

LXplore: An NLP-Based Tool for Distilling Learning Analytics and Learning Design Instruments out of Scientific Publications

Atezaz Ahmad¹^a, Jan Schneider¹^b, Daniel Schiffner¹^c, Esad Islamovic²
and Hendrik Drachsler¹^d

¹DIPF — Leibniz Institute for Research and Information in Education, Frankfurt, Germany

²Goethe University, Frankfurt, Germany

Keywords: Information Retrieval, Natural Language Processing, Learning Analytics, Indicators, Metrics, Learning Activities, Learning Design.


Abstract: Each year, the amount of research publications is increasing. Staying on top of the state of the art is a pressing issue. The field of Learning Analytics (LA) is no exception, with the rise of digital education systems that are used broadly these days from K12 up to Higher Education. Keeping track of the advances in LA is challenging. This is especially the case for newcomers to the field, as well as for the increasing number of LA units that consult their teachers and scholars on applying evidence-based research outcomes in their lectures. To keep an overview of the rapidly growing research findings on LA, we developed LXplore, a tool that uses NLP to extract relevant information from the LA literature. In this article, we present the evaluation of LXplore. Results from the evaluation show that LXplore can significantly support researchers in extracting information from relevant LA publications as it reduces the time of searching and retrieving the knowledge by a factor of six. However, the accurate extraction of relevant information from LA literature is not yet ready to be fully automatized and some manual work is still required.


1 INTRODUCTION


Over the past decade, the field of Learning Analytics (LA) has been widely cultivated and plays a crucial role in the development and advancement of education. Consequently, it triggered interest among the sister communities in the learning sciences, for example, didactics in STEM education (Kubsch et al., 2022), psychometrics (Drachsler and Goldhammer, 2020), neuroscience (Calle-Alonso et al., 2017). This broad interest in LA results in three challenges practitioners and researchers are facing. First, the increasing usage of LA in other fields also increases the **heterogeneity** of LA and an even more diverse publication landscape. With such an increase in heterogeneity, it is difficult for researchers and practitioners to identify evidence-based practices because they are scattered over many scientific outlets (Saqr et al., 2022). Furthermore, the terminologies used for LA


are also very heterogeneous, e.g., indicators and metrics are widely used interchangeably in the literature (Ahmad et al., 2022). The second challenge of LA concerns its **pedagogical alignment** and a lack of alignment with pedagogical models (Bakharia et al., 2016; Macfadyen et al., 2020). For example, LA can track a lot of data (digital traces) throughout the learning process. Yet, it remains difficult or even unclear how to effectively present these results back as meaningful LA indicators, so that third parties can improve the process of learning (Chatti et al., 2014; Ahmad et al., 2022). The third challenge consists of the growing use of LA within and beyond these communities. This increases the number of publications in the field, resulting in an **information overflow** (Bakharia et al., 2016; Martin et al., 2016) state. Therefore, it becomes increasingly difficult to stay on top of the state-of-the-art. For example, as per Google Scholar Metrics, and the time of writing, there are over 49,500 reviewed articles related to the keywords of *Learning Analytics* published between 2020 and 2023.

The work by (Ahmad et al., 2022; Ahmad et al., 2022) presented a solution to the pedagogical alignment and heterogeneity challenges through a frame-

^a <https://orcid.org/0000-0003-4894-8258>

^b <https://orcid.org/0000-0001-8578-6409>

^c <https://orcid.org/0000-0002-0794-0359>

^d <https://orcid.org/0000-0001-8407-5314>

work aimed at aligning LA and learning design (LD) in a sustainable manner. Adhering to this framework, the Open Learning Analytics Indicator Repository (OpenLAIR) (Ahmad. et al., 2022) provides a list of empirically tested LA indicators and their metrics for common learning activities that are part of a typical LD of a course. It enables course creators, educators, practitioners, and researchers to make informed decisions concerning LA and LD for their course designs and/or dashboards. The information presented by OpenLAIR consists of an ontology that connects LD events with LA indicators and metrics through LD-LA activities. The main elements of this ontology are **LD events** (e.g., receiving information or imitating an expert), **LD-LA activities** (e.g., reading or writing), **LA indicators** (e.g., feedback or engagement), and their corresponding **LA metrics** (e.g., time or initiative). In the remainder of the paper, we refer to these as *LD-LA Instruments*. Aligned with the third challenge mentioned above, a limitation of OpenLAIR is to keep up to date with the quickly evolving state of research. Manually extracting relevant information from the literature is time-consuming and difficult. To overcome this problem, we created LExplore, a natural language processing (NLP) extension for OpenLAIR, which aims to support the extraction of these *LD-LA Instruments* from LA articles.

In this paper, we present the evaluation of LExplore. We tested and evaluated it in two different scenarios. The first one is *automatic extraction*, where we compare the results of the NLP extension against the already human-harvested data. The second one consists of the *semi-automatic extraction*, where we ask experts in LA (N=10) to use LExplore to extract *LD-LA Instruments*. The presented work was guided by the following research questions:

RQ1: To what extent can we use LExplore to automatically and reliably extract *LD-LA Instruments* from a LA publication corpus?

In order to answer RQ1, we compare the analytics results from LExplore against human-identified results to calculate its accuracy by calculating the precision, recall, and f-score of the trained classifiers.

RQ2: To what extent can LExplore support the automatic or semi-automatic extraction of *LD-LA Instruments*?

To answer RQ2, we conducted user tests where we applied the System Usability Scale (SUS) (Brooke, 1986) and the Technology Acceptance Model (TAM) (Davis, 1985) to explore the perceived usability, ease of use, and usefulness of LExplore for practitioners and researchers.

2 RELATED WORK

The widespread use of LA in sister communities has made the field increasingly heterogeneous which makes it difficult to get an overview of the field and identify best practices. This challenge was already identified in 2016 by the Learning Analytics Community Exchange (LACE) project¹. The LACE project developed an LA evidence hub that aimed to provide an overview of effective and ineffective LA studies according to four propositions; whether they improve and support learning outcomes, improve learning support and teaching, are used widely, and are used ethically (Ferguson and Clow, 2017). However, in 2019 this initiative stopped due to the high human effort to maintain the evidence hub. In the meantime, the heterogeneity of the field increased even further. Moreover, there is still a need for a better alignment and presentation of *LD-LA Instruments* and pedagogy, to address the general pedagogical alignment problem concerning LA. For example, LA can track a lot of data about the learner and their environment but it is not clear how to use this data to identify relevant LA indicators that support the educational aims and competencies of the students (Chatti et al., 2014). With the increasing rollout of LA at higher education institutions also the amount of LA support units is growing. Those LA units consult their teachers and scholars on applying evidence-based research outcomes in their lectures. In the field of assessment, validated items and assessment instruments reliably measure the stage of knowledge of a student. Compared to LA, we are not at the stage to have a virtual shell with validated instruments that we can pull out and apply to a new learning context. There is a high demand for a tool that provides an overview of evidence-based *LD-LA Instruments* that the scholars can successfully apply in their lectures (Bakharia et al., 2016). A study by (Saqr et al., 2022), examines if and to what extent frequently used LA indicators of success predictions are portable across a homogeneous set of courses. Still, it is a challenge to find and present a suitable LA approach for an activity or a construct. The heterogeneous nature of LA and the terminologies used to represent them are reasons for this. For example, basic terms such as *indicators* and *metrics* are not formally defined and are mostly used interchangeably in LA publications. Consequently, finding a consensus concerning best LA practices and approaches is not trivial (Ahmad et al., 2022). To address these challenges, we have been creating OpenLAIR, which is organized according to the common LD-LA activities and provides an overview of evidence-based LA in-

¹<https://bildungsserver.de/bisy.html?a=8924&spr=1>

dicators that have been applied to these LD activities in the past. A current limitation of OpenLAIR is the capability of staying up to date with the quickly developing state-of-the-art on LA. Therefore, we aim to overcome this shortage with NLP technology; otherwise, it will suffer the same fate as the LACE evidence hub.

Keeping OpenLAIR up to date is not a unique challenge in the field of LA. Similar initiatives are emerging in other fields too. In the field of Biomedicine, new studies are published at an unmanageable pace. Every minute two papers are published on average in the field of Medicine (Fiorini et al., 2018). Due to the abundance and inconsistency of scientific literature, the field is currently facing a scientific crisis. Hence, it is increasingly strenuous, time-consuming, and challenging to find relevant articles for a query and extract the relevant metadata from the content of the paper(s) (Voytovich and Greenberg, 2022).

Automatically keeping track of the new knowledge being published is challenging, as the data provided in the articles is unstructured. Handling unstructured data is challenging because it is essential to pre-process it before the data can be utilized (Baviskar et al., 2021). We argue that applications of Artificial Intelligence (AI) such as Machine Learning (ML), Deep Learning, and NLP play a crucial role in automatizing the process of handling and processing unstructured data. Especially, NLP handles linguistic data or unstructured data efficiently by offering a variety of powerful and reliable methods (Ceylan, 2022). New advancements in AI have been made possible with AI toolkits. For example, TensorFlow provides a library of state-of-the-art models that can be used to create various scalable AI-powered applications. Azure Machine Learning Studio and IBM Watson Studio are further examples of AI toolkits.

The use of these toolkits leads to the creation of applications capable of sorting articles based on relevance. One example is ASReview (Hindriks, 2020), an open-source tool that helps researchers and practitioners to get an overview of the relevant publications and filter irrelevant articles during the first phase of a systematic literature search. Likewise, Litstudy (Heldens et al., 2022) is a Python-based package that is used to analyze scientific literature from the comfort of a Jupyter notebook. It provides metadata using visualizations, network analysis, and NLP for the selected scientific publications. Another example is Arxivbox², which is an AI-based web interface for browsing major computer vision and machine learning conference papers. EduBERT (Clavié and Gal,

2019) is another example of an NLP-based tool used to improve the classification of forum texts.

In the field of medicine, some NLP applications already handle the continuous stream of advances reported in the scientific literature. For example, SNPcurator (Tawfik and Spruit, 2018) is used to help in automatically extracting Single-Nucleotide Polymorphisms (SNP) associations of any given disease and its reported statistical significance. SNPcurator is based on NLP and text mining that further helps healthcare professionals in locating appropriate information from biomedical articles quickly and effectively. Similarly, BioBERT (Lee et al., 2020) is a pre-trained Bio-medical NLP model trained exclusively on large-scale Bio-medical corpora. It helps to understand complicated bio-medical texts. ExECT (Extraction of Epilepsy Clinical Text) (Fonferko-Shadrach et al., 2019) is yet another example that is able to extract epilepsy information from free texts in clinic letters. It further helps by storing patient data in a structured format. These NLP models enable the automated detection of contradictions in Bio-medical literature and may lower the inconsistency in the literature.

3 BACKGROUND

Both, the heterogeneity and alignment challenges are diminished by OpenLAIR. Nevertheless, both rely on an up-to-date representation. To address the information overflow challenge in LA and keep OpenLAIR up to date, we propose an NLP tool named LxExplore.

3.1 OpenLAIR

OpenLAIR³ is a web application that presents users with a structured approach for selecting evidence-based LA indicators for educational practice so that they can get an informed idea of how to implement LA in their courses based on their LD. The tool presents *LD-LA Instruments* in a structured and categorized manner, which are connected and aligned to LD-LA activities and LD events. OpenLAIR is aimed to support different types of users. Teachers can use this already tested/existing knowledge to select relevant learning activities that may lead students to understand the topic/course better. Researchers, architects, or programmers can use this knowledge to design a LA indicators dashboard using the metrics provided by OpenLAIR. Currently, the information presented by OpenLAIR is based on the literature review

²<https://github.com/ankanbhunia/arxivbox>

³<https://edutec-tool.github.io/>

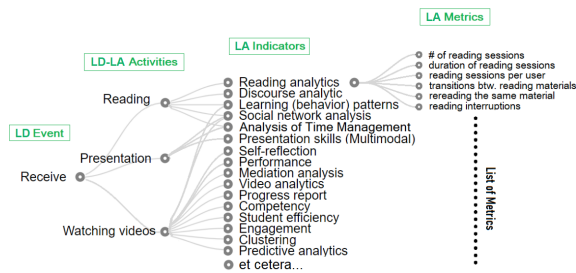


Figure 1: Tree view example of our JSON dataset (Ahmad et al., 2022).

in (Ahmad et al., 2022), which has been manually extracted out of 161 LA papers over ten years (2011-2020). The results of the literature review (Ahmad et al., 2022) suggested that learning activities in LA and LD can be aligned and create a common ground between LA and LD. Therefore, the authors looked for LD-LA activities in LA literature and harvested *LD-LA Instruments* manually.

The information displayed in OpenLAIR is deployed in a MongoDB Atlas⁴ cloud database that contains eight JSON documents based on the eight LD events (Verpoorten et al., 2007). Each event has *LD-LA activities* (e.g., reading), whereas each activity has LA indicators (e.g., reading analytics) and their metrics (e.g., number of views/keystrokes/reading sessions) (see Figure 1) assigned. The data stored in the database (DB) is hierarchical, where one LA indicator can be a subset of many learning activities (e.g., the indicator ‘learning behavior patterns’ lies under the *LD-LA activities* reading and watching videos) and one learning activity can be a subset of other LD events (e.g., the learning activity ‘group work’ lies under the LD events ‘create’, ‘practice’ and ‘debate’).

Apart from the OpenLAIR evaluation study (Ahmad et al., 2022) and its literature review (Ahmad et al., 2022), we have used and tested this tool in a few scenarios. One is brainstorming sessions to identify and recommend meaningful learning activities in Moodle courses. Another example is workshops tailored to designing and recommending LA indicators and their metrics for LA dashboards (Ahmad et al., 2023; Karademir et al., 2022).

Further, refer to Figure 2 for more details regarding OpenLAIR connection, working, and placement with the current and proposed LExplore system.

3.2 LExplore

LExplore⁵ is an NLP-based ML extension for OpenLAIR developed in Python (backend) and Streamlit⁶ (frontend) that extracts information from a Portable Document Format (PDF) LA research article and further identifies and categorizes keywords (*LD-LA Instruments*) from the extracted content based on instruments specified in its classifier. Figure 2 presents the system architecture of LExplore extending OpenLAIR. LExplore has four main functionalities: text extraction, pre-processing, training the classifier, and displaying the extracted and processed results out of new (untagged) documents (see Figure 2).

First, LExplore reads the PDF articles extracting the content/text from them. Next, it converts them into pickle⁷ format (a Python object serialization that converts an object into a byte stream).

The pre-processing phase consists of four sub-steps: 1. The extracted text is tokenized (to a list of words). 2. The stop words that do not contain any meaning (e.g., the, it, etc.) and special characters and numbers are removed. 3. The words are converted to their base form by removing affixes called stemming (e.g., eating, eats, eaten to eat). LExplore offers two stemming algorithms PorterStemmer⁸ (which works best with any type of word and the one we used for training) and WordNetLemmatizer⁹ (which works best with verbs and particularly nouns). 4. In the final preprocessing step, words are assigned labels to indicate their semantic role (relationship) with the help of our JSON dataset (Data.json used in OpenLAIR) (see Figure 2). All articles are linked to the LD-LA-Instruments in the database using a reference number. This reference number is utilized by the NLP classifier while assigning the labels.

The third phase consists of training the classifier. To train this classifier, the tool matches instruments/words from the DB of OpenLAIR to the extracted content/text from the previous steps and learns from it. The NLP classifier uses the Naive Bayes algorithm, which is a supervised ML learning algorithm also used for solving text classification problems (Zhang and Li, 2007; Gültekin and Bayat, 2022).

Lastly, the user provides an unknown LA article to display results. The tool processes its content and with the help of the classifiers, the predicted results are shown to the user. This ‘prediction’ value is a log-likelihood threshold indicating the accuracy (0 mean-

⁴<https://www.mongodb.com/atlas/database>

⁵<https://github.com/atezaz/LExplore>

⁶<https://streamlit.io/>

⁷<https://docs.python.org/3/library/pickle.html>

⁸<https://tartarus.org/martin/PorterStemmer/>

⁹https://www.nltk.org/_modules/nltk/stem/wordnet.html

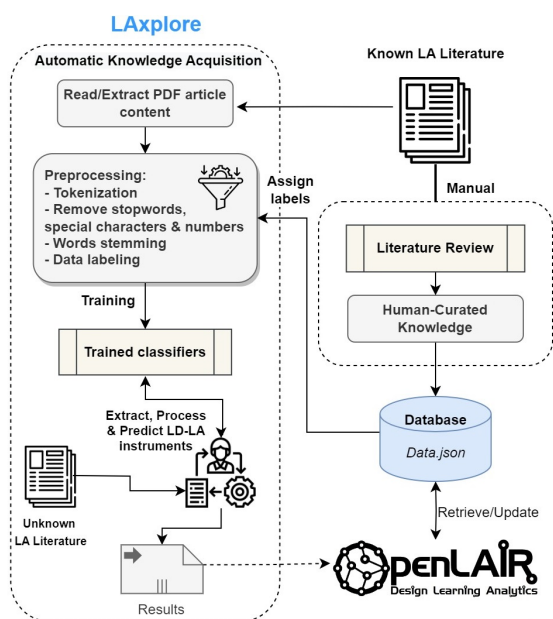


Figure 2: LAxplore system architecture with OpenLAIR.

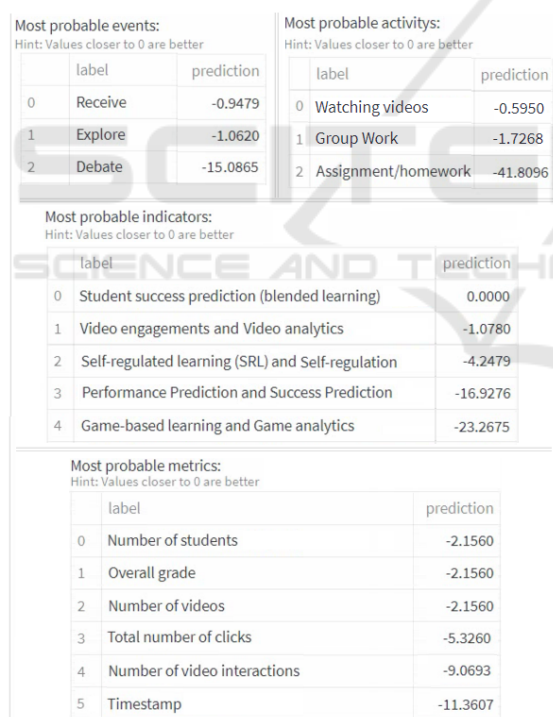


Figure 3: Example of LAxplore results view.

ing perfect accuracy) (see Figure 3).

4 METHODOLOGY

We evaluated LAxplore with the main aim to identify how such a tool can support the extraction of *LD-LA Instruments* from new LA literature. Furthermore, we assessed usability, ease of use, usefulness, and model accuracy.

For training the LAxplore’s classifiers, we used the same list of peer-reviewed LA publications (N=161) and data set harvested in OpenLAIR presentation and evaluation study (Ahmad. et al., 2022) and in the literature review (Ahmad et al., 2022) (see Figure 2), which contains articles related to Technology Enhanced Learning (TEL) from the Learning Analytics and Knowledge Conference (LAK) series from 2011 to 2020, publications in the Journal of Learning Analytics (JLA), Proceedings of the European Conference for Technology Enhanced Learning (ECTEL) from 2012 to 2020, IEEE Transactions on Learning Technologies, as well as special issues for LA in the Journal of Computer Assisted Learning (JCAL).

We conducted two different evaluations: *automatic extraction* and *semi-automatic extraction*. In automatic extraction, we trained two classifiers with two different year ranges. Classifier 1 was trained with *eight-years* (2011-2018) of publications (N=116). Classifier 2 was trained by adding two more years (*ten-years*) (2011-2020) of publications (total N=161). To test the classifier, we selected ten random, but still relevant articles from the proceedings of LAK 2021 and harvested the *LD-LA Instruments* manually (two persons involved). We tested the accuracy of both classifiers against the manually harvested data. Lastly, we calculate the recall, precision, and F-score in order to answer our first research question (RQ1).

For the evaluation of the semi-automatic extraction, we recruited ten researchers (four Postdocs and six PhD candidates) in LA and Technology Enhanced Learning. In this second evaluation, we focused on the usability, usefulness, and ease of use of LAxplore. Here, participants had to perform two tasks. The first task was to manually extract *LD-LA Instruments* from an LA article. In the second round, participants were presented with LAxplore and asked to extract of the *LD-LA Instruments* for a different article with the help of the tool. After each task participants filled a small survey regarding their experience of extracting this information (difficulty, motivation, etc.).

The materials used for this semi-automatic evaluation were two LA articles (Lim et al., 2021; Günther, 2021) that were randomly selected from a set of articles used in validating the classifier. Half of the participants worked with article (Lim et al., 2021) for the first task and with article (Günther, 2021) in

Table 1: *LD-LA Instruments* extraction results for two different training set of classifiers.

LD-LA Instruments	Classifier training set	Precision	Recall	F-score	Log-likelihood threshold
LD Events	1. 2011-2018	0.80	0.86	0.82	0 to -10
	2. 2011-2020	0.87	0.92	0.90	0 to -10
LD-LA activities	1. 2011-2018	0.84	0.75	0.79	0 to -10
	2. 2011-2020	0.96	0.82	0.89	0 to -10
LA Indicators	1. 2011-2018	0.67	0.63	0.65	0 to -10
	2. 2011-2020	0.77	0.73	0.75	0 to -10
LA Metrics	1. 2011-2018	0.69	0.85	0.76	0 to -20
	2. 2011-2020	0.72	0.78	0.75	0 to -20

Table 2: Survey items after tasks one and two.

Items after Task one	Mean (SD)	Items after Task two	Mean (SD)	T-test
The task was easy to do.	4.1 (2.38)	After using the tool, the task was easy to do.	6.1 (1.1)	0.0266
The task was boring to do.	5.2 (1.62)	After using the tool, the task was boring to do.	2.6 (1.65)	0.0022
I think there is a need for a Tool for doing such tasks or making them easy.	6.5 (0.71)	The Tool helped me to do the task and make the task easy and exciting.	6.2 (0.75)	0.2961

*7 (strongly agree) and 1 (strongly disagree)

*SD = standard deviation

the second task; the other half performed the task with the opposite order of articles. The questionnaire used after the second task included a System Usability Scale (SUS) (Brooke, 1986) and a Technology Acceptance Model (TAM) (Davis, 1985) questionnaires that helped us to evaluate the usability, usefulness, and ease of use of LExplore to support the semi-automatic extraction of *LD-LA Instruments* out of LA articles (RQ2).

5 RESULTS

Table 1 contains the results of the *automatic extraction*, where human-identified instruments (superclass) are compared against the results of both classifiers, and the mean (M) values are used for calculating the precision, recall, and F-score. Our results show that Classifier 1 performs better than Classifier 2. The increase of data (*LD-LA Instruments*) by including more years (2011-2020) increases the quality. Both classifiers extracted LA events exceptionally well, with Classifier 2 scoring higher (see F-score in Table 1). The reason for that is there are only eight learning events to predict from and a larger number of articles used for training. Therefore, the probability of predicting the right LD events is higher. The same applies to *LD-LA activities* and LA Indicators, where Classifier 2 outperformed the first one. Although, LA metrics prediction results (F-

score) exhibit no such improvement between the first and second classifiers. We reason that the number of metrics increased dramatically in the last years and there is a lack of common terminologies for LA metrics. The log-likelihood threshold column presents the range of confidence values used for considering the results of both classifiers (Collins, 2013). The output grid of LExplore also presents the prediction score/value alongside each *LD-LA instrument* (see Figure 3), where the prediction value closer to zero is better. During our evaluation, we considered the proposed range to be useful for adequate results. The tool ignores the results whose prediction value (log-likelihood threshold) is below -45.

Regarding the *semi-automatic extraction*, table 2 summarizes the comparison of the survey results. It shows that the task was easier and less tedious with the help of LExplore. Results from these surveys also indicate a need for such a tool. This is also supported by the results of the SUS and TAM surveys.

The average time for task one (harvesting instruments manually) was recorded as 27 minutes. The average time for task two was about four and a half minutes including the validation of the results. Without ratifying the results, the authors recorded that one can extract the *LD-LA Instruments* in less than a minute.

To measure the LExplore usability score we use SUS (Brooke, 1986). Table 3 presents the mean values for each SUS item. To calculate the SUS score, the item values (such as items no. 2, 4, 6, 8 & 10)

Table 3: System usability score for LAXplore.

SUS items	Mean (SD)
1. I think that I would like to use LAXplore frequently.	6.2 (1.03)
2. I found LAXplore unnecessarily complex.	1.6* (0.7)
3. I thought LAXplore was easy to use.	5.6 (1.9)
4. I think that I would need the support of a technical person to be able to use the LAXplore.	2.2* (1.62)
5. I found the various functions in LAXplore were well integrated.	5.8 (1.55)
6. I thought there was too much inconsistency in LAXplore.	1.8* (0.92)
7. I would imagine that most people would learn to use LAXplore very quickly.	6.4 (0.7)
8. I found LAXplore very cumbersome to use.	1.3* (0.48)
9. I felt very confident using LAXplore.	5.7 (0.82)
10. I needed to learn a lot of things before I could get going with LAXplore.	1.7* (0.67)
SUS mean score (Percentage)	5.61 (80.14%)

*7 (strongly agree) and 1 (strongly disagree)

*These values need to be inverted for SUS mean score (e.g., 1.6 to 5.4)

*SD = standard deviation

are inverted for the calculation of a mean value (see Table 3 last row value). LAXplore receives a score of 80.14%. According to Bangor et al., (Bangor et al., 2008) SUS acceptability scale, 80% is an excellent adjective rating and falls into an acceptable range.

As a second measure, we used an adapted TAM (Davis, 1985) to determine the usefulness and ease of use (c.f. table 4). The overall TAM usefulness mean score is 6.05 and the TAM ease of use mean score is 6.27 out of 7, which we consider a good overall system rating.

6 DISCUSSION AND CONCLUSION

For our first research question (RQ1) concerning the automatic extraction of *LD-LA Instruments*, the results presented show that our tool is capable of extracting *LD-LA Instruments* up to an acceptable level of precision. The LAXplore classifiers performed better in harvesting LD events and LD-LA activi-

ties. As discussed in the result section, one reason for that is the limited number of LD events (N=8) and LD-LA activities (N=40). The classifiers performed also good with LA indicators (N=135), where many of these indicators share similar goals/names and also work in similar ways, e.g., ‘predictive analytics’ repeated 62 times, ‘self-regulated learning’ appeared 24 times, and metrics (N >1000) (Ahmad et al., 2022; Ahmad. et al., 2022). Increasing data, i.e. using more publications, when training the classifier improved the quality in almost every category apart from LA metrics, where the F-score nearly remains the same. We argue that this is due to different wording/terminology/synonyms for a similar metric/measurement in the articles. For example, ‘materials used or resources used’, ‘test score or quiz score’, ‘final grade or final score or GPA’, ‘assignment or homework’ etc. are just a few examples out of many. We found only one incident where a LA indicator is being introduced called ‘lecture videos thermal analytics’ by using multimodal data, where the authors investigate the usage of thermal imaging for understanding students’ cognitive load (Srivastava et al., 2020). In contrast, in the majority of LA articles, the authors used different LA metrics (measurements) for a similar indicator (Ahmad et al., 2022); for example, self-regulated learning has been approached differently in a study of (Saint et al., 2020) and (Kia et al., 2020).

To sum up, the precision of LAXplore classifiers is good with scores ranging from 77% to 90%. However, we do not consider it good enough to be used unsupervised. We envision that with new advances in NLP, reducing the noise in the data discussed above, and more training data, the precision will increase to allow for a fully automated process.

The aim of RQ2 is to explore to what extent the semi-automatic extraction support users in the process of extracting *LD-LA Instruments*. In our scenario, results show that LAXplore reduces the time of extraction of *LD-LA Instruments* on average by a factor of six and makes this process easier and less tedious. In terms of usability, LAXplore got an excellent evaluation based on the acceptability scale of (Bangor et al., 2008). Thus, we assign LAXplore high usability. Participants were able to quickly use and learn to operate the tool. The system usability is proven to be a valid method and a critical predictor of actual system use and user experience (Drew et al., 2018; Peres et al., 2013; Vlachogianni and Tselios, 2022). Extending this, we use TAM to evaluate new technology adoption. The results (see table 4) show remarkable ratings. LAXplore is considered useful for the extraction of *LD-LA Instruments*. Regarding TAM’s

Table 4: TAM usefulness and ease of use for LExplore.

TAM usefulness items	Mean (SD)	TAM ease of use items	Mean (SD)
In case I need to do the same task, using this Tool in my job/work would enable me to accomplish tasks more quickly.	6.4 (0.97)	In case I need to do the same task, learning to operate this Tool would be easy for me.	6.3 (0.67)
In case I need to do the same task, using this Tool would improve my job/work performance.	5.7 (1.34)	In case I need to do the same task, I would find it easy to get this Tool to do what I want it to do.	6.3 (0.82)
In case I need to do the same task, using this Tool in my job/work would increase my productivity.	6.3 (0.82)	In case I need to do the same task, my interaction with this Tool would be clear and understandable.	6.4 (0.7)
In case I need to do the same task, using this Tool would enhance my effectiveness on the job/work.	5.8 (1.55)	In case I need to do the same task, I would find this Tool to be flexible to interact with.	6.1 (0.99)
In case I need to do the same task, using this Tool would make it easier to do my job/work.	5.9 (0.99)	In case I need to do the same task, It would be easy for me to become skillful at using this Tool.	6.1 (0.74)
In case I need to do the same task, I would find this Tool useful in my job/work.	6.2 (0.92)	In case I need to do the same task, I would find this Tool easy to use.	6.4 (0.7)
TAM usefulness mean score	6.05	TAM ease of use mean score	6.27

*SD = standard deviation

perceived ease of use, the tool is easy and forthright to be handled independently. Despite some uncertain reports regarding its theoretical assumptions, TAM is still a popular, frequently cited, and used model for evaluating the system's usefulness and ease of use (Md Lazim et al., 2021; Al-Emran and Granić, 2021; Chuttur, 2009). It is commonly applied in the context of information technology (Al-Emran et al., 2018). Consequently, we state that LExplore can be used for the semi-automatic extraction of *LD-LA Instruments* out of new LA literature..

We conclude, backed by the result of our study, that LExplore supports the semi-automatic extraction of *LD-LA Instruments*. This is an indicator of how AI and especially NLP can help to keep track of the continuous advances in scientific fields such as LA field by highlighting existing research practices in a repository (i.e., OpenLAIR).

6.1 Limitations

This study has two main limitations. First, there could be some margin of human lapses or slips in the data harvesting for the comparison with classifier results. In a few cases, the results of the classifiers were considered true positives if the results were nearly similar for example, a metric such as 'dashboard access' was considered as 'view dashboard/page' or an indicator 'monitoring' as 'self-regulation' or vice versa. The classification was considered false positive when

a result is very far away from the actual instrument and does not make any sense. The latter is currently hard to quantify. Second, if a LA article does not provide enough background information or measurements (metrics) that are used to create or evaluate the presented LA indicator, the results of the classifier may not be reliable. To handle such articles, the tool should be able to provide appealing results based on its content. Therefore, different NLP models are required that also include a semantic representation.

6.2 Future Work

Currently, this NLP solution serves the purpose of extracting *LD-LA Instruments* out of LA articles. Nevertheless, the solution is far from ideal and requires expert validation to verify results. Therefore, we consider it important to provide more accurate results and reduce or remove expert validation. Lastly, the tool works best on LA articles that contain an implementation, method, and description of *LD-LA Instruments*. Therefore, using and setting up ASReview (Hindriks, 2020) should help identify such relevant LA articles that can be used/consumed by LExplore.

REFERENCES

Ahmad, A., Kiesler, N., Schiffner, D., Schneider, J., and Wollny, S. (2023). Caught in the lifelong learn-

- ing maze. helping people with learning analytics and chatbots to find personal career paths. *International journal of information and education technology*, 13(3):423–429.
- Ahmad, A., Schneider, J., Griffiths, D., Biedermann, D., Schiffner, D., Greller, W., and Drachsler, H. (2022). Connecting the dots—a literature review on learning analytics indicators from a learning design perspective. *Journal of Computer Assisted Learning*, n/a(n/a):1–39.
- Ahmad, A., Schneider, J., Weidlich, J., Di Mitri, D., Yau, J. Y., Schiffner, D., and Drachsler, H. (2022). What indicators can i serve you with? an evaluation of a research-driven learning analytics indicator repository. In *Proceedings of the 14th International Conference on Computer Supported Education - Volume 1: CSEDU*, pages 58–68. INSTICC, SciTePress.
- Al-Emran, M. and Granić, A. (2021). Is it still valid or outdated? a bibliometric analysis of the technology acceptance model and its applications from 2010 to 2020. In Al-Emran, M. and Shaalan, K., editors, *Recent Advances in Technology Acceptance Models and Theories*, pages 1–12. Springer International Publishing, Cham.
- Al-Emran, M., Mezhuyev, V., and Kamaludin, A. (2018). Technology acceptance model in m-learning context: A systematic review. *Computers & Education*, 125:389–412.
- Bakharia, A., Corrin, L., de Barba, P., Kennedy, G., Gašević, D., Mulder, R., Williams, D., Dawson, S., and Lockyer, L. (2016). A conceptual framework linking learning design with learning analytics. In *Proceedings of the Sixth International Conference on Learning Analytics and Knowledge, LAK '16*, page 329–338, New York, NY, USA. Association for Computing Machinery.
- Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human-Computer Interaction*, 24(6):574–594.
- Baviskar, D., Ahirrao, S., Potdar, V., and Kotecha, K. (2021). Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions. *IEEE Access*, 9:72894–72936.
- Brooke, J. (1986). System usability scale (sus): a quick-and-dirty method of system evaluation user information. *Reading, UK: Digital equipment co ltd*, 43:1–7.
- Calle-Alonso, F., Cuenca-Guevara, A., de la Mata Lara, D., Sánchez-Gómez, J. M., Vega-Rodríguez, M. A., and Sánchez, C. J. P. (2017). Neurok: A collaborative e-learning platform based on pedagogical principles from neuroscience. In *CSEDU (1)*, pages 550–555, Porto, Portugal. scitepress.
- Ceylan, C. (2022). *Application of Natural Language Processing to Unstructured Data: A Case Study of Climate Change*. PhD thesis, Massachusetts Institute of Technology.
- Chatti, M. A., Lukarov, V., Thüs, H., Muslim, A., Yousef, A. M. F., Wahid, U., Greven, C., Chakrabarti, A., and Schroeder, U. (2014). Learning analytics: Challenges and future research directions. *eled*, 10(1).
- Chuttur, M. (2009). Overview of the technology acceptance model: Origins, developments and future directions. *elibrary*, 6:24.
- Clavié, B. and Gal, K. (2019). Edubert: Pretrained deep language models for learning analytics. *arXiv*, n/a:4.
- Collins, M. (2013). The naive bayes model, maximum-likelihood estimation, and the em algorithm. *Lecture Notes*, n/a:21.
- Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. PhD thesis, Massachusetts Institute of Technology.
- Drachsler, H. and Goldhammer, F. (2020). Learning analytics and eassessment—towards computational psychometrics by combining psychometrics with learning analytics. In Burgos, D., editor, *Radical Solutions and Learning Analytics: Personalised Learning and Teaching Through Big Data*, pages 67–80. Springer Singapore, Singapore.
- Drew, M. R., Falcone, B., and Baccus, W. L. (2018). What does the system usability scale (sus) measure? In Marcus, A. and Wang, W., editors, *Design, User Experience, and Usability: Theory and Practice*, pages 356–366, Cham. Springer International Publishing.
- Ferguson, R. and Clow, D. (2017). Where is the evidence? a call to action for learning analytics. In *Proceedings of the Seventh International Learning Analytics and Knowledge Conference, LAK '17*, page 56–65, New York, NY, USA. Association for Computing Machinery.
- Fiorini, N., Canese, K., Starchenko, G., Kireev, E., Kim, W., Miller, V., Osipov, M., Kholodov, M., Ismagilov, R., Mohan, S., et al. (2018). Best match: new relevance search for pubmed. *PLoS biology*, 16(8):e2005343.
- Fonferko-Shadrach, B., Lacey, A. S., Roberts, A., Akbari, A., Thompson, S., Ford, D. V., Lyons, R. A., Rees, M. I., and Pickrell, W. O. (2019). Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the exact (extraction of epilepsy clinical text) system. *BMJ Open*, 9(4):7.
- Gültekin, G. and Bayat, O. (2022). A naïve bayes prediction model on location-based recommendation by integrating multi-dimensional contextual information. *Multimedia Tools and Applications*, 81(5):6957–6978.
- Günther, S. A. (2021). The impact of social norms on students' online learning behavior: Insights from two randomized controlled trials. In *LAK21: 11th International Learning Analytics and Knowledge Conference, LAK21*, page 12–21, New York, NY, USA. Association for Computing Machinery.
- Heldens, S., Sclocco, A., Dreuning, H., van Werkhoven, B., Hijma, P., Maassen, J., and van Nieuwpoort, R. V. (2022). litstudy: A python package for literature reviews. *SoftwareX*, 20:101207.
- Hindriks, S. (2020). A study on the user experience of the asreview software tool for experienced and unexperienced users. M.S. thesis, Utrecht University.

- Karademir, O., Ahmad, A., Schneider, J., Di Mitri, D., Jivet, I., and Drachler, H. (2022). Designing the learning analytics cockpit—a dashboard that enables interventions. In *Methodologies and Intelligent Systems for Technology Enhanced Learning, 11th International Conference 11*, pages 95–104. Springer International Publishing.
- Kia, F. S., Teasley, S. D., Hatala, M., Karabenick, S. A., and Kay, M. (2020). How patterns of students dashboard use are related to their achievement and self-regulatory engagement. In *Proceedings of the Tenth International Conference on Learning Analytics and Knowledge, LAK '20*, page 340–349, New York, NY, USA. Association for Computing Machinery.
- Kubsch, M., Czinczel, B., Lossjew, J., Wyrwich, T., Bednorz, D., Bernholt, S., Fiedler, D., Strauß, S., Cress, U., Drachler, H., Neumann, K., and Rummel, N. (2022). Toward learning progression analytics — developing learning environments for the automated analysis of learning using evidence centered design. *Frontiers in Education*, 7:605.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Lim, L.-A., Gasevic, D., Matcha, W., Ahmad Uzir, N., and Dawson, S. (2021). Impact of learning analytics feedback on self-regulated learning: Triangulating behavioural logs with students' recall. In *LAK21: 11th International Learning Analytics and Knowledge Conference, LAK21*, page 364–374, New York, NY, USA. Association for Computing Machinery.
- Macfadyen, L. P., Lockyer, L., and Rienties, B. (2020). Learning design and learning analytics: Snapshot 2020. *Journal of Learning Analytics*, 7(3):6–12.
- Martin, F., Ndoye, A., and Wilkins, P. (2016). Using learning analytics to enhance student learning in online courses based on quality matters standards. *Journal of Educational Technology Systems*, 45(2):165–187.
- Md Lazim, C. S. L., Ismail, N. D. B., and Tazilah, M. D. A. K. (2021). Application of technology acceptance model (tam) towards online learning during covid-19 pandemic: Accounting students perspective. *Int. J. Bus. Econ. Law*, 24(1):13–20.
- Peres, S. C., Pham, T., and Phillips, R. (2013). Validation of the system usability scale (sus) sus in the wild. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 57, pages 192–196, Los Angeles, USA. SAGE Publications Sage CA: Los Angeles, CA, SAGE journals.
- Saint, J., Gašević, D., Matcha, W., Uzir, N. A., and Pardo, A. (2020). Combining analytic methods to unlock sequential and temporal patterns of self-regulated learning. In *Proceedings of the Tenth International Conference on Learning Analytics and Knowledge, LAK '20*, page 402–411, New York, NY, USA. Association for Computing Machinery.
- Saqr, M., Jovanovic, J., Viberg, O., and Gašević, D. (2022). Is there order in the mess? a single paper meta-analysis approach to identification of predictors of success in learning analytics. *Studies in Higher Education*, 47(12):2370–2391.
- Srivastava, N., Nawaz, S., Lodge, J. M., Velloso, E., Erfani, S., and Bailey, J. (2020). Exploring the usage of thermal imaging for understanding video lecture designs and students' experiences. In *Proceedings of the Tenth International Conference on Learning Analytics and Knowledge, LAK '20*, page 250–259, New York, NY, USA. Association for Computing Machinery.
- Tawfik, N. S. and Spruit, M. R. (2018). The SNPcurator: literature mining of enriched SNP-disease associations. *Database*, 2018. bay020.
- Verpoorten, D., Poumay, M., and Leclercq, D. (2007). The eight learning events model: A pedagogic conceptual tool supporting diversification of learning methods. *Interactive Learning Environments*, 15(2):151–160.
- Vlachogianni, P. and Tselios, N. (2022). Perceived usability evaluation of educational technology using the system usability scale (sus): A systematic review. *Journal of Research on Technology in Education*, 54(3):392–409.
- Voytovich, L. and Greenberg, C. (2022). Natural language processing: Practical applications in medicine and investigation of contextual autocomplete. In Staartjes, V. E., Regli, L., and Serra, C., editors, *Machine Learning in Clinical Neuroscience*, pages 207–214. Springer International Publishing, Cham.
- Zhang, H. and Li, D. (2007). Naïve bayes text classifier. In *2007 IEEE international conference on granular computing (GRC 2007)*, pages 708–708, Fremont, CA, USA. IEEE.