Emma-Leigh Hodge
The University of Waikato
Rahat Hasan
The University of Waikato
Eden Poihipi
The University of Waikato
Rebecca Barker
The University of Waikato

MOTIVATING STUDENTS AND STAFF

The Robots are Coming for your Students' Feedback

The Data Analytics team at The University of Waikato gathers student feedback (as rich qualitative data) but manual analysis of these comments poses a time challenge for reporting. To address this, we explored the possibility of condensing qualitative information by leveraging natural language processing (NLP) technology, specifically Google's NLP sentiment analysis. We employed a robust coding framework to test the validity of NLP-coded student feedback, analysing 1000 comments from the University's 2021 course evaluations. Results show a statistical correlation between our sentiment analysis and NLP, offering promising evidence for NLP's efficacy in providing accurate, high-level insights into student feedback sentiment.

DOI 10.18573/conf2.g

Hodge, E-L., Hasan, R., Poihipi, E., and Barker, R. 2024. The Robots are Coming for your Students' Feedback. In: Bowell, T., Pepperell, N., Richardson, A., and Corino, M-T. (eds.). *Revitalising Higher Education: Insights from Te Puna Aurei LearnFest 2022.* Cardiff: Cardiff University Press, pp. 56-62.

Introduction

As the Evaluations and Data Analytics team at the University of Waikato, we collect student feedback (i.e. evaluations) to enhance teaching practices and our students' learning. Some student feedback is collected as quantitative ratings of Likert-scale questions, measuring various aspects of the teaching and learning experience, such as whether a paper (i.e. a course) was well-organised or the learning objectives were reached (University of Waikato 2023). Quantitative rating data of this sort can be readily described in aggregate or presented in graphics. However, a substantial portion of students' feedback comprises detailed qualitative data that is less easily communicated en masse (Weis & Willems 2017: 223-243). While our quantitative analyses of student feedback provide staff with useful information about how a paper has gone overall, it tends to be students' qualitative responses that provide the most insight about specific things that could be improved. Furthermore, students' critical comments about a course often provide the starting point for thinking through how the required improvement could be achieved (e.g. Tucker 2015).

Although student feedback can provide individual staff with invaluable insight into their teaching practices (Tucker 2015; Zaitseva, Tucker & Santhanam 2022), the substantial time investment required for the careful and manual analysis of these comments at scale has thus far significantly limited the ability to aggregate insights from students' qualitative feedback. For our team's specific circumstances, condensing and aggregating the qualitative data would be an advantageous first step for three key reasons. First, it enables information from student evaluations to be more easily communicated to those not directly involved in the delivery of a course without sacrificing student confidentiality. For example, if multiple students in a course, or across courses in a programme, are raising similar concerns about specific teaching quality issues, it may be important for a supervisory staff member to see that professional development or other support is needed. Second, the volume of data and the in-depth qualitative analysis (e.g. thematic analysis) required to synthesise insights from qualitative evaluations data outweigh the staffing resources generally available to central evaluations teams. In recent years at Waikato, for example, we have received up to 40,000 pieces of qualitative feedback from our three main evaluation periods each year. Finally, the literature underscores the importance of university processes that are responsive to students' experiences (both good and bad) and recognise that students are active contributors who enhance the quality of education at such institutions (e.g. Knight et al. 2022; Mertens 2019). The best practice for engaging students in the process of evaluation and improving an institution's teaching and learning systems is through 'closing the loop'. In this context, closing the loop means collecting, responding to, and communicating a summary of their feedback and the subsequent institutional responses to the student body (Tschirhart & Pratt-Adams 2019).

When communicating the findings of student evaluations, it is important that student confidentiality is maintained and that data is presented succinctly. One common method for representing open-ended feedback and maintaining confidentiality is employing a text analytics word cloud, particularly when many student responses are available. However, word clouds tend to be based on the frequency at which keywords were used in feedback, which means that the valence behind those words is not always clear (e.g. Knight et al. 2022). Further, students may misspell terms or use colloquialisms that pose a problem for frequency-based measures (Kukich 1992; Sag et al. 2002).

An alternative option is to conduct an analysis of each piece of feedback to determine the overall sentiment—in other words, carry out a sentiment analysis (Zhang et al. 2023). While sentiment analysis can be done manually, artificial intelligence (AI) tools have become available that enable sentiment analysis using Natural Language Processing (NLP). Existing literature suggests that NLP can provide a resource-efficient approach that is increasingly being used to analyse textual content from students (e.g. Graesser & McNamara 2012; Ormerod & Harris 2010; Ormerod, Patel & Wang 2023). Popular evaluation software packages also include AI-powered text analytics to condense qualitative data (e.g. *Blue: The people insights platform* (Explorance 2023); *NVivo* (Lumivero 2023)).

Current Study

At our university (University of Waikato) we run frequent evaluations of the papers offered and of the contributions of teaching staff. Among other questions, the evaluations include open-ended questions to students about what they think was done well and what could be improved. Student responses to the open-ended questions often contain useful feedback but are difficult and time-consuming to aggregate for reporting purposes and aggregation is essential for retaining student confidentiality. We compared human-versus-NLP sentiment analyses to investigate the viability of leveraging NLP to revitalise our evaluations reporting. Therefore, with the assistance of a Bluenotes Faculty Grant and funding from the University of Waikato's Summer Research Scholar programme, we proposed that NLP sentiment analysis could serve as a first-step summary of student responses. Furthermore, we sought to identify an analytic process that would require minimal coding or pre-training so that, if successful, we could widely disseminate the approach to other educational institutions.

Method

Test Population
We opted to analyse the feedback provided by students during the two largest trimesters in 2021. The total population consisted of 35,925 pieces of student feedback. For both

trimesters, there was a larger number of responses regarding feedback of the paper (A = 7,692; B = 16,357) compared to feedback about individual teaching staff (*A* = 4,549; *B* = 7,327). Within these subsets of feedback data (e.g. *A Trimester and Paper Feedback*), the positively-primed question was answered more often than the corresponding negatively primed question, as shown in Table 1. Broadly, the positively-primed question asks students what they thought was done well in a paper or by a teacher, while the negatively-primed question asked students to reflect on things that could be improved upon.

	A Trimester		B Trimester	
	Positive Question	Negative Question	Positive Question	Negative Question
Paper	4017	3675	8499	7858
Teacher	2535	2014	4073	3254

Table 1. Frequencies of qualitative feedback in the sampling population.

Stratified sampling

For all coding samples, we drew evaluation data from the population, stratified so that each sample contained an equal number of comments from (1) A and B trimesters; (2) paper feedback and specific teacher feedback; and (3) each of the two standard open-ended questions. We sampled comments in this way to control for any impact resulting from the differences in the question wording. These sampling parameters were also used to ensure sufficient statistical power for later comparisons across categories in the full dataset: to test, for example, whether student sentiment tended to be more or less positive in their evaluation of papers compared to when they consider specific teaching staff.

Analysis

Prior to compiling the full sample of student feedback, we ran an a priori power analysis using G*Power (Faul et al. 2007). Assuming that A-trimester versus B-trimester feedback was similar in sentiment, we established that a sample size of above 800 was required for our planned 2-x-2 ANOVA testing given an alpha level of .05, a minimum power level of .80, and a small effect size (f = .10). Therefore, we determined that our final dataset would consist of one thousand student responses.

Developing the coding framework

To determine and then iteratively test a coding framework, we drew samples of comments from the population that was not a part of the full study dataset. We coded and discussed an initial set of 50 comments among the research team to establish a draft framework for our content analysis (Lauzen & Dozier 2005; Widnall et al. 2020). It was at this stage that we decided a five-point Likert scale would be appropriate for the sentiment ratings (*Sentiment*) and that a categorical measure of whether the feedback was constructive would be of use (*Constructivity*). A second sample of comments (evenly split, n = 200) was then coded in line with the draft framework. We added a third measurement intended to complement the sentiment and the constructivity measures: a frequency count of personal comments rather than comments focused on someone's teaching abilities (*Personal Comments*: e.g. instances where students stated their dislike of a certain accent or of a teacher's clothing). In this process, we also identified comments that were typical of each sentiment rating, which we collated into the coding guide shown in Table 2.

A second coder was given 25% of the test sample (n = 50) to code according to the framework outlined in Table 2, as well as according to the constructivity and personal categories.

Process

For summarising university evaluations, where results can be very consequential for teaching staff in particular, it is important to ensure that any analytic approach is sufficiently valid (Mertens 2019). Validity is the methodological concern of measurement accuracy (O'Leary-Kelly & Vokurka 1998). In our case, we were interested in whether the sentiment ratings applied by an NLP algorithm were as accurate (i.e. valid) as those determined by people. Based on our power analysis, our student scholar initially coded the sentiment of a sample containing one thousand pieces of student feedback, then re-coded 10% of the dataset a second time (n = 100) to attain intra-coder reliability. Other research team members also independently coded a random selection of 100 comments from the dataset (i.e. 10%) to check inter-coder reliability. For both types of reliability, we calculated the percentage agreement. In line with others' research, we considered an agreement of 80% or higher as acceptable to then use our content analysis as a benchmark (Lauzen & Dozier 2005).

We then generated 'Negative', 'Neutral', or 'Positive' codes for the same sample (N = 1000) using Google's out-of-the-box NLP sentiment analysis tool (Google 2023). Google's NLP is a pre-trained machine-learning model capable of sentiment analysis which we utilised without any additional training specific to our purposes (Wang 2019). In some contexts, researchers have produced promising results without extensive context-specific training (e.g. Uppaal, Hu & Li 2023), which, if accurate, can provide further time and resource efficiencies for analysts (Zang et al. 2023). We compared human coders' sentiment ratings to those of an untrained (i.e. without specific training) NLP, hoping that our procedure would be generalisable to other institutions. As others have done, we tested the NLP-derived sentiment ratings against the benchmark of ratings applied by human coders (e.g. Socher et al. 2013).

Table 1. Frequencies of qualitative feedback in the sampling population.

	A Trimester		B Trimester	
	Positive Question	Negative Question	Positive Question	Negative Question
Paper	4017	3675	8499	7858
Teacher	2535	2014	4073	3254

Sentiment Analysis of Course Evaluation Comments

Code	Example Comments for each Code	Explanation for each Code
1 = strongly disagree	'Poorly organised paper' 'It would have been nice to be taught rather than watching old YouTube videos'	The comment contains nothing positive and is focused purely on negative aspects.
2 = disagree	'This is a difficult paper to do online and I would recommend not running it as an online option, the teacher was enthusiastic and great on a whim' 'The lab workshops were helpful, however it is a lot to learn in one three hour session'	The comment focuses mainly on the negative but has a positive aspect to it.
3 = neutral / uncertain	'Offering workshops and practise [sic] tests' 'Online tools' 'Feedback, enthusiasm'	Comments that you cannot determine as being positive or negative without reference to which question (i.e. things that went well / improvements). If unsure on comment also code as a 3 to allow additional context.
4 = agree	'The teacher has a friendly approach to most students' 'The labs were good when I understood what was happening, but sometimes they felt rushed over'	The comment focuses mainly on the positive but it has an aspect which is negative. This could be a word similar to 'most', 'usually', 'sometimes' or it could be one negative sentence in an overall positive comment.
5 = strongly agree	'Lectures were very well structured, tutorials helped us to go over the content for the week' 'The lecturer was very helpful with responses to assignment questions, the specific assignment focused workshops were very helpful' 'Open communication line with content being explained clearly, described everything well and was enthusiastic about the paper' 'Overall great paper'	The comment is fully positive with no negative aspects. There is enough context in the comment to understand that it is wholly positive.

Table 2. Final version of the sentiment-coding framework.

Findings and Discussion

We found encouraging evidence for the application of NLP to accurately summarise qualitative evaluation feedback from students. For the paper feedback, three-point ratings attributed by coders were statistically related to the ratings generated by the NLP, X^2 (4) = 342.12, p < .001, Φ = .90. In Figure 1, we show the relationship between the two sets of sentiment ratings. Most of the time, when our team coded feedback as positive, so too did the NLP (96.49%). To a slightly lesser extent, when we gave a negative sentiment rating, the NLP most often did as well (72.52%). Furthermore, our content analysis ratings were validated by two independent coders, who agreed with the primary coder's ratings of paper feedback in 92% and 84% of cases, respectively.

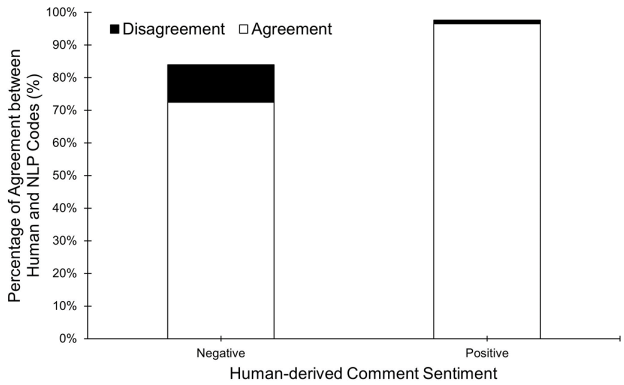

Figure 1. Percentage Agreement for Paper Feedback between Human- and NLP-ratings that Agreed (white) and Disagreed (black) as a function of Feedback Valence.
Note: the bars may not equal 100% as both the human and NLP categorised comments into three categories (negative, neutral, and positive) of which the neutral codes are not presented.

In the case of feedback about teaching staff, three-point sentiment raters from our coders were also significantly related to those of the NLP, X2 (4) = 226.62, p < .001, Φ = .68. We include Figure 2 to display the proportion of agreement when our human-coders considered the feedback to be negative (left: *Coder Agreement* = 60.44%) compared to positive in valence (right: *Coder Agreement* = 91.34%). We found sufficient inter-rater agreement between each of the independent coders and the ratings given by the primary coder (87% and 82% agreement). Overall, the Paper and the Teacher feedback ratings follow a similar trend; agreement between the human-given ratings and the NLP's is high, particularly when the feedback is positive.

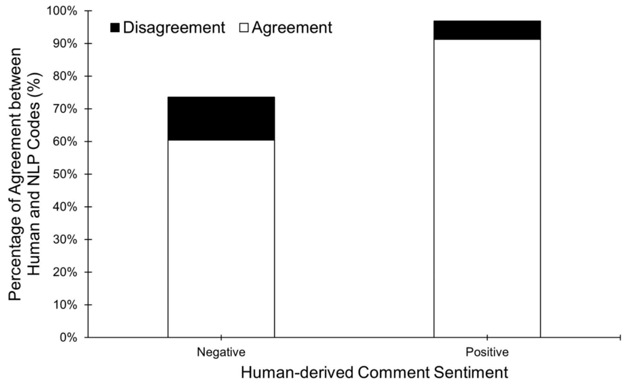

Figure 2. Percentage Agreement for Teacher Feedback between Human- and NLP-ratings that Agreed (white) and Disagreed (black) as a function of Feedback Valence.
Note: the bars may not equal 100% as both the human and NLP categorised comments into three categories (negative, neutral, and positive) of which the neutral codes are not presented.

To explore the relationship between the frequency with which the NLP gave positive ratings compared to the human coders, we conducted a 2-x-2 ANOVA. The complete model was statistically significant, F (3, 1926) = 15.36, p < .001. We found a main effect of comment type that feedback about teachers was more likely to be positive (M = 2.37, 95% CI [2.31, 2.42) than feedback about a paper (M = 2.17, 95% CI [2.12, 2.22]), F (1,1926) = 27.10, p < .001. We also found a small but significant main effect of the coder type, F (1, 1926) = 14.40, p < .001: specifically, the NLP-generated coding tended to be more positive (M = 2.34, 95% CI [2.29, 2.39]) than codes generated by the primary human coder (M = 2.20, 95% CI [2.15, 2.25]). There was no interaction effect between the subject of the feedback (i.e. paper vs. teachers) and the coder type (i.e. NLP vs. human), F (1, 1926) = 2.98, p < .085. Our analysis also showed that ratings from the NLP were more stable across subject type (M_{Diff} = .13) than the ratings applied by our team's primary coder (M_{Diff} = .26), although only marginally.

Evaluation responses in which students made personal comments were relatively few (*Paper Feedback* = 1.80%; *Teacher Feedback* = 6.41%). The content relating to specific teachers, where personal comments are more likely, amounts to one personal comment for approximately every 16 pieces of substantive critique. Further, not all the personal comments were negative. In fact, 34% of the personal comments were favourable, and another 3% were neutral in valence. Our general categorisation of whether comments were constructive was confounded by the incidence of neutral comments, where coders did not have enough contextual information to judge the sentiment. For the paper feedback, 99% of the comments that we categorised as unconstructive were also rated as neutral in sentiment. In the teacher feedback, 97% of unconstructive comments were neutrally coded. It is likely that the neutral point on our sentiment scale is measuring a similar construct to the constructive category, and, therefore, the sentiment scale alone would suffice in a study replication.

As others have done, we found that much of the student feedback required a neutral rating due to an absence of information (Socher et al. 2013). For our study, we ensured that the human coders had the same limited amount of information as the NLP would, as we were primarily testing to what extent the NLP could generate meaningful data independent of human support. Although our concern was that the neutral rating would be overly prevalent in the NLP dataset, the neutral rating was more often applied by human coders for both the paper feedback (*NLP* = 20%; *Human* = 30%) and the teacher feedback (*NLP* = 16%; *Human* = 31%). As neither coding exercise included the questions to which students were responding, that the human coders categorised approximately 30% of feedback as 'neutral' was to be expected.

Given the potential impact of negative feedback on university teaching staff, particularly if the feedback is personal rather than constructive, a risk-averse approach to aggregating evaluations into a quasi-quantitative measure of sentiment is warranted (Graesser & McNamara 2012; Zaitseva, Tucker & Santhanam 2022). Often such a risk-averse approach is understood to mean one with significant active human oversight rather than an algorithmic approach (Khurana et al. 2023). In our research, however, we found that the use of NLP algorithms to summarise student feedback may, in fact, produce more favourable results. Further research regarding the precise circumstances in which an NLP tends toward more positive appraisals is needed, however, prior to any institutional-level use for communicating evaluation results. An avenue for such research is in applying the Summarize and Score (SASC) method to the analysis of university feedback. In general terms, the SASC method uses NLP to analyse the sentiment of texts and concurrently generate an explanation of the otherwise 'black box' process used to arrive at that sentiment rating (Singh et al. 2023). The greater level of transparency that the SASC method can provide would be beneficial for disseminating information from the evaluations to both staff and students, thereby helping to 'close the loop'.

In summary, our analysis revealed a statistically significant relationship between our team's sentiment coding and the NLP sentiment analysis. Consistently, when we identified a comment as positive, the NLP algorithm also classified it as positive. Notably, our findings indicated that, on average, the NLP tended to assign a higher proportion of positive codes compared to our human coders. These results constitute promising evidence for the future use of NLP to convey accurate top-level insights from large qualitative datasets more effectively. Specifically, these results support the use of NLP to revitalise our approach to evaluating the wealth of useful feedback provided by students.

References

Explorance. 2023. Blue: The People Insights Platform. [Computer program]. Available at https://explorance.com/products/blue/ [Accessed 24 September 2023].

Faul, F., Erdfelder, E., Lang, A-G. and Buchner, A. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39: 175-191.
doi.org/10.3758/BF03193146

Google. 2023. Natural Language AI. [Computer program]. Available at https://cloud.google.com/natural-language [Accessed 24 September 2023].

Graesser, A.C. and McNamara, D.S. 2012. Automated analysis of essays and open-ended verbal responses. In: Cooper, H, Camic, P M, Long, D L, Panter, AT, Rindskopf D and Sher, K J (eds.). *APA handbook of research methods in psychology, Foundations, planning, measures, and psychometrics*. Washington, DC: pp. 307-325.
doi.org/10.1037/13619-017

Khurana, D., Koli, A., Khatter, K. and Singh, S. 2023. Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 82: 3713-3744.
doi.org/10.1007/s11042-022-13428-4

Knight, G., Ordidge, H., Timothy, A. and Davis, M. 2022. Change agents: The impact of a student partnership on the educational practice of a diverse higher education engineering faculty. *International Journal for Students as Partners*, 6: 27-43.
doi.org/10.15173/ijsap.v6i2.4947

Kukich, K. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24: 377-439.
doi.org/10.1145/146370.146380

Lauzen, M.M. and Dozier, D.M. 2005. Maintaining the double standard: Portrayals of age and gender in popular films. *Sex Roles*, 52: 437-446. doi.org/10.1007/s11199- 005-3710-1

Lumivero. 2023. NVivo. [Computer program]. Available at https://lumivero.com/products/nvivo/ [Accessed 24 September 2023].

Mertens, D.M. 2019. *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. 5th ed. SAGE Publications.

O'Leary-Kelly, S.W. and Vokurka, R.J. 1998. The empirical assessment of construct validity. *Journal of Operations Management*, 16: 387-405.
doi.org/10.1016/S0272-6963(98)00020-5

Ormerod, C.M. and Harris, A.E. 2010. Neural network approach to classifying alarming student responses to online assessment: arXiv:1809.08899, 20 September 2018. Available at https://doi.org/10.48550/arXiv.1809.08899 [Accessed 25 September 2023].

Ormerod, C.M., Patel, M. and Wang, H. 2023. Using language models to detect alarming student responses: arXiv:2305.07709, 12 May 2023. Available at http://arxiv.org/abs/2305.07709 [Accessed 25 September 2023].

Sag, I.A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. 2002. Multiword expressions: A pain in the neck for NLP. In: Gelbukh, A (ed.). *Computational Linguistics and Intelligent Text Processing*. Berlin, Heidelberg: Springer. pp. 1-15. doi.org/10.1007/3-540-45715-1_1

Singh, C., Hsu, A.R., Antonello, R., Jain, S., Huth, A.G., Yu, B. and Gao, J. 2023. Explaining black box text modules in natural language with language models: arXiv:2305.09863, 17 May 2023. Available at https://doi.org/10.48550/arXiv.2305.09863 [Accessed 25 September 2023].

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y. and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, Seattle, USA, on 18-21 October 2013, pp. 1631-1642.

Tschirhart, C. and Pratt-Adams, S.D. 2010. Closing the loop: An evaluation of student-led module feedback at one UK higher education institution. *International Journal for Students as Partners*, 3: 78-90. doi.org/10.15173/ijsap.v3i2.3554

Tucker, B.M. 2015. *The student voice: Using student feedback to inform quality in higher education*. Open Access Published Ph.D. Thesis. Curtin University.

University of Waikato. 2023. *The evaluation questions*, nd. Available at https://www.waikato.ac.nz/students/teaching-and-learning/student/student-evaluations/the-evaluation-questions/ [Accessed 22 April 2024].

Uppaal, R., Hu, J. and Li, Y. 2023. Is Fine-tuning Needed? Pre-trained Language Models Are Near Perfect for Out-of-Domain Detection: arXiv:2305.13282, 22 May 2023. Available at https://doi.org/10.48550/arXiv.2305.13282 [Accessed 25 September 2023].

Wang, S. 2019. *Predicting unethical behavior from interview responses: Machine learning models versus human judges*. Published thesis (PhD), Massachusetts Institute of Technology.

Weis, D. and Willems, H. 2017. Aggregation, validation, and generalization of qualitative data - Methodological and practical research strategies illustrated by the research process of an empirically based typology. *Integrative Psychological and Behavioral Science*, 51: 223-243. doi.org/10.1007/s12124-016-9372-4.

Widnall, E., Grant, C.E., Wang, T., Cross, L., Velupillai, S., Roberts, A., Stewart, R., Simonoff, E. and Downs, J. 2020. User perspectives of mood-monitoring apps available to young people: Qualitative content analysis. *JMIR mHealth and uHealth*, 8: e18140- e18140. doi.org/10.2196/18140

Zaitseva, E., Tucker, B. and Santhanam, E. (eds.) 2022. *Analysing Student Feedback in Higher Education: Using Text-Mining to Interpret the Student Voice*. 1st ed. Oxon and New York: Routledge.

Zhang, W, Deng, Y, Liu, B, Pan, S J and Bing, L 2023 Sentiment analysis in the era of large language models: A reality check. arXiv:2305.15005. doi.org/10.48550/arXiv.2305.15005

Te Puna Aurei LearnFest 2022 was an online conference delivered by The University of Waikato in partnership with Cardiff University on 23 and 24 November 2022.

Bowell, T., Pepperell, N., Richardson, A., and Corino M-T. (eds.) 2024. *Revitalising Higher Education: Insights from Te Puna Aurei LearnFest 2022*. Cardiff: Cardiff University Press.

doi.org/10.18573/conf2

Cardiff University Press | Gwasg Prifysgol Caerdydd

First published 2024

Cardiff University Press
Cardiff University, PO Box 430
1st Floor, 30-36 Newport Road
Cardiff CF24 0DE

cardiffuniversitypress.org

Publication design and layout by
Academic Visual Communication

ISBN: 978-1-9116-5350-9 (PDF)