EasyChair Preprint

№ 5669

Distinguishing the Themes Emerging from Masses of Open Student Feedback

Timo Hynninen, Antti Knutas, Maija Hujala and Heli Arminen

June 3, 2021

Distinguishing the Themes Emerging from Masses of Open Student Feedback

Timo Hynninen*, Antti Knutas**, Maija Hujala*** and Heli Arminen***
*South-Eastern Finland University of Applied Sciences / Department of Information Technology, Mikkeli, Finland
**LUT University / LUT School of Engineering Science, Lappeenranta, Finland
***LUT University / LUT School of Business and Management, Lappeenranta, Finland
timo.hynninen@xamk.fi, antti.knutas@lut.fi, maija.hujala@lut.fi, heli.arminen@lut.fi

***Abstract*—Student feedback is one of the key methods for assessing the quality of teaching in higher education. Feedback is often collected using both Likert-type scales and open-ended questions. However, open-ended text answers are a difficult resource to utilize because of the manual work involved in qualitative analysis, and it is a challenge to gain insight of the underlying themes or issues behind the feedback. This paper presents a study in which we create and analyze topic models from open-ended student feedback. First, 6087 individual student evaluations were collected from university courses between two academic years, from 2016 to 2018. Then, topic models from the feedback texts were created using the Latent Dirichlet Allocation method with the R programming language and environment for statistical computing. After analyzing the resulting topic models, six categories of feedback were distinguished: 1) Positive comments about arrangements, 2) dissatisfaction in the teaching, 3) comments about course arrangements and deadlines, 4) lack of student motivation, 5) interest in the topic and understanding the material, and 6) comments about interesting, rewarding but challenging courses. Finally, this paper discusses the topic modelling results to provide an insight into the automatic analysis of student feedback.**

I. Introduction

Student evaluations of teaching are an integral part of quality control in higher education, as they have been commonly used to evaluate both teaching material and teachers themselves [1], [2]. Student evaluations are often collected using Likert-type scales, from which numeric data is easy to collect, process, analyze, and present as an indicator of quality. Unfortunately, not only are student evaluations a hard resource to use to their full potential, the validity of evaluations as a measurement of quality is questionable [3], and evaluations do not necessarily reflect about students learning [4]. In addition to numeric data feedback, forms often contain open-ended questions prompting the students to leave free-form feedback. However, these open text feedbacks are more challenging to analyze in large numbers, as they would require work-intensive qualitative analysis.

This paper presents a study on the use of open-ended student evaluations collected in large quantities. Specifically, the objective of this study is to apply a topic modelling method as a systematic, automated way to analyze open text feedbacks in the masses. The main research question this paper addresses is, **what can be learned from topic models of open student feedback**? The main research question is further divided into sub-questions, which are listed as follows.

- What themes can be recognized from the topic model, and how can the emerging topics be described?
- How do the discovered themes relate to course arrangements, course topics, student motivation, teaching methods, and the teachers competence?
- Which themes or topics can be addressed by course staff and how?

The working hypothesis is that particular themes and development agendas can be distinguished from masses of open-text feedback by using statistical analysis methods, such as topic modelling. This would benefit the oversight of higher level education on an institution-wide level or a study programme wide level.

Rest of the paper is structured as follows. Section 2 presents related research on the text-mining and analysis of student feedback. Section 3 presents the research method and the data collection procedure. In section 4 the results from the Latent Dirichlet Allocation (LDA) [5] analysis are presented. Finally, section 5 discusses the results, and section 6 concludes.

II. Related Work

Analyzing the open texts in student evaluations of teaching is by no means a novel concept. Alhija & Fresko [6] investigated what can be learned from students written comments by performing a manual content analysis for 3067 collected feedbacks. The study distinguished three major domains of feedback (the course, the instructor, and context of instruction), in addition to individual content areas within the domains (such as course content, assignments, teaching style, scheduling and student composition). The study also concluded that comments tend to be more positive than negative and general in their nature. In a similar vein, Brockx et al. [7] also analysed comments left in 2029 student feedback surveys by using manual inspection and coding. The authors conclude that positive comments deal with the combination between theory and practice, whereas negative comments focus on the evaluation and context of the course.

TABLE I
CORRELATION MATRIX OF THE TOPIC PROBABILITIES (N = 6087)

	Topic 1	**Topic 2**	**Topic 3**	**Topic4**	**Topic 5**	**Topic 6**
Topic 1	1					
Topic 2	-0.3071*	1				
Topic 3	-0.3164*	-0.1009*	1			
Topic 4	-0.1586*	-0.1088*	-0.2192*	1		
Topic 5	-0.0810*	-0.2573*	-0.2451*	-0.1311*	1	
Topic 6	-0.0458*	-0.3314*	-0.1779*	-0.3339*	-0.1698*	1

* $p<0.05$

Grebennikov & Shah [8] analysed 78800 comments from study programme feedback surveys from 2001 to 2011 using a text analysis tool. The analysis yielded 26 different comment categories, which could be broadly grouped into five main domains. The use of text analytics tools for analysing student feedback has previously been explored in several studies with smaller sample sizes, for example by Kabanoff et al. [9], as Santhanam et al. [10], and Stupans et al. [11].

III. METHODS

The student feedback data used in this study come from the feedback surveys collected at LUT University during the academic years 2016-17 and 2017-18. In the university in question, a development project of the feedback questionnaire started in 2015. Therefore, the questionnaire has been subject to a number of revisions during the last few years. The first questionnaire (2016-17) had one broad open-ended question: "*Other feedback about the course (for example, ways to enhance learning during the course)*". The second questionnaire (2017-18) had five more focused open-ended questions: "*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:*", and "*What factors promoted my learning and how could learning be supported better?*" In addition, both questionnaires included several 5-point Likert-scale questions about, for example, students motivation, teaching methods, workload, and perceived learning. The link to the surveys was distributed via email to all enrolled students of all courses approximately half a week after the course had ended. Responding to the surveys was voluntary and anonymous.

The total number of student feedback questionnaires collected was 9148 in 2016-17 and 8092 in 2017-18. For the topic modelling, we restricted the sample to the responses that contained at least one answer to the open-ended questions written in Finnish. The resulting sample size for topic modelling was 6087 including 2445 surveys collected in 2016-17 and 3642 surveys collected in 2017-18.

We used the LDA topic modelling [5] as a statistical text mining method [12] to distinguish recurring themes from student feedback. The underlying mechanism in LDA is a probabilistic Bayesian network model, in which each document is characterized by certain topics, and each topic is defined by a specific set of words, which co-occur with a certain probability. To summarize, the topics of each document are defined by a set of words that often appear together. The algorithm is further explored by their inventors in [12]. A summary of the algorithm's process is presented as follows from a list originally published by Chaney et al. in [13].

1) For K topics, choose each topic distribution β_k. (Each β_k is a distribution over the vocabulary.)
2) For each document in the collection:
   (a) Choose a topic assignment z_n from ϕ_d.(Each z_n is a number from 1 to K.)
   (b) For each word in the document
   i. Choose a topic assignment z_n from ϕ_d. (Each z_n is a number from 1 to K.)"
   ii. Choose a word w_n from the topic distribution $\beta_z n$. (Notation $\beta_z n$ selects the z_nth topic from step 1.)

For the analysis, we used a modified version of the NAILS script [14], which utilizes the topicmodels R package [15] and visualized with the LDAvis library [16]. Semantic coherence, a quality value for deciding the number of topic models [17], was calculated using the R stm library [18]. LDA-based topic modeling is a commonly used method for text analysis and equivalent methods have been used to statistically analyze scientific texts in earlier studies [19], [20], [21]. Our analysis process is as follows.

1) Download student feedback data.
2) Sort the feedback by metadata such as course type and language.
3) Select the subset to be analyzed; in this case Finnish language responses.
4) Preprocessing:
   i) Use the R textmining [22] library's Finnish stopwords list
   ii) Stem the words using the snowball stemmer library [23]
5) i) Assign each row of feedback into a single document unit.
   ii) Break the word content of document into unigrams.
6) Run the NAILS script[1] [14] on the data, which performs the following:

[1]Available in GitHub at https://github.com/aknutas/nails

TABLE II
DESCRIPTION OF THE TOPICS

Topic	Name	Keywords
1	Positive comments about arrangements	Exam, good, exercises, lectures, moodle, material, better, homework
2	Dissatisfaction in the teaching	Student, teacher, every, return, some, questions, own, even, entire, point, help, get, done
3	Comments about course arrangements and deadlines	Practical assignment, time, group, same, period, workload, weekly assignment, instruction, other
4	Lack of motivation and support	Task/assignment, doing, always, exercises, weekly, right
5	Interest in the topic, understanding the material	Topic, more, little, less, pass/through, part, time, example, felt
6	Comments about interesting, rewarding but challenging courses	Topic, good, learning, lecturer, interesting, fitting, motivation

i) Calculates the optimal number of topics with R stm library [18].
ii) Performs LDA topic modeling with R topicmodels package [15].
iii) Visualizes the topicmodel with the R LDAvis library [16].

7) Naming the themes:
i) Manual inspection of results, and
ii) Assigning theme names based on the topic models' keywords.

IV. RESULTS

Initially the LDA algorithm was run multiple times with different parameters in order to discover most suitable models for the analysis. In total we created a series of 26 models with a range of 4 to 30 topics, which were then evaluated with the semantic coherence quality measure. Several local quality value maximums were selected for qualitative evaluation and after final evaluation a six topic model was chosen. A map of the topics and the relative distances between them are visualized in Figure 1. The inter-distance map shows that topics in the chosen model are relatively well distinct, even though some topics are more close to each other.

In addition to the inter-topic distances using the Jensen-Shannon divergence method (see Figure 1), we assessed topic similarities using Pearson correlation coefficients (see Table I). As shown the correlations are quite low ($|r|$= 0.0458 - 0.3339) indicating that the topics in the model are not strongly correlated with each other (the probability of a topic belonging in a certain topic is not too dependant on other topics). LDA is a probabilistic model and we used the probabilities returned by the algorithm to assign each observation to its most likely topic. The number of feedbacks assigned to one topic varies between 13% and 21% (see Table III): Topic 1 has the most number of observations (20.8% of the total). Topic 6 (19.25%), Topic 3 (17.1%) and Topic 2 (15.62%) are the next largest categories. Topic 5 (13.91%) and Topic 4 (13.31%) contained a similar number of observations.

The topic modeling results and the themes we discovered are summarized in Table II. Next we present the descriptions of the topics, based on the keywords and the authors' quick assessment of the most central texts within a topic: Because LDA is a probabilistic method, the texts with the highest probability to belong in one topic can be treated as canonical examples of what themes the topic contains.

As depicted in Figure 1, the words occurring in Topics 1, 3, and 4 are close to each other. The common theme for these topics were course arrangements. Common keywords for these topics were all related to course arrangements and organization, such as "assignment," "exercises," or "homework."

However, during a manual inspection of the most central feedback texts, we found that the Topics 1 and 3 contained mostly positive or neutral feedback, while Topic 4 contained negative feedback regarding the course arrangements. Topic 1 was generally the most positive feedback category, while Topic 3 contained discussion about stress or hurry, and Topic 4 discussed poor course support or lack of motivation. All topics acknowledged positive things about the studied subject.

Topic 2 contained very specific keywords, from which it was hard to categorize the topic as positive or negative. Again, upon further analysis of the texts, manual inspection revealed that Topic 2 mostly discussed the dissatisfaction with teaching and the students professional relationship with the teacher. Also, the Topic 2 was most concentrated on what the student might have needed, but did not receive.

Topics 5 and 6 discussed the subject of the course or the material. Again, these topics were similar in their keywords, so further manual inspection was required. Generally both topics acknowledged that the subject or material was interesting or beneficial, but Topic 6 also contained discussion about challenging topic or positive stress.

TABLE III
FREQUENCIES AND PERCENTAGES OF THE TOPICS

Topic	f	%
1	1266	20.80
2	951	15.62
3	1041	17.10
4	810	13.31
5	847	13.91
6	1172	19.25
Total	6087	100

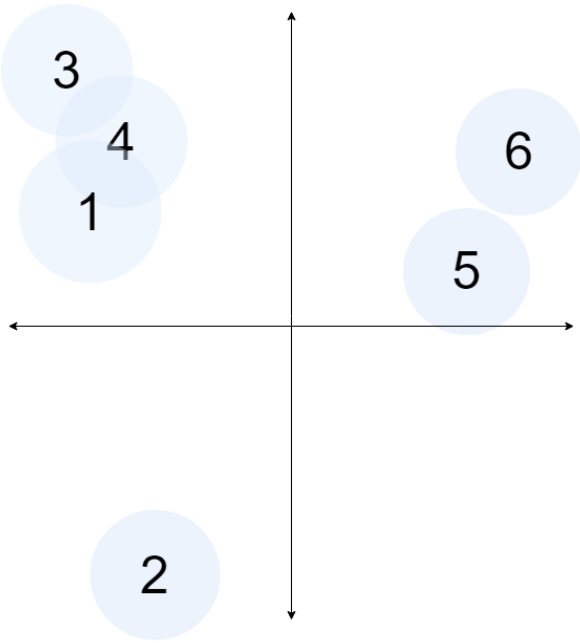

Fig. 1. The intertopic distance map visualizing the topic-term relationships using the LDAvis library [16]

V. Discussion

Our study demonstrates that this kind of analysis can be used to supplement quantitative student feedback evaluations. We identified several themes relevant to analyzing and improving teaching and learning outcomes, such as themes related to materials, course arrangements, student motivation, and course schedules. These themes can be used to give more context and pinpoint issues that need improvement.

The themes distinguished in this study are in line with the types of feedback presented by Alhija & Fresko [6], with the key difference that we could not see any disposition towards positive comments being in the majority in our results. In contrast to Brockx et al. [7], whose study suggests negative comments focus on course evaluation and context, we could distinguish one topic of positive comments within these themes and course arrangements.

Next, we address the individual research questions. To answer the first research question *what themes can be recognized from the topic model, and how can the emerging topics be described*, we can state that several different themes could be identified, from motivating subjects to student critique of course arrangements.

For the second research question *how do the discovered themes relate to course arrangements, course topics, student motivation, teaching methods, and the teachers competence?* We found several topics, each of which is a theme connected to one aspect of teaching, including motivation, teaching arrangements and professionalism.

Finally, for the third research question *which themes or topics can be addressed by course staff and how?* Each of the discovered topics provide actionable feedback for the teaching staff, or for the use in curriculum design. The themes are important as a single student feedback text alone might not be useful but if, for example, several students give feedback about the lack of support, this can indicate that some course arrangements should be altered or communicated better to the students.

All in all, to answer our main research question guiding the process of this study *what can be learned from topic models of open student feedback*, we conclude that analyzing open text feedback using statistical or machine learning methods can be used to as a source for course (or program) wide qualitative information. However, the interpretation of this automated analysis may be difficult: Even though some emergent themes behind the feedback can be distinguished, some topics are harder to utilize in practice.

The main contribution of this study is the constructed topic model which distinguishes themes arising from open student feedback. The used data set is similar in size in comparison to previous work, such as [6] and [7]. Some themes were related to the student's perception of the quality of teaching, similarly to the study by Kabanoff et al. [9]. The study also adopted the practice of using manual analysis to some degree to ensure the results are meaningful and contextually relevant, as suggested by Santhanam et al. [10].

VI. Conclusion

Our research goal was to examine which kind of themes emerge from large datasets of free-form text student feedback. We accomplished this by processing 6078 observations of unique student feedbacks using the LDA algorithm with the R programming language and environment for statistical computing. With this method, we discovered six topics discussing different aspects of teaching and learning. The generalized results contribute to existing knowledge about the types of feedback students leave in their course evaluations.

There are, of course, limitations concerning this study which need to be addressed. The topic model with six themes was chosen using the semantic coherence measure, which is a statistical method for determining the quality of the model. A more thorough approach would have required a qualitative evaluation of alternative models. However, we did choose a relatively low number of topics which suggests that the resulting themes can be generalized. In addition, we calculated the Pearson correlation coefficients to verify that the topics are distinct from each other.

Furthermore, the analysis of the resulting topic model was mainly based on the descriptive keywords (produced by the LDA algorithm) and an informal inspection by the researchers. For this reason, there is a possibility that some topics description may be misrepresented in our presentation of the results. We, therefore, must acknowledge the risk of researcher bias, even though we try to limit it by working in a group of researchers, and the fact that the analysis is based on the results of a statistical method. In order to properly dismiss this risk, we propose

a thematic analysis approach that enables a systematic, qualitative evaluation of the topic modelling results to be carried out in future work. Additionally, the robustness and generalizability of using the topic modeling method on student feedback requires further study.

Other future research avenues could include the automatic processing of open feedback, or longitudinal studies with feedback collected from several successive years. The dataset in this study was also limited to one language only (Finnish), meaning that there is more work to be done in analyzing the responses which were left out here.

REFERENCES

[1] 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 & Evaluation in Higher Education*, vol. 25, no. 4, pp. 397–405, 2000.

[2] F. Zabaleta, "The use and misuse of student evaluations of teaching," *Teaching in Higher Education*, vol. 12, no. 1, pp. 55–76, 2007.

[3] P. Spooren, B. Brockx, and D. Mortelmans, "On the validity of student evaluation of teaching: The state of the art," *Review of Educational Research*, vol. 83, no. 4, pp. 598–642, 2013.

[4] 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," *Studies in Educational Evaluation*, vol. 54, pp. 22–42, 2017.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[6] F. N.-A. Alhija and B. Fresko, "Student evaluation of instruction: what can be learned from students written comments?" *Studies in Educational Evaluation*, vol. 35, no. 1, pp. 37–44, 2009.

[7] B. Brockx, K. Van Roy, and D. Mortelmans, "The student as a commentator: students' comments in student evaluations of teaching," in *Procedia: social and behavioral sciences*, 2012, vol. 69, pp. 1122–1133.

[8] L. Grebennikov and M. Shah, "Student voice: using qualitative feedback from students to enhance their university experience," *Teaching in Higher Education*, vol. 18, no. 6, pp. 606–618, 2013.

[9] B. Kabanoff, A. Richardson, and S. Brown, "Business graduates' perceptions of the quality of their course: A view from their workplace," *Journal of Institutional Research*, vol. 12, no. 2, p. 1, 2003.

[10] E. Santhanam, B. Lynch, and J. Jones, "Making sense of student feedback using text analysis–adapting and expanding a common lexicon," *Quality Assurance in Education*, vol. 26, no. 1, pp. 60–69, 2018.

[11] I. Stupans, T. McGuren, and A. M. Babey, "Student evaluation of teaching: A study exploring student rating instrument free-form text comments," *Innovative Higher Education*, vol. 41, no. 1, pp. 33–42, 2016.

[12] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, p. 77, Apr. 2012.

[13] A. J.-B. Chaney and D. M. Blei, "Visualizing topic models," in *Sixth international AAAI conference on weblogs and social media*, 2012.

[14] A. Knutas, A. Hajikhani, J. Salminen, J. Ikonen, and J. Porras, "Cloud-based bibliometric analysis service for systematic mapping studies," in *Proceedings of the 16th International Conference on Computer Systems and Technologies*. ACM, 2015, pp. 184–191.

[15] B. Grün and K. Hornik, "topicmodels: An R package for fitting topic models," *Journal of Statistical Software*, vol. 40, no. 13, pp. 1–30, 2011.

[16] C. Sievert and K. Shirley, *LDAvis: Interactive Visualization of Topic Models*, 2015, r package version 0.3.2. [Online]. Available: https://CRAN.R-project.org/package=LDAvis

[17] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 262–272.

[18] M. E. Roberts, B. M. Stewart, and D. Tingley, *stm: R Package for Structural Topic Models*, 2018, r package version 1.3.3. [Online]. Available: http://www.structuraltopicmodel.com

[19] J. Kasurinen and A. Knutas, "Publication trends in gamification: A systematic mapping study," *Computer Science Review*, vol. 27, pp. 33–44, 2018.

[20] B. Penzenstadler, A. Raturi, D. Richardson, C. Calero, H. Femmer, and X. Franch, "Systematic mapping study on software engineering for sustainability (SE4s)," in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. ACM, 2014, p. 14.

[21] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles." ACM Press, 2011, p. 448.

[22] I. Feinerer, K. Hornik, and D. Meyer, "Text mining infrastructure in r," *Journal of Statistical Software*, vol. 25, no. 5, pp. 1–54, March 2008. [Online]. Available: http://www.jstatsoft.org/v25/i05/

[23] M. Bouchet-Valat, *SnowballC: Snowball Stemmers Based on the C 'libstemmer' UTF-8 Library*, 2019, r package version 0.6.0. [Online]. Available: https://CRAN.R-project.org/package=SnowballC