Received 20 September 2022
Revised 8 January 2023
2 March 2023
5 March 2023
10 March 2023
Accepted 12 March 2023

Exploring the influence of student emotions and professor behaviour on course ratings: a quantitative analysis

Krzysztof Rybinski
Department of Business and International Relations, Vistula University, Warsaw, Poland and European Humanities University, Vilnius, Lithuania

Abstract

Purpose – This paper aims to investigate the relationship between student emotions, professors' performance and course ratings and difficulty.

Design/methodology/approach – Natural language processing models are used to extract six basic emotions and several categories of professors' harmful performance from nearly one million student reviews randomly selected from the website ratemyprofessors.com. These features are used in regression analysis to analyse their relationship with numerical ratings of course quality and course difficulty.

Findings – Negative emotions and bad performance by professors are detected more often for low-rated courses and courses perceived as more difficult by students. Positive emotions are seen for highly rated and less challenging courses.

Practical implications – This paper shows that natural language processing tools can be used to enhance and strengthen the quality assurance processes at universities. The proposed methods can improve the often-contested student evaluation of teaching practices, help students make better and more informed choices about their courses and assist instructors to better tailor their teaching approaches and create a more positive learning environment for their students.

Originality/value – This paper presents a novel analysis of how student emotions and poor performance by professors, derived automatically from teacher evaluations by students, affect course ratings. Results also lead to a novel hypothesis that the student–course emotional match or student tolerance of bad behaviour by professors can affect the performance of students and their chances of completing their degree.

Keywords Student evaluation of teaching, Student emotions, Teacher bad behaviour, Course ratings, Natural language processing, Zero-shot learning

Paper type Research paper

1. Introduction

Student evaluation of teaching (SET) is one of the most contested practices in higher education, as shown by the voluminous literature and many meta-studies. Even so it has become an increasingly important metric of quality at the individual, institutional and system levels (Darwin, 2020) and is regularly conducted at more than 16,000 higher education institutions globally (Cunningham-Nelson *et al.*, 2019). It affects faculty promotion, tenure and wage rises. Open platforms that collect student feedback, such as

Competing interests: No potential conflict of interest was reported by the author.
Data availability statement: The data used in this article is available online at https://doi.org/10.5281/zenodo.7097271.
Funding details: This work has not been supported by any funding.

Quality Assurance in Education
Vol. 31 No. 3, 2023
pp. 436-451

0968-4883
DOI 10.1108/QAE-09-2022-0171

ratemyprofessors.com in the USA or whatuni.com in the UK, provide valuable and trusted information for prospective and current students (Chiang, 2017) and are frequently consulted by those choosing a course (Boswell and Sohr-Preston, 2020).

Extensive literature shows that teacher characteristics are essential in determining learning outcomes and strongly influence teaching evaluations. Hanushek *et al.* (2019) show, for example, that the cognitive skills of teachers are strongly related to the performance of students across 31 countries. Pounder (2014) argues that teachers displaying transformational classroom leadership behaviours stimulate academic motivation and effort, engage students in instructional activities in their own time, engender self-efficacy and facilitate cognitive and affective learning. Ran and Xu (2019) and Feld *et al.* (2020) analyse how various teacher characteristics impact student learning outcomes and student performance. One way that teacher performance is influential is through the emotions of students, as Goetz *et al.* (2013) showed, for example, that a supportive presentation style was conducive to students exhibiting positive emotions like enjoyment and pride, while high levels of excessive demand from lessons were associated with negative emotions like anxiety, anger and helplessness. Mainhard *et al.* (2018) investigate how the interpersonal quality of teaching accounts for the variability in student emotions and conclude that teachers may be even more important for student emotions than the previous research showed.

As discussed, SET is a commonly used tool to evaluate the quality of courses and constitutes one of the key pillars of the university quality assurance system. However, the typical practice is to base the course quality assessment on the numerical, Likert scale, part of the SET survey. At the same time, systematic processing and understanding of the narrative portion of the SET surveys are less common. These outdated quality assurance practices do not reflect the profound evolution of new young generations. They exhibit growing awareness, increasingly see themselves as customers (Toth and Bedzsula, 2021) and provide frequent instant feedback on social networks through text with emoticons.

There are no technical barriers to reforming the SET as advances in automated text analysis in the past decade allowed researchers to study the corpora of student evaluations of teaching. These studies applied finding keywords and phrases (Subtirelu, 2015, 2017; Park, 2019; Murray *et al.*, 2020), performed sentiment analysis (Chou *et al.*, 2020; Okoye *et al.*, 2022) or topic modelling (Azab *et al.*, 2016). Recurrent neural networks have been used to analyse 154,000 instructor reviews (Onan, 2019), and they achieved higher accuracy than the conventional machine learning classifiers on the sentiment analysis task. The most recent deep learning models based on the transformers architecture have been trained to predict course ratings in a corpus of over one million student reviews in the USA and UK (Rybinski and Kopciuszewska, 2021) and achieved good accuracy.

Several works discuss the correlation between the sentiment of the student reviews and the course ratings (Chou *et al.*, 2020; Okoye *et al.*, 2022) or the perceived course difficulty (Felton *et al.*, 2008). However, for the first time in the literature, this paper applies an advanced natural language processing (NLP) method called zero-shot learning to extract more complex student emotions rather than sentiment from the large corpus of one million student reviews. The negative emotions, such as *fear* or *anger* that students may have about a professor or a course, are often caused by bad performance on the part of the professor, such as weak teaching skills, poor communication or harassment. Consequently, several types of bad or incompetent performance by professors are also extracted from the student reviews.

The findings of this study have implications for instructors in higher education. By understanding the emotions and experiences of their students as reflected in their course

evaluations, instructors can better tailor their teaching approaches and create a more positive learning environment for their students. Moreover, the findings of this study may contribute to improvements in the professional development of instructors, aimed at addressing any areas for improvement identified through the analysis of student evaluations. Overall, the examination of student emotions and professors' actions in this study has the potential to contribute to the broader goal of improving student learning in higher education.

The novel approach presented in this paper enables the systematic analysis of large volumes of qualitative data and combines this information with traditional Likert scale SET scores to offer fresh insights into the learning and teaching processes.

2. Theory, data and research hypotheses

2.1 Student evaluation of teaching

Several meta-studies of SET covering more than 500 papers indicate why SET is one of the most contested practices in higher education (Emery *et al.*, 2003; Spooren *et al.*, 2013; Linse, 2017; Kreitzer and Sweet-Cushman, 2021; Heffernan, 2021; Rybinski and Kopciuszewska, 2021). While there are some studies showing that there are minimal or no biases in the judgment of students towards their professors (Abbas *et al.*, 2022; Okoye *et al.*, 2022), vast majority of studies shows that teaching evaluations are affected by factors that are unrelated to the quality of the teaching or learning outcomes. Biases include the positive correlation between students' grades and their evaluations and bias against non-majority faculty for their gender, skin colour and politics or religion. Moreover, student evaluations are influenced by racist, sexist and homophobic prejudices, including increasingly abusive comments directed mostly to women (Adams *et al.*, 2021) and those from marginalised groups. Student ratings are shown to be biased about discipline and subject area, and they are heavily influenced by student demographics and the academic culture of the teaching process. Even when universities have systematic processes to collect student feedback using a range of mechanisms, the data to inform improvements is not used effectively, and students remain sceptical about the real impact of SET on the quality of teaching (Shah *et al.*, 2017). Students will even punish instructors who implement many well-known teaching techniques, encouraging instructors to increase SET scores by sacrificing the learning process (Crumbley *et al.*, 2001).

Many of the issues listed above, and the list is by no means comprehensive, arise because university administrators use the numerical part of SET. However, they make only limited use of the narrative part in which textual answers are given to open questions in the SET surveys. One possible reason is that educators have reservations about the use of the textual part of SET, even when such data is provided in the form of visualisation reports (Cunningham-Nelson *et al.*, 2021). However, the recent advances in NLP have changed the game. Such models have a deep understanding of the language semantic structures, understand the different meanings of the exact words depending on the context and achieve super-human accuracy on many language tasks (Wang *et al.*, 2019). With their help, valuable and reliable data can be extracted from student textual reviews and used in economic and machine learning models, as shown in this paper. Chung Sea Law (2010) finds that using student surveys can extract useful student experience information and has the potential to transform the quality-monitoring mechanisms and help shift the focus of quality assurance activities more to the enhancement-led views.

2.2 Student reviews

The website ratemyprofessors.com is the most extensive open-access collection of student reviews globally, with more than 20 million collected since 1999, most of them for US

universities and colleges and some for Canadian ones. The data have been analysed by numerous papers, some of them in recent years, covering anywhere from a few thousand to as many as 7.9 million student reviews, with some key and recurring observations.

The numerical data from ratemyprofessors.com have been found to exhibit various biases, such as racial minority faculty being evaluated worse than white faculty (Reid, 2010; Murray *et al.*, 2020), which has implications for tenure and promotion. Faculty who have tended to receive higher ratings are young, male, in the humanities and arts, rather than lower-ranked science and engineering faculty and holding the rank of full professor (Boehmer and Wood, 2017; Rosen, 2018; Murray *et al.*, 2020). Factors associated with lower ratings were course difficulty and whether the student reviews mentioned accent or teaching assistants (Murray *et al.*, 2020). Student reviews exhibited prior rating bias (Ackerman and Chung, 2017), while positive correlations were observed between the course ratings and instruction quality and easiness (Boehmer and Wood, 2017; Rosen, 2018). Student reviews were negatively affected by the age of the professors once they turned forty, but this effect disappears for those who are considered attractive by student raters (Stonebraker and Stone, 2015). However, Flegl and Rosas (2019) demonstrate that the effect of experience dominates the effect of gender and, in some areas, also the effect of age.

Despite these biases, the website has been shown to offer valuable insights for students, making them more engaged in their chosen course (Reber *et al.*, 2017) and affecting their choice of courses and professors (Boswell and Sohr-Preston, 2020). Furthermore, the majority of the students surveyed expressed trust in the ratings featured on the website (Chiang, 2017), despite the many biases documented by the literature cited earlier. Moreover, the discussion section shows that the effect of student emotions is by order of magnitude larger than the biases discussed above.

Each student review on the website ratemyprofessors.com provides rich metadata, but the information collected for this research was the text of the review, the name of the university or college, the date of the review, the course/professor quality rating and the score for the perceived level of course difficulty. Course/professor quality ratings are in the range from 1 – worst to 5 – best, with the possible values scraped from the ratemyprofessor.com website 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5 and 5. Perceived course difficulty scores are in the range from 1 – very easy to 5 – very difficult, with the possible values scraped from the ratemyprofessors.com website 1, 2, 3, 4 and 5.

The data were collected in April 2020, and the scrapers collected a random sample of nearly one million reviews for the years 1999–2020. A limit of no more than 20 reviews per course or professor was imposed, and if there were more reviews than that, then the most recent 20 were collected. The data cover more than 69,000 teaching faculty.

Figure 1 shows that the random sample scraping from the ratemyprofessors.com website collected between two thousand and ten thousand reviews each month, with peaks in May and December. There are some periods where student ratings deviate from the sample mean, which is taken into account by using annual and monthly dummy variables. Ratings are also becoming more dispersed towards the end of the analysed period.

Previous research has shown (Boehmer and Wood, 2017; Rosen, 2018; Murray *et al.*, 2020) that students tend to give higher ratings to easier courses. Table 2 presents the joint distribution of the course/professor rating and the course difficulty. As expected, the lowest ratings, where the rating = 1, are most often given to challenging courses with a difficulty level of 5, with a substantial margin over classes with a difficulty score of 4, at 61,484 cases against 26,646. Interestingly, the opposite relationship does not hold. Although the most accessible courses with a difficulty level of 1 enjoy the maximum rating much more often

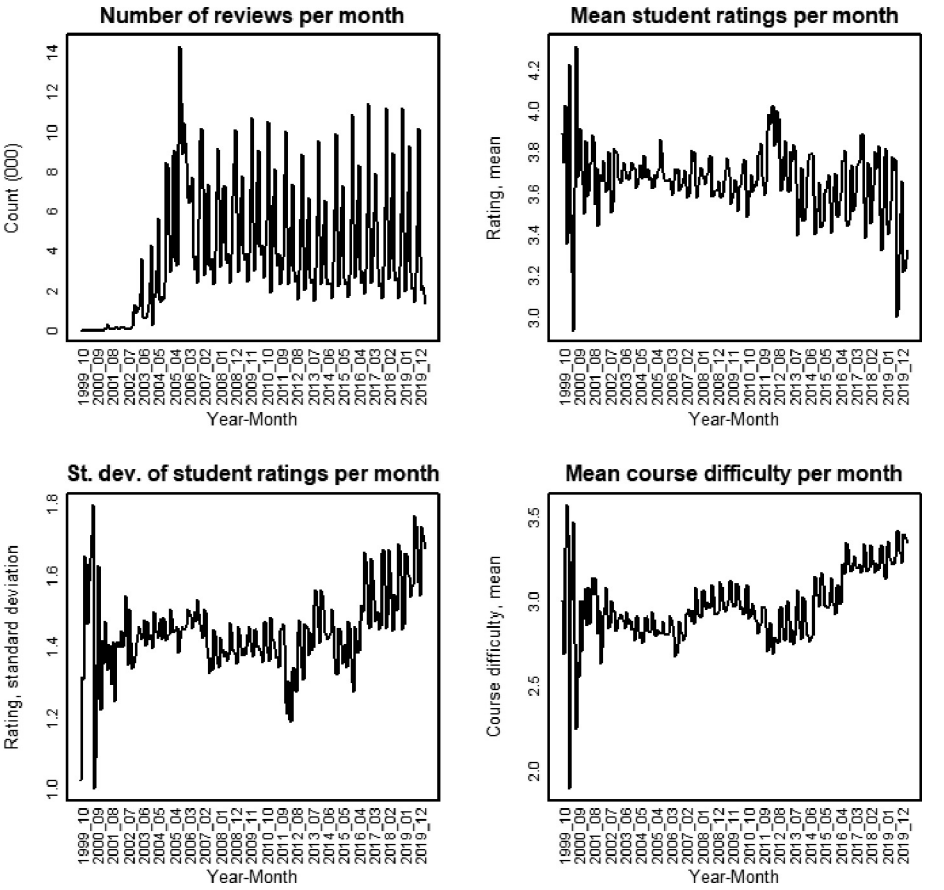

Figure 1.
Monthly statistics of student reviews scraped from the ratemyprofessors.com website

than the most difficult ones, at 93,377 against 14,264, the highest number of best course ratings of 5 were given to 100,266 courses with a medium level of difficulty of 3.

2.3 Extracting the student emotion and bad professor performance features with natural language processing models

This paper uses two NLP models based on deep neural networks to extract features from student reviews. In both cases, the neural network model is DistilBERT, which was introduced in Sanh *et al.* (2020). The model has 66 million parameters and is based on the larger BERT model (Devlin *et al.*, 2019). Its number of parameters is reduced by 40% from the BERT model, and it is 60% faster but retains 97% of BERT's accuracy. In simplified terms, the original BERT model was trained as follows. First, the text is split into words, which are converted into numbers or tokens. This stage is called tokenisation. The model uses a vocabulary of 30,000 word-token pairs. These numbers are then converted into embeddings, which are numerical vectors with a length of 768, and they are fed into the neural network. Then 15% of the tokens are randomly masked, and the neural network is trained to predict the masked tokens. This process is called self-supervised learning. Each

model prediction is compared with the masked tokens, the loss function is computed and backpropagation is used to recalculate the model weights based on the value of the loss function. Training is conducted on the BookCorpus (Zhu *et al.*, 2015), containing 800 million words, and on English Wikipedia, containing 2.5 billion words.

As a result of the training process, BERT learns contextual embeddings for words, meaning that words with similar meanings also have similar embedding vectors. If we calculate, for example, the expression for the word embeddings "king" – "man" + "woman", then the result will be the embedding vector representing the word "queen". In this way, the model learns the representation of the English language. The model can then be fine-tuned or trained on a much smaller corpus for specific language tasks, such as sentiment analysis or feature extraction, as explained below.

2.3.1 Model 1 for extracting student emotions. The DistilBERT model introduced in Sanh *et al.* (2020) was fine-tuned on a corpus of Twitter posts to detect six types of emotion: sadness, joy, love, anger, fear, and surprise. An emotion label describes each tweet, and the model learned to predict these labels as a supervised learning task. The corpus of tweets is available at https://huggingface.co/datasets/viewer/?dataset=emotion and the model is available at https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion. Each student review is treated as a mix of these six emotions. Model 1 calculates the probabilities of the text representing each of the emotions. These probabilities, which sum to one, are used in the subsequent analysis presented in the Estimation methods section. It should be noted that emotions in the text are conveyed through various means, including the choice of words, sentence structure, overall text structure, and even capitalisation and punctuation. During the fine-tuning process, the model learns to associate these text properties with specific emotions, as expressed through emoticons in Twitter messages. As a result, the fine-tuned neural network weights (parameters) are dependent on the corpus on which the model was fine-tuned.

Table 1 in the Supplementary materials provides random examples of student reviews and the emotion probabilities assigned by model 1

There are several advantages to detecting emotions from text, as opposed to just detecting sentiment:

- Greater granularity: Emotion detection allows for a more fine-grained analysis of text, as it can identify specific emotions rather than just the overall sentiment. This can provide a more detailed understanding of the emotional content of a text.
- More meaningful insights: By identifying specific emotions, it is possible to gain a deeper understanding of the attitudes and feelings of the writer.
- More accurate analysis: In some cases, detecting emotion may be more accurate than just detecting sentiment. For example, a text that expresses both positive and negative emotions may be difficult to classify using sentiment analysis, but emotion detection could identify both the positive and negative emotions present.
- Greater flexibility: Emotion detection algorithms can be customised to identify specific emotions that are relevant to a particular application or domain. This allows for greater flexibility in adapting the analysis to specific needs.

Model 1 was applied to extract six basic emotions from students' textual reviews, and the following research hypothesis was formulated:

H1. Negative emotions of students that are extracted by the natural language processing models are associated with lower course ratings and higher course

difficulty scores. Positive emotions are associated with higher course ratings and lower course difficulty scores.

2.3.2 Model 2 for extracting bad performance by the professors. The DistilBERT model (Sanh *et al.*, 2020) was fine-tuned on a multi-nli corpus that can be used for zero-shot classification of a text for any set of labels. The model is available at https://huggingface.co/typeform/distilbert-base-uncased-mnli; the data corpus is available at https://huggingface.co/datasets/multi_nli. Zero-shot classification means that the model has a very good understanding of the language and can perform a given text classification task without being trained on a domain-specific corpus. Model 2 is used to detect bad performance by the teaching faculty. The following labels describe bad performance by teachers: "poor or bad teaching skills of the professor", "unfair or improper grading", "poor or bad interpersonal skills or qualities of the professor", "extremely or very difficult tests or exams", "harassment or abuse or improper conduct" and "review or opinion on the course or professor". The last label represents a general comment about teaching faculty that does not fit the five previous examples of bad teaching. Model 2 assigns one of these labels to each student review using the highest probability, and this dummy variable is later used in the econometric analysis. Table 2 in the supplementary materials provides examples of student reviews and the probabilities assigned by Model 2 to all the categories of bad performance by teachers.

Model 2 was applied to extract professors' bad performances from students' textual reviews, and the following research hypothesis was formulated:

H2. Bad performances of professors that are extracted by the natural language processing models are associated with lower course ratings and higher course difficulty scores.

3. Estimation method

Course/professor ratings and perceived course difficulty are used separately as dependent variables. The independent variables are the emotion scores and the bad performance scores. Because of the multicollinearity shown in Figure 2, one emotion variable "joy" and one dummy variable for "general commentary" are dropped, leaving ten independent variables. The model also includes year and month dummy variables. Regressions for emotions and bad performance are estimated separately, as specified below.

The formulation of regression equations requires justification. Student emotions formed during the course and professor actions during the course influence the overall course rating, which explains the structure of equations (1) and (2). Similarly, evidence of deficient performance by a professor, such as a boring teaching style, may lead to a perception that the course is difficult, as postulated in equation (4). But the causality in equation (3) cannot be readily established. In this case, the interpretation would be what emotion patterns are typical for courses perceived as difficult or easy:

$$\text{course\_rating}_i = \alpha + \beta_1 \text{anger}_i + \beta_2 \text{fear}_i + \beta_3 \text{love}_i + \beta_4 \text{sadness}_i + \beta_5 \text{surprise}_i + \varepsilon_i \quad (1)$$

$$\text{course\_rating}_i = \alpha + \beta_1 \text{bad\_teaching}_i + \beta_2 \text{poor\_skills}_i + \beta_3 \text{unfair\_grading}_i + \beta_4 \text{harassment}_i + \beta_5 \text{diff\_exams}_i + \varepsilon_i \quad (2)$$

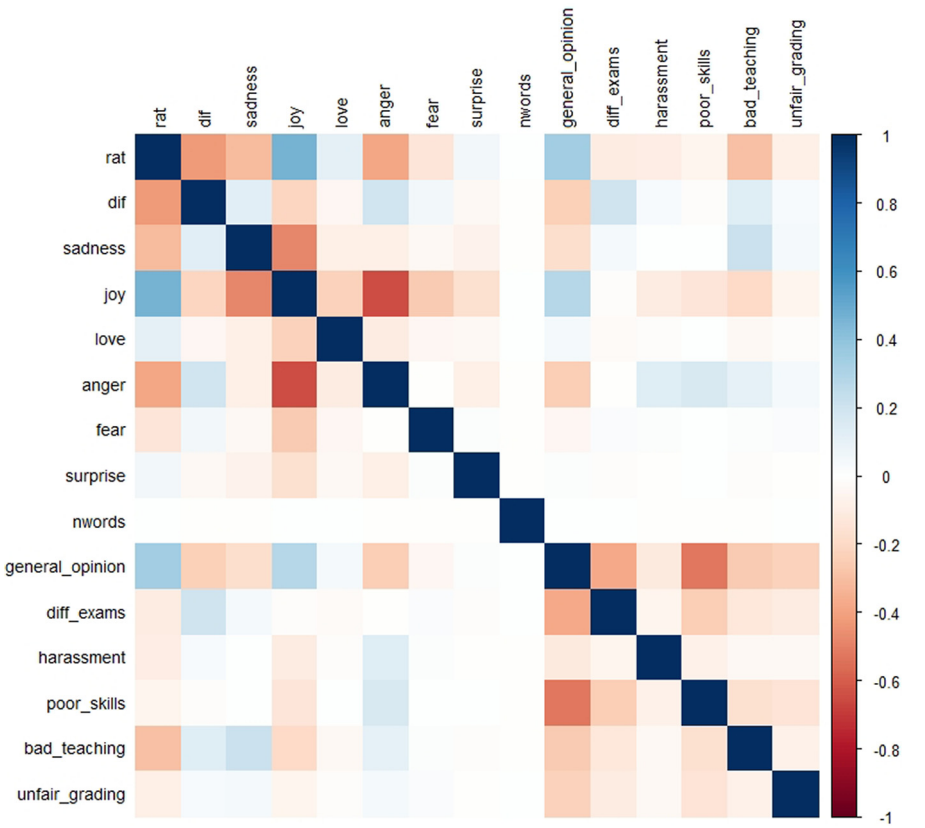

Figure 2. Heatmap with Pearson correlations for the student review rating (rat), the perceived course difficulty (dif), and the derived student emotions and bad performance by professors. $N = 975{,}860$

$$\text{course\_difficulty}_i = \alpha + \beta_1 \text{anger}_i + \beta_2 \text{fear}_i + \beta_3 \text{love}_i + \beta_4 \text{sadness}_i + \beta_5 \text{surprise}_i + \varepsilon_i \quad (3)$$

$$\text{course\_difficulty}_i = \alpha + \beta_1 \text{bad\_teaching}_i + \beta_2 \text{poor\_skills}_i + \beta_3 \text{unfair\_grading}_i + \beta_4 \text{harassment}_i + \beta_5 \text{diff\_exams}_i + \varepsilon_i \quad (4)$$

An extended version of Models 1 and 2 has also been estimated, with $\text{course\_difficulty}_i$ added as an independent variable to account for the course difficulty bias documented in the literature.

In equations (1)–(4), course_score_i and $\text{course\_difficulty}_i$ are the numerical values of the course/professor rating and course difficulty score for the ith student review. anger_i is the anger score for the ith student review, and other student emotions are treated similarly. bad_teaching_i is the dummy variable equal to 1 if the ith student review was classified as revealing bad teaching performance, that is, Model 2 assigned the highest probability to the category "poor or bad teaching skills of the professor". Similar coding applies for the other categories of bad performance. i = 1,2, . . ., 975,860.

4. Results

4.1 How student emotions and bad performance by faculty are related to the teacher rating and course difficulty

The estimated linear regression models [equations (1)–(4)] positively validate *H1* and *H2* research hypotheses. As Table 3 shows, all the emotions and bad performance are significant and have the expected signs. High scores for "anger" (estimate and significance of −1.765***), "fear" (−1.621***) and "sadness" (−1.879***) reduce the course rating and high scores for "love" (0.262***) improve the rating. "Surprise" also reduces the course rating (−0.036***), but there were no prior expectations about the sign. All five bad performance dummy variables reduce the course rating, and "bad teaching" has the highest impact at −2.027***. It means that when the course/professor review is classified as "bad teaching" performance, the course/professor score falls on average by more than two points on a 1–5 scale, which is a considerable distance to fall.

Interestingly, the second-highest impact is for "harassment" at −1.586***. Among the most common words in the reviews classified as harassment are "rude", "condescending", "hate" and "insult". When students experience this type of bad behaviour by a professor, they seem to punish that professor dearly in their reviews. The adjusted R^2 in course rating regressions is 0.29 for student emotions and 0.17 for professor harmful performance. It is quite surprising that student emotions and professors' bad behaviour extracted from the textual review can explain such a significant part of the course rating variance, given that no other features related to courses, professors or students characteristics are included in the regressions.

Additionally, it is possible to test the sensitivity of the results with respect to the course difficulty bias. Students tend to give higher evaluations to a professor when they receive good grades in the course. To examine whether this student grading/course difficulty bias affects the reported results, the course difficulty score was included as an explanatory variable in regressions with course/professor rating as the dependent variable. These results are reported in Table 3, and Models 1a and 2a should be compared with Models 1 and 2, respectively. The course difficulty/student grading bias is clearly present in the data, with courses perceived as more difficult being rated lower by 0.3–0.4 points for every one-point increase in perceived course difficulty.

However, even when controlling for this bias, the estimates for student emotions and professor behaviours maintain the same sign and remain highly significant in Models 1a and 2a. The magnitude of all estimates, except for the surprise emotion, is somewhat lower, though.

This robustness exercise demonstrates that the relationships between student emotions, professors' poor performance and course ratings are robust and are not affected by course difficulty bias.

Emotions and harmful performance are less helpful for explaining course difficulty, as the adjusted R^2 of these regressions is lower at 0.06–0.08. Emotions produce a better fit for the course/professor rating, while bad performances perform marginally better in the course difficulty regressions. All estimates have the expected signs. Negative emotions and improper professor behaviour are correlated with courses perceived as more difficult, and high values of "love" emotion indicate less difficult courses. The magnitude of estimates is much smaller than in the course rating regressions.

5. Discussion

This paper presents a novel analysis of how student emotions and poor performance by professors, automatically derived from student evaluations of teachers, affect course ratings.

Recently developed advanced NLP models, which have demonstrated excellent performance in many text processing tasks (Acheampong *et al.*, 2021), were used to extract these features from almost one million textual student reviews.

Both research hypotheses stated in this paper are positively validated. Negative emotions and poor performance by professors were detected more frequently for low-rated courses and courses that are perceived as more challenging by students. Positive emotions were associated with highly rated and less challenging courses. These results demonstrate that the NLP techniques applied to extract both types of features from student reviews are valid and can provide reliable insights.

The strong link documented in this paper between student emotions and the bad performance of teachers on one side and course ratings on the other calls for further research. Araujo *et al.* (2018) show that children with caregivers who demonstrate higher-quality interactions have significantly better communication, problem-solving and fine motor skills. It would be important and informative to verify how the student emotions and the bad performance by professors derived from student textual reviews affect the development, learning outcomes and subsequent professional careers of students. While it would be difficult to analyse at the individual level, given the anonymity of student feedback, such analysis could be done for aggregated data.

The automated feature extraction methods presented in this paper can be used by students to make better and more informed choices about their courses. Dillon and Smith (2019) find a strong complementarity in degree completion between student abilities and college quality. Rocconi *et al.* (2020) document that students flourish in academic environments that match their personality, while de Boer and Van Rijnsoever (2022) argue for extending the student selection process by testing their personality traits. One possible hypothesis along this line of research would be that the student–course emotional match or student tolerance of bad performance by professors can affect the performance of students and their chances of completing their degree.

Students who have negative feelings about their course or are irritated by the improper performance of their professor are more likely to attend the class irregularly or to drop out of the course, although such a decision also depends on student characteristics such as Big Five personality traits. As Liu and Loeb (2021) show, the ability of teachers to engage students in class and to show up for class improves the likelihood of the students graduating. It is essential information for at-risk students who may fail their course because of their negative feelings or improper performance of their professor. The current systems for evaluating faculty completely ignore this dimension.

Additionally, the model with emotion and professors' activity variables jointly included as independent variables was estimated (Table 3 in Supplementary materials). The estimates maintain the same sign and are significant.

The models presented in this paper have one obvious limitation. They do not include controls for the student's year of study or gender, for the course type or for the professor's characteristics of age, position, ethnicity and so forth. The literature discussed has shown that the numerical course ratings posted on ratemyprofessors.com exhibit many biases, with minority faculty receiving lower ratings, for example. The features for student emotions and bad performance by professors are extracted from the textual reviews, and it is possible that some textual student reviews are affected by bias. Unlike the numerical ratings, though, which by their construction are general, the textual reviews are specific, refer to what happened and give particular examples of bad performance by professors, as shown in Tables 1 and 2 in the Supplementary materials (e.g. "Was misgraded a few times" and "She spend most of the class telling stories barely related the course description"). It makes the

likelihood of there being bias in the student emotions extracted or the bad performance identified very low. However, the regressions use the numerical course ratings that exhibit such bias as dependent variables. Research has shown that these biases are statistically significant but their magnitude is small. For example, a dummy indicating that the professor is white raises the professor's average rating by 0.118 over that of a non-white professor (Murray *et al.*, 2020). The magnitude of other effects, like the professor's accent, rank or course type, never exceeds 0.2. These bias effects are small next to the effects of student emotions or bad performance by professors, which in absolute terms are in the range of 0.5–2.7, so they are larger by order of magnitude. Therefore, the regressions without controlling for professor, course or student characteristics generate reliable estimates.

6. Conclusions

The paper demonstrates that advanced NLP tools can enhance the often-contested SET practices, help students make better and more informed choices about their courses and assist instructors in better tailoring their teaching approaches to create a more positive learning environment for their students. This paper also identifies several avenues for future research. Multilingual NLP models such as those presented in this paper can understand over 100 languages and are available in an open-source format, meaning that every university has the opportunity to collect textual SET data and use zero-shot learning to analyse student emotions, poor teaching performance and other relevant factors. University SET data can be enhanced with characteristics of professors and courses, and Models 1–4 can be fine-tuned to control for many of the biases documented in the literature. The zero-shot learning method can also be applied to extract other useful features from students'

Table 1.
Data summary of the student reviews scraped from the website ratemyprofessors.com

Data characteristics	Value(s)
Total number of scraped reviews	975,860
Review length in alphanumeric words: minimum/median/maximum/standard deviation	0/44/176/21
Number of universities or colleges with at least 20 reviews	2,306
No of student reviews for each university or college: minimum/median/maximum	1/92.5/8712

Note: Authors' own work

Table 2.
Distribution of reviews by course/teacher rating and course perceived difficulty

Rating ↓/Perceived difficulty →	1	2	3	4	5
1	8,623	5,972	18,042	26,646	61,484
1.5	2,787	3,446	8,155	12,203	12,615
2	5,178	6,589	15,221	22,930	15,713
2.5	3,504	5,587	9,866	11,779	7,377
3	7,540	11,665	22,500	20,855	8,979
3.5	6,934	12,534	14,673	11,892	4,576
4	16,584	32,250	39,694	28,486	7,723
4.5	18,398	31,016	32,734	20,958	5,772
5	93,377	86,961	100,266	61,512	14,264

Note: Authors' own work

Table 3. Results of linear regressions

Independent variables	Dependent variable - course rating Model 1 nobs = 975,860 Estimate		Standard error	Dependent variable - course rating Model 1a nobs = 975,860 Estimate		Standard error	Dependent variable - course difficulty Model 3 nobs = 975,860 Estimate		Standard error
(Intercept)	4.211	***	0.002	5.164	***	0.003	2.746	***	0.002
anger	−1.765	***	0.004	−1.497	***	0.004	0.774	***	0.004
fear	−1.621	***	0.012	−1.402	***	0.011	0.632	***	0.010
love	0.262	***	0.006	0.241	***	0.006	−0.062	***	0.007
sadness	−1.879	***	0.005	−1.646	***	0.005	0.673	***	0.005
surprise	−0.036	***	0.010	−0.066	***	0.009	−0.085	***	0.010
course_difficulty				−0.347	***	0.001			
m_02							−0.010	•	0.006
m_05	0.011	**	0.004	0.010	**	0.003			
m_08	0.016	**	0.005	0.014	**	0.005			
y_2000	−0.376	**	0.121	−0.341	**	0.106			
Adj. R2	0.2903			0.3804			0.0603		
	Model 2			Model 2a			Model 4		
(Intercept)	4.205	***	0.002	5.235	***	0.003	2.644	***	0.002
bad_teaching	−2.027	***	0.005	−1.658	***	0.005	0.948	***	0.005
poor_skills	−0.694	***	0.003	−0.577	***	0.003	0.300	***	0.003
unfair_grading	−1.065	***	0.006	−0.866	***	0.006	0.511	***	0.006
harassment	−1.586	***	0.012	−1.340	***	0.011	0.631	***	0.011
diff_exams	−0.901	***	0.004	−0.533	***	0.004	0.944	***	0.004
course_difficulty				−0.390	***	0.001			
m_02							−0.010	•	0.006
m_05	0.015	***	0.004	0.013	***	0.004			
m_08	0.016	**	0.006	0.014	**	0.005			
y_2000	−0.476	***	0.133	−0.427	***	0.116			
Adjusted *R*2	0.1721			0.2828			0.0834		

Notes: [1] the unit of observation is a single student review; [2]*p*-value significance levels: *** < 0.001, **< 0.01; *< 0.05, •< 0.1; *p*-values based on robust standard errors; [3]m_0* – denotes the monthly dummy for February, May and August; y_2000 is a dummy variable for student reviews posted in the year 2000; Authors' own work

textual reviews. Finally, statistical inference models may be used to investigate causal relationships between professor behaviours and student emotions.

References

Abbas, A., Haruna, H., Arrona-Palacios, A., Camacho-Zuñiga, C., Núñez-Daruich, S., Enríquez de la O, J. F., Castaño-Gonzalez, R., Escamilla, J. and Hosseini, S. (2022), "Students' evaluations of teachers and recommendation based on course structure or teaching approaches: an empirical study based on the institutional dataset of student opinion survey", *Education and Information Technologies*, doi: 10.1007/s10639-022-11119-z.

Acheampong, F., Nunoo-Mensah, H. and Chen, W. (2021), "Transformer models for text-based emotion detection: a review of BERT-based approaches", *Artificial Intelligence Review*, Vol. 54 No. 8, pp. 5789-5829, doi: 10.1007/s10462-021-09958-2.

Ackerman, D. and Chung, C. (2017), "Is RateMyProfessors.com unbiased? A look at the impact of social modelling on student online reviews of marketing classes", *Journal of Marketing Education*, Vol. 40 No. 3, pp. 188-195, doi: 10.1177/0273475317735654.

Adams, S., Bekker, S., Fan, Y., Gordon, T., Shepherd, L., Slavich, E. and Waters, D. (2021), "Gender bias in student evaluations of teaching: 'Punish[ing] those who fail To do their gender right'", *Higher Education*, Vol. 83 No. 4, pp. 787-807, doi: 10.1007/s10734-021-00704-9.

Araujo, M., Dormal, M. and Schady, N. (2018), "Childcare quality and child development", *Journal of Human Resources*, Vol. 54 No. 3, pp. 656-682, doi: 10.3368/jhr.54.3.0217.8572r1.

Azab, M., Mihalcea, R. and Abernethy, J. (2016), "Analysing RateMyProfessors evaluations Across institutions, disciplines, and cultures: the Tell-Tale signs of a good professor", in Spiro, E. and Ahn, Y-Y. (Eds). *Social Informatics*, Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 438-453.

Boehmer, D. and Wood, W. (2017), "Student vs. faculty perspectives on quality instruction: Gender bias, 'hotness', and 'easiness' in evaluating teaching", *Journal of Education For Business*, Vol. 92 No. 4, pp. 173-178, doi: 10.1080/08832323.2017.1313189.

Boswell, S. and Sohr-Preston, S. (2020), "I checked the prof on ratemyprofessors: effect of anonymous, online student evaluations of professors on students' self-efficacy and expectations", *Social Psychology of Education*, Vol. 23 No. 4, pp. 943-961, doi: 10.1007/s11218-020-09566-y.

Chiang, K. (2017), "Students' perspectives on RateMyProfessors.com: an empirical investigation of perception and attitude", *International Journal of Social Media And Interactive Learning Environments*, Vol. 5 No. 1, pp. 21-31, doi: 10.1504/ijsmile.2017.10006127.

Chou, S., Luo, J. and Ramser, C. (2020), "High-quality vs low-quality teaching", *Journal of International Education In Business*, Vol. 14 No. 1, pp. 93-108, doi: 10.1108/jieb-01-2020-0007.

Chung Sea Law, D. (2010), "Quality assurance in post-secondary education: the student experience", *Quality Assurance in Education*, Vol. 18 No. 4, pp. 250-270, doi: 10.1108/09684881011079125.

Crumbley, L., Henry, B.K. and Kratchman, S.H. (2001), "Students' perceptions of the evaluation of college teaching", *Quality Assurance in Education*, Vol. 9 No. 4, pp. 197-207, doi: 10.1108/EUM0000000006158.

Cunningham-Nelson, S., Baktashmotlagh, M. and Boles, W. (2019), "Visualizing student opinion Through text analysis", *IEEE Transactions on Education*, Vol. 62 No. 4, pp. 305-311, doi: 10.1109/te.2019.2924385.

Cunningham-Nelson, S., Laundon, M. and Cathcart, A. (2021), "Beyond satisfaction scores: visualising student comments for whole-of-course evaluation", *Assessment and Evaluation in Higher Education*, Vol. 46 No. 5, pp. 685-700, doi: 10.1080/02602938.2020.1805409.

Darwin, S. (2020), "From the local fringe to market Centre: analysing the transforming social function of student ratings in higher education", *Studies in Higher Education*, Vol. 46 No. 9, pp. 1978-1990, doi: 10.1080/03075079.2020.1712690.

de Boer, T. and Van Rijnsoever, F. (2022), "In search of valid non-cognitive student selection criteria", *Assessment and Evaluation in Higher Education*, Vol. 47 No. 5, pp. 783-800, doi: 10.1080/02602938.2021.1958142.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019), "BERT: Pre-Training of deep bidirectional transformers for language understanding", ArXiv, doi: 10.48550/arXiv.1810.04805

Dillon, E. and Smith, J. (2019), "The consequences of academic match between students and colleges", *Journal of Human Resources*, Vol. 55 No. 3, pp. 767-808, doi: 10.3368/jhr.55.3.0818-9702r1.

Emery, C.R., Kramer, T.R. and Tian, R.G. (2003), "Return to academic standards: a critique of student evaluations of teaching effectiveness", *Quality Assurance in Education*, Vol. 11 No. 1, pp. 37-46, doi: 10.1108/09684880310462074.

Feld, J., Salamanca, N. and Zolitz, U. (2020), "Are professors worth it? The Value-Added and costs of tutorial instructors", *Journal of Human Resources*, Vol. 55 No. 3, pp. 836-863, doi: 10.3368/jhr.55.3.0417-8752R2.

Felton, J., Koper, P., Mitchell, J. and Stinson, M. (2008), "Attractiveness, easiness and other issues: student evaluations of professors on ratemyprofessors.com", *Assessment and Evaluation in Higher Education*, Vol. 33 No. 1, pp. 45-61, doi: 10.1080/02602930601122803.

Flegl, M. and Rosas, L.A.A. (2019), "Do professor's age and gender matter or do students give higher value to professors' experience?", *Quality Assurance in Education*, Vol. 27 No. 4, pp. 511-532, doi: 10.1108/QAE-12-2018-0127.

Goetz, T., Lüdtke, O., Nett, U., Keller, M. and Lipnevich, A. (2013), "Characteristics of teaching and students' emotions in the classroom: Investigating differences across domains", *Contemporary Educational Psychology*, Vol. 38 No. 4, pp. 383-394, doi: 10.1016/j.cedpsych.2013.08.001.

Hanushek, E., Piopiunik, M. and Wiederhold, S. (2019), "The value of smarter teachers. International evidence on teacher cognitive skills and student performance", *Journal of Human Resources*, Vol. 54 No. 4, pp. 857-899, doi: 10.3368/jhr.54.4.0317.8619R1.

Heffernan, T. (2021), "Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching", *Assessment and Evaluation in Higher Education*, Vol. 47 No. 1, pp. 144-154, doi: 10.1080/02602938.2021.1888075.

Kreitzer, R. and Sweet-Cushman, J. (2021), "Evaluating student evaluations of teaching: a review of measurement and equity bias in SETs and recommendations for ethical reform", *Journal of Academic Ethics*, Vol. 20 No. 1, pp. 73-84, doi: 10.1007/s10805-021-09400-w.

Linse, A. (2017), "Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees", *Studies in Educational Evaluation*, Vol. 54, pp. 94-106, doi: 10.1016/j.stueduc.2016.12.004.

Liu, J. and Loeb, S. (2021), "Engaging teachers: Measuring the impact of teachers on student attendance in secondary school", *Journal of Human Resources*, Vol. 56 No. 2, pp. 343-379, doi: 10.3368/jhr.56.2.1216-8430R3.

Mainhard, T., Oudman, S., Hornstra, L., Bosker, R. and Goetz, T. (2018), "Student emotions in class: the relative importance of teachers and their interpersonal relations with students", *Learning and Instruction*, Vol. 53, pp. 109-119, doi: 10.1016/j.learninstruc.2017.07.011.

Murray, D., Boothby, C., Zhao, H., Minik, V., Bérubé, N., Larivière, V. and Sugimoto, C. (2020), "Exploring the personal and professional factors associated with student evaluations of tenure-track faculty", *Plos One*, Vol. 15 No. 6, p. e0233515, doi: 10.1371/journal.pone.0233515.

Okoye, K., Arrona-Palacios, A., Camacho-Zuñiga, C., Achem, J.A.G., Escamilla, J. and Hosseini, S. (2022), "Towards teaching analytics: a contextual model for analysis of students' evaluation of teaching through text mining and machine learning classification", *Education and Information Technologies*, Vol. 27, pp. 3891-3933, doi: 10.1007/s10639-021-10751-5.

Onan, A. (2019), "Mining opinions from instructor evaluation reviews: a deep learning approach", *Computer Applications in Engineering Education*, Vol. 28 No. 1, pp. 117-138, doi: 10.1002/cae.22179.

Park, M. (2019), "What's important: an exploratory analysis of student evaluations About physics professors on RateMyProfessors.com", *Journal of College Science Teaching*, Vol. 48 No. 4, pp. 36-44, available at: www.nsta.org/whats-important-exploratory-analysis-student-evaluations-about-physics-professors

Pounder, J. (2014), "Quality teaching through transformational classroom leadership", *Quality Assurance in Education*, Vol. 22 No. 3, pp. 273-285, doi: 10.1108/QAE-12-2013-0048.

Ran, F.X. and Xu, D. (2019), "Does contractual form matter? The impact of different types of non-tenure track faculty on college students' academic outcomes", *Journal of Human Resources*, Vol. 54 No. 4, pp. 1081-1120, doi: 10.3368/jhr.54.4.0117.8505R.

Reber, J.S., Ridge, R.D. and Downs, S.D. (2017), "Perceptual and behavioral effects of expectations formed by exposure to positive or negative ratemyprofessors.com evaluations", *Cogent Psychology*, Vol. 44 No. 1, p. 1338324, doi: 10.1080/23311908.2017.1338324.

Reid, L.D. (2010), "The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com", *Journal of Diversity in Higher Education*, Vol. 3 No. 3, pp. 137-152, doi: 10.1037/a0019865.

Rocconi, L.M., Liu, X. and Pike, G.R. (2020), "The impact of Person-Environment fit on grades, perceived gains, and satisfaction: an application of holland's theory", *Higher Education*, Vol. 80 No. 5, pp. 857-874, doi: 10.1007/s10734-020-00519-0.

Rosen, A. (2018), "Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors.com data", *Assessment and Evaluation In Higher Education*, Vol. 43 No. 1, pp. 31-44, doi: 10.1080/02602938.2016.1276155.

Rybinski, K. and Kopciuszewska, E. (2021), "Will artificial intelligence revolutionise the student evaluation of teaching? A big data study of 1.6 million student reviews", *Assessment and Evaluation in Higher Education*, Vol. 46 No. 7, pp. 1127-1139, doi: 10.1080/02602938.2020.1844866.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2020), "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter", ArXiv, doi: 10.48550/arXiv.1910.01108.

Shah, M., Cheng, M. and Fitzgerald, R. (2017), "Closing the loop on student feedback: the case of Australian and Scottish universities", *Higher Education*, Vol. 74 No. 1, pp. 115-129, doi: 10.1007/s10734-016-0032-x.

Spooren, P., Brockx, B. and Mortelmans, D. (2013), "On the validity of student evaluation of teaching", *Review of Educational Research*, Vol. 83 No. 4, pp. 598-642, doi: 10.3102/0034654313496870.

Stonebraker, R. and Stone, G. (2015), "Too old to teach? The effect of age on college and university professors", *Research in Higher Education*, Vol. 56 No. 8, pp. 793-812, doi: 10.1007/s11162-015-9374-y.

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*, Vol. 44 No. 1, pp. 35-62, doi: 10.1017/s0047404514000736.

Subtirelu, N.C. (2017), "Exploring the intersection of gender and race in evaluations of mathematics instructors on ratemyprofessors.Com", in Friginal, E. (Ed.), *Studies in Corpus-Based Sociolinguistics*, Routledge, New York, available at: www.taylorfrancis.com/chapters/edit/10.4324/9781315527819-9/exploring-intersection-gender-race-evaluations-mathematics-instructors-ratemyprofessors-com-nicholas-close-subtirelu

Toth, Z. and Bedzsula, B.P. (2021), "What constitutes quality to students in higher education? An empirical investigation of course-level student expectations", *Quality Assurance in Education*, Vol. 29 Nos 2/3, pp. 116-134, doi: 10.1108/QAE-07-2020-0088.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S.R. (2019), "SuperGLUE: a stickier benchmark for General-Purpose language understanding systems", *arXiv*, 1905.00537. 10.48550/arXiv.1905.00537.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A. and Fidler, S. (2015), "Aligning books and movies: towards story-like visual explanations by watching movies and reading books", arXiv, 10.48550/arXiv.1506.06724.

Supplementary material

The supplementary material for this article can be found online.

Corresponding author

Krzysztof Rybinski can be contacted at: rybinski@rybinski.eu