Ullmann, T., Rienties, B. (2021). Using text analytics to understand open ended student comments at scale: Insights from four case studies. **Assessing and Enhancing Student Experience in Higher Education**. In M. Shah, J.T.E. Richardson, A. Pabel, & B. Oliver (Eds). Palgrave Macmillan.

Using text analytics to understand open-ended student comments at scale: Insights from four case studies

Thomas Ullmann & Bart Rienties
Institute of Educational Technology, Open University UK, MK76AA, Milton Keynes.

Students write tens of thousands of open-ended comments in student evaluation questionnaires, which are collected as part of institutional and national surveys. Often as part of a quality enhancement strategy, teachers analyse these comments in order to gain insights into student perspectives and guide revisions of modules. While institutions have access to enormous amounts of qualitative data, to date limited efforts have been made to analyse and disseminate these data, which could be used by academics and administrative leaders to identify areas of good practice and areas needing improvement. This chapter will examine several innovative uses of qualitative data with automated text analytics (i.e., natural language processing) used to assess and enhance the student experience. Using four case studies from the Open University UK, we will discuss the affordances and limitations of such methods. We found strong differences in quality and quantity of contributions to student comments based upon individual and disciplinary factors.

Keywords: student comments, text analytics, student experience, surveys

Introduction

A key concern for most higher education institutions (HEIs), teachers, and readers of this book in particular is whether students, or learners in general, are "satisfied" with their learning experience (Kember & Ginns, 2012; Li, Marsh, & Rienties, 2016; Li, Marsh, Rienties, & Whitelock, 2017; Richardson, 2013). These days most institutions are using some form of student experience evaluation instruments to monitor and/or to improve the teaching and learning experience (Arbaugh, 2014; Boring, Ottoboni, & Stark, 2016; Rienties, 2014). In particular in a UK higher education context student evaluation scores are important, as HEIs are ranked on a yearly basis using surveys, such as the National Student Survey (NSS, Langan & Harris, 2019; Richardson, 2013).

The analysis of student evaluation data allows institutions and teachers to unpack what students liked and did not like in terms of teaching and learning, and perhaps search for unobserved patterns and underlying information in learning processes (Arbaugh, 2014; Li et al., 2017). For a detailed overview of the concepts of student evaluation approaches and how to measure student evaluation data, we refer to elsewhere in this book (see Chapter XXX). Recently there is a wealth of big data studies that have shown that understanding how students go about their studying and their orientations to their learning environment will affect their satisfaction rating (Arbaugh, 2014; Langan & Harris, 2019; Rienties & Toetenel, 2016).

For example, one study measuring which factors predicted learner satisfaction and academic performance amongst 48 MBA online and blended learning courses in the US, Arbaugh (2014) found that learners' behaviour impacted student satisfaction and academic performance. In a follow-up study linking 116,646 students' evaluations with learning design characteristics of 422 blended and online courses, Li et al. (2017) found six theoretical blocks of core constructs which impacted student satisfaction (i.e., module design, presentation of course, learner characteristics, learner characteristics linked to respective course, learner

history, workload). In a recent study of 1.8 million completed NSS surveys, Langan and Harris (2019) found a stable structure of the NSS instrument, while at the same time an increasing harmonisation of scores across disciplines and universities, thereby questioning how such evaluation results should be used to compare degrees, disciplines, or even institutions.

There is now a wealth of both small-data (Dommeyer, Baum, Hanna, & Chapman, 2004; Gamliel & Davidovitz, 2005) as well as big-data studies (Boring et al., 2016; Langan & Harris, 2019; Li et al., 2017; Rienties & Toetenel, 2016) that have explored how student evaluation scores can be used to understand and perhaps even compare the student experience, and to measure learning on a module, course, qualification, degree level, or even across institutions. However, relatively few studies (e.g., Grebennikov & Shah, 2013; McDonald, Moskal, Goodchild, Stein, & Terry, 2020; Shah, 2019; Shah, Nair, & Richardson, 2017; Zaitseva, Milsom, & Stewart, 2013) have explored how qualitative data from student evaluation tools could potentially be used to get detailed insights of the lived experiences of students.

Typically, in student evaluation instruments 2-4 open questions are included whether/why/how (or not) students enjoyed a particular course (Kember & Ginns, 2012; Shah, 2019; Shah et al., 2017). Students often contribute dozens of comments each year to student evaluation questionnaires. Many teachers have reported that these qualitative comments are extremely useful to know what students liked, and disliked, and often are used to inform pedagogical and learning design changes (Kember & Ginns, 2012; Moskal, Stein, & Golding, 2015; Rienties, 2014). However, very few institutions use these qualitative data on a large scale, as the process of the manual analysis of the student comments becomes intractable with growing student responses (Richardson, 2005), as the manual analysis is time-consuming and resource-intensive.

In this chapter, building on previous work (Clow et al., 2019; Ullmann et al., 2018) we aim to apply a big data perspective to the analysis of the tens of thousands of comments received by the Open University UK (OU) as part of student evaluation feedback. It uses automated empirical text analysis methods to gauge the 'hot' topics that students talk about in an academic year, and it evaluates the sentiment that students express towards these topics. By demonstrating the potential affordances and limitations of big data analyses of qualitative student evaluation data, we hope to contribute to an emerging discussion how institutions might use rich and detailed insights from qualitative student evaluation data on a large scale.

Using text analytics approaches to understand qualitative student evaluation data

At present most teachers analyse student comments 'long-hand' by reading each comment in turn. When analysing student comments, teachers may focus initially on categorising and ordering the seemingly diverse range of statements written by their students. In line with good-practice guidelines (HEFCE, 2016; Kember & Ginns, 2012; Moskal et al., 2015), often the goal of teachers is to arrive at a set of key themes or topics that can help to provide focus for the improvement of a course as well as to better understand the "strengths" and "weaknesses" of the course. These analyses are based on the students' comments in front of them, yet it can be difficult to know if these comments are representative of the group who responded to a student evaluation.

Analysing these comments using conventional qualitative methods is often time consuming, in particular for large online courses, and potentially subject to bias in identifying themes, especially if being done quickly. Indeed, some recent studies have highlighted the potential of automated analysis of student comments (Grebennikov & Shah, 2013; McDonald et al., 2020; Shah, 2019; Zaitseva et al., 2013). For example, Zaitseva et al. (2013) compared data on student evaluations from Year 1, Year 2, and NSS over a period of three years at one UK HEI and collated 148,928 comments using automated semantic analysis, leading to integrated concept maps of themes and clusters of concepts. Furthermore, McDonald et al. (2020) developed a proof of concept of automated text analysis by comparing 1,168 comments on two open comment questions with Quantext. The authors found some support that automated analysis could lead to comparable results to manual coding. Finally, Shah (2019)

compared 16,582 comments from on-campus and online students in an Australian context using Leximancer and found substantially different experiences and identified themes between these two groups of students.

In other words, with big qualitative data researchers could start to explore the power of natural language processing (NLP). There are several text analytics approaches that could potentially help teachers to make sense of these large amounts of qualitative data, in a similar way to common manual content analysis approaches. For example, one potential text analytics approach is to use *keyword analysis* (Coughlan, Ullmann, & Lister, 2017; Ullmann, 2015b, 2017), which finds words/themes that students talk about unusually often, or perhaps less than expected (Rayson, 2008). In addition, a *dictionary-based approach* (e.g., Ullmann, 2015a) uses a list of words that are closely associated with a theme. If a student evaluation comment contains a respective dictionary word then it is annotated with the category label, mimicking the manual annotation task. A main merit of this dictionary approach is that there are some large off-the-shelf dictionary approaches that could be adopted. Alternatively, teachers could work together with specialists to specify words or themes of interest that could help to identify pedagogically relevant issues.

Furthermore, an increasing number of researchers use so-called *sentiment analysis* (Jeonghee, Nasukawa, Bunescu, & Niblack, 2003; Leong, Lee, & Mak, 2012; Wen, Yang, & Rosé, 2014; Zaitseva et al., 2013). In the simplest case the category classes of a sentiment framework are positive and negative. Once the text is annotated, the frequency counts for each category can be used as input for statistical tests, such as the X^2-test (Ullmann et al., 2018). For example, using sentiment analysis Zaitseva et al. (2013) found that first-year students moved from more affectively oriented goals to learning and goal reaffirmation in their second year, and finally focussed more on achievement and outcome-oriented learning in their third year.

As illustrated in Table 1, there are a range of text analytics options available to mimic typical tasks of the manual content analysis. While there are now several studies that have started to use NLP to explore student evaluation comments (e.g., McDonald et al., 2020; Shah et al., 2017; Zaitseva et al., 2013), few studies have explored how "representative" these comments are, and whether unique voices could be potentially "hidden" in these comments. In this chapter we aim to highlight several examples using four cases. Each case study shows exemplary a text analytics technique for each of the typical tasks of a manual content analysis as shown in Table 1. Yin (2009) emphasised that a case study approach investigates a phenomenon in-depth and in its natural context. While we do not claim that these case studies are best-case or exemplary best practice, by highlighting how we work in practice with teachers we hope that readers might be persuaded to potentially explore similar approaches.

Table 1 Mapping between manual content analysis and automated text analysis techniques

Manual content analysis	***Text analysis used in here***
Exploring themes	Keyword analysis
Annotation of text according to a framework	Dictionary-based approach Sentiment analysis
Statistical analysis based on manually annotated data	Statistical analysis based on automatically annotated data

Source: Clow et al. (2019)

In case study 1 we aim to use keyword analysis in conjunction with dictionary approaches to explore themes that students are commenting on in terms of student evaluations for a large range of online courses. Beyond what students are describing, in case study 1 we will explore whether text mining could help to provide insights and whether qualitative comments provided by students are "representative" of the wider student voice (or not). In other words, to what extent do some students tend to answer the open-text comments more than others? As is highlighted by a wealth of studies individual differences and context can directly or indirectly influence learning processes, outcomes, and lived experiences by

students (Arbaugh, 2014; Li et al., 2017; Richardson, Mittelmeier, & Rienties, 2020). Are there discernible variations between groups of students by gender, socio-economic status or disability for example? Should it matter that we may not have as many comments from some sub-groups of students?

In case study 2 we aim to explore whether the sentiments expressed by students in student evaluations are similar or different across courses. Using sentiment analysis (Leong et al., 2012; Socher et al., 2013; Wen et al., 2014; Zaitseva et al., 2013) we will explore whether there are perhaps some courses that receive mostly "positive" comments, while others might mainly receive "negative" comments. By combining these student evaluation feedback with other metrics and perspectives from teachers, this potentially could be useful for quality enhancement.

In case study 3 we will explore an example of statistical analysis of automatically annotated data, whereby we will compare the expressed experiences of students following a mathematics and statistics qualification versus students who are following different qualifications but take part in the same maths and stats courses. One might assume that disciplinary differences might positively or negatively influence how students perceive the learning experience in multi-disciplinary and/or interdisciplinary courses (Borrego & Newswander, 2010; Rienties & Héliot, 2018), and text mining approaches allow us to contrast these potential differences.

Finally, in case study 4 we aim to contrast the student evaluation comments of high versus low performing students. By comparing which keywords high and low performing students are more (or less) often using, and by linking these with actual student comments, we will explore whether students' experiences of courses are indeed "universal", or whether substantial differences might be visible. Overall, these four cases are selected to illustrate some of the potential affordances and limitations of using automated text analysis approaches, which by definition would be very time consuming or difficult to implement using manual content analysis.

Method and approaches used

Context and setting

This study took place at the OU, a distance-learning institution with an open-entry policy and the largest university in the UK. Open-entry means that a person can study towards most degrees without a typical prior qualification, as it is often required by other HEIs. In the past thirty years, the OU has consistently collected learner feedback to further improve the learning experience and learning designs (Ashby, Richardson, & Woodley, 2011; Richardson, 2005, 2006). In line with other learner satisfaction instruments, at the OU the Student Experience on a Module (SEaM) questionnaire was implemented, which has been tested and validated by a range of studies (e.g., Li et al., 2016; Li et al., 2017; Nguyen, Rienties, Toetenel, Ferguson, & Whitelock, 2017; Rienties & Toetenel, 2016). At the OU each course is taught by several lecturers (also referred to as tutors). Each tutor is responsible for a small group of 15-20 students. This report presents initial analysis of over 50,000 comments from a sample of large undergraduate courses. These comprise a substantial dataset on which to trial the techniques outlined below.

Instrument: Student Experience on a Module (SEaM)

In addition to 40+ closed questions, the SEaM questionnaire contains four open-ended questions that asks students about their study experience. Answers to three of these questions were used here – the fourth concerns tutors and was only viewable by the respective tutor. The relevant three SEaM questions were:

- *Question 1:* What aspects of teaching materials, learning activities or assessment did you find particularly helpful to your learning?

- *Question 2:* What aspects of teaching materials, learning activities or assessment did you find not particularly helpful to your learning? We would welcome any further suggestions or comments to consider for future editions of the module.
- *Question 3:* Do you have any other comments to add about your study experience on this module?

The first two questions asked about what was helpful (positive aspects), or not so helpful (negative aspects). The third question was a catch-all question to capture all other concerns. These questions gave students the opportunity to provide feedback, opinion and comment in their own words about their course and, more broadly, their experience of studying at the OU. Free-text response questions are considered more effective than closed-answer questions in capturing the nuance and richness of the student voice (HEFCE, 2016; Richardson, Slater, & Wilson, 2007; Shah et al., 2017). Participation in SEaM was voluntary so not all OU students responded (Li et al., 2016; Li et al., 2017) and, from the students that did respond, not all wrote comments. Inspection of the SEaM responses showed that students did not always directly answer the question asked, and might write about issues that were not directly related to a particular question.

Data analysis

Traditionally, student comments from the SEaM survey were manually analysed by OU staff. There are certain restrictions to this approach, which we sought to overcome with a set of automated methods that mimic largely the manual content analysis approach. The approach adopted in this book chapter is a form of NLP (Jeonghee et al., 2003; Socher et al., 2013; Ullmann, 2015b, 2017; Ullmann, Wild, & Scott, 2012; Wen et al., 2014). This is a method that can be applied to large datasets and thus allows us to perform analysis on student comments from many courses. However, a limitation of this NLP approach is that it lacks an understanding of written text as well as the contexts in which students left their comments. The position of this chapter, therefore, is that the automated analysis cannot replace the manual analysis, but it can provide means to assist the sense making of student comments.

The analysis was based on answers to the 2013-2017 version of the SEaM survey. In early 2018 a new version was introduced and this made changes to two of the three questions under investigation. OU ethics policy permits the use of SEaM data for university data analysis, and before the survey students were explicitly told about the data protection policy of the OU and that their data can be used for research and quality enhancement purposes (Open University UK, 2014). The analysis of open-text comment data presented below conformed to these stated uses and data were anonymised before use.

Results

Case study 1: What are students talking about?

In order to get an understanding of what students were talking about in terms of student evaluation comments in selected courses in the period 2013-2017, we first explored quantitatively how often students contributed to open-ended questions. Secondly, we explored whether there was a potential free-text item non-response bias in terms of gender, socio-economic status, and disability in terms of contributing to student evaluation comments. Thirdly, we explored what students were actually talking about using NLP approaches by highlighting the top 10 unique (or lack of) contributions in 2015 and 2016.

Length of student comments

Our analysis showed that over 2.5 million words were used by students in the *subsample* of 50K+ comments used in this analysis. Furthermore, whilst the number of comments declined from question 1 to question 3, the number of words per question increased. On average students that wrote a comment to question 1 wrote 40 words, 48 for question 2 and 67 for question 3. To put this in context, given a reading speed of 250 words per minute it would take two months to read all these comments (16 working days to read all answers to question 1 and 2, and nearly 7 days for question 3). Obviously, it would take even longer to code, analyse and interpret these comments. Therefore, putting qualitative student evaluation data into a

common dataset and use text analytics approaches could potentially be a cost-saving exercise for both teachers and managers.

Who are the students that comment?

So how much variation in commenting could we identify in answers to SEaM? On average, based upon historical response data, a course can expect that 24% of students that were invited to participate would answer closed-questions in the survey, as well as leave at least one comment. In fact, 83% of those students that participated in SEaM left at least one comment, so in a way only 20% of the student voices are typically "heard" in student evaluations at the OU. Of course, a crucial follow-up question was whether the 83% of commenting students were different from the 17% of students who filled out the questionnaire but did not write a comment?

Female students commented significantly more often than male students ($X^2(1) = 150.76$, $p < .001$), see Table 2. The odds of writing a comment were 1.5 times higher if a student was female compared to male students, although females did not necessarily write more words relative to men. Small but significant differences were found when comparing the analyses per faculty, where for some faculties there was a significant difference in terms of gender, and for others not.

Table 2 Results of Chi-squared analysis to determine whether non-response significantly varies for gender, socio-economic status and disability per faculty

	Gender		***Socio-economic status***		***Disability***	
Faculty	*$X^2(1)$*	***p***	*$X^2(1)$*	***P***	*$X^2(1)$*	***p***
All	150.76	.00**	18.36	.00**	0.99	.32
Faculty1	19.27	.00**	8.52	.00**	0.65	.42
Faculty2	8.57	.00**	0.02	.88	1.46	.23
Faculty3	37.15	.00**	7.86	.01**	0.87	.35
Faculty4	2.42	.12	4.24	.04*	0.67	.41

*** p <.01. *p<.05.* Source: Ullmann et al. (2018)

In general, students from a low socio-economic background commented less often than students from a higher socio-economic background ($X^2(1) = 18.36$, $p < .001$). The odds of commenting were 1.27 times higher if the student came from a higher socio-economic background. Furthermore, students from low socio-economic status wrote on average shorter comments (Mean = 103, Mdn = 74) than students from a higher socio-economic status (Mean = 113, Mdn = 82, Wilcoxon rank-sum = 17035319, $p < .001$, $r = -.03$). Finally, there was no significant difference between students who declared a disability and those who did not ($X^2(1) = 0.99$, $p < .32$).

In other words, this analysis showed that students who wrote comments were different from students who elected not to give comments and only responded to the multiple-choice questions of the survey.

What are students talking about?

In line with Zaitseva et al. (2013), mere analysis of the most frequent words falls short when comparing free-text responses from year to year. Our analysis showed that several words appeared year after year in the top frequent list, such as module, course, tutor, tutorials, materials, books, and activities. Looking only at the top frequent words did not show any meaningful differences between academic years. Therefore, the log-likelihood ratio (LL) statistic was used to discern topics that students talk about every year from topics that are particular for a given academic year (Coughlan et al., 2017; Ullmann, 2015b, 2017).

Table 3 shows a cut-down list of (noun and multi-noun) words ordered by LL. This statistic took into account the frequency of each words, but also the overall frequency of words to adjust for data sets with different word sizes. The greater the LL, the more unusual frequent was the occurrence of that word in a year compared to all other years. For example, the noun word “face-to-face tutorials” was the word with the highest LL. This means that is was mentioned unusually frequently in 2016 compared to all other years. In 2016 it was used 560 times relative to 358 times on average in previous years. This could be related to a policy change at the OU, whereby the mode of delivery of tutorials was changed.

For 2016, an inspection of the word list shows that students talked unusually often about '*face-to-face tutorial(s)', tutor(s), face-to-face, tutorial(s), and 'online tutorials'*. Also, they talked about the *system or booking system*. Dates and distances seemed to be a topic indicated by words such as *year, date(s), miles, and distance,* which mostly refer to *specific tutorial events* and *distances* to these events. As indicated in Table 3, for 2015 the list of relatively 'overused' words suggested a different focus from 2016. They talked about *practise,* about *module resources,* such as for example *module book, networking book, textbook, websites, wiki,* and *e-portfolio*. Students also talked about certain subjects, such as *linux, robotics, networking, software, business,* and *education*, but also about values, such as *diversity, learning,* and *benefits,* and relative to previous years were less likely to talk about DVDs (which were phased out in 2015). In other words, these automated approaches could identify main themes and new emerging themes. These topics could provide an entry point to the manual analysis of student comments on a module level by teachers or analysts.

Table 3 Top 10 of excerpt of what students are talking about in 2016 vs 2015

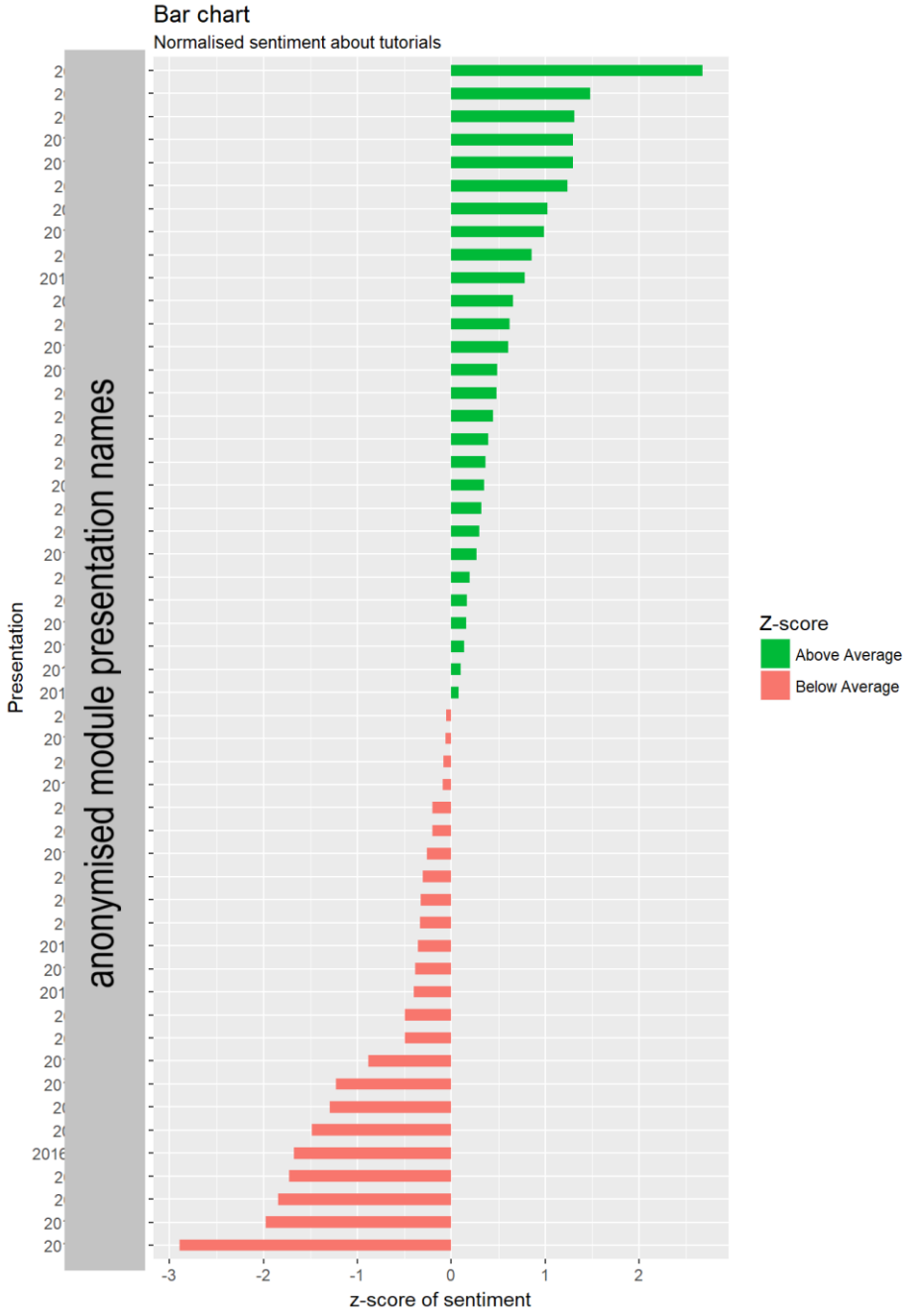
Term	Actual WC	Reference WC	LL	use
2016				
Face-to-face tutorials	560	1432	209.62	+
Tutors	434	1493	68.39	+
Face-to-face	117	254	60.67	+
System	111	253	52.64	+
Tutorials	1197	5242	49.27	+
Year	330	1191	42.73	+
booking system	32	32	41.88	+
online tutorials	140	403	39.18	+
Course	1345	8804	35.78	-
Date	185	632	29.94	+
2015				
Dvds	180	1074	18.72	-
Dvd	103	657	15.83	-
Course	1873	8804	14.68	-
Practise	26	49	10.23	+
module book	42	105	7.89	+
Networking	47	122	7.75	+
Linux	57	158	7.26	+
Choice	84	485	7.01	-
Face-to-face tutorials	284	1432	6.83	-
networking book	20	41	6.54	+

+ more frequent use in respective year. - less frequent use in respective year
Source: Ullmann et al. (2018)

Case study 2 Determining sentiment of SEaM across modules

Sentiment analysis combined with a dictionary-based text analysis offered a method of estimating the balance of positive and negative comments made by students. Sentiment analysis of the comments could provide extra information, which scores derived from the closed questions alone cannot provide, as they did not tell us why students chose to rate their experience in such way. Figure 1 shows a selection of high-population courses and the (course level) aggregated mean sentiment scores for different presentations of those courses, using the sentiment analysis tool developed by Socher et al. (2013). The means of each course in this visualisation were normalised in form of Z-scores. Z-scores showed how far, relative to the other courses included in the analysis, the course presentation was from the mean of other included courses. Most of the mean scores for individual courses centred around the global mean, but there were several courses with substantially higher Z-scores.

Figure 1 Normalised sentiment by respective course presentations



Source: Ullmann et al. (2018)

This sentiment analysis approach could assist in supporting teachers and taking decisions about which modules comments need further analysis or additional focus. By using an integrated approach of learning design, learning analytics data, and appropriate support (Rienties et al., 2016), several lessons could be learned when analysing courses with high positive sentiment, as well as understanding why students indicated more negative sentiment towards other courses.

Case study 3: Comparing comments from different disciplines in one core qualification
In many qualifications, students from different disciplines and/or specialisations follow a common set of combined courses (Clow et al., 2019; Rienties & Héliot, 2018). Several authors have argued that these combined qualifications could lead to specific issues for some groups of students (Borrego & Newswander, 2010; Rienties & Héliot, 2018). Indeed, one OU department was interested in potential differences between students studying courses towards their primary qualification in maths and statistics, and students studying the same courses but working towards a different qualification aim. The analysis included open comments from 22 courses from this respective department, which revealed ten top keywords as indicated in Table 4.

Table 4 Log-likelihood ratio of student comments of Qualification X vs non-Qualification X students

Term	Actual same qual.	Reference diff. qual.	LL-ratio	Use (Actual compared to reference)
unit	115	56	12.83	+
course	169	197	8.48	-
concept	48	69	7.63	-
knowledge	26	41	6.04	-
bit	38	54	5.74	-
answer	47	23	5.17	+
revision	40	20	4.11	+
feedback	24	35	4.06	-
module	455	447	4.02	-
issue	25	36	4.01	-

Source: Clow et al. (2019)

Following this analysis, a random selection of comments was inspected manually. This provided insights into the context in which the terms were being used and helped with the interpretation of results. Terms used much more frequently by same qualification students were ‘unit’, ‘answer’ and ‘revision.’ More frequent mention of ‘unit’ indicated that students might think more at the level of individual units of courses. Inspection of the open comments showed that they mostly used ‘unit’ in the context of the perceived quality of specific units in a course, or to a lesser degree how units ideally should interlink. Compared to students studying towards

a different qualification, students studying towards the maths and statistics qualification provided significantly more feedback specific to individual units. In contrast, students from different qualifications seemed to prefer more generic terms when talking about their learning, choosing words such as course or module, and expressed remarks about the observed quality of the whole course. The word 'concept' referred mostly to difficulty a student had with a concept, although in some cases it was used to express praise for how well the course writers explained a concept. This could indicate that department-qualification students were being more specific in their feedback and thereby, potentially, providing more ‘actionable’ feedback to the module team. By using text mining approaches of student evaluation comments teachers could potentially compare how students from different disciplines in their course react to their approach, and whether (or not) there is a need for specific personalisation for respective disciplines.

Case study 4: Keywords of high and low performing students

In our final case, we compared differences in one faculty who were interested in the differences regarding the student experience of high and low performing students. Inspection of the top keywords in Table 5 showed that high performing students mentioned the terms 'content', 'feedback', 'link', 'group', 'change', and 'interaction' unusually often when compared with low performing students. High performing students used the word 'content' mostly to refer to the perceived quality of the course content. High performing students were found to be very critical when it came to broken content. This could be seen from how they talked about the term 'link'. Students used the word 'link' mostly in the context of complaining about broken hyperlinks. They also used 'link' to a lesser degree to highlight certain connections that they have made while studying, such as a link between theory and practice, links between topics of the course, links to previous learning.

Table 5 Log-likelihood ratio of high vs low performance student comments

Term	Actual	Reference	LL-ratio	Use
content	83	1	18.17	+
help	62	28	17.13	-
teaching material	45	22	15.19	-
feedback	130	6	13.69	+
none	47	21	12.63	-
link	67	2	10.01	+
problem	32	15	9.73	-
experience	102	32	9.21	-
group	48	1	8.68	+
change	46	1	8.16	+
guide	63	21	7.06	-
interaction	41	1	6.87	+

Source: Clow et al. (2019)

Students used the word 'feedback' mostly to refer to the feedback they received from their tutor, but also to the written feedback about their marked assignments. For high performing students, feedback was relatively more important in order to understand how they performed compared to the expectations set by teachers. High performing students also debated more often about the ways of working with other students, using words such as 'group', or 'interaction'. They used it to refer to their experiences of group work, to talk about tutor groups (online and face-to-face), or collaborative exercises. It seemed that high performing students were more sensitive to problems related to group work compared to the low performing students. According to the respective faculty contact person, collaborative activities had not been essential and not part of the assessment for the specific set of courses. Possibly, high performing students might attempt these activities and therefore might have more of a say about them compared to low performing students that focussed more on the materials that were essential for the course and therefore did not attempt those additional collaborative activities.

In contrast, the terms 'help', 'problems' and 'guide' were more often used by low performing students. Students used the word 'help' mostly to refer to the help that students received from the OU, such as help from tutors, learning materials, and study guides, but also addressing the lack of help. The word 'problem' was used to note down various problems of students, such as problems with the course, but also problems with study skills. Finally, low performing students often used the word 'none'. This may seem a strange observation, however, 'none' was often used to indicate that a student had nothing to comment on. Perhaps this indicated that low performing students felt they were less willing to contribute to student evaluation, or had less time to comment. This was interesting as the use of the term ‘none’ would appear to go counter to an expectation that low performing students should have more to comment on. One potential reason for this contradiction might be that low performing students might not have the skills to verbalise their study experience, and this might have resulted instead in a short answer.

Discussion

This chapter aimed to illustrate how text analytics approaches could be used to better understand the qualitative comments from student evaluations in four unique case studies. There is an emerging body of literature that has used various forms of text analytics to explore student comments from student evaluations (e.g., McDonald et al., 2020; Shah, 2019; Zaitseva et al., 2013), although to the best of our knowledge few studies have focussed on how context and individual differences might influence these processes and outcomes. All the analyses above considered tens of thousands of students' comments from large courses over the last four years at the Open University UK (OU). Given that it is an intractable problem to manually analyse large amounts of student data, the four case studies showed some of the affordances and limitations of automated methods to analyse student comments.

All case studies were based on a requirement analysis of stakeholders from all faculties of the OU. The variety of case studies showed that text analytics can cater for a range of scenarios that are important for faculty staff. In our experience, many of these requirements can be approached analytically with one or more of the techniques of the manual content analysis shown in Table 1. In this chapter we showed three concrete text analytics techniques that can be related to the manual tasks. Although these text analytics techniques were relatively simple and are likely to be outperformed by more sophisticated text mining techniques, they also proved to be versatile enough to respond to many of the requirements of faculty staff. We recognise that this work is explorative in nature and work-in-progress, however preliminary feedback from teachers and senior managers at respective faculties indicated that such results from text analytics approaches of student comments could potentially be valuable (Clow et al., 2019).

These automated approaches can guide efforts to further analyse trends in comments, which can look to identify the reasons behind the prominence of these topics, why students talk so

much about them, and whether the comments suggest specific areas for attention or actions that can be taken in response. Furthermore, by adding sentiment analysis researchers and teachers can identify topics where the student respondents changed their views over time. Knowing this can support us to target these topics and to evaluate the reasons behind such changes.

One particular concern across the four case studies was that one cannot talk about THE student voice (i.e., a voice representative for all students). As highlighted in each of the case studies, some groups of learners were more inclined to contribute in terms of quantity and quality of student comments relative to other groups of learners, which is in line with previous studies (Shah, 2019). Perhaps to the unobserved eye of a teacher this might positively or negatively influence which pedagogical redesigns might be needed. We have shown that individual differences, such as gender and socio-economic status, as well as cognitive performance of students can influence (non)commenting behaviour. This is important to bear in mind when making generalisations about the whole student cohort, whether based on manual or automated analysis of SEaM comments.

Furthermore, another important lesson from the four case studies was that automated methods of analysis can identify which topics are more or less common in a particular year and compare against programme, faculty or university norms. This potentially could be powerful to determine which courses, qualifications, and/or disciplines receive high and positive student comments, while others might provide indications for further improvement. However, an important lesson to keep in mind is that several big data analyses have shown no relation between student satisfaction and grades (Nguyen et al., 2017; Rienties & Toetenel, 2016), so managers and teachers wanting to make firm decisions about an "under" or "overperforming" course need to be extremely careful.

As highlighted in this chapter, although there is a potentially rich treasure trove of quantitative and qualitative data in student evaluations, only a minority of students (and a subset of the entire student population) tend to complete and respond to student evaluations (Li et al., 2017; Rienties & Toetenel, 2016). In an extreme case a teacher who is unaware of these generalisability issues might "optimise" his/her course based upon student evaluation comments from mainly highly educated, high performing women who tend to write more, are specifically, more sophisticated and critical, and for example would want to have fewer graded collaboration activities and hints. These collaboration activities might be perceived as bringing their average grade down, and as they are well-prepared, they would not need hints, and might actually want more detailed follow-up materials beyond the learning outcomes. However, this might adversely impact "lower performing" students from lower socio-economic status, who are less likely to comment on student evaluations, and who might need these pedagogical interventions to keep them on board. In the near future, one option to explore would be to give teachers feedback about which type of student might have said X, Y, or Z, and which types of students commented mainly on A or B, thereby providing more pedagogically informed options for teachers to meet the diverse needs of their learners. At the same time this may lead to anonymisation and ethical issues, which need careful scrutiny.

Acknowledgement

This paper is in part based upon two Data Wrangler reports (Clow et al., 2019; Ullmann et al., 2018) that were published under a creative common license at the OU. We are extremely grateful for all the OU staff who have helped to support the building of the dictionary approaches, as well as their kind suggestions which data could be useful to explore. Many thanks to the Text Analytics of Student Comments (TASC) Initiative in IET for its support of this work.

Biography

Dr Thomas Ullmann is a Lecturer in the Institute of Educational Technology at The Open University and an Honorary Lecturer in the Department of Computer Science at the University

of Warwick. With more than 15 years of experience, he is an expert in the evaluation of Educational Technology in Higher Education. His current research interests include the evaluation of online and distance teaching and learning, educational text analytics, and the automated analysis of reflective thinking in writings. Thomas is the lead of the Text Analytics of Students Initiative delivering text analytics to the whole of the University.

Thomas published over 30 publications in high-ranked journals, conferences, and workshops and over 45 research deliverables, internal research reports, quality enhancement reports, and scholarship work. Thomas co-organised the Workshop series on Awareness and Reflection in Technology-Enhanced Learning held in conjunction with the European Conference on Technology Enhanced Learning. Thomas regularly presents his work at conferences.

Dr. Bart Rienties is Professor of Learning Analytics and head of Academic Professional Development at the Institute of Educational Technology at the Open University UK. As Associate Director he leads a group of academics who provide university-wide academic professional development and innovation courses and conduct evidence-based research of how professionals learn. As educational psychologist, he conducts multi-disciplinary research on work-based and collaborative learning environments and focuses on the role of social interaction in learning, which is published in leading academic journals and books. His primary research interests are focussed on Learning Analytics, Professional Development, and the role of motivation in learning. Furthermore, Bart is interested in broader internationalisation aspects of higher education.

References

Arbaugh, J. B. (2014). System, scholar, or students? Which most influences online MBA course effectiveness? *Journal of Computer Assisted Learning, 30*(4), 349-362. doi: 10.1111/jcal.12048

Ashby, A., Richardson, J. T. E., & Woodley, A. (2011). National student feedback surveys in distance education: an investigation at the UK Open University. *Open Learning: The Journal of Open, Distance and e-Learning, 26*(1), 5-25. doi: 10.1080/02680513.2011.538560

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*.

Borrego, M., & Newswander, L. K. (2010). Definitions of interdisciplinary research: Toward graduate-level interdisciplinary learning outcomes. *The Review of Higher Education, 34*(1), 61-84. doi: 10.1353/rhe.2010.0006

Clow, D., Coughlan, T., Cross, S., Edwards, C., Gaved, M., Herodotou, C., . . . Ullmann, T. (2019). Scholarly insight Winter 2019: a Data wrangler perspective. Milton Keynes: Open University UK.

Coughlan, T., Ullmann, T. D., & Lister, K. (2017). *Understanding Accessibility as a Process through the Analysis of Feedback from Disabled Students*. Paper presented at the W4A'17 International Web for All Conference, New York. http://oro.open.ac.uk/48991/

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education, 29*(5), 611-623. doi: 10.1080/02602930410001689171

Gamliel, E., & Davidovitz, L. (2005). Online versus traditional teaching evaluation: mode can matter. *Assessment & Evaluation in Higher Education, 30*(6), 581-592. doi: 10.1080/02602930500260647

Grebennikov, L., & Shah, M. (2013). Student voice: using qualitative feedback from students to enhance their university experience. *Teaching in Higher Education, 18*(6), 606-618. doi: 10.1080/13562517.2013.774353

HEFCE. (2016). Review of information about learning and teaching and the student experience. Results and analysis of for the 2016 pilot of the National Student Survey. London: HEFCE.

Jeonghee, Y., Nasukawa, T., Bunescu, R., & Niblack, W. (2003, 19-22 Nov. 2003). *Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques.* Paper presented at the Third IEEE International Conference on Data Mining.

Kember, D., & Ginns, P. (2012). *Evaluating teaching and learning*. New York: Routledge.

Langan, A. M., & Harris, W. E. (2019). National student survey metrics: where is the room for improvement? *Higher Education, 78*(6), 1075-1089. doi: 10.1007/s10734-019-00389-1

Leong, C. K., Lee, Y. H., & Mak, W. K. (2012). Mining sentiments in SMS texts for teaching evaluation. *Expert Systems with Applications, 39*(3), 2584-2589. doi: 10.1016/j.eswa.2011.08.113

Li, N., Marsh, V., & Rienties, B. (2016). Modeling and managing learner satisfaction: use of learner feedback to enhance blended and online learning experience. *Decision Sciences Journal of Innovative Education, 14*(2), 216-242. doi: 10.1111/dsji.12096

Li, N., Marsh, V., Rienties, B., & Whitelock, D. (2017). Online learning experiences of new versus continuing learners: a large scale replication study. *Assessment & Evaluation in Higher Education, 42*(4), 657-672. doi: 10.1080/02602938.2016.1176989

McDonald, J., Moskal, A. C. M., Goodchild, A., Stein, S., & Terry, S. (2020). Advancing text-analysis to tap into the student voice: a proof-of-concept study. *Assessment & Evaluation in Higher Education, 45*(1), 154-164. doi: 10.1080/02602938.2019.1614524

Moskal, A. C. M., Stein, S. J., & Golding, C. (2015). Can you increase teacher engagement with evaluation simply by improving the evaluation system? *Assessment & Evaluation in Higher Education, 41*(2), 286-300. doi: 10.1080/02602938.2015.1007838

Nguyen, Q., Rienties, B., Toetenel, L., Ferguson, F., & Whitelock, D. (2017). Examining the designs of computer-based assessment and its impact on student engagement, satisfaction, and pass rates. *Computers in Human Behavior, 76*(November 2017), 703-714. doi: 10.1016/j.chb.2017.03.028

Open University UK. (2014). Ethical use of Student Data for Learning Analytics Policy. Retrieved 23 June 2016, 2016, from http://www.open.ac.uk/students/charter/essential-documents/ethical-use-student-data-learning-analytics-policy

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics, 13*(4), 519-549. doi: 10.1075/ijcl.13.4.06ray

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education, 30*(4), 387-415. doi: 10.1080/02602930500099193

Richardson, J. T. E. (2006). Investigating the relationship between variations in students' perceptions of their academic environment and variations in study behaviour in distance education. *British Journal of Educational Psychology, 76*(4), 867-893. doi: 10.1348/000709905X69690

Richardson, J. T. E. (2013). The National Student Survey and its Impact on UK Higher Education. In M. Shah & C. S. Nair (Eds.), *Enhancing Student Feedback and Improvement Systems in Tertiary Education* (Vol. 5, pp. 76–84). Abu Dhabi, UAE: Commission for Academic Accreditation.

Richardson, J. T. E., Mittelmeier, J., & Rienties, B. (2020). The role of gender, social class and ethnicity in participation and academic attainment in UK higher education: an update. *Oxford Review of Education, 46*(3), 346-362. doi: 10.1080/03054985.2019.1702012

Richardson, J. T. E., Slater, J. B., & Wilson, J. (2007). The National Student Survey: development, findings and implications. *Studies in Higher Education, 32*(5), 557-580. doi: 10.1080/03075070701573757

Rienties, B. (2014). Understanding academics' resistance towards (online) student evaluation. *Assessment & Evaluation in Higher Education, 39*(8), 987-1001. doi: 10.1080/02602938.2014.880777

Rienties, B., Boroowa, A., Cross, S., Kubiak, C., Mayles, K., & Murphy, S. (2016). Analytics4Action Evaluation Framework: a review of evidence-based learning analytics interventions at Open University UK. *Journal of Interactive Media in Education, 1*(2), 1-12. doi: 10.5334/jime.394

Rienties, B., & Héliot, Y. (2018). Enhancing (in)formal learning ties in interdisciplinary management courses: a quasi-experimental social network study. *Studies in Higher Education, 43*(3), 437-451. doi: 10.1080/03075079.2016.1174986

Rienties, B., & Toetenel, L. (2016). The impact of learning design on student behaviour, satisfaction and performance: a cross-institutional comparison across 151 modules. *Computers in Human Behavior, 60*, 333-341. doi: 10.1016/j.chb.2016.02.074

Shah, M. (2019). Making the student voice count: using qualitative student feedback to enhance the student experience. *Journal of Applied Research in Higher Education, ahead-of-print*(ahead-of-print). doi: 10.1108/JARHE-02-2019-0030

Shah, M., Nair, C. S., & Richardson, J. T. E. (2017). Chapter 8 - Accessing Student Voice: Using Qualitative Student Feedback. In M. Shah, C. S. Nair, & J. T. E. Richardson (Eds.), *Measuring and Enhancing the Student Experience* (pp. 91-101): Chandos Publishing.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). *Recursive deep models for semantic compositionality over a sentiment treebank.* Paper presented at the Proceedings of the 2013 conference on empirical methods in natural language processing, Seattle, Washington.

Ullmann, T. (2015a). *Automated detection of reflection in texts. A machine learning based approach.* (PhD), Open University UK, Milton Keynes. Retrieved from http://oro.open.ac.uk/45402/

Ullmann, T. (2015b). *Keywords of written reflection - a comparison between reflective and descriptive datasets*. Paper presented at the Proceedings of the 5th Workshop on Awareness and Reflection in Technology Enhanced Learning, Toledo, Spain. http://ceur-ws.org/Vol-1465/paper8.pdf

Ullmann, T. (2017). *Reflective Writing Analytics - Empirically Determined Keywords of Written Reflection*. Paper presented at the Seventh International Learning Analytics & Knowledge Conference, Vancouver, Canada. http://oro.open.ac.uk/48840/

Ullmann, T., Lay, S., Cross, S., Edwards, C., Gaved, M., Jones, E., . . . Rienties, B. (2018). Scholarly insight Spring 2018: a Data wrangler perspective. Milton Keynes: Open University.

Ullmann, T., Wild, F., & Scott, P. (2012). *Comparing automatically detected reflective texts with human judgements.* Paper presented at the 2nd Workshop on Awareness and Reflection in Technology Enhanced Learning Saarbrucken, Germany.

Wen, M., Yang, D., & Rosé, C. P. (2014). *Sentiment Analysis in MOOC Discussion Forums: What does it tell us.* Paper presented at the 7th Educational Data Mining Conference.

Yin, R. K. (2009). *Case study research: Design and methods* (5 ed.). Thousand Oaks: Sage.

Zaitseva, E., Milsom, C., & Stewart, M. (2013). Connecting the dots: using concept maps for interpreting student satisfaction. *Quality in Higher Education, 19*(2), 225-247. doi: 10.1080/13538322.2013.802576