ASSESSMENT & EVALUATION IN HIGHER EDUCATION
2023, VOL. 48, NO. 3, 377–389
https://doi.org/10.1080/02602938.2022.2081668

First, do no harm: automated detection of abusive comments in student evaluation of teaching surveys

Samuel Cunningham, Melinda Laundon, Abby Cathcart, Md Abul Bashar and Richi Nayak

Queensland University of Technology, Brisbane, Queensland, Australia

ABSTRACT

Student evaluation of teaching (SET) surveys are the most widely used tool for collecting higher education student feedback to inform academic quality improvement, promotion and recruitment processes. Malicious and abusive student comments in SET surveys have the potential to harm the wellbeing and career prospects of academics. Despite much literature highlighting abusive feedback in SET surveys, little research attention has been given to methods for screening student comments to identify and remove those that may cause harm to academics. This project applied innovative machine learning techniques, along with a dictionary of keywords to screen more than 100,000 student comments made via a university SET during 2021. The study concluded that these methods, when used in conjunction with a final stage of human checking, are an effective and practicable means of screening student comments. Higher education institutions have an obligation to balance the rights of students to provide feedback on their learning experience with a duty to protect academics from harm by pre-screening student comments before releasing SET results to academics.

KEYWORDS
Student evaluation of teaching; text analysis; abusive comments; course evaluation; machine learning

Introduction

Student evaluation of teaching (SET) surveys, though contested, are the most widely used tool to obtain higher education students' feedback on learning, teaching and the student experience (Spooren, Brockx, and Mortelmans 2013; Heffernan 2022). SET surveys provide valuable opportunities for student voice, information for educators and higher education institutions to inform improvements to teaching, curriculum and quality assurance, and data for academic recruitment, promotion and performance management processes (Alderman, Towers, and Bannah 2012). In the Australian context, SET surveys also ensure that universities meet the requirements of the *Higher Education Standards Framework (Threshold Standards)* (2021), which mandates that all students must be given regular opportunities to provide feedback on their educational experience. Yet tensions remain between the value of student voice, educator and institution access to feedback, and the risk of harm to educators who receive abusive or malicious comments.

SET surveys often consist of a combination of Likert scales and open-text questions. Quantitative data from the Likert-scale questions are easy to analyse, aggregate and report, which is why they are more often subject to analysis than the free-text comments provided by

CONTACT Samuel Cunningham ✉ sam.cunningham@qut.edu.au

students. However, this focus on quantitative SET data means that the importance of students' qualitative, free-text feedback can be overlooked. Student comments can be lost or not considered when only aggregated quantitative data are analysed and reported to educators and administrators. There has been a recent debate about the ethical obligations of universities to better analyse and consider students' free-text comments. Given that universities actively seek feedback through SET surveys, students have a reasonable expectation that their comments will be read and acted on by relevant people (Santhanam et al. 2021, 3).

Many students carefully consider the comments they provide through SET surveys and try to provide constructive feedback to help educators to improve the student experience (Brockx, Van Roy, and Mortelmans 2012). These free-text comments, when written in a considered and constructive way, can provide feedback for educators and administrators to consider and potentially implement future improvements. Studies have shown that free-text comments have more influence on educators' reflection on and development of improvements to their teaching than quantitative student survey data (Stupans, McGuren, and Babey 2016; Nawaz et al. 2022). Recent research has found that a qualitative analysis of SET comments can produce different findings than a quantitative analysis of the same dataset (Gelber et al. 2022). There is also evidence that university leaders make better strategic decisions when they consider qualitative student feedback as part of quality assurance processes (Grebennikov and Shah 2013).

Although there is clearly a value in student feedback that is considered and constructive, in recent years there has been growing concern about the prevalence of abusive comments in open-text SET surveys. Analysis of 30,684 SET comments from Curtin University in 2014 concluded that most students do not abuse the privilege of giving anonymous feedback (Tucker 2014), and less than 1% of comments were deemed to be abusive or unprofessional. While the proportion of abusive comments may be small, receiving an abusive comment makes some educators less likely to read or engage with subsequent student feedback (Cunningham-Nelson, Laundon, and Cathcart 2021). Abusive comments also have the potential to cause great harm to academics' mental health and career prospects. For example, a scoping review on the causes of occupational stress among academics in Australia and New Zealand identified the increased scrutiny of SET results by management and an absence of respect in student feedback as key stressors (Lee et al. 2022). Researchers have also noted that abuse is mostly directed at women or educators from marginalised groups (Chávez and Mitchell 2019).

The vast majority of SET surveys enable students to provide feedback that is anonymous. Researchers have found that while there can be key benefits to anonymity, including giving participants the confidence to provide feedback, anonymity can also have a disinhibiting effect on bad behaviour (Joinson 2001), and there can be tensions between anonymity and socially responsible behaviour (Guo and Yu 2020). A recent study of anonymous abusive comments received by educators from an Australian university categorised examples of what they termed non-constructive SET comments. Their typology identified five themes: '(i) allegations; (ii) insults; (iii) comments about appearance, attire, and accent; (iv) projections and blame; and (v) threats and punishment' (Lakeman et al. 2021, 5). The authors note that these types of comments contribute to occupational stress, impact wellbeing and, in some cases, may be considered libellous (Lakeman et al. 2021). Links have also been made between anonymous student comments and cyber aggressive behaviours including insults, verbal taunts and malicious teasing and humiliation (Page et al. 2021). Others have highlighted the discriminatory nature of SET comments and concluded that universities that continue to rely on them to evaluate teachers and courses are actively harming marginalised groups, including women and people from minority ethnic groups (Heffernan, 2022).

Since SET results are often released at the busiest time of the teaching period where educators are devoting time to marking and student support, these comments can be especially damaging. Detecting and removing unacceptable comments before they are read by staff members will help reduce this negative impact on educators. It may be argued that institutions who undertake SETs have a duty of care to check for abusive comments before releasing results to staff. However, few institutions scan for and remove abusive comments. Heffernan (2022) found that 21% of academics surveyed said that their institution filtered or censored comments before release. Although there is a lack of published evidence about why institutions do not screen all SET comments, private correspondence with evaluations managers from several Australian universities suggests that resource constraints and technology limitations are key factors.

Universities have attempted to address this issue by developing policies that seek to remind students of their obligations to refrain from abusive or discriminatory behaviour. These codes of conduct are often framed around behaviour on campus and are usually disconnected from SET protocols, relying on students to voluntarily regulate their behaviour or for the behaviour to be reported by others for action to be taken. Other institutions have a profanity blocker built into their survey tool, although these often fail to keep up with creative and innovative terms of abuse and emerging norms for online speech (Song et al. 2022).

There is a growing concern about the wellbeing of both staff and students from both psychosocial and occupational health and safety perspectives (Wray and Kinman 2020; Dodd et al. 2021; Heffernan 2022). SETs can be seen as both a potential cause of stress for staff, a vehicle for harm and a dataset that may be used to identify students who are at risk.

This paper seeks to explore ways for the benefits of using free-text comments to be realised without compromising academics' wellbeing. There is widespread evidence that academics have reason to be wary of open-text SET comments, which can be abusive and biased. SET comments often remain on educators' permanent records, are likely to be seen by their supervisors or colleagues and may impair their promotion and recruitment prospects and influence performance reviews. By developing a method using both machine learning and a dictionary approach to screen for unacceptable student comments, we are moving closer to an institutional framework that can protect educators from these consequences.

Background and context

As student evaluation surveys are widely used, an automatic method of detecting comments which are likely to be unacceptable would be of wide benefit in the higher education sector. The Australian National Tertiary Education Union (NTEU) released a report in 2018 stating that six out of 10 respondents had experienced disrespectful or abusive comments through student evaluation surveys. Staff members who are part of minority groups are also more likely to receive abusive comments, with women, non-binary and Aboriginal and Torres Strait Islander people all more likely to receive unacceptable comments (Heffernan 2022). Unacceptable student comments have been identified as an under-researched topic and a key area of concern for academic educators. Research also indicates that some academics do not even read student survey comments because of their fear that they will encounter abusive or unacceptable comments. This prevents them from engaging with and acting on constructive student comments (Cunningham-Nelson, Laundon, and Cathcart 2021).

Ideally, higher education institutions would resource evaluations staff members to read all student comments before they are reported to academics or used in performance and quality management processes. However, this manual process is too time-consuming, not practical and not scalable. It would also slow down the process of educators having access to the feedback provided by students, potentially hindering improvements to units or teaching. The majority of students do not provide offensive comments (Tucker 2014), and this is often used as a rationale for institutions to have limited or no screening. It also makes the task of identifying the small number of unacceptable comments even more difficult.

A recent study highlights that developments in machine learning now enable the automatic detection of misogynistic tweets on Twitter (Bashar, Nayak, and Suzor 2020). This methodology has been successful in identifying these comments in short tweets written online through Twitter. This project applied similar principles to the usually short free-text comments written by students in SET surveys. Analysis of SET comments across the previous three years at Queensland University of Technology (QUT) showed that the average length of these comments was 235 characters. This is comparable with Twitter's maximum character limit of 280 characters. Research shows a clear benefit of text analytics and natural language processing on analysing the free-text comments written by students (Cunningham-Nelson, Baktashmotlagh, and Boles 2019).

An edited volume on analysing student feedback in higher education (Zaitseva, Santhanam, and Tucker 2021) noted that there remains limited published work using systematic analysis of free-text comments in evaluation data. Early studies focussed on manual coding and analysis techniques (Alhija and Fresko 2009) and keyword analysis (Scott 2005; Grebennikov and Shah 2013). More recent studies have examined the use of software, including Leximancer (Shah and Pabel 2019) and WordStat (Santhanam et al. 2021). Prior studies focus on extracting useful insights from free-text comments rather than on detecting unacceptable comments. There is scant published work in which unacceptable or abusive comments are detected in relation to student evaluations, with Hum, Wuetherick, and Jang (2021) briefly addressing how their approach to text analysis of SET surveys revealed some 'critical issues that merited or required immediate intervention' (p. 186). Santhanam et al. (2021, 169) concluded that while there is growing interest in text analysis of qualitative SET data and agreement on its value for quality improvement, many of the approaches are resource-intensive, and there is a lack of consideration of how these can feasibly be integrated into institutional reporting and quality assurance processes.

The current research was undertaken at QUT, a large Australian metropolitan-based university with more than 50,000 students. In 2020 QUT undertook a wide-ranging review of its evaluation strategy. The purpose of the review was to promote staff and student wellbeing, enhance reliability and validity, and improve staff and student confidence in the integrity of survey results by:

- identifying responses that raise (immediate) concerns about the risk of harm to self or others
- identifying unacceptable comments
- identifying unambiguous cases of misread scale
- identifying erroneous comments about design, teaching or curricula
- educating students in professional communication.

The new strategy was implemented in 2021, using the Qualtrics Survey and Reporting Tool as the technology platform underpinning a redeveloped SET. QUT's SET survey, the Student Voice Survey, consists of two components. In the first portion of the survey, students are presented with five Likert-scale questions about the subject/unit and two open-text questions: 'What aspects of this unit were done well?' and 'What aspects of this unit could be done better?' The second component of the survey allows the students to provide feedback to each educator who taught them in the subject (including, for example, the coordinator, lecturer, tutor and/or demonstrator). They are presented with one open-text question and one Likert question for feedback.

A key element of the revised evaluation strategy was to develop protocols that sought to protect staff and student interests and balance the right of students to provide feedback with the responsibilities of the university to protect staff from the risk of harm. As part of this process, a definition of unacceptable comments was determined by the University, following extensive consultation with staff and students. QUT Evaluation Protocols (2021) define an unacceptable comment as including:

a. Harassing, threatening or abusive language, including profanities and language that is intimidating or discriminatory. Discriminatory comments include comments relating to sexual orientation, gender identity, intersex status, disability, ethnicity, marital status, nationality, age, religion and/or political persuasion.
b. Personal attacks relating to appearance or other matters unconnected to units, teaching or the learning experience.

This research aimed to explore how a combined dictionary and innovative machine learning approach might assist in detecting unacceptable comments. Human research ethics approval for this research (Approval No. 2000000997) was obtained through QUT.

Methodology

This project utilised more than 100,000 student comments given in QUT's Student Voice Survey during 2021. All the comments were run through a combined dictionary-based and machine learning approach. The dictionary approach was deployed throughout the duration of the survey, and the machine learning approach was utilised on completion of the survey. For both components, it is important that they can detect all potentially unacceptable comments. While this can result in falsely mislabelling some comments as unacceptable, a higher rate of false positives is preferred to false negatives to minimise the risk of missed abusive comments.

Both methods were used in a supervised way, meaning that any comment identified as potentially unacceptable was checked by a person to verify whether the decision made by the dictionary-based or machine learning approach was correct. All comments flagged were manually checked by an evaluations staff member to determine whether they met the QUT definition of an unacceptable comment or raised concerns about a risk of harm to the student or other people (although this article focuses on the results pertaining to unacceptable comments). QUT's survey system (Qualtrics) also contains an in-built sensitive data policy that alerts students while they are typing if their comment contains a common profanity. Students can continue to submit the response if they wish. The Qualtrics profanity list contains fewer terms than the dictionary that was used in this project to scan SET comments after they were submitted.

Following the publication of highly publicised research on abusive comments in SET surveys, the research team added another step to the project, whereby both the dictionary approach and the machine learning approach were applied to the abusive student comments published in research by Lakeman et al. (2021) and Düddul (2021). These comments are shocking in nature and have been widely condemned as examples of the risk of harm that abusive student comments pose to academics. This provided an opportunity to test whether the proposed approach would identify these comments from other universities as being unacceptable.

The process presented in Figure 1 begins with the launch of the Student Voice Survey. The dictionary approach is undertaken during the live survey period. Comments are automatically compared against the existing dictionary, and any comments containing words from the dictionary are then flagged automatically for manual review by a staff member. If the staff member deems the comment as unacceptable, actions such as removing or redacting parts of the comment are taken. Once the survey has closed, all comments are run through a developed machine learning algorithm. Comments which are identified by the machine learning algorithm as potentially unacceptable are also manually checked by a staff member, and action is taken on those verified as unacceptable. The remaining comments are then released to educators in the institution. It is important to note that the remaining comments may contain a small number of unacceptable comments if they were not identified by the above process. Once receiving their comments, educators then have the opportunity to identify those that they believe are unacceptable and request a review by an evaluations staff member.

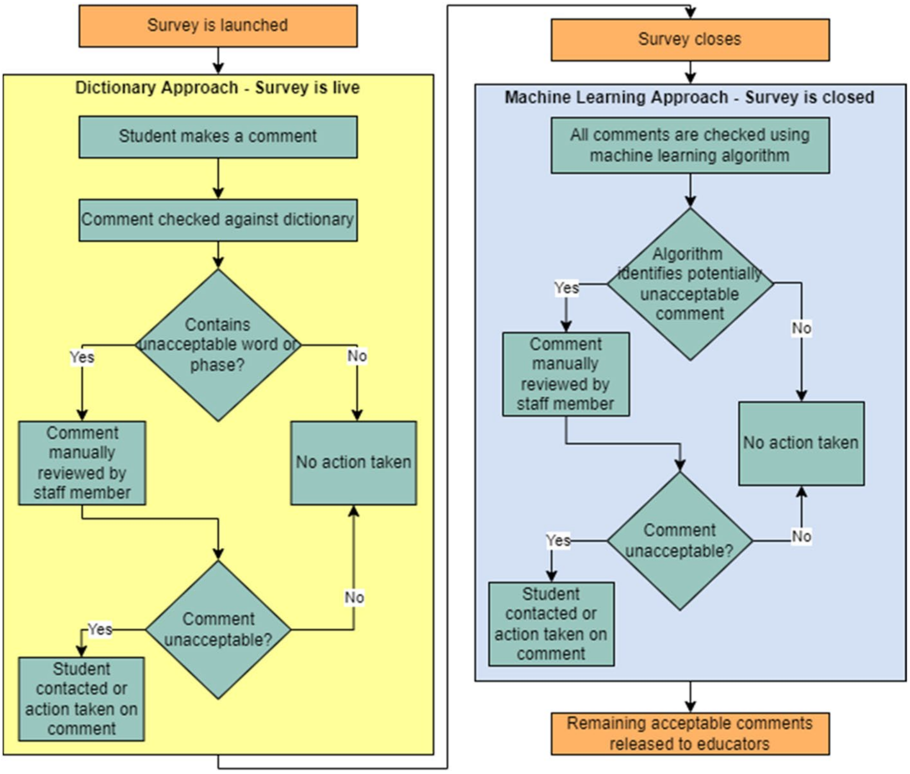
Survey is launched
Dictionary Approach - Survey is live
Student makes a comment
Comment checked against dictionary
Contains unacceptable word or phase?
Yes
No
Comment manually reviewed by staff member
No action taken
Comment unacceptable?
Yes
No
Student contacted or action taken on comment
Survey closes
Machine Learning Approach - Survey is closed
All comments are checked using machine learning algorithm
Algorithm identifies potentially unacceptable comment
Yes
No
Comment manually reviewed by staff member
No action taken
Comment unacceptable?
Yes
No
Student contacted or action taken on comment
Remaining acceptable comments released to educators

Figure 1. The process used for detecting unacceptable comments.

Part 1: Live screening of comments—dictionary approach

The first part of the methodology used to identify unacceptable comments was a dictionary approach. The dictionary approach involves matching individual words or phrases to an existing list or dictionary of words. The most critical element of the dictionary approach is this list of terms. This list of terms includes a wide range from multiple sources. It is important to include specific terms in this list that need to be detected but balance this with being cautious about selecting words that can commonly appear. A larger dictionary will, of course, identify more potentially unacceptable comments, but it also means that more comments will need to be manually reviewed, requiring additional staff resources. Other aspects that must be considered include words of different forms such as ‘run’ or ‘running’. The character ‘*’ is used to represent a wildcard, so ‘run*’ would capture both terms. Quotation marks can also be used in the dictionary to ensure only exact terms are matched.

The process for developing a dictionary was iterative. One key source for the dictionary terms was existing lists of profanities. Many of these lists are available online and often used as part of surveys in general (Sood, Antin, and Churchill 2012). In creating the dictionary for this research, several available online lists were utilised. The created dictionary also included a set of words relating to the potential risk of harm to self and others. These terms were again sourced from a range of online sources, as well as existing survey software lists.

The profanity and risk-of-harm terms formed the basis of a general dictionary; however, it was important that the dictionary created was both context-specific for student evaluations and the institution’s local context. The QUT definition of unacceptable comments was used to support these context-specific terms. Previous comments identified as being unacceptable were used as a basis for identifying some of these terms, as well as open public websites such as

Rate My Professors (ratemyprofessors.com). New terms were added to the dictionary as they were identified, in recognition that language constantly evolves. Terms not relating to learning and teaching were added at this stage, such as comments on physical appearance.

The dictionary approach was used to scan for unacceptable comments while the survey was active. This 'live' scanning was important if, for example, a staff member is threatened in a written student comment, because relevant action can be taken swiftly to avoid potential harm and to support the educator. Second, the regular checking of unacceptable comments enables the opportunity to prompt students to edit a comment before the survey closes. This mitigates the negative outcomes of an educator reading an unacceptable comment while also providing the opportunity for students to learn about professional conduct. As stated earlier, QUT uses Qualtrics to deploy its SET survey. Qualtrics' TextIQ feature supports using dictionaries to identify these comments in a live manner. Thus, this first part of the process was able to detect and address a number of unacceptable comments before the survey closed.

Part 2: Post-screening of comments—machine learning approach

The machine learning model was used to screen all given student comments once the survey closed. This machine learning model is based on an architecture known as RoBERTa (Robustly Optimized BERT Pretraining Approach). The key benefit that the RoBERTa model presents is that it can be pre-trained on large amounts of text comments in other contexts and then adjusted to suit a specific domain (Liu et al. 2019). Figure 2 shows the model used, which consists of two key components: the pre-training stage and training stage.

The pre-trained portion involved using three datasets, including Wikipedia articles (D1), IMDb movie reviews (D2) and QUT student comments (D3). The Wikipedia dataset contains over 160GB of text data and was designed to give the model an understanding of the English language as a whole. The IMDb movie reviews were used to further improve the model, as these reviews are often of a short nature and include personal opinions (Maas et al. 2011). The pre-trained model was further fine-tuned with QUT student comments to understand the language patterns used in these surveys, however the comments were not identified as being acceptable or not. The QUT student dataset (D3) contains a total of 268 K comments, out of a total of 335,237 QUT student comments from 2018 to 2021. The remaining unseen comments were used to evaluate the model for its accuracy performance.

The regularising process then allowed the model to be applied to a specific context; in this case, QUT student comments were categorised as either acceptable or unacceptable based on the QUT definition of unacceptable comments. In the training stage, a small data set containing 522 unacceptable and an equal amount of acceptable comments were used to train the model. As unacceptable comments are not a frequent occurrence, only 652 examples of unacceptable

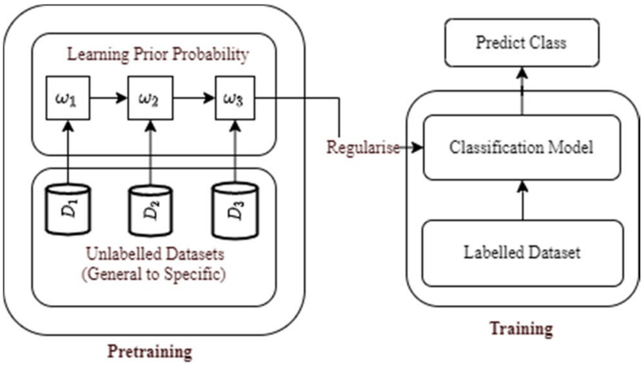

Figure 2. Machine learning model.

comments were presented in the complete dataset. This signified the importance of the pre-training step, as machine learning models often require hundreds of thousands of examples to form useful predictions.

Results and discussion

The dictionary and machine learning approaches complemented each other and served different purposes, identifying different sets of comments. The methodology presented was applied to comments in QUT's Student Voice Survey across 2021. Table 1 shows the number of comments in total and the potentially unacceptable comments identified by each method.

Across the subject and teacher components of the Student Voice Survey, a total of 105,483 comments were screened using this automated process. With this process, only 7.5% of the comments had to be reviewed manually across 2021. With adjustments made to both the dictionary and machine learning approach at the end of Semester 1, fewer (6.2%) comments were required to be manually reviewed in Semester 2. Manual review by evaluations staff members subsequently identified 100 comments as meeting the QUT definition of unacceptable. This means that 100 abusive or risk-of-harm comments were removed before SET results were released to educators and their supervisors.

The dictionary approach provided a live mechanism for responding to comments made by students throughout the survey periods. It also created a mechanism to educate students by inviting them to revise comments before the survey closed. Students were notified that their comments had been flagged via a screening process, reminded of their obligations under the QUT Student Code of Conduct (QUT, 2019) and given an opportunity to revise unacceptable comments before the survey closed. Across 2021, 80 students were given the opportunity to revise their comments. Approximately 50% of students took up this opportunity.

A review of the dictionary terms was completed at the end of Semester 1, with some phrases being added that were identified as missing from the dictionary and some phrases removed that were determined to be not useful. The dictionary needs regular updating as language changes over time. It is also clear from the results presented in Table 1 that changes in the dictionary can reduce the number of false positives. Semester two shows the potential number of unacceptable comments reduced to half of the original amount, while the number of unacceptable comments identified increased.

The machine learning method complemented the dictionary approach by identifying comments through an algorithmic approach. This allowed the identification of comments to go beyond matching a list of existing words to identify words used in a combination or context that may constitute an unacceptable comment. This means that a wider range of comments was detected. As the machine learning component is currently more time-consuming than the dictionary approach, it is conducted after the survey closes and therefore does not allow for students to be given the opportunity to amend their comments during the survey period. However, the importance of the machine learning approach is clear, with an additional 10 comments in Semester 1 and 42 comments in Semester two being actioned before results were released to staff. This signifies the importance of both parts of the process.

When the survey feedback is released to educators, they are able to request removal of unacceptable comments if they believe they were missed by the screening. Such requests are considered

Table 1. Comparison between dictionary and machine learning approaches.

	Total number of comments screened	Potentially unacceptable comments identified by dictionary approach	Potentially unacceptable comments identified by machine learning approach	Number of comments manually deemed unacceptable
Sem-1	53,017	2,654	2,028	36
Sem-2	52,466	1,292	1,938	64

in line with the definitions within the published Evaluation protocols. In 2021, with the automated screening process implemented, a total of 15 removal requests were made, with three being determined to fit the QUT definition of unacceptable, and thus removed. This is significantly reduced, compared to the 112 requests and 80 removals the year prior. Each unacceptable comment received by an educator represents potential harm caused, so this had a clear impact on educators.

This process highlighted the importance of always having a human in the loop to make the final judgement of whether a comment was acceptable or not. The human is able to review the comment in the particular university context and in relation to the university's unacceptable comment definition.

Testing the approach

To further test the approach's reliability, abusive comments published in recent published articles were screened using both the dictionary and machine learning approaches. The comments and the results of each approach are included in Table 2. Testing the approaches against these comments presented two opportunities. The first was to test how well the current processes were working, and second, missed terms could be added to the dictionary and example comments could be used to improve the training of the machine learning model. This meant that both methods would be improved for future detection.

From the 17 comments presented in Table 2, 11 were identified by the dictionary approach, and all 17 were identified by the machine learning approach. Regardless, there is a need for the continual updating of these processes, such as adding new words to the dictionary and providing these comments as additional training data for the machine learning algorithm. It is important the dictionary approach remains, as there are certain words or phrases that are clearly unacceptable, but may not have been previously used in student feedback, and therefore will not be part of the machine learning training process.

Limitations and future research considerations

Language is fast-moving, the meaning of words can change, and novel words are often added to everyday vocabulary. Language can also have different meanings when read by different audiences and backgrounds. For the dictionary approach, this is especially important, and the dictionary must be regularly reviewed and updated.

The machine learning approach learns from previous examples of unacceptable comments. These comments are in the minority of SET comments, so it can be difficult to identify the substantial number of sample phrases often required to train machine learning models. This creates an opportunity for the higher education sector to come together in sharing data to train the model. It also emphasises the symbiotic nature of the dictionary and machine learning approaches working together, particularly considering the machine learning model requires additional training.

Making the decision of what is or is not an unacceptable comment is also challenging and requires human intervention. There is a spectrum of comments that may be critical, harsh or negative, and comments that are abusive, discriminatory and unacceptable. It is important to maintain a consistent approach to avoid some of this subjectivity and to promote staff confidence in the methods and outcomes. Some of these comments also contain allegations that are unverifiable. The importance of context and educator characteristics in establishing whether comments are unacceptable means that it is essential for educators to have avenues to request the review and removal of comments that they consider unacceptable.

This research highlights several critical areas of inquiry requiring further research. The most pressing need identified is in relation to comments raising concerns about a risk of harm to the student, staff members or other students. Detecting comments of concern was also an

Table 2. Unacceptable comments presented in recent publications.

Comment in the literature and source	Identified by dictionary approach	Identified by machine learning approach
You look like something the cat dragged in. (Lakeman et al. 2021)	No	Yes
People who's [sic] mother tongue is not English should not be employed as lecturers. (Lakeman et al. 2021)	No	Yes
She is really rude which is why everyone hates her. (Lakeman et al. 2021)	No	Yes
You are a cultural Marxist, your Wokeness undermines everything you do. Not all your students are left wing nut jobs like you. You seriously need to lose some weight. (Lakeman et al. 2021)	No	Yes
What the fuck did you think you were doing to take a couple of days off for your grandmother's funeral when we had an assignment due? (Lakeman et al. 2021)	Yes	Yes
That fucking dyke bitch failed me she's fucking useless that's why I failed. (Lakeman et al. 2021)	Yes	Yes
I'd like to shove a broom up her arse. (Lakeman et al. 2021)	Yes	Yes
She should be stabbed with a pitchfork. (Düddul 2021)	Yes	Yes
If I was X, I would jump off the tallest building and kill myself if I was that dumb. (Düddul 2021)	Yes	Yes
Stupid old hag needs a good fucking. (Düddul 2021)	Yes	Yes
This bitch should be fired immediately. Why is someone this ugly allowed to teach? She better be careful I never see her in the car park. She needs to get a better fashion pick. Her clothes are hideous. (Düddul 2021)	Yes	Yes
I couldn't concentrate because I couldn't tell if the teacher was a man or woman. (Düddul 2021)	Yes	Yes
This lecturer has no empathy for students not supporting the LGBTQ ideology. (Düddul 2021)	No	Yes
She looks like a man professor not a woman one (Düddul 2021)	Yes	Yes
He made me uncomfortable because gays and lesbianism are against my religion (Düddul 2021)	Yes	Yes
There are only two genders, men and women! (Düddul 2021)	Yes	Yes
You look like 13 year old boy but the brain of a woman power bullshit and your [sic] a germ. (Düddul 2021)	Yes	Yes

unanticipated outcome of Hum, Wuetherick and Jang's approach to analysing student comments (2021). While QUT and other universities do in practice screen for concerning words that may indicate (for example) suicidal ideation or violence, there is little published work on analysing student comments to detect comments raising concerns about a risk of harm to the student or other people.

Another future research avenue is the implications of live screening comments and providing students with opportunities to amend their comments. The potential educative benefits of this approach will be examined in a future paper.

While this paper investigated the effect of screening comments and preventing abusive comments reaching educators, it is too early to determine the impact of this work in building staff confidence in engaging with student feedback.

Concluding remarks

Universities have an obligation to give students the opportunity to provide feedback to educators on their learning experience. Here we have argued that universities have a comparable obligation to protect their employees from the risk of harm. As our review of the literature has highlighted, unacceptable SET comments are a problem for the sector. By providing academic staff with SET results that include abusive or discriminatory feedback, the sector is exposing them to potential harm with consequences for their wellbeing and creating a climate of distrust that discourages further engagement with student feedback.

Some universities have processes whereby educators may request the removal of unacceptable comments when they identify them; however, by that stage, harm may have already occurred. Although it is not yet possible to identify all unacceptable comments, this paper presented a two-fold approach to detecting unacceptable comments before they are shown to educators. The live dictionary approach aims to identify these comments as the survey is underway and provides students with the opportunity to reconsider their feedback and provide it in a more constructive and professional way. The machine learning method aims to identify any unacceptable comments after the survey has closed. Our results show that neither approach in isolation is sufficient to identify a comprehensive range of unacceptable comments. Language is evolving, and dictionaries will always struggle to keep up with creative forms of profanity, abuse and discrimination. Similarly, as machine learning can only be as good as the training model, there will always be comments that the university would deem to be unacceptable that have not been included in the training because they have not been identified previously. At the end of the process, there is still—and we would argue always will be—a need for human intervention. A dictionary and machine learning approach simply reduces the number of comments a human will need to read.

Some might suggest that this multi-method approach is excessive when only a small number of unacceptable comments are made by students. It should be remembered, however, that every single unacceptable comment that is removed prevents an individual from being exposed to a personal attack in their workplace. To date, this project has prevented 100 abusive comments from reaching academic staff members. Moreover, by ensuring that students know when their comments are unacceptable to the university, we are also creating an opportunity for them to reflect further on their professional conduct.

Universities have a duty of care to protect educators from harm, and the approach outlined here leverages the power of machine learning to develop a SET-screening process that goes some way to meeting that obligation.

Acknowledgements

The authors acknowledge both the contributions made by QUT students in providing feedback via the Student Voice Survey. We would also like to thank Dr Le Hoa Phan for her expert advice that informed our review of evaluation and Ashleigh Quaill, Sarah Stevens and Tina Graham for their support in implementing this screening process at QUT.

ORCID

Sam Cunningham http://orcid.org/0000-0002-3110-5281
Melinda Laundon http://orcid.org/0000-0001-7160-4942
Abby Cathcart http://orcid.org/0000-0001-7003-1273
Md Abul Bashar http://orcid.org/0000-0003-1004-4085
Richi Nayak http://orcid.org/0000-0002-9954-0159

Conflict of Interest

The authors declare no potential conflict of interest.

References

Alderman, L., S. Towers, and S. Bannah. 2012. "Student Feedback Systems in Higher Education: A Focused Literature Review and Environmental Scan." *Quality in Higher Education* 18 (3): 261–280. doi:10.1080/13538322.2012.730714.

Alhija, F. N.-A., and B. Fresko. 2009. "Student Evaluation of Instruction: What Can Be Learned from Students' Written Comments?" *Studies in Educational Evaluation* 35 (1): 37–44. doi:10.1016/j.stueduc.2009.01.002.

Bashar, M. A., R. Nayak, and N. Suzor. 2020. "Regularising LSTM Classifier by Transfer Learning for Detecting Misogynistic Tweets with Small Training Set." *Knowledge and Information Systems* 62 (10): 4029–4054. doi:10.1007/s10115-020-01481-0.

Brockx, B., K. Van Roy, and D. Mortelmans. 2012. "The Student as a Commentator: Students' Comments in Student Evaluations of Teaching." *Procedia - Social and Behavioral Sciences* 69: 1122–1133. doi:10.1016/j.sbspro.2012.12.042.

Chávez, K., and K. M. W. Mitchell. 2019. "Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity." *PS: Political Science & Politics* 53 (2): 270–274.

Cunningham-Nelson, S., M. Baktashmotlagh, and W. Boles. 2019. "Visualizing Student Opinion through Text Analysis." *IEEE Transactions on Education* 62 (4): 305–311. doi:10.1109/TE.2019.2924385.

Cunningham-Nelson, S., M. Laundon, and A. Cathcart. 2021. "Beyond Satisfaction Scores: Visualising Student Comments for Whole-Of-Course Evaluation." *Assessment & Evaluation in Higher Education* 46 (5): 685–700. doi:10.1080/02602938.2020.1805409.

Dodd, R. H., K. Dadaczynski, O. Okan, K. J. McCaffery, and K. Pickles. 2021. "Psychological Wellbeing and Academic Experience of University Students in Australia during COVID-19." *International Journal of Environmental Research and Public Health* 18 (3): 866. doi:10.3390/ijerph18030866.

Düddul, P. 2021. " Read the Student Survey Responses Shared by Academics and You'll See Why Professor Hambling is Justified in Burning Hers." *The Conversation*. https://theconversation.com/read-the-student-survey-responses-shared-by-academics-and-youll-see-why-professor-hambling-is-justified-in-burning-hers-167897

Gelber, K., K. Brennan, D. Duriesmith, and E. Fenton. 2022. "Gendered Mundanities: Gender Bias in Student Evaluations of Teaching in Political Science." *Australian Journal of Political Science*: 1–22. doi:10.1080/10361146.2022.2043241.

Grebennikov, L., and M. Shah. 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.

Guo, K. H., and X. Yu. 2020. "The Anonymous Online Self: Toward an Understanding of the Tension between Discipline and Online Anonymity." *Information Systems Journal* 30 (1): 48–69. doi:10.1111/isj.12242.

Heffernan, T. 2022. "Abusive Comments in Student Evaluations of Courses and Teaching: The Attacks Women and Marginalised Academics Endure." *Higher Education*: 1–15. doi:10.1007/s10734-022-00831-x.

Heffernan, T. 2022. "Sexism, Racism, Prejudice, and Bias: A Literature Review and Synthesis of Research Surrounding Student Evaluations of Courses and Teaching." *Assessment & Evaluation in Higher Education* 47 (1): 144–154. doi:10.1080/02602938.2021.1888075.

Higher Education Standards Framework *(Threshold Standards)*. 2021. (Cth) (Austrl.)

Hum, G. B. Wuetherick, and Y. Jang. 2021. "Supporting Practical Use and Understanding of Student Evaluations of Teaching through Text Analytics Design, Policies, and Practices." In *Analysing Student Feedback in Higher Education*, edited by E Zaitseva, B Tucker, and E Santhanam, 180–191. Routledge.

Joinson, A. 2001. "Self-Disclosure in Computer-Mediated Communication: The Role of Self-Awareness and Visual Anonymity." *European Journal of Social Psychology* 31 (2): 177–192. doi:10.1002/ejsp.36.

Lakeman, R., R. Coutts, M. Hutchinson, M. Lee, D. Massey, D. Nasrawi, and J. Fielden. 2021. "Appearance, Insults, Allegations, Blame and Threats: An Analysis of Anonymous Non-Constructive Student Evaluation of Teaching in Australia." *Assessment & Evaluation in Higher Education*: 1–14. doi:10.1080/02602938.2021.2012643.

Lee, M., R. Coutts, J. Fielden, M. Hutchinson, R. Lakeman, B. Mathisen, D. Nasrawi, and N. Phillips. 2022. "Occupational Stress in University Academics in Australia and New Zealand." *Journal of Higher Education Policy and Management* 44 (1): 57–71. doi:10.1080/1360080X.2021.1934246.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv Preprint arXiv:1907.11692*

Maas, A. L. R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. 2011. "Learning Word Vectors for Sentiment Analysis." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon.

National Tertiary Education Union 2018. *Staff Experience of Student Evaluation of Teaching and Subjects/ Units*. NTEU Policy and Research Unit. https://www.nteu.org.au/library/download/id/9058.

Nawaz, R., Q. Sun, M. Shardlow, G. Kontonatsios, N. R. Aljohani, A. Visvizi, and S.-U. Hassan. 2022. "Leveraging AI and Machine Learning for National Student Survey: Actionable Insights from Textual Feedback to Enhance Quality of Teaching and Learning in UK's Higher Education." *Applied Sciences* 12 (1): 514. doi:10.3390/app12010514.

Page, A., J. Charteris, and School of Education, University of Newcastle. 2021. "Student Evaluations of Teaching and Student Cyberaggression: The Impact of Keyboard Warriors in Tertiary Education." *Journal of Education and Humanities* 4 (1): 96–123. doi:10.14706/JEH2021415.

Queensland University of Technology. 2021. "QUT Evaluation Protocols." Unpublished internal document.

Queensland University of Technology. 2019. "QUT Student Code of Conduct." https://www.mopp.qut.edu.au/E/E_02_01.jsp

Santhanam, E. B. Lynch, J. Jones, and J. Davis. 2021. *From Anonymous Student Feedback to Impactful Strategies for Institutional Direction*. Routledge

Scott, G. 2005. "Accessing the Student Voice." In *Higher Education Innovation Program and the Collaboration and Structural Reform Fund, Department of Education, Science and Training*. Canberra: Commonwealth of Australia.

Shah, M., and A. Pabel. 2019. "Making the Student Voice Count: Using Qualitative Student Feedback to Enhance the Student Experience." *Journal of Applied Research in Higher Education* 12 (2): 194–209. doi:10.1108/JARHE-02-2019-0030.

Song, Y., Q. Lin, K. H. Kwon, C. H. Y. Choy, and R. Xu. 2022. "Contagion of Offensive Speech Online: An Interactional Analysis of Political Swearing." *Computers in Human Behavior* 127: 107046. doi:10.1016/j.chb.2021.107046.

Sood, S. J. Antin, and E. Churchill. 2012. "Profanity Use in Online Communities." CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Spooren, P., B. Brockx, and D. Mortelmans. 2013. "On the Validity of Student Evaluation of Teaching: The State of the Art." *Review of Educational Research* 83 (4): 598–642. doi:10.3102/0034654313496870.

Stupans, I., T. McGuren, and A. M. Babey. 2016. "Student Evaluation of Teaching: A Study Exploring Student Rating Instrument Free-Form Text Comments." *Innovative Higher Education* 41 (1): 33–42. doi:10.1007/s10755-015-9328-5.

Tucker, B. 2014. "Student Evaluation Surveys: Anonymous Comments That Offend or Are Unprofessional." *Higher Education* 68 (3): 347–358. doi:10.1007/s10734-014-9716-2.

Wray, S., and G. Kinman. 2020. "The Psychosocial Hazards of Academic Work: An Analysis of Trends." *Studies in Higher Education*: 1–12.

Zaitseva, E. E. Santhanam, and B. Tucker. 2021. "Discovering Student Experience: Beyond Numbers through Words." In *Analysing Student Feedback in Higher Education*, edited by E Zaitseva, B Tucker and E Santhanam, 1–16. Routledge.