IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 10, 2021, accepted September 25, 2021, date of publication September 29, 2021, date of current version October 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3116425

Palaute: An Online Text Mining Tool for Analyzing Written Student Course Feedback

NIKU GRÖNBERG[1], ANTTI KNUTAS[1], TIMO HYNNINEN[2], AND MAIJA HUJALA[3]
[1]Department of Software Engineering, LUT University, 53850 Lappeenranta, Finland
[2]Department of Information Technology, South-Eastern Finland University of Applied Sciences, 50100 Mikkeli, Finland
[3]School of Business and Management, LUT University, 53850 Lappeenranta, Finland

Corresponding author: Antti Knutas (antti.knutas@lut.fi)

This work was supported in part by the Finnish Ministry of Education under Grant "Smart learning environments and their content production."

ABSTRACT Collecting student feedback is commonplace in universities. These surveys usually include both open-ended questions and Likert-type scale questions but the answers to open questions tend not to be analysed further than simply reading them. Recent research has shown that text mining and machine learning methods can be utilized to extract useful topics from masses of open student feedback. However, to our knowledge, not many off-the-shelf applications exist for processing open-ended student feedback automatically. Additionally, the use of text mining tools may not be available to all educators, as they require in-depth knowledge of text-mining, data analysis, or programming tools. To address this gap the current study presents a tool (Palaute) for analyzing written student feedback using topic modeling and emotion analysis. The utility of this tool is demonstrated with two real-life use cases: First, we analyze student feedback data collected from courses in a software engineering degree programme, and then feedback from all courses organized in a university. In our experiments, the analysis of open-ended feedback revealed that on certain software engineering course modules the workload is perceived as heavy, and on some programming courses the automatic code grader could be improved. The university-wide analysis produced indicators of good teaching quality, such as interesting courses, but also some concrete improvement points like the time given to complete course assignments. Therefore, the use of the tool resulted in actionable improvement points, which could not have been identified using only numeric feedback metrics. Based on the demonstrated utility, this paper describes the design and implementation of our open-source tool.

INDEX TERMS Curriculum development, educational technology, student evaluation of teaching, text processing.

I. INTRODUCTION

In universities the most common way to evaluate the quality of teaching is to analyze feedback collected from the students [1]–[9]. However, student evaluations of teaching (SET) as a measure of teaching quality is limited at best. First, education research has shown that SET is not a reliable metric for teaching quality, as student ratings of teaching and student learning are not related [7], [10], [11]. Second, while feedback questionnaires usually comprise of both Likert scales and open-ended questions, written feedback is often left unused [12] mostly due to the manual analysis being laborious [13].

The associate editor coordinating the review of this manuscript and approving it for publication was Biju Issac.

The current study focuses on the added value provided by open-ended, written student feedback. The automatic analysis of open-ended feedback using text mining and machine learning tools is a recent trend in higher education research (see for example [13]–[30]). Extant literature has shown that analysing open-ended feedback can uncover insights that could not be distinguished using quantitative evaluations only.

Qualitative open-ended questions have the advantage over Likert-type questions by allowing the respondent more freedom in their answers, in addition to allowing answers that were not expected in the survey design [31]. This is especially useful in student feedback surveys, where the open-ended questions allow the respondent to point out individual pain points or positive aspects of the course. Closed-ended

questions give a direction, but they only provide as detailed information as is specifically asked in the question.

The open-ended questions require human interpretation, and especially coding of the answers is a laborious task [31]. This is not an issue with low student numbers, but interpreting the feedback becomes very costly and unreasonable as the course participant count rises to hundreds or even thousands. Similarly, drawing conclusions from student feedback on an institution or organizational unit level can be difficult for the same reason.

In this study, a tool was created (Palaute - plot, analyze, learn, and understand topic emotions [1]) to better address the demand for written student feedback analysis. The goal was to create a tool that would improve the workflow of addressing student feedback by summarizing and generating insights from the data. The additional benefit of using Palaute is that it allows much larger data sets than is easily feasible with manual coding. This means that multiple data sets from different years from the same course can be combined and analysed easily, as well as, programme-wide analyses can be conducted, or analyses of large MOOCs. Combining the written feedback from all of the courses of a study programme should give new, actionable insights about the health of the programme that are based on qualitative SET data.

This study contributes to the field of SET by creating a novel artefact that combines multiple SET analysis steps into a single tool. The process follows the design science research approach by providing an artefact and evaluating its usefulness. Thus, the following research questions were formulated:

RQ1 What can be learned from the written student feedback with the tool?

RQ2 How does the tool benefit the user?

The rest of this paper is organized as follows: Literature and relevant studies are presented in Section II. The research process and artefact requirements are specified in Section III. Implementation and used analysis method are detailed in Section IV, followed by two demonstrations in Section V and the evaluation results are discussed in Section VI. Lastly, the main takeaways from this study are summarized in Section VII.

II. RELATED WORK

Text mining has been used in education analysis as a part of the field of educational text mining [32]. The diverse approaches include online forum [33] and VLE analysis [34], modeling student teamwork [35], MOOC diagnostics [36], and extracting course improvement suggestions [18].

Student written feedback analysis, a specific branch of educational text mining, has seen attention in recent research as well. Multiple different techniques have been shown to work with the evaluation of teaching data, including sentiment analysis [1], [37]–[40], Latent Dirichlet Analysis (LDA) [22], [33], rule-based classification [18], and key phrase extraction [41]. Diverse tools have been created to automate the listed, including Sobek [42] for text mining, Leximancer [14] for visualization, and a tool for extracting improvement suggestions [18]. Furthermore, workflows to combine text mining with qualitative approaches have been proposed, for example, by Hujala *et al.* [13].

There exist fewer tools that are aimed at streamlining and automating the process for the tools. While some tools for analysis have been published, few to none exist to support an approach that combines LDA text mining, supports a following thematic analysis, and additional sentiment analysis to add emotional valence analysis to the discovered themes. The tool introduced in this paper aims to address the research gap in having tools the streamline multi-step processes, such as one proposed in Hujala *et al.* [13]. The new tool implements a process for combining thematic analysis with LDA for analyzing themes in large student evaluation of teaching datasets, building on a line of research introduced by Finch *et al.* [43] and adding depth compared to analyses based solely on LDA ([18], [34], [44]).

III. ARTEFACT DESIGN GOALS

The main goal of this paper is to design and evaluate an artefact that solves a task in a problem domain using scientific principles. In the process, we follow the design science research (DSR) process, as defined by Peffers [45]. Design science research is an approach that aims to solve an issue in specific a problem domain using an iterative design process that applies the latest knowledge from related fields of science [46]. During the process, an artefact is produced and evaluated, its validity established by the utility in solving the issue [47]. At the same time, the process of applying and testing underlying kernel theories will provide new evidence or knowledge to the state of the art scientific knowledge base [48], [49].

The main goal is to address research and a practical solution gap for automated support for evaluating large SET data sets in a manner that would be feasible for lecturers of large courses or directors of degree programmes, following an LDA and thematic analysis-based process originally established in [13]. Furthermore, we investigate what other kinds of analyses can be provided to add value to the analysis outcomes, such as sentiment analysis.

The artefact design is accomplished in two stages: Iterative design by researchers and SET analysts, and feedback from practitioners. The utility of the artefact will be evaluated based on the research questions laid out in section I. Scientific rigour, which separates design science from everyday design, is accomplished by grounding the findings in established scientific literature and methods.

The goal of the artefact is to extract meaningful information from large text corpora to the user. To accomplish this, the tool must first allow the user to input the data. Then, the data must be preprocessed, analyzed and visualized to the user, so that the insights can be highlighted from the data. The varying structure of the survey instrument used in

[1] "Palaute" is also Finnish for "feedback."

TABLE 1. Functional requirements.

ID	Description	Reasoning
FR1	User can input data to the tool	Users should be able to insert the data they want to analyze
FR2	The tool works with Finnish and English	The university offers courses in both Finnish and English, therefore the surveys and their results are also in Finnish or English
FR3	User can select the columns from the data for the analysis	The university course survey is not static, so the tool should work with multiple different structures of data
FR4	User is presented with default options for using the tool	The tool should be easy and fast to use, so the user should be able to run the analysis immediately after uploading the data
FR5	User can modify the options used in the analysis	Topic modeling is highly dependent on the options, so the user should be allowed to tweak them for more accurate analysis
FR6	The tool summarizes text data	The goal of the tool is to make understanding the answers to open questions easier. This can be accomplished by creating summaries of the text data
FR7	The tool creates insights from data	The tool should create insights from the data that are hard to find by reading the texts
FR8	The tool allows for data exploration	Users should be able to explore the data in multiple ways
FR9	The analysis results are visualized to the user	Results should be visualized in multiple ways as a way to communicate them to the user
FR10	The user can filter the results	Giving more tools to users allows them to perform the actions they want. In other words, the tool is flexible

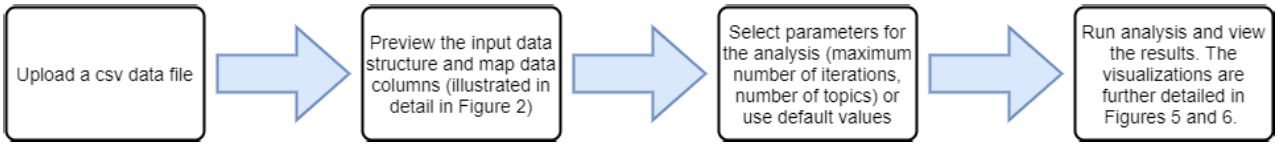


FIGURE 1. Illustration of the workflow using Palaute.

this university means that the tool must be able to handle different kinds of data, as having the user manually structure the data into a specific format would break the workflow. The functional requirements for the artefact are listed in Table 1.

IV. ARTEFACT IMPLEMENTATION

The artefact performs topic modeling, sentiment analysis and emotion analysis on data sets of varying kinds. This core functionality of the artefact is built on two R packages STM (structural topic model) and Syuzhet. Topic modeling is done using the STM package by [50]. Sentiment and emotion analysis is done using the Syuzhet package by [51]. The Syuzhet package contains multiple lexicons for sentiment analysis and NRC lexicon [52] for emotion analysis. Syuzhet also allows using custom lexicons.

The source code of Palaute is licensed as GNU general public license v3.0 (GPLv3) and can be found at Zenodo, an open research artefact repository.[2] A Docker file can also be downloaded to run the tool with minimal setup.[3]

The workflow of analyzing text feedback is illustrated in Figure 1. Using the tool does not require previous knowledge about text mining. The user is needed only to upload a data file (in csv format) and select which columns should be included in the analysis.

A. ANALYSIS AND MODELING METHODS

1) TOPIC MODELING

Latent Dirichlet Allocation (LDA) is a probabilistic model frequently used in text mining. LDA assumes that each document in the corpus is a random mixture of different topics, and distribution over words characterizes each topic [53], [54]. In other words, the corpus contains unknown topics that are spread out in multiple documents, and a group of words characterizes each topic. Words can also belong to multiple topics with varying probabilities.

The topic count is defined by the user beforehand, meaning LDA always generates as many topics as is specified. There have been solutions for finding the best amount of topics, like running the LDA multiple times with different topic counts and optimizing the perplexity of the model [54], [55].

Structural topic model (STM) improves upon LDA by including document-level metadata in the analysis. In addition to taking in the bag-of-words representation of the corpus, STM can also take in document-level covariates. This means that, for example, in surveys, quantitative data like gender or age of the respondent can be included as a covariate in the model. Lucas *et al.* [56] and Roberts *et al.* [50], [57] demonstrated that including covariate information does account for better results as the variance in topic prevalence is reduced.

Another improvement of STM over LDA is the explicit estimation of correlation between topics [56]. In other words, STM estimates how different topics relate to each other. This allows for visualization of the topic correlations, which can be helpful in getting a deeper understanding of the corpus-level structure of the topics.

2) SENTIMENT AND EMOTION ANALYSIS

Sentiment analysis is a text mining method used to understand the feelings or thoughts of the writer from the text [58].

[2]https://doi.org/10.5281/zenodo.3826074
[3]https://github.com/Nikug/Palaute

Early methods categorized documents or individual sentences into positive, negative or neutral. More recent aspect-based methods categorize sentiments based on a more fine-grained spectrum [59]. For example, the NRC emotion lexicon [52] distinguishes eight sentiment categories based on the eight basic emotions.

Sentiment analysis can be done on three levels: document, sentence, entity, or aspect [60]. Documents can contain multiple different sentiments. For example, in a course evaluation survey, a student might complain about difficult group work while praising the lecturer for explaining the subject well. In this case, it is hard to assign a positive or negative sentiment to the document. This problem continues in the sentence level as multiple differing sentiments can also be expressed in a single sentence, for example, ''The lectures were great but too long''. In this case, ''lectures were great'' is a positive sentiment, but ''lectures were too long'' is a negative sentiment, and both sentiments focus on the same target, ''lectures''. Therefore, it makes sense to analyze sentiments on the entity or aspect level; otherwise, all the expressed sentiments cannot be accurately identified [60].

In addition to sentiments, emotions, like sadness, anger and joy, can also be identified from text. Emotion analysis follows the same procedures as sentiment analysis, but emotion analysis has a different classification goal. Identifying sentiments and emotions from text are treated as separate problems, although sentiments can be identified from the emotions [61].

Tabak and Evrim [62] compared emotion lexicons and their effects on emotion analysis. These lexicons included the National research council Canada (NRC) word-sentiment association lexicon, EmoSenticNet (ESN), DepecheMood (DPM) and Topic based DepecheMood (TDPM). The lexicons contain different emotions and words based on those emotions, for example, NRC contains the eight emotions from Plutchik's wheel and two sentiments (positive, negative). In contrast, ESN contains six emotions (joys, sadness, disgust, anger, surprise, fear), and DPM and TDPM are built with eight emotions (happy, sad, angry, afraid, annoyed, inspired, amused, don't care). For comparison, matching emotions were selected from NRC and ESN, while DPM and TDPM were mapped to match the emotions of NRC and ESN. Overall, NRC and DPM performed the best in classifying emotions from news headlines.

After reviewing the literature, we used the emotion lexicon created at NRC by Mohammad *et al.* [52], and Mohammad and Turney [63] and translated it to over 20 languages, including Finnish. The lexicon contains classifications for positive or negative sentiments; and eight emotions (joy, trust, sadness, anger, surprise, fear, anticipation, disgust) commonly called Plutchik's wheel [64].

B. THEMATIC ANALYSIS

The last step in the process is lightweight thematic analysis, where the practitioner (educator, administrator) or a researcher overviews the analysis outcomes and assigns one or several themes to the analysis outcomes. Thematic analysis is a 'qualitative research method for identifying, analysing and reporting patterns (themes) within the data [65, p.79] and has been used widely, including in student feedback analysis [66]. It is essentially an iterative, qualitative method for reviewing data that aims toward increased abstraction.

In the Palaute system, the primary data sources for lightweight qualitative analysis are 1) keywords in each topic are presented, and as a novel feature, the system also presents 2) most characteristic answers for each topic. This approach presents the best of both worlds: Full responses are more rich in meaning than keywords [67], the analysis is based on topic probabilities as recommended by Finch *et al.* [43], and the algorithm-based sampling and reading allow for efficient analysis [13].

A lightweight, practitioner-oriented and partially automated thematic analysis process, as shortened from [13], proceeds as follows.

1) Reading ten to twenty most characteristic responses from each topic and topic keywords, as generated by the LDA topic-modelling process
2) Generating initial codes for each row, using either a data grounded or a theory-driven approach
3) Defining and naming themes

C. UTILIZING R LIBRARIES FOR ANALYSIS

1) TOPIC MODELING

The STM package contains a function for calculating the topic model. The topic model can be calculated using only the documents, but there can also be metadata in the form of covariates. The first type of covariates is prevalence covariates [50]. Prevalence covariates are external data that can be used in the calculation of topic prevalence. For example, in the context of course evaluation surveys, a Likert-type question about the workload of the course can be used as a prevalence covariate.

The second type of covariates is content covariates [50]. Content covariates affect the words used in a topic, and in the current implementation of STM, content covariates create strict groups of documents so that each document can only belong to a single group.

Topical content covariates change the STM model a lot since the documents are forced into groups [50]. In the context of course evaluation surveys, it could be used with some Likert-type questions that would significantly affect the vocabulary used in the topics. The survey questions could also be included as content covariates as it would make sense that different questions are answered differently.

The artefact has support for using both covariate types, although, as a limitation of the STM package, there can be only one content covariate, but multiple prevalence covariates are supported. Each of the data columns has the possibility to be either a document, prevalence covariate, content covariate or be excluded from the analysis. This means that different combinations of covariates and documents can be tested without having go to Excel or other tools to change the structure of

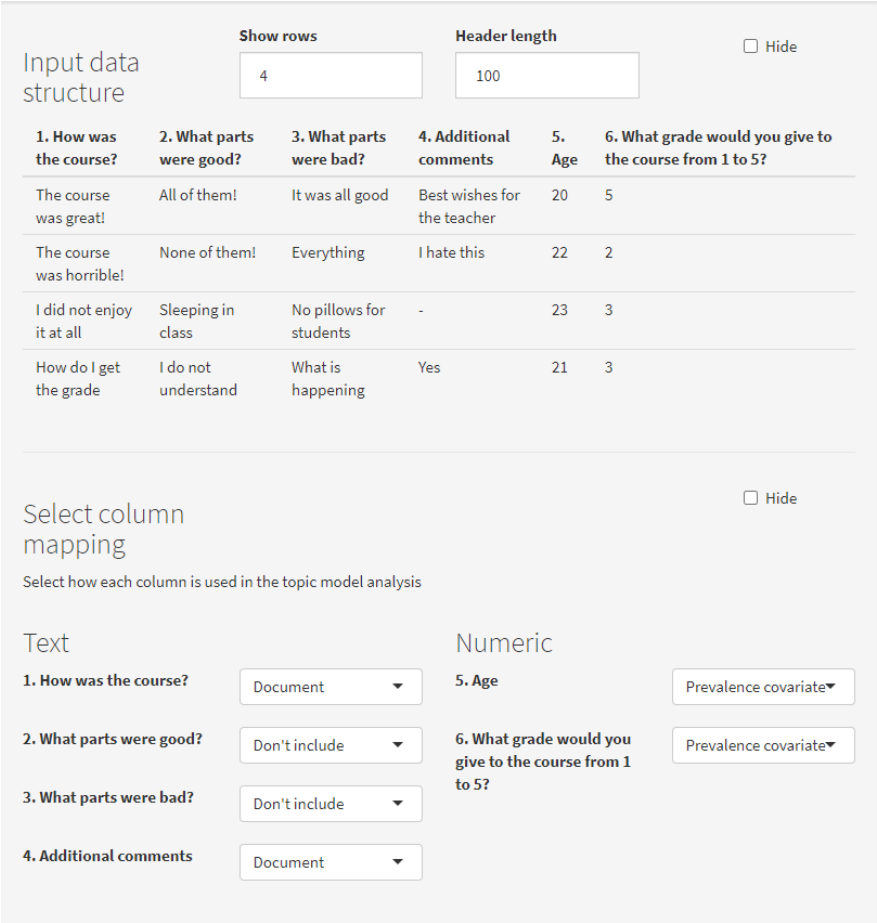


FIGURE 2. Example of column mapping.

data manually. This also allows the tool to work without any limitations on how the columns should be ordered, named or how many columns there should be. Figure 2 shows what the mapping in Palaute looks like with a short example data set with six questions.

STM package also contains tools for selecting the best model and the computationally best number of topics [50] using the semantic coherence algorithm [68]. Semantic coherence is related to the concept of pointwise mutual information, and it has been shown that the metric correlates well with a human judgment of topic quality [68]. The semantic coherence metric is commonly used as the standard evaluation option in popular analysis libraries, including stm [50] and topicmodels [69]. The artefact contains a function that trains multiple models for each number of topics and evaluates them based on semantic coherence and exclusivity. Based on this automatic evaluation, the system automatically proposes the number of topics with the highest quality values to the user.

2) SENTIMENT AND EMOTION ANALYSIS

Sentiment analysis and emotion analysis are performed using the NRC lexicon simply by matching the words in the data to the lexicon words and adding up the sentiment values for each matched word. This analysis does not consider the order of the words, the context of the words, negations, nor emphasis, but it should still yield a general sense of the data.

The sentiment analysis and emotion analysis are performed on the whole data set as a summary of the corpus. For individual topics, representative documents are selected, and the sentiment and emotion analysis are run with only the selected documents. There are multiple ways of making this selection of documents, but the current implementation is that the artefact selects the documents exclusively, meaning each document is added to the corpus of the topic that has the highest prevalence in that document. Dividing the documents exclusively among the topics makes sure that each document is used in the overall analysis only once, as multiple topics sharing the same documents would make the topics more similar to each other.

D. VISUALIZATION

The results of the text analysis can be visualized in multiple ways. LDA Topics are visualized using LDAvis by Sievert and Shirley [70]. LDAvis uses the Jensen-Shannon divergence to calculate the inter-topic distances from the

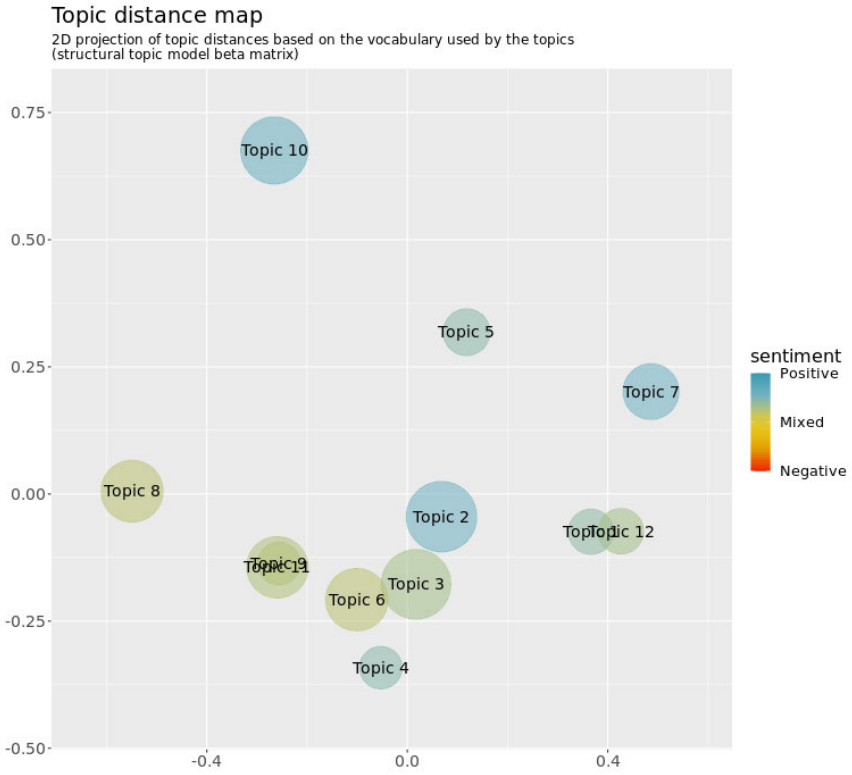


FIGURE 3. Example of an inter-topic distance plot.

word-topic probability matrix, which is then reduced to two dimensions to be shown as a two-dimensional plot. Each topic is displayed as a circle, with the area of the circle being proportional to the topic proportion.

Palaute adds to the LDAvis plot by expressing the sentiment of the topic as a color. The inter-topic distances are calculated from the STM model's beta matrix, which contains the log values of the word probabilities by the topic. As STM uses logarithmic values of the word probabilities exponent function must be applied to the values in the beta matrix before the inter-topic distances can be calculated.

The sizes of circles are proportional to the topic proportions, but this does not mean overlapping circles should be interpreted as sharing similar words proportional to the overlap. Instead, the distance between the topics is the measure of topic similarity, meaning they use similar vocabulary. Another important note is that since the plot is a two-dimensional representation of a higher dimensional construct, information is lost as the distances are projected two-dimensionally. Dimensional scaling is done using classical multidimensional scaling. The dimension scaling algorithm tries to keep the inter-topic distance similar when reducing dimensions, but there is information that is lost. So, just because two topics are close to each other, it does not necessarily mean they should be merged as one, although this should be the case. An example of this type of plot can be seen in Figure 3.

Theta matrix of the STM model contains the document topic proportions by topics. This matrix can be visualized to show what documents belong to which topics and how much of that document belongs to the other topics. In the artefact, this is done by creating a scatter plot of the documents, where the color of the document is based on the highest topic prevalence, as is the size of the circle. So, larger circles have a larger portion of them dedicated to a single topic. The Barnes-Hut variant of t-Distributed stochastic neighbor embedding (t-SNE) was used to dimensionally scale the data down to two dimensions [71].

Documents that have similar topic proportions cluster together in this plot. When documents are highly cohesive in the sense that they belong mainly to one topic, it causes clear clusters of documents to emerge in the plot to represent the topics. When the documents contain multiple topics more evenly, then the topics are not represented as single clusters. When the documents share similar topic proportions, they tend to share similar vocabulary, meaning semantically similar documents also cluster together. Topic labels are placed on the mathematical means of the document coordinates. The circles can be clicked, which shows that document, in addition to information about the document topic proportions. Figure 4 shows the example of topic-document relation of the data set with 12 topics.

The artefact contains a page with detailed information about each topic. An example of this can be seen in Figure 5.

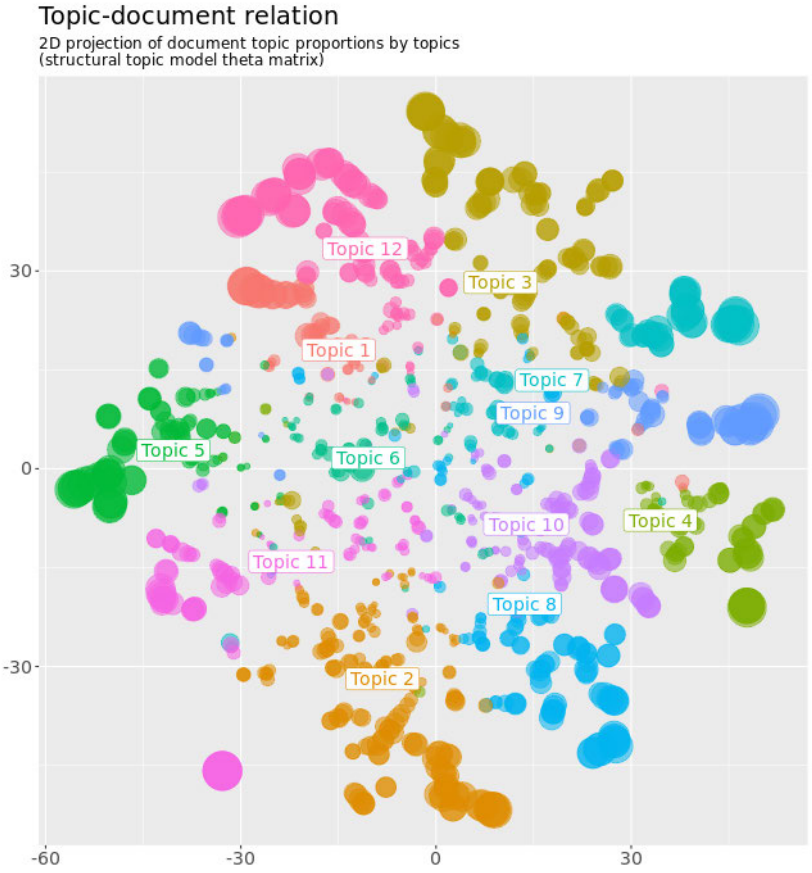


FIGURE 4. Example of a topic-document relation plot.

A similar panel to Figure 5 is generated for each topic and the details page contains all these panels. The user has the option to hide each of the smaller sections inside the panel using a filtering panel. There are also options for sorting the emotion analysis results in descending or alphabetical order, as it can be easier to do comparisons between topics when the results are in the same order. The sentiment is shown as a single bar. The number of shown keywords and documents can be changed by the user. Keywords are selected in the same way as in the inter-topic distance plot, and the documents are selected in the order of highest topic prevalence. This information should aid the user to understand what the topic is about by its vocabulary and example documents. The sentiment and emotions give additional insights about how, in this case, the survey respondents feel about the specific topic. For example, if the examination was too hard in the course, and it is a recurring theme in the survey answers, it should end up as a topic that is negative and has a vocabulary that uses emotionally negative words.

V. APPLYING AND EVALUATING THE ARTEFACT

The following subsections present two examples of how Palaute can be used to analyze written feedback data. In both examples, Palaute is demonstrated with student feedback data collected through two slightly different course feedback questionnaires: one for the academic year 2016–17 and the other for 2017–18.The feedback questionnaires contained both Likert-type scales and open-ended questions. The first questionnaire (2016–17) comprised seven Likert-scale questions, one open-ended question and seven background questions. The second questionnaire (2017–18) comprised five Likert-scale questions and five open-ended questions, one for each Likert-scale question, were included as well. The open-ended questions are presented in Table 2. The first example analyzes data collected from a whole degree programme, and in the second example the data is collected from all courses in the whole university.

A. ANALYZING FEEDBACK ON DEGREE PROGRAMME LEVEL

First, Palaute is demonstrated with student feedback data collected from courses in a software engineering degree programme. Only responses that contained answers to open-ended questions written in Finnish were included. The dataset is a total of 36 courses with 742 responses.

Responses to the open-ended questions were collapsed to a single column. For example, if the respondent answered to four open-ended questions, the answers were mapped to four

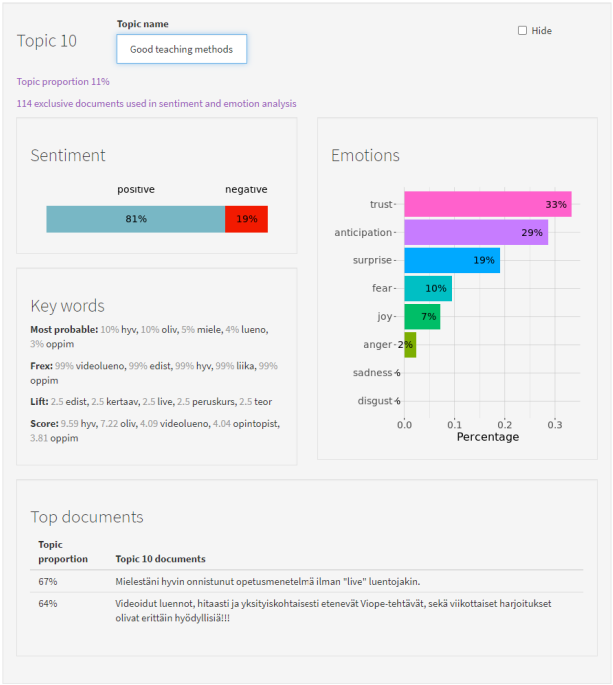


FIGURE 5. Example of a summary page.

TABLE 2. Open-ended questions used in demonstrations.

Sample	Open-ended questions
2016-2017	Other feedback about the course (for example, ways to enhance learning during the course)
2017-2018	What factors affected my level of motivation? What factors affected how much I invested in my learning? What factors affected the workload? My feedback regarding the teaching methods? What factors promoted my learning and how could learning be supported better?

rows, each with their matching Likert-type answers. Only full rows were included, meaning a row is dropped if the document is empty or one or more of the covariates are empty.

The model was run with 11 to 15 topics with 500 maximum iterations. This yielded a model that converged at 414 iterations with 12 topics. The topics, labelled using a machine-supported thematic analysis process introduced in [13], are listed in Table 3.

According to sentiment analysis, the feedback tended to be positive. However, lightweight thematic analysis of the most characteristic responses from each topic highlighted suggestions for improvement instead of praise.

Figure 6 shows that trust and anticipation are the most matching emotions.

Going over the topics, topics 1, 3, 5 and 7 include feedback on, for example, quizzes, assignments, exercises and exams. A total of 28 % of the feedback falls into these topics.

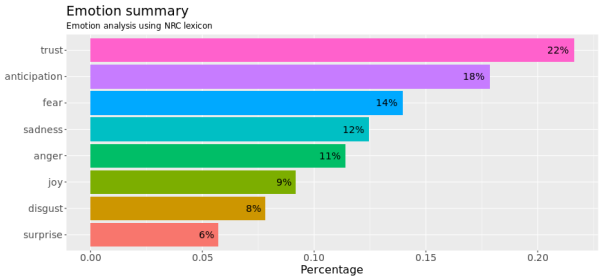


FIGURE 6. Emotion analysis summary from the degree programme wide analysis.

Topic 4 relates to teaching methods and it is the only completely positive topic. This topic is also relatively large at 13%. The courses and their topics are also deemed interesting, which is shown in topic 2 (10% of the feedback) and its documents.

Topic 6 contains suggestions from students. For example, some instructions could be made clearer and some additional topics could be taught in the lectures. The suggestions are mostly not critical of the current methods, and only suggest ways of further improving the courses.

Topics 8, 9 and 10 relate to workload, timing and schedule issues. Topic 8, labelled as ''Low motivation due to heavy workload'', deals with the heavy workload, affecting the students' motivation negatively. Some other reasons for low motivation were also mentioned, like lack of interest in a mandatory course. Topic 9 highlights the hurry the students face with their studies, and the SE courses are just too much work. Topic 10 deals with various timing and schedule issues. For example, evaluation and feedback from some exercises were delayed, which was not liked, there was not enough time to do some exercises, and the workload was sometimes too much.

Finally, topic 11 is related to course material (5% of the feedback) and topic 12 (4%) contains other, miscellaneous comments. There is a variety of short positive comments that are not highly connected with each other.

B. ANALYZING FEEDBACK ON INSTITUTION-WIDE LEVEL

Next, we analyze feedback data collected from all courses in the university. The data set contains a total of 6087 student course evaluations.

We ran the analysis similarly to the previous experiment. This time the analysis yielded a model that converged at 26 iterations with 6 topics. The topics are listed in Table 4.

Overall, as the feedback comments have come from students taking courses in different fields, the summary is more generic in comparison to the degree programme wide feedback analysis. Based on the inter-topic distance and emotion analysis presented in Figure 7, most of the feedback is positive or mixed, and no topic is primarily negative. Topics 3 and 6 are very close to each other, while the other topics are more distinct from each other. Topic 1 is the most mixed in terms of sentiment, while the others are mostly positive.

TABLE 3. Labelled topics from the degree programme wide analysis.

Topic	Proportion	Sentiment	Most probable keywords	Label
1	7%	76% positive	course, tasks, weekly quizzes, could, bad	Feedback on course exercises and improvement suggestions
2	10%	69% positive	course, good, topic, very, interest, difficult, thing	Course topics are thought to be interesting
3	8%	64% positive	exam, assignment, course, week, interest, more	Feedback on exams, practical assignments, and grading items in general
4	13%	86% positive	good, task, topic, interest, weekly task, material, lecture	Good teaching methods and other praise
5	7%	71% positive	weekly task, exercises, course, assistant	Feedback on course exercises and assignments
6	6%	76% positive	course, interest, learning, exam, weekly tasks, tasks	Feedback on the content and improvement suggestions for the course tasks
7	8%	65% positive	task, quizzes, questions, example, solution	Some weekly tasks or quizzes are unclear
8	11%	88% positive	many, really, weekly tasks, assignment, help	Low motivation due to heavy workload
9	14%	66% positive	task, assignment, workload, much, lecture, credit point	Heavy workload and too fast pace
10	9%	56% positive	course, team, example, moodle, assignment, good	Interesting course topic but the scheduling or workload could be easier
11	5%	84% positive	course, teacher, thing, material, example	Comments about the course material
12	4%	73% positive	time, course, student, grade, assignment	Miscellaneous comments

TABLE 4. Labelled topics from the university wide analysis.

Topic	Proportion	Sentiment	Most probable keywords	Label
1	12%	50% positive	exam, learning, return, weekly tasks, question	Mandatory attendance on courses. Feedback about content and course arrangements
2	17%	71% positive	topic, lecture, student, week, group	Feedback on the course workload. Course arrangements were too broad
3	18%	71% positive	course, really, example, practical assignment, very	Student motivation on this course was high. Content was interesting
4	17%	74% positive	tasks, good, more, moodle, teacher	Improvement suggestions for course assignments
5	16%	67% positive	time, much, practice, suitable	There was too little time to complete course assignments
6	20%	72% positive	lecture, good, could, course	Feedback about lectures, their content, and lecture arrangements

Inspecting the feedback in the different topics, Topic 1 consists of comments and suggestions on course arrangements. In particular, mandatory attendance in classes emerged as a common subject. Overall, 12% of the feedback fell into this topic.

Topics 2 and 5 are close to Topic 1 in the inter-topic distance matrix. Topic 2 consisted of feedback on courses, their workload, and the scope of course arrangements. Topic 5, too, contained feedback on course assignments, and in particular, the workload and time given to complete assignments.

Topics 3 and 6 shared the most similarities in their vocabulary. Topic 6 contained feedback about lectures, their content, and course arrangement. Topic 3 contained feedback on similar issues, with the distinction that there were also comments on the students' motivation.

Finally, topic 4 was the most distinct in terms of vocabulary used in the feedback. Comments in this topic generally contained some constructive criticism, and improvement suggestions for the course.

VI. DISCUSSION

Based on our design requirements, we created an online tool for analyzing written course feedback, Palaute. It is an online service, meaning no external software needs to be installed. Hence, the tool makes the analysis workflow more accessible to educators who are not proficient with data mining tools.

Through creating a design and demonstrating it as an instantiated artefact in a naturalistic environment [72] (also known as 'in the wild' in HCI), we demonstrate the usefulness of the concept and therefore the validity of our design ideas. Two contributions to the field are as follows:

- **Feasibility for practitioners.** The field of student evaluation of teaching and text mining has discussed the feasibility of speeding up topic modeling [67], since automated machine learning always needs validation [73]. A process that combines thematic analysis and LDA was proposed in [13] and is now demonstrated in this paper. With Palaute, a practitioner can analyze the findings from, for example, an institution-wide feedback survey in less than a quarter of a working day.
- **Combined analysis process.** As one of the major contributions to the field, combining topic modeling with sentiment and emotion analysis in CSE is a novel combination of methods that have not been widely explored in the literature. Single approaches to LDA and sentiment analysis have been studied extensively (e.g. [21], [22], [25]). However, the practice of how to

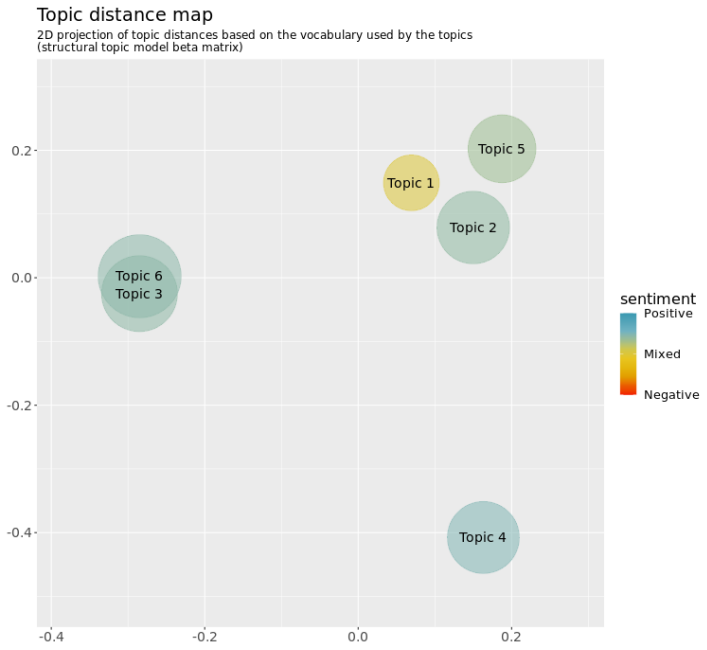


FIGURE 7. Inter-topic distance plot of written feedback collected from all courses in the university.

combine these analyses into a single practical workflow has been less explored.

VII. CONCLUSION

Topic modeling and emotion analysis can be used in the educational context as a way of creating summaries of the data. Palaute is a tool that was created to accomplish that task. While the analysis of feedback requires some understanding of text mining, the tool significantly streamlines the analysis process compared to generic text analytics tools.

The goal of this study was to create an artefact that can be used to analyze written course feedback. The evaluation of the tool was done in two experiments: Analyzing feedback within one degree programme in software engineering, and analyzing the feedback collected from all courses arranged in the university.

To answer the first research question: ''What can be learned from the written student feedback with the tool?'', we can learn the most popular points students make in the written feedback. On the degree programme level this was the heavy workload, issues and frustrations with the automatic code checker, large problems with the UI course, and that there was also a lot of praise for the SE courses. Furthermore, the university-wide analysis produced indicators of good teaching quality, such as interesting courses, but also some concrete improvement points like the time given to complete course assignments.

To answer the second research question: ''How does the tool benefit the user?'', the main benefit of Palaute is the user interface that it provides to the complex methods that are used under the hood. Performing topic modeling, emotion analysis, and visualizing the results is not trivial, so automating this process is useful. Topic modeling allows thousands of documents to be summarized quickly, which would be very time consuming if done manually. Palaute is useful in understanding the structure of the data, and the graphical user interface makes the whole process of analyzing the data much easier than having to write the code for the analysis. As the tool highlights what students think is wrong with the course, action can be taken to solve these issues, which should improve the course. This can have an impact on, for example, the dropout rates in courses. Topic modeling groups together similar comments, so the overall themes reflect what most students think is important. Thus the tool highlights points from large data sets that can be acted on to improve the teaching.

There remains future work and limitations in the field. One of the major remaining issues in establishing the workflow is integration: Currently, users of the system still need to download and format the data, analyze it in Palaute, and then download it back for further use. A more comprehensive suite, or alternatively a plugin system, would allow processing the feedback in the system where it was originally gathered. The second main limitation in the tool is the support for qualitative and thematic analysis, or mainly the support for the researcher's own notes or tags. The current expectation is that the results are exported and qualitative analysis is finalized on the desktop. However, as future research, it would be beneficial to investigate research that could enable collaborative notes or tags to the dataset through the tool. This was out of scope in the current study by intention, in order to focus on a specific research issue.

REFERENCES

[1] D. W. Jordan, "Re-thinking student written comments in course evaluations: Text mining unstructured data for program and institutional assessment," Ph.D. dissertation, Dept. Educ., California State Univ., Long Beach, CA, USA, May 2011. [Online]. Available: http://scholarworks.csustan.edu/handle/011235813/46

[2] M. Shevlin, P. Banyard, M. Davies, and M. Griffiths, "The validity of Student evaluation of teaching in higher education: Love me, love my lectures?" *Assessment Eval. Higher Educ.*, vol. 25, no. 4, pp. 397–405, Dec. 2000.

[3] F. Zabaleta, "The use and misuse of Student evaluations of teaching," *Teach. Higher Educ.*, vol. 12, no. 1, pp. 55–76, Feb. 2007. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/13562510601102131

[4] D. E. Clayson, "Student evaluations of teaching: Are they related to what students learn?: A meta-analysis and review of the literature," *J. Marketing Educ.*, vol. 31, no. 1, pp. 16–30, Apr. 2009.

[5] A. Hoel and T. I. Dahl, "Why bother? Student motivation to participate in Student evaluations of teaching," *Assessment Eval. Higher Educ.*, vol. 44, no. 3, pp. 361–378, Apr. 2019.

[6] D. Kember, D. Y. P. Leung, and K. P. Kwan, "Does the use of student feedback questionnaires improve the overall quality of teaching?" *Assessment Eval. Higher Educ.*, vol. 27, no. 5, pp. 411–425, Sep. 2002.

[7] P. Spooren, "On the credibility of the judge: A cross-classified multilevel analysis on students' evaluation of teaching," *Stud. Educ. Eval.*, vol. 36, no. 4, pp. 121–131, 2010.

[8] P. Spooren and F. Van Loon, "Who participates (not)? A non-response analysis on students' evaluations of teaching" *Proc.-Social Behav. Sci.*, vol. 69, pp. 990–996, Dec. 2012.

[9] S. L. Wallace, A. K. Lewis, and M. D. Allen, "The state of the literature on student evaluations of teaching and an exploratory analysis of written comments: Who benefits most," *College Teaching*, vol. 67, no. 1, pp. 1–14, 2019.

[10] B. Uttl, C. A. White, and D. W. Gonzalez, "Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related," *Stud. Educ. Eval.*, vol. 54, pp. 22–42, Sep. 2017.

[11] P. Spooren, B. Brockx, and D. Mortelmans, "On the validity of student evaluation of teaching: The state of the art," *Rev. Educ. Res.*, vol. 83, no. 4, pp. 598–642, Dec. 2013. [Online]. Available: http://journals.sagepub.com/doi/10.3102/0034654313496870

[12] F. N.-A. Alhija and B. Fresko, "Student evaluation of instruction: What can be learned from students' written comments?" *Stud. Educ. Eval.*, vol. 35, no. 1, pp. 37–44, Mar. 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0191491X09000066

[13] M. Hujala, A. Knutas, T. Hynninen, and H. Arminen, "Improving the quality of teaching by utilising written student feedback: A streamlined process," *Comput. Educ.*, vol. 157, Nov. 2020, Art. no. 103965.

[14] I. Stupans, T. McGuren, and A. M. Babey, "Student evaluation of teaching: A study exploring student rating instrument free-form text comments," *Innov. Higher Educ.*, vol. 41, no. 1, pp. 33–42, Feb. 2016. [Online]. Available: http://link.springer.com/10.1007/s10755-015-9328-5

[15] M. Shah and A. Pabel, "Making the student voice count: Using qualitative student feedback to enhance the student experience," *J. Appl. Res. Higher Educ.*, vol. 12, no. 2, pp. 194–209, Jul. 2019.

[16] L. Grebennikov and M. Shah, "Student voice: Using qualitative feedback from students to enhance their university experience," *Teach. Higher Educ.*, vol. 18, no. 6, pp. 606–618, Aug. 2013.

[17] S. Gottipati, V. Shankararaman, and S. Gan, "A conceptual framework for analyzing students' feedback," in *Proc. IEEE Frontiers Educ. Conf. (FIE)*, Oct. 2017, pp. 1–8.

[18] S. Gottipati, V. Shankararaman, and J. R. Lin, "Text analytics approach to extract course improvement suggestions from students' feedback," *Res. Pract. Technol. Enhanced Learn.*, vol. 13, no. 1, p. 6, Dec. 2018. [Online]. Available: https://telrp.springeropen.com/articles/10.1186/s41039-018-0073-0

[19] S. Cunningham-Nelson, M. Baktashmotlagh, and W. Boles, "Visually exploring sentiment and keywords for analysing student satisfaction data," in *Proc. 29th Australas. Assoc. Eng. Educ. Conf.*, 2018, p. 132.

[20] S. Cunningham-Nelson, M. Baktashmotlagh, and W. Boles, "Visualizing student opinion through text analysis," *IEEE Trans. Educ.*, vol. 62, no. 4, pp. 305–311, Nov. 2019.

[21] S. Pyasi, S. Gottipati, and V. Shankararaman, "SUFAT—An analytics tool for gaining insights from student feedback comments," in *Proc. IEEE Frontiers Educ. Conf. (FIE)*, Oct. 2018, pp. 1–9.

[22] S. Unankard and W. Nadee, "Topic detection for online course feedback using LDA," in *Proc. Int. Symp. Emerg. Technol. Educ.* Cham, Switzerland: Springer, 2019, pp. 133–142.

[23] Z. M. Ibrahim, M. Bader-El-Den, and M. Cocea, "Mining unit feedback to explore students' learning experiences," in *UK Workshop on Computational Intelligence*. Cham, Switzerland: Springer, 2018, pp. 339–350.

[24] S. Baddam, P. Bingi, and S. Shuva, "Student evaluation of teaching in business education: Discovering student sentiments using text mining techniques," *J. Bus. Educ. Scholarship Teach.*, vol. 13, no. 3, pp. 1–13, 2019.

[25] T. Hynninen, A. Knutas, M. Hujala, and H. Arminen, "Distinguishing the themes emerging from masses of open student feedback," in *Proc. 42nd Int. Conv. Inf. Commun. Technol., Electron. Microelectron.*, Opatija, Croatia, May 2019, pp. 557–561. [Online]. Available: https://ieeexplore.ieee.org/document/8756781/

[26] T. Hynninen, A. Knutas, and M. Hujala, "Sentiment analysis of open-ended student feedback," in *Proc. 43rd Int. Conv. Inf., Commun. Electron. Technol. (MIPRO)*, Oct. 2020, pp. 755–759.

[27] K. F. Hew, X. Hu, C. Qiao, and Y. Tang, "What predicts Student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach," *Comput. Educ.*, vol. 145, Feb. 2020, Art. no. 103724.

[28] A. Onan, "Mining opinions from instructor evaluation reviews: A deep learning approach," *Comput. Appl. Eng. Educ.*, vol. 28, no. 1, pp. 117–138, Jan. 2020.

[29] X. Peng and Q. Xu, "Investigating learners' behaviors and discourse content in MOOC course reviews," *Comput. Educ.*, vol. 143, Jan. 2020, Art. no. 103673.

[30] D. F. Sengkey, A. Jacobus, and F. J. Manoppo, "Implementing support vector machine sentiment analysis to Students' opinion toward lecturer in an Indonesian public university," *J. Sustain. Eng., Proc. Ser.*, vol. 1, no. 2, pp. 194–198, Sep. 2019.

[31] G. Vinten, "Open versus closed questions—An open issue," *Manage. Decis.*, vol. 33, no. 4, pp. 27–31, May 1995. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/00251749510084653/full/html

[32] R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text mining in education," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 9, no. 6, p. e1332, 2019.

[33] M. Morgan, A. Nylén, M. Butler, A. Eckerdal, N. Thota, and P. Kinnunen, "Examining manual and semi-automated methods of analysing MOOC data for computing education," in *Proc. 17th Koli Calling Int. Conf. Comput. Educ. Res.*, Nov. 2017, pp. 153–157.

[34] D. Buenano-Fernandez, W. Villegas-Ch, and S. Lujan-Mora, "Using text mining to evaluate student interaction in virtual learning environments," in *Proc. IEEE World Eng. Educ. Conf. (EDUNINE)*, Mar. 2018, pp. 1–6.

[35] R. Vivian, K. Falkner, N. Falkner, and H. Tarmazdi, "A method to analyze computer science students' teamwork in online collaborative learning environments," *ACM Trans. Comput. Educ.*, vol. 16, no. 2, pp. 1–28, Mar. 2016.

[36] Y. Nie, H. Luo, and D. Sun, "Design and validation of a diagnostic MOOC evaluation method combining AHP and text mining algorithms," *Interact. Learn. Environ.*, vol. 29, no. 2, pp. 315–328, 2020.

[37] S. Ahmad, A. Gupta, and N. K. Gupta, "Automated evaluation of students' feedbacks using text mining methods," *Int. J. Recent Technol. Eng.*, vol. 8, no. 4, pp. 337–342, Nov. 2019. [Online]. Available: https://www.ijrte.org/wp-content/uploads/papers/v8i4/D6846118419.pdf

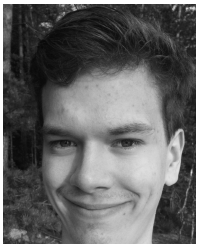
[38] F. de Paula Santos, C. P. Lechugo, and I. F. Silveira-Mackenzie, "'Speak well' or 'complain' about your teacher: A contribution of education data mining in the evaluation of teaching practices," in *Proc. Int. Symp. Comput. Educ. (SIIE)*, 2016, pp. 1–4.

[39] A. Koufakou, J. Gosselin, and D. Guo, "Using data mining to extract knowledge from student evaluation comments in undergraduate courses," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2016, pp. 3138–3142.

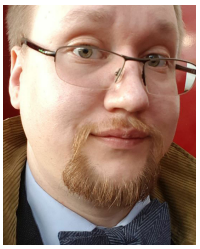
[40] C. Pong-Inwong and K. Kaewmak, "Improved sentiment analysis for teaching evaluation using feature selection and voting ensemble learning integration," in *Proc. 2nd IEEE Int. Conf. Comput. Commun. (ICCC)*, Oct. 2016, pp. 1222–1225.

[41] T. Sliusarenko, L. Harder Clemmensen, and B. Kjær Ersbøll, "Text mining in students' course evaluations: Relationships between open-ended comments and quantitative scores," in *Proc. 5th Int. Conf. Comput. Supported Educ.* Aachen, Germany: SciTePress, 2013, pp. 564–573. [Online]. Available: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0004384705640573

[42] E. Reategui, M. Klemann, D. Epstein, and A. Lorenzatti, "Sobek: A text mining tool for educational applications," in *Proc. Int. Conf. Data Science (ICDATA)*, 2011, p. 1.
[43] W. H. Finch, M. E. Hernández Finch, C. E. McIntosh, and C. Braun, "The use of topic modeling with latent Dirichlet analysis with open-ended survey items," *Transl. Psychol. Sci.*, vol. 4, no. 4, p. 403, 2018.
[44] D. Buena no-Fernandez, M. González, D. Gil, and S. Luján-Mora, "Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach," *IEEE Access*, vol. 8, pp. 35318–35330, 2020.
[45] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *J. Manage. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007. [Online]. Available: http://www.tandfonline.com/doi/full/10.2753/MIS0742-1222240302
[46] A. R. Hevner, "A three cycle view of design science research," *Scandin. J. Inf. Syst.*, vol. 19, no. 2, p. 4, 2007.
[47] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quart.*, vol. 1, pp. 75–105, Mar. 2004.
[48] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS Quart.*, vol. 37, no. 2, pp. 337–355, Feb. 2013.
[49] A. Knutas, Z. Pourzolfaghar, and M. Helfert, "The role and impact of descriptive theories in creating knowledge in design science," in *Proc. Int. Conf. Comput.-Hum. Interact. Res. Appl.* Cham, Switzerland: Springer, 2017, pp. 90–108.
[50] M. E. Roberts, B. M. Stewart, and D. Tingley, "STM: An R package for structural topic models," *J. Stat. Softw.*, vol. 91, no. 2, pp. 1–40, 2019. [Online]. Available: http://www.jstatsoft.org/v91/i02/
[51] M. L. Jockers. (2015). *Syuzhet: Extract Sentiment Plot Arcs From Text*. [Online]. Available: https://github.com/mjockers/syuzhet
[52] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," 2013, *arXiv:1308.6242*. [Online]. Available: http://arxiv.org/abs/1308.6242
[53] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[54] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Process. Mag.*, vol. 55, no. 4, pp. 77–84, Nov. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2133806.2133826
[55] X. Wang and D. H.-L. Goh, "Components of game experience: An automatic text analysis of online reviews," *Entertainment Comput.*, vol. 33, Mar. 2020, Art. no. 100338. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1875952119301211
[56] C. Lucas, R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley, "Computer-assisted text analysis for comparative politics," *Political Anal.*, vol. 23, no. 2, pp. 254–277, 2015. [Online]. Available: https://www.cambridge.org/core/product/identifier/S1047198700011736/type/journal%5Farticle
[57] M. E. Roberts, B. M. Stewart, and E. M. Airoldi, "A model of text for experimentation in the social sciences," *J. Amer. Stat. Assoc.*, vol. 111, no. 515, pp. 988–1003, 2016. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01621459.2016.1141684
[58] S. Tedmori and A. Awajan, "Sentiment analysis main tasks and applications: A survey," *JIPS*, vol. 15, pp. 500–519, Oct. 2019.
[59] J. Tao and X. Fang, "Toward multi-label sentiment analysis: A transfer learning based approach," *J. Big Data*, vol. 7, no. 1, p. 1, Dec. 2020. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0278-0
[60] M. Hu and B. Liu, *Mining Summarizing Customer Reviews*. USA: ACM, 2004.
[61] A. Kumar, A. Ekbal, D. Kawahra, and S. Kurohashi, "Emotion helps sentiment: A multi-task model for sentiment and emotion analysis," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.
[62] F. S. Tabak and V. Evrim, "Comparison of emotion lexicons," in *Proc. HONET-ICT*, Oct. 2016, pp. 154–158.
[63] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," 2013, *arXiv:1308.6297*. [Online]. Available: http://arxiv.org/abs/1308.6297
[64] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories Emotion*. Amsterdam, The Netherlands: Elsevier, 1980, pp. 3–33.
[65] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Res. Psychol.*, vol. 3, no. 2, pp. 77–101, 2006.
[66] A. Poulos and M. J. Mahony, "Effectiveness of feedback: The students' perspective," *Assessment Eval. Higher Educ.*, vol. 33, no. 2, pp. 143–154, 2008.
[67] G. Brookes and T. McEnery, "The utility of topic modelling for discourse studies: A critical evaluation," *Discourse Stud.*, vol. 21, no. 1, pp. 3–21, Feb. 2019.
[68] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. empirical methods natural Lang. Process.*, 2011, pp. 262–272.
[69] K. Hornik and B. Grün, "Topicmodels: An R package for fitting topic models," *J. Stat. Softw.*, vol. 40, no. 13, pp. 1–30, 2011.
[70] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proc. Workshop Interact. Lang. Learn. Vis. Interfaces*. Baltimore, MD, USA: Association for Computational Linguistics, 2014, pp. 63–70. [Online]. Available: http://aclweb.org/anthology/W14-3110
[71] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, Oct. 2014. [Online]. Available: http://jmlr.org/papers/v15/vandermaaten14a.html
[72] J. Pries-Heje, R. Baskerville, and J. R. Venable, "Strategies for design science research evaluation," in *Proc. ECIS*. USA: AIS, 2008, p. 87.
[73] L. Hagen, "Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?" *Inf. Process. Manage.*, vol. 54, no. 6, pp. 1292–1307, Nov. 2018.



NIKU GRÖNBERG received the master's degree in software engineering. He worked in the "Smart learning environments and their content production" Research Project at LUT University as a Research Assistant during the research project. He is currently working in the software industry.



ANTTI KNUTAS is currently working with the Department of Software Engineering, LUT University, as an Assistant Professor in software construction. He is also heading the bachelor's degree programme in software engineering in his institution. In addition to research at LUT, he has previously worked with the Science Foundation Ireland Research Centre for Software with Dublin City University and the PONG Labs, University of Milan. He has more than 50 papers in his field. His current research interests include human factors in software engineering, civic tech, and computer-supported collaboration.

TIMO HYNNINEN currently works as a Senior Lecturer of information technology with the South-Eastern Finland University of Applied Sciences. He is also the Head of the software engineering degree programme. His main research interests include software testing and quality assurance, computing education, and computing applications for education. Previously, he has also worked on video games research in academia and software development in the industry.

MAIJA HUJALA is currently working with LUT School of Business and Management as an Associate Professor in industry and data analysis. She has published 38 peer-reviewed scientific publications. Her current research interests include data science in education, especially in student evaluation of teaching. Her previous research interests include the acceptability of wind power and structural changes in the global pulp and paper industry. She is experienced in quantitative data analysis and interdisciplinary research.