Assessing university-wide student experience initiatives: an application of natural language processing, association tests and dashboard visualisation

Data Analytics Major Project and Placement MOD007894

Assessment Element 011 – Dissertation

SID 2057296

17 August 2022

List of figures

List of tables

Declaration

I, Emma Schubart, declare that the work in this dissertation titled “Assessing university-wide student experience initiatives: an application of natural language processing, association tests and dashboard visualisation” is carried out by me. This work has not been submitted to Anglia Ruskin University or any other educational institution for the award of a degree or educational qualification. I also declare that the information published in this dissertation has been obtained and presented in accordance with academic rules and ethical conduct. Any information obtained from other sources has been properly referenced.

Acknowledgement

I would like to thank my Major Project supervisor, Dr. Silvia Cirstea for her instruction, insight and constant support throughout the duration of my Master's course. I am also grateful to my module leaders, Dr. Mahdi Maktabdar, Dr. Mahmud Hasan and Dr. Faraz Janan for their instruction and guidance.

Table of Contents

Abstract

The Student Experience Team within Student & Library Services at Anglia Ruskin University (ARU) delivers a range of student experience initiatives. One initiative is the Welcome Buddy scheme, which is a volunteering scheme connecting current students, who act as "Welcome Buddies" to new students to help them transition into ARU. The Student Experience Team seeks to inform and improve the administration of their Welcome Buddy scheme via data-driven solutions. However, robust analysis of the unstructured data they have collected is beyond their capacity. This paper addresses the significant data analytics needs of the Student Experience Team by applying statistical, machine learning, and natural language processing techniques to demographic, survey (including both open-ended and closed-ended questions) and chat space data collected by the Student Experience Team on its Welcome Buddy scheme. More specifically, this paper assesses associations between Welcome Buddy performance scores and relevant demographic characteristics, analyses chat data from the "Ask ARU" chat spaces of Welcome Buddies and their assigned new student cohort for formality, warmth and helpfulness metrics, conducts sentiment analysis (both supervised and unsupervised) on open-ended exit survey questions from new students, and produces a Power BI dashboard for the ARU Student Experience Team to visualise the impact and future direction of the Welcome Buddy scheme. Results indicate a high level of student satisfaction with the Welcome Buddy scheme. Results also indicate that high Welcome Buddy performance is significantly associated with Welcome Buddies who are female, older, and engage in informal, highly helpful and warm communication within their chat spaces. These findings demonstrate the utility of supervised and unsupervised sentiment analysis, descriptive statistics, and regression models in the evaluation of student ambassador schemes at universities.

Introduction

As public-facing teams and departments undertake to improve operations via data-driven solutions, they are faced with the challenge of retrieving actionable insights from unstructured categorical data, often in the form of survey and chat space data. For this reason, robust analyses of such data are often underutilised. While various natural language processing techniques have been applied to text data across industries, from patient experience survey data in hospitals [1,2], to automobile insurance fraud detection [3], to human resource development [4], to market structure surveillance of sedan cars and diabetes drugs [5] to student evaluations in higher education settings [6,7,8,9], there have not been applications of categorical data analysis within the context of student experience within university-wide ambassadorial schemes nor student ambassador performance within such schemes. This paper applies supervised and unsupervised sentiment analysis and association testing to a student ambassador scheme at Anglia Ruskin University (ARU).

The Student Experience Team within Student & Library Services at ARU delivers a range of student experience initiatives, focused on improving student engagement, retention, and success across ARU. One initiative is the Welcome Buddy scheme, which is a volunteer, ambassadorial scheme connecting current students, who act as "Welcome Buddies", to new students to help them transition into ARU and their course. The Student Experience Team recruits and trains approximately 600 current students each year who act as a friendly face to every new level 4 UG student (and level 3 ARUC students) on their course for the duration of the new student cohort's first trimester. The results of this paper address the need of the

Student Experience Team to gain actionable insights from multiple data sources which pertain to the Welcome Buddy scheme. In so doing, this paper will lead and promote data-driven decision making within the administration of the Welcome Buddy scheme across ARU campuses. Sentiment, survey data and demographic analyses indicate a high level of student satisfaction among new students with the Welcome Buddy scheme. Results also indicate that high Welcome Buddy performance is significantly associated with Welcome Buddies who are female, older, and engage in informal, highly helpful and warm communication within their chat spaces. These findings demonstrate the utility of supervised and unsupervised sentiment analysis, descriptive statistics, and regression models in the evaluation of student ambassador schemes at universities.

Literature Review

Most of the analyses within this paper involved the interrogation of categorical data. Powers and Xie define categorical variables as variables which "can be measured using only a limited number of values or categories" as opposed to continuous variables, which can theoretically take on an infinite number set of values.[10] This paper analyses survey data, including both closed-ended and open-ended questions, chat space data and demographic data. In order to prepare these types of non-numeric data for analysis, it must be transformed into a numeric representation. This literature review focuses on two different types of transformations: encoding and embedding.

Encoding

Encoding transforms categorical data into vectors, at which point the categorical data can be analysed quantitatively. There are a number of methods to encode categorical data. One-hot encoding is a well-known encoding technique. Encoding under this technique depends upon the number of levels within the categorical variable of interest. For example, if a discrete categorical variable x has n distinct values, then the One-hot encoding of this variable yields a vector where every component is a zero expect for the ith component, which has the value 1. The encoded binary features are considered dummy variables. A clear disadvantage of this technique lies in the instance of high cardinality, which can create storage problems. However, this can be overcome with sparse vector and matrix representations, which Python's Scikit-learn One-hot encoder does by default. Another issue is the low information yield from this technique, since the Euclidean distance between these features (which are represented as zeros and ones) can tell us only if the features are the same or different, and we cannot ascertain how similar or how different they are.

Another encoding technique is label encoding, in which the encoded equivalents of the categorical data "are determined by how we arbitrarily choose to assign integer values to them."[11] The disadvantages of this technique include the potential to reduce the accuracy of algorithms, since arbitrarily encoded data can prevent algorithms from accurately assessing the differences between variables. Although never explicitly addressed, this drawback of label encoding was perhaps inadvertently illustrated by Potdar et. al, where label encoding is likely the explanation for the poor classification accuracy of artificial neural network models applied to categorical data.[12] However, this method can be well-suited for ordinal categorical data, in which case the encoded integer labels can correspond to the ordinality of the data without increasing dimensionality.

Target encoding is another option for encoding categorical data. This method encodes the categorical "target" variable with the mean values of that target variable. According to Python's scikit-learn documentation, the categorical target features "are replaced with a blend of posterior probability of the target given particular categorical value and the prior probability of the target over all the training data".[13] This method of encoding can be applied to any number of feature values without increasing the dimensionality. The disadvantage of this method lies in the potential for data leakage, leading to overfitting, since the categorical features are replaced with their means, and this can generate a strong correlation between these two features. This method is not advisable for use with categorical data in which there are strong interaction effects.

Embedding

Slightly different from encoding techniques are word embedding techniques, which semantically represent text data as vectors. Even though both types of data transformation techniques involve the numeric representation of non-numeric data, unlike encoded vectors, word embeddings are derived from the semantic relationships within data and the embedded values are neither meaningless nor arbitrary.

Related embedding techniques include the well-known contributions by Mikolov et al. across two papers. Mikolov et al. developed techniques for using categorical data as input for deep learning algorithms. Their first contributions were two techniques for word embeddings, or real-valued vector representation of words: skip-gram and continuous bag of words (CBOW).[14] The skip-gram algorithm learns word embeddings with a loss function that measures how well skip-gram predicts the words that surround the target word. Under the skip-gram architecture, the closer a surrounding word is to the target word, the heavier it is weighed. The CBOW algorithm learns word embeddings with a loss function that measures how well CBOW predicts the target word, given the words that surround the target word. Under the CBOW architecture, the order of surrounding words does not influence prediction. In a subsequent paper, Mikolov et al. introduced their word2vec algorithm.[15] Word2vec uses an unsupervised learning process whereby unlabeled data is trained via artificial neural networks to produce a distributed representation of words using either the CBOW or continuous skip-gram architecture. Mikolov et al. found that CBOW is faster while skip-gram does a better job for infrequent words. An advantage of word2vec is that unlike other embedding and encoding methods, the vector size is not necessarily equal to the number of unique words in the corpus. Instead, the size of the vector can be manually specified according to the corpus size and project type. This flexibility is particularly beneficial for large data.

Another popular word embedding technique is the TF-IDF vectorizer. The TF-IDF value is a statistic which calculates the "weight" or importance of each word within a document with respect to the words within the document(s) as a whole. The TF-IDF vectorizer is the product of two statistics, the term frequency and the inverse document frequency. A word receives a high TF-IDF value when it appears with high frequency in a single document and low frequency across the corpus. A vector consisting of elements equal to the number of unique words in all documents is created for each document. The TF-IDF vectorizer is an unsupervised learning technique. The TF-IDF vectorizer is appealing because it allows for the identification of words which are highly discriminative for documents within the corpus. However, the TF-IDF vectorizer is not effective as a dimensionality reduction technique nor is it informative regarding inter- or intra-document statistical structure.[16]

One response to the shortcomings of the TF-IDF vectorizer was the development of latent semantic indexing (LSI).[17] LSI uses singular value decomposition to identify statistical co-occurrences of words that appear together across a corpus of documents. These co-occurrences can be highly informative with regard to the topics within the words and documents. This approach can achieve significant dimensionality reductions in large corpa. Deerwester et al. posit that the derived features of LSI, linear combinations of the original TF-IDF features, can capture synonymy and polysemy between words. Since LSI uses the bag-of-words architecture, it does not take the order of words, nor the order of documents, into account. In other words, this model assumes the exchangeability of the words and documents.

Blei et al. propose Latent Dirichlet Allocation (LDA) in response to the exchangeability assumption upon which LSA is predicated. LDA was first developed by Pritchard et al. in the domain of population genetics [18] and was applied to machine learning by Blei et al.[16] Blei et al. refer to the de Finetti theorem which states that any collection of exchangeable random variables has a probability distribution that is a "mixture" distribution of independent and identically distributed sequences of random variables. "Mixture" here means a weighted average.[19] Blei et al. posit that in order to properly consider exchangeable representations for documents and words, mixture models that capture the exchangeability of both words and documents need to be applied. Thus, Blei et al. propose the LDA model. Their iteration of LDA is a generative statistical model that explains observations through unobserved groups, with each group explaining why some parts of the data are similar. In the case of identifying topics occurring within a set of documents, or topic modelling, LDA “represent[s] a collection of documents as a distribution of topics where we infer the topic distribution using the distribution of words in documents”.[11] In short, LDA seeks to project the features in higher dimensional space onto a lower dimensional space in order to lower dimensional costs.

Bengio et al. also propose an embedding technique focused on avoiding the “curse of dimensionality” which refers to the exponential increase in computational efforts required for data analysis as dimensionality increases. Bengio et al. propose a distributed representation learning technique for mapping words that appear in text documents to a collection of vectors, which they refer to as a distributed representation for words.[20] Their model associates a distributed word feature vector (a real-valued vector in $\mathbb{R}^m$) with each word in the vocabulary, expresses the joint probability function of word sequences in terms of the feature vectors of these words in the sequence and simultaneously learns the word feature vectors and the parameters of that probability function. This model is especially notable for its demonstration of how neural networks can be used to implement a function that computes the probability of a sequence of words.

FastText, another word embedding technique, is a library developed by Facebook’s AI Research lab for efficient learning of word representations and text classification.[21] FastText supports supervised (classifications) and unsupervised (embedding) representations of words and sentences utilising the CBOW or skip-gram architectures. FastText has the same objective as word2vec (learning vector representations of words), but the key difference is that word2vec trains with words to learn word embeddings and fastText trains with character n-grams, where words are represented by the sum of the character n-gram. A disadvantage of fastText is that its granular training unit (n-grams as opposed to words) increases the size and processing time of the model. However a clear advantage of this technique

is that it gives the model the ability to predict different variations of words that are not directly in its own vocabulary.

Many of these embedding techniques support transfer learning. In transfer learning, a machine applies knowledge it gained from being trained on one one task to a different but related task. Bengio et al. define transfer learning as “the ability of a learning algorithm to exploit commonalities between different learning tasks in order to share statistical strength, and transfer knowledge across tasks”.[22] For example, Kratzwald et al. propose the sent2affect algorithm, which uses transfer learning to classify the emotional content of various samples of texts from different datasets.

Another transfer learning model is Bidirectional Encoder Representations from Transformers (BERT), which is a transformer-based machine learning technique for natural language processing pre-training.[23] Unlike word2vec, BERT is context-specific. This means that while word2vec generates a single word embedding representation for a given word in the vocabulary regardless of different contexts of each occurrence of the word, BERT takes into account the context for each occurrence of a given word. This extremely powerful model processed almost every single English-based Google Search query in October 2020.[24]

Sanh et al. propose DistilBERT, which is a small, fast, cheap and light Transformer model based on the BERT architecture.[25] Sanh et al. demonstrate that knowledge distillation performed during the pre-training phase of the DistilBERT model reduces the size of a BERT model by 40%, retains 97% of its language understanding capabilities and is 60% faster than a BERT model.

This literature review identified encoding and word embedding techniques that are relevant to the objectives of analysing relatively small amounts of data ($n < 1000$) in the form of surveys, chat spaces and demographic variables. The existing literature indicates that there are a number of techniques available of varying degrees of versatility to address some of the challenges of categorical data analysis.

Methods

The section will be organised into three subsections which cover the three data analysis tasks of this project:

1. student participant exit survey data analysis
2. chat space data analysis
3. demographic data analysis

Each subsection will include an exhaustive explanation of the research framework, data collection and analysis method. All data was collected and analysed in accordance with The General Data Protection Requirement (2016) and Data Protection Act (2018).

student participant exit survey

After the culmination of the Welcome Buddy scheme at the end of the new student cohort’s first trimester, the Student Experience Team issues an exit survey to all students to assess the performance of the scheme. Students are incentivized to complete the survey with the chance to win an Amazon voucher.

This project analysed survey responses from the September 2021 trimester and January 2022 trimester issues of the identical exit survey. Once combined, there were 127 survey responses in total.
First the survey data was split into open-ended and closed-ended survey questions. The high-level architecture of the survey data analysis is illustrated in Figure 1. The table below contains each original survey question, its associated responses, categorical variable type, feature extraction technique, and the abbreviation for each survey question/variable. From here on, each variable will be referred to by its abbreviation for ease of reference.

survey question	**variable abbreviation**	**possible responses**	**variable type**	**feature extraction technique**
campus selection	campus	☐ Cambridge ☐ Chelmsford ☐ Peterborough	nominal	One-hot encoding
How did you find out about the Welcome Buddy scheme?	how_became_aware	☐ email from the Welcome Buddy team ☐ email notification from your Welcome Buddy ☐ email when you accepted your place at ARU ☐ ARU webpages ☐ ARU Facebook group ☐ Open Day ☐ I didn’t receive any information ☐ Other	nominal	One-hot encoding
When did your Welcome Buddy first contact you?	first_contact	☐ during Welcome Week ☐ before I arrived ☐ once teaching had started ☐ They still haven't contacted me	nominal	This variable was not encoded.
Please select the most common way that you used to	access_method	☐ links in the auto emails ☐ the ARU app ☐ Going directly to	nominal	This variable did not undergo cross tabulation and was not encoded.

access your Ask ARU chat space with your Welcome Buddy.		the Ask ARU website in a web browser ☐ I didn't use my chat space		
Did you meet up with your Welcome Buddy (in person or online) in your first few weeks at ARU?	met_up	☐ No - I didn't feel I needed this support ☐ No - the Welcome Buddy didn't offer ☐ No - we couldn't figure out a convenient time for us to meet ☐ Yes - we met up in an online video call ☐ Yes - we met up in person ☐ Other	nominal	binary encoding (grouped by “yes” and “no” responses, “other” responses were omitted from cross tabulation)
In what ways did your Welcome Buddy help you to settle in to life as a student at ARU?	how_helped	open-ended	nominal	word2vec word embedding (for unsupervised sentiment analysis); binary encoding (for supervised sentiment analysis)
One of the reasons for running the buddy scheme through Ask ARU, was so that you could meet other new students as well as your Welcome Buddy. How helpful did you find this?	meet_new	☐ I loved that I could chat to other new students ☐ I would rather have been able to talk to my buddy one to one only ☐ I don't feel strongly either way	ordinal	label encoding
Are you friends with any of the new students	friends	☐ No - none of them ☐ Yes - a few of	nominal	binary encoding (grouped by “yes” and “no” responses)

that are in the same group chat as you?		them ☐ Yes - all of them ☐ Yes - just one of them		
Are there any ways that we could improve the Welcome Buddy scheme?	suggestions	open-ended	nominal	This variable was not transformed, it underwent manual topic extraction.
Are there any other comments you would like to make about the Welcome Buddy scheme?	comments	open-ended	nominal	This variable was not transformed, it underwent manual topic extraction.

Table 1: Survey questions, variable name, response options, variable type and feature extraction technique

closed-ended questions

The analysis objective for the closed-ended questions was to describe variable distributions and interrogate relationships between variables. There were 127 survey responses.

The nominal variables underwent One-hot encoding. One-hot encoding was chosen for the nominal variables because of its ease of use, efficient running time and its conductivity to cross tabulation and independence testing. The ordinal variable (meet_new) underwent label encoding. Label encoding was chosen for its ability to efficiently capture the ordinality of this variable. The variable meet_new is perhaps not overtly ordinal, but there is an implied scale across the three possible responses. “I loved that I could chat to other new students” is the strongest positive statement regarding the helpfulness of the social element of the Welcome Buddy scheme, “I would rather have been able to talk to my buddy one to one only” reflects a rejection of the helpfulness of the social element, and “I don't feel strongly either way” acts as the middle ground. As such, the meet_new variable was label encoded on a 1-3 scale with 3 representing the strongest positive response, “I loved that I could chat to other new students”.

For the purpose of independence testing, the variable met_up was both encoded as a binary variable in order to capture more concise information within this variable. In order to reduce the complexity of this variable, which contained 7 possible responses, (“No - I didn't feel I needed this support” , “No - the Welcome Buddy didn't offer”, “No - we couldn't figure out a convenient time for us to meet”, “Yes - we met up in an online video call”, “Yes - we met up in person”, “Other”), I only took into account “Yes” and “No” responses and encoded these as 1 and 0 respectively.

After preprocessing of the closed-ended questions, the survey responses were analysed with descriptive statistics. This was accomplished by cross tabulating variables to compare and analyse responses against each other. Since the variables are categorical, the chi-squared test for independence was applied to determine the relationship between variables and subgroups of variables.

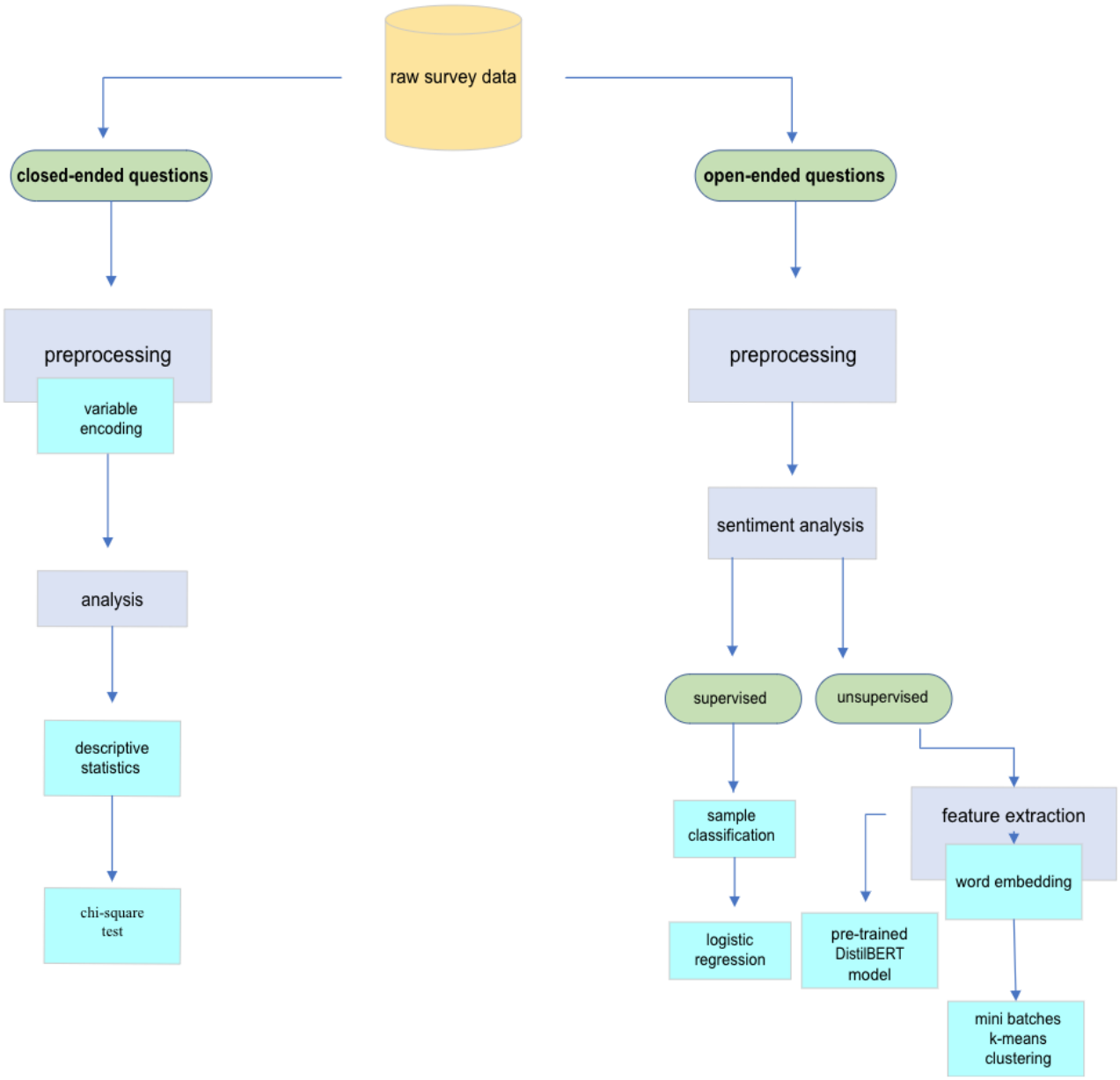

Figure 1: high-level architecture of survey data analysis

open-ended questions

The analysis objective for the open-ended questions was sentiment analysis and topic extraction. The "how_helped" variable underwent both supervised and unsupervised sentiment analysis. This dissertation employs both lexicon-based and machine learning-based sentiment analysis.

Preprocessing the free-text consisted of vocabulary reduction. Vocabulary reduction involved lowercasing the text, removing stop words and other words and phrases which could artificially skew the sentiment analysis and tokenization. For example, all instances of "Welcome Buddy" and "buddy" were removed from the text, as these words could be algorithmically misinterpreted. However, "welcome" was not removed from the text, since a response including the word "welcome" is meaningful for sentiment

analysis. Negations were also handled carefully during vocabulary reduction. After expanding all contractions (“don’t” to “donot”, “won’t” to “willnot”, etc.), all nouns and adjectives were joined to the negation (if present) immediately preceding them. For example, after vocabulary reduction, “ wasn’t helpful” was retained as the text token “wasnot_helpful.” Vocabulary reduction was achieved by utilising the Python library, Natural Language Toolkit (NLTK) and regular expressions because both tools facilitate efficient and comprehensive natural language processing.

After preprocessing, three different sentiment analyses (two unsupervised and one supervised) were conducted on the text within the how_helped variable. The first sentiment analysis was an unsupervised, pre-trained DistilBERT model. DistilBERT was selected for its ease of use and computational speed. The second unsupervised sentiment analysis utilised the word2vec embedding model from the Python Gensim library. Since the vocabulary of the how_helped variable was relatively small after preprocessing, (687 unique text tokens, 130 responses) the parameters of the word2vec model were set accordingly:

sentences = responses
workers = 1 (number of threads to use while training)
size = 350 (number of dimensions of the embedding; 350 was chosen after trial and error as this size maximised the intuitive accuracy of the embeddings which was examined by extracting the words most similar to a given word, according to the word2vec model)
min_count = 1 (minimum word count occurrence of words to consider when training, words with less than this occurrence count will be ignored by the model)
window = 4 (maximum distance between a target word and words around the target word)
sg = 1 (specifies training algorithm; CBOW = 0, skip gram = 1)
sample = $1x10^{-3}$ (controls how much subsampling occurs)

After generating the word2vec vectors for each response, principal component analysis was applied to the vectors to reduce the high-dimensional word vectors to two-dimensional scatter plots in order to visualise the word embeddings. Then the vectors were clustered with a Mini-batches K-means clustering algorithm. This variant of the K-means algorithm uses random samples of the input data to reduce the time required during training. The Mini-batches K-means algorithm was evaluated with the Silhouette Coefficient. The optimal number of clusters was determined to be two.

The supervised sentiment analysis required the manual creation of a new binary variable, “sentiment” and the manual assignment of positive (1) or negative (0) sentiment values to each individual response within the “how_helped” variable. I also conducted supervised sentiment analysis of the responses within the “how_helped” variable. The sentiment scores were assigned according to the rubric below.

sentiment	score	characteristics
positive	1	conveyed demonstrably favourable feedback about the helpfulness of the Welcome Buddy scheme
negative	0	conveyed dissatisfaction with the helpfulness of the Welcome Buddy scheme
neutral	omitted	did not express favourable or unfavourable feedback about the helpfulness of the Welcome Buddy scheme

Table 2: supervised sentiment classification rubric for responses within "how_helped" variable

During this sample classification process, any neutral responses were removed, thereby reducing noise within the data ahead of the supervised sentiment analysis. Then during the preprocessing stage, the text was cleaned with the same vocabulary reduction and tokenization function applied during unsupervised sentiment analysis. After cleaning, there were 116 responses and 660 unique text tokens. The cleaned text data was saved as a new variable, "new_how_helped". Then all responses within the "new_how_helped" variable were embedded with the TF-IDF vectorizer. Its computational speed, ease of use, and the fact that the TF-IDF vectorizer weights word tokens according to relevance made it the appropriate embedding tool for this size dataset. After setting the predictor variable ("new_how_helped") and outcome variable ("sentiment"), the data was split into test and train sets, where test_size = 0.3. The manually assigned sentiments of the responses were classified with a logistic regression. The regression model was evaluated with precision and recall metrics, F-score and the area under a receiver operating characteristic (ROC) curve.

The "suggestions" and "comments" variables also consisted of free-text, open-ended data. These variables contain student input about the Welcome Buddy scheme. In order to extract information from these variables, I manually went through the responses within both variables to extract the most recurrent topics, as well as topics that were less redundant but which I deemed to be valuable points for consideration by the Student Experience Team.

chat space data analysis

The chat space data was retrieved from ARU TopDesk. The chat spaces consisted of the chat communications from the "Ask ARU" chat spaces between Welcome Buddies and their assigned group of new students. A function was created to extract only the messages written by the Welcome Buddy from each chat space. Only this text data was the subject of the chat space analysis. Chats were removed from analysis if the members moved the chat to an external platform (e.g. WhatsApp). After all removals there were 555 samples of chat space data.

The objective of the chat space data analysis was first to extract formality, helpfulness and warmth metrics and second to analyse the relationship between these three metrics and the Welcome Buddy performance score. Formality, helpfulness and warmth metrics were assigned manually. I created the rubric below to methodically assess and assign formality, helpfulness and warmth scores to each Welcome Buddy. Each score was awarded irrespective of engagement from new students in the chat.

score	**formality**	**helpfulness**	**warmth**
1	extremely casual tone, uses slang, emojis and/or symbols in place of words, doesn't always communicate in complete sentences	unresponsive to questions asked by students	distant tone, not particularly welcoming
2	familiar tone, communicates in mostly complete sentences, minimal slang	acknowledges and answers queries but doesn't extend themselves beyond what is directly asked by students	welcomes students to campus but does not go beyond friendliness expected of Welcome Buddies
3	highly professional tone, always communicates in complete sentences	makes concerted effort to act as a resource for new students irrespective of student engagement	extremely welcoming and open, makes concerted effort to be encouraging and kind

Table 3: formality, helpfulness and warmth scoring rubric

After assigning formality, helpfulness and warmth scores to each chat, I conducted association analyses between these three metrics and the overall performance score.

demographic analysis

The demographic data was retrieved from ARU TopDesk. The demographic data that was interrogated for this project includes the campus, university school, university faculty, academic course name, student ID number, compound name, gender, date of birth, ethnicity, nationality and overall performance scores of each Welcome Buddy who participated in the scheme September 2021 trimester and/or January 2022 trimester. The Welcome Buddy performance score is awarded by the Student Experience Team after checking the Welcome Buddy chat spaces twice throughout the trimester to ensure Buddies are following through on their training and delivering the expectations of the scheme. The performance score is on a 1-5 scale, where 1 is the highest score indicating optimal performance. See Supplementary table 1 for the grading criteria.

During preprocessing, any overall score missing values were imputed with the mean of the scores assigned by the Student Experience Team to the particular Welcome Buddy earlier in the trimester. There were a total of 8 missing scores, 7 of which appeared to be the result of an oversight by the Student

Experience Team and 1 was the result of a Welcome Buddy leaving the university before the end of the term. In order to create an “age” variable, a function was applied which took the “date of birth” variable as input and calculated the current age of each Welcome Buddy. All the variables, besides gender which was encoded as a binary variable, were encoded as dummy variables.

The objective of the demographic analysis was to ascertain how relevant variables contributed to Welcome Buddy performance within the Welcome Buddy scheme. This was achieved first with descriptive statistics and then through independence testing. Value counts were extracted in order to visualise campus, gender, age, ethnicity, nationality, university school, university faculty, academic course, and the overall Welcome Buddy performance score distributions. Then the relationships between the following variables were interrogated: gender and campus, overall performance score and campus, faculty and overall score, course and overall score, gender and overall score, and school and gender.

In order to interrogate the association between Welcome Buddy gender and age and Welcome Buddy performance score I ran an ordered logit model. Then since descriptive statistics revealed that females are overrepresented across Welcome Buddy participants, I implemented a multiple linear regression model where all demographic variables were predictors and performance score was the outcome variable. More specifically, the predictors consisted of university school, age, campus, gender, formality, warmth and helpfulness scores.

Results

Like Methods, this section will be divided into three subsections:

1. student participant exit survey data analysis
2. chat space data analysis
3. demographic data analysis

student participant exit survey

The exit survey respondents were the members of the new student cohorts who took part in the Welcome Buddy scheme in the September 2021 and January 2022 trimesters. The responses to the closed-ended survey questions underwent descriptive statistical analysis. Particularly informative variables underwent independence testing. The results of the student participant exit survey are organised by variable.

campus

Figure 2 shows the number of student respondents from each ARU campus.

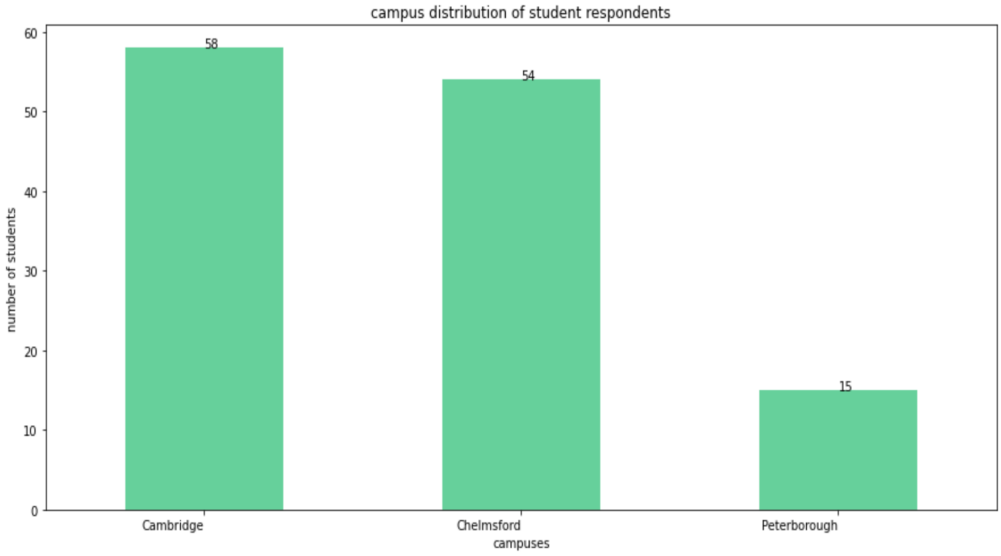

Figure 2: barplot of campus distribution of student respondents

There were 58 respondents from the Cambridge campus, 54 from the Chelmsford campus and 15 from the Peterborough campus. The Cambridge campus has over 10,000 students overall, the Chelmsford campus has over 6,000 students and the Peterborough campus has around 2,000 students.

how_became_aware

The how_became_aware variable contains 7 unique values ("email from the Welcome Buddy team", "email notification from my Welcome Buddy", "email when I accepted my place at ARU", "ARU webpages", "ARU Facebook group", "Open Day", "I didn't receive any information", "Other"), however, respondents had the option to select any number of available responses. As such, in the case of multiple selections, each unique response option should be understood as one of the ways that a respondent became aware of the Welcome Buddy scheme, not the only way they were made aware. Table 4 shows all of the response combinations, in descending order of frequency. The response with the highest frequency (selected by 32 respondents) was "email when I accepted my place at ARU". As this table makes clear, the response combinations to this question are not highly informative. For example, the single response "email from the Welcome Buddy team" has the second highest selection frequency, while the answer combination of "Open Day" and "email notification from my Welcome Buddy" is tied for the lowest selection frequency.

	how_became_aware
Email when I accepted my place at ARU	32
Email from the Welcome Buddy team	25
Email from the Welcome Buddy team,Email notification from my Welcome Buddy	16
Email when I accepted my place at ARU,Email from the Welcome Buddy team,Email notification from my Welcome Buddy	9
Open Day	8
Email notification from my Welcome Buddy	7
Email when I accepted my place at ARU,Email from the Welcome Buddy team	6
I didn't receive any information	3
Open Day,Email when I accepted my place at ARU,ARU Webpages,Email from the Welcome Buddy team,Email notification from my Welcome Buddy	3
Email when I accepted my place at ARU,ARU Webpages,Email from the Welcome Buddy team,Email notification from my Welcome Buddy	3
ARU Webpages,Email from the Welcome Buddy team,Email notification from my Welcome Buddy	2
Email when I accepted my place at ARU,ARU Facebook group,Email from the Welcome Buddy team	2
Email when I accepted my place at ARU,ARU Webpages,Email from the Welcome Buddy team	2
Other	1
Email from the Welcome Buddy team,Email notification from my Welcome Buddy,Other	1
Email when I accepted my place at ARU,ARU Facebook group	1
ARU Webpages	1
Open Day,Email from the Welcome Buddy team	1
Open Day,ARU Facebook group	1
Email when I accepted my place at ARU,Email notification from my Welcome Buddy	1
Email when I accepted my place at ARU,ARU Facebook group,Email from the Welcome Buddy team,Email notification from my Welcome Buddy	1
Open Day,Email notification from my Welcome Buddy	1

Table 4: frequency of unique "how_became_aware" responses

In order to gain a more informative picture of the how_became_aware variable, I extracted the percentages of respondents who selected each response (regardless of multi-selection) subset by campus, displayed in Figure 3 as stacked bar plots. The distribution of each response across the three campuses is summarised in the table below. Each cell should be interpreted as the percentage of respondents from campus X who selected the response Y.

Percentage of respondents per campus who were made aware via each method

	Cambridge	Chelmsford	Peterborough
email from the Welcome Buddy team	55.2%	53.7%	66.7%
email notification from my Welcome Buddy	27.6%	40.7%	40.0%
email when I accepted my place at ARU	58.6%	42.6%	20.0%
ARU webpages	12.1%	7.4%	0%

ARU Facebook group	6.9%	1.9%	0%
Open Day	8.6%	9.3%	26.7%
I didn't receive any information	3.4%	1.9%	0%
Other	3.4%	0%	0%

Table 5: Percentage of respondents per campus who were made aware via each method

first_contact

This variable asked respondents when their Welcome Buddy first contacted them. As illustrated by Figure 3, most students reported that their Welcome Buddy first contacted them during Welcome Week.

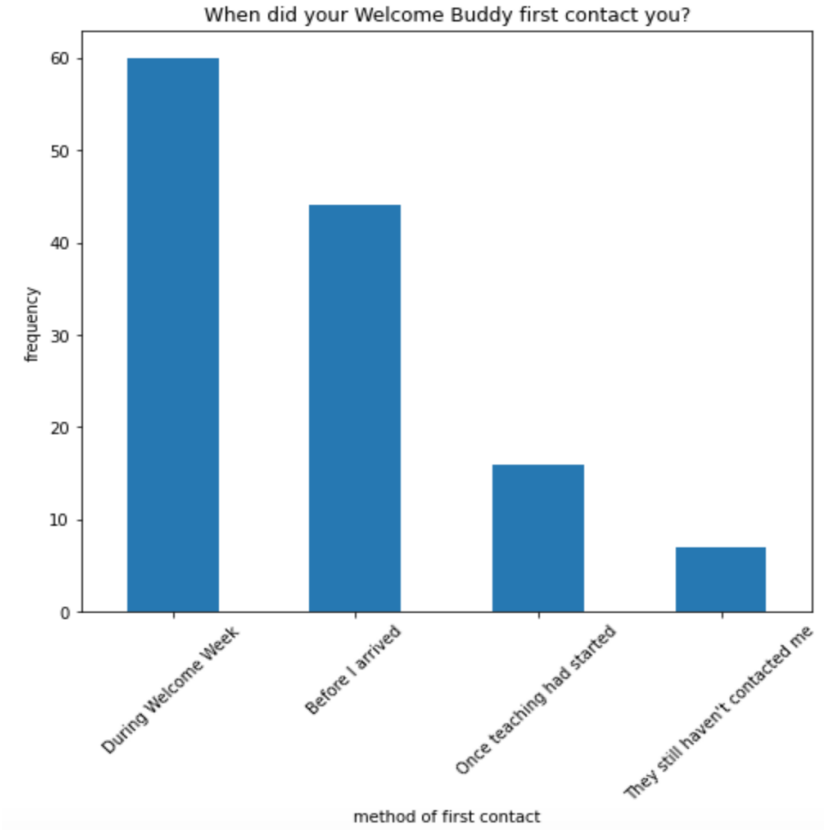

Figure 3: frequency of each method of first contact by Welcome Buddy

Table 6 and Figure 4 show the response breakdown per campus.

first_contact	first_contact			
campus	Before I arrived	During Welcome Week	Once teaching had started	They still haven't contacted me
Cambridge	29.31	51.72	12.07	6.90
Chelmsford	38.89	42.59	14.81	3.70
Peterborough	40.00	46.67	6.67	6.67
All	34.65	47.24	12.60	5.51

Table 6: percentage of respondents per campus who were first contacted via each method

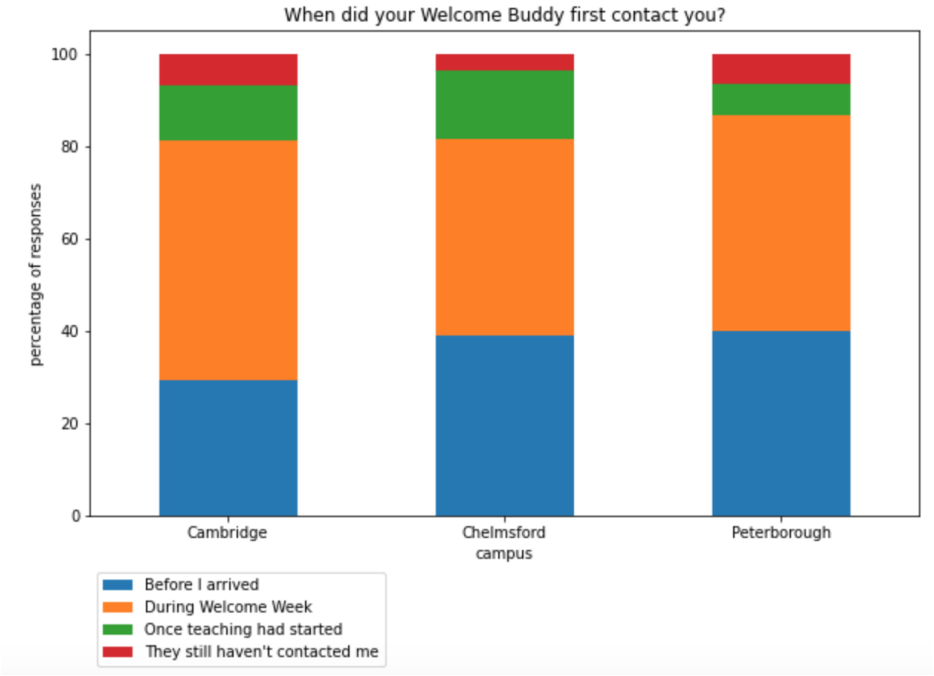

Figure 4: percentage of each method of first contact by Welcome Buddy per campus

Most students were contacted before arriving on campus. A chi-squared test indicated that there is no statistically significant relationship between respondents' campus membership and when their Welcome Buddy first contacted them ($p = 0.8598$).

met_up

This variable ascertains if and how students met up with their Welcome Buddy in the first few weeks of classes. The raw counts of respondents who selected each method is displayed in Figure 5 below.

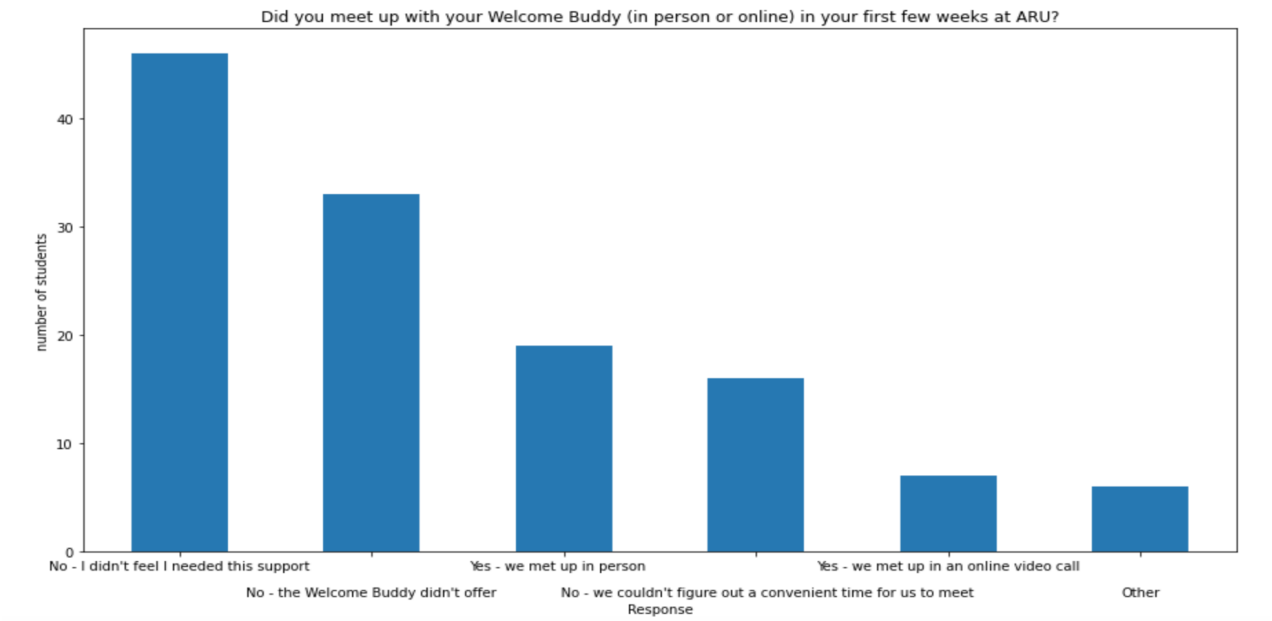

Figure 5: frequency counts of whether respondents met up with their Welcome Buddy

The two most frequent responses indicate that most students did not meet up with their Welcome Buddy in the first few weeks of classes as a result of either the Welcome Buddy's lack of engagement or the student's lack of interest. The met_up responses are stratified by campus in Figure 6 below.

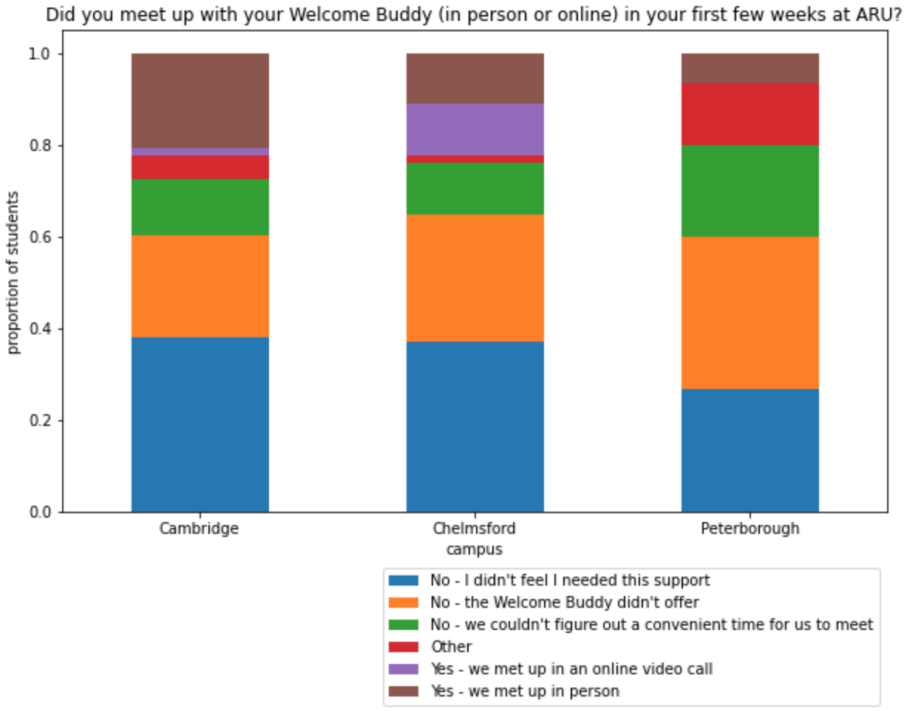

Figure 6: whether respondents met up with their Welcome Buddy by proportion of respondents per campus

Nearly identical percentages of students from the Cambridge and Chelmsford campuses reported they did meet with their Welcome Buddy (either online or in-person) during the beginning of classes. 22.41% of Cambridge respondents met with their Welcome Buddy and 22.22% of Chelmsford respondents met with their Welcome Buddy. There was not a statistically significant relationship between the campus respondents and whether or not they met up with their Welcome Buddy ($p = 0.2190$).

By way of exploratory analysis, I encoded this variable as a binary variable, splitting along "yes" and "no" responses, and omitting "other" in an effort to detect any relationship between campus and a low-cardinality version of met_up. In order to interrogate this relationship I also had to encode campus as dummy variables. However, a chi-squared test revealed no significant relationship between campus and the binarized met_up variable (met_up & Cambridge $p = 0.5994$; met_up & Chelmsford $p = 0.7850$; met_up & Peterborough $p = 0.1999$).

meet_new

The meet_new variable contains information about whether or not respondents felt that the Welcome Buddy scheme helped them meet new students. Figure 7 displays the percentage of respondents from each campus who selected each response.

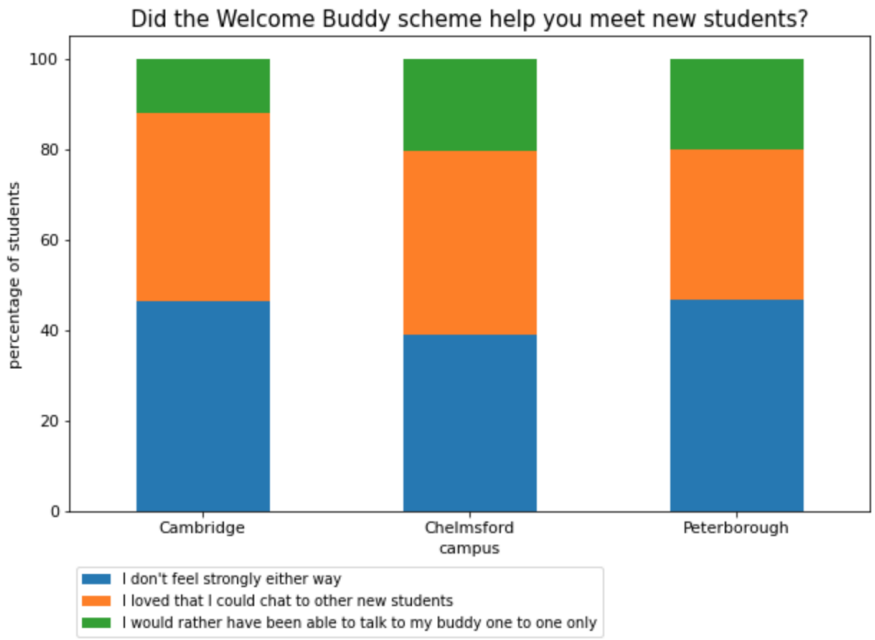

Figure 7: whether respondents met new students through the Welcome Buddy scheme by percentage of respondents per campus

Most students reported they "don't feel strongly either way" about the friend-making capacity of the Welcome Buddy scheme.

I ran a chi-squared test to interrogate the relationship between campus membership and whether or not respondents felt the scheme helped them make friends. There was not a statistically significant relationship between these two variables ($p = 0.7519$), indicating that there is no evidence of dependence between campus and whether the scheme helped respondents make friends.

friends

The friends variable contains responses to the question, “Are you friends with any of the new students that are in the same group chat as you?”. The group chat in question refers to the “Ask ARU” chat pace in which Welcome Buddies could message their new student group. The stacked bar plot below displays the percentages of respondents from each campus who selected each response.

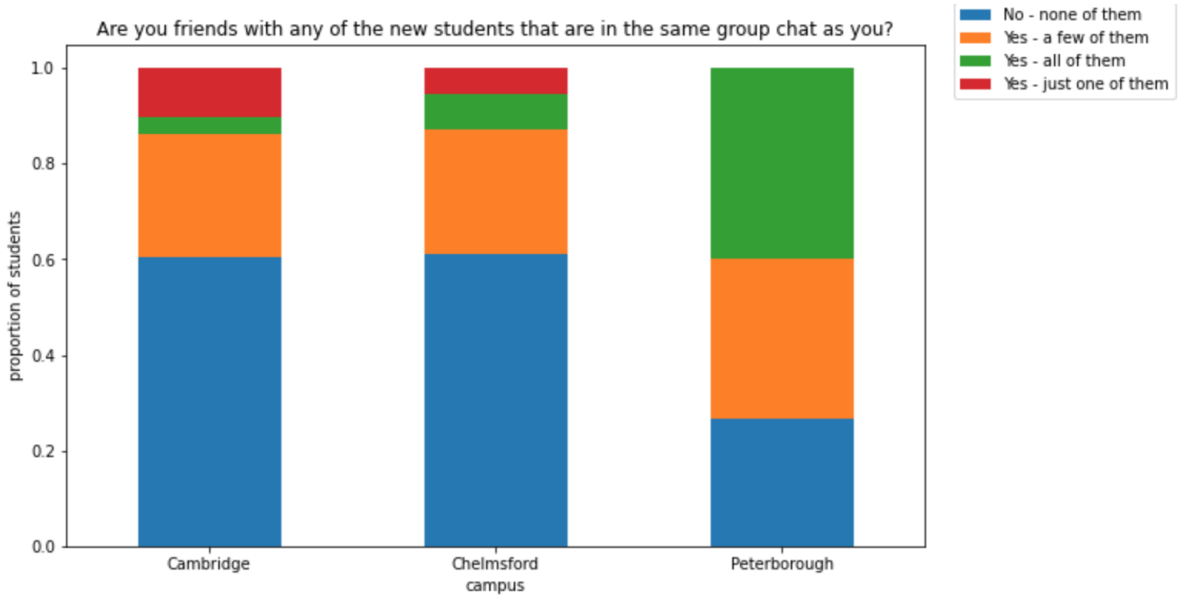

Figure 8: whether respondents are friends with any new students in their group chat, by proportion of respondents per campus

Across campuses, most students reported that they are not friends with any of the students in their Welcome Buddy group chat, with the exception of Peterborough. In Peterborough, all students reported either that they were not friends with any of the students in their group chat or were friends with at least a few of them. The chi-squared test indicated a statistically significant relationship between respondents’ campus and whether they are friends with any new students in their group chat ($p = 0.0010$). This indicates dependence between campus and whether or not students in the Welcome Buddy scheme became friends with each other.

To further explore this relationship, I binarized the “friends” variable (1 = “yes”, 0 = “no”). Once the responses were grouped into two responses instead of four, the significant relationship between “campus” and “friends” disappeared in Cambridge ($p = 0.4464$) and Chelmsford ($p = 0.3875$). However the relationship was significant for students in Peterborough ($p = 0.0124$). This is likely the result of the size of the Peterborough campus. The student population on the Peterborough campus is smaller than those of Cambridge and Chelmsford so it is likely that far more students in Peterborough know each other. As a result, Peterborough students are more likely to be friends with a higher percentage of the students on their campus than students in Cambridge and Chelmsford, where the campuses are larger. Smaller campuses often foster tighter-knit student bodies than larger campuses because the campus social life rests on the participation of relatively fewer individuals on smaller campuses. Therefore, the “friends” variable may not be capturing the social impact of the Welcome Buddy scheme, but the social differences that are the result of different campus sizes.

how_helped

The “how_helped” variable contains the free text responses to the survey question which asked respondents how the Welcome Buddy scheme helped them settle into life at ARU. This variable underwent supervised and unsupervised sentiment analysis. The first unsupervised analysis was with a pretrained DistilBERT model. This model indicated sentiment homogeneity within the responses, assigning the following sentiment scores:

“sadness” score: 0.0098,
“joy” score: 0.9342,
“love” score: 0.0043,
“anger” score: 0.0417,
“fear” score: 0.0079,
“surprise” score: 0.0021

According to DistilBERT, over 93% of text responses expressed joy, indicating overwhelmingly positive feedback from respondents about the helpfulness of the Welcome Buddy scheme.

The next unsupervised sentiment analysis was achieved first by vectorizing the text with a word2vec model. One of the capabilities of word2vec is that after vectorization, it can extract the words within a corpus which are the most contextually similar to a given target word within the corpus. A target word can be any word within the model’s vocabulary and is the reference point for extractions like vector similarity. According to this word2vec model, the following words were the top five most similar words to the target word “helpful”: “queries”, “fundamentals”, “websites”, “knowing”.

In order to visualise the word2vec word embedding, I applied principal component analysis. A scatterplot of the word2vec is seen in Figure 9 below.

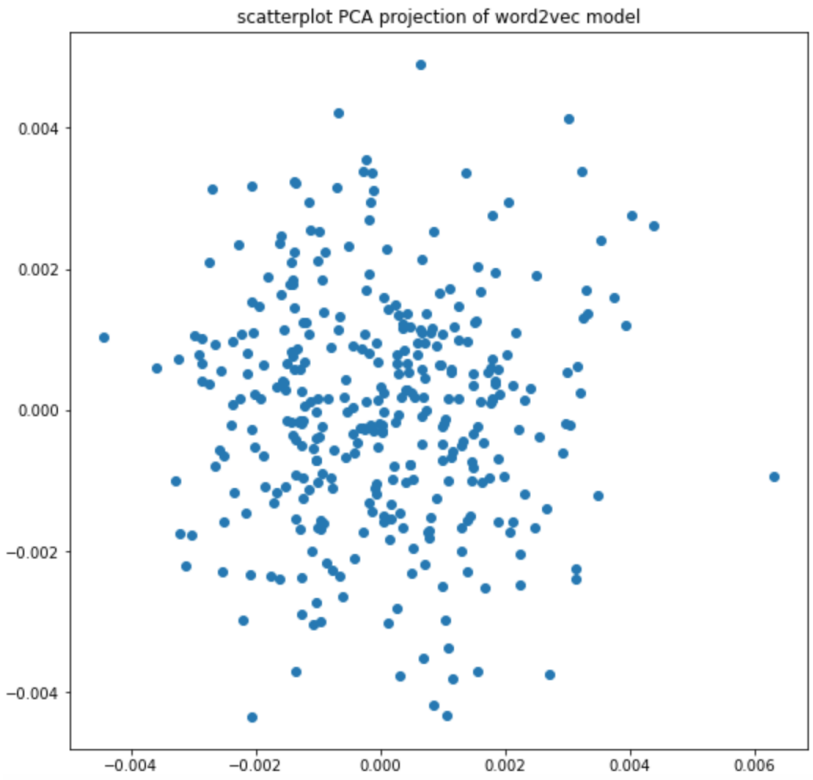

Figure 9: scatterplot projection of responses after word2vec embedding

Using a Mini-batches K-means clustering algorithm, clustered the vectors into two clusters. However, as is evident from the scatterplot above, there is a lack of clear cluster structures within the vectors. Instead, they largely form a single cluster. This was confirmed by the clustering algorithm, which placed 125 vectors in the first cluster and 5 vectors into the second cluster. In order to evaluate the clustering algorithm, I used the Silhouette coefficient. This coefficient is an evaluation metric frequently used in problems where ground truth labels are not known. It is calculated using the mean intra-cluster distance and the mean nearest-cluster distance, and the coefficient value lands between -1 and 1. Well defined clusters result in coefficient values closer to 1 while incorrect clusters will result in values closer to -1. The Silhouette coefficient (S) was 0.40, indicating high performance. When the number of clusters was increased, the coefficient value consistently decreased (when k = 3, S = 0.10; when k = 4, S = 0.08), indicating two clusters was the optimal clustering scheme for this data.

After the assignment of binary sentiment scores, I extracted the percentages of positive and negative responses within the “how_helped” variable. The breakdown of positive and negative responses is displayed in Figure 10.

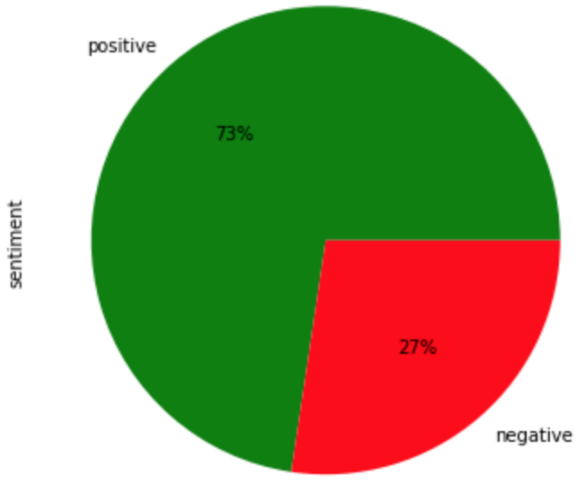

Figure 10: percentage of positive and negative responses after manual classification

73% of the responses were positive and 23% were negative. The number of positive and negative sentiments stratified by campus is displayed in Figure 11.

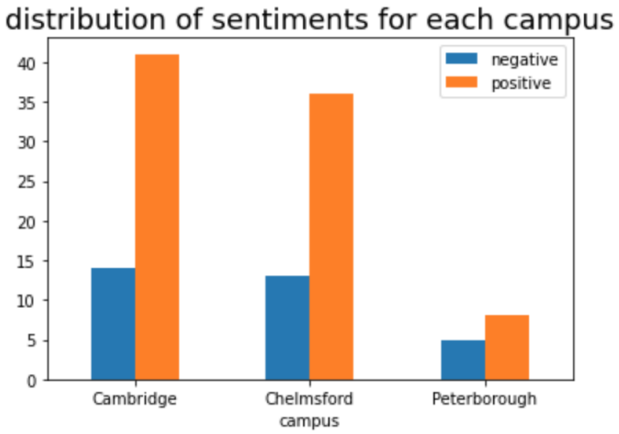

Figure 11: campus breakdown of positive and negative responses after manual classification

After undergoing cleaning and embedding with a TF-IDF vectorizer, a logistic regression classified the sentiment of the manually assigned responses. The logistic regression classifier had an accuracy of 77% on the test set. The performance of the logistic regression was further evaluated with a confusion matrix, classification report (which includes accuracy, precision, F1-score and support calculations) and the area under a ROC curve. The confusion matrix reported 1 true positive, 8 false positives, 0 false negatives and 26 true negatives. The classification report is deployed in table 7.

	precision	recall	f1-score	support
0.0	1.00	0.11	0.20	9
1.0	0.76	1.00	0.87	26
accuracy			0.77	35
macro avg	0.88	0.56	0.53	35
weighted avg	0.83	0.77	0.70	35

Table 7: evaluation metrics for supervised sentiment classification

The F1-score indicates that this classifier is better at identifying positive responses than negative ones. However, since the dataset is imbalanced (there are far more positive responses than negative responses) the performance of the logistic regression is difficult to interpret. Nonetheless, the classifier is highly sensitive which is desirable for this particular classification problem. The area under a ROC (.91 for the logistic classifier, versus the no skill curve which has an area of .50) indicates a high performing model.

```
No Skill: ROC AUC=0.500
Logistic: ROC AUC=0.910
```

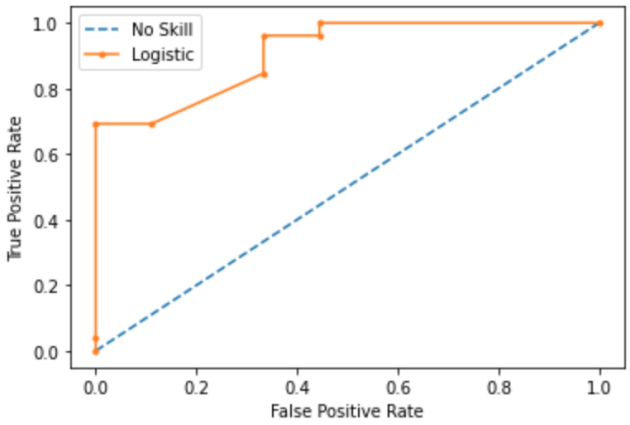

Figure 12: area under the ROC curve, evaluation model of performance of logistic regression

suggestions & comments

These two variables contain free-text, open-ended input from students about the Welcome Buddy scheme. After manually extracting the recurring topics from both the "suggestions" and "comments" variables, I extracted the results displayed in Figure 13. Figure 13 consists of the most popular topics and the associated complaints, consequences and suggestions alluded to within the responses to both survey questions.

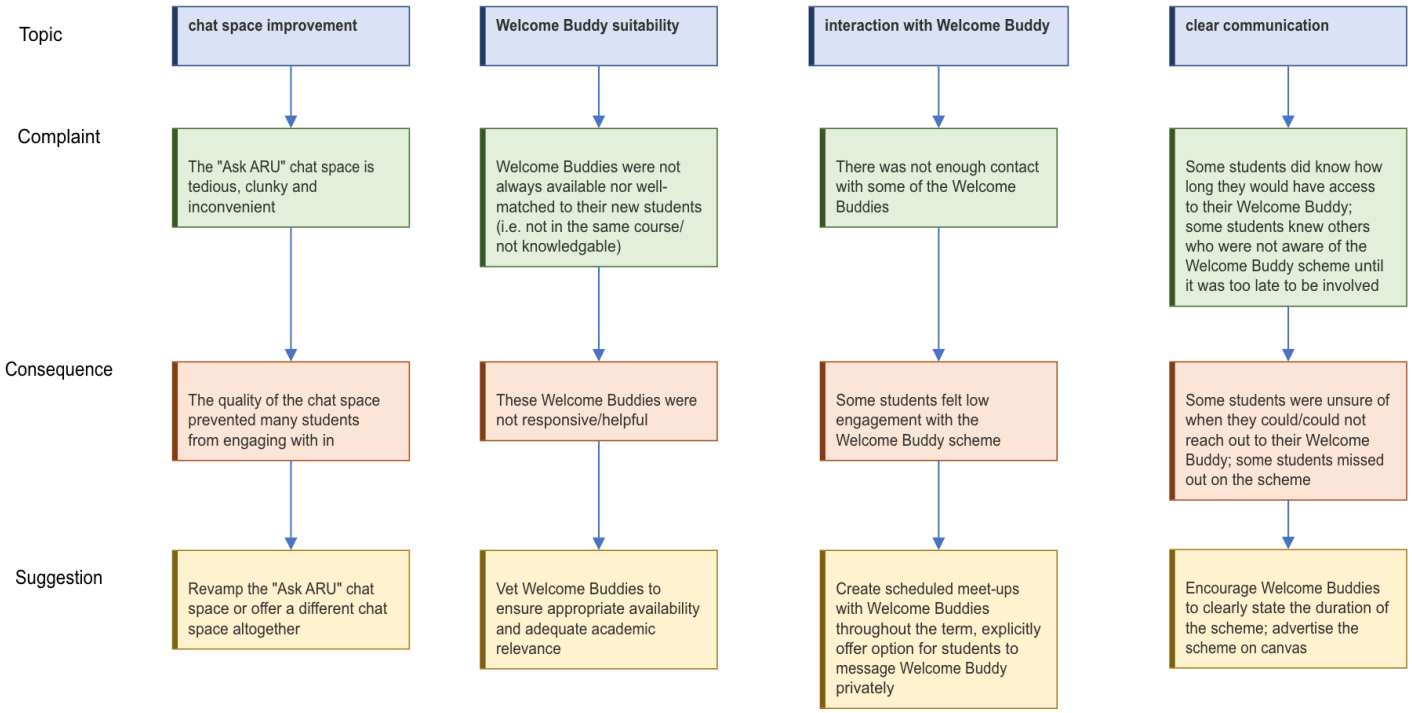

Figure 13: input from student respondents

chat space data analysis

The chat space analysis consisted of interrogation of the chat content sent by Welcome Buddies to their new student cohort within the "Ask ARU" chat space. After extracting messages sent by Welcome Buddies and manually assigning each Welcome Buddy a formality, warmth and helpfulness scores based on the content of their messages, I extracted the correlation coefficients between formality, warmth and helpfulness scores and each Welcome Buddy's overall performance score, which was assigned by the Student Experience Team. The correlation between formality and overall score ($r = 0.0627$) was not significant ($p = 0.14$). The correlations between warmth and overall score ($r = -0.4325$) and helpfulness ($r = -0.4178$) were both significant ($p < 0.05$). These two strong correlation coefficients indicate that as a Welcome Buddy's chat content becomes warmer and more helpful, their overall performance score improves (note, a lower performance score from the Student Experience Team indicates a better performance).

demographic analysis

The demographic analysis involved interrogation of various demographic variables of Welcome Buddies who participated in the scheme in the September 2021 trimester and/or the January 2022 trimester. This subsection will first cover descriptive statistics of demographic variables of interest and will conclude with an analysis of the regression models run on the demographic variables.

Descriptive statistical analysis

campus

As seen in Figure 14, the number of Welcome Buddies per campus resembles the trend among the student respondent survey data.

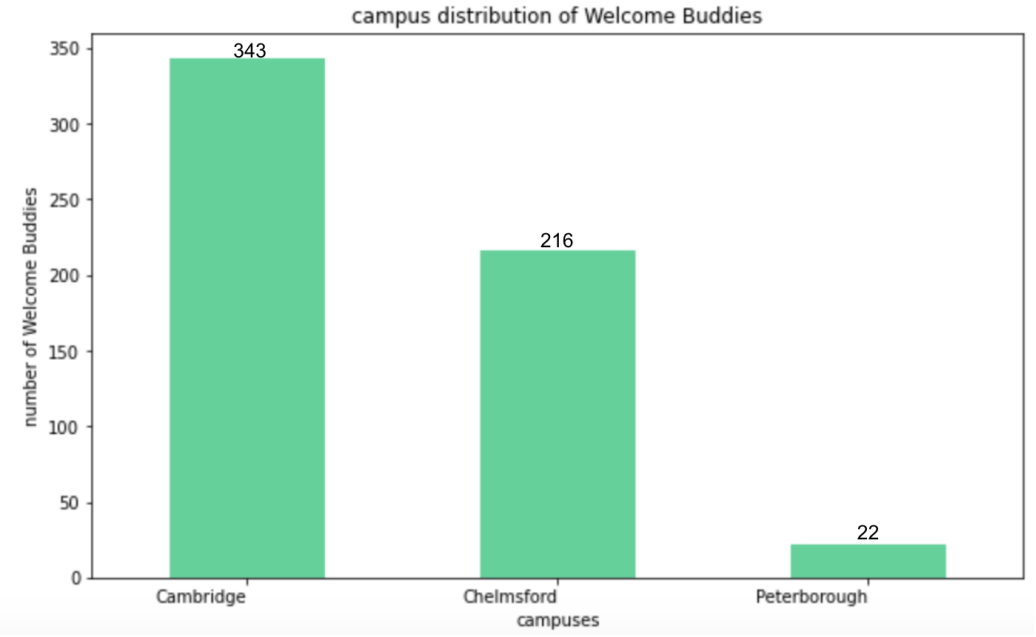

Figure 14: number of Welcome Buddies per campus

gender

Figure 15 demonstrates that the gender breakdown of Welcome Buddies is dominated by female students (452 female, 118 male). Note, any students whose gender identification information was not available on ARU TopDesk were omitted from this analysis (n = 571).

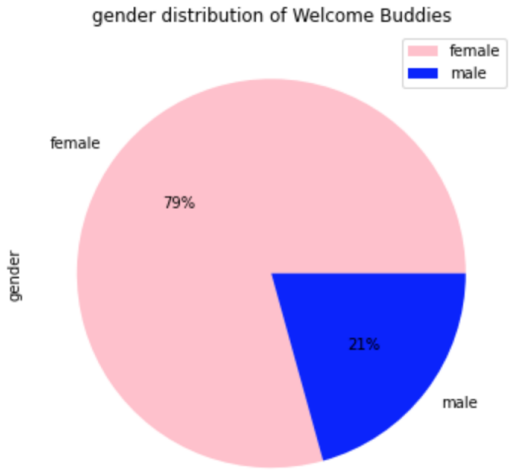

Figure 15: gender of Welcome Buddies

Figure 16 demonstrates the gender breakdown of Welcome Buddies across campuses. In Cambridge the gender breakdown of Welcome Buddies was 77% female and 23% male, in Chelmsford 81% of Welcome Buddies were female and 19% were male and in Peterborough 100% of Welcome Buddies were female.

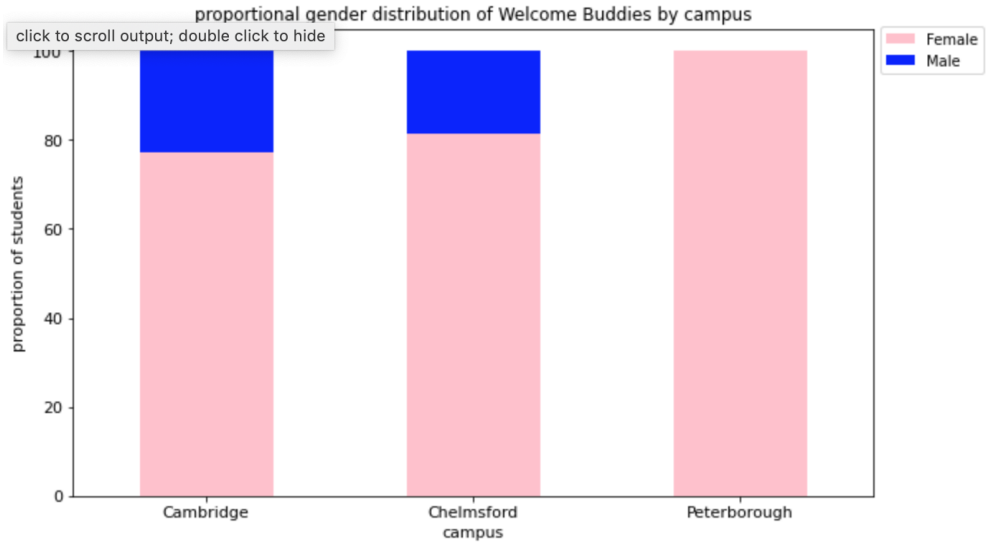

Figure 16: gender of Welcome Buddies by campus

age

The ages of Welcome Buddies is displayed in Figure 17, which demonstrates a wide distribution of Welcome Buddy ages. Any Welcome Buddies whose ages were not available on TopDesk were omitted from this analysis (n = 570). The mean age was 24.7 years old and the median age was 21.0 years old. Most Welcome Buddies were 25 years old and younger (76.5%).

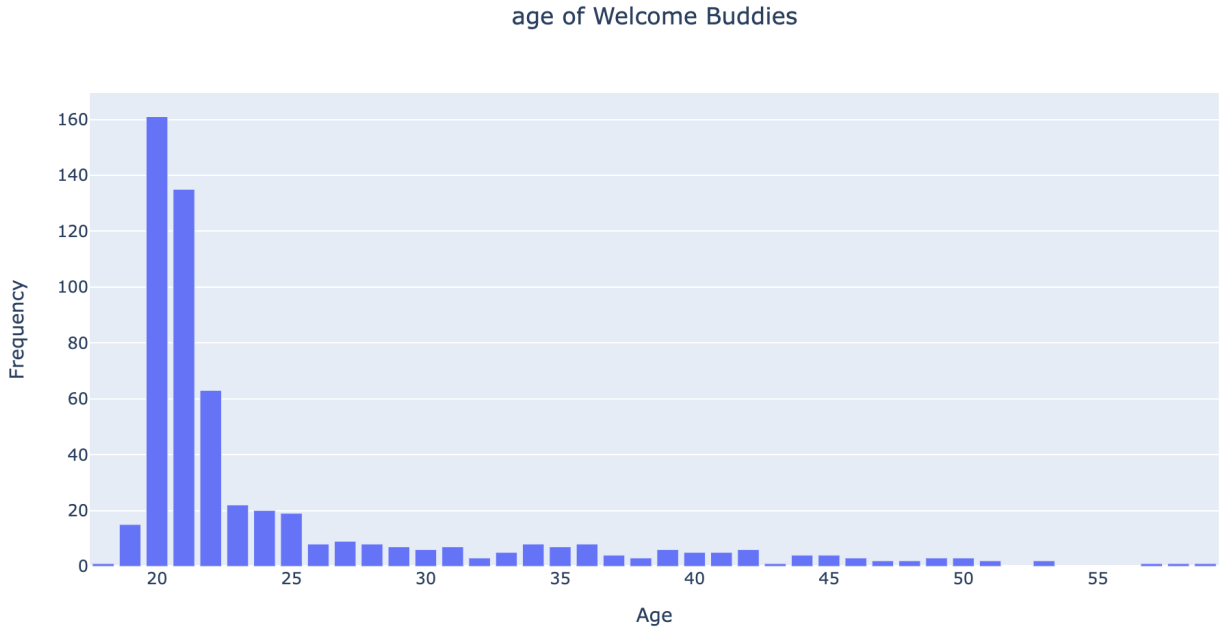

Figure 17: age counts of Welcome Buddies

ethnicity

The ethnic breakdown is displayed in Figure 18 which demonstrates that the majority of Welcome Buddies were white. Any Welcome Buddies whose ethnic information was not available were omitted from this analysis (n = 563).

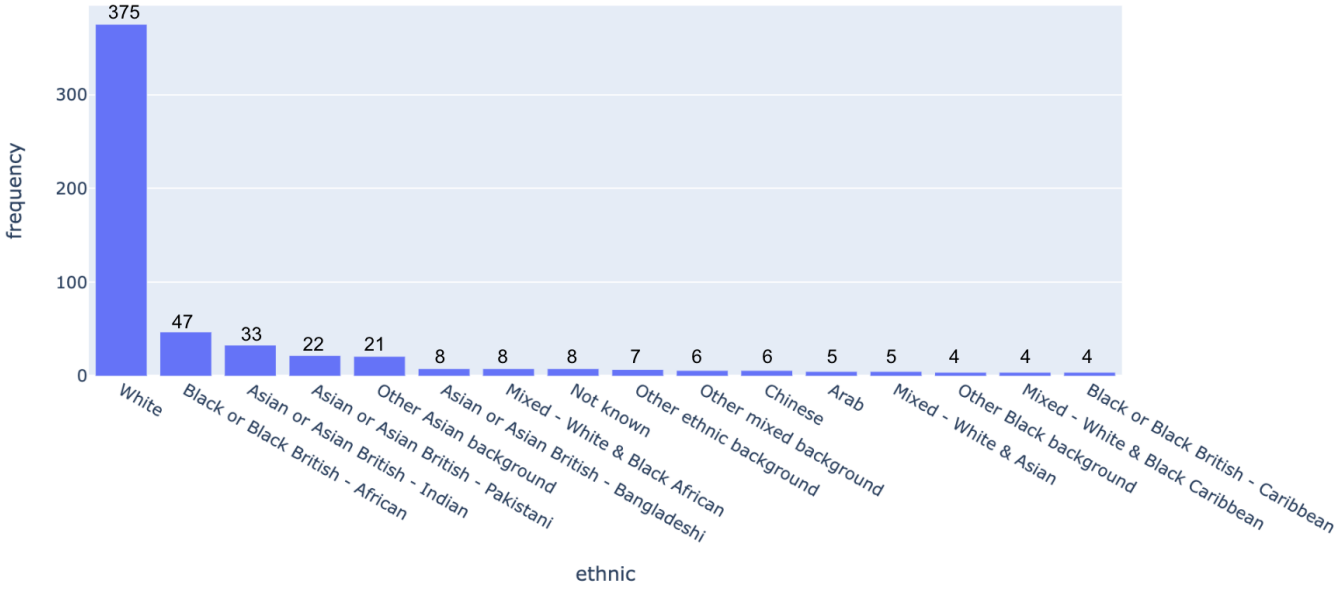

Figure 18: ethnic counts of Welcome Buddies

nationality

Figure 19 demonstrates the nationality breakdown of Welcome Buddies. The five most frequent nationalities were UK national (393), Portuguese (20), Indian (16), Italian (16) and Romanian (15).

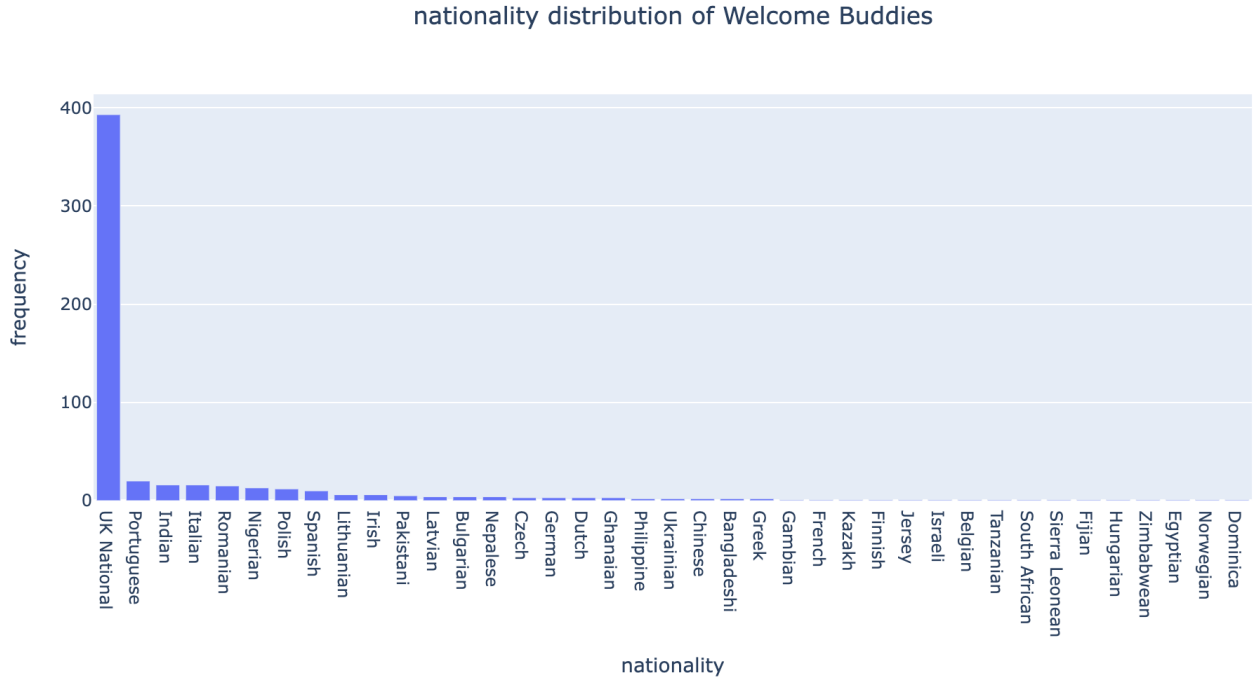

Figure 19: nationality counts of Welcome Buddies

faculty

Figure 21 displays distribution of Welcome Buddies across faculties. The ARU faculties consist of the Faculty of Health, Education, Medicine and Social Care (HEMS), Faculty of Science and Engineering (SE), Faculty of Arts, Humanities and Social Sciences (AHSS), Faculty of Business and Law (BL). Most Welcome Buddies are from within HEMS.

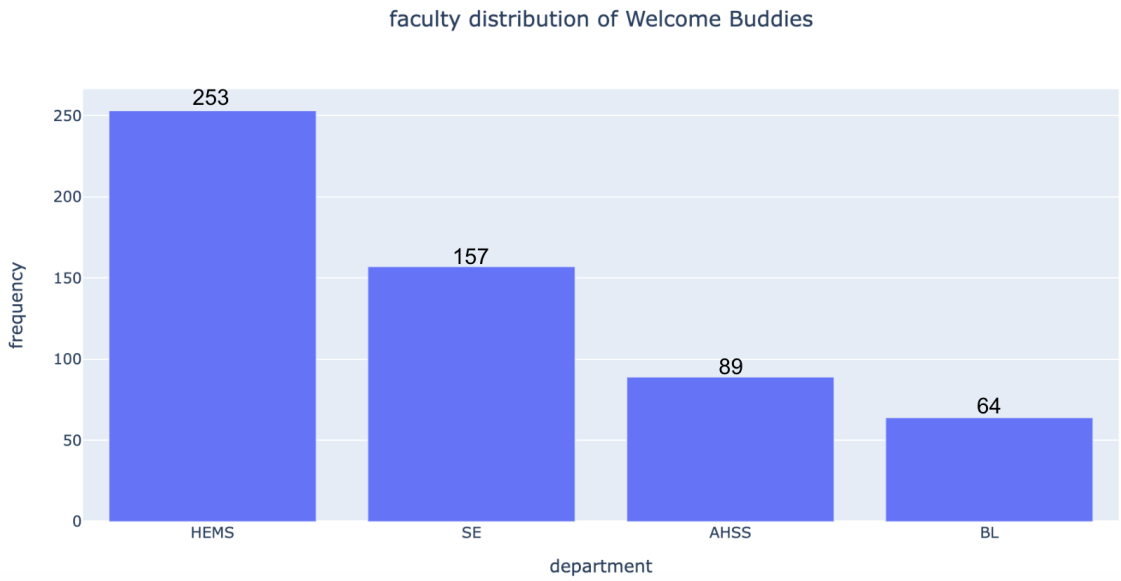

Figure 20: faculty membership counts of Welcome Buddies

course

Interrogating the distribution of Welcome Buddies by course revealed that the five courses with the most Welcome Buddies included Midwifery (48), Primary Education Studies (37), Nursing - Adult (37), Paramedic Science (32) and Biomedical Science (30). Since the number of courses across Welcome Buddies was too large for succinct visualisation (n = 90), I split this variable by frequency. Supplementary tables 2 and 3 list courses from which there are at least 5 Welcome Buddy representatives and the courses from which there are less than 5 Welcome Buddy representatives, respectively. The histogram in Figure 21 displays the course count frequency distribution. Analysis of the "course" variable revealed that most courses (58) had less than 5 Welcome Buddy representatives.

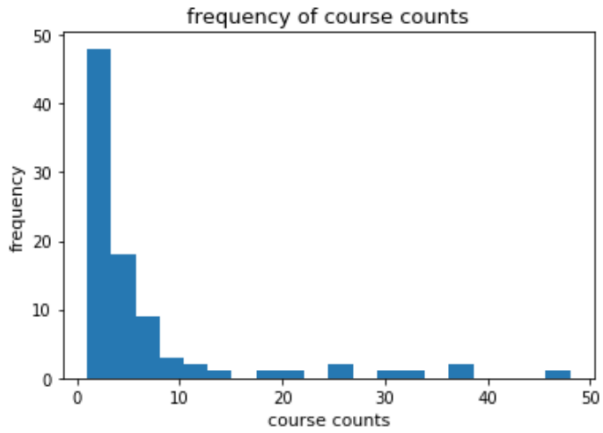

Figure 21: histogram of course count frequencies

school

With the exception of the Cambridge School of the Creative Industries, the names of schools are consistent across ARU campuses. Interrogating the school distribution of Welcome Buddies across campuses revealed that across campuses, the school with the highest number of Welcome Buddy representatives was the School of Nursing and Midwifery. The school with the lowest number of Welcome Buddy representatives was the School of Engineering and the Built Environment. Figure 22 displays the school distribution of Welcome Buddies.

.

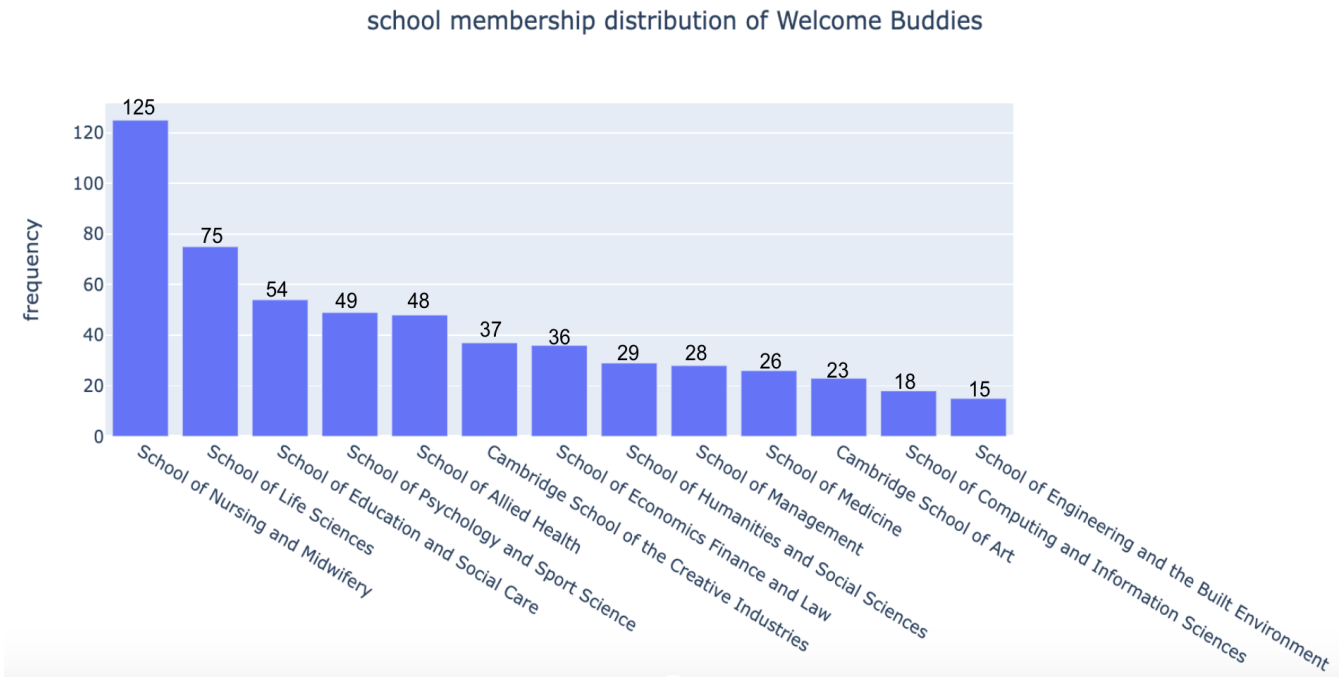

Figure 22: school counts Welcome Buddies

overall performance score

The overall performance score assigned by the Student Experience Team to each Welcome Buddy falls within a 1 to 5 scale, where the highest score is a 1 and the lowest is a 5. The distribution of overall scores indicates that the score with the highest frequency is a score of 3, as indicated by Figure 23.

Figure 23: distribution of Welcome Buddies' performance scores

Regression analysis

In order to assess the relationship between multiple demographic variables and Welcome Buddies' overall performance scores, I ran a number of regression models on the data. Since the females were overrepresented within the population of Welcome Buddies, the association between most individual demographic variables and overall performance score would likely be confounded. For example, extracting the correlation between course and overall performance score in isolation from any other predictor variables was untenable, since the courses with highest frequencies of Welcome Buddy representatives were female-dominated courses across ARU campuses. The overrepresentation of female Welcome Buddies in the courses with the highest Welcome Buddy representatives is demonstrated in Table 8.

course name	**number of Welcome Buddy representatives**	**number of female Welcome Buddy representatives**
Midwifery	48	48
Nursing (Adult)	37	35
Primary Education Studies	37	31

Table 8: number of Welcome Buddies per course vs. number of female Welcome Buddies per course

Predictably, the same trend exists across university schools as well.

school name	number of Welcome Buddy representatives	number of female Welcome Buddy representatives
School of Nursing and Midwifery	125	120
School of Life Sciences	75	62
School of Education and Social Care	54	47

Table 9: number of Welcome Buddies per school vs. number of female Welcome Buddies per school

In order to assess associations between these highly dependent variables, I conducted different types of regression analyses. First I ran an ordinal regression in which Welcome Buddy gender and age were predictors and overall performance score was the outcome variable. This regression model indicated that there is no significant association between age and overall performance score, and that women are more likely to receive lower (better) performance scores than men (gender ordinal coefficient = -0.6743, p-value = 0.00044).

To further interrogate Welcome Buddy performance, I ran a multiple linear regression model on the following predictors: age, gender, campus, university school, chat-space formality, chat-space helpfulness and chat-space warmth. This model treated the overall performance score as a continuous variable. The forest plot in Figure 24 displays the coefficients and 95% confidence intervals.

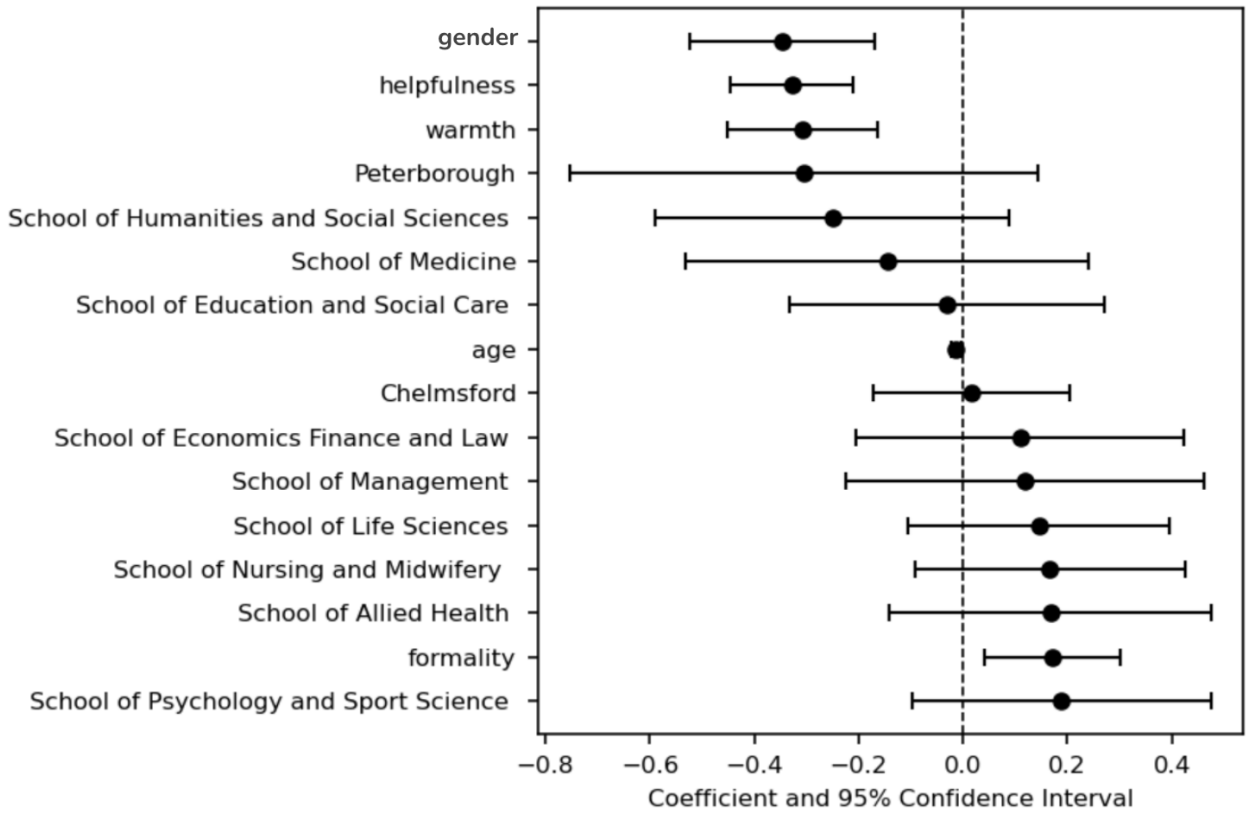

Figure 24: forest plot of coefficients and 95% confidence intervals of predictors from multiple linear regression model

Chat-space formality (coefficient = 0.1709, p-value = 0.011463), chat-space helpfulness (coefficient = -0.3271, p-value = $1.0x10^{-7}$), chat-space warmth (coefficient = -0.3080, p-value = $2.91x10^{-5}$), gender (coefficient = -0.3470, p-value = $1.257x10^{-4}$) and age (coefficient = -0.0136, p-value = 0.0058) are all significantly associated with overall performance score. This indicates that Welcome Buddies are more likely to have high overall performance if they are older, female and are informal, highly helpful and warm in their chat-space communications with new students.

Dashboard

In order to centralise the key findings which would be particularly informative for the Student Experience Team, I created a PowerBI dashboard, displayed in Figure 25. The dashboard includes visualisations of the age, campus, performance ratings, gender, university school and formality/helpfulness/warmth score distributions of Welcome Buddies. Ideally, the Student Experience Team will use this dashboard to inform the future direction of the scheme. The visualisation dashboard will also be used to report to Senior members of the Student and Library Services Management team. The dashboard is also attached as Supplementary file 4.

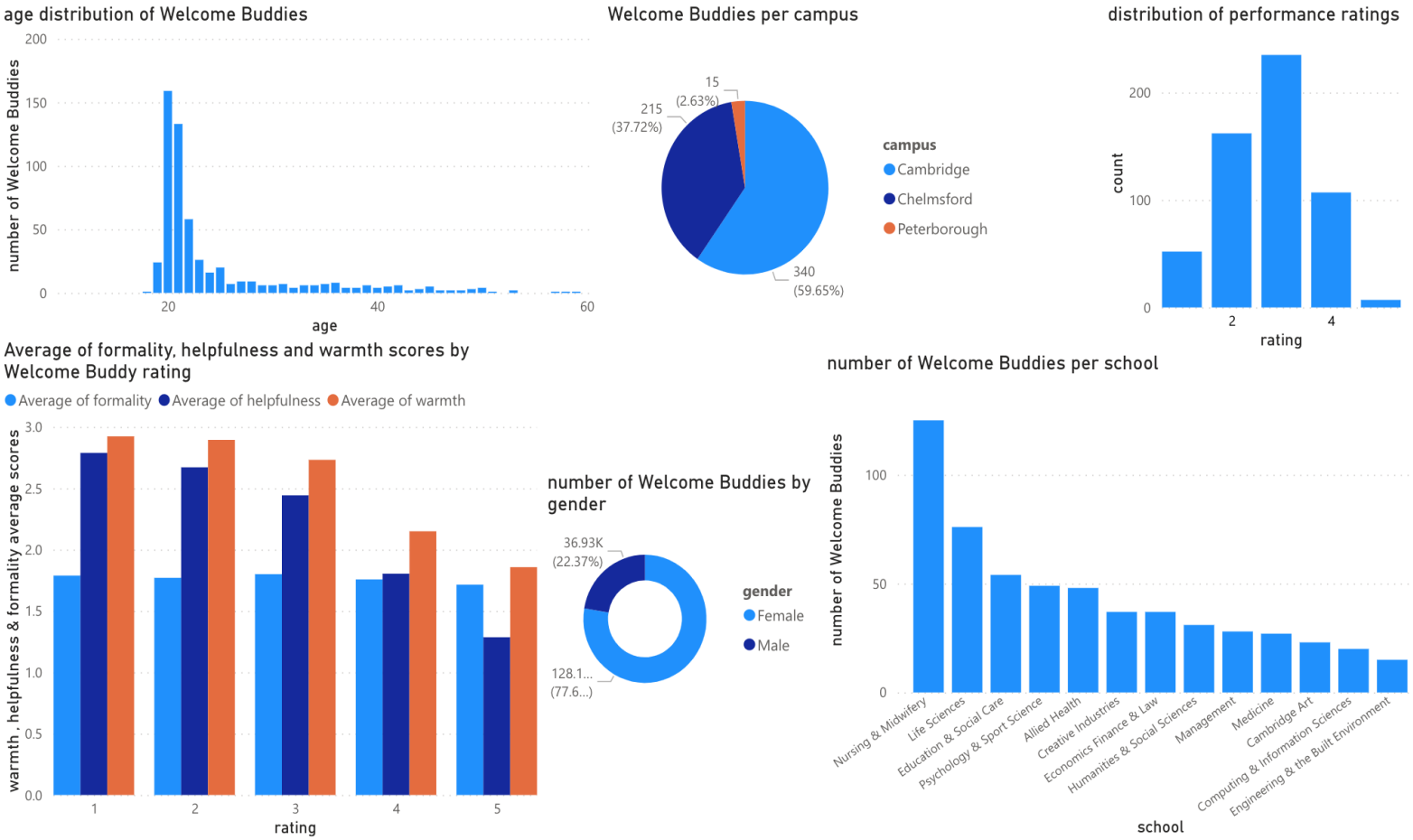

Figure 25: dashboard summarising key analyses for Student Experience Team

Discussion

student participant exit survey

The lack of significant relationships between “first_contact”, “met_up”, “meet_new” and which ARU campus a student attends are positive findings for the Student Experience Team. This indicates that there are probably not any campus-specific factors which are affecting these aspects of the administration of the Welcome Buddy scheme. However, responses within the “friends” variable are dependent on campus. 56.96% of student respondents reported that they are not friends with any of the students in their group chat. These responses can be confounded by the COVID-19 pandemic, which certainly inhibited social activities for (especially new) students across universities. Nonetheless, responses to this survey question could also indicate that students were not necessarily well grouped, and/or that students did not feel comfortable/inclined to interact with each other in the chat spaces. The “met_up” variable is also potentially confounded by the fact that respondents were engaging with the Welcome Buddy scheme during the COVID-19 pandemic. At that time, the Student Experience Team encouraged Welcome Buddies to offer video calls in lieu of face-to-face meetings, even though many Welcome Buddies and new students also elected to meet each other in-person. The context of the pandemic indicates that the responses contained within the “met_up” variable should not be extrapolated to Welcome Buddy trends beyond the setting of the COVID-19 pandemic. This survey question would have been more informative if it included an option for respondents to indicate that they did not meet up as a result of the pandemic.

All analyses of free-text data (i.e. the open-ended responses within the “how_helped”, “comments” and “suggestions” variables and chat space data) were complicated by the widespread use of euphemistic, highly polite language. This extremely polite communication style is likely a result of British social and cultural tendencies. For this reason, the results of analyses of free-text data should be interpreted with a British communication style in mind.

The supervised and unsupervised sentiment classifications of the “how_helped” variable revealed the homogeneity of responses to the survey question asking respondents how the Welcome Buddy scheme helped them settle into life at ARU. All three sentiment analyses indicated that the majority of respondents felt positively about the Welcome Buddy scheme.

A notable result of analysis of survey responses is the fact that students’ experiences with the scheme seemed to be consistent across ARU campuses. Any differences in student experiences with the scheme (e.g. “friends” variable) seemed to be the result of differences in campus dynamics (e.g. total student population size), not administrative differences in the scheme between campuses. This conclusion reflects positively on the Student Experience Team’s administration and coordination of the Welcome Buddy scheme across campuses.

chat space data analysis

The chat space findings provide the Student Experience Team with actionable insights with which to train future Welcome Buddy cohorts. Emphasising informal, warm and helpful communication styles could

improve new students' experience with the scheme, thereby furthering the Student Experience Team's mission to improve student engagement, retention, and success across ARU.

The Student Experience Team requested that I also assess customer (student) satisfaction within the chat spaces. However, such an analysis was not feasible because only positive feedback from students was expressed within the chat spaces, if any feedback was shared at all. In order to make analysis of student satisfaction possible in the future, the Student Experience Team could issue a survey pertaining to the chat space communications to all new students within each group chat at the end of the trimester.

demographic analysis

Demographic analyses revealed that Welcome Buddies were overwhelmingly female. The Student Experience Team may want to focus recruitment efforts on male students in order to increase the diversity of their Welcome Buddy representatives. Analysis of the courses with Welcome Buddy representatives also revealed that most courses had less than 5 Welcome Buddy representatives. Perhaps the Student Experience Team may want to focus recruitment efforts within these under-represented courses, unless the overall enrollment within these courses is already low and therefore proportional to the low number of Welcome Buddy representatives.

Additionally, a demographic variable which would have been interesting to interrogate would have been whether or not Welcome Buddies and new students were commuter students. However, this information was not available at the time of this project.

Conclusion

This project addresses the significant data analysis needs of the Student Experience Team in order to enable the Team to make data-driven solutions about the future directions of the Welcome Buddy scheme. There were three data sources which were the subject of analysis: student participant exit survey, chat space and demographic data. This project had four objectives:

1. Analyse student responses within open-ended and closed-ended survey data
2. Extract formality, helpfulness and warmth scores for each Welcome Buddy from the "Ask ARU" chat spaces of Welcome Buddies and their assigned new student cohort
3. Interrogate the diversity of Welcome Buddies across demographic categories and associations between Welcome Buddy performance scores and relevant demographic characteristics
4. Create a dashboard for the ARU Student Experience Team to visualise the impact and future direction of the Welcome Buddy scheme

Through the use of supervised and unsupervised sentiment analysis, various encoding techniques, and regression models, results indicate:

1. high satisfaction among new students with the Welcome Buddy scheme.
2. high Welcome Buddy performance is significantly associated with Welcome Buddies who are female, older, and engage in informal, highly helpful and warm communication within their chat spaces.

These findings demonstrate the utility of supervised and unsupervised sentiment analysis, descriptive statistics, and regression models in the evaluation of student ambassador schemes at universities.

As this project is a Master's project which was the result of only a week-long placement with the Student Experience Team at ARU, the computational scope of the analyses was limited. Therefore, the analyses implemented in this project are by no means novel. Nonetheless, this project implements standard, well-researched and appropriately applied statistical, machine learning and natural language processing tools and techniques in a novel context. As a future direction, I would implement more advanced and computationally intensive natural language algorithms in order to gain more nuanced insights from free-text data.

Supplementary Materials

Supplementary Table 1: Chat space rubric

Chat space checks – criteria. September 2021

Having offered the scheme to new students, we check the chat spaces twice throughout the trimester to ensure buddies are following through on their training and delivering the expectations of the scheme which are:

- Post an initial welcome message in their chat ASAP once we notify them that it's ready.
- Welcome any subsequent students who are added to their chat.
- Answer any questions that come through – if in doubt, email our team.
- Organise a face to face meet up (up to them if it's video call or in person)
- Engage for the whole trimester by continuing to post messages at least once a month until December

Rating	TW3/4	TW8/9	TW12
1	They are doing everything required and are being generally brilliant (regardless of engagement from their new students)	They are checking in with their students regularly still and going above and beyond	Above and beyond, amazing!

2	They are just missing one element from above	They are checking in with their students once a month	Have written a really good message since our last check in
3	They are doing okay	They have posted only one message since the last check in	Have written something since our last check in but very brief / haven't done anything since but were doing well up until that point
4	They have posted an initial welcome message and nothing since	They haven't done anything since the last check in	If they were a 4 last time and haven't done anything since
5	They haven't even posted an initial welcome message	n/a	n/a

Chat space checks – criteria. January 2022

Based on a system of continual improvement, we updated our criteria for the most recent round of the Welcome Buddy scheme and brought forward our checks. We found from the September scheme that TW3/4 was too late for the first check – if a buddy was rated a 5 at that point, it was very unlikely that new students would engage if they finally posted 4 weeks in.

*Due to Covid restrictions, buddies weren't encouraged to offer a face to face meet up, instead this was changed to a video call.

Rating	TW1	TW6	TW12

1	They are doing everything required and are being generally brilliant (regardless of engagement from their new students)	They are checking in with their students regularly still and going above and beyond	Above and beyond, amazing!
2	Reached out several times and offered a video call OR Reached out several times but hadn't yet offered a video call	Reached out several times since our last check in	Have written a really good message since our last check in
3	Reached out once and offered a video call	They have only reached out once since the last check in	Have written something since our last check in but very brief / haven't done anything since but were doing well up until that point
4	They have posted an initial welcome message but didn't offer a video call	They haven't done anything since the last check in	If they were a 4 last time and haven't done anything since
5	They haven't even posted an initial welcome message	n/a	n/a

Supplementary table 2: courses with 5 or more Welcome Buddy representatives

Supplementary table 3: courses with less than 5 Welcome Buddy representatives

Supplementary file 1: survey data analysis script

Supplementary file 2: sentiment analysis script

Supplementary file 3: demographic analysis script

Supplementary file 4: dashboard

References

1. Cammel, S. A. *et al.* How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. *BMC Med. Inform. Decis. Mak.* **20**, 97 (2020).
2. Abirami, A. M. & Askarunisa, A. Sentiment analysis model to emphasize the impact of online reviews in healthcare industry. *Online Information Review* **41**, 471–486 (2017).
3. Wang, Y. & Xu, W. Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decis. Support Syst.* **105**, 87–95 (2018).
4. Arnulf, J. K., Dysvik, A. & Larsen, K. R. Measuring semantic components in training and motivation: A methodological introduction to the semantic theory of survey response. *Hum. Resour. Dev. Q.* **30**, 17–38 (2019).
5. Netzer, O., Feldman, R., Goldenberg, J. & Fresko, M. Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science* **31**, 521–543 (2012).
6. Newman, H. & Joyner, D. Sentiment Analysis of Student Evaluations of Teaching. in *Artificial Intelligence in Education* 246–250 (Springer International Publishing, 2018). doi:10.1007/978-3-319-93846-2_45.
7. Nkomo, L. M. & Daniel, B. K. Sentiment Analysis of Student Engagement with Lecture Recording. *TechTrends* **65**, 213–224 (2021).
8. Giang, N. T. P., Dien, T. T. & Khoa, T. T. M. Sentiment Analysis for University Students' Feedback. in *Advances in Information and Communication* 55–66 (Springer International Publishing, 2020). doi:10.1007/978-3-030-39442-4_5.
9. Rani, S. & Kumar, P. A Sentiment Analysis System to Improve Teaching and Learning. *Computer* **50**, 36–43 (2017).
10. Powers, D. & Xie, Y. *Statistical Methods for Categorical Data Analysis*. (Emerald Group Publishing, 2008).

11. Hancock, J. T. & Khoshgoftaar, T. M. Survey on categorical data for neural networks. *Journal of Big Data* **7**, 1–41 (2020).
12. Potdar, Pardawala & Pai. A comparative study of categorical variable encoding techniques for neural network classifiers. *Int. J. High Risk Behav. Addict.*
13. Target Encoder — Category Encoders 2.5.0 documentation. https://contrib.scikit-learn.org/category_encoders/targetencoder.html.
14. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv [cs.CL]* (2013).
15. Mikolov, Sutskever & Chen. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.*
16. Blei, Ng & Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*
17. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **41**, 391–407 (1990).
18. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
19. Gilboa, I. De Finetti's Theorem. in *Theory of Decision under Uncertainty* 89–93 (Cambridge University Press, 2010). doi:10.1017/ccol9780521517324.009.
20. Bengio. Bengio Y., Ducharme R., Vincent P., Jauvin C. *A neural probabilistic language model, Journal of.*
21. Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. Bag of Tricks for Efficient Text Classification. *arXiv [cs.CL]* (2016).
22. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
23. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]* (2018).
24. Schwartz, B. Google: BERT now used on almost every English query. *Search Engine Land*

https://searchengineland.com/google-bert-used-on-almost-every-english-query-342193 (2020).

25. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv [cs.CL]* (2019).
26. Gallan, A. S., Girju, M. & Girju, R. Perfect ratings with negative comments: Learning from contradictory patient survey responses. *Patient Experience Journal* **4**, 15–28 (2017).
27. Bahja, M. & Lycett, M. Identifying patient experience from online resources via sentiment analysis and topic modelling. in *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies* 94–99 (Association for Computing Machinery, 2016). doi:10.1145/3006299.3006335.
28. Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S. & Prendinger, H. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems* vol. 115 24–35 (2018).
29. Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A. & Donaldson, L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *J. Med. Internet Res.* **15**, e239 (2013).