2	0	Post
13	Winnie The Pooh	NA		Hi Phoenix, what aspects of multimodal learning	7	2	0	0	0	Reply

Figure 1. A Screenshot of the dataset constructed from the first five days of the course

The slack channel generated discussion data from 13 participants and 4 course designers. The dataset contains a total of 99 text messages, which comprised a total of 32 individual threads. The mean number of messages per participant was 5.8, with a median of 3. The highest number of messages was 18 – produced by one of the participants, closely followed by one of the instructors with 16 messages. The minimum message length was 52 characters, a median of 997 characters. The longest participant response was 4466 (for illustration purposes, the text on page two of this proposal is about 3000 characters).

Although our dataset is small for large-scale text analysis, this pilot demonstrates that the LA community can focus on generating ethical and shareable data – generating a resource that can grow over time and support replication efforts in automation of text, particularly around theoretical constructs. We hope to engage LA researchers in a conversation around which of the fields and what format will be most appropriate, to provide sufficient meta-data and document the dataset for release and public use by the LA community, via the website of the Society for Learning Analytics and Research. We also hope to spur discussions around future data collection efforts and a public store.

REFERENCES

Buckingham Shum, S. & Luckin, R. (2019). Learning Analytics and AI: Politics, Pedagogy and Practices. *British Journal of Educational Technology*.

Dietze, S., George, S., Davide, T., & Drachsler, H. (2016). Datasets for learning analytics. *Journal of Learning Analytics* 3(2), 307-311.

Drachsler, H., Dietze, S., Herder, E., d'Aquin, M., & Taibi, D. (2014). The learning analytics & knowledge (LAK) data challenge 2014. In *LAK14* (pp. 289-290). ACM.