

RESEARCH TITLE:

Application of Natural Language Processing (NLP) for Sentiment Analysis and Topic Modelling of Postgraduate Student Engagement Data

AUTHORS:

**Dr Etain Kiely
Dr Cormac Quigley
Ontiretse Ishmael &
Ikechukwu Ogbuchi**

ORGANISATION/INSTITUTION:

Atlantic Technological University

Foreword

The StudentSurvey.ie Steering Group is pleased to publish the results of five research projects analysing the qualitative data generated by the free-text response questions in StudentSurvey.ie and PGR StudentSurvey.ie. The results contained within this report make up one part of this research series.

Five projects were funded by research bursaries offered by StudentSurvey.ie in October 2022. The aim of the bursary awards was to promote greater ownership and encourage wider use of the StudentSurvey.ie and PGR StudentSurvey.ie data. Proposals for the analysis of the qualitative data emerging from StudentSurvey.ie and PGR StudentSurvey.ie were invited from members of the research community within the participating institutions, as well as commercial data analysis companies. The projects were completed in May 2023.

Each project is an independent project undertaken by qualified and experienced researchers on behalf of StudentSurvey.ie. Each project took a unique approach. The commonalities between all five projects are that they all utilised well-grounded methodologies, offer mechanisms for replication of the analysis in future years, and are innovative and authentic.

These results are the first of their kind for StudentSurvey.ie and PGR StudentSurvey.ie and we hope they are the first of many research projects involving the qualitative results of these surveys.

What are StudentSurvey.ie and PGR StudentSurvey.ie?

StudentSurvey.ie (the Irish Survey of Student Engagement) is an annual national survey of student engagement among first year undergraduate, final year undergraduate and taught postgraduate students in higher education institutions in Ireland.

PGR StudentSurvey.ie (the Irish Survey of Student Engagement for Postgraduate Research Students) is a biennial national survey of student engagement among Masters by Research students and PhD students in higher education institutions in Ireland.

Both surveys are designed to focus on student engagement, namely the amount of time and effort that students put into meaningful and purposeful educational activities, and the extent to which institutions provide such opportunities and encourage students to engage with them. The data collected reflect students' self-reported perceptions of their experiences.





Application of Natural Language Processing (NLP) for Sentiment Analysis and Topic Modelling of Postgraduate Student Engagement Data



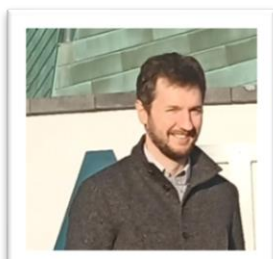
Ollscoil
Teicneolaíochta
an Atlantaigh

Atlantic
Technological
University

Atlantic Technological University



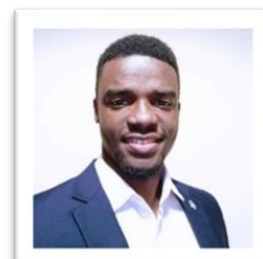
**Dr. Etain
Kiely**



**Dr. Cormac
Quigley**



**Ontiretse
Ishmael**



**Ikechukwu
Ogbuchi**

Final Report - May 2023

**Postgraduate Research Students (PGR)
StudentSurvey.ie Report**

Foreword

Note From the Research Team

Our research active team are delighted to have the chance to apply machine learning skills to contribute to the analysis of the Irish postgraduate research student data. We acknowledge the vital role research and education plays in our diverse society enabling innovative contributions that can empower future generations.

We aim to encourage researchers to validate and expand upon our findings, promoting a collaborative and robust education research community. A comprehensive how to guide is available in Appendix 2, detailing the implementation of NLP analysis with large qualitative datasets. We are grateful to the community of research students for contributing their perspectives, we hope this analysis captures your voices through sentiment analysis and opinion mining.



Recipients of International Award for "Responsible AI" (Dataiku Award)

Note From ATU President, Orla Flynn



I am delighted to see the contribution of researchers from Atlantic Technological University (ATU) to this important agenda. Along with other newly-established technological universities, the development and growth of our research activities will be a main focus for us in the years ahead. This report highlights the importance of sustaining and improving our research environment to support postgraduate researchers. It emphasizes the need for enhanced student funding, high-quality skill acquisition experiences, and well-resourced work environments. The innovative use of machine learning approaches offers new findings and broadens the evidence base for a data driven approach in education research. This report will add to the national discourse around providing adequate support and funding for researchers, aligned to "*Impact 2030: Ireland's new research and innovation strategy*".

The study offers new insight into the need for funding as an enabler for researchers to contribute to their fields of study. It highlights the value of structured training and development opportunities to empower researchers and nurture talent. Further evidence also supports the need for a supportive and well-resourced work environment which supports inclusivity and collaboration. I hope this report will spark meaningful discussions and actions to create a vibrant research ecosystem, and I thank and commend the authors, and all involved.

Dr Therese Montgomery Dr John Healy



Special thanks to Donal McGinty for providing the computational resources, Dr John Healy for offering support and expertise in the emerging field of Natural Language Processing (NLP) and Dr Therese Montgomery for support in preparing the report.

Suggested citation: When referencing this document please use the following citation:

Kiely, E., Quigley, C., Ishmael, O., & Ogbuchi, I. (2023). Application of Natural Language Processing (NLP) for Sentiment Analysis and Topic Modelling of Postgraduate Student Engagement Data. In PGR StudentSurvey.ie Qualitative Data Analysis Report. Dublin.



Executive Summary

Introduction

This study uses Natural Language Processing (NLP) to explore the sentiment and prevalent topics within the PGR (Postgraduate Student Survey) StudentSurvey.ie, analysing data from 2018, 2019 and 2021. The study focuses on two free response questions to extract insights into what experiences postgraduate students find most valuable about their research programme (Q10) and what can be improved (Q11). The study uses sentiment analysis to examine how different aspects of the student experience relate to the overall positivity reported by postgraduate students.

Innovative Methodology

This report demonstrates the application of Natural Language Processing (NLP) techniques in analysing a large qualitative dataset. Sentiment scoring, which involves determining the emotional tone of each response, was conducted using Microsoft Azure, and statistical analysis was performed to explore the relationship between sentiment scores and characteristic factors. Latent Dirichlet Allocation (LDA), a topic modelling technique, was utilized to uncover key topics in the entire dataset and specific subgroups of respondents. Finally, the report examined specific responses, highlighting the student voice, emotional tone, and perspective by selecting and presenting respondent's illustrative quotes.

This methodology has allowed us to broaden the evidence base for aspects known to be important to the postgraduate research experience. Beyond this, we have also identified novel areas of importance, by leveraging topic modelling to reveal latent thematic similarities and patterns. This can be achieved even when different language or terminology is used to describe related concepts. To facilitate transparency and reproducibility, the data has been visualized with an interactive dashboard (see Appendix 1). A comprehensive step-by-step guide illustrates on how to implement this novel approach using both open source and black box solutions (see Appendix 2). The integration of NLP and machine learning offers additional capability and functionality to researchers to expand sample sizes, reduce bias, shorten project time frames, enhance data exploration, and strengthen the rigor of education research.

Insights and Key Findings

The findings relate to the two guiding questions, the sentiments expressed in response are predominantly neutral or positive with only 11% of responses classified as negative for Q10, and 26% negative for Q11.

Q10 Insights:

What aspects/ elements of your research degree programme are most valuable?

The crucial importance of the supervisor's role in the postgraduate research experience is substantiated by an extended body of evidence. As well as being the dominant topic in the respondent's data, sentiment analysis captured that the relationship has effects across most aspects of the student experience. The study also shows that a good supervisor – student relationship affects the positive sentiment of the respondents most valuable experience, even for those who did not mention the relationship directly. Institutional acknowledgment of the significance of this relationship would be demonstrated by supporting and resourcing environments (time and space) that cultivate these relationships. The respondents emphasize the critical value of skill development, both in formal research-specific training and transferrable team and communication skills development. Positive sentiment scores for the most valuable experiences are correlated with the evidence of opportunities to develop these skills. In terms of tangible demonstration of the acquisition of transferable and subject specific skills, the publication of papers and presentation at conferences is strongly linked to a positive most valuable experience. Respondents who have positive working relationships, clarity on university processes, coupled with sufficient funding and good life balance report more positive best experience.



Illustrative Quotes of Sentiment

 Positive Sentiment 0.98	The findings are illustrated by a positively classified sentiment: □ <i>"First and foremost, supervision. Secondly, the funding available and the opportunities it affords me to pursue my research and avail of additional supports and services. Thirdly, the research community is very inclusive and feel part of a network within the IoE."</i>
 Negative Sentiment 0.98	In contrast a negatively classified sentiment shows the difficulty and hardship a student endures if these environments and conditions are not in place: □ <i>The courses on offer are very poor. One of my supervisors never has time to see me and no interest in my research. My other supervisor is moving away and since that decision was made is also not spending much time on my project anymore (despite being an excellent supervisor beforehand). My stipend is not enough to live and I am in a constant financial struggle. With all this together I am still giving all I can, but it is very hard to hold all that pressure and getting no support.</i>

Q10. Statistical Analysis of Factors

Statistical analysis of the sentiments identified four factors as being statistically important for determining the positivity of students' most valuable experiences.

- **Gender:** Female respondents have a significantly more positive response to Q10 than male respondents. Latent topics within the female researchers' responses differ to those of their male counterparts.
- **Accommodation Type:** Students who live in their own home had significantly less positive responses to Q10 compared to those who rent. There is an association between residence type and study mode, with a higher proportion of students living in their own accommodation studying part-time.
- **Mode of Study (Full/Part time):** Students who are studying part time had significantly less positive responses to Q10 compared to those studying full time. Topic modelling identified that part time students have separate needs when compared to full time students and that these are not observed if topic modelling is only carried out on the responses as a whole.
- **Stipend Funding:** respondents who received a stipend expressed a more positive sentiment than those who did not. Topic modelling shows that adequate funding is crucial for students and impacts on all aspects of their personal and academic life.

Q11 Insights:

What aspects of your research degree experience could be improved?

The study identified funding, effective communication systems, and university life balance as key topics for suggested development. This is consistent with the findings of the study conducted by Share et al. (2020) which highlighted the need for adequate funding as a critical factor to maintaining a good quality of life and maximizing the research experiences.

Improved communication processes are needed to seamlessly integrate students into university environments, including awareness of necessary forms and systems for payment, travel, and gaining skills in lecturing and lab work. Induction and personal development plans did not show a significant relationship with positive responses to question 10. This suggests they are not adequate on their own and there is a need for ongoing, frequent, and timely communication with researchers throughout their study. A well-resourced work environment goes beyond the physical space, to promoting a collaborative and supportive peer network. Absence of such an environment was associated with feelings of loneliness and isolation among respondents. By addressing concerns related to appropriate funding, communication, collaborative environments, universities can potentially improve the satisfaction and creative productivity of research students. Covid did not emerge as a specific topic in the 2021 data. Its impact was referenced in terms of restricted access to facilities and peer support network however it was acknowledged as a transient issue and was not addressed directly by either Q10 and Q11.



Introduction



This study uses data collected in 2018, 2019 and 2021 by the PGR StudentSurvey.ie (Irish Survey of Student Engagement for Postgraduate Research Students; *Suirbhé na hÉireann ar Rannpháirtíocht na Mac Léinn do Mhic Léinn Taighde Iarchéime*).

The survey asks students directly about their experiences of higher education in Ireland, including their academic, personal, and social development. The survey is open to Masters by Research (NFQ 9) and PhD students (NFQ 10) for completion. Twenty-one higher education institutions participated in 2021 survey; seven additional questions were added in response to the COVID pandemic. The study will investigate the variation between previous years data and new themes emerging also against the backdrop of Covid pandemic for the 2021 data.

Qualitative data provides rich descriptions of the respondents' thoughts, feelings, and lived experiences (Daniel & Harland, 2018) however the heterogeneous nature of the data which tends to have high volume and complexity can be time and labour-intensive. This study aims to contribute to the existing body of knowledge derived from previous Studentsurvey.ie qualitative research summarised in Table 1. These studies employed the use of semi-automated approaches to address the onerous task of managing large qualitative datasets. The limitations identified in these studies include challenges of managing missing data and the resource-intensive nature of thematic analysis, while also noting the how researcher bias can lead to variation in results between years. This study utilises Natural Language Processing (NLP) to mitigate researcher bias and offers a repeatable year-to-year methodology for examining patterns in the data.

Table 1 Summary of Published Reports of Student Survey.ie

Research	Authors	Qualitative Method	Data Cohort	Year Data
Study 1	Dr. Stephen Erskine & David Harmon	Meta-analysis and Frequency of Character	Undergraduate & taught postgraduate	2016 - 2020
Study 2	Dr Darragh McCashin & Christine Boyd	Thematic analysis	Undergraduate & taught postgraduate	2020
Study 3	Dr. Seán Lacey	Word Frequency analysis Text Coding and Reference Extraction	Postgraduate data	2018-2019
Study 4	Dr Michelle Share, Dr Caitríona Delaney & Dr Rory Mc Daid	Thematic analysis Content analysis	Postgraduate data	2019
Study 5	Dr Sarah Diffley (Lead), Dr. Lydia Devenney, Dr. Isobel Cunningham & Dr. Vicky O'Rourke	Thematic and Content analysis	Undergraduate & taught postgraduate	2021

In this study, NLP is used to automate and enhance qualitative analysis providing an efficient tool for unstructured data (Chen et al., 2020). This is achieved by replacing manual processes in tasks like data pre-processing, theme identification, and categorization, reducing time and bias while providing valuable insights from large textual datasets. This method does not solely rely on the researcher's reasoning and abilities for analysis (Tierney, 2012) but instead employs internal resources such as dictionaries, thesaurus, templates, and libraries to identify keywords within the data, providing more objective results than researcher induction



alone. Guetterman et al. (2018) propose that NLP not only establishes a foundation for quicker qualitative coding but also serves to validate qualitative findings.

Objectives

The study seeks to achieve the following objectives.

1. Examine NLP machine learning performance comparing algorithms and accuracy scores.
2. Apply the NLP model to 2018/2019, 2021 PGR StudentSurvey.ie data.
3. Visualise insights using interactive dashboard to explore sentiments and opinions.
4. Disseminate findings through written report.
5. Provide a guide and tools for researchers on how to replicate and implement NLP.

The report will utilize these objectives as signposts to structure the study's report navigating the reader through the method, key insights using dashboard visuals and offer a reference guide on how to undertake and replicate NLP research in education settings.

Method

As an initial investigation, NLP was utilized to analyse the word frequency in relation to respondents' perceptions of the most valuable aspect of research degrees. Figure 1 presents word cloud representations illustrating how the NLP study replicates the analysis conducted by Dr Sean Lacey from MTU REVEAL study (2021). Both studies capture “supervisor”, followed by “skills” as the most frequent words for Q10 using 2018/2019 data.

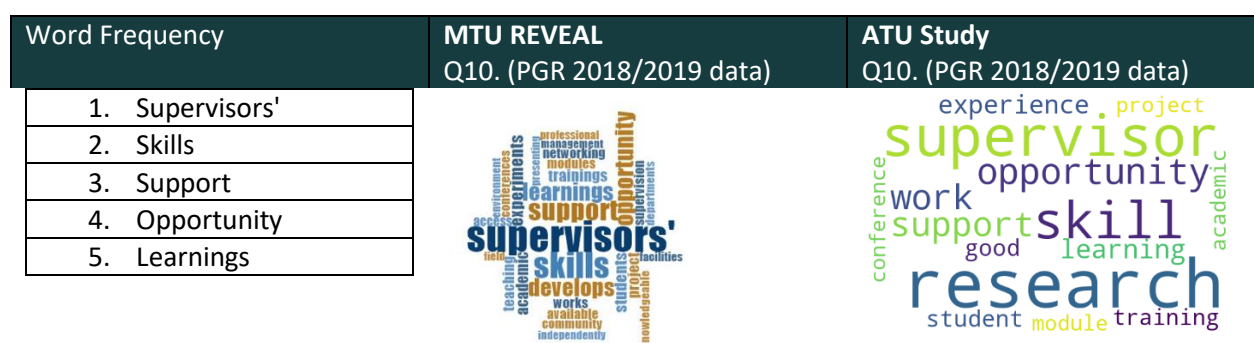


Figure 1: Word clouds showing similarity for Q10 most valuable aspect of research degree.

Question 11 of the PGR studentsurvey.ie investigated how the research experience could be improved. The MTU (2021) study identified the need for additional student support as most prevalent response word. Figure 2 illustrates that NLP consistently replicated the same top 5 frequency counts.

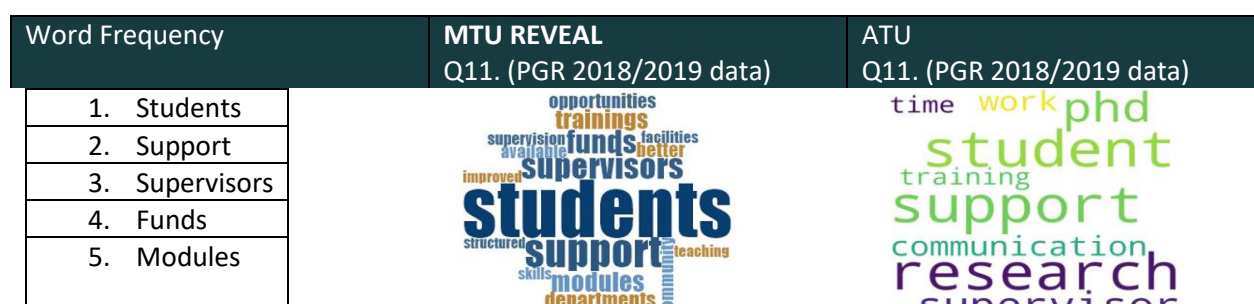


Figure 2: Word cloud similarity of MTU and ATU analysis for Q11 aspects to be improved.



Processing of responses to Q10 and Q11

The study utilized Microsoft Azure Power BI's text analytics features to conduct data analysis, enabling the extraction of sentiment scores. In addition, the pycaret library was employed to extract topics using Latent Dirichlet Allocation (LDA). Figure 3 outlines the systematic approach adopted however it should be noted that it is not necessary to complete sequentially, these are numbered here for the ease of explaining how the findings will be presented in this study.

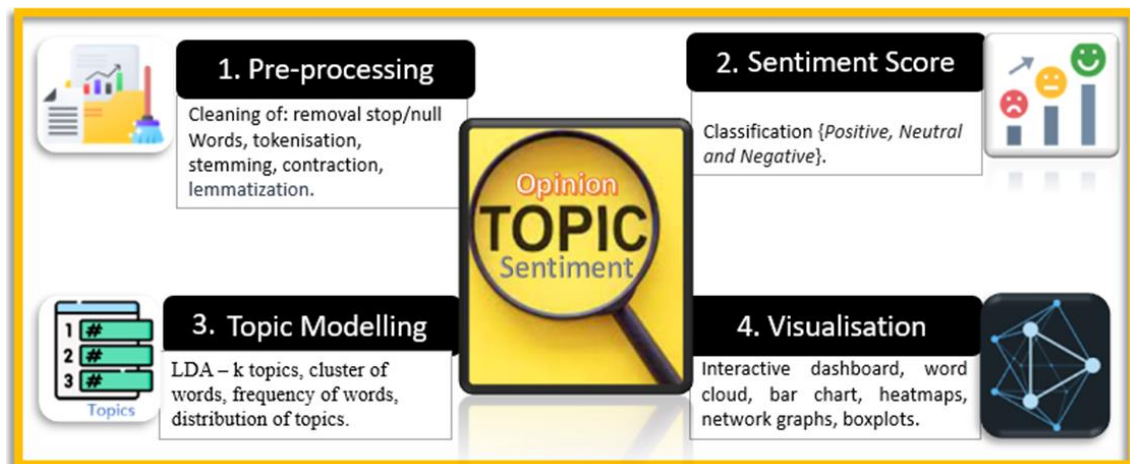


Figure 3 NLP Approach/Methods Applied to Student Survey Data

Step 1: Pre-processing

Power BI facilitated the removal of stop words from the textual data, contributing to noise reduction in the corpus. The data was cleaned by eliminating stop words, performing tokenization to identify units of meaning, and reducing words to their word stems through lemmatization, which considers part of speech and applies morphological analysis. These cleaning techniques aimed to enhance the quality of the corpus and improve classifier accuracy (Lim & Buntine, 2014; Lin & He, 2009; Malema et al., 2022; Rakhmanov & Schlippe, 2022). The survey was completed by a group of respondents who used the Irish language, and their responses were translated for analysis. However, the original format of the Irish language responses was retained for the interactive interface, allowing for a more authentic and inclusive experience for users.

Step 2: Sentiment Analysis

NLP technique is used to identify the emotional tone in text (Bansal et al., 2022) and the application of sentiment analysis (also referred to as opinion mining) enables analysis of student feedback (Dalipi et al., 2021; Kastrati et al., 2021). Power BI's built-in text analytics tool was utilized to extract sentiment scores. This determined the sentiment associated with each piece of text which was classified positive, negative, or neutral. The choice of Microsoft Azure Text Analytics was justified by comparing the sentiment scores obtained from three different sentiment analysers. For the sake of convenience, the combination of Microsoft Azure Text Analytics and Power BI is referred to as "PowerBI" in this report.

Step 3: Topic Modelling

NLP topic modelling using Power BI integrated with Pycaret is an open-source machine learning library in Python which facilitates the estimation of the optimal number of topics and the training of the LDA model (Latent Dirichlet Allocation). The LDA model is commonly used for sentiment analysis on platforms like Twitter (Lim & Buntine, 2014). It assumes that topics and their associated words follow a Dirichlet distribution, combining probability distributions over observable and hidden variables (Anusha & Singh, 2015).



Step 4: Interactive Visualisation

Insights are visualised using an interactive dashboard enabling users to drill into specific language or category codes. Typical visuals can include word cloud, network graphs, radial graphs, heatmaps, boxplots, infographics and sentiment gauges.

Statistical Analysis

To facilitate the evaluation process, standard parametric and non-parametric tests were conducted using Minitab (statistical analysis software) . These statistical tests allowed for the rigorous analysis of the data, providing insights and supporting the validity of the findings.

By incorporating these tools and techniques, the study aimed to conduct a comprehensive and robust analysis of the textual data. It harnessed the strengths of Power BI for data visualization, Pycaret for NLP topic modelling, and Minitab for statistical analysis, ensuring a comprehensive understanding of the data and providing valuable insights.

Two free text open-ended questions 10 and 11 on the survey are investigated.

PGR Student Survey Question	Labelling Protocol
Q10. What aspects/ elements of your research degree programme are most valuable?	denoted as " Q10: Most Valued"
Q11. What aspects of your research degree experience could be improved?	denoted by "Q11: Improvements".

The total number and response rate to the two key questions Q10: "Most Valued" and "Q11: Improvements" is presented in Table 4. In 2021, 3541 postgraduate students of the 11,195-research student population responded to the survey in Ireland. Of those students 61% answered Q10 and Q11. This was an increase on the response rate of 2018 and 2019 for these two open ended questions.

Table 2 Yearly Number and Rate of Response for the Q10: MOST VALUED and Q11: IMPROVEMENTS

Fieldwork Year for Survey	Number and Rate of Response	
	Q10: MOST VALUED	Q11: IMPROVEMENTS
2018	1603 (54%)	1571 (53%)
2019	1423 (52%)	1391 (51%)
2021	2174 (61%)	2168 (61%)
TOTAL	5200 (56%)	5130 (55%)

The two questions generated a total of 928,522 data characters across the years 2018, 2019 and 2021 as presented in Table 3.

Table 3 Data characters for Q10 and Q11 for years 2018, 2019, 2021.

PGR Question	Number of characters
Q10: Most Valued	383,364
Q11: Improvements	545,158
Total	928,522



Participants

A total of 9245 students completed the survey between the years 2018, 2019 and 2021. The demographic data for this cohort is presented in Figure 4 which illustrates the accompanying interactive PGR dashboard. This provides a high-level overview of the demographic breakdown across gender, mode of study, area of study, country of domicile, NQAI level, funding support and type of residency.

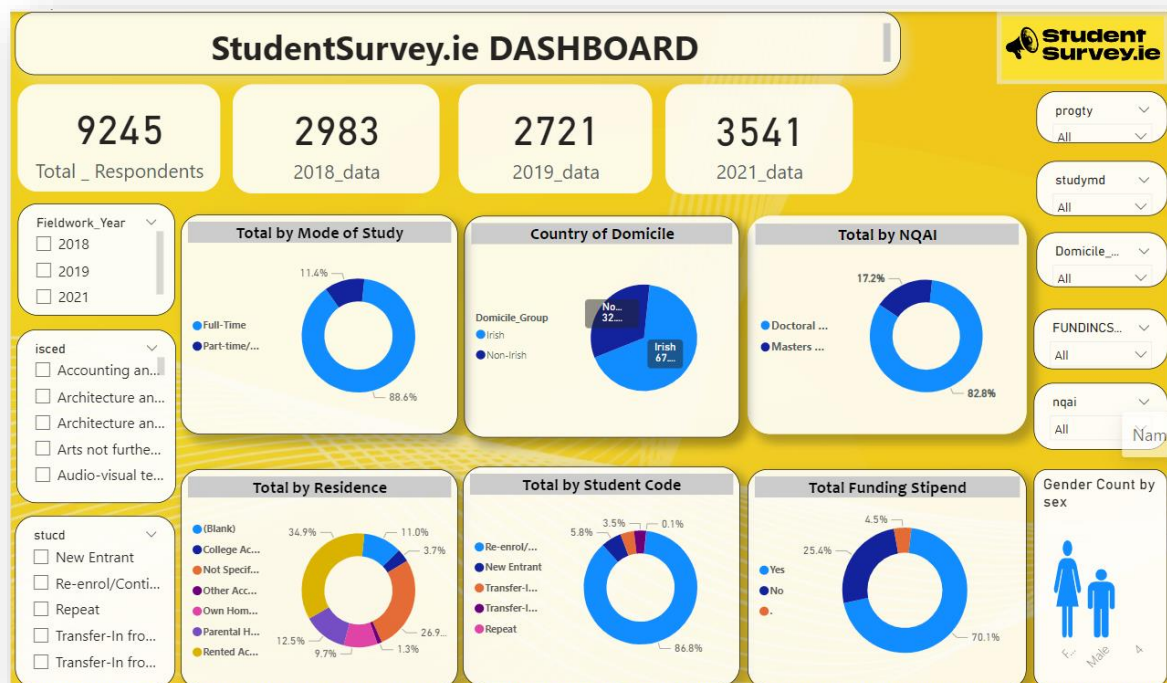


Figure 4: Screenshot of interactive dashboard which can be used to dynamically explore the demographic breakdown of PGR respondents.

Objective 1: Examine NLP machine learning models performance and accuracy scores.

Sentiment analysis, also known as opinion mining is an NLP technique to identify the emotional tone behind a given piece of text (Bansal et al., 2022). Sentiment analysis through NLP assigns a sentiment score to text on a continuous scale from entirely negative to entirely positive. The actual scores are arbitrary and vary depending on the package used. In this document we will normalise all scores to a scale from plus one to minus one with the classifications shown in Table 4. Sentiment categories were defined using the fine-grained sentiment style categorization (microsoft.,2023)

Table 4 Sentiment scores and the classification system.

Sentiment Gauge	Score Range	Sentiment Category
	0.5 to 1	Very Positive
	0.1 to 0.5	Positive
	-0.1 to 0.1	Neutral
	to -0.1 to -0.5	Negative
	-0.5 to -1	Very negative



There are several software packages available for sentiment analysis which vary in the level of complexity in how they approach the analysis. At the simplest level, punctuation and stop words are removed and the remaining words are each given a predetermined score leading to an overall sentiment. This approach has very obvious shortcomings since it lacks the ability to detect context however it is entirely transparent in how it operates. At the other end of the scale, AI based packages ingest complete sentences or paragraphs and use a trained neural network, in essence a black box, to determine the sentiment score. This has a far greater sensitivity to context and meaning and results in more accurate sentiment scoring at the cost of explainability. It is also highly dependent on the material used to train the algorithm and may inadvertently show bias if a particular dialect or regional variation is underrepresented in the training material (Bordoloi & Biswas, 2023; Paper, 2021)

In this study we investigate the use of three different packages TextBlob (Hazarika et al., 2020), Vader (Hutto et al. 2014) and PowerBI (G.Singh et al. 2023) and examine the output scores to look at consistency. Both Textblob and VADER are open source and are transparent in how they parse, transform and analyse text. Power BI contains an inbuilt analytics engine, which is black box in nature. The nature of the underlying engine which performs the analysis is not described beyond that it is a “*pretrained machine learning model*” which has been trained with an unshared dataset reported to be largely using consumer reviews (“Transform and shape data in Power BI Desktop for AI insights,” 2023).

Sentiment scores for the responses to Q10 were generated using each package and compared. **Error! Reference source not found.**4 shows the scores generated by the different packages plotted on a 3D scatter plot and that they are weakly correlated with a high level of significance. The pairwise correlation shows that there is a poor level of agreement between the scoring of any individual response. The objective of our use of sentiment analysis is not to measure individual answers, rather groups or subgroups. Therefore, while increased accuracy of sentiment analysis will lower the number of responses needed to find a statistical significance (in essence reducing the noise of the observed sentiment signal), errors in determining the sentiment of a single answer are not problematic provided that the classification is correct in enough cases to provide a true signal.

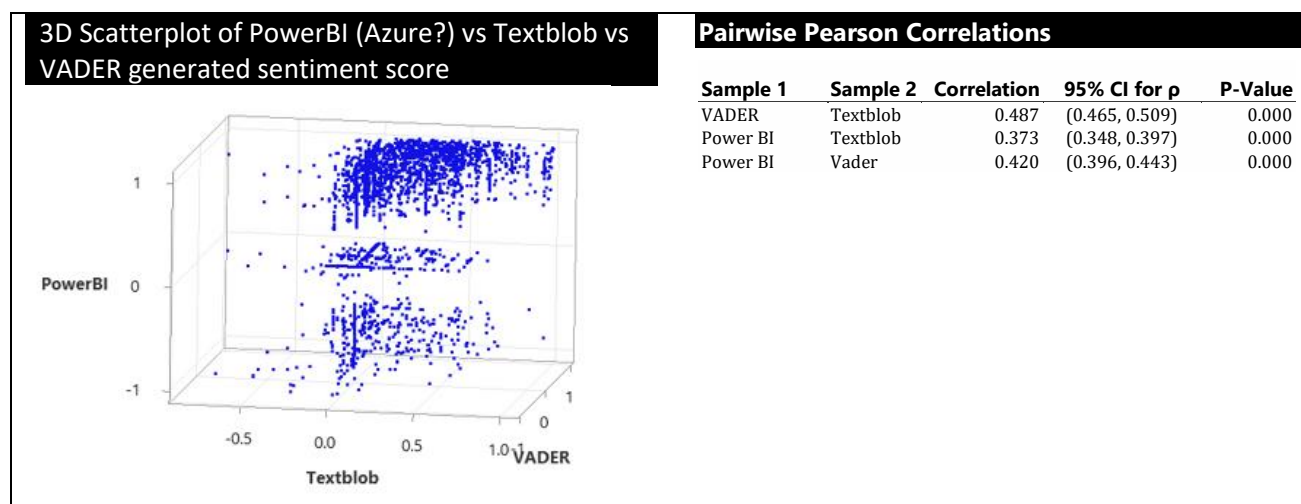


Figure 4 3D scatterplot of sentiment scores generated by PowerBI, VADER and Textblob for the responses to Q10: MOST VALUED and the pairwise correlations of the values

It is worth commenting that despite the text of Question 10 “*What aspects/ elements of your research degree programme are most valuable?*” asking for most valuable elements, students often used it as an opportunity to report issues, negative experiences, or failings which they experienced. While one would expect all positive or at least neutral responses if the answers given were strictly responses to the question posed. As such, the



sentiment analysis allows for an additional level of insight as it discriminates between responses and allows for topic analysis on a more granular basis.

To conduct a thorough assessment, the research team randomly selected 400 PGR responses from the overall survey and manually classified the sentiment as positive, negative, or neutral. Our objective was to test three sentiment analysis packages: Texblob, Vader, and PowerBI. The generated sentiment scores for each package; were compared with sentiments assigned by manually reviewing the responses. PowerBI achieved the highest accuracy at 60.26%, followed by Vader at 39.74%, and Texblob at 21.79%. On the basis that the Azure Power BI seemed to consistently perform better (based on the subjective opinion of the researchers when examining the qualitative statement), PowerBI was deemed the most appropriate model for this study Table 5 provides an illustrative sample of responses (n=20) for Q10 taken illustrating very positive, neutral and very negative responses and their associated scores from three sentiment scoring packages.

- ❑ **Examples 1 and 2** provide examples of the superior ability of the Power BI package to determine sentiment, even where the individual words may not be particularly positive.
- ❑ **Example 4** illustrates how the removal of punctuation (in this case an ASCII emoji) makes a large difference to the perceived sentiment leading to different classification by the three packages.
- ❑ **Example 7** shows how single words are assigned sentiment values. This can be problematic however in most cases a neutral value is correctly assigned.
- ❑ **Examples 14 to 19** reveals how VADER and Textblob can fail to recognise negative sentiment while they are correctly identified by PowerBI.
- ❑ **Example 20** shows a statement which could subjectively be considered neutral.

Table 5 Sample of responses to Q10: MOST VALUED and colour coded associated scores comparing sentiment score.

# True	Response to Question 10 (Q10: Most Valued)	Textblob	Vader	PowerBI
1 Positive 	"Doing it! Learning how look up research, research methods and then analysis methods has been invaluable. I had an 'instinct' about the worth of the project i am engaged in, and researching it in an academic way is helping me to describe and articulate this. I feel my research has the potential to be useful to others when it is finally complete/"	0.12	0.84	1.00
2	"The staff within the department are so generous with their time and knowledge. Before covid hit it was a wonderful collegial experience. I've learned so much from every staff member including administration staff. The string philosophical underpinning within the programme and the department was such a new experience and really opened up my research possibilities and thinking. Wonderful guest speakers invited in".	0.36	0.91	1.00
3	"Working with my supervisor has been the most valuable part of my programme, and the most consistent. Even with the lockdown, they have been fabulously supportive and have gone above and beyond to make sure that I am both working well and taking care of myself. They are the reason I stayed on, and there doesn't seem to be an upper limit on how much I learn from them.	0.31	0.91	1.00
4	Work in the lab. this is where magic happens ;)	0.38	0.23	1.00
5	Lots of opportunities to disseminate my research and network. Fantastic office, facilities, and I.T services. Easy access to the library, and other university resources. Great careers team nearby. Nice staff and fellow postgraduates. All very approachable, and understanding. I love the University sports facilities - the gym, clubs, and societies. I value the relationship with my supervisors.	0.40	0.97	1.00
6	The GRC is brilliant. My supervisor and departmental support are excellent. The ability to participate in research culture pre-COVID was also brilliant.	0.93	0.95	1.00
7 Neutral 	Training opportunities	0.00	0.38	0.00
8	Developing Transferable skills,	0.00	0.00	0.00



9	<i>The opportunity to teach other students.</i>	0.13	0.42	0.00
10	<i>Independence, supervisor support, modules provided, teaching experience</i>	0.00	0.40	0.00
11	<i>Too early to say</i>	0.10	0.00	0.00
12	<i>Interest</i>	0.00	0.46	0.00
13	<i>opportunity to take part in a specialist training, contact with other researchers, networking, computational facilities, mentoring</i>	0.13	0.42	0.00
14	<i>The part where I got almost enough money to put up with poor working conditions, lack of a social life and increasing costs of antidepressant medication arising from same. The rest just leaves a bad taste in my mouth after the appalling treatment I received.</i>	0.24	0.89	-1.00
15	<i>The courses on offer are very poor. One of my supervisors never has time to see me and no interest in my research. My other supervisor is moving away and since that decision was made is also not spending much time on my project anymore (despite being an excellent supervisor beforehand). My stipend is not enough to live and I am in a constant financial struggle. With all this together I am still giving all I can, but it is very hard to hold all that pressure and getting no support.</i>	0.03	-0.89	-1.00
16	<i>It used to be the ability to access resources otherwise unavailable to the general public. Unfortunately these are no longer available due to COVID.</i>	-0.04	-0.32	-1.00
17	<i>Being able to manage unsatisfying work and being unsupported in my work. Persevering</i>	0.50	-0.40	-0.99
18	<i>My supervisor is definitely most knowledgeable in my research field which excites me and motivates me to improve and succeed. However, our poor work relationship makes me regret ever starting my programme. When campus was open, the facilities available to students were most valuable (research hub, desk space, community contacts etc.) these unfortunately are difficult to replicate virtually</i>	0.00	0.67	-0.98
19	<i>The only thing I've learnt is academics lack social intelligence to the point in notably damaged my mental health, felt like no one had my back. You can only rely on yourself, The most valuable thing I learnt is how strong I am to persevere.</i>	0.17	0.71	-1.00
20	<i>The natural development of resilience to work stresses. It is completely overwhelming in the beginning (unimaginable amount of work to do in short timeframe). Once you get through that phase you learn to relax and realise that everything can't be done in a day.</i>	0.20	0.34	-0.99

Sentiment scores were generated by PowerBI using the text analytics functions. Some limitations worth noting are that as PowerBI sentiment analysis is a black box product, no user editable settings are available, and it is subject to updates. As a result, the exact scores provided may not be replicable however, as there are over 5000 responses, the analysis is robust to minor changes in scoring.

Objective 2: Apply NLP to 2018, 2019 and 2021 PGR StudentSurvey.ie data

This section explores the findings and insights of the sentiment analysis. Statistical tests are used to establish which factors return a statistically significant change in the observed sentiment of the responses to Q10 or Q11. This is then further investigated using topic modelling.

Covid pandemic was alluded to by respondents in the 2021 dataset this will be further explored later in the report using topic modelling. Table 6 presents the breakdown between positive, negative, and neutral responses to Q10 and Q11 for 2018, 2019 and 2021. In both cases, 41% of the responses are neutral.



Table 6 Descriptive statistics of PowerBI generated Sentiment Scores.

Question #	N	Mean	Median	Negative	Neutral	Positive
Q10. Most Valued	5,192	0.31352	0	595	2,111	2,486
Q11 Improvements	5,170	0.04467	0	1,356	2,134	1,680

Further statistical analysis of the scores shown in Figure 5, plotting of sentiment and response length for Q10 shows that a large number of neutral short responses exist and that the distribution of the scores is not normal. Since the sentiment scores do not follow a normal distribution, non-parametric tests were employed to compare the median values and identify any statistically significant differences. Removal of the short responses (i.e. less than 50 characters) was investigated and the resulting distribution of sentiment scores was checked. No change in distribution was observed, in all cases the resulting cases are not normally distributed. Since the arbitrary removal of short answers may induce a bias in the results, all answers were included in the sentiment analysis.

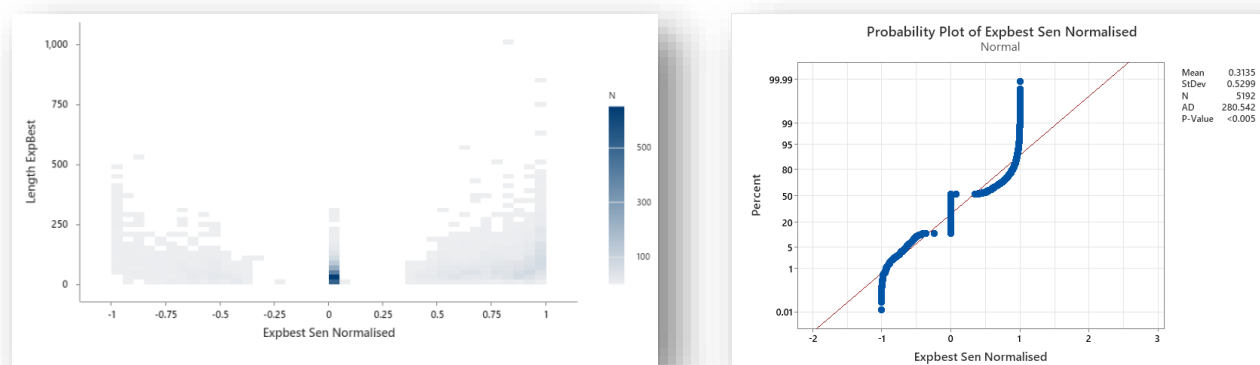


Figure 5 a) Binned Scatterplot of the response length vs sentiment b) sentiment distribution plotted against an expected normal distribution.

The Kruskal-Wallis H-Test was used where the data met the underlying assumptions required. Where the distribution of sentiment scores between groups was substantially different Mood's Median Test was used. Post-hoc testing was completed where needed, by pairwise comparison.

Initially, the data was compared to see if there was a statistically significant difference between the sentiment scoring between each year. The results in Table 7 show that there is no significant difference between the sentiment by year. This allowed sentiment scores for all years to be analysed together where appropriate to allow for greater statistical power.

Table 7 Kruskal-Wallis test of the sentiment scores for responses to Q10 compared by year

Fieldwork Year	N	Median	Mean Rank	Z-Value
2018	1605	0	2659.9	2.04
2019	1408	0	2589.0	-0.22
2021	2179	0	2554.7	-1.71
Overall	5192		2596.5	



Null hypothesis

H_0 : All medians are equal

Alternative hypothesis

H_1 : At least one median is different.

Method	DF	H-Value	P-Value
Not adjusted for ties	2	4.61	0.100
Adjusted for ties	2	4.94	0.085

On this basis, 13 different factors relating to various aspects of the student populations characteristics, including demographic characteristics, educational context, funding sources, and supervisory arrangements were explored. Table 8 and Table 9 show the results of the analysis using Kruskal-Wallis testing for the sentiment scores of responses to Q10 and Q11 respectively.

The results in Table 8 and Table 9 provide several insights. From an overview perspective, the sentiment students express in their responses to Q10 are much more dependent on other factors measured in the survey than responses to Q11. This outcome may be attributed to the question formulation, with Q11 requiring a less personal response. No significant difference was found for the sentiment of responses to Q11 when grouped by any of the factors examined.

This provides a useful control when considering potential causes of the differences in sentiment due to the factors explored. For example, there is a strongly significant difference in the sentiment expressed by male and female respondents to Q10 but not Q11. The cause of the difference in sentiment for Q10 is not because one gender expresses a generally more positive sentiment than the other as a result of their gender.

Table 8: The p values determined by Kruskal-Wallis H-test examining responses to Q10 for given factors.

Factor	N	Groups	p†	p adjusted for ties
Fieldwork Year	5192	3	0.1	0.085
Type of institution	5192	3	0.413	0.388
Year of Study	5192	12	0.966	0.956
Gender	5190	2††	<0.0005***	<0.0005***
Domiciliary of Origin (Two categories)	5192	2	0.436	0.42
Residence Type	4575	6	0.032*	0.023*
Student Code (Identifiers for distinct student types)	5192	5	0.706	0.677
Mode of Study	5192	2	0.047*	0.039*
Programme Type	5189	2	0.828	0.822
ISCED Level One (subject description)	5192	11	0.216	0.169
Stipend	5051	2	<0.0005***	<0.0005***
Self-funded	5189	2	<0.0005***	<0.0005***
Number of Supervisors	5164	3	0.063	0.051

†No adjustment for multiple comparisons has been made when declaring significance for these p values

†† One category was removed from analysis as it contained too few responses to allow for statistical analysis



Table 9: The p values determined by Kruskal-Wallis H-test examining responses to Q11 for given factors.

Factor	n	Groups	p ⁺	p ⁺ adjusted for ties
Fieldwork Year	5170	3	0.980	0.978
Type of institution	5170	3	0.242	0.218
Year of Study	5170	12	0.066	0.043
Gender	5169	2 ⁺⁺	0.336	0.318
Domiciliary of Origin (two categories)	5170	2	0.549	0.534
Residence Type	4558	6	0.102	0.079
Student Code	5170	5	0.750	0.723
Mode of Study	5170	2	0.252	0.234
Program Type	5165	2	0.515	0.500
ISCED Level One (subject description)	5170	11	0.146	0.107
Stipend	5025	2	0.274	0.256
Self-funded	5169	2	0.896	0.892
Number of Supervisors	5144	3	0.324	0.298

⁺No adjustment for multiple comparisons has been made when declaring significance for these p values

⁺⁺ One category was removed from analysis as it contained too few responses to allow for statistical analysis

For Q10 four factors show statistical significance in their association with the sentiment of the responses. These factors gender, mode of study (part/full time), stipend funding and accommodation type are further explored using topic modelling of respondents' opinions in the relevant section.

The question “Overall Experience: How would you evaluate your entire research experience at this institution?” was compared separately as it does not specifically relate to an individual attribute. This provides a useful insight into the relationship between best experience and overall experience. A statistically significant relationship between these questions and the sentiment in Q10 was found (Moods median test, $p = 0.005$) but not Q11 (Moods median test, $p = 0.067$). Pairwise comparison is shown in Figure 6 and demonstrates that the sentiment of response to Q10 has the expected relationship with those responding excellent, expressing the most positive sentiment in response to Q10. This result, though not surprising, does validate the sentiment analysis approach.

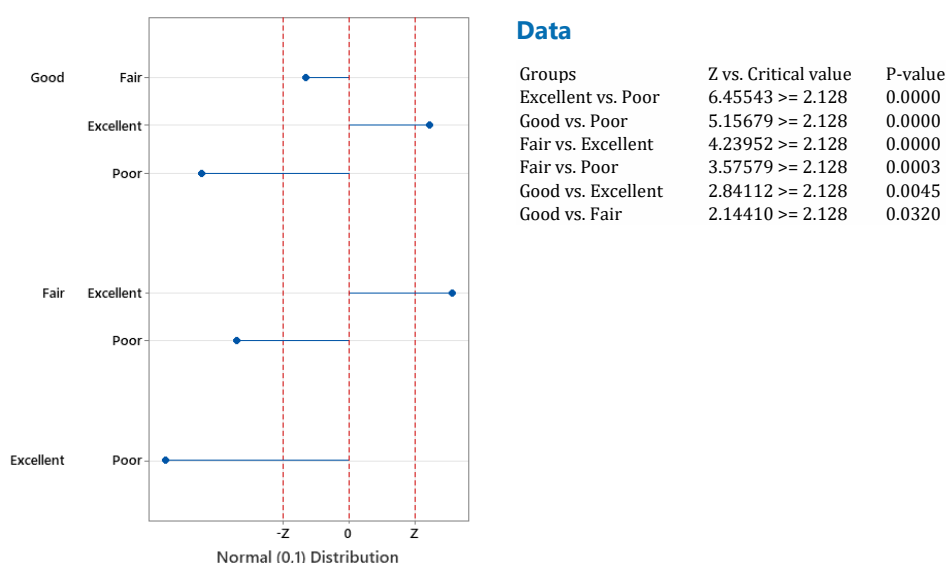


Figure 6 Pairwise comparison of sentiment of response to Q10 and rating of overall research experience



Topic Modelling

Topic modelling was carried out on the responses to Q10 in line with the methods set out earlier. This approach provides a clear visualization of the prevalent topics and the associated words, offering insights into the focus areas addressed by the respondents in their answers to Q10. The topic modelling revealed one dominant topic and three further main topics within the student responses.

The dominant topic emerging from responses to Q10 relates to supervisor – student relationship. The results of this analysis are presented in Figure 7. The utilization of a word cloud and distribution table to showcase that the dominant topic within question 10 is represented by the terms "Supervisor" and "Research". The analysis assumes that question 10 covers multiple topics, and each topic can be identified by a distribution of words. By visually representing the word cloud and presenting the distribution table, the study highlights the dominant themes. The LDA topic modelling analysis reveals a prominent subtopic related to supervisors in the dataset, indicated by a score of 32.25. This high score highlights the significance and prevalence of this subtopic compared to others identified in the analysis. The frequent occurrence of terms and words associated with supervisors provides valuable insights highlighting the value of the working relationship between the student and supervisor is what matters. Illustrative quotes draw attention to the negative impact when a supervisor support is lacking.

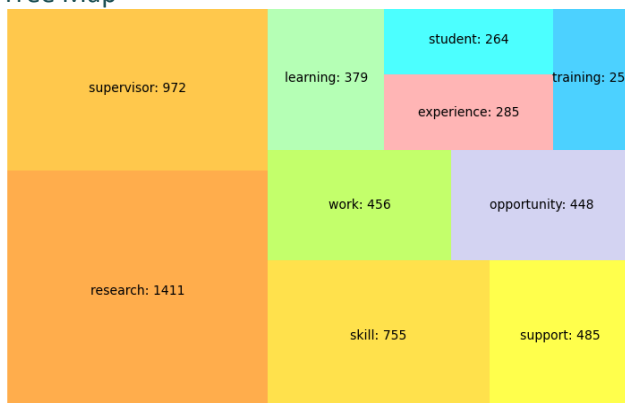


	Sentiment Score: 0.93 Experience and expertise of my supervisor - International connections developed through my supervisor - Learning to become an independent researcher
Negative 	Sentiment Score: -0.9998 "Lack of community in the department lack of support from supervisor. The feeling of isolation I don't know who to turn to for help". Sentiment Score: -0.98 My supervisor is definitely most knowledgeable in my research field which excites me and motivates me to improve and succeed. However, our poor work relationship makes me regret ever starting my programme. When campus was open, the facilities available to students were most valuable (research hub, desk space, community contacts etc.) these unfortunately are difficult to replicate virtually.

Figure 7 Dominant Topic Supervisor with illustrative quotes of positive and negative sentiments

The tree diagram and semantic network graph visualizations in Figure 8 provide valuable insights into the dominant themes and interrelationships within the topic distribution analysis. The tree diagram showcases the hierarchical structure and proportions of topics, with larger rectangles representing more significant or influential topics. On the other hand, the semantic network graph focuses on illustrating the relationships between topics, highlighting the connections and associations between different clusters. These visualizations again illustrate the importance of supervisors and research topics in shaping the overall topic distribution, aiding in the identification of key factors and themes in the data.

Tree Map



Semantic Network graph

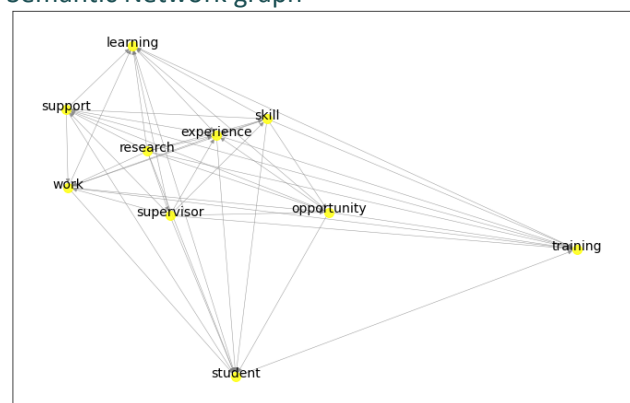


Figure 8 Graphical representations of the topic modelling completed with LDA

The identification of the supervisor as a dominant topic aligns with the findings from previous thematic analysis and word frequency analysis conducted. Using the sentiment analysis of the response to Q10, we can explore how this student supervisor relationship relates to the students reported most valuable experience. Table 6 shows statements relating to supervision and research found to have a statistically significant relationship with the sentiment expressed in response to Q10.

Table 6 Statistical analysis of responses to supervisor support and feedback using Moods Median test

Statement relating to supervisor	Code	P value	Significant
My supervisor(s) provides the appropriate level of support for my research	SSUPP	0.013	*
I have regular contact with my supervisor(s), appropriate for my needs	SCONT	0.014	*
My supervisor(s) provides feedback that helps me to direct my research activities	SFEED	0.003	*



Pairwise comparison of each was carried out. Where the large majority of answers are either mostly or definitely agree, no difference can be established for the other responses however, this does not suggest a relationship either way (i.e. it cannot be inferred that students who respond with a mostly disagree have a more positive sentiment, rather just that there is insufficient data). The analysis suggests that there is a significant and positive relationship between the perceived level of support provided by supervisors, the frequency of contact, and the effectiveness of feedback in directing research activities and the sentiment observed when the respondents describe the most valuable aspect of their research studies. Importantly, this is also true and equally strong in significance when responses to Q10 including supervis* are excluded (where * includes any ending to the word). This excludes the possibility that responses which mention supervisor are generally more positive and suggests that a positive student supervisor relationship results in an overall better sentiment about students most valuable experience, whether that best experience directly involves the supervisor. This adds further significance to the weight to the importance of the student – supervisor relationship.

Additional subtopics were identified through topic modelling. These topics relate to Topic 1: “*Transferable Skills*”, Topic 2: “*Research Skills Training*”, Topic 3: “*University life balance and satisfaction*” as illustrated in Figure 8. This dashboard enables users to use the slide the gauge to explore the polarity of responses in each sub topic.

Figure 9 Screen shot of interactive Q10 Topic representations with illustrative quotes and sentiment slider.

In identifying and exploring this subtopic of transferrable skills, researcher students are aware of the need to develop these skills and that they place a high relative value on opportunities to develop them. Both the provision of opportunities to engage in formal development such as taught modules and the informal opportunities available through work such as teaching are acknowledged by respondents.



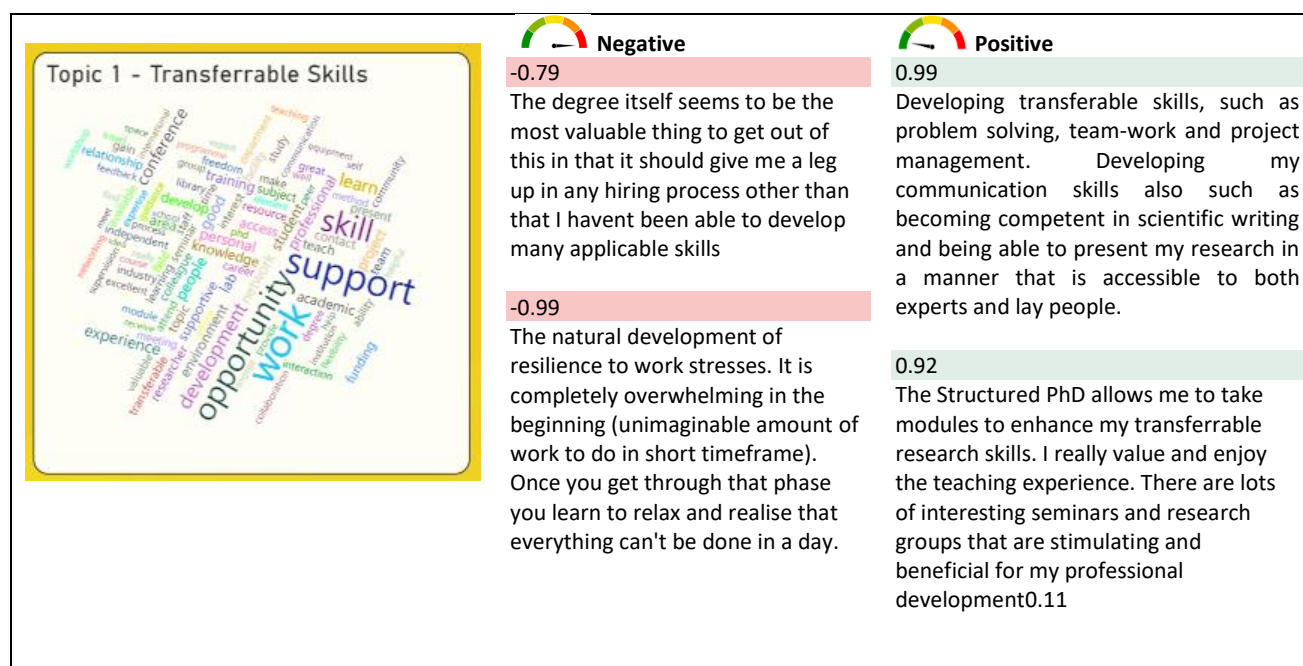


Figure 10 Word cloud for Topic 1 - Transferrable Skills with illustrative quotes.

The identification of transferable skills as a key topic within the space of most valuable experience is further borne out by the examination of the relationship of the sentiment of the responses to Q10 and their relationship to their actual and perceived development of their transferable skills as measured by other questions in the survey. Table 7 shows the five questions related directly to this topic for which there is a statistically significant relationship to the sentiment expressed in Q10.

Table 7 Relationship of Sentiment of Q10 to responses to statements relating to Transferrable Skills and Training

Statement relating to Transferrable Skills and Training	Code	P value	Significant
My skills in critically analysing and evaluating findings and results have developed during my programme	RSANALYSE	0.001	*
My confidence to be creative or innovative has developed during my programme	RSINNOV	<0.0005	*
My ability to manage projects has developed during my programme	TSKPROJ	0.001	*
My ability to communicate information effectively to diverse audiences has developed during my programme	TSKCOMM	0.002	*
Working as part of a team	DOTTEAM	0.002	*
I have frequent opportunities to discuss my research with other research students	RCOPP1	0.369	No

The analysis specifically suggests that as well as being widely acknowledged as a significant part of students' research experience, the perception of developing transferable skills has a direct impact on the positivity expressed by students in how they reflect on their most valuable experience.

This highlights the significance of transferrable skills not just as key to career development but also as impacting on students' perception of their research experience. Specifically, the feeling of developing skills in critical analysis and evaluation, being creative or innovative, managing projects, communicating effectively to diverse audiences and working as part of a team positively impact this perception. These findings underscore the positive influence of ensuring research environments offers opportunities to enhance these skills, equipping participants with valuable abilities that can be utilized in diverse contexts beyond their current



research project. Interestingly, the opportunity to discuss research with other students is not related to their most valuable experience and suggests that teamwork is perceived as more than just working in an environment which has other researchers.

In terms of tangible demonstration of the acquisition of transferable and subject specific skills, the publication of papers and presentation at conferences is strongly linked to a positive most valuable experience. While these are also markers of success in other ways they are strongly related to the development of both subject specific and transferable skills.

Table 8 Statements relating to dissemination

Statement relating to research/training	Code	P value	Significant
Presenting a paper or poster at an academic research conference	DOPRESENT	0.002	*
Submitting a paper for publication in an academic journal or book	DOPUBLISH	0.014	*

Topic 2 – Research Skills Training

The training received by postgraduate students is highly valued. In combination with the theme of transferable skills, it shows that students undertaking post graduate research value formal training. Figure 11 shows the word cloud and illustrative quotes showing how training is viewed positively and the lack of structured training provided by the university has negative sentiment.

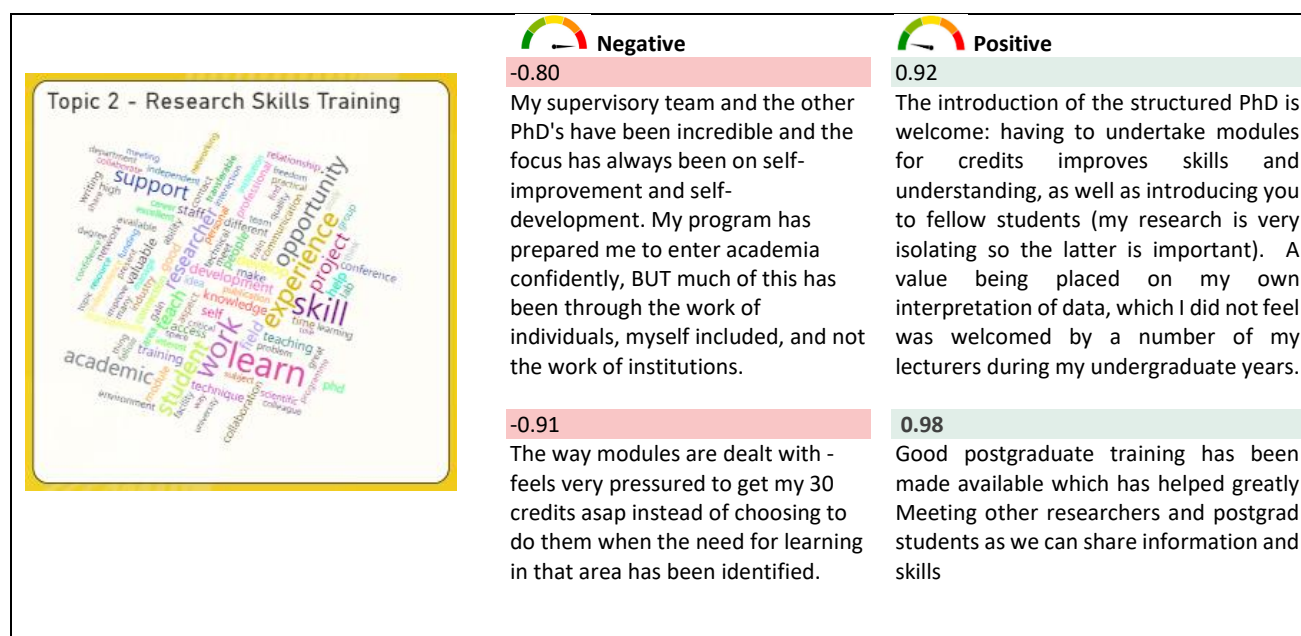


Figure 11 Word cloud for Topic 2 –Research Training and Skills Development with illustrative quotes

In addition, this is further evidenced by the sentiment analysis of Q10 when looking at whether students undertake training and receive career advice. Availing of training and receiving career advice results in a more positive response to Q10. Interestingly, agreeing on a training or development plan has no discernible impact. This may be worth exploring in further studies when considering the utility of this activity as presented in Table 9.



Table 9 Statements relating to Training

Statement relating to research/training	Code	P value	Significant
Receiving training to develop my research skills	DORESSKILL	0.003	*
Receiving training to develop my other transferable skills	DOTRANSK	0.019	*
Receiving advice on career options	DOCAREER	0.014	*
Agreeing a personal training or development plan	DOPERDEV	0.946	ns
Taking part in a placement or internship	DOINTERN	0.832	ns

Topic 3 – University Life Balance/Satisfaction

The final topic which comes out of the responses to Q10 is one of work life balance and work satisfaction. This has different factors such as institutional supports, clarity of purpose and meaningful work. Associations with factors with sentiment indicating that the awareness of student supports, overall life satisfaction, and satisfaction within the institution impact on the sentiment expressed by the respondents.

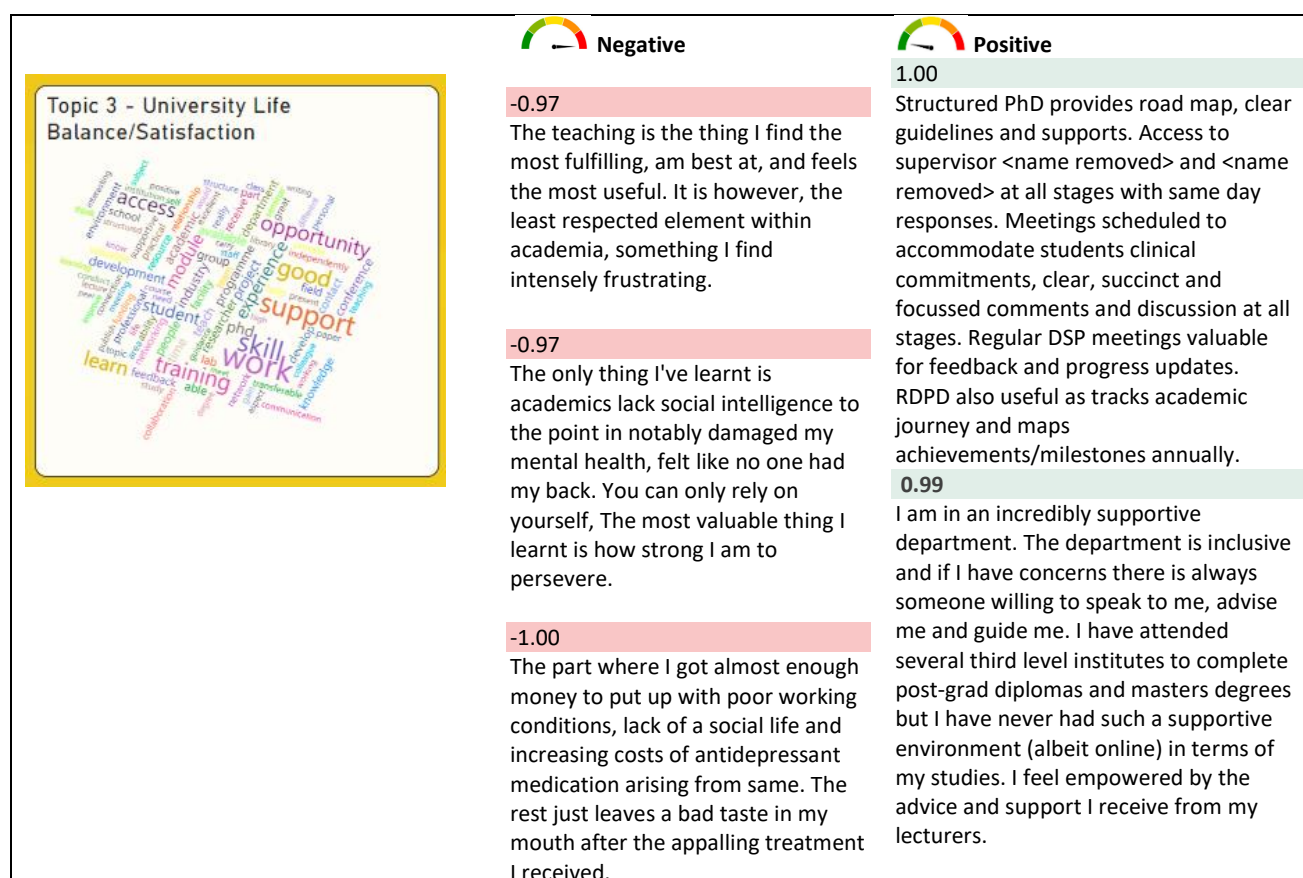


Figure 12 Word cloud for Topic 3 – University and Work Life Balance with illustrative quotes

These are reflected in the relationship of the sentiment scores of responses to Q10 to questions about each of these factors. Table 10 shows that having a clear understanding of the expectations which they are under, both in terms of direct assessment and more generally as a research student is important. An awareness of student supports within the environment is also positively associated with a higher sentiment score for the reported most valuable experience.



Table 10 Statements relating to assessment and supports

Statements relating to assessment and supports	Code	P value	Significant
I understand the required standard for my thesis	PASTD	0.008	*
The final assessment procedures for my research degree are clear to me	PAPROC	0.003	*
I understand my responsibilities as a research degree student	RSSTU	0.002	*
How aware are you of the various student supports available? (Recreation, healthcare, counselling, etc)	RSSUPPORTS	0.013	*

The sentiment score of the best experience is also related to the students overall life satisfaction and satisfaction within the institute. That is to say, students who are more satisfied with their lives report more positive experiences. This finding is not surprising but reassuring as it speaks to the validity of this analytical process.

Table 11 Statements relating to Personal and University Life

Statements relating to Personal and University Life	Code	P value	Significant
I am satisfied with my life nowadays (Do you agree or disagree with the following statements about your personal outlook?)	POLIFENOW	0.009	*
I am satisfied with my life within my institution nowadays (Do you agree or disagree with the following statements about your personal outlook?)	POLIFEINSTNOW	0.002	*

Topic modelling within subgroups

When conducting the statistical analysis of the sentiment scores, it was noted that four factors showed a statistically significant relationship to the sentiment of the responses to Q10. In this section we explore these factors in turn, looking at differences in the topics expressed by each group.

Factor 1 - Gender

Exploring sentiment distribution within each gender category allows for a comparison of the sentiments expressed by males and females. This identifies potential variations or trends in sentiment based on gender, providing valuable insights into the emotional perceptions and experiences of different gender groups.

The breakdown of Table 12 provides an overview of the gender demographics for groups across the 3 years of data.

Table 12 Breakdown of respondents by gender

Fieldwork Year for Survey	Gender		
	FEMALE	MALE	Undeclared/ other
2018	1711	1272	
2019	1621	1100	
2021	2091	1448	2
TOTAL (9245)	5423	3820	2



It should be noted that two respondents did not declare their gender in 2021 data. To fully capture the opinions and sentiment the responses to Q10 are presented in Table 13. One respondent expressed that they highly valued transferrable skills when it comes to their university experience. Another emphasized the importance of critical thinking and possessing a broad knowledge base to effectively navigate the complexities of their academic discipline.

Table 13 Undeclared Gender

Year	Gender	Sentiment	Quote
2021	Undeclared	Neutral 	Transferable skills
2021	Undeclared	Positive 	Critical thinking and having a broad awareness of the theory and practicalities of my subject

Figure 13 illustrates the sentiment responses categorized by gender, specifically distinguishing between male and female respondents. Of the remaining **9243** participant, 5190 responded and these can be categorized into two groups based on their gender (male and female), there is no need for a post-hoc test.

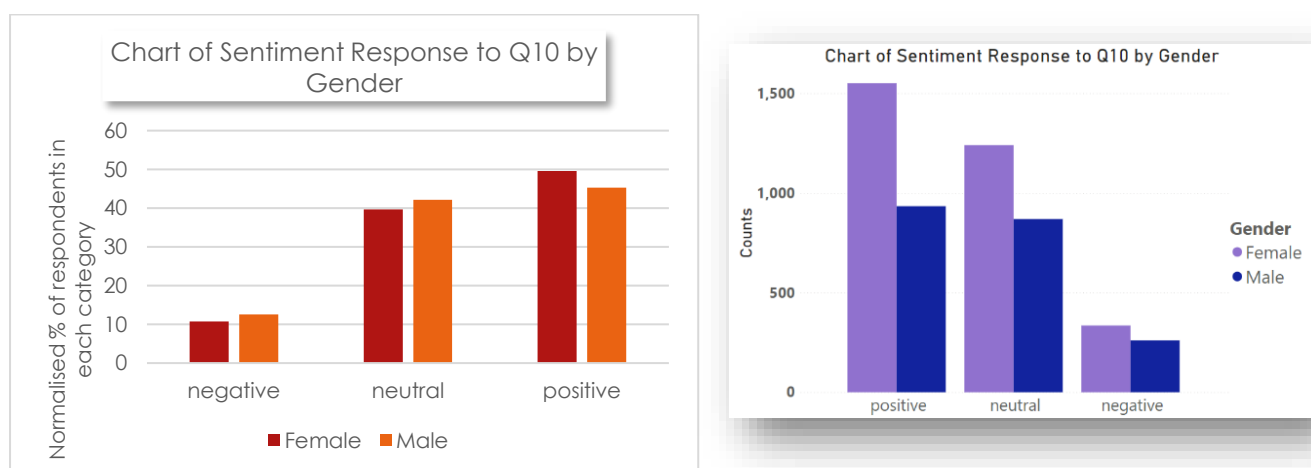


Figure 13 A chart showing sentiment responses by gender categorised into Male and Female

Table 14 shows the result of the Kruskal-Wallis test which compares the sentiment scores between two groups, Female and Male. **The test shows that female respondents have a significantly more positive response to Q10 ($p < 0.0005^{***}$) than male respondents.**

Table 14 Kruskal-Wallis Test: Q10 Sentiment versus Gender

Gender	N	Median	Mean Rank	Z-Value	Method	DF	H-Value	P-Value
Female	3127	0	2661.1	3.88	Not adjusted for ties	1	15.07	0.000
Male	2063	0	2496.1	-3.88	Adjusted for ties	1	16.15	0.000
Overall	5190		2595.5					



The interpretation of this finding requires further study however, some insights are possible. One hypothesis would be that female respondents use more positive language; however, this can be ruled out as no difference in sentiment is seen in the response to Q11 between male and female respondents. It may be that female students have more positive most valuable experiences, or that female students simply report more positive experiences when considering which elements of research degree programme are most valuable.

Further analysis using topic modelling illustrated in Figure 14 highlights the importance of relationships and opportunities for females in developing their research work and advancing career skills. The absence of this highlights how discouraged and isolated females can feel when compared with the confidence gained through making meaningful research contribution illustrated by females with positive quotes.

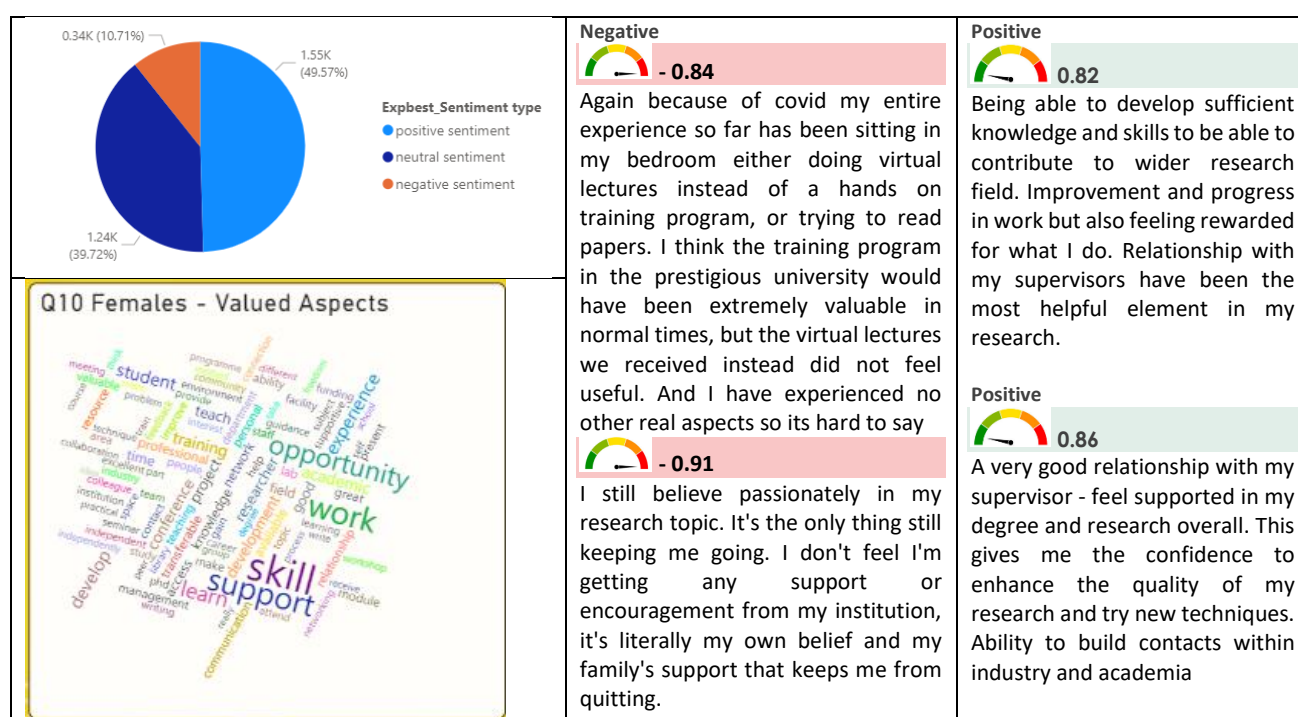
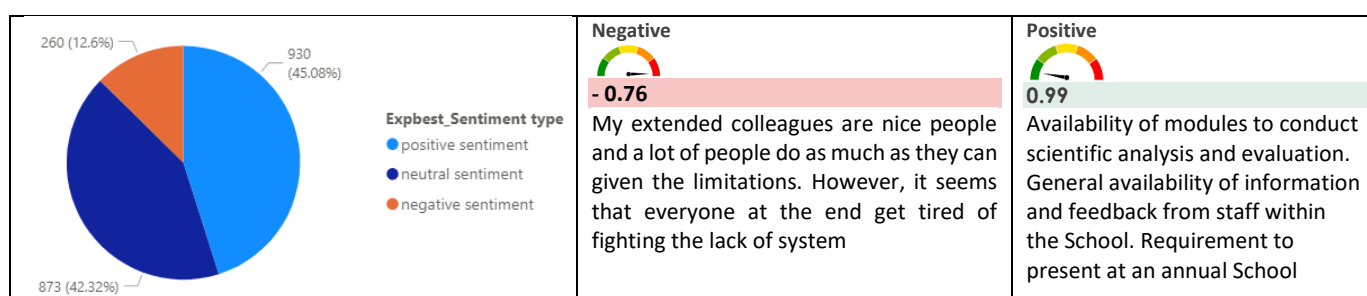


Figure 14 Word cloud and sentiment score breakdown for female respondents with illustrative quotes

Positive topics that emerged for males shown in Figure 15 also focused on relationship to develop peer and support networks while negative comments recognised the limitations of systems, facilities and opportunities.



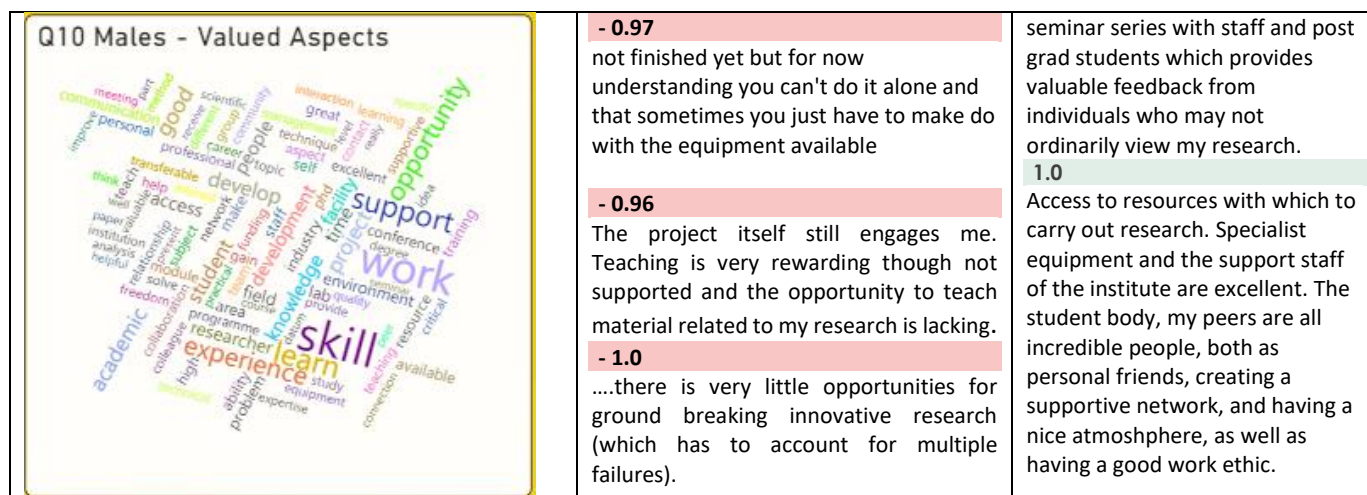


Figure 15 Word cloud and sentiment score breakdown for male respondents with illustrative quotes

The complementary PGR interactive dashboard allows users to explore notable differences in sentiment distribution based on these characteristics. For example shown in Figure 16, in 2021 data of 192 full time males studying natural sciences PhD programme who responded to Q10, 40% conveyed a positive sentiment.

Figure 16 Decomposition Network of Full Time Males Science Students and Sentiment scores.

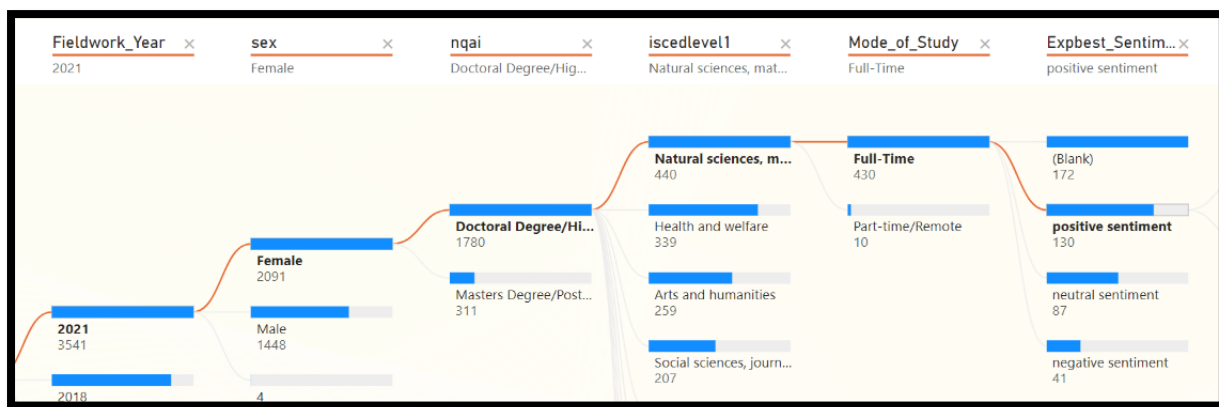


Figure 17 Decomposition Network of Full Time Females Science Students and Sentiment scores.

For a full summary of the interactive decomposition network is available in Appendix 1.

Factors 2 & 3: Accommodation Type and Study Mode

Initial analysis of accommodation type included all responses except for blank ones. The analysis showed a significant result with a p-value of 0.032*, indicating a difference in sentiment based on residence type. The subsequent analysis excluded the "not specified" responses and still yielded a statistically significant result with a p-value of 0.023*. Pairwise comparison revealed that the only significant difference was between **students who own their own home and those in rented accommodation**, with a p-value of 0.003**.

Students who own their **own home** had **significantly less positive responses** compared to those who rent. Based on the topic analysis, this is a counterintuitive result however it has a number of plausible explanations, some of which are deserving of additional study. Firstly, a Chi-Square test indicated an association between residence type and study mode, with a **higher proportion of students living in their own accommodation studying part-time**.

Part time study is strongly associated with a reduced sentiment score for responses to Q10 (Table 15). On this basis, for the purposes of assessing the impact of accommodation, full and part time students were examined separately.

Table 15 Statistical analysis of sentiment scores for Q10 by mode of study and bar chart showing distribution

Mode_of_Study	N	Median	Mean Rank	Z-Value	Method	DF	H-Value	P-Value
Full-Time	4615	0	2611.1	1.99	Not adjusted for ties	1	3.96	0.047
Part-time/Remote	577	0	2479.5	-1.99	Adjusted for ties	1	4.24	0.039
Overall	5192		2596.5					

Additionally, based on national demographic data, students who own their own house are likely to be of significantly different demographic profile to those who rent. The possibility of these demographic factors being important in determining the experience of postgraduate researchers is worth of further study. Figure 18 shows the distribution of students across accommodation types by study mode.



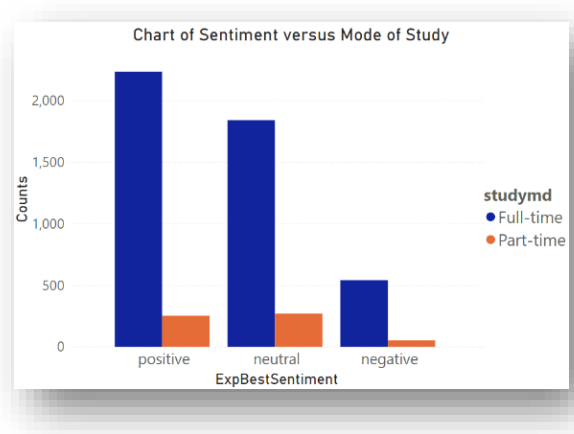
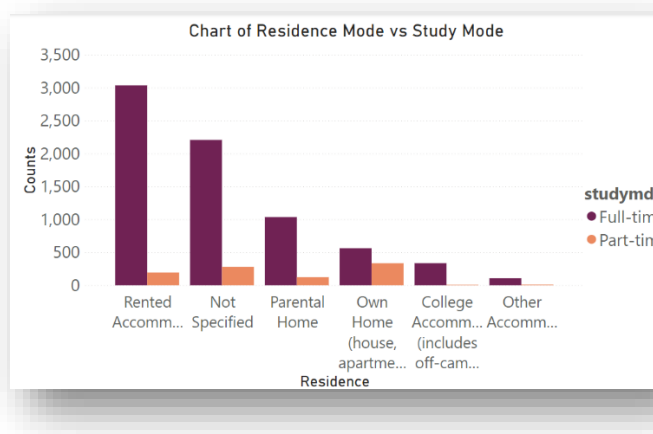
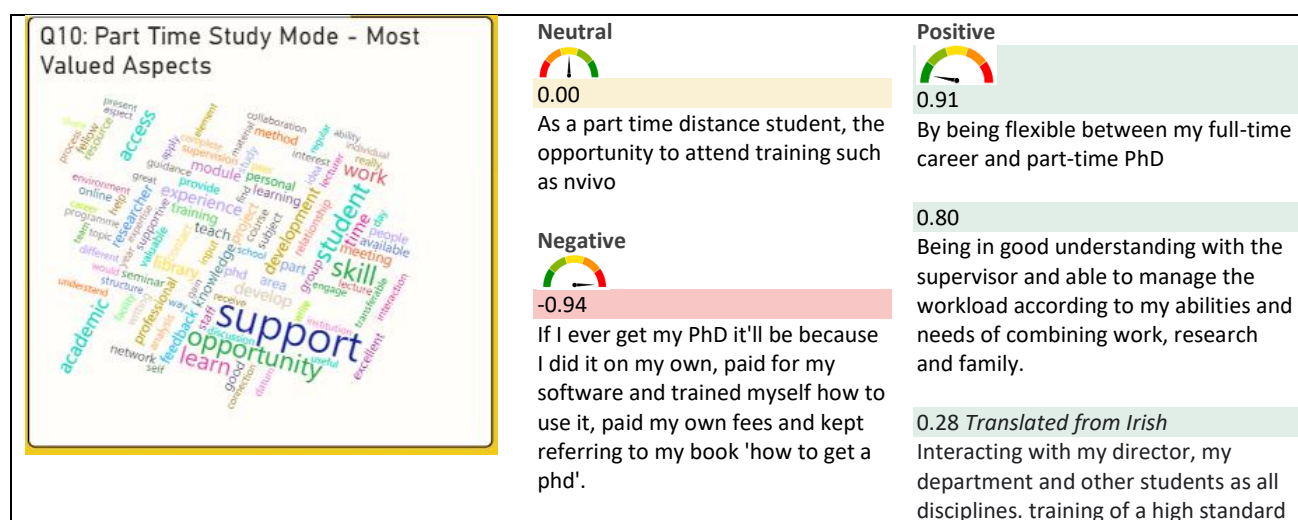


Figure 18 The distribution of students between accommodation types vs study mode

Taking the full-time and part-time students were separately, no significant difference was observed in sentiment for part time students when separated by the factor of accommodation type. This may simply be due to the low number of part time students resulting in an insufficient sample size.

When **full time students** were considered separately, responses to Q10 by those in their **own accommodation** were shown to be less positive than those in **rented accommodation** ($p = 0.0012^{**}$), college accommodation ($p = 0.0151^{*}$) and the parental home ($p = 0.0159^{*}$). The same possible explanations still apply, with the national demographic of homeowners significantly different to those who are renting.

Full-time and part-time students are likely to have a very different experiences completing research. Figure 19 show the results of topic modelling used to gain insight into the different experiences and challenges faced by these two distinct student cohorts. Part time students refer to the importance of flexibility to manage family and work life, have access to campus facilities, training and feeling supported and as part of a community of researchers. These topics as pertinent for full time students where the topics are broadly reflective of those coming through from the topic modelling of all responses. This is unsurprising given that part time students are a minority (approx. 12% of respondents) and suggests that particular attention may be required for the part time cohort who have unique needs which are not reflected by the general cohort. This suggests that to ensure an inclusive and representative research environment, attention should be paid to specific subsets of students and this is an area for further study.

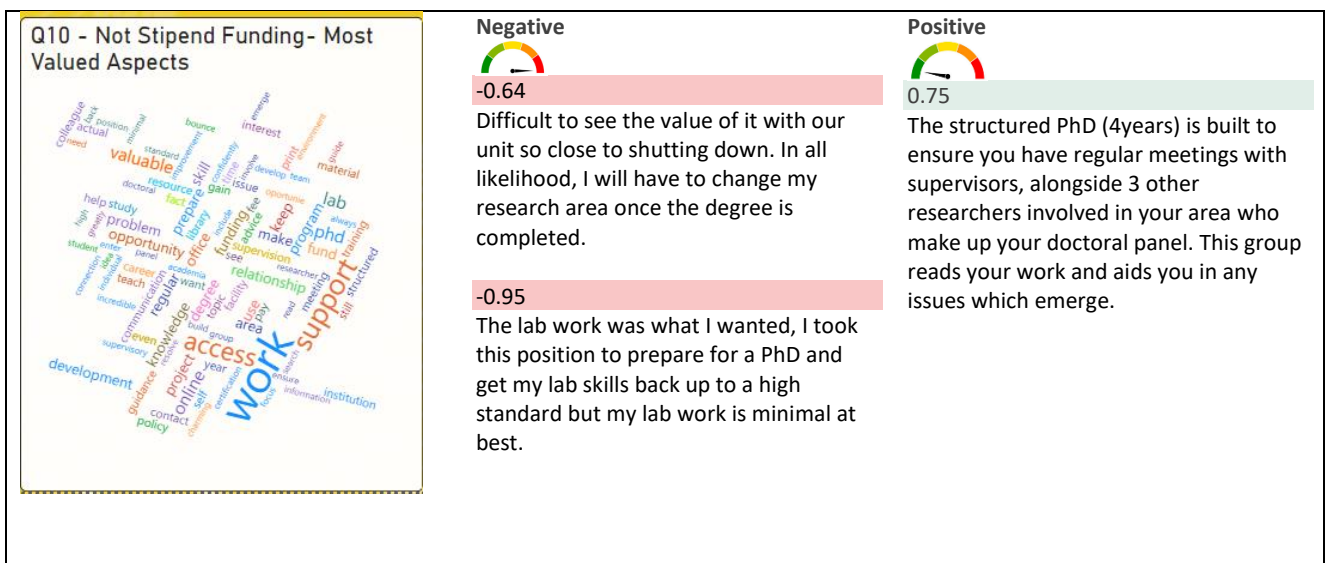


[illegible]

Figure 19 Word cloud for full and part time respondents with illustrative quotes

Factor 4: Stipend Funding

The final factor shown to cause a statistical difference in the sentiment of the responses relates to whether they are in receipt of a stipend and whether they are self-funded. That funded students have a more positive most valuable experience is in line with what would be expected. Given the current cost of living crisis, it might be expected that the topics which dominate unfunded students will start to also affect funded students to some extent. Based on the topic modelling, the funded students place a greater emphasis on skill acquisition and seem to be more confident in their position as students. Unfunded students on the other hand place a much greater emphasis on support and academic experiences conveying the importance of being able to pursue areas of interest, avail of research and teaching opportunities and support a lifestyle.



Furthermore, the topic modelling analysis revealed that certain factors explored in Q10 exhibited statistically significant relationships with the responses provided in Q11. This indicates that the topics identified in Q10 were also evident in the responses given to Q11, highlighting the interconnectedness and relevance of these factors in the dataset. We can as a result show that many of the topic areas identified in Q11 can be shown have an actual effect on the most valuable experiences of the student cohort as assessed through sentiment analysis. This provides further substantiation for these topics as matters, the addressing of which, would have a positive impact on the postgraduate research experience.

The utilization of treemaps, along with semantic network graphs, enables researchers to efficiently evaluate the relative importance and distribution of categories within the dataset as illustrated in Figure 21. This provides a visual overview of the contributions of different elements and their interconnectedness, enhance the understanding of the dataset's composition, supporting the identification of key areas for further investigation such as funding and structured training programmes.

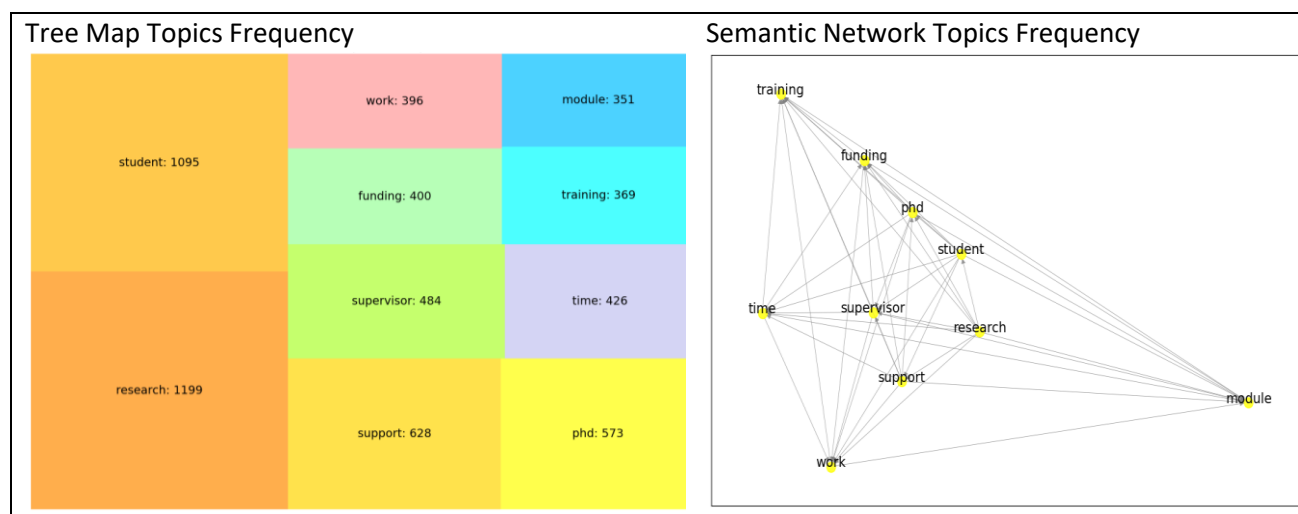


Figure 21 Topic frequency and semantic network for responses to Q11

Dominant Topic - Funding

This sophistication of LDA to extract topics that may not be immediately evident or intuitive goes beyond simple keyword-based categorization and uncovers more nuanced and complex relationships between words and documents. LDA is particularly useful in discovering latent or emerging topics, as well as identifying overlapping or related themes that may exist within the dataset. With Q11, the distribution values are explored for understanding the dataset's composition, these values represent the probability of a topic appearing in a document. For example, a value of 12.96 for the topic "funding" indicates its dominance, while a value of 8.64 for "communication" suggests its importance. By examining these values, researchers gain quantitative insights into topic significance, aiding analysis, theme identification, and uncovering underlying patterns.

The dominant topic of funding indicates that respondents expressed a need for better financial support and resources to enhance their experience. This includes increased funding opportunities, scholarships, or grants to maintain a reasonable quality of life as illustrated in Figure 22.



post graduate research experience and given its dominance in Q11, is one which many researchers feel needs to be addressed.

Other Topics: Suggested Improvements - Communication, Research Environment

Topic 1: Communication



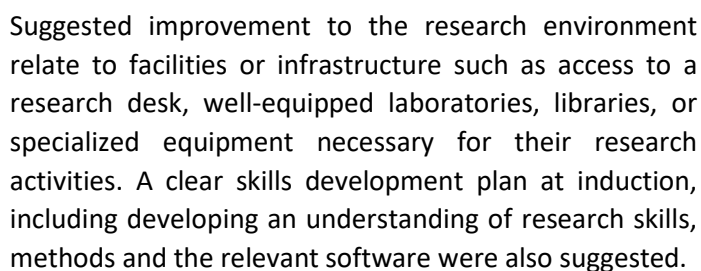
Topic 1 - communication relates to respondents' sentiments regarding the importance of effective communication channels and strategies within their research or academic environment. The suggested improvements include clearer communication about roles and responsibilities, improved information dissemination on procedural requirements and collaboration opportunities. Figure 23 shows illustrative quotes for this topic.

illustrative Quotes	
Positive 	0.89 My Roles/responsibilities explained at the beginning in relation to teaching and personal progress. - What progress I need to make and when by. - What to priorities along the way to make myself employable as soon as I finish. give postgraduate guide sheet beginning help regard contact go get set finance system would made life much less stressful
Neutral 	0.04 I would have really liked to be more involved in multidisciplinary discussions within my institution (every month I find out about something really interesting I had no idea existed, and I have been here for many years) - I would have really liked to have had a better initiation to my the PhD process, from the side of my department and from my supervisors and these were either non-existent or very weak. Which made me loose very precious time. - Other countries have a better centralized support to PhD students related to a) the management of their scholarship funds; to support them in finding more if necessary (in or out of the institution) b) the procedural aspects of the PhD such as forms and other relevant paper work and they also provide access to the internal network of academics as well as tools or materials that might be useful for research.
Negative 	-0.85 1) I would like to receive more support from my supervisor, including support to publish in academic journals 2) I would like to have more adequate funding system, because my scholarship is for 3 years, but the programme itself is 4. That's absurd. 3) I would like to be informed about conferences , funding , exchanges, publication opportunities in other universities that is relevant to my research. Because 20% of my time I spend on searching and looking for these on my own -0.99 More transparency about how the university system works. I was without a paycheck for three months, as HR did not inform me directly that my paperwork had expired. They just stopped processing my claims.

Figure 23 Word cloud for communication as a topic within Q11 with illustrative quotes



Topic 2 uncovered several latent topics that relate to the university research environment capturing the range of sentiments relating to appropriate work environment, induction and training opportunities. Suggested improvements highlight the importance of feeling valued and treated with respect by the university ensuring they have an appropriate positive working environment, facilities, and support structures. Respondents expressed a desire for better training opportunities to develop project planning, or collaborative initiatives to optimize their research experience. Figure 24 provides illustrative quotes.



The lack of funding, resources and facilities requires students to undertake additional paid employment and work from spaces that do not encourage a positive collegial work practice.

0.99	A clearer understanding of the degree process and required milestones through orientation and supervisor input, more opportunities to connect and engage with other researchers in less formal contexts to feel more connected and supported and to brainstorm, more training regarding research methods and proper research processes, more support
0.99	A clear skills development plan at induction, including developing an understanding of research skills, methods and the relevant software. I think students could benefit from having a week or two weeks of full skills training when beginning their research. It would also create opportunities to develop stronger links with other researchers starting at the same time
0.98	- Availability of working areas for graduate students - More social interaction with the department and other graduate students (esp. in my department) - More social spaces in the arts block, especially for eating lunch - No paying for printing and especially scanning (being able to create pdfs of book chapters and physical journal articles available would really improve my research experience)
-0.99	Lack of facilities, e.g., prototyping labs, design, and ideation spaces. Lack of connection between departments
-0.99	There is very limited workspace available in my institution, and when there is, the computers and wifi dont work. As a result, I find myself forced to work at home on my laptop which is very difficult.

Figure 24 Word cloud for university environment as a topic within Q1 1 with illustrative quotes

While covid does not emerge as a topic in the data its impact is relayed in the 2021 data relating to the context of reduced access to facilities and peer support network. Illustrative quotes offer a snapshot of the sentiments for the 2021 data however this no longer persists as an issue in postgraduate research.

Male full-time respondent: *Academic writing, Learning to construct scientific arguments. I also value the social aspect of the research degree programme - it has affected me that I haven't seen the other postgraduate students in person in 11 months.*

Male full-time respondent: *The collaborative mentality of scientists in the Moyne. We all look out for each other and help each other all the time. I'm constantly learning from being around my peers. Unfortunately*



that has been greatly reduced due to covid. Its very understandable though. I'd rather see less people in work than know that our hospitals are even more of a war zone than they already are.

Female full-time respondent: *It's difficult to say now due to covid and plans are messed up and it's extremely difficult to work from home while I need to be in the institute to collect data. However, I am grateful for the opportunity of having a funded PhD. I do love research in the lab.*

Discussion and Conclusions

By applying NLP techniques such as sentiment analysis and LDA topic modelling to the survey data, it has been possible to both gain new insight into what is required for students to have a positive experience when carrying out postgraduate research as well as find new evidence to support well established paradigms in postgraduate student support (Share et al., 2021; Lacey, 2021).

It is hoped that this study will be a timely contribution to the national discourse as according to the Government of Ireland's publication "Impact 2030: Ireland's new research and innovation strategy" (2022), submissions are being encouraged on the supports the State should provide to PhD researchers. The anticipated recommendations of a national review of State supports for PhD researchers include raising their pay to €25,000, aiming to retain and support research talent in the context of rising inflation and cost-of-living challenges.

It is clear that the supervisor-student relationship is a defining aspect of all postgraduate research. Topic modelling shows that it remains a dominant theme when students report their experiences. Sentiment analysis shows that the effect of the supervisor – student relationship extends far beyond the relationship itself and has a significant effect on the students' positivity around their most valuable experiences (Eisenbeiss, Knippenberg, & Boerner 2008; Ryan & Deci 2000). The perceived level of supervisor support is significantly and positively related to respondents' sentiment scores, highlighting the importance of frequent contact and effective feedback in guiding research activities. A positive working relationship with supervisors contributes to more positive and valuable experiences for researchers.

Research skills development is significantly associated with reporting a positive most valuable experience. Skills such as presenting work at conferences, submitting papers for publication, understanding thesis standards, clarity of assessment procedures, receiving research skills training, and understanding responsibilities as a research degree student all contribute to a more positive research experience (Thomas, L., & Quinn, J. 2007; Lovitts, B. E. 2005). Respondents report this as a valuable experience in Q10. when structures are in place within universities and suggest it as a development in Q11 where it absent.

The exploration of the subtopic pertaining to transferrable skills offers interesting insights, revealing students are highly aware of the importance of developing skills that can be applied in diverse contexts beyond their current research projects. The findings prioritize both formal and informal opportunities to enhance them such as engaging in taught modules and teaching activities. By fostering an environment that promotes the acquisition of transferrable skills, research institutions contribute to the holistic growth and professional development of research students, enabling them to thrive in various domains and make meaningful contributions to their fields.

The sentiment analysis conducted on responses across 13 different factors revealed statistical significance in the association between sentiment and factors such as gender, mode of study, stipend funding, and accommodation.



The factors included:

- ☐ **Gender:** Female respondents displayed a significantly more positive response to Q10 compared to male respondents.
- ☐ **Accommodation Type:** Students who live in their own home exhibited significantly less positive responses compared to those who rent.
- ☐ **Mode of Study (Full/Part-time):** Students studying part-time had significantly less positive responses compared to those studying full-time. There was also an association between residence type and study mode, with a higher proportion of students living in their own accommodation studying part-time.
- ☐ **Stipend Funding:** Respondents who received a stipend expressed a more positive sentiment than those who did not.

While many statistically significant relationships between the sentiment of the best student experience and a range of factors have been established, there are a number of areas where the sample size is not sufficient. It is recognised as a limitation in this study, perhaps trends in future years with a larger sample set and an improved sentiment classifier may prove to be a fruitful source of new insight. The use of NLP and statistical analysis of the resulting scores is likely to continue to reveal new insights in future years. These include areas such as the impact of supervisor panel size, support in teaching/demonstrating and working environments.

Suggested developments emerging as topics from the Q11 highlight the need for adequate funding so students can maintain a good quality of life and maximise research experiences. Better communication processes are required for seamless integration with university environments for example awareness of forms and systems to get paid, travel, and gain experience lecturing. Induction and personal development plans showed no significant relationship with positivity of responses to question 10. This suggests that communication needs to be ongoing, frequent, and timely. The importance of well-resourced work environment was highlighted not only as a physical space or facility to undertake research (lab and desk based) but also as an environment to foster collaborative work and create a support peer network. In its absence comments relating to loneliness and isolation were evident in the data. In Ireland, funding for postgraduate research has witnessed several trends. There has been a notable increase in funding opportunities, with various scholarships and schemes available like the Irish Research Council and initiatives like the Marie Skłodowska-Curie Actions offering support to students. However, the funding landscape remains highly competitive, with limited resources and rigorous selection processes (Government of Ireland., 2022). It is hoped this study will contribute to the national discourse by using NLP approaches to highlight research students' sentiments and opinions.

A summary of how to interact with the dashboard is provided in Appendix 1 and a step-by-step guide to implement NLP research is presented in Appendix 2.



References

- Ali, M. (2020). Topic Modelling in Power BI using PyCaret. Towards Data Science. Retrieved from <https://towardsdatascience.com/topic-modeling-in-power-bi-using-pycaret-54422b4e36d6>
- Anusha, A., & Singh, S. (2015). Is that twitter hashtag worth reading. ACM International Conference Proceeding Series, 10-13 (August), 272–277. <https://doi.org/10.1145/2791405.2791526>
- Bordoloi, M., & Biswas, S. K. (2023). Sentiment analysis: A survey on design framework, applications and future scopes. In Artificial Intelligence Review (Issue 0123456789). Springer Netherlands. <https://doi.org/10.1007/s10462-023-10442-2>
- Chen, X., Xie, H., Cheng, G., Poon, L. K. M., Leng, M., & Wang, F. L. (2020). Trends and features of the applications of natural language processing techniques for clinical trials text analysis. Applied Sciences (Switzerland), 10(6). <https://doi.org/10.3390/app10062157>
- Dalipi, F., Zdravkova, K., & Ahlgren, F. (2021). Sentiment Analysis of Students' Feedback in MOOCs: A Systematic Literature Review. Frontiers in Artificial Intelligence, 4(September), 1–13. <https://doi.org/10.3389/frai.2021.728708>
- Daniel, B. K., & Harland, T. (2018). Qualitative data analysis. Higher Education Research Methodology, 15(1), 98–110. <https://doi.org/10.4324/9781315149783-10>
- Effrosynidis, D., Symeonidis, S., & Arampatzis, A. (2017). A Comparison of Pre-processing Techniques for Twitter Sentiment Analysis. In J. Kamps, G. Tsakonas, Y. Manolopoulos, L. Iliadis, & I. Karydis (Eds.), Research and Advanced Technology for Digital Libraries. TPD. Lecture Notes in Computer Science, 10450. Cham: Springer.
- Eisenbeiss, S. A., Knippenberg, D. V., & Boerner, S. (2008). Transformational leadership and team innovation: Integrating team climate principles. Journal of Applied Psychology, 93(6), 1438-1446.
- Gottipati, S., Shankararaman, V., & Lin, J. (2018). Latent dirichlet allocation for textual student feedback analysis.
- Government of Ireland. (2022, August 2). Impact 2030: Ireland's New Research and Innovation Strategy. Retrieved from <https://www.gov.ie/en/publication/27c78-impact-2030-irelands-new-research-and-innovation-strategy/>
- Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vinod Vydiswaran, V. G. (2018). Augmenting qualitative text analysis with natural language processing: Methodological study. Journal of Medical Internet Research, 20(6). <https://doi.org/10.2196/JMIR.9702>
- Hazarika, D., Konwar, G., Deb, S., & Bora, D. (2020). Sentiment Analysis on Twitter by Using TextBlob for Natural Language Processing (pp. 63-67). doi: 10.15439/2020KM20.
- Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225. <https://doi.org/10.1609/icwsm.v8i1.14550>
- Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. (2021). Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study. Applied Sciences, 11(9). <https://doi.org/10.3390/app11093986>
- Lacey, S. (2021). REVEAL – PGR Roadmap to StudEntSurVey.ie Qualitative Data Analysis – PostGraduate. Munster Technological University.
- Lim, K. W., & Buntine, W. (2014). Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. CIKM 2014 - Proceedings of the 2014 ACM International Conference on Information and Knowledge Management, 1319–1328. <https://doi.org/10.1145/2661829.2662005>
- Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. International Conference on Information and Knowledge Management, Proceedings, 375–384. <https://doi.org/10.1145/1645953.1646003>
- Lovitts, B. E. (2005). Being a good course-taker is not enough: A theoretical perspective on the transition to independent research. Studies in Higher Education, 30(2), 137-154.



- Magana Vsevolodovna, R. (2021). Natural Language Processing with PyCaret and Power BI. Retrieved from <https://ruslanmv.com/blog/NLP-with-Pycaret-and-Power-BI>
- Malema, G., Ishmael, O., Okgetheng, B., & Rammidi, G. (2022). Setswana Parts of Speech Tagging : Indirect Relative. *International Journal of Computational Linguistics (IJCL)*, 13(4), 48–56.
- Microsoft. (2003). What Is Sentiment Analysis? Microsoft Dynamics 365. Retrieved from <https://dynamics.microsoft.com/en-ie/ai/customer-insights/what-is-sentiment-analysis/>
- PyCaret. (2020). Installation. Retrieved from <https://pycaret.readthedocs.io/en/stable/installation.html>
- Rakhmanov, O., & Schlippe, T. (2022). Sentiment Analysis for Hausa: Classifying Students' Comments. *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, June, 98–105. <https://aclanthology.org/2022.sigul-1.13>
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68-78.
- Share, M., Delaney, C., & McDaid, R. (2021). PGR StudentSurvey.ie Qualitative Data Analysis Report. StudentSurvey.ie. Dublin.
- Singh, G., Kumar, A., Singh, J., & Kaur, J. (2023). Data Visualization for Developing Effective Performance Dashboard with Power BI. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)* (pp. 968-973). Uttarakhand, India. doi: 10.1109/ICIDCA56705.2023.10100169.
- Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, 110, 298–310.
- Thomas, L., & Quinn, J. (2007). First-generation entrants to higher education: Family, background, and educational achievements. *Higher Education Quarterly*, 61(4), 411-429.
- Tierney, P. J. (2012). A Qualitative Analysis Framework Using Natural Language Processing and Graph Theory.



Appendices

Appendix 1

Objective 4: Visualise insights using interactive dashboard.

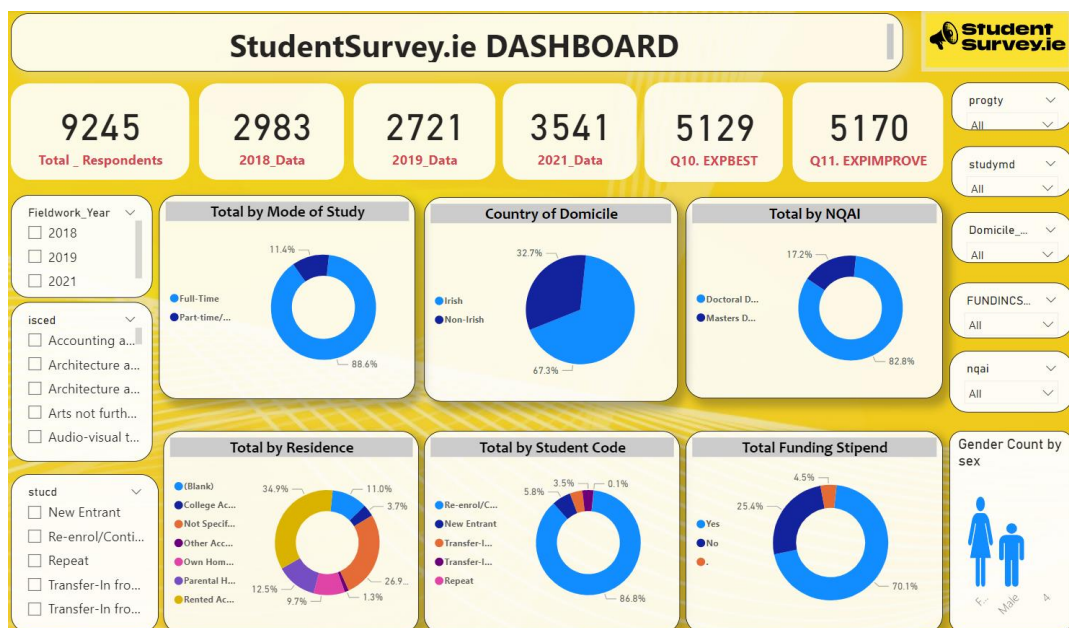
The interactive dashboard consists of a collection of 7 reports, each designed to provide users with a comprehensive understanding of sentiment and opinions related to Q10 (Q10: MOST VALUED) and Q11 (Q11: IMPROVEMENTS). The reports utilize various visualizations, filters, and slicers to enhance user exploration and analysis.



This makes it easier to identify trends and patterns, and to gain insights into data that might not be immediately apparent.

The dashboard offers four main report dashboards:

Demographics: This dashboard report focuses on demographic information, allowing users to examine sentiment and opinions based on different characteristics such as gender, education level, or any other relevant demographic factors. Visualizations in this tab help users identify patterns and trends specific to various demographic segments.

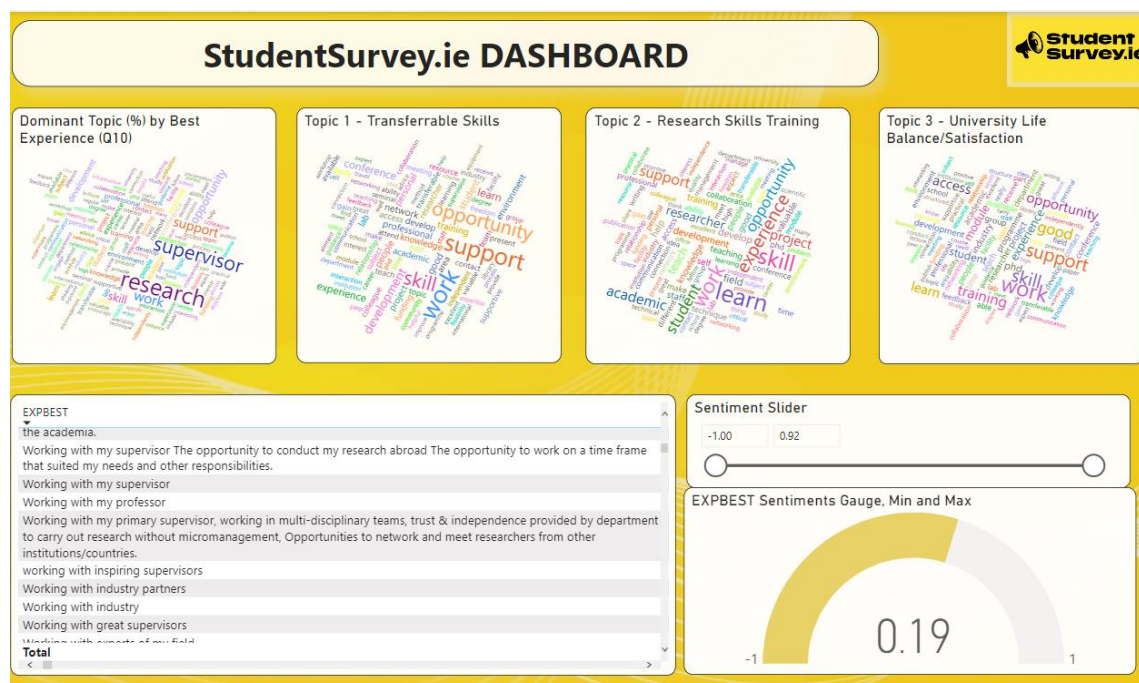


The dashboard provides statistics for different years. It presents demographic details like education levels (iscd), study codes (studycd), program types (progty), study modes (part-time/full-time), and domicile (Irish/non-Irish). The Funding dropdown menu displays the number of respondents funded with stipends or without. Lastly, the dashboard includes information about gender, including the number of females, males, and those with non-binary gender identification.

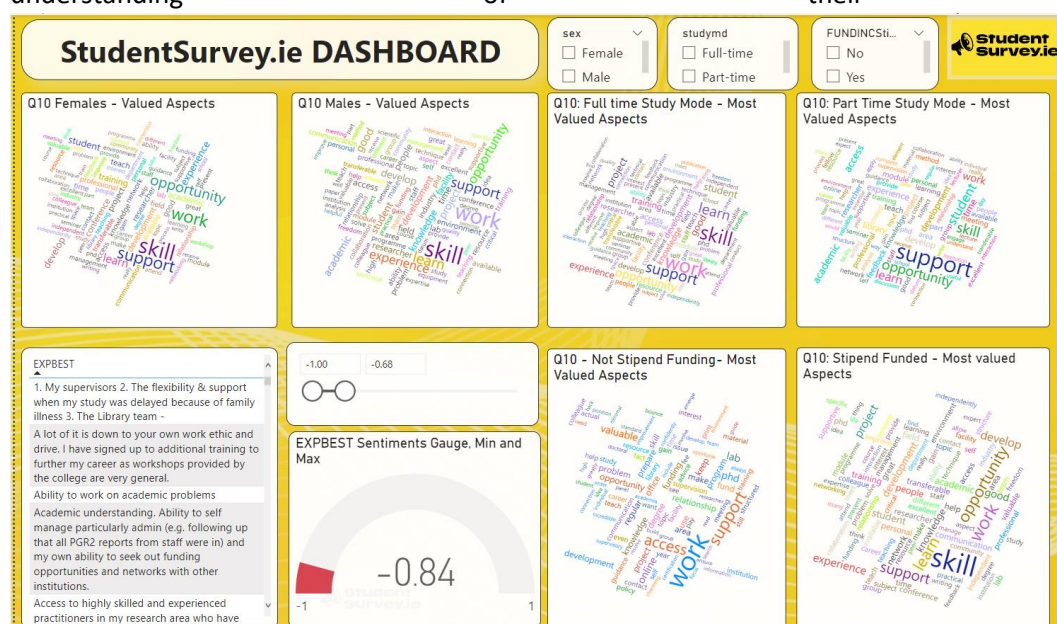


Q10 Topic

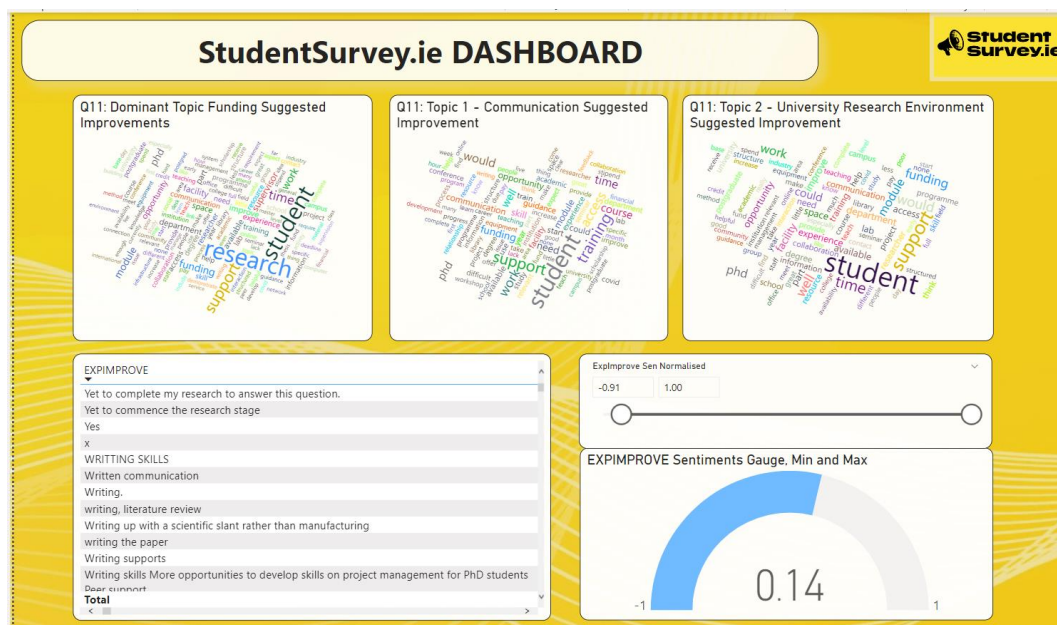
PyCaret was used to generate topics and analyse the data for Q10, while Power BI incorporated gauges and sliders to extract comments based on categories, ranging from negative to positive. This combination of tools enabled comprehensive analysis and visualization of the survey responses in an interactive and user-friendly manner.



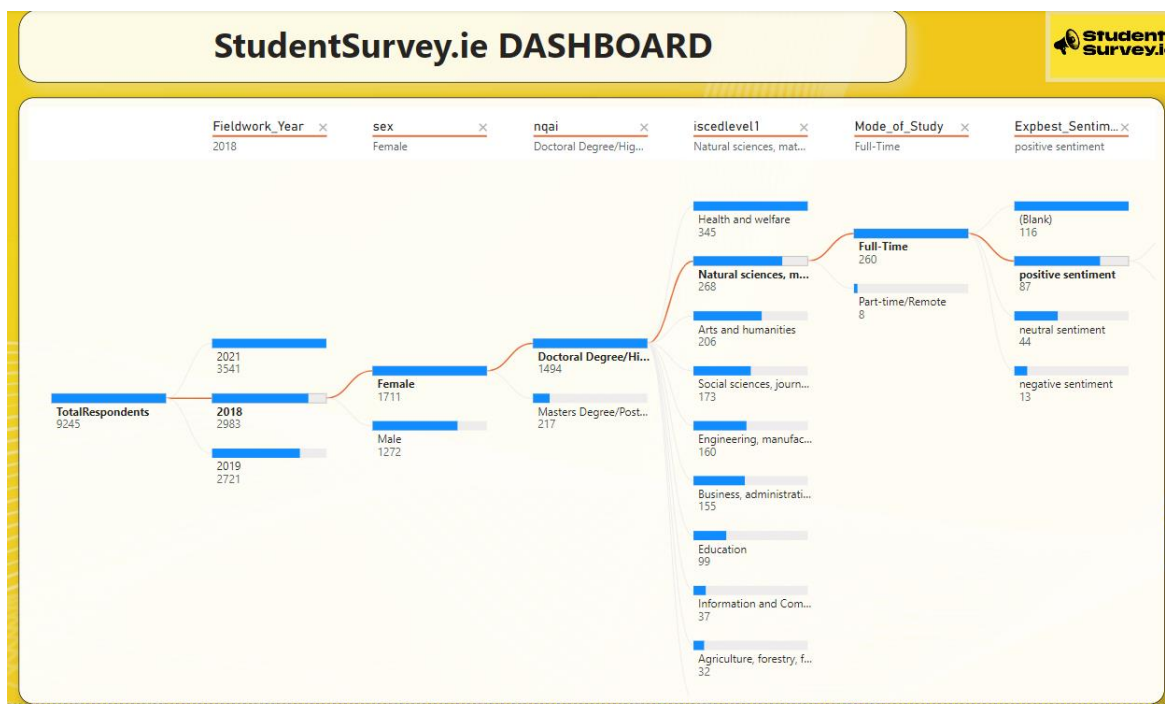
Q10 Statistical Factors of Significance: the dashboard report explores statistical analysis to identify the factors that hold significance for respondents regarding Q10. This analytical approach aids in the identification of key drivers or influencers that shape the sentiment expressed in response to Q10. By delving into the statistical analysis of these factors, users can uncover valuable insights and make informed decisions based on a deeper understanding of their significance.



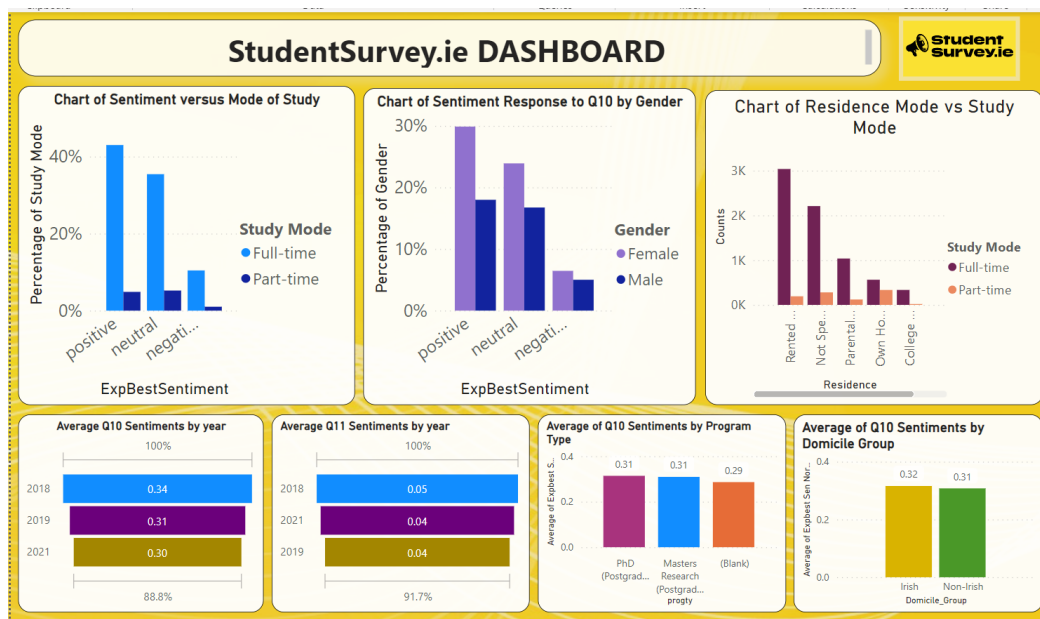
Q11 Topics and Sentiment Data: dashboard report explores Q11, examining the topics discussed and sentiment expressed by respondents. Users can explore visual representations that depict sentiment distributions, such as pie charts or stacked bar charts, to gain insights into the overall sentiment distribution and variations across different topics.



Decomposition Network: The decomposition network in the context of LDA and sentiment analysis can provide insights into how positive and negative sentiments are distributed among different respondent characteristics. It allows for the visualization of sentiment patterns based on specific attributes such as gender, NQAI level, or any other demographic or categorical variable available.



Interactive charts also offer an overview of the average sentiment by study mode, gender, accommodation.



This can be published so the data is accessible and available to a wider range of stakeholders, including researchers, policymakers, and the general public while adhering to privacy and security consideration of the data.

Appendix 2

Objective 5: How to Guide for Researchers

An Intro of Pycaret

PyCaret is an open-source, low-code machine learning library in Python that revolutionizes the machine learning workflow. With its end-to-end functionality and model management capabilities, PyCaret accelerates the cycle of machine learning experiments. Pycaret automate various tasks, such as data preprocessing, feature selection, model training, hyperparameter tuning, and model evaluation. Pycaret offering Natural Language Processing (NLP) capabilities, making it a comprehensive tool for tasks involving text analysis and language understanding. (PyCaret, 2020; Magana Vsevolodovna, 2021; Ali, 2020).

FEW STEPS TO INSTALL PYCARET

'Pycaret', with Power BI

create an anaconda environment and Install PyCaret

```
# create a conda environment
conda create --name yourenvname python=3.8

# activate conda environment
conda activate yourenvname

# install pycaret
pip install pycaret

# create notebook kernel connected with the conda environment
python -m ipykernel install --user --name yourenvname --display-name "display-name"
```

Example



Anaconda Prompt (Anaconda3)

```
(base) C:\Users\G00425457>conda create --name studentsurveypowerbi python=3.8
```

Install Language Model

Anaconda Prompt (Anaconda3)

```
(base) C:\Users\G00425457> conda activate studentsurveypowerbi
```

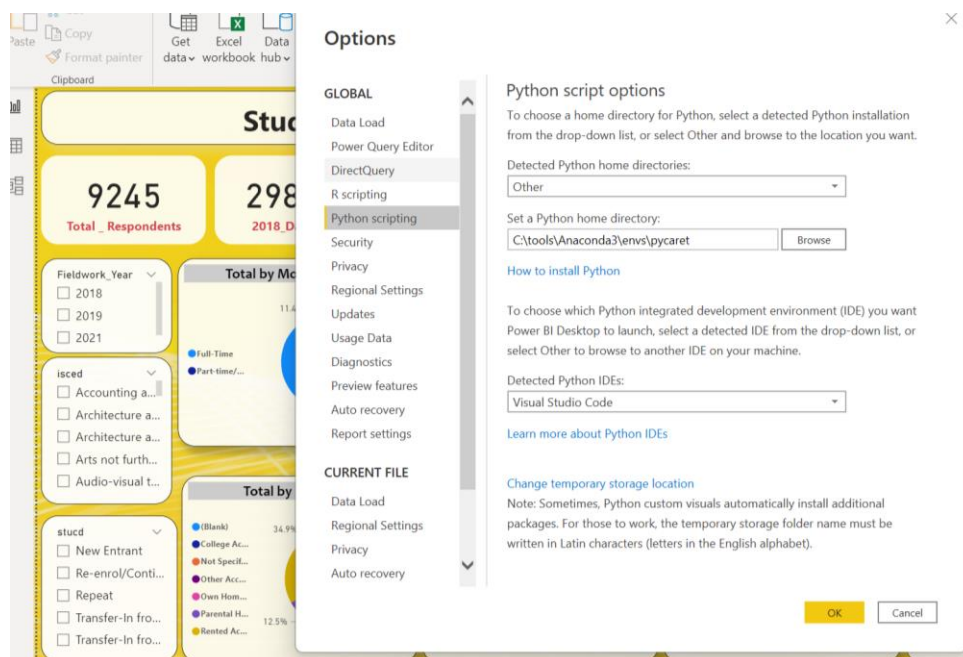
Download English Language Model:

Anaconda Prompt (Anaconda3)

```
(base) C:\Users\G00425457> python -m spacy download en_core_web_sm
```

```
(base) C:\Users\G00425457>python -m textblob.download_corpora
```

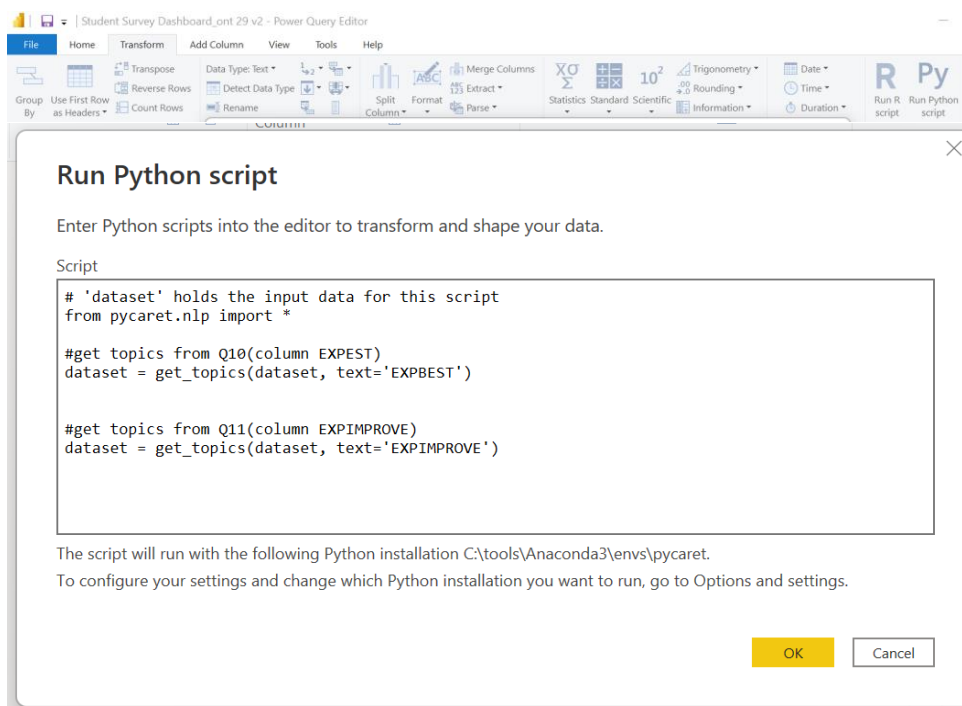
navigate to Options and Settings > Options > Python scripting in Power BI. In the 'Detected Python home directories' option, choose 'Other', and in the 'Set a Python home directory' dropdown menu, select the corresponding location (PyCaret, 2020; Magana Vsevolodovna, 2021; Ali, 2020).



PowerBi

In Power BI, locate '*PY run python script*'. The option to run Python scripts is found under the "Transform data" tab.





1. Loading required packages

The python packages below are required for our analysis. If any of the packages are missing or you run into an error, use the "!pip install package-name" command in a cell to install the missing package as shown in the error message.

```
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
import nltk
from nltk.probability import FreqDist
from matplotlib.patches import FancyArrowPatch
import seaborn as sns
import plotly.graph_objects as go
from collections import defaultdict
from mpl_toolkits.mplot3d import Axes3D
import squarify
from nltk.sentiment import SentimentIntensityAnalyzer
import plotly.graph_objs as go
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer
import string
from textblob import TextBlob
import string
```

```
#download vader lexicon
nltk.download('vader_lexicon')
```

###2. Data Loading



Data that contains text for sentiment analysis is loaded into the notebook using the code sections below. The previews of the head(top 5 rows), tail(bottom five rows) and the sample(random 10 rows) show a quick overview of the data. Because Data is in excel format, we specify the name of the excel sheet our data is on within the excel file and we specify the columns we want loaded with the sentiments using the "usecols" attribute.

```
# load the text data into a pandas DataFrame
```

```
df = pd.read_excel('PGR StudentSurvey.ie data 2018, 2019, 2021.xlsx',  
sheet_name='PGR_StudentSurveyie_data', usecols=['EXPBEST','EXPIMPROVE','Fieldwork_Year',  
'sex','Domicile_Group','nqai','progty'])
```

3. Data Cleaning

Here we clean off rows of data with missing values as this may introduce errors while generating sentiments and do not really add value to our analysis. This is done using the dropna() function which loops through the entire data and removes rows with missing values which are denoted as NaN

```
# Remove the rows with NaN values  
df = df.dropna()
```

4. Generating Sentiment Score For Select Columns

Here is the fun part, now that our data has been loaded and cleaned, we proceed with generating sentiment scores for every row. Three sentiment analysis methods were explored. In the code sections below we show the Nltk (Vader) sentiment and the textblob sentiment. TextBlob is a Python library that can be used to process textual data. It is good for sentiment analysis, tokenization and sentence correction.

In the code sections below, we defined a function get_sentiment that takes an input referred in the code sections as "text", runs the specified sentiment analyser on it and returns sentiment as a score. We have done this first for textblob and secondly for nltk(vader) for two columns EXPBEST and EXPIMPROVE representing Q10 and Q11 respectively and we have saved the sentiment scores generated in new filenames referring to the sentiment analyser used.

```
#Generating sentiment scores using TextBlob  
#define function first  
def get_sentiment(text):  
    sentiment = TextBlob(text).sentiment.polarity  
    return sentiment  
  
# apply function to both columns and store score in new column names  
df['EXPBEST_TextBlob_Sentiment'] = df['EXPBEST'].apply(get_sentiment)  
#do the same for EXPIMPROVE  
df['EXPIMPROVE_TextBlob_Sentiment'] = df['EXPIMPROVE'].apply(get_sentiment)  
  
# preview the dataframe to see  
df.sample(5)
```



```
#NLTK Sentiment Score generation
#store the loaded SentimentIntensityAnalyser as sia
sia = SentimentIntensityAnalyzer()

#define function for nltk sia
def get_nltk(text):
    sentiment = sia.polarity_scores(text)['compound']
    return sentiment

# apply the function on all
df['EXPBEST_Sentiment_nltk'] = df['EXPBEST'].apply(get_nltk)
#do the same for EXPIMPROVE
df['EXPIMPROVE_Sentiment_nltk'] = df['EXPIMPROVE'].apply(get_nltk)

# preview the data
df.sample(5)
```

5. Storing and exporting table with generated sentiment scores

You can export the generated sentiments easily using the `to_csv` command and giving it a file name as demonstrated in the code sections below. Both Vader and Textblob generate sentiment scores between -1 and 1, where -1 would be at the extreme end of a negative comment, 1 would be a very positive comment and 0 would be on the neutral side. These open source analysers are not perfect and may need some human input to test its accuracy. This is why we took a smaller sample of the original file and labelled it ourselves to compare the best ones for our survey analysis.

```
df.to_csv('Table with generated sentiment scores.csv')
```



Appendix 3

Table of statistical significance values for Mood's Median test of Sentiment Score assigned to Q10 and other questions in the survey

Question	Code	p value ¹	Significance ²
¹ Moods median test of response to given question and sentiment of Q10			
² (* = significant, ns = not significant, correction for multiple comparisons has been made)			
		p	
Receiving advice on career options	DOCAREER	0.014	*
Do you agree or disagree that the teaching / demonstration you delivered enhanced your overall research experience?	DOENHANCE	0.973	N s N
Taking part in a placement or internship	DOINTERN	0.832	s
Agreeing a personal training or development plan	DOPERDEV	0.946	ns
Receiving training to develop my research skills	DORESSKILL	0.003	*
Do you agree or disagree that you have been given appropriate support and guidance for your teaching / demonstration?	DOSUPPORT	0.03	ns
Receiving training to develop my other transferable skills	DOTRANSK	0.019	*
I understand the requirements and deadlines for formal monitoring of my progress	PADEAD	0.18	ns
I received an appropriate induction / orientation to my research degree programme	PAIND	0.653	ns
The final assessment procedures for my research degree are clear to me	PAPROC	0.003	*
I understand the required standard for my thesis	PASTD	0.008	*
There is someone in my institution I can talk to about my day-to-day problems (Do you agree or disagree with the following statements about your personal outlook?)	PODTPROBL	0.03	*
I am satisfied with my life within my institution nowadays (Do you agree or disagree with the following statements about your personal outlook?)	POLIFEINSTNOW	0.002	*
I am satisfied with my life nowadays (Do you agree or disagree with the following statements about your personal outlook?)	POLIFENOW	0.009	*
I feel that my research degree programme is worthwhile (Do you agree or disagree with the following statements about your personal outlook?)	PORESDEGWORTH	0.024	*
I am satisfied with my work-life balance (Do you agree or disagree with the following statements about your personal outlook?)	POWLBALANCE	0.67	ns
The research ambience in my department stimulates my work	RCAMB	0.194	ns
I have frequent opportunities to discuss my research with other research students	RCOPP1	0.369	ns
I have opportunities to become involved in the wider research community, beyond my department	RCOPP2	0.037	ns
My department provides access to a relevant seminar programme	RCSEM	0.056	ns



There is adequate provision of computing resources / facilities	RICOMFAC	0.096	ns
There is adequate provision of library facilities (including physical / online resources)	RILIBFAC	0.097	ns
I have access to the specialist resources and facilities necessary for my research	RISPECFAC	0.102	ns
I have a suitable working space	RIWSPC	0.805	ns
My skills in critically analysing and evaluating findings and results have developed during my programme	RSANALYSE	0.001	*
Other than my supervisor(s), I know who to approach if I am concerned about any academic aspect of my research degree pr	RSCONTACT	0.094	ns
My institution values and responds to feedback from research degree students	RSFEEDBACK	0.673	ns
My confidence to be creative or innovative has developed during my programme	RSINNOV	0	*
My understanding of research integrity (e.g. rigour, ethics, transparency, attributing the contribution of others) has	RSINTEGRITY	0.012	*
I understand my responsibilities as a research degree student	RSSTU	0.002	*
I am aware of my supervisor	RSSUP	0.042	ns
How aware are you of the various student supports available? (Recreation, healthcare, counselling, etc)	RSSUPPORTS	0.013	*
My skills in applying appropriate research methodologies, tools and techniques have developed during my programme	RSTOOLS	0.109	ns
I have regular contact with my supervisor(s), appropriate for my needs	SCONT	0.014	*
My supervisor(s) provides feedback that helps me to direct my research activities	SFEED	0.003	*
My supervisor(s) provides the appropriate level of support for my research	SSUPP	0.013	*
My supervisor(s) help me to identify my training and development needs as a researcher	STDEV	0.071	ns
My ability to communicate information effectively to diverse audiences has developed during my programme	TSKCOMM	0.002	*
I have developed contacts or professional networks during my programme	TSKCONTACT	0.694	ns



