Natural Language Processing of Student's Feedback to Instructors

A Systematic Review

Sunar, Ayse Saliha; Khalid, Md Saifuddin

Published in:
IEEE Transactions on Learning Technologies

Link to article, DOI:
10.1109/TLT.2023.3330531

Publication date:
2024

Document Version
Peer reviewed version

Natural Language Processing of Student's Feedback to Instructors: A Systematic Review

Ayse Saliha Sunar, Md Saifuddin Khalid

***Abstract*—Course developers, providers and instructors gather feedback from students to gain insights into student satisfaction, success and difficulties in the learning process. The traditional manual analysis is time-consuming and resource-intensive, resulting in decreased insights and pedagogical impact. To address the problems, researchers use natural language processing techniques that apply the fields of machine learning, statistics and artificial intelligence to the feedback datasets for various purposes. These purposes include predicting sentiment, opinion research, insights into students' views of the course, and so on. The aim of this study is to identify themes and categories in academic research reports that use natural language processing for student feedback. Previous review studies have focused exclusively on sentiment analysis and specific techniques such as machine learning and deep learning. Our study put forward a comprehensive synthesis of various aspects, from the data to the methods used, to the data translation and labelling efforts, and to the categorisation of prediction/analysis targets in the literature. The synthesis includes two tables that allow the reader to compare the studies themselves and present the identified themes and categorisations in one figure and text. The methods, tools and data of 28 peer-reviewed papers are synthesised in 20 categories under six themes: aim and categorisation, methods and models, and tools and data (Size and Context, Language, and Labelling). Our research findings presented in this paper can inform researchers in the field in structuring their research ideas and methods, and in identifying gaps and needs in the literature for further development.**

***Index Terms*—natural language processing, students' feedback, classroom intervention, sentiment prediction, category prediction**

I. Introduction

STUDENT feedback has always been a very insightful source for lecturers and course designers to understand what students need, what has been helpful and useful and what has not, and when and how to intervene in the learning process when student feedback makes it necessary. Qualitative feedback provides insights for improving curriculum/course content, staff quality, assessment, learning support, teaching methods, teaching and learning resources, course management and the learning environment [1]. Lecturers usually analyse feedback manually by identifying themes, labelling them as codes, categorising them, reporting them with and without using learning theory-based organising categories, and highlighting problems [2]. However, when the volume of comments increases, for example with large classes or open online settings, manual analysis becomes time-consuming and there is a risk that important comments are overlooked. Student feedback is inherently subjective, and quick analysis of anonymous responses can lead to misinterpretation. Therefore, automatic analysis of feedback is becoming increasingly important in education. The students' qualitative text feedback is analysed by applying concepts of opinion mining [3], sentiment analysis [4], [5], and language models [6].

Natural language processing (NLP) applications, which enable machines to understand spoken or written human language, are helping to develop models for analysing student feedback, predicting student satisfaction, performance, etc., and identifying important feedback that requires immediate action. For example, Google Forms offers a template as an exit ticket at the end of the class for a quick feedback. The service then creates a workbook sheet showing all the answers submitted by students but no visualisation or alert system.[1] For a meaningful analysis of the feedback, it would be helpful to have a quick visualisation and immediate decision making for the instructors to intervene in a timely manner. However, this task is particularly difficult when the responses collected are written texts. It is time-consuming and error-prone for the teaching team to go through each comment every week or after each semester. Considering large classrooms or online teaching with a large number of participants, it is important to automatically analyse the comments and summarise what students thought and who needs immediate help. Natural language processing provides convenient methods to assess the sentiment in students' comments, extract a summary and red flag those comments that need attention. However, the tools and methods vary depending on the audience, language, quality and type of dataset, so there is no single solution for all classrooms.

There are initiatives by researchers to understand the phenomena in related literature. Some reviews focus on the bibliometric analysis of the literature. For example, Ahadi et al. [7] provide a bibliometric analysis of research on text mining in education. The authors provide statistics on authors' affiliations, publication details, citations, topics used and so on. While this method provides a particular perspective on the trend in a particular field, the core of our research is to analyse the literature from the perspective of the educational context and the technique of the methods used. For example, Ulfa et al. [8] analysed 12 articles between 2014 and 2019 to investigate the analysis of online student feedback using sentiment analysis. The term *sentiment* is used in the study

A. S. Sunar is with Dept of Computer Engineering, Bitlis Eren University, Turkey and Dept of Computer Science of University of Warwick, UK.

M. S. Khalid is with Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark.

[1]https://sites.google.com/a/mail.brandman.edu/edsu-533-classroom-tutorial/create-an-exit-ticket-using-google-forms

as an umbrella term for polarity, emotion and opinion mining. In particular, the authors focus on what algorithms are used with what goal for implementing sentiment analysis for student feedback in online learning. Our study differs from [8], we focus on all types of classroom settings, i.e. online, face-to-face, hybrid and not only on purpose and method but also on aspects related to data and tools.

Another systematic review has been conducted by Kastrati et al. [4] to systematically classify the research and results of the application of natural language processing, deep learning, and machine learning solutions for sentiment analysis in the education domain. The authors identified 92 articles published between 2015 and 2020. The authors especially focus on identifying challenges and trends in the literature as a contribution. The authors present the publications by year, rank, and publisher first. Then, the model, method and evaluation criteria of the algorithms used. The authors, similarly to our research, focus on the data though, they only consider data sources for categorisation. Similarly, Dalipi et al. [9] address sentiment analysis, but focus only on applications in MOOCs and not necessarily on natural language applications. Although the authors cite examples of machine learning and natural language processing, as well as other statistical and unspecified methods, there is no existing review on natural language applications on student feedback data.

In the review on the trends and challenges in implementing NLP methods for educational feedback analysis, Shaik et al. [10] synthesised methods such as sentiment annotation, entity annotation, text summarisation, and topic modelling to address NLP-related challenges such as sarcasm, domain-specific language, ambiguity and aspect-based sentiment analysis in education. In a table, the article synthesises the identified methods and programming packages Python, Java and R to solve NLP-related problems but irrespective of the type of data. The authors reviewed not only work in education but also research that they believe can easily adapt the method to education. So, with the aim of more nuanced analysis and focused empirical papers on one activity, that is, qualitative feedback to teachers from students.

Our study aims to provide researchers and learning technologists with a systematic overview of aims, methods, tools, data features, and pedagogical considerations in natural language processing in the context of student feedback to instructors. From the perspective of pedagogical and didactic design, the aims, results and reflections in the contributions of the empirical studies have great potential for various applications of natural language processing in assessment tools in education and training contexts. We, therefore, propose themes and categorisation with the respective features of studies.

The research questions in this study are as follows:

1) What are the most common goals cited in the literature for applying natural language processing techniques to students' feedback?
2) How can the existing literature be analysed to find common patterns with different aspects of the research?
3) How can the literature in the generic categories be synthesised based on the factors identified in question 2?

II. Methodology

This systematic literature review is conducted according to Creswell's five-step procedure for literature search and analysis [11, pp. 1-81].

- Identify key terms for your literature search
- Search for literature on a topic by consulting different types of materials and databases, including those available in an academic library and on the Internet
- Critically evaluate and select the literature for your review
- Organise the literature you have chosen by abstracting or noting the literature and making a visual diagram of it
- Write a literature review that includes summaries of the literature for your research report

A. Identify key terms

The keywords selected after several test searches that yielded relevant papers from Google Scholar are "education", "NLP", "natural language processing", "student feedback" and "course evaluation".

B. Locate the literature

Initially, the Google Scholar, Web of Science, and ERIC databases were selected because they cover a wide range of articles related to educational research and the application of NLP.

Second, abstracts and full texts were searched using different combinations of keywords. Google Scholar returned thousands of articles, but only the relevant articles on the first 10 pages were selected for screening. Web of Science and ERIC provided 9 and 19 articles respectively. A total of 128 full-text peer-reviewed articles were located. After duplicates were removed, 119 articles remained.

C. Critically evaluate and select literature

In the first round of screening, articles dealing with the provision/generation of automated feedback for students or with peer feedback were excluded. After the first round of elimination, 57 articles remained. The second round of screening was done by quickly reading the full text to identify synonymous words and find relevance that was not previously considered. Then the articles about literature reviews, conceptual frameworks or very preliminary results that provided little or no knowledge about their objective, model, results, and dataset used for evaluation and comparison were also removed. Missing information on one or two aspects is included stating the missing information in our analysis. Finally, 28 articles were selected for qualitative and quantitative synthesis.

Figure 1 summarises the search and the evaluation of the suitability or exclusion of papers based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart [12].

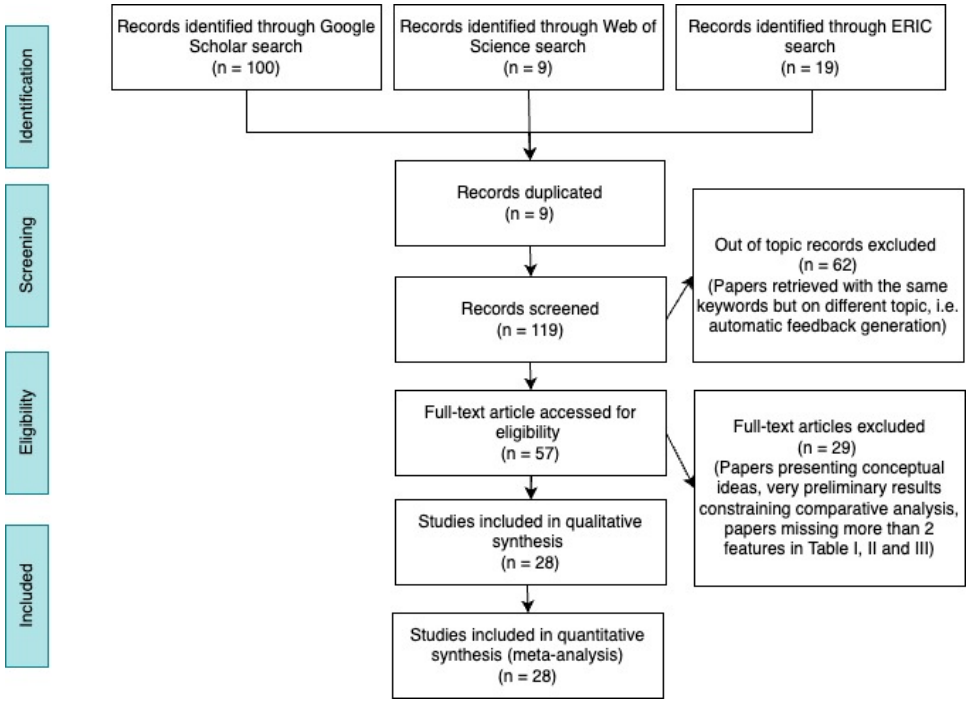

Fig. 1. Methodological procedure for the selection of the literature sample.

D. Organise the literature and visualise it

In Section III Analysis and Results, the 28 full-text articles are analysed by stating the aim, the procedure of data collection, the methods of analysis and the results, as suggested by Craswell [11]. The constant comparative method [13] is used for the qualitative analysis and synthesis of these articles. We gather the themes or variables about the features related to the objective, methodology and technology used in the studies as well as the data-related features including language, translation and collection process as these are the core features for the implementation of NLP.

The coding agreement was given attention to in two levels: 1) by the same author, the themes and concepts applied across the articles and constantly compared while applying both pruning and paring down processes, 2) between-author agreement, which was achieved through discussion. We identified synonyms and carried out reciprocal translation, examined possible grouping among the themes to sensitise ourselves to major patterns [14]. The themes are presented by Tables I, II and III. Then, the relationship between the features, trends in a single feature, and similarities and differences in the data are analysed. In the end, four themes are identified, which are: *According to Aim and Categorisation of Prediction, According to Method and Models, According to Tools,* and *According to Data Language, Size and Labels.* Then, categories under these themes are identified and explained in the following section with examples from the literature.

III. Qualitative Synthesis

This analysis and synthesis includes 28 peer-reviewed papers published between 2013 and 2022. Of these, ten (36%) are journal publications, while the other eighteen (64%) are conference papers. Figure 2 shows the distribution of papers over the years. More than half of the papers (15 out of 28, %54) are from the last three years.

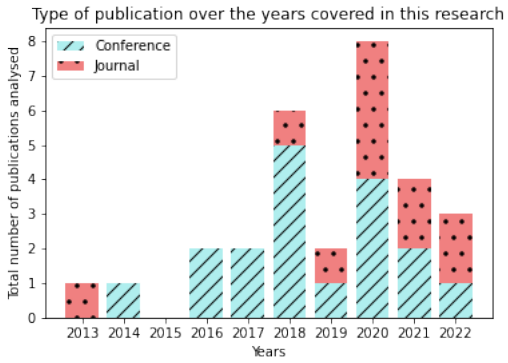

Fig. 2. Distribution of literature by year and type of publication.

The reviewed empirical studies on NLP techniques applied to students' feedback are synthesised into different categories and topics based on the characteristics such as objectives, methods and instruments used, type of data collection, and extent of data used. Table I and Table II provide an overall of each reviewed paper, summarising the objective, the methods or models used, the tools used, the categories applied for prediction and analyses, the target and original language if the data-set is translated, and the type of labelling (if any). Table III contains information on the characteristics of the dataset of each paper reviewed and summarises the resource, educational context (e.g. face-to-face or online), time of data collection, type of data, language and size. If the information

cannot be extracted from the paper, it is stated as NA (not available).

Based on the analysed characteristics of the studies as summarised in Table I, Table II and Table III, a qualitative synthesis is presented in the following four themes and twenty categories.

A. According to Aim and Categorisation

The aim of the studies varies. It is identified that some terms are used interchangeably. For example, in some studies, the terms *polarity*, *sentiment* or *opinion* were used for the same kind of sentiment prediction. In this paper, polarity and sentiment are used to represent the scale of sentiment in a comment from negative or -1 to +1; or opinion is used for sentences that contain subjective ideas, advice and reflections.

The studies are divided into six categories:

1) Sentiment prediction: The majority of the included literature (10 papers, 35.7%) applies sentiment analysis to classify comments as negative, neutral or positive; occasionally, some researchers only consider positive and negative classes and omit the neutral comments. The main goal of sentiment prediction is usually to identify ineffective use during the teaching practice to be improved. In some research, the weight of sentiment is revised with external factors. For example, Nikolovski et al. [22] calculates student objectivity score in self-evaluation which affects the sentiment score. In this study, the authors identify an interesting result that students give higher grades even if they have negative emotions in the free text. Also, some studies like Dhanalakshmi et al. [18] manually categorise the comments and then predict the sentiment, which are also categorised in this title as it does not make category predictions. Gutierrez et al. [21] identifies 99 features that students use for teachers related to the polarity of their sentiment, i.e. the word *support* is positive, the word *should* is negative.

2) Category and rating prediction complimented with sentiment: The second widely targeted objective, almost the same proportion as the sentiment prediction category (9 papers, 32.1%) is the use of sentiment prediction with the prediction of predefined categories. In the papers, the categories vary, although the focus is usually on pedagogical design and content, instructor, facility and assessment. For example, Nguyen et al. [25] and Nguyen et al. [26] used four themes such as *curriculum, lecturers, facilities,* and *others*. Ngoc et al. [30] and Edalati et al. [31] used similar categories such as *content, instructor, design, general,* and *structure*. As Sindhu et al. [32] focuses specifically on comments about teachers, the defined six categories also focus on pedagogy and teaching skills as *teaching pedagogy, behaviour, knowledge, assessment,* and *experience*. In contrast to the contributions in this category, the last example in the literature, Gottipati et al. [35] generates 16 topics such as: *faculty interaction, faculty engagement, faculty feedback and approachableness, faculty fairness and preparation, faculty presentation, course content, course skills, course value usefulness and challenges, course projects,* and *assignments*. When used in conjunction with sentiment prediction, category prediction can help teachers, course designers and administration more precisely diagnose the parts that need improvement.

3) Emotion prediction/analysis: While sentiment prediction scales the polarity of the comment, whether it is an opinion, advice, emotion or just a statement, emotion prediction focuses on the emotion itself in the comments. Emotion recognition helps stakeholders to describe and analyse the emotions of course participants, usually towards the institute, instructor, and course. For example, Marcu and Danubianu [15] aims to analyse the students' emotions towards the school. To categorise the emotion, they use two models, Plutchik and Ekman, as *anger, anticipation* (Plutchik only), *disgust, fear, joy, sadness, surprise* and *trust* (Plutchik only). In the other study, Sadriu et al. [19] use the Parrot model to identify seven emotions as: *love, joy, surprise, anger, sadness, fear* and *neutral*. Another interesting study is that Pham et al. [16] predicts sentiment along with predicting the extent, implying the use of feeling words to show the strength of the feeling.

4) Opinion mining: Opinion mining is used to extract sentences that reflect opinions, suggestions, advice, etc. It differs from sentiment analysis in that it does not scale the polarity of opinion, but detects opinions among positive or negative sentences. For example, Gottipati et al. [41] develop a rule-based machine learning technique with part-of-speech tagging (PoS tagging or PoS tagging or POST), called grammatical tagging, to extract the course improvement suggestions given by students in their feedback. Four categories are used for the PoS tagging: *positive statement, negative statement, suggestion* and *none*. For a better use of the results, Pyasi et al. [43] develop a dashboard generating an Excel sheet with visual reports of summaries that include sentiments and suggestions as an output to help the course instructors and developers. Similarly, the oldest study in the literature is Rashid et al.'s research [42] dated back to 2013 aims at extracting features and opinion words by extracting sequential and association pattern rules in the sentences. However, the rules extracted within the research by [42] could be insufficient for a different dataset.

5) Lexicon creation: Lexicon in machine learning is a set of vocabulary related to a specific domain or language. There are many free big-size lexicons in different languages, however, they are not necessarily sufficient in educational contexts and languages other than English. For example, Almosawi and Mahmood [28] aims to create an Arabic lexicon from students' feedback with the sentiment (positive and negative) scores. Other studies also use and edit already available lexicons for better results even though their aim is not to create a lexicon. For example, Nguyen et al. [26] proposes a framework with the designed facts and rules for representation and computations using already labelled two datasets in Vietnamese; one is for Vietnamese full names and one is for student course feedback corpus to predict sentiment and classify the review topic of texts in Vietnamese. While the proposed framework performs poorly than the literature, the results are promising for a local language. In some studies, the used lexicons are modified to accomplish the prediction aim even though the researchers do not aim at creating a new lexicon. For example, Gottipati et al. [35] revise TextBlob, which is trained with movie reviews

TABLE I
INSTRUMENTATION OF THE STUDIES IN THE LITERATURE (1/2).

Study	Aim	Method/Model	Tools	Categorisation / Prediction	Translation	Labelling labour
[15]	Emotion analysis	Plutchik, Ekman	Orange tool, Tweet profiler widget	8 emotions: anger, anticipation, disgust, fear, joy, sadness, surprise and trust	Translated to English	Machine
[16]	Sentiment and magnitude prediction	Google NLP	Google cloud-based Natural Language Processing API	3 sentiments: positive, negative, neutral; magnitude: from -1 to 1	Already in English	Machine
[17]	Create bidirectional feedback system considering both student and teacher feedback	Multiple Linear Regression	NCSS	Re-weighted feedback value from 0 to 1	NA	NA
[18]	Sentiment prediction in pre-defined categories	SVM, Naive Bayes, K Nearest Neighbor and Neural Network classifier	Rapid Miner	2 sentiments: Positive and Negative	Only English answers included	Human
[19]	Emotion and sentiment prediction	Parrot model for emotions; SentiWord, Emoticon, Improved Polarity classifiers	Python TextBlob library and MonkeyLearn API	7 emotions: love, joy, surprise, anger, sadness, fear and neutral; 3 sentiments: positive, negative, neutral	Albanian Google translation to English	Human
[20]	Sentiment prediction	Logistic, Multilayer perceptron, Simple logistic, SVM, Logistic model trees, Random Forest and Naive Bayes classifier	Python NLTK, WEKA	2 sentiments: positive and negative	Human revised Google translation to English	Human
[21]	Sentiment prediction	SVM	R	3 sentiments: positive, negative, neutral	NA	Human
[22]	Student objectivity prediction in self-evaluation	BERT	NA	2 sentiment: positive and negative	Translation to English	Machine
[23]	Sentiment prediction	An Ensemble model with Naive Bayes, Logistic Regression, Support Vector Machine, Decision Tree and Random Forest	NA	3 sentiments: positive, negative, neutral	Google translation to English	Human
[24]	Sentiment prediction	BiNB, BiSVM, LSTM, L-SVM, D-SVM, LD-SVM, Dependency Tree-LSTM	NA	3 sentiments: positive, negative, neutral	No translation	Already labelled
[25]	Sentiment and category prediction	Naive Bayes, Maximum entropy, LSTM, bi-LSTM	Datumbox, Java DeepLearning4j	4 Topics: curriculum, lecturers, facilities, others; 3 sentiments: positive, negative, neutral	No translation	Already labelled
[26]	Sentiment and category prediction	Facts and rules extraction with Naive Bayes	SWI-Prolog	4 Topics: curriculum, lecturers, facilities, others; 3 sentiments: positive, negative, neutral	No translation	Already labelled
[27]	Sentiment prediction	Naive Bayes, Maximum Entropy, SVM	Datumbox, libSVM	3 sentiments: positive, negative, neutral	No translation	Human
[28]	Create an Arabic language lexicon with sentiments	Naive Bayes, Support Vector Machine, k-Nearest Neighbors	Python Scikit	2 sentiments: positive, negative	No translation	Human
[29]	Predict student rating from textual comments	BERT	Python Pytorch	Grade from 1 to 5	Already in English	Human
[30]	Sentiment and category prediction	BERT	Python Scikit	5 categories: instructor, content, structure, design, general; 3 sentiments: positive, negative, neutral	Already in English	Machine
[31]	Sentiment and category prediction	AdaBoost, SVM, Random Forest, Decision Tree, Stochastic Gradient Descent, 1D-CNN, BERT	Python Auto-sklearn, AutoKeras	5 aspects: content, instructor, design, general, structure; 3 sentiments: positive, negative, neutral	Already in English	Human
[32]	Sentiment and category prediction	LSTM	OpenNLP, Python NLTK	6 Aspects: teaching pedagogy, behaviour, knowledge, assessment, experience; 3 sentiments: positive, negative, neutral	Only English answers included	Human

TABLE II
INSTRUMENTATION OF THE STUDIES IN LITERATURE (CONTINUED - 2/2).

Study	Aim	Method/Model	Tools	Categorisation / Prediction	Translation	Labelling labour
[33]	Correlation analysis between sentiment and feedback scores	VADER algorithm	Algorithmia.com	4 sentiments: positive, negative, neutral, compound score; 5 feedback: positive, structure, negative, irrelevant, other	Already in English	Human
[34]	Analyse the language used in description of poorly rated teachers	Latent Dirichlet Allocation	pyLDAviz, Python topic modelling library Gensim	None	Already in English	Human
[35]	Sentiment and category prediction	Latent Dirichlet Allocation	Python Text Blob, Polarity Analyser, Django, JavaScript D3 for visualisation	16 topics: faculty interaction, faculty engagement, faculty feedback and approachableness, faculty fairness and preparation, faculty presentation, course content, course skills, course value usefulness and challenges, course projects, and assignments; 2 sentiments: positive, negative	Already in English	Human
[36]	Sentiment prediction	Naive Bayes, Complement Naive Bayes, Maximum Entropy, SVM	NA	3 sentiments: positive, negative, neutral	Already in English	Human
[37]	Sentiment prediction	Random Forest, SVM	Python Scikit-learn, Text Analytics API 4 by Microsoft, Alchemy Language API 5, Aylien Text API	3 sentiments: positive, negative, neutral	Already in English	Human
[38]	Sentiment prediction	LSTM based Salp Swarm Algorithm, SVM, LR, NB	TextBlob, Twitter API, Amazon EC2, Google Visualization, Google Charts, Google Sites, Google spreadsheets, Google Closure, Google Analytics	3 sentiments: positive, negative, neutral	NA	Human
[39]	Sentiment prediction	Sentiment Analysis Lexicon for English (SALE)	Crawler 4j, JSoup Parsing	3 sentiments: positive, negative, neutral	Already in English	Human
[40]	Sentiment prediction	SVM, Naive Bayes, Complement NB, Maximum Entropy	NA	3 sentiments: positive, negative, neutral	Already in English	Human
[41]	Opinion mining	General Linear Model, SVM, Ctree, Decision Tree	Django, D3	4 categories: positive statement, negative statement, suggestion, none	Already in English	Human
[42]	Extract frequent features and opinion words	Rule extraction with Apriori, GSP	WEKA, GoTagger	None	NA	No labelling

to be used in their model due to its lack of contrasting conjunctions and suggestive words. Similarly, Nasim et al. [37] proposes a lexicon-based sentiment prediction method. The authors exploit the MPQA (Multi-Perspective Question Answering) lexicon by editing it according to the student-teacher context. For example, the words labelled as negative such as fine, lecture, and miss are corrected as positive by the authors. The authors highlight that the models outperformed when trained with the revised lexicon.

6) Statistical and mathematical analysis: The studies in this category proposes an analysis system using mathematical and statistical approaches to make meaningful insights into course evaluation or descriptive analysis. For example, Ekbote and Inamdar [17] use multiple linear regression to re-weight the feedback value (grade between 1 to 5) considering teachers' feedback on students and Cumulative Grade Point Average (CGPA) scores together and analyse the course evaluation from a bidirectional feedback system lenses. Lundqvist et al. [33] examines the correlation between sentiment (*positive, negative, neutral, compound score*) and feedback (*positive, structure, negative, irrelevant, other*). On the contrary, Valcarcel et al. [34] analyse the language used in the description of poorly rated teachers to find common patterns. The authors identify that the students use distinctly different language to address teaching-related complaints and behaviour perceived as unfit for teachers.

B. According to Method and Models

Researchers have applied state-of-the-art NLP models and methods to understand the written comments by students with the aims mentioned previously. According to the method and models embraced, we can categorise the studies into three and in our sample, they are almost equally distributed:

1) Baseline machine learning models: Baseline algorithms refer to machine learning algorithms which are simple models to establish minimum expected performance on a dataset [44]. These studies apply commonly used simple models to their datasets regardless of language or size. Ten of the total papers (36%) fall into this category. The most applied baseline models are Support Vector Machine (SVM), Naive Bayes (NB), k-Nearest Neighbour(kNN), Decision Tree (DT), Random Forest (RF), or their slightly different applications such as BiSVM, Ctree.

2) Deep learning models: Even though Deep Learning is not a new method, its use has recently become widespread due to emerging technologies in computer processing capabilities and a large amount of data available, which also leads to its applications in educational context [45]. We also observe in the literature analysis, nine of the papers (32%) exploit deep learning models to improve their already developed models or compare deep learning models with baseline machine learning models in predicting sentiment and/or category. The most used deep learning algorithms are long short-term memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT). However, using a deep learning model does not always guarantee higher performance. For example, Rybinski and Kopciuszewska [29] propose an NLP model to evaluate teaching by developing a BERT model processing the written text comments to predict the rating given by the reviewer. While the model is successful in predicting ratings, it hardly reaches 51% when predicting the review topic. Similarly, the BERT model developed by Ngoc et al. [30] performs better though, the performance between the BERT and non-deep learning models trained with term frequency-inverse document frequency (TF-IDF) is not big. Not all studies build their deep learning model but some studies, e.g. Sadriu et. al [19], use ready tools like Monkey Learn that implement a deep learning model in their algorithm.

3) Mathematical statistical approach: nine papers (32%) apply some mathematical and statistical models such as linear regression to calculate the weights for classification or use a statistical approach to create rules for detecting association amongst the text. For example, Nguyen et al. [26] create rules and facts by calculating percentages and probabilities of certain sentence structures in the Vietnamese language. Another widely used statistical model, Latent Dirichlet Allocation, which is especially used for topic modelling is used by Valcarcel et al. [34] and Gottipati et al. [35].

C. According to Tools

The researchers supporting the open science initiative[2] enable transparent and accessible knowledge produced through science. Our findings show that almost all studies in the selected literature use open-source tools that are mostly fully free of charge, unlike tools like MatLab which is very commonly used for data analysis. Therefore, we divided the research by tools into two: open-source and closed-source tools. The exact tools used by each study are listed in Table I and Table II. For this study, we could not extract the necessary information about which tools were used in six studies so they are extracted in this category ([22], [23], [24], [36], [40]).

1) Open-source tools: Twenty of the papers (87%) use open-source tools that allow fully accessible or provide open licences for certain features or academic use. It is seen in NLP studies that researchers tend to use open libraries mainly in Python and Java or Google-based platforms. Python and its famous libraries NLTK, scikit, pytorch and keras are the most commonly used tools in the literature analysed. Some researchers also prefer platforms such as Django, which provides a framework for using Python. The platforms WEKA and Datumbox, which are both written in Java and allow the user to access a range of machine learning algorithms, are also commonly used by researchers for NLP applications. Apart from these, there are a small number of different tools such as R or Orange. In addition, there are some cases where free licences are used for academic use, which is usually a paid tool, such as Rapid Miner.

2) Closed-source tools: Only three of the studies (13%) used paid tools, which sometimes allow free trial use with quota or for a limited time. The studies especially applying statistical regression models used closed-source statistical software such as NCSS in [17]. Some other tools like GoTagger used in Rashid et al. [42] provide free use for trial purposes which are classified as closed-source tools too.

D. According to Data Language, Size and Labels

How data is collected and pre-processed is the most initial and important part to inform the model that will be developed for understanding students' opinions and attitudes and applying interventions when designed.

The data may or may not have been collected for the purpose of machine analysis of the written comments, or it may have been retrieved from online websites. Consequently, the data used in the literature vary, from the size of the data to the timing of the collection to the type of questions and responses. Table III shows the characteristics of the data collected and the table is prepared for the purpose of more accurate generic categorisation.

The analysis in the literature shows that the data source and the educational setting, i.e. face-to-face, online or hybrid, as well as the timing of data collection, e.g. at the end or during the course, do not really influence the choice of method and model. However, the language of the original data and the data size are a guide to the choice of model, method and tool.

The data has been collected in different forms such as only text, only demographics and grades, only ratings and Likert, or a mix of these types. The collected data type is important because it is a proxy while identifying features and building the method. Since this review focuses on natural language processing applications, all the included studies collected at least text data. We are not going to use this feature for categorisation.

The majority of the studies use data collected during or end of face-to-face classes. Only a few of them use data collected through online resources such as websites to let users rate their teachers. On this kind of website, the instructors are rated by

[2]https://www.unesco.org/en/open-science

TABLE III
DATA CHARACTERISTICS OF THE STUDIES IN THE LITERATURE.

Study	Data resource	Educational Context	Time collected	Data type	Language	Data size
[15]	School survey	Face-to-face	End of semester	Text	Romanian	191 high school students
[16]	Online forum and QA sessions	Hybrid	During course	Text	English	3630 comments over 300 PG students on Google+
[17]	Course survey	NA	End of semester	Text, Rating (teacher and student feedback), CGPA, Test scores	NA	NA
[18]	Course survey	Face-to-face	End of semester	Text, Rating, Demographics	Arabic English	6433 responses from 6 courses
[19]	Course survey	Online	End of semester	Text	Albanian	624 paragraphs of opinions expressed by 114 undergrad student
[20]	Course survey	Face-to-face	End of semester	Text, Rating	Myanmar English	3000 undergraduate, master and PhD students from several courses
[21]	Teacher survey	Face-to-face	End of semester	Text	Spanish	1040 comments from undergraduate students
[22]	Course survey	Face-to-face	End of semester	Text, Rating	Macedonian	Over 400K grade and over 70K free text from undergrad students
[23]	Faculty survey	Face-to-face	End of semester	Text	English Filipino	1822 textual comments
[24]	Course survey	Face-to-face	End of semester	Text, Rating	Vietnamese	Over 16K reviews from HE students between 2013-2016
[25]	Course survey	Face-to-face	End of semester	Text, Rating	Vietnamese	Over 16K reviews from HE students between 2013-2016
[26]	Course survey	Face-to-face	End of semester	Text, Ratings	Vietnamese	Over 26K Vietnamese full names in UIT-ViNames dataset & Over 16K reviews from HE students in UIT-VSFC
[27]	Course survey	Face-to-face	End of semester	Text	Vietnamese	5K comments collected in two years from HE students
[28]	Course survey	Face-to-face	End of semester	Text, Demographics	Arabic	4812 feedback in both southern Iraqi dialect and the modern Arabic language) from 802 HE students
[29]	Online review of professors, course, and universities in the UK and US	Various	Various	Text, Rating	English	1,6 million reviews
[30]	Course reviews	Online	End/During online course	Text	English	Around 22K reviews gathered from 15 courses on Coursera
[31]	Online review forum	Online MOOCs	End/During online course	Text	English	Over 21K reviews from 15 different computer science courses on Coursera
[32]	Course survey	Face-to-face	End of semester	Text	Urdu English	2000 of 5000 comments specifically given for teachers
[33]	Online forum & Feedback	Online MOOC	During course	Text	English	More than 25000 online posts on FutureLearn
[34]	Online review of K12 teachers	Various	Various	Text, Ratings	English	Around 360K reviews about K-12 schools originating in the US
[35]	Course survey	NA	End of semester	Text, Ratings	English	Over 153K comments about 183 courses and 334 faculty during 4 years in Singapore
[36]	Real-time feedback & End of unit	Face-to-face	During course	Text	English	Over 1K posts from undergraduate and postgraduate students
[37]	Course survey	Face-to-face	End of semester	Text	English	Over 1,2K comments from HE students
[38]	Course survey & Tweets	Hybrid	End of semester	Text, Objective-type questions	NA	450 evaluation from 990 students in 7 HE courses; Over 16K tweets
[39]	Online forums and chats	NA	During course	Text	English	Over 11K replies in 1.6K from Blackboard forums and WhatsApp from 4 HE courses & NA size of course surveys (for result comparison)
[40]	Online forums	NA	During course	Text	English	Over 1K comments from HE students on Facebook
[41]	Course survey	Face-to-face	End of semester	Text, Ratings	English	Over 5K comments from 7 undergraduate courses collected from two terms in a year
[42]	Faculty survey	NA	End of semester	Text	English	NA data size from 5 HE courses

different students who may have attended different classes, online, face-to-face or hybrid. Therefore, it is not possible to identify the class setting in each study.

Categorisation in terms of data can be done in many ways. Within our approach, the studies altogether can be divided into 9 data-related categories under the data theme.

Based on the findings, the literature is categorised into four according to data size and source.

If the data is collected from face-to-face classes and only in a year, the data size is usually up to a few hundred. However, if the data is collected over the years or more than one course or retrieved from online rating websites, the size exceeds thousands.

Based on Size and Context

1) Big/Small size educational data collected through institutional settings: The majority of studies (20 papers, 71.4%) use institutional surveys as the data source. These surveys are usually conducted at the end of each semester or during the semester by the institution itself. The surveys are mostly collected through online survey tools and in some online courses, the data is collected through the institution's online platform, for instance, Pham et al. [16] collected data from an online forum and QA sessions during their hybrid postgraduate course. Some surveys assess courses, institutions and lecturers together, while others include feedback to lecturers only, such as Gutierrez et al. [21], Lalata et al. [23] and Rashid et al. [42].

2) Big size educational data collected through MOOCs/Online courses: Three studies (Ngoc et al. [30], Edalati et al. [31] and Lundqvist et al. [33]), which makes up the 10.7% of the literature, exploit the data set collected through open-ended questions in course evaluation surveys and online forums on the MOOC platforms Coursera and FutureLearn.

3) Big size educational data collected on social media: In some studies (14.3%), social media data was collected during the course. For example, Rybinski and Kopciuszewska [29] collected data from social media websites that allow users to rate their professors and universities. In the study, over 1.6 million reviews were collected from the websites ratemyprofessors.com and whatuni.com. Similarly, Valcarcel et al. [34] collected 360K comments on the website RateMyTeacher.com, which rates and evaluates teachers in K-12 schools in the US. In another educational context, Masood et al. [39] collected data from WhatsApp along with Blackboard forums during the course. The authors intend to use more comments from various social media tools such as Facebook, Instagram and Snapchat.

4) Non-educational data: For the application of NLP methods, some studies used non-educational data from social media or already available labelled datasets. For example, El-Demerdash et al. [38] used over 16K random tweets to train their sentiment analysis algorithms. Nguyen et al. [26] used the UIT -ViNames dataset with over 26K Vietnamese full names along with the course data.

The language of data naturally is not always English, however, available technologies do not always work well with other languages, especially with the languages used by small populations. Therefore, not all studies develop a multilingual model or a model compatible with a language other than English. Since the choice of language has a great impact on tool selection, we also categorised the data based on the language and translation of data. Based on the findings, the literature is categorised into three according to the target data language. Note that the information regarding the data could not be extracted from four of the studies ([17], [21], [38], [42]).

Based on Language

5) Data set already in English: More than half of the studies (58.4%) use the comments originally written in English. Amongst them, 12 studies use data already collected in English and 2 studies included comments that were written only in English in their multilingual dataset.

6) Machine or manual translation to English: Three of the studies translated the data into English using Google Translate, (Sadriu et al. [19], Lwin et al. [20] and Lalata et al. [23]) sometimes later revised by human whereas another two studies (Marcu and Danubianu [15] and Nikolovski et al. [22]) manually translated the data into English. In total, 20.8% of the studies translated the comments that are written in a language rather than English.

7) Dataset in languages other than English: Some researchers create a lexicon in a specific language rather than English or build a machine learning model compatible with many languages enabling them not to translate the data into English. For example, Nguyen et al. [24], Nguyen et al. [25], Nguyen et al. [26] and Giang et al. [27] used original dataset in Vietnamese; Almosawi and Mahmood [28] collected data in Southern Iraqi dialect and the modern Arabic language. In total, five of the studies (20.8%) applied their techniques to the data in a language other than English.

Having understood the general description of data, the most important step is to pre-process the data as the machine learning models learn from the pre-processed data to make meaningful conclusions about the teaching and learning. If a study uses a supervised learning method in their study, then the study needs a labelled dataset. In pre-processing data, data labelling is an important step to identify categories. The labelling labour could be carried out by language/domain professionals or the researchers themselves. For this categorisation, research does not need labelling and therefore it is not considered. So, six papers ([17], [24], [25], [26], [42]) are excluded from the categorisation as they either use already labelled data or provide no information regarding the labelling process. Based on the findings, the literature is categorised into two according to data labelling labour.

Based on Labelling Labour

8) Human labelling: The majority (19 papers, 82.6%) of labelling has been done manually and a few of them has used the already labelled data. For example, Giang et al. [27] build a guideline for human annotators which explains rules for labelling a Vietnamese sentence as positive, negative, or neutral. Two linguistic experts separately labelled the data and when they have a disagreement, the guideline has been revised for future labelling work that would be carried out by students who are not an expert. Another method for human labelling is that, for instance, Rybinski et al. [29] use the overall rating

the comments given by the comment author as a label for the text, i.e. quality rating scaled from 1 to 5 is the label for the review written by students about the class. Some researchers such as Lwin et al. [20] use machine labelling for the numeric data by clustering with k-means algorithm while using human labelling for textual data.

9) Machine labelling: The rest (4 papers, 17.4%) of the studies carried out the labelling with the aid of a machine, which implies that the algorithms used in the models automatically label the data. Machine learning algorithms such as BERT or Google NLP are used to predict the categories implying sentiment. For example, Marcu and Danubianu [15] use the already designed categories for emotions by the Plutchik and Ekman algorithms and let the algorithms label each comment. The evaluation of a model's performance is usually calculated by comparing how much the machine label matches with the pre-labels. In this case, however, the authors evaluated the success of the model by comparing both results generated by the algorithms. Ekman model classified the 191 comments of a Romanian school into five categories of emotions: anger, disgust, fear, joy, sadness and surprise. Plutchik model classified two emotions in addition to Ekman's: Trust and Anticipation. With 36.64% model accuracy, 30% of the data that was marked with these feelings, which were also probabilistically associated with joy by the Ekman model.

Nikolovski et al. [22] creates a layer in their model using the BERT algorithm to label the comments with their sentiment scores. They then fine-tune these labels by training with the groups categorised by the subject of the comments.

IV. Discussion and Conclusion

Our motivation for embarking on this review is to help researchers in the learning technology field, including ourselves, to design research in NLP applications for analyzing students' feedback to teachers by understanding the state-of-the-art techniques and their context and conditions for their application. After a careful selection process explained in Section II Methodology, a qualitative analysis and synthesis process was conducted in order to answer three research questions as explained below.

- What are the most common goals cited in the literature for applying natural language processing techniques to students' feedback?

The aims of the 28 studies examined are divided into six categories: 1) sentiment prediction, 2) category and rating prediction complimented with sentiment, 3) prediction and analysis of emotions, 4) opinion mining, 5) lexicon construction, and 6) statistical and mathematical analysis. The aim of most studies is either to analyse sentiment or to predict categories or ratings of comments. While the analysis and prediction of emotion follow these categories, it is also observed that researchers use sentiment and emotion interchangeably. In this study, when only positive, negative or neutral comments are examined, they are categorised as sentiment. For other feelings such as anger, hate, love, etc., the target is categorised as emotion. Given the nature of the natural language of university students, who are primarily adolescents, the comments may contain many modern, informal expressions and abbreviations that cannot be readily understood by the machine and reflect adolescents' perceptions of the education system, which may be perceived as unorthodox by the course designers preparing the questions. These types of challenges hinder algorithms in processing knowledge and may require the creation of lexicons that provide phrases and sentences in the appropriate language and with the appropriate educational content. In the literature analysed, there are a few studies that use different lexicons for one language or context, but only one aims to create a lexicon in a language other than English in an educational context. Lexicon creation and commentary analysis should build on existing theories of learning and pedagogy. For example, when applying self-determination theory, student feedback is categorised into 1) self-efficacy (Expectancy component): The student's belief in their ability to perform the task) 2) Intrinsic value (Value component: The student's goals and beliefs about the importance and interest of the task) 3) Emotion (Effective component): The student's emotional reactions to the task). In addition, the framework included technological barriers, teacher barriers, location and physical environment, and other factors were considered.

- How can the existing literature be analysed to find common patterns with different aspects of the research?

The existing literature reviews focus either on natural language processing techniques aimed at online learning or on models, tools and assessment criteria used to identify challenges on this topic. It became clear that there is no single way to categorise the papers to provide meaningful guidance on implementation, as the literature is very diverse. Therefore, we analysed each paper based on two aspects: 1) the instruments used in the literature by collecting information on the aim, method/model, tools, categorisation/prediction, translation, and labelling, and 2) the data characteristics used in the studies in the literature by collecting information on the data source, educational context, time of data collection, data type, language, and finally data size. This analysed information is presented in Tables I, I, and III. After each paper has been analysed against the identified factors, they are brought together to extract the common patterns that will be answered in the final research question.

- How can the literature in the generic categories be synthesised based on the factors identified in question 2?

As we specifically targeted researchers building their research setting, we identified six themes to be considered and addressed: (1) the aim and categorisation of the research, (2) the method and model to be used, (2) the tools to be used depending on the method, (3) the size and context of the data, (5) the language of the data and, where appropriate, (6) the labelling work of the data. All the papers collected in this study are analysed and categorised into groups for each theme. These groups identified under each theme and their associated papers are shown in Figure 3 with the white boxes. The categories are expected to provide a mapping of the literature as part of the protocol and decisions applied in the research on the

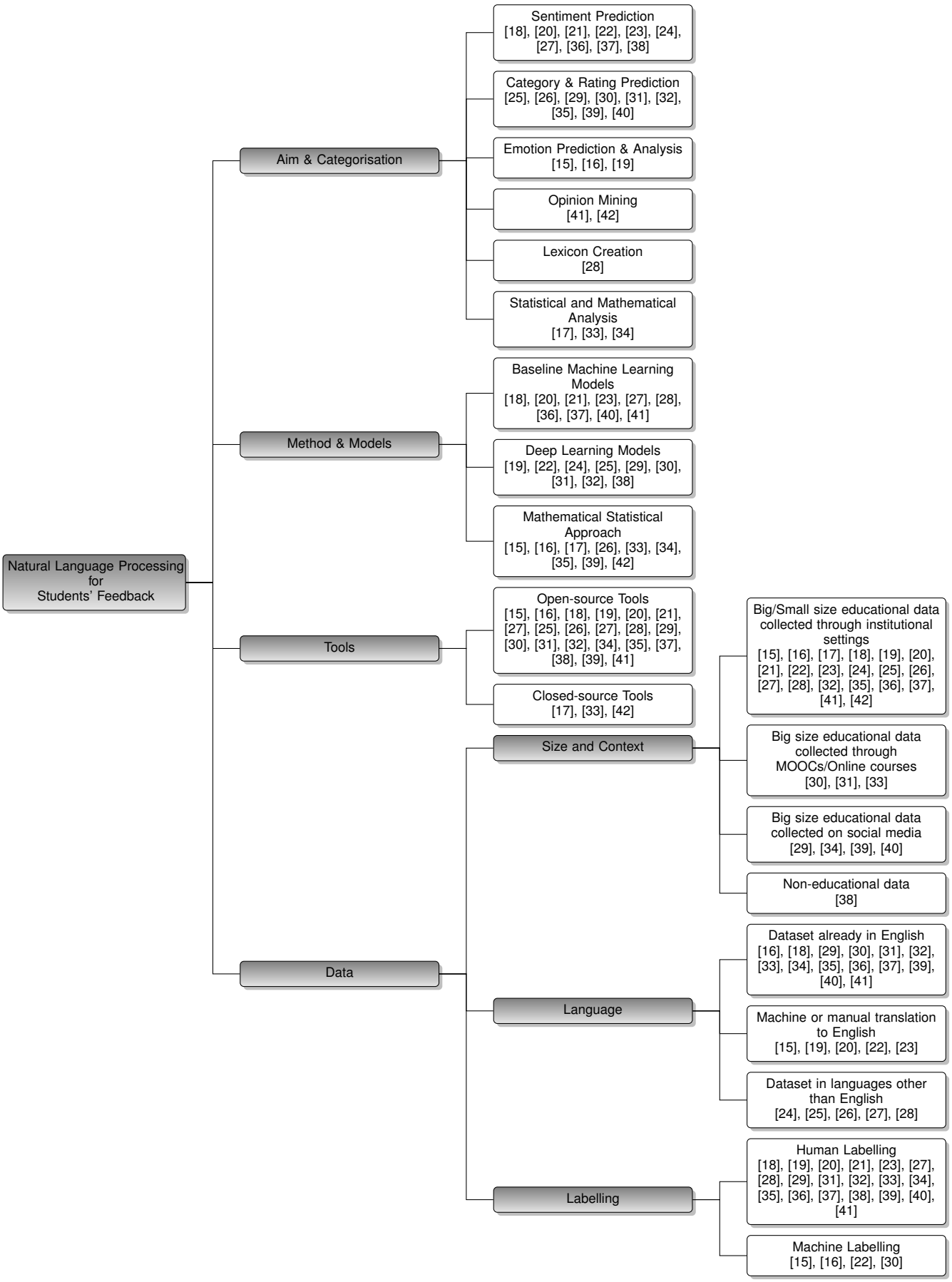

Fig. 3. Depict of generic categorisation of the literature.

application of NLP for analysing and providing insights on students' feedbacks to teachers.

Following the categories in Figure 3, except for studies with a few missing cells in the tables as they are not provided in the papers, the readers of this research can track all the categorisation for a specific paper. For example, a paper aiming for emotion prediction (let us take [15] as an example) applies a mathematical and statistical approach by using open-source tools with the data which is collected through institutional settings in a local language, therefore translated to English, and automatically labelled with the help of algorithms.

A. Scope of Future Work

This section provides recommendations and suggestions for NLP-based learner feedback analysis systems for education and training contexts. Figure 3 shows the mapping of the contribution of the existing literature and the scope of future work as follows: (1) Only one of the 28 articles included had the aim of 'lexicon building', which was to be done by applying concepts and perspectives grounded in learning theory, e.g. self-determination theory [46], students' perceived level of control [47], social interaction [48], interest in the material[49], clarity of goals [50], exposure to new concepts [51], intrinsic motivation [52], flow[53] ,trust[52], emotional and other technological, situational and teacher-related factors [2]. Moreover, lexicons of one language can only be translated into other languages if one has a distinct cultural and colloquial experience of communication. (2) Only one article deals with non-educational data, and the remaining studies deal with large or small educational data collected in educational institutions, MOOCs, online courses, and social media. Research on data from the work environment, physical training and other contexts of vocational education needs to be further developed. (3) Only five of the mapped articles analysed datasets in languages other than English and the majority (i.e. 19 articles) either collected or translated the datasets into English. Thus, more NLP studies should be conducted on non-English datasets. (4) Only four of the 23 studies reported using machine labelling of texts in English. Therefore, further research should be conducted in NLP studies to perform reliable machine labelling in both English and other languages. In addition, the theory-based concepts mentioned for lexicon creation are also applicable to labelling.

The scopes of future research under the above four points also apply to all kinds of qualitative feedback and assessment through NLP in education and training contexts. It is important to emphasise that the application of learning theories, psychology and subject-specific pedagogy (e.g. Science and Geography) should inform NLP study protocols.

In addition to this identified future work, consider replicable applications of the findings in the literature for future online or face-to-face courses. Developing tools for course developers and instructors to quickly collect feedback and immediately process comments would be a very useful application for practice. For example, Pyasi et al. [43] are developing a dashboard that uses the method suggested by Gottipati et al. [41] to create a visualisation for course instructors. To amplify the impact of these research studies, other researchers, projects and companies should consider a cross-platform, multilingual model for processing student feedback.

B. Limitations

The challenges, therefore the limitations, can be divided into two: access to full-text papers and lack of necessary information in papers.

The papers that are already open-accessed or their pre-print versions openly shared were easily accessed. The institutions of the authors provided access to some of the databases so that those papers were also easily accessed. A few of the papers that are not included in these groups cannot be accessed, therefore, are excluded from the research.

The second limitation is that some papers have not presented the information that is necessary for our categorisation approach. For example, some studies have not mentioned the language of the data or the number and level of the students that the data is gathered. If the information can be extracted from a second source, i.e. name of the databases or a screenshot of a database so that we can see the language, this research takes this secondary information into account. However, if the information cannot be extracted in any way, it is indicated as NA (not available) in the tables.

Also, even though the reviewed literature is the use of natural language processing in education, the focus of the literature is generally the technical success of the developed models. However, the impact of the developed models in the real life classroom implementations was not reported in the literature. Lastly, in order to successfully develop natural language processing applications in education, the pedagogical applications and theories and the advancement in natural language processing should inform each other and the results should be reported.

REFERENCES

[1] C. Steyn, C. Davies, and A. Sambo, "Eliciting student feedback for course development: the application of a qualitative course evaluation tool among business research students," *Assessment & Evaluation in Higher Education*, vol. 44, no. 1, pp. 11–24, 2019.

[2] M. S. Khalid, S. A. Chowdhury, and M. Parveen, "A theoretical framework to analyze students' formative feedback on classroom teaching," vol. 0, number: ICEdDE. [Online]. Available: http://www.dpi-proceedings.com/index.php/dtssehs/article/view/33699

[3] M. Misuraca, G. Scepi, and M. Spano, "Using opinion mining as an educational analytic: An integrated strategy for the analysis of students' feedback," *Studies in Educational Evaluation*, vol. 68, p. 100979, 2021.

[4] Z. Kastrati, F. Dalipi, A. S. Imran, K. Pireva Nuci, and M. A. Wani, "Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study," *Applied Sciences*, vol. 11, no. 9, p. 3986, 2021. [Online]. Available: https://doi.org/10.3390/app11093986

[5] K. Sangeetha and D. Prabha, "Sentiment analysis of student feedback using multi-head attention fusion model of word and context embedding for lstm," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 4117–4126, 2021.

[6] M. Masala, S. Ruseti, M. Dascalu, and C. Dobre, "Extracting and clustering main ideas from student feedback using language models," in *International Conference on Artificial Intelligence in Education*. Springer, 2021, pp. 282–292.

[7] A. Ahadi, A. Singh, M. Bower, and M. Garrett, "Text mining in education—a bibliometrics-based systematic review," *Education Sciences*, vol. 12, no. 3, p. 210, 2022.

[8] S. Ulfa, R. Bringula, C. Kurniawan, and M. Fadhli, "Student feedback on online learning by using sentiment analysis: A literature review," in *2020 6th international conference on education and technology (ICET)*. IEEE, 2020, pp. 53–58.

[9] F. Dalipi, K. Zdravkova, and F. Ahlgren, "Sentiment analysis of students' feedback in moocs: A systematic literature review," *Frontiers in artificial intelligence*, vol. 4, p. 728708, 2021.

[10] T. Shaik, X. Tao, Y. Li, C. Dann, J. McDonald, P. Redmond, and L. Galligan, "A review of the trends and challenges in adopting natural language processing methods for education feedback analysis," *IEEE Access*, vol. 10, pp. 56 720–56 739, 2022.

[11] J. W. Creswell, *Educational research: Planning, conducting, and evaluating quantitative*. Prentice Hall Upper Saddle River, NJ, 2002, vol. 7.

[12] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group*, "Preferred reporting items for systematic reviews and meta-analyses: the prisma statement," *Annals of internal medicine*, vol. 151, no. 4, pp. 264–269, 2009. [Online]. Available: https://doi.org/10.1136/bmj.b2535

[13] J. Hewitt-Taylor, "Constant comparative analysis: A method of analysing qualitative data," *Nursing Standard*, vol. 15, no. 42, pp. 39–42, 2001.

[14] J. Belur, L. Tompson, A. Thornton, and M. Simon, "Interrater reliability in systematic review methodology: exploring variation in coder decision-making," *Sociological methods & research*, vol. 50, no. 2, pp. 837–865, 2021.

[15] D. Marcu and M. Danubianu, "Sentiment analysis from students' feedback: a romanian high school case study," in *2020 International Conference on Development and Application Systems (DAS)*. IEEE, 2020, pp. 204–209. [Online]. Available: https://doi.org/10.1109/DAS49615.2020.9108927

[16] T. D. Pham, D. Vo, F. Li, K. Baker, B. Han, L. Lindsay, M. Pashna, and R. Rowley, "Natural language processing for analysis of student online sentiment in a postgraduate program," *Pacific Journal of Technology Enhanced Learning*, vol. 2, no. 2, pp. 15–30, 2020. [Online]. Available: https://doi.org/10.24135/pjtel.v2i2.4

[17] O. Ekbote and V. Inamdar, "A realistic mathematical approach for academic feedback analysis system," in *Proceeding of International Conference on Computational Science and Applications: ICCSA 2019*. Springer, 2020, pp. 127–137.

[18] V. Dhanalakshmi, D. Bino, and A. M. Saravanan, "Opinion mining from student feedback data using supervised learning algorithms," in *2016 3rd MEC international conference on big data and smart city (ICBDSC)*. IEEE, 2016, pp. 1–5.

[19] S. Sadriu, K. P. Nuci, A. S. Imran, I. Uddin, and M. Sajjad, "An automated approach for analysing students feedback using sentiment analysis techniques," in *Pattern Recognition and Artificial Intelligence: 5th Mediterranean Conference, MedPRAI 2021, Istanbul, Turkey, December 17–18, 2021, Proceedings*. Springer, 2022, pp. 228–239. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-04112-9_17

[20] H. H. Lwin, S. Oo, K. Z. Ye, K. K. Lin, W. P. Aung, and P. P. Ko, "Feedback analysis in outcome base education using machine learning," in *2020 17th International conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON)*. IEEE, 2020, pp. 767–770.

[21] G. Gutiérrez, J. Ponce, A. Ochoa, and M. Álvarez, "Analyzing students reviews of teacher performance using support vector machines by a proposed model," in *Intelligent Computing Systems: Second International Symposium, ISICS 2018, Merida, Mexico, March 21-23, 2018, Proceedings 2*. Springer, 2018, pp. 113–122.

[22] V. Nikolovski, D. Kitanovski, D. Trajanov, and I. Chorbev, "Case study: Predicting students objectivity in self-evaluation responses using bert single-label and multi-label fine-tuned deep-learning models," in *ICT Innovations 2020. Machine Learning and Applications: 12th International Conference, ICT Innovations 2020, Skopje, North Macedonia, September 24–26, 2020, Proceedings 12*. Springer, 2020, pp. 98–110.

[23] J.-a. P. Lalata, B. Gerardo, and R. Medina, "A sentiment analysis model for faculty comment evaluation using ensemble machine learning algorithms," in *Proceedings of the 2019 International Conference on Big Data Engineering*, 2019, pp. 68–73.

[24] V. D. Nguyen, K. Van Nguyen, and N. L.-T. Nguyen, "Variants of long short-term memory for sentiment analysis on vietnamese students' feedback corpus," in *2018 10th International conference on knowledge and systems engineering (KSE)*. IEEE, 2018, pp. 306–311.

[25] P. X. Nguyen, T. T. Hong, K. Van Nguyen, and N. L.-T. Nguyen, "Deep learning versus traditional classifiers on vietnamese students' feedback corpus," in *2018 5th NAFOSTED Conference on information and computer science (NICS)*. IEEE, 2018, pp. 75–80.

[26] K. V. Nguyen, T. Van Huynh, and A. G.-T. Nguyen, "A novel perspective of text classification by prolog-based deductive databases," in *Advances and Trends in Artificial Intelligence. From Theory to Practice: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part II 34*. Springer, 2021, pp. 138–148.

[27] N. T. P. Giang, T. T. Dien, and T. T. M. Khoa, "Sentiment analysis for university students' feedback," in *Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 2*. Springer, 2020, pp. 55–66.

[28] M. M. Almosawi and S. A. Mahmood, "Lexicon-based approach for sentiment analysis to student feedback," *Webology (ISSN: 1735-188X)*, vol. 19, no. 1, 2022.

[29] K. Rybinski and E. Kopciuszewska, "Will artificial intelligence revolutionise the student evaluation of teaching? a big data study of 1.6 million student reviews," *Assessment & Evaluation in Higher Education*, vol. 46, no. 7, pp. 1127–1139, 2021.

[30] T. V. Ngoc, M. N. Thi, and H. N. Thi, "Sentiment analysis of students' reviews on online courses: A transfer learning method," in *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 2021, pp. 306–314.

[31] M. Edalati, A. S. Imran, Z. Kastrati, and S. M. Daudpota, "The potential of machine learning algorithms for sentiment classification of students' feedback on mooc," in *Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 3*. Springer, 2022, pp. 11–22.

[32] I. Sindhu, S. M. Daudpota, K. Badar, M. Bakhtyar, J. Baber, and M. Nurunnabi, "Aspect-based opinion mining on student's feedback for faculty teaching performance evaluation," *IEEE Access*, vol. 7, pp. 108 729–108 741, 2019.

[33] K. Lundqvist, T. Liyanagunawardena, and L. Starkey, "Evaluation of student feedback within a mooc using sentiment analysis and target groups," *International Review of Research in Open and Distributed Learning*, vol. 21, no. 3, pp. 140–156, 2020.

[34] C. Valcarcel, J. Holmes, D. C. Berliner, and M. Koerner, "The value of student feedback in open forums: A natural language analysis of descriptions of poorly rated teachers." *Education Policy Analysis Archives*, vol. 29, no. 79, p. n79, 2021.

[35] S. Gottipati, V. Shankararaman, and J. Lin, "Latent dirichlet allocation for textual student feedback analysis," in *Proceedings of the 26th International Conference on Computers in Education ICCE 2018: Manila, Philippines, November 28-30. Research Collection School Of Information Systems.*, 2018, pp. 220–227.

[36] N. Altrabsheh, M. Cocea, and S. Fallahkhair, "Learning sentiment from students' feedback for real-time interventions in classrooms," in *Adaptive and Intelligent Systems: Third International Conference, ICAIS 2014, Bournemouth, UK, September 8-10, 2014. Proceedings*. Springer, 2014, pp. 40–49.

[37] Z. Nasim, Q. Rajput, and S. Haider, "Sentiment analysis of student feedback using machine learning and lexicon based approaches," in *2017 international conference on research and innovation in information systems (ICRIIS)*. IEEE, 2017, pp. 1–6.

[38] A. A. El-Demerdash, S. E. Hussein, and J. F. Zaki, "Course evaluation based on deep learning and ssa hyperparameters optimization," *CMC-Computers Materials & Continua*, vol. 71, no. 1, pp. 941–959, 2022.

[39] K. Masood, M. A. Khan, U. Saeed, M. A. Al Ghamdi, M. Asif, and M. Arfan, "Semantic analysis to identify students' feedback," *The Computer Journal*, vol. 65, no. 4, pp. 918–925, 2022.

[40] M. A. Ullah, "Sentiment analysis of students feedback: A study towards optimal tools," in *2016 International Workshop on Computational Intelligence (IWCI)*. IEEE, 2016, pp. 175–180.

[41] S. Gottipati, V. Shankararaman, and J. R. Lin, "Text analytics approach to extract course improvement suggestions from students' feedback," *Research and Practice in Technology Enhanced Learning*, vol. 13, pp. 1–19, 2018.

[42] A. Rashid, S. Asif, N. A. Butt, and I. Ashraf, "Feature level opinion mining of educational student feedback data using sequential pattern mining and association rule mining," *International Journal of Computer Applications*, vol. 81, no. 10, 2013.

[43] S. Pyasi, S. Gottipati, and V. Shankararaman, "Sufat-an analytics tool for gaining insights from student feedback comments," in *2018 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2018, pp. 1–9.

[44] S. Z. Li and A. Jain, Eds., *Baseline Algorithm*. Boston, MA: Springer US, 2009, pp. 60–60. [Online]. Available: https://doi.org/10.1007/978-0-387-73003-5_538

[45] I. Duru, A. S. Sunar, S. White, and B. Diri, "Deep learning for discussion-based cross-domain performance prediction of mooc learners

grouped by language on futurelearn," *Arabian Journal for Science and Engineering*, vol. 46, pp. 3613–3629, 2021.
[46] E. L. Deci and R. M. Ryan, "Self-determination theory," *Handbook of theories of social psychology*, vol. 1, no. 20, pp. 416–436, 2012.
[47] B. J. Zimmerman, "Self-regulated learning and academic achievement: An overview," *Educational psychologist*, vol. 25, no. 1, pp. 3–17, 1990.
[48] S. Gibbons, V. Scrutinio, and S. Telhaj, "Teacher turnover: Does it matter for pupil achievement? cep discussion paper no. 1530." *Centre for Economic Performance*, 2018.
[49] K. R. Wentzel and D. B. Miele, *Handbook of Motivation at School*. Routledge, 2016.
[50] P. Gollwitzer and G. B. Moskowitz, "Goal effects on action and cognition," in *Social psychology: Handbook of basic principles*. Guilford Press, 1996, pp. 361–399.
[51] M. Sakaki, A. Yagi, and K. Murayama, "Curiosity in old age: A possible key to achieving adaptive aging," *Neuroscience & Biobehavioral Reviews*, vol. 88, pp. 106–116, 2018.
[52] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary educational psychology*, vol. 25, no. 1, pp. 54–67, 2000.
[53] P. H. Mirvis, "Flow: The psychology of optimal experience," 1991.

Ayse Saliha Sunar, PhD received her MSc from Nagoya University, Japan, in Information Systems and her PhD from the University of Southampton, UK, where she researched MOOCs. She is an Assistant Professor at Bitlis Eren University, Turkey, and held a postdoctoral research position at the Josef Stefan Institute, Slovenia from 2019-2021. She has worked as a researcher in EU Horizon projects and coordinated Erasmus+ projects focused on developing recommendation systems for open educational resources using learning analytics and machine learning technologies. Her research interests include learning analytics, MOOCs, machine learning, technology-enhanced learning, and natural language processing and its applications in education.

Md Saifuddin Khalid, PhD is an Associate Professor in Learning Technology and Digitalization at the Department of Applied Mathematics and Computer Science (DTU Compute) and the leader of the Centre for Digital Learning Technology (learnT) at the Technical University of Denmark. Khalid's goal is to play the role of a change agent to solve or circumvent the barriers of developing, integrating, and adopting digital systems for learning, education, and training by solving decision-making dilemmas at both individual and organisational levels. He teaches digital learning technology and entrepreneurship, user experience (UX), interaction design, service innovation, human-centred information systems design, and statistics.