Where's the harm? Screening student evaluations of teaching for offensive, threatening or distressing comments

Matthew J. Gibson, Justin Luong
School of Computer Science and Engineering, University of New South Wales, Sydney, Australia

Hanbit Cho
School of Psychology, University of New South Wales, Sydney, Australia

Bryan Moh, Simone Zanin
School of Computer Science and Engineering, University of New South Wales, Sydney, Australia

Mentari Djatmiko, R. Zach Aandahl
Pro Vice-Chancellor Education Portfolio, University of New South Wales, Sydney, Australia

Student evaluation surveys provide educational institutions with important feedback regarding the student experience of teaching and courses; however, qualitative comments can contain offensive, insulting or threatening content. Large educational institutions generate thousands of comments per academic term; therefore, manual screening processes to find potentially harmful comments are not generally feasible. We developed a methodology for semi-automated screening of student comments that incorporates a machine learning decision support system and a detailed psychological assessment protocol. In a case study at a large public Australian university, our system identified 4,258 out of 62,049 (6.9%) comments as potentially harmful and requiring further review. Feedback from stakeholders demonstrates that this methodology is useful in reducing staff workload and could be broadly applied to different settings.

Implications for practice or policy:

- Educational institutions can adopt this methodology to dramatically decrease the number of working hours required to screen harmful free-text comments.
- Researchers can use the proposed psychology-based assessment as an example of how to develop a protocol to categorise comments.
- Educators and researchers can use this case study to follow best practices to develop their own decision support system that implements free-text comment classifiers.

Keywords: offensive comments, student experience, survey screening, student evaluation, teaching

Introduction

We propose a decision support system guided by a new assessment protocol for text categorisation of concerning comments. We consider the assessment protocol as psychology-based as the intended outcome is the improvement of teacher mental well-being. For this system we combine heuristic and machine learning models to automatically code potentially concerning student experience comments into a reduced sized subset that can be manually reviewed by appropriate staff. The use of machine learning to automate classification of potentially harmful free-text comments dramatically lowers the manual screening burden on staff. We demonstrate the efficacy of the decision support system by screening two terms of free-text comments at a major Australian university over the 2019 academic year.

Student evaluation surveys are a common and important method of evaluating teaching quality and effectiveness at universities and other educational institutions (Spooren et al., 2013). These surveys are typically designed to elicit student feedback on how courses can be improved and how lecturers or tutors could further develop their teaching skill. However, these instruments may have detrimental side effects. For example, although infrequent, some comments are offensive, insulting or potentially threatening (Tucker, 2014). Manual screening or filtering processes can be difficult to implement due to the sheer volume of student comments (up to 100,000 comments can be generated each term at a public research university). This lack of scrutiny over student surveying can lead to pernicious unintended consequences.

One major concern is the psychological distress educators might experience from receiving abusive comments. How unprofessional or abusive comments from students may affect academic staff is an understudied area of research, although it fits into a pattern of student bullying of teachers (May & Tenzek, 2018) and contra-power harassment (Lampman et al., 2009). The psychological effects of abusive comments have been extensively researched in closely related topics such as workplace bullying and cyberbullying. Indeed, one of the most common ways employees are bullied is by being the target of slander, scorn and belittlement (Vartia, 2001). Evidence from various studies has shown that workplace bullying is associated with lower self-esteem (Saks & Ashforth, 1997); higher levels of depression, cardiovascular disease (Kivimäki et al., 2003); anxiety, stress, burnout (Bowling & Beehr, 2006); negative affect, intention to quit and reports of numerous physical symptoms (Djurkovic et al., 2003); see Bowling and Beehr (2006) for a review. Furthermore, a previous study, which examined the relationship between traditional bullying and cyberbullying, shows that they have much overlap, especially regarding their consequences (Kowalski & Limber, 2013). In other words, cyberbullying and traditional bullying closely resemble each other in terms of effect on the victim. Finally, universities may have a student code of conduct that defines acceptable behaviour and forbids students from harassment and bullying.

In addition to educators being exposed to risks associated with bullying, there may be other negative consequences from releasing comments without a pre-screening process. As student feedback is usually provided to the educators for them to read at their own leisure, in our view previous negative experiences with this system may lead them to come up with their own preventative measures. For instance, similar to how employees who are bullied are more likely to leave their jobs (Djurkovic et al., 2003), educators may also be inclined to completely disengage from the system by simply not accessing it. As a result, educators may end up missing out on feedback that may help improve their teaching skills.

Research into abusive language detection including hate speech, cyberbullying and personal attacks has been largely focused on social media (Mehdad & Tetreault, 2016; Nobata et al., 2016; Wulczyn & Dixon, 2017). Additionally, researchers have applied text analysis on social media text corpuses to identify posts containing suicidal ideation (De Choudhury et al., 2016; Ji et al., 2018; O'Dea et al., 2015; Shing et al., 2018). In contrast, text analysis on student experience surveys has largely focused on analytics that can help improve teaching processes (Abd-Elrahman et al., 2010; Ibrahim et al., 2018; Stupans et al., 2016). Bullying, occupational stress and social anxiety are prevalent for both students and staff in modern university environments (Malik & Björkqvist, 2019; Pörhölä et al., 2019; Vaill et al., 2020), and electronic feedback systems like student experience surveys can be abused by bullies in a manner similar to online trolling on social media. A recent study analysing quantitative survey responses has found evidence of bias against non-English speaking teachers and women (Fan et al., 2019).

Free-text responses in large-scale survey instruments can provide valuable information; however, acquiring this information may require a labour-intensive coding process. We explored an alternative semi-automated solution with the aim of reducing the human effort required to screen harmful free-text comments. The main technical difficulty in implementing a text analytics decision support system is acquiring a sufficient quantity of consistently annotated text data. We propose an assessment protocol that assists annotators in categorising comments in consistent manner suitable for text analytics. Further, we show how the unstructured survey comments when combined with the annotations can be used for predictive classification of comments from new surveys. We adopted a standard approach: preprocessing the text into a bag of words form and then classifying with a linear support vector machine (Joachims, 1998).

Our methodology can be adapted by educational institutions to dramatically decrease this labour-intensive process. Screening for harmful free-text comments often involves many working hours and can require multiple staff to ensure screening fidelity. We reduced the necessary work hours for this task by a factor of 14 making a reasonable task for small, specialised teams assisting survey management. Our case study illustrates how researchers and educators can follow best practice to develop a decision support system that implements free-text comment classifiers. Although our methodology focused on specifically improving the teacher assessment experience (with the view, this would improve the overall educational outcomes for both teachers and students), our proposed psychology-based assessment can be used by researchers in a number of different fields as a framework for developing novel context-specific free-text screening protocols.

Educational institutions have a duty of care to provide a safe learning environment for both students and educators. There is a need for a methodology to be set in place to avoid the potential risk of harm in the current feedback system. Our work is a confluence between previous research on text analytics on student experience surveys and research into the identification of abusive language and suicidal ideation in social media. In this paper, we propose a decision support system to assist in the screening of comments at a major Australian university, evaluate the effectiveness of our approach and discuss the insights that our system gives into the underlying data.

Materials and methods

We developed a decision support system composed of three components that independently screen, identify and flag potentially harmful comments for further scrutiny by a human reviewer. Each of the three components use standard text analytic classification techniques to either flag a comment as potentially harmful or no threat. After comments have been labelled by the decision support system, the human reviewer screens the potentially harmful comments and labels them according to the assessment protocol. The design of the assessment protocol drove the design of the text classifiers in the decision support system; therefore, we present and outline the assessment protocol structure first.

We establish four main categories a priori to classify student comments: no-threat, cyberbullying, cyberthreat and self-harm. Aside from no-threat, each category is further divided into low or high severity. Cyberbullying involves dissemination of harmful or cruel speech or engaging in other forms of social cruelty (bullying) via the web. In contrast, cyberthreats are direct threats or materials disseminated via the web that indicate that the author of the cyberthreat may engage in an act of violence. Self-harm diverges from cyberthreat and cyberbullying specifically with regards to where the harm is directed. Whereas cyberthreat and cyberbullying comments intend harm towards others, self-harm comments intend harm towards oneself (Willard, 2007). In our experience, educators found self-harm comments from students highly distressing; the four main categories were established after careful review of previous comments that teachers had identified as concerning.

We define comments that are more explicit, direct and less ambiguous as high severity, while comments that are less explicit, indirect and more ambiguous as low severity. Comments that are clearly not offensive or threatening or have no indication of self-harm are categorised as no-threat. A summary of the classification system with specific examples of severity by category can be found in Table 1.

Table 1
An assessment protocol to assist reviewers when labelling student comments

Category	Description	Related to	Examples
No threat	The comments are highly relevant to the educational aspect of the course and these comments aim to be constructive.	1) The general aspects of the course (e.g., schedule, topics). 2) Academic or administrative staff.	1) "I thought the lecture was boring." 2) "I would have liked the tutor to have more energy."
Low severity cyberbullying	The comments can be considered offensive to some people. Whether the comments were made to intentionally cause emotional distress or harm to the target is ambiguous.	1) Religion, Physical or Mental aspect, socioeconomic status, politics, race, and gender.	1) "Needs to learn how to speak English." 2) "The tutor is really sexy."
High severity cyberbullying	The comments include profanity and/or is clearly offensive. It is clear that the comments were made to intentionally cause emotional distress or harm to the target.		1) "The tutor should stop being a bitch." 2) "Can't stand that retard."
Low severity cyberthreat	The comments convey that the target(s) should be physically harmed via indirect action or cause resulting in serious injuries or loss of life.	1) Academic and administrative staff or property. Note: Comments may also have similar characteristics of Cyberbullying (e.g., use of profanity).	1) "I hope he gets hit by a bus." 2) "I wish he killed himself."
High severity cyberthreat	The comments convey that the target(s) will be physically harmed via direct action by the commenter, resulting in serious injuries or loss of life.		1) "I'm going to murder him." 2) "I'm going to destroy the building."
Low severity self-harm	The comments indicate that the commenter has desire to inflict physical harm to him/herself potentially resulting in serious injuries or loss of life.	1) The commenter as the target (victim). Note: Comments referring to non-life-threatening injuries should be ignored (e.g., "I wish I broke my leg for exam.").	1) "I want to kill myself." 2) "I wish I could die tomorrow."
High severity self-harm	The comments indicate that the commenter has full intention to inflict physical harm to him/herself potentially resulting in serious injuries or loss of life.		1) "I'm going to hang myself." 2) "I'm going to drown myself."

The decision support system is comprised of three components that work independently to label student experience comments. Each component is derived from a unique data source. The first two components are heuristic classifiers that flag student experience comments based on an n-gram list where n ∈ [1, 2, 3] (Cavnar & Trenkle, 1994). The n-gram lists were extracted using text analytics from two external data sources; the first data source contains toxic comments directed towards others, while the second data source contains comments that are indicative of intention of self-harm. The third component is a text classifier trained on an annotated set of historical student experience comments. The final decision support system is comprised of the two text classifiers targeting toxic comment and self-harm n-gram lists and the text classifier trained on historical student comment data.

Empirical examination of student experience comments guided the formulation of this framework; we found that student comments were, on the whole, more reserved than social media comments. The two n-gram lists are intended to identify more toxic outliers that are too sparse to train effectively using student experience data, while the historical student comment text classifier helps identify more subtle yet still potentially harmful comments.

We adopt a simple screening process where the corpus of student experience comments is first processed by the decision support system. The output from the decision support system is a subset of the original corpus containing only the flagged comments. This list of flagged comments is then passed to a human reviewer for final arbitration on the severity of the comments. The entire process for the decision support system and human review is illustrated in Figure 1.

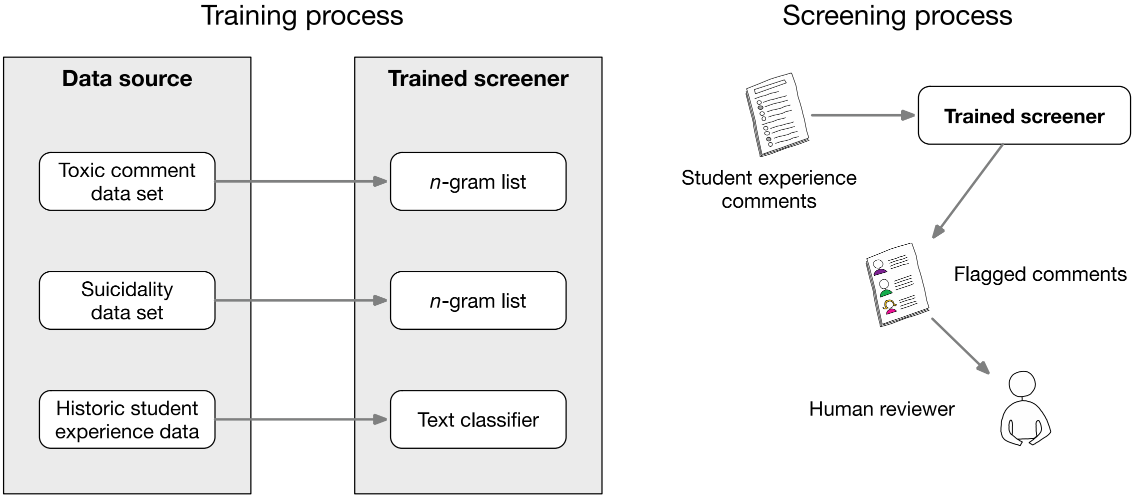

Figure 1. A schematic overview of the training and screening process.
Note. The left side of the figure shows the decision support system training process. The right side shows the decision support system being systematically used in conjunction with a human reviewer.

The same preprocessing procedure is used for all three components of the decision support system. In particular stop words (e.g., “the”, “and”, “of”, ...), white space and numbers are removed. In addition, all characters are transformed to lower case, and each word is replaced with its most basic conjugate form. The comments are weighted using a bag of words model with term frequency-inverse document frequency weighting (Ramos, 2003). Term frequency-inverse document frequency weightings account for the increased importance of terms appearing in problematic comments as well as the decreased importance of terms appearing in all comments.

Our system used two external component data sources. For the first component n-grams are extracted from the Kaggle Wikipedia Toxic Comments Challenge (Wulczyn et al., 2017). This corpus contains 115,737 comments from Wikipedia users, of which 11.7% contain problematic comments annotated into six distinct categories: toxic, severe toxic, obscene, threat, insult and identity hate. The n-grams were extracted from the subset of comments that were flagged with any of the six problematic categories. We refer to this component as Kaggle. For the second component (Reddit), n-grams were extracted from annotated data containing suspected suicidal ideation from the Reddit topic “SuicideWatch” (Ji et al., 2018). The Reddit corpus contains 3,549 suicidal ideation samples and was obtained via request from the authors of the original study. We refer to this component as Reddit. The data in both of these components were anonymised, and personal information was removed by the data set curators.

We drew upon institutional student experience survey (SES) data from a large public Australian university (referred to as the university) for the third and final component. The data for this component was extracted from the university data warehouse and anonymised according to university ethics and governance policies. Major Australian public research universities can have up to 60,000 students, over 90% of whom are enrolled in coursework degrees (Department of Education, Skills and Employment, 2022). The university conducts an internal online student survey for the vast majority of courses every term. In particular, we analysed qualitative responses to the free-text survey question “This person’s teaching could be improved by ...”, where a student response concerned one of the student’s lecturers or tutors. This question was selected from the survey instrument as it solicits a critical response towards an individual. Due to the sensitive information contained in the data set, all data were strictly anonymised, no demographic information was included, and all comments were de-identified.

To annotate, 50,000 responses were randomly sampled without replacement over the two primary academic terms of 2017. This corpus is divided into 50 individual annotation units similar to the methodology employed by Shing et al. (2018). Each annotation unit was initially analysed independently by two annotators, where each of the annotators indicated whether a comment fell into any category in Table 1 besides no-threat. A third annotator annotated the subset of comments where the initial two annotators disagreed (without visibility of the previous annotations). The final annotated value was the consensus between a minimum of two annotators.

The annotated student responses were then used to train a linear support vector machine classifier (SVM) (Joachims, 1998), as this method of classification is effective in identifying abusive language in text (Davidson et al., 2017). The classes for the SVM are unbalanced; comments that belong to the minority class (threatening comments) are infrequent. The problematic comments were oversampled using the synthetic minority oversampling technique (Chawla et al., 2002). An appropriate oversampling balance level was heuristically determined using the Kaggle toxic comment data set (Wulczyn et al., 2017) with methodology similar to (Chen et al., 2018). We are interested in identifying as many potential problematic comments as possible while reducing the Type II error rate, as the screening output will be followed by a final review. As such, recall (true positive / [true positive + false negative]) was optimised on the minority class. Five-fold cross-validation was employed when training and evaluating the performance of the SVM classifier.

The decision support system is composed of the three components (Kaggle, Reddit, and SES), where each of the components screen student comments independently and augment the comment data with a Boolean indicator on whether or not the comment was found to be potentially harmful. This approach allows reviewers to concentrate first on comments that have the highest mutual agreement from all three components.

Results

Over two academic terms, the decision support system classified 4,258 out of 62,049 free-text student comments (6.9%) as potentially harmful resulting in approximately fourteen times less human review work. Of the 4,258 potentially harmful comments, 8.5% were labelled problematic after review: 349 cyberbullying, 4 cyberthreatening and 10 self-harm. Agreement of indicators from all three components in the decision support system was associated with a high positive detection rate (48.4%).

We applied the decision support system to student responses for the free-text question “This person’s teaching could be improved by ...” from the first two academic terms at the University in 2019. The survey instrument used was a standard student experience survey, which is administered internally using experience management software near the end of each academic term. The free-text comments for the two academic terms were extracted from the experience management software and saved in a plain text comma-separated file format. Term 1 elicited 37,215 responses to the teaching question, while Term 2 solicited 24,834 responses to the teaching question for a total of 62,049 responses. The lower number of responses in Term 2 is a trend that is observed annually, where the overall response rate is highest in the first term before dropping substantially over the course of the academic year.

Out of the 62,049 responses, the decision support system identified 2,477 potentially harmful comments from Term 1 and 1,781 potentially harmful comments from Term 2 for a total of 4,258 (6.9%) potentially harmful responses. A comment was considered potentially harmful if any of the three components in the decision support system identified that comment. The final subset of 4,258 potentially harmful comments was annotated with an indicator variable corresponding to each of the three components of the decision support system.

A human reviewer manually labelled the 4,258 results from the decision support system according to the psychology-based assessment protocol described in Table 1. The reviewer was allowed to sort the potentially harmful comments according to the highest agreement amongst the decision support system components. Additional reviewers validated these results after the initial labelling was complete. The human labelling process identified 363 comments that fell into categories beyond no-threat. The 363 comments represent around 8.5% of the comments that the decision support system flagged for review. Agreement of indicators from all three components in the decision support system provided the highest positive detection rate (48.4%) after confirmation from the human reviewer.

Table 2
Count and rate of problematic comments identified by the decision support system

SES	Kaggle	Reddit	Self-harm	Cyberbully	Cyberthreat	No-threat	Rate
Y	Y	Y	1	58 (3)	2 (1)	65	0.484
Y	N	Y	3	59	0	151	0.291
Y	Y	N	0	31 (1)	0	96	0.244
N	Y	Y	3	15	0	163	0.0904
Y	N	N	0	156 (1)	1	2243	0.0651
N	Y	N	0	14	0	354	0.0380
N	N	Y	3	16	1	823	0.0232

Note. The table shows the number of self-harm, cyberbullying, cyberthreat and no-threat comments based on the combination of indicators from the three components in the decision support system (SES, Kaggle and Reddit). Numbers in parentheses indicate the number of comments that were high severity for that given category. The rate of problematic comments to all comments is displayed in the last column.

Table 2 shows the breakdown of comments identified by the decision support system components and the category that the comments were labelled under using the psychology-based assessment protocol. Each row of Table 2 shows the categorical breakdown and the threat rate for a given level of agreement between the three classifiers (Kaggle, Reddit and SES). For a given category (e.g., cyberbully), if there is a number in parentheses in the table, it indicates how many of the total number of identified comments were of high severity. For example, in the top row where all three indicators agree, there were 58 instances of cyberbullying identified by the human labeller, and of those 58, there were 3 that were found to be of high severity. The categories in Table 2 are not mutually exclusive (i.e., a comment could be considered both cyberbullying and cyberthreatening at the same time). After analysing the data from the two academic terms, we identified only a single comment that fell under two categories.

During the labelling process, we screened the comments to see if there were any outstanding subgroups of interest. We found that a large number of the comments identified by the decision support system (471 or 11%) contained references to either “accent” or “English”. Even though these comments formed a substantial portion of those that were identified, a majority of this subgroup of comments (441 or 94%) were labelled by the human reviewer as no-threat. Many of comments in this subgroup were short and lacked the context to be classified as bullying or threatening. We did not identify any other substantial subgroups of interest during the labelling process.

Discussion

Automated methods for screening student experience comments can dramatically reduce staff workload and help identify harmful free-text comments before they are released to teachers. Research on the efficacy of student evaluation surveys dates back to the 1920s, and there is an established consensus that these surveys are a worthwhile means of evaluating teaching at higher education institutions (Wachtel, 1998). Student evaluation of teaching surveys are widely used as a component of faculty assessment in the United States of America, Australia, and Europe. A 2010 survey of academic deans in the United States of America (Miller & Seldin, 2014) showed that 94% of institutions always used systematic student ratings. Additionally, a 2014 survey conducted by (Vasey & Carroll, 2016) had 54% of respondents state that their institution required frequent use of online student evaluations. There is active international academic discussion concerning student evaluation of teaching surveys (Iyamu & Aduwa-Oglebaen, 2005; Mittal & Gera, 2013; Yin et al., 2014). The use of online forums as a method of bullying is also widespread internationally in educational contexts (Bhat et al., 2017; Garrett, 2014; Odora & Matoti, 2015; Safaria, 2016; Woudstra et al., 2018; Zhou et al., 2013). Although the decision support system we describe was applied to student comments from an Australian university, the methodology is general and can be easily adapted to use across a wide range of international educational institutions.

Allowing free-text responses can elicit valuable feedback, but these instruments must be reliable and not allowed as a channel for abuse. Student evaluations of teaching are potentially contentious (Boring et al., 2016; Cramer & Alexitch, 2000; Spooren et al., 2013) and allowing online abuse in evaluation surveys will only further this perception. Online student evaluations of teaching easily allow the application of text analytics to free-text comments, and methods similar to ours can be used to classify this unstructured data. An ideal solution to detecting problematic comments would be to manually screen every comment per term. Unfortunately, this would come at the expense of human effort; from empirical observation, screening several hundred comments took a dedicated annotator about 1 hour. The total time to complete annotations for an academic year could potentially take 40 to 60 days of full-time work.

The quantity of problematic comments we discovered from screening two terms of data was small (363 out of 62,049 or 0.59%) and only slightly larger than Tucker's (2014) previous manual examination of comments at a different Australian university (0.19% unprofessional or abusive comments). Our empirical examination of problematic student comments found them less overtly offensive compared to comments found in the Reddit and Kaggle pseudonymous social media corpuses. This observation is in line with findings from previous research; negative student evaluations of teaching tend to use less intense language when compared to positive evaluations (Stewart, 2015). The rate of problematic comments identified was very high (48.4%) when all three components of the components in the decision support system agreed and over 24% when any two of the components agreed. By using ensemble voting between the three components trained with different types of problematic comments, reviewers can focus first on the comments identified by a majority of those components.

The vast majority of identified problematic comments were cyberbullying (96.1%). The bullying in the comments tended to follow a theme, where the student in question would treat the teacher evaluation survey as an anonymous complaint form. We found that a typical cyberbullying comment from a student lacked constructive criticism and tended to exclusively blame the teacher for their negative learning experience. Profanity was rare, but attacks on the teacher's perceived job competency were very common, for example, "They should just quit teaching".

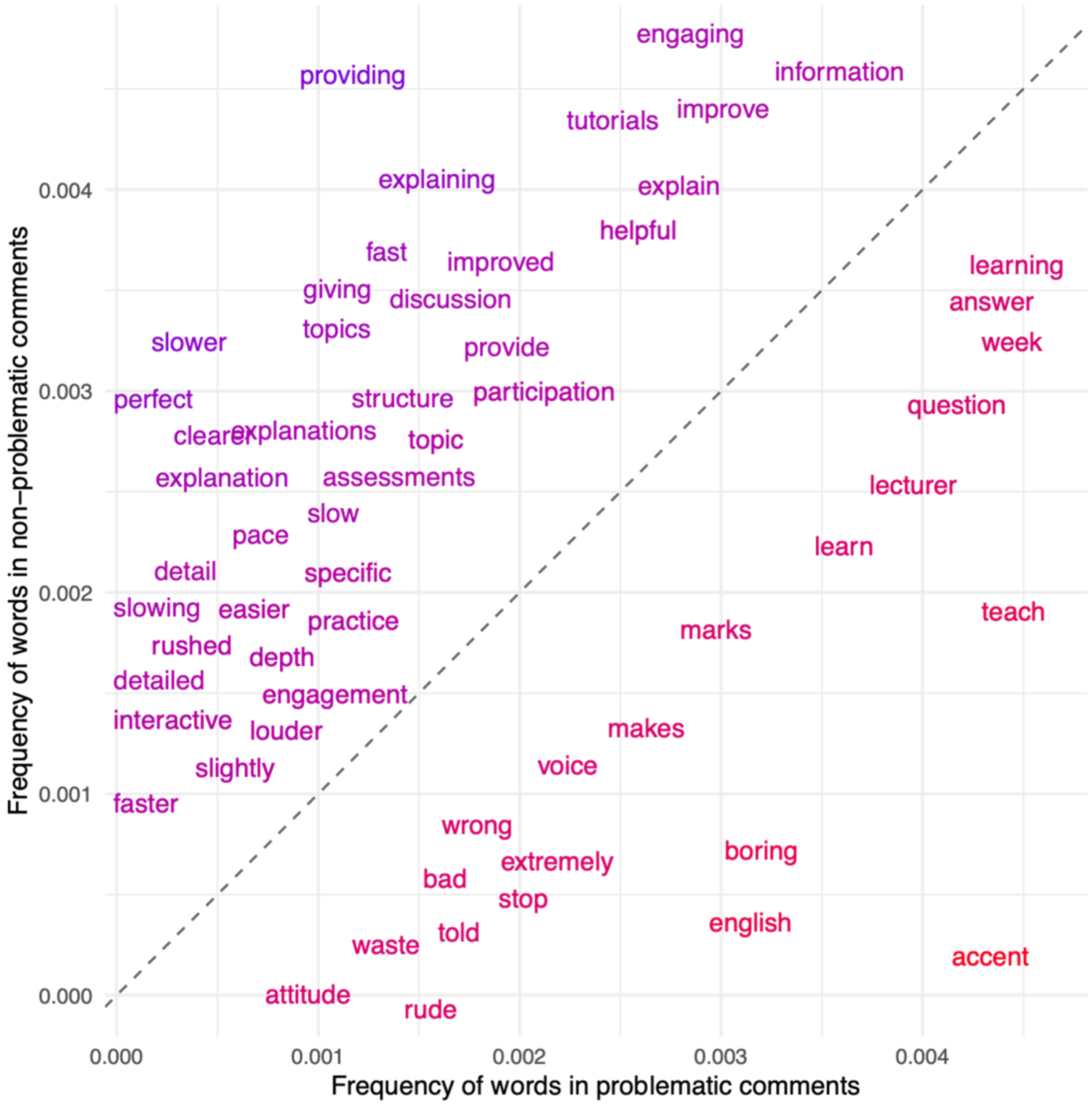

Figure 2. Representative words indicating problematic and non-problematic comments
Note. The figure shows the frequency of words in comments classified as potentially harmful (by the decision support system) versus the frequency of words in all other comments for Terms 1 and 2 in 2019. Words are displayed on the figure if they appear in both categories of comments and the difference between frequencies is greater than 0.001.

In our examination of student comments that were flagged by the decision support system, we found that 11% of those comments referenced either “accent” or “English”. The majority of these comments were very brief, generally non-offensive but might be considered culturally insensitive, for example, “had trouble understanding their accent”. The less overtly offensive characteristic in these comments agrees with recent research into large platform online ratings of professors by students (Subtirelu, 2015). Additionally, a recent study on psychometric responses to student evaluations of teaching at the same institution as our study found evidence of potential bias against non-English speaking background teachers (Fan et al., 2019). Around 38% of the university’s teaching staff are from a non-English speaking background (Fan et al., 2019), and around a quarter of the students are international. This suggests that there are frequent occurrences where one or both of teaching staff and student originate from (potentially different) non-English language backgrounds. Communication is a primary component of teaching (Frymier & Houser, 2000; Rasmussen, 2001), and misunderstanding between the parties involved could easily lead to frustration. This frustration combined with a lack of empathy for cultural differences could explain the relatively high frequency of comments targeting accent and English proficiency.

This is illustrated by Figure 2, which shows words from the comments plotted by their frequency in problematic comments against their frequency in non-problematic comments. The further a word is off the diagonal, the more exclusive that word is to the corpus on the corresponding axis. The frequency of words in problematic comments follows the x-axis; the words “boring”, “English”, “teach” and “accent” all tend towards that axis. This corresponds with our empirical examination of potentially problematic comments: students would frequently complain about their teacher’s ability to engage the class, teach correctly or the English proficiency and/or accent of the teacher.

Because university survey instruments typically capture psychometric evaluations of teaching staff alongside the qualitative text comments, a potential future direction would be to incorporate both of these sources of information when screening comments. The combined use of qualitative and quantitative survey data is not widespread, and there is active discussion about the merits of this approach (Gergen, 2015; Jackson, 2015; Landrum & Garza, 2015). Additionally, student and staff demographic information (where available) could potentially help identify problematic comments as minority groups may suffer higher rates of cyberbullying (Wensley & Campbell, 2012).

Our case study offers points of best practice that extend widely to the application of text analytics in educational technology. Machine learning tools for text analytics are widespread, well documented and mature; there are a number of easy-to-use libraries, such as WEKA (Hall et al., 2009) and scikit-learn (Pedregosa et al., 2011). These libraries include functions to convert text to an amenable representation for classification such as term frequency-inverse document frequency. Simple heuristic classifiers are a good starting point as the heuristic rule can later be incorporated as a classifier for features in a formal machine learning context. This being said, there is a jump in project complexity as soon as machine learning–based classifiers are incorporated. These classifiers are sensitive to input data quality and have a requirement that the annotation process is high quality and consistent. We found that consistency is difficult between multiple annotators even when annotation standards are trained and practised together. The annotation protocol described in Table 1 was instrumental in achieving a satisfactory outcome; several revision cycles of the protocol with feedback were required. Rigorous annotation was the most time-consuming part of this process, and we suggest that potential applicators distinguish early on if they wish to use text analytics to automate standard business processes or if they want to conduct a full analysis for further research purposes.

Text data is unstructured and opaque to conventional statistical analysis; however, the text analytics tools to process this data into a format suitable for analysis are now accessible to individuals with modest analytical expertise. More sophisticated techniques such as topic modelling and text summarisation offer opportunities to condense and analyse large amounts of text data. Although we adopted a supervised learning approach for classification, an unsupervised or partially supervised learning system could be rapidly deployed in practice. An ideal solution for educational institutions would be an accurate unsupervised classifier available as an open-source software tool.

Although our study discovers a rate of abuse that is approximately similar to that in Tucker (2014), it is not a comprehensive solution for identifying abuse. Even with our ensemble approach, the decision support system identified a large proportion of false positive offensive comments. We made a decision early on to tune the system to minimise Type II errors at the expense of Type I errors. Future research could carefully optimise a system for both types of errors, which would further reduce manual screening labour.

Higher education institutions have a duty of care responsibility to both student and staff psychological well-being and safety. The use of online communication has numerous benefits but exposes students and staff to the risk of cyberbulling and other abuse. Early identification with an automated screening system allows for proactive intervention and support to involved parties. Frameworks like the decision support system described in this paper help deal with emergent problems from the rapid adoption of online technologies and the growing number of staff and students at educational institutions.

Conclusion

Overall, the decision to construct a software-based decision support system was useful for the staff involved in the project. The automated process improved over previous manual inspection and allowed systematic identification of problematic comments. Although the need to enforce testing and provide evaluation metrics was quite intensive and well beyond what was required to train the decision support system, a less rigorous approach to training would be both practical and accessible for other institutions. For researchers contemplating similar work, an incremental text coding process would allow the relevant annotation data to be built up slowly over several rounds of surveys.

Text analytics projects are not widely conducted in educational research although they can be very useful. Our method shows that these projects can be done in an iterative fashion that is approachable for researchers developing their expertise. Researchers could start with simple classifiers and combine them with more complicated machine learning models that learn from the data. This approach is not redundant as the models can be combined via bagging (majority voting) to further boost the performance. By using a human-in-the-loop, we do not need the full complexity of a completely automated solution, but we can effectively extend the reach of the worker or researcher.

Online survey instruments allow universities to collect and disseminate feedback and constructive criticism from students. Survey evaluation is a crucial part of closing the loop: instruction and learning, student evaluation followed by feedback and revision of teaching (Powney & Hall, 1998). To keep this process robust and healthy, it is vital for instructors and students to engage in a constructive dialogue. Early detection of potentially distressing or offensive comments should encourage instructors to stay engaged throughout the evaluation process.

References

Abd-Elrahman, A., Andreu, M., & Abbott, T. (2010). Using text data mining techniques for understanding free-style question answers in course evaluation forms. *Research in Higher Education Journal*, *9*, Article 10520. https://www.aabri.com/manuscripts/10520.pdf

Bhat, C. S., Ragan, M. A., Selvaraj, P. R., & Shultz, B. J. (2017). Online bullying among high-school students in India. *International Journal for the Advancement of Counselling*, *39*(2), 112–124. https://doi.org/10.1007/s10447-017-9286-y

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Bowling, N. A., & Beehr, T. A. (2006). Workplace harassment from the victim’s perspective: A theoretical model and meta-analysis. *Journal of Applied Psychology*, *91*(5), 998–1012. https://doi.org/10.1037/0021-9010.91.5.998

Cavnar, W. B., & Trenkle, J. M. (1994). *N-gram-based text categorization* (Technical Report). U.S. Department of Energy. https://www.osti.gov/biblio/68573

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, H., McKeever, S., & Delany, S. J. (2018) A comparison of classical versus deep learning techniques for abusive content detection on social media sites. In S. Staab, O. Koltsova, & D. Ignatov (Eds.), *Lecture notes in computer science: Vol. 11185. SocInfo 2018*: *Social informatics* (pp. 117–133). Springer. https://doi.org/10.1007/978-3-030-01129-1_8

Cramer, K. M., & Alexitch, L. R. (2000). Student evaluations of college professors: Identifying sources of bias. *Canadian Journal of Higher Education*, *30*(2), 143–64. https://doi.org/10.47678/cjhe.v30i2.183360

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512–515. https://ojs.aaai.org/index.php/ICWSM/article/view/14955

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2098–2110). Association for Computing Machinery. https://doi.org/10.1145/2858036.2858207

Department of Education, Skills and Employment. (2022). *2020 Student summary tables* (D22/70092) [Data set]. https://www.dese.gov.au/higher-education-statistics/resources/2020-student-summary-tables

Djurkovic, N., McCormack, D., & Casimir, G. (2003). The physical and psychological effects of workplace bullying and their relationship to intention to leave: A test of the psychosomatic and disability hypotheses. *International Journal of Organization Theory & Behavior*, *7*(4), 469–497. https://doi.org/10.1108/IJOTB-07-04-2004-B001

Fan, Y., Shepherd, L., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLOS ONE*, *14*(2), e0209749. https://doi.org/10.1371/journal.pone.0209749

Frymier, A. B., & Houser, M. L. (2000). The teacher-student relationship as an interpersonal relationship. *Communication Education*, *49* (3), 207–219. https://doi.org/10.1080/03634520009379209

Garrett, L. (2014). The student bullying of teachers: An exploration of the nature of the phenomenon and the ways in which it is experienced by teachers. *Aigne*, *5*(1), 19–40. https://aigne.ucc.ie/index.php/aigne/article/view/1476

Gergen, K. J. (2015). The quantitative/qualitative distinction: Blessed are the impure. *Qualitative Psychology*, *2*, 210–213. https://doi.org/10.1037/qup0000034

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18. https://doi.org/10.1145/1656274.1656278

Ibrahim, Z. M., Bader-El-Den, M., & Cocea, M. (2018). A data mining framework for analyzing students feedback of assessment. In C. Glahn & L. Dirckinck-Holmfeld (Eds.), *Proceedings of the 13th EC-TEL Doctoral Consortium* (pp. 1–7). CEUR-WS. http://ceur-ws.org/Vol-2294/DCECTEL2018_paper_13.pdf

Iyamu, E. O., & Aduwa-Oglebaen, S. E. (2005). Lecturers' perception of student evaluation in Nigerian universities. *International Education Journal*, *6*(5), 619–625. http://ijdri.com/iej/2005/2005dec.pdf

Jackson, M. R. (2015). Resistance to qual/quant parity: Why the 'paradigm' discussion can't be avoided. *Qualitative Psychology*, *2*(2), 181–198. https://doi.org/10.1037/qup0000031

Ji, S., Yu, C., Fung, S., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. *Complexity*, Article 6157249. https://doi.org/10.1155/2018/6157249

Joachims T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Lecture notes in computer science: Vol 1398. Machine learning* (pp. 137–142). Springer. https://doi.org/10.1007/BFb0026683

Kivimäki, M., Virtanen, M., Vartia, M., Elovainio, M., Vahtera, J., & Keltikangas-Järvinen, L. (2003). Workplace bullying and the risk of cardiovascular disease and depression. *Occupational and Environmental Medicine, 60*(10), 779–783. https://doi.org/10.1136/oem.60.10.779

Kowalski, R. M., & Limber, S. P. (2013). Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of Adolescent Health*, 53(1), S13–S20. https://doi.org/10.1016/j.jadohealth.2012.09.018

Lampman, C., Phelps, A., Bancroft, S., & Beneke, M. (2009). Contrapower harassment in academia: A survey of faculty experience with student incivility, bullying, and sexual attention. *Sex Roles*, *60*(5), 331–346. https://psycnet.apa.org/doi/10.1007/s11199-008-9560-x

Landrum, B., & Garza, G. (2015). Mending fences: Defining the domains and approaches of quantitative and qualitative research. *Qualitative Psychology*, *2*(2), 199–209. https://doi.org/10.1037/qup0000030

Malik, N. A., & Björkqvist, K. (2019). Workplace bullying and occupational stress among university teachers: Mediating and moderating factors. *Europe's Journal of Psychology*, *15*(2), 240–259. https://doi.org/10.5964/ejop.v15i2.1611

May, A., & Tenzek, K. E. (2018). Bullying in the academy: understanding the student bully and the targeted 'stupid, fat, mother fucker' professor. *Teaching in Higher Education*, *23*(3), 275–290. https://doi.org/10.1080/13562517.2017.1379482

Mehdad, Y., & Tetreault, J. (2016). Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 299–303). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-3638

Miller, J. E., & Seldin, P. (2014). Changing practices in faculty evaluation. *Academe*, *100*(3), 35–38. https://www.aaup.org/article/changing-practices-faculty-evaluation

Mittal, S., & Gera, R. (2013). Student evaluation of teaching effectiveness (SET): An SEM study in higher education in India. *International Journal of Business and Social Science*, *4*(10). https://ijbssnet.com/journals/Vol_4_No_10_Special_Issue_August_2013/35.pdf

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 145–153). Association for Computing Machinery. https://doi.org/10.1145/2872427.2883062

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions*, *2*(2), 183–188. https://doi.org/10.1016/j.invent.2015.03.005

Odora, R. J., & Matoti, S. N. (2015). The nature and prevalence of cyber bullying behaviors among South African high school learners. *International Journal of Educational Sciences, 10*(3), 399–409. https://doi.org/10.1080/09751122.2015.11890362

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pörhölä, M., Almonkari, M., & Kunttu, K. (2019). Bullying and social anxiety experiences in university learning situations. *Social Psychology of Education*, *22* (3), 723–742. https://doi.org/10.1007/s11218-019-09496-4

Powney, J., & Hall, S. (1998). *Closing the loop: The impact of student feedback on students' subsequent learning*. Scottish Council for Research in Education.

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning, 242*(133–142).

Rasmussen, J. (2001). The importance of communication in teaching: A systems-theory approach to the scaffolding metaphor. *Journal of Curriculum Studies*, *33*(5), 569–582. https://doi.org/10.1080/00220270110034369

Safaria, T. (2016). Prevalence and impact of cyberbullying in a sample of Indonesian junior high school students. *The Turkish Online Journal of Educational Technology*, *15*(1), 82–91. http://www.tojet.net/articles/v15i1/1519.pdf

Saks, A. M., & Ashforth, B. E. (1997). Organizational socialization: Making sense of the past and present as a prologue for the future. *Journal of Vocational Behavior*, *51*(2), 234–279. https://doi.org/10.1006/jvbe.1997.1614

Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé, H., III, & Resnik, P. (2018). Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology* (pp. 25–36). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0603

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, *83*(4), 598–642. https://doi.org/10.3102/0034654313496870

Stewart, M. (2015). The language of praise and criticism in a student evaluation survey. *Studies in Educational Evaluation*, *45*, 1–9. https://doi.org/10.1016/j.stueduc.2015.01.004

Stupans, I., McGuren, T., & Babey, A. M. (2016). Student evaluation of teaching: A study exploring student rating instrument free-form text comments. *Innovative Higher Education*, *41*(1), 33–42. https://doi.org/10.1007/s10755-015-9328-5

Subtirelu, N. (2015). "She does have an accent but…": Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com. *Language in Society*, *44*(1), 35–62. https://doi.org/10.1017/S0047404514000736

Tucker, B. (2014). Student evaluation surveys: Anonymous comments that offend or are unprofessional. *Higher Education*, *68*(3), 347–358. https://doi.org/10.1007/s10734-014-9716-2

Vaill, Z., Campbell, M., & Whiteford, C. (2020). Analysing the quality of Australian universities' student anti-bullying policies. *Higher Education Research & Development*, *39*(6), 1262–1275. https://doi.org/10.1080/07294360.2020.1721440

Vartia, M. A.-L. (2001). Consequences of workplace bullying with respect to the well-being of its targets and the observers of bullying. *Scandinavian Journal of Work, Environment & Health*, *27*(1), 63–69. https://doi.org/10.5271/sjweh.588

Vasey, C. & Carroll, L. (2016). How do we evaluate teaching? *Academe, 102*(3), 34–9. https://www.aaup.org/article/how-do-we-evaluate-teaching

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, *23*(2), 191–212. https://doi.org/10.1080/0260293980230207

Wensley, K., & Campbell, M. (2012). Heterosexual and nonheterosexual young university students' involvement in traditional and cyber forms of bullying. *Cyberpsychology, Behavior, and Social Networking*, *15*(12), 649–654. https://doi.org/10.1089/cyber.2012.0132

Willard, N. E. (2007). *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress* (2nd ed.). Research Publishers LLC.

Woudstra, M. H., van Rensburg, E. J., Visser, M., & Jordaan, J. (2018). Learner-to-teacher bullying as a potential factor influencing teachers' mental health. *South African Journal of Education*, *38*(1), Article 1358. https://doi.org/10.15700/saje.v38n1a1358

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1391–1399). Association for Computing Machinery. https://doi.org/10.1145/3038912.3052591

Yin, H., Lu, G., & Wang, W. (2014). Unmasking the teaching quality of higher education: Students' course experience and approaches to learning in China. *Assessment & Evaluation in Higher Education*, *39*(8), 949–970. https://doi.org/10.1080/02602938.2014.880107

Zhou, Z., Tang, H., Tian, Y., Wei, H., Zhang, F., & Morrison, C. M. (2013). Cyberbullying and its risk factors among Chinese high school students. *School Psychology International*, *34*(6), 630–647. https://doi.org/10.1177/0143034313479692

Corresponding author: R. Zach Aandahl, zach.aandahl@utas.edu.au

Please cite as: Gibson, M. J., Luong, J., Cho, H., Moh, B., Zanin, S., Djatmiko, M., & Aandahl, R. Z. (2022). Where's the harm? Screening student evaluations of teaching for offensive, threatening or distressing comments. *Australasian Journal of Educational Technology, 38*(2), 35-48. https://doi.org/10.14742/ajet.6133