Developing Natural Language Processing tools to enhance Learning Analytics in Higher Education

Third Year BSc Data Science Project Report
CS350

Mohammad Sadiq Habib

Supervisor

Dr. Martyn Parker

University of Warwick
April, 2024

Developing Natural Language Processing tools to enhance Learning Analytics in Higher Education

Mohammad Sadiq Habib

Abstract

Student feedback provides valuable insights into the learning experience of an educational system. These reviews represent an opportunity to identify strengths and areas for improvements within different educational topics as expressed by the primary learning recipients themselves. However, comments provided in learning reviews are qualitative in nature, making them unstructured and difficult to analyse systematically using computers.

This project explores qualitative data within higher education by deploying and developing natural language processing tools that can analyse grammar, words, sentences, meaning (semantics), sentiments, topics, and themes from student reviews. The product created is an automated data dashboard that presents the results as the data passes through the model diagnostic pipeline. The analysis validates the output and cross-correlates it with quantitative data about the same context, preferably from the same survey.

The results show that the data engineering principles and model pipeline developed in this project create tangible opportunities to enrich the analysis of student reviews. More specifically, deploying DistilBERT model for sentiment analysis, LDA algorithm for topic modelling, LexRank and Luhn models for extractive text summarisation, and T5 model for abstractive text summarisation in this order enables the qualitative data analysis to be rigorous, scalable, reconcilable, and receptive to stakeholder feedback.

Keywords— Learning Analytics, Natural Language Processing, Deep Learning, Large Language Models, Transformer Architecture, Text Classification, Text Summarisation, Topic Modelling

Acknowledgements

I would like to thank my supervisor Dr. Martyn Parker for his support and continued guidance throughout this project.

Contents

Chapter 1

Introduction

1.1 Background

The ability of computers to process human text and language has increased by leaps and bounds over the last couple of decades with the emergence of more developed algorithms to interpret language, as well as significant advancements in the processing capabilities of computers [25], [29], [31]. More specifically, the development of the Graphical Processing Unit (GPU) has become one of the most important advances in computing technology as it enables computers to carry out a very large number of tasks in a fraction of the time by separating the problem into a number of parallel tasks [48]. Moreover, from an algorithmic perspective, the development of artificial intelligence (AI) techniques has enabled computers to process and learn from information in a more human-like manner whilst leveraging this GPU-driven ability to process and store large amounts of data [31]. More specifically, there exist both machine learning (ML) and rules-based strategies within AI that give computers the ability to interpret language. Whilst the latter employs many-to-one lexicon relationships and rules through which computers process data, ML enables computers to learn from the data during training [31]. One branch of ML that has experienced significant breakthroughs recently is transformer-based deep learning, where computers process information using nodes similarly to the human brain [29].

Whilst computers employ rules-based strategies to process information, natural language is anything but systematic. In fact, the study of meaning in itself represents a comprehensive branch of language study known as semantics [49]. Moreover, the linguistic elements that add to or even create meaning are extensive in nature but can predominantly be categorised into two main frameworks: morphology, which studies the internal structure of words, and pragmatics, which studies how context contributes to meaning [6]. Both elements attempt to reconcile the infinite number of permutations when writing a text in order to provide meaning. Given the advances in the capacity and language-processing ability of computers, this has given rise to a large number of opportunities to process the internal structure of words as well as the context when attempting to interpret human language in most industries. Education is one prominent sector in which these advances in technology can and must be leveraged. As a pivotal cornerstone of societal growth, deploying state-of-the-art machines and algorithms in education can significantly improve the learning experience of students, i.e. enhance learning analytics, and in turn generate positive ramifications for society [17].

1.2 Definitions

Natural Language Processing (NLP)

Natural Language Processing, also referred to as NLP from hereon, develops methods that gives computers the ability 'to understand text and spoken words in much the same way human beings can' [47].

Learning Analytics Learning Analytics in Higher Education adopts a data-driven approach to enhance learning motivation, engagement, retention, and benefits such as career prospects, problem-solving ability, identifying strengths and weaknesses, and constructing a plan of action to consolidate these characteristics. These objectives serve to ultimately develop a student's 'knowledge acquisition, skill development, and cognitive gains' [17].

Diagnostic Analysis

Diagnostic analysis refers to the outcome and process through which trends, topics, and themes are identified.

Prescriptive Analytics

Prescriptive analytics aims to generate recommendations and solutions that directly address the diagnoses identified in a particular study.

1.3 Motivation

1.3.1 Problem Statement

Within the Higher Educational space, there is a need to enhance learning analytics due to the presence of a multitude of factors, as enumerated below.

Problem 1: The emergence of higher-education alternatives within the last decade has increased the need to optimise learning in higher-education. More particularly, the development of work apprenticeship programmes, as well as the rapid rise of online academies and courses within the educational technology (EdTech) space have provided viable alternative pathways to higher education. A study conducted by Team Multiverse in February 2023 found that people are four times more likely to think that work 'apprenticeships offer young people better job prospects than university' (44% to 11%) [45]. Consequently, enhancing learning analytics provides a way for such institutions to remain at the forefront of education and retain their unique selling proposition.

Problem 2: Institutions have more data than ever before, and the pipelines that have been put in place to collect and collate data are only going to be useful if they can help to monitor performances and innovate. Within education, virtual learning environments such as Moodle provide a plethora of additional data fit for exploration, providing lots of new opportunities to learn about student learning patterns. These provide valuable insights that can be analysed and prescribed from.

Problem 3: The emergence of new tools to aid teaching and learning offers many solutions that can be deployed in combination. Developing learning analytics approaches therefore provides an excellent pathway to explore how to use these tools and in what combinations to enrich the learning experience of students.

Table 1.1: Examples of new tools and developments to aid teaching and learning processes [38]

Tool	Examples
Administration Support in Teaching	Tabula
Virtual Learning Environments	Moodle, Blackboard
Online Bulletin Board/Collaboration Tools	Padlet
Lecture Capture Recordings	Echo360
Online Assessment Tools	Questionmark Perception
Record Ongoing Professional development	Mahara

1.3.2 Gap in Existing Provision

Current learning analytics tools, whilst helpful, have numerous drawbacks that are limiting their contributions in the educational space.

Gap 1: Current learning analytics tools are predominantly quantitative-oriented, whilst qualitative data analysis is primarily conducted by humans without leveraging advances in technology and statistics [17]. The issues with this manual approach are two-fold. First, there is an inability to diagnose from large qualitative datasets through a data mining process that employs computers with strong processing power and heavy-duty GPUs [48]. Among other NLP techniques, data mining would enable an in-depth sentiment and thematic analysis to derive meaningful patterns and insights from the datasets. Moreover, manually sorting through the qualitative data inhibits a cross-correlation analysis between the quantitative and qualitative data. Not only is it an inefficient use of the data, but it also restricts the diagnostic and prescriptive capabilities of the analytics tools used.

Gap 2: Within the learning analytics space, current tools overemphasise on data visualisation as opposed to conducting rigorous analysis and establishing cause and effect relationships or even strong associations and/or correlations. Not only does this structure underplay the "complexity of teaching and learning processes", but it also runs the risk of producing incorrect prescriptive analytics given the lack of statistical rigour and evidence-based judgement that can give proper meaning to the data [17]. Where traditional modelling techniques such as linear regression are used, it is argued that these over-simplify the learning process, which is a multi-dimensional and non-linear problem [17].

Gap 3: Arguably one of the biggest gaps in current learning analytics tools is the apparent overemphasis on analytics rather than learning. Moreover, the underlying objective behind learning analytics has predominantly been to prevent "non-completion in higher education institutions" [17]. However, the latter produces an imbalanced focus on improving grades and "non-completion metrics" as opposed to enhancing learning engagement, retention, experience, motivations, and possibilities [17].

1.4 Objectives

This project ultimately aims to develop and deploy tools that employ NLP techniques to enhance learning analytics by producing justifiable diagnostic analysis and prescriptive analytics on qualitative data and respond to the current problems (1.3.1) by addressing the aforementioned gaps (1.3.2). The following core objectives help the project to accomplish this goal and create tangible value by enhancing learning motivation, retention, engagement, and benefits (1.2).

Core Objectives

1. Create a preprocessing pipeline to engineer the qualitative data. This stage includes extracting, cleaning, transforming, and storing the data.
2. Explore the need for this project, identify the gaps with the current provision, and establish the key NLP concepts that that will directly address these concerns.
3. Develop tools based on the aforementioned NLP concepts and deploy these models to carry out an in-depth diagnosis of the qualitative data.
4. Create a validation pipeline to evaluate both models and findings. This pipeline should include a review of the respective underlying objectives of each model, a cross-correlation with the quantitative data, and a comparison with the opinions of stakeholders such as students themselves.

Chapter 2

Theory

2.1 Data Preprocessing

The common theme in all the datasets is their qualitative form. The ultimate goal is to convert the text into some form where it can be processed and understood by the processing models and visualised to enhance understanding, and this requires preprocessing. Different models require different degrees of preprocessing, and the preprocessing techniques employed by each algorithm are stated in their respective subsections herein Chapter 2. From an overarching perspective, the complete preprocessing algorithm takes a text and can carry out the following processes on it: text normalisation, word normalisation, stop word removal, and tokenisation. The NLP preprocessing methods deployed in this project are stored in the Preprocessor module within the models package of the project repository (3.4). This section explores the different preprocessing techniques and their functionality.

2.1.1 Normalisation

Normalisation transforms the qualitative data into a canonical form where it can be processed by the algorithms [4]. The goal is to normalise both the text and the words, enabling the model to treat similar words such as 'The' vs 'the', and 'helping' vs 'help' vs 'help.' together rather than individually. The two categories of normalisation therefore are text normalisation and word normalisation. While text normalisation is a general technique that includes processes such as separating words from punctuation marks and converting all text into lower-case, word normalisation is a NLP-specific process that attempts to normalise words with the same logical semantics, e.g. changing 'playing' to 'play'. Stemming and lemmatisation are two word normalisation techniques, and have the following differences over one another.

- Stemming adopts a systematic approach, such as implementing rules that eliminate 'ing' and 'ed' from 'explaining' or 'enjoyed' [4]. It is less computationally expensive then lemmatisation, but might output a word that does not exist, for instance producing 'bor' when stemming from 'bored'.

- Lemmatisation is a more intelligent but computationally expensive method that normalises words in such a way that it will always return a dictionary-compliant word [4]. It may not be suitable for exploratory analysis on very large datasets. A suitable dataset size varies on the processing power of one's computer and can be determined empirically.

2.1.2 Stop Word Removal

A straightforward yet key aspect of preprocessing is to eliminate stop words that establish pragmatics and contribute to the semantics of a sequence of words, but which do not have significant stand-alone meaning. Examples of stop words are 'a' and 'the' [31]. The text transformed after this step mostly contains nouns, verbs, and adjectives.

2.1.3 Tokenisation

Tokenisation is the process by which a longer piece of text is separated into smaller, identifiable units called tokens. These units can be words, subwords, or characters [33]. Tokenisation creates a vocabulary of words, subwords, or letters, from the qualitative dataset. The splitting process makes the input text more manageable for a model and can improve both efficiency and results [31]. It is noteworthy that some processing models such as certain deep learning models take as input the text at token level, which implies that there is a need to optimise the way that the text is broken down. For this reason, this project considers numerous tokenisation techniques.

1. **Word Tokenisation**: This technique splits the input text into different words. While it is correct by definition, it can lead to a very large vocabulary generated by the corpus. A corpus refers to the entire input text used for training. This resource-heavy vocabulary must then be embedded for the models to process [16]. Additionally, word tokenisation is not suitable for processing wrongly-spelled words. While one contingency plan is to limit the vocabulary size to the n most frequently-used words, where n is a positive integer input argument, the model may consequently not learn some important words. It will instead represent them as Out Of Vocabulary (OOV) [31].

2. **Character Tokenisation**: This technique splits a text into its characters, and then represents the words by these characters. While this creates a very efficient vocabulary with no OOV or misspelled words (the vocabulary in this case is simply a list of characters such as the letters and punctuation marks in use), the downside is that a character such as 'b' has no meaning on its own. Character tokenisation may also generate a very large sequence length for a given text, leading to complications down the line when processing the tokens [31].

3. **Subword Tokenisation**: Also known as n-gram tokenisation, subword tokenisation attempts to address issues with both word tokenisation and character tokenisation by only splitting rarely used words into smaller, meaningful subwords [16]. For instance, 'algorithmic' is split into 'algorithm' and 'ic', which helps the model process the term 'algorithmic' as a combination of the root word 'algorithm' and suffix 'ic'. Likewise, word tokenisation partitions the term 'oceanic' into 'ocean' and 'ic', further highlighting the syntactic situations where the suffix 'ic' is employed to the processing model [31]. It is important to note that transformer based

deep learning processing models predominantly adopt subword tokenisation to construct their vocabulary of words as means to maintain their dynamism [35].

Finally, the process of tokenisation gives each token a unique id in the vocabulary. The most frequent tokenisers used in the NLP space are **WordPiece** and **Byte-Pair Encoding**, which are both n-gram tokenisers used by deep learning models (2.3.2) such as DistilBERT, GPT-2, and RoBERTa [31].

2.2 Word Embedding

The goal of preprocessing the text, including the tokenisation stage, was to create a computationally-efficient vocabulary with respect to space and time complexity. Word embedding gives a numerical representation to every term in this linguistic vocabulary so that it can be fed to a processing model such as a Neural Network. This process is known as word embedding, or simply embedding, after which each data point, i.e. text, or token, i.e. word, subword, letter, or punctuation mark, is given a n-dimensional vector representation where n is a carefully selected positive integer to balance model plausibility and parsimony [31]. The vectors can be manipulated mathematically as a means of performing operations on the text, which makes it important for the embedding to capture the semantic and syntactic relationships.

In this vein, it is crucial to note that words with similar meanings are expected to have similar vectors in this embedding space. Moreover, the embedding space can also constitute of vectors that represent all types of text sequences, such as sentences, words, subwords, letters, or punctuation marks [31]. This process is not uniquely applicable to word tokens, as the term 'word embedding' may suggest. For clarity, the term embedding, and not word embedding, is used moving forwards. The embedding process involves the following two main considerations.

1. The numerical representation of the tokens must represent the meaning of the tokens. For instance, there must be a clear distinction between the vector representing 'teach' and 'learn' [21].
2. The numerical representation of the tokens must represent the context and relationship of that token in a sequence of tokens that form a word and sequence of words. This is sometimes referred to as distributional semantics, and ensures the construction of Beginning of Sentence (BOS) tokens, End of Sentence (EOS) tokens, and everything in between [21].

There are two classes of embeddings, namely static and contextualised embedding.

2.2.1 Static Embedding

Static embedding gives a vector representation to each data point or token in the corpus. Common techniques include Word2Vec, GloVe, Skip-Gram, FastText and Bag of Words (BoW) [31]. The sentiment and thematic analysis models, where applicable, only employ the BoW embedding technique. This technique is introduced herein this section.

Bag of Words (BoW)

Bag of Words is an embedding algorithm that creates a vector space where the dimensions represent the canonical form of the normalised nouns, adjectives, and verbs present in the data. Each sentence is then represented by a vector enumerating the number of times that noun, adjective, or verb is present in it [44]. Consider two sentences, namely 'the teacher is great' and 'teacher and classes are engaging'. The preprocessing algorithm would include word normalisation, text normalisation, and stop word removal functionality. This will transform the sentences to 'teacher great' and 'teacher class engage' respectively. As a result, for a dataset consisting of these two sentences, the BoW algorithm will produce a 4-dimensional vector representing the words teacher, class, great, and engage. Sentence 1 and 2 would consequently be represented by the vectors (1,0,1,0) and (1,1,0,1) respectively.

A comprehensive evaluation of this embedding technique is determined by its use-case. For the purposes of this project therefore, the BoW model is evaluated in the context of sentiment and thematic analysis of text.

Disadvantages of BoW

1. BoW only considers the number of occurrences of the non-stop words and not their order [44]. If a particular data point lists factors by order of importance, for instance, BoW will give equal weighting to all the listed factors.

2. BoW cannot consider compound words such as 'least helpful' in tandem, and would actually enlist 'least' and 'helpful' as two separate, unrelated words [44]. The inability to consider multi-words and consequently text pragmatics will create forced errors in the embedding process.

3. Sparsity and overfitting: BoW employs word tokenisation, treating every word as a separate dimension. Each document might only use a handful of these words, leading to an embedding where a large number of the feature values in each vector are 0 [31]. This makes the vectors sparse, and their high dimensionality generates high model parsimony (complexity). BoW is therefore associated to overfitting and excessive complexity [44].

Advantages of BoW

1. Easy to build on: BoW can be easily built on to enhance model processing and understanding capabilities (2.2.1).

2. BoW gives a good indication of the subjects in a particular text. By enumerating the occurrences of particular words, each embedding can decipher the most frequently used words and consequently the broader topic in question, for instance sports or politics [31].

3. Ease of use: BoW is a simple embedding algorithm that is easy and intuitive to understand.

Term Frequency - Inverse Document Frequency (TF-IDF)

The canonical form of multiple static embedding techniques can be developed to enhance feature representation of the text. For instance, to add weighting functionality to BoW to any feature value i, multiply it by the number of documents (N) and divide it by the number of documents in which that word appears (n). This is known as the term frequency-inverse document frequency (TF-IDF) algorithm.

$$TF - IDF_i = BoW_i \times \frac{N}{n}$$

Through the above equation, more weighting is allocated to nouns, adjectives, and verbs as opposed to words with little semantics such as 'can' and 'not'. The latter would otherwise have the greatest frequency and therefore weighting despite having little stand-alone meaning [44].

2.2.2 Contextualised Embedding

Contextualised embedding models dynamically generate vector representations for each token in the corpus based on that token's usage in a particular context. Unlike static embedding techniques, they can distinguish between the semantics of polysemous words or homonyms, i.e. words with similar spellings but different meanings such as 'bank' [27]. This dynamic process is enabled by the training mechanism of deep-learning models that employ architectures such as transformers. To understand their functionality, it is important to first go through the basics of a neural network.

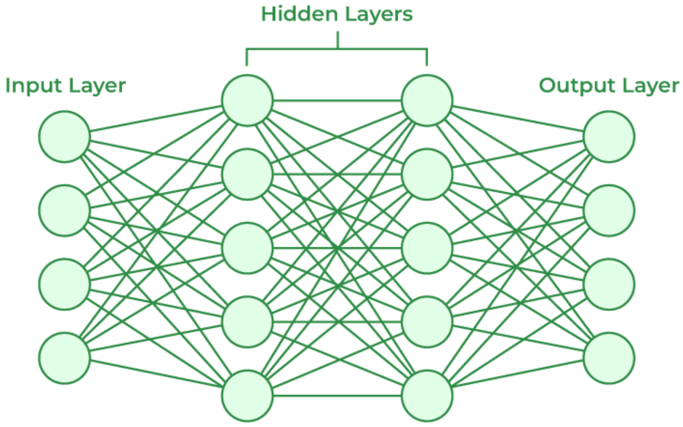


Figure 2.1: Neural Network Architecture [30]

Neural networks are the underlying technology behind deep learning models that allow information to be processed through layers of nodes. These artificial nodes aim to give the model a brain-like ability to process information [31]. As illustrated in Figure 2.1, every neural network consists of at least 3 layers of nodes, namely an input layer, an output layer, and one or more hidden layers. A deep-learning neural-network will usually have three or more hidden layers [39]. Each individual node contains input data, weights, a bias or threshold, and an output. In NLP, there are two key steps in training a deep-learning network for any particular task, these are known as pre-training and fine-tuning via transfer learning [46].

Pre-training

Tokenisation and embedding are two key structures when pre-training transformer-based deep learning models. The numerical representations of each token after the tokenisation and embedding process enable the model to capture the context and relationships of a sequence of words through the attention mechanism [28]. In order for a neural network to process information optimally in its multiple layers of nodes, it goes through the backpropagation cycle which consists of the following stages [32].

1. Forward pass: Data is fed to the input layer of nodes, where it is multiplied by corresponding initialised weights [31]. Based on this product, an activation function determines the output of the node which is then input to the next node in procedural like manner.

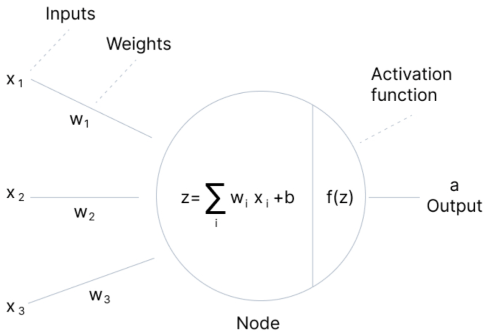


Figure 2.2: Activity in a node [25]

   Common activation functions include but are not restricted to binary step functions, linear functions, sigmoid functions, softmax functions, and rectified linear units [31].

2. Error calculation: The process for each node and subsequently each layer of node continues until an output is generated by the output layer, which is then compared to the intended output. This comparison takes place in the form of a loss function, and could take the form of Mean Squared Error, Mean Absolute Error, or Log Loss, the latter of which is primarily used for classification purposes [31].

3. Backward pass: This step updates the value of the weights employed at each node. Using the error value calculated in the previous step, the gradient of the error is propagated back through the network in descending order of node layers, i.e. starting from the output layer all the way through the hidden layers. To update the weights of the nodes based on their contribution to the error calculation, this step employs the chain rule to find the derivative of the loss function with respect to every node's weight. The value of the derivative indicates the magnitude of error rectification that a change in weight would produce. The learning rate hyperparameter also contributes in determining the size of the weight updates, since a smaller learning rate updates the weights by a smaller amount [32].

4. Weights update: The values of the weights are subsequently updated as part of a process known as gradient descent, where the weights update in the opposite direction of the gradient until the loss function is minimised.

Backpropagation is a cycle since the four aforementioned stages are repeated either for a pre-determined number of iterations, until the loss function returns a value below a desired threshold, or until the model stops improving beyond a certain percentage after every iteration [31].

Fine-tuning

Fine-tuning involves updating the weights and biases of a pre-trained model using a context-specific dataset that is relevant to the particular task that the model is being employed for. The benefits of this step are two-fold, namely leveraging the pre-training knowledge base of the model as well as enhancing its performance for a particular task [32]. The fine-tuning step is elaborated upon in Sections 2.3.2 and 2.4.2, where deep learning models based on the transformer architecture are fine-tuned for classification and clustering with varying degrees of success.

Transformers

The transformer architecture has become a very prominent mechanism in the NLP space since it enables neural networks to employ a self-attention mechanism through which they can capture long-range dependencies when embedding text [31]. This was first identified and published by Vaswani et al., 2017, in a paper titled ‘Attention Is All You Need’ [11]. The diagram below illustrates the architecture of a transformer network.

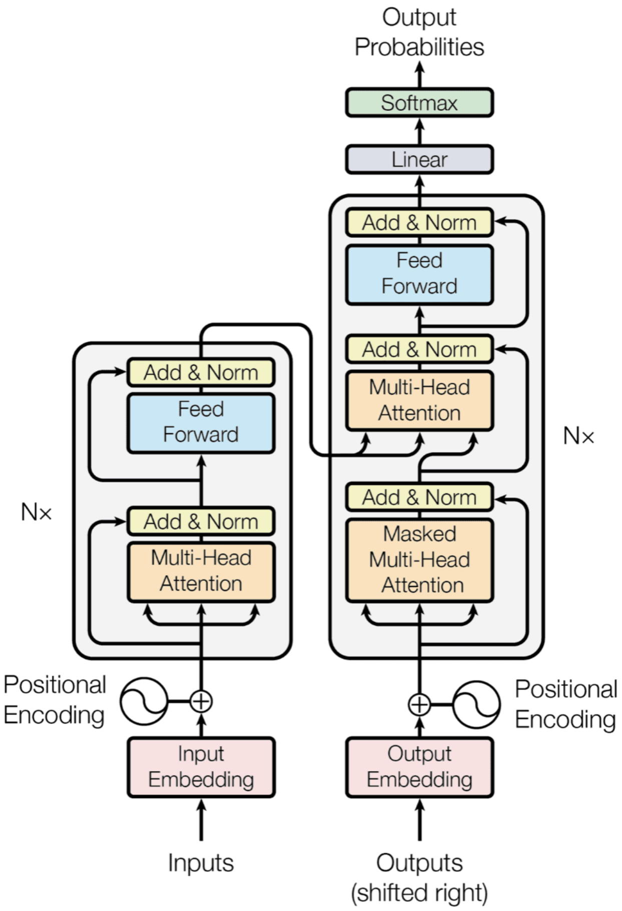


Figure 2.3: Architecture of a Transformer Network [25]

Transformers usually contain an encoder and a decoder, both of which are composed of layers of feed forward networks with self-attention. A transformer processes the whole input text in parallel, which allows it to consider the entire corpus when learning about the context of each token at each node. The **encoder** helps to achieve this by converting the input sequence into a weighted sum of tokens alongside the activation function transformations using the self-attention mechanism and feed-forward network respectively [25]. The former in particular enhances its understanding of the language and ability to capture context [31]. The **decoder** works in the opposite way, generating a new token by token sequence based on the output of the encoder [11]. As the self-attention mechanism is materialised, transformer-based networks can overcome the following limitations.

1. A network processing text sequentially such as a recurrent neural network (RNN) typically experiences vanishing gradient during backpropagation [31]. More specifically, due to the sequential nature of processing, the influence of earlier inputs diminishes exponentially during the gradient descent stage that optimises the loss function. This makes it difficult for RNNs to capture long-term relationships in a text, especially if a particular token references a sequence of text from earlier or later in a big document [43]. Transformers, on the other hand, enable models to uniquely focus on the relevant and crucial pieces of the input sequence.

2. Currently, a network processing text sequentially can also be modified to 'selectively store and retrieve information over long sequences, making it effective for capturing long-term dependencies' [43]. Unlike transformers, however, these sequence-to-sequence models cannot process the input in parallel and must consider every piece of the input sequence when generating an output sequence (decoding) based on the encoded representation. This makes networks such as Long Short-Term Memory (LSTM) computationally inefficient when compared to transformers [31].

One deep learning model based on the transformer architecture is BERT, which stands for Bidirectional Encoder Representations from Transformers and was introduced by Google researchers Devlin et al., 2019, in a paper titled 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding' [12]. BERT is trained on a large corpus of english data and has a two-fold training strategy, namely Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). More specifically, through MLM, the model randomly masks 15% of the words and then attempts to unmask the terms. The model must also predict whether or not two randomly selected sentences follow each other in the data as part of NSP training [12]. In this vein, at each processing phase, BERT considers the whole data corpus in parallel and not sequentially. This makes it bi-directional and able to capture long-range dependencies between words that are referred to later or referenced from earlier [41].

Large Language Models

Deep-learning networks based on the transformer architecture represent the building blocks of all large language models (LLMs) such as GPT-3, T5, and BERT [31]. This relationship stems from the ability to utilise the transformer-based model for a wide range of language tasks by simply adding final layer(s) to the pre-trained network in the fine-tuning phase. This allows the model to produce a wide array of results that leverage its language-understanding capabilities, and is explored further in Sections 2.3.2 and 2.4.2. This concludes the argument that, when provided with the required computational resources and sufficiently large training dataset, a transformer architecture-based approach produces more accurate results than other neural networks [23].

2.3 Sentiment Analysis

Sentiment analysis is an NLP technique that enables classification of the different inputs (reviews) into sentiment categories such as positive, negative, and neutral [31]. This enables each input review to be diagnosed and identified as an issue, strength, or neutral feedback. When performing sentiment analysis, a desirable model would be able to consider the challenges and nuances of the human language. The latter includes but is not limited to synonyms, homonyms, metaphors, irony, and other figures of speech. This project evaluates two approaches to sentiment analysis, namely lexicon-based and deep-learning.

2.3.1 Lexicon-based Approach

A lexicon-based approach in NLP analyses the sentiment of a dataset via a many-to-one pre-determined set of rules. To put this into context, take the example of WordNet, which is a thesauri that categorises words by synonyms and hypernyms (e.g. a panda 'is a' carnivore 'is a' placenta 'is a' mammal) [31]. A lexicon-based sentiment analysis employing WordNet would assign sentiments to certain base words in the thesauri, and then use the WordNet synonym and hypernym framework to determine the features and sentiment of the overarching input text in relation to those base words. Since a lexicon-based approach requires no training and can accept text inputs, this process only requires preprocessing steps until but not including tokenisation [31].

The main advantage of a lexicon-based approach is its time complexity given the pre-determined nature of rules. However, whilst lexicon-based analysis can be relatively simple to implement and interpret, it has several drawbacks as enumerated below.

1. It may struggle to categorise complex, ambiguous, or colloquial text data that includes ironies and metaphors given the linguistic and non-contextualised nature of the methodology.
2. A many-to-one relationship fails to consider the nuances that exist between similar words, especially when it comes to adjectives. For instance, 'a good lecturer' is not necessarily equivalent to 'a proficient teacher' [23]. Since WordNet associates 'proficient' and 'good' to the same word 'good', this can create challenges down the line when it comes to interpreting adjectives.
3. Thesauri resources such as WordNet are difficult to maintain as current terminologies must be retained and new terminologies added, making it difficult to keep the thesauri up to date.

2.3.2 Transformer-based Approach

From an overarching perspective, a deep learning approach to sentiment analysis employs the neural networks explored in Section 2.2.2. More specifically, the goal is to deploy a fine-tuned version of a pre-trained transformer-based model to determine the sentiment of a text. The fine-tuned model employed for the purposes of this project is titled ‘distilbert-base-uncased-finetuned-sst-2-english’ and is developed by Hugging Face [42]. The base model upon which fine-tuning is carried out is DistilBERT.

DistilBERT is a smaller version of the BERT model. It reduces the size of BERT by 40% but still retains “97% of its language understanding capabilities” using the student-teacher theory [27]. In other words, the distilled version of BERT attempts to replicate BERT’s behaviour while employing a model compression technique that reduces the model parsimony whilst still achieving one of the desired backpropagation objectives (2.2.2). To deploy DistilBERT for sentiment analysis under text classification, it suffices to add a layer on top of the pre-trained model to classify the input as positive, negative, or neutral. This allows the sentiment analysis classification model to leverage the transformer-based network’s dynamic and comprehensive grasp of large amounts of text compared to a lexicon-based approach [31].

A distilled version of BERT has been employed over other large language models such as GPT-3 for the following reason. GPT-3, whilst still employing a transformer architecture, is trained on an auto-regressive basis where the network’s backpropagation cycle aims to improve the model’s ability to predict the next value in a sequence based on the previous values in that sequence [40]. This naturally makes it uni-directional and unable to capture long-range dependencies that stem further along a text. GPT-3 nonetheless naturally excels at text generation with no fine-tuning given its pre-training predictive modelling strategy. On the other hand, as explained in Section 2.2.2, BERT is bi-directional, making it more desirable for classification.

2.4 Thematic Analysis

Thematic analysis models enable the process to diagnose strengths, issues, themes, and trends from the data. Diagnoses lead directly to the prescriptive analytics since they point out the main themes of a strength or issue that must be acted upon or rectified respectively. Thematic analysis follows sentiment analysis in the workflow to enable strengths and areas for improvements to be identified from the different sentiment categories respectively. There are multiple thematic analysis techniques in NLP, some of which are explored herein this section.

2.4.1 Keyword Extraction

The goal behind keyword extraction is to extract the most relevant terms that represent themes or topics of a particular text. Quantifying relevance can be subjective, and it is important to establish its meaning. For instance, relevance could refer to the most frequently used words. Alternatively, it could refer to terms that have a higher than average probability of being sequentially co-occurrent, with stronger connections for frequently used pairs. There are multiple libraries in Python that provide keyword extraction capabilities, such as Wordcloud, SpaCy, and TextRank [36]. In this product, the Visualiser module has been designed to provide functionality for two keyword extraction methods, namely word clouds and word frequency graphs.

Both keyword extraction models output the most frequently used terms for the positive, negative, and neutral reviews of each course respectively, but differ in output visualisation. At the most fundamental level, the goal behind the two models is to give an insight into the most frequently used terms for each sentiment category in the literal sense of the term. Nonetheless, while both can be developed to further consolidate their advantages, their predominant limitations cannot be addressed without incorporating other, non-keyword extraction based, thematic analysis techniques.

Advantages of Keyword Extraction Output

1. Visual appeal: Interpreting wordclouds is intuitive as illustrated in the figure below.

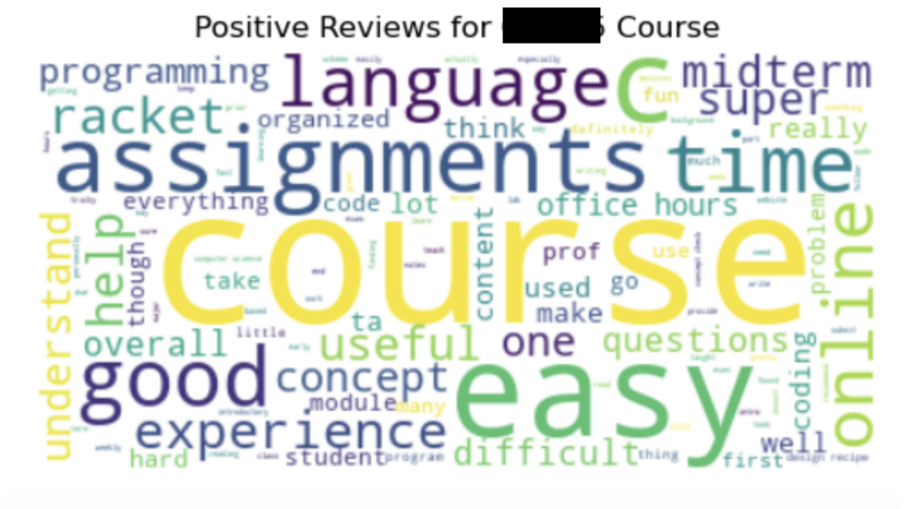


Figure 2.4: Wordclouds for an example set of course reviews

For each sentiment category, word clouds enable one to determine the most frequently used words based on their prevalence in the image. This gives overarching insights into the potential topics of discussion in the reviews, which could be relevant and adequate in certain use-cases such as presentations.

2. Bird's-eye view: Keyword extraction models give an overview into the occurrence of terms in the data. This helps to suggest topics and skims over the feelings related to them. In this case, 'course', 'assignment', 'experience', and 'language' are the key terms discussed by students for this particular course.

Limitations of Keyword Extraction Output

1. Limited associative capabilities: While word frequency graphs can give a more quantitative overview on the terms used in the data when compared to word clouds, they still cannot provide an analytical insight into the occurrence of terms in the data.

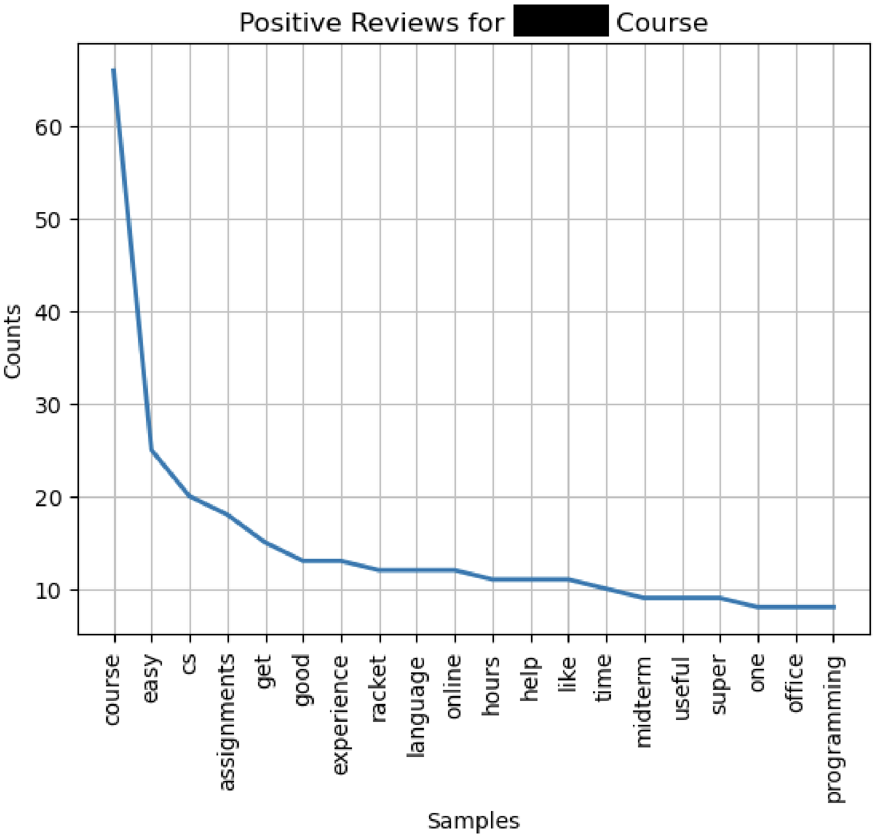


Figure 2.5: Word frequency graph for an example set of course reviews

More specifically, keyword extraction techniques struggle to establish rigorous associations between adjectives, verbs and nouns. For instance, some of the prevalent nouns identified in the graph above are 'course', 'assignment', 'experience', and 'language', and the set of prevalent adjectives are 'easy', 'good' and 'useful'. Exactly what is easy, good, and useful is not clear, and this limits the model's diagnostic abilities.

2. Lack of scope for causality analysis: Since keyword extraction graphs cannot establish associations, it becomes difficult to establish diagnoses and their cause from the dataset in order to achieve core objective three of the project (1.4).

It is for this reason that wordclouds and word frequency graphs have limited applicability in the product. They in fact embody one of limitations in the current provision, namely the overemphasis on data visualisation and the lack of rigour in analysis (1.3.2).

2.4.2 Topic Modelling

Topic modelling consist of different types of algorithms that attempt to fetch topics from a text that aren't necessarily related to the frequency of words. According to Jurafsky, 2023, in his book 'Speech and Language Processing', topic modelling is a more powerful NLP technique than keyword extraction as it can incorporate concepts from statistics, linear algebra, or machine learning within its architecture [31]. Common topic modelling techniques involve Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and transformer-based deep learning (BERTopic). While LSA achieves a very similar output to Keyword Extraction, the other techniques use concepts of conditional probability, bayesian statistics, or neural networks to study topics such as subject relationships and word correlations [24]. The Python libraries employed for topic modelling in this project include Gensim and BERTopic. This section explores numerous topic modelling algorithms and justifies the choice of algorithm for the product pipeline from a theoretical perspective.

BERTopic

BERTopic is a topic modelling algorithm developed using a multitude of NLP concepts such as deep learning, dimensionality reduction, clustering, tokenising, and weighting scheme. In the first place, it leverages the embeddings of the transformer-based BERT model, following which it deploys numerous techniques to model the topics in a text. Whilst different models can be employed for each concept, BERTopic's default workflow employs the following models, and more information can be found in the paper 'BERTopic: Neural topic modeling with a class-based TF-IDF procedure' by Maarten Grootendorst, 2022 [19].

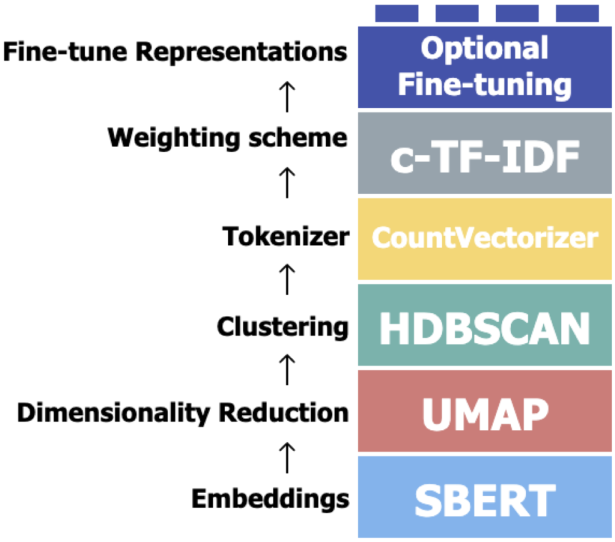


Figure 2.6: BERTopic Default Model Architecture Score

1. **SBERT** for **Embedding**: The text is firstly given a vector representation in an embedding space using Sentence-Bert (SBERT), enabling the algorithm to capture the semantics and relationships captured by the transformer-based model BERT fine-tuned for sentence similarity tasks [19].

2. **UMAP** for **Dimensionality Reduction**: To enable both an efficient and accurate modelling of topics, Uniform Manifold Approximation and Projection (UMAP) is deployed to reduce the dimensionality of the embeddings space while retaining as much information about the data. UMAP connects points close to each other in the space using edges, following which it construct a low-dimensional embedding of the data that aims to:

   (a) preserve relationship between vertices that share an edge;

(b) embody the shape of the existing surfaces on which the data lies [25].

3. **HDBSCAN** for **Clustering**: The key step to model the topics is clustering through Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), which is a hierarchical density-based technique. More specifically, HDBSCAN identifies clusters in a hierarchical framework from the reduced embeddings, and the optimal number of clusters is selected using cluster stability analysis. The latter is a nuanced subject which attempts to find the weighted sum of "Validity Index" values of clusters based on the intra-cluster density rather then distance [25].

4. **CountVectorizer** for **Tokeniser**: CountVectorizer is a technique that generates a Bag of Words (BoW) embedding representation for the documents in a text, including the pre-processing steps (see Section 2.2.1 for more details on BoW) [19].

5. **C-TF-IDF** for **Weighting Scheme**: Section 2.2.1 introduced the TF-IDF algorithm which allocates more weights to nouns, adjectives, and verbs in a document. Cluster TF-IDF builds upon this canonical version by considering the tokens identified in the previous step. The key difference is that c-TF-IDF aims to measure the importance of each token to a cluster instead of to the whole text. C-TF-IDF concatenates the documents in a cluster and measures the importance of different words for that topic. This step enables different clusters to be distinguished from one another by determining cluster-word representations [19].

6. **Fine-tuning**: This optional step incorporates smaller documents that consist of topics represented by keywords and documents and which can enhance the topic modelling ability of the algorithm. Some powerful techniques in this category include KeyBERT and KeyBERTInspired [25].

Advantages of BERTopic

1. Embedding accuracy: Since BERTopic employs a fine-tuned version of the embedding algorithm BERT, it can leverage the latter's grasp of language and context driven by the transformer architecture without having to preprocess the text, which potentially alters meaning [25].

2. Number of topics accuracy: BERTopic inherently determines the number of topics in a dataset as opposed to considering the variable as a hyperparameter [19]. This allows BERTopic to leverage the long-range dependency capture of the transformer-based embedding algorithm to appropriately identify the optimal number of topics in the data.

Disadvantages of BERTopic

1. The main disadvantage of BERTopic is the contextualised embedding step which requires the processing computer to have powerful processors such as heavy-duty GPUs [19].

2. BERTopic assigns every document to a single topic rather than a distribution of topics, even though it does not necessarily assume that the topics are independent of each other [18].

3. Since students often discuss multiple topics in their review, and this requires prior data engineering transformations for each document to ensure BERTopic's successful deployment in theory. This is explored in Section 3.3.1.

4. BERTopic produces many outliers and therefore data points which are discarded compared to other topic modelling techniques such as LDA, which is explored below [18].

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation is a generative statistical model that employs the posterior distribution iteratively to find the limits of two distributions, namely the distribution of keywords per topic and the distribution of topics per document. According to Blei et al., 2003, in a paper titled 'Latent Dirichlet Allocation' and published in the Journal of Machine Learning Research, the underlying intuition behind the algorithm is that any qualitative data point contains multiple topics with different probabilities and each topic consists of multiple key words, again with different probabilities [2]. For instance, one student review could elaborate on 'teaching' as a key strength and 'lecture recordings' as a positive side note, with the distribution of keywords for each topic consisting of the support ('teacher', 'engage', 'proactive') and ('recordings', 'accessible', 'considerate') respectively. The algorithm is part of the class of unsupervised machine learning models and works as follows.

1. State the desired number of topics, call it x.

2. Define two hyperparameters, namely:

    (a) α**:** alpha represents the parameter of the Dirichlet distribution of topics per document. The higher the value of alpha, the higher the number of dominant topics per document.

    (b) η**:** eta represents the parameters of the Dirichlet distribution of key words per topic. Likewise, a higher eta value indicates that each topic consists of more key words [2].

3. Use the preprocessing and embedding process to create a token vocabulary and embedding from the corpus. The generally accepted tokenisation and embedding techniques for LDA are word tokenisation and bag of words embedding [20].

4. Randomly assign an integer value in the range $[0, x]$ to each token. Use this random generation to determine the two initial conditional probabilities.

    (a) $Pr(word\ w \mid topic\ t)$

    (b) $Pr(topic\ t \mid document\ d)$

    This groups words and documents into the predefined number of clusters, i.e. topics.

5.
```
for $d \in document$ do
    for $w \in words$ do
        $p(word\ w\ with\ topic\ t) \leftarrow p(topic\ t \mid document\ d) \times p(word\ w \mid topic\ t)$
    end for
end for
```

Repeat steps 4 and 5 for a given number of iterations, and the resulting probabilities should converge [20].

Since the entire algorithm is dependent on the number of clusters, i.e. topics, hyperparametrised in the input, it is important to feed the correct number of topics into the algorithm. For this reason, the **coherence score** C_ν is calculated for different topic numbers using Normalised Pointwise Mutual Information (NPMI). This evaluation technique is supported by Syed and Spruit, 2017, in a paper authored by and published in the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA) [10].

$$NPMI(w, w') = \frac{\log \frac{Pr(w,w')+\epsilon}{Pr(w) \times Pr(w')}}{-\log(Pr(w, w') + \epsilon)}$$

The NPMI attempts to determine the probability that two keywords w and w' belonging to the same topic can also be found in the same document [10]. The error term ϵ prevents the model from encountering a log 0 error, in which case two keywords from the same topic have no probability of being in the same document. The NPMI is calculated for a range of topic numbers, following which the model with the highest coherence score is selected.

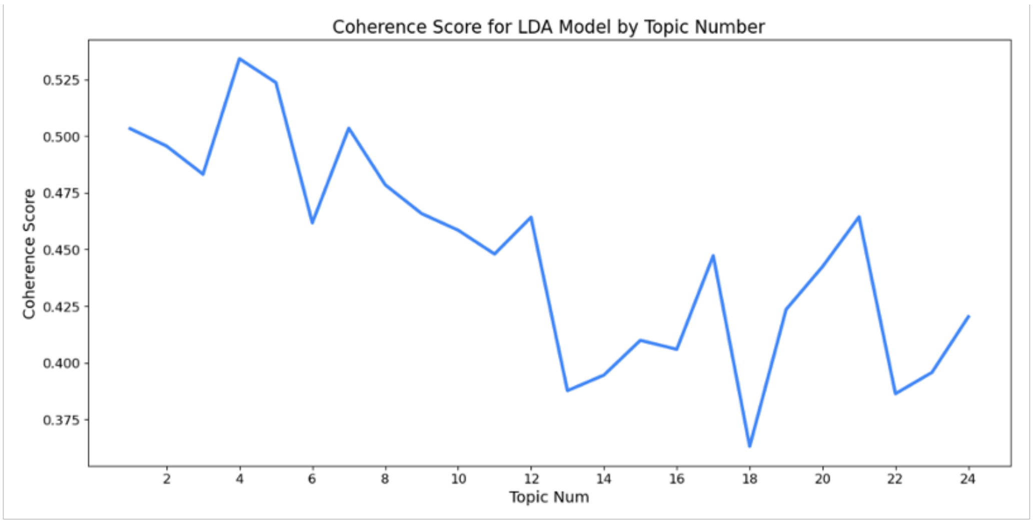


Figure 2.7: Coherence Score C_ν Graph

Advantages of LDA

1. Analytical Rigour: LDA is able to associate each document/review/data-point in the dataset to a particular topic. The latter represents the mode of the topic distribution for that document/review/data-point [2]. This enables a key stage of output validation where all the reviews associated to each topic can be verified to ensure they indeed belong to the same topic.

2. Clustering: LDA is a clustering algorithm that does not require the data to indicate the topics beforehand. This makes it appropriate to deploy on very large datasets and is especially constructive if the dataset has already been categorised into some valid structure. For instance, having already categorised the data by sentiment, clustering generates valid outputs for each sentiment category that represent the strengths or areas for improvements respectively.

3. Quantification: The algorithm generates LDA tokens and assigns each token to the different clusters. By bringing these tokens together, it is possible to calculate the percentage of LDA tokens taken by each topic. Quantifying at this stage of the thematic analysis enables different datasets to be compared to each other as part of a cross-correlation analysis.

Disadvantages of LDA

1. The main disadvantages of this model stem from the use of the BoW technique, and these have been discussed in Section 2.2.1.

2. One main assumption that is consistent but not unerring is that documents with the same topic employ similar words [20]. It is a feature that could compromise the results if a broad topic contains multiple documents where the words are not similar.

3. The hyperparameters need to be carefully selected as they can interfere with the results if they are not correctly reconciled with the dataset. For instance, as mentioned earlier, the value of alpha determines the number of dominant topics per document.

2.4.3 Extractive Text Summarisation

The goal behind Text Summarisation is to summarise a large amount of text into a smaller number of words and sentences containing key information from the larger text. Extractive text summarisation is one technique that 'extracts' key information from a larger piece of text before grouping them into a summary. Sumy and NLTK are libraries on Python that provide extractive text summarisation functionality using different methodologies [37].

Three extractive text summarisation techniques are Luhn, LexRank, and Latent Semantic Analysis. Luhn creates a sentence ranking hierarchy using a word frequency criterion, LexRank employs a graphical-based approach, and Latent Semantic Analysis employs the singular value decomposition statistical technique [37]. Since these methods are all extractive, they do not generate new text. Instead, they attempt to produce the most relevant sentences that are characteristic and illustrative of the whole text. They attempt to select the most relevant sentences, but differ in their methodology of quantifying this relevance. Amongst LexRank, Luhn, and LSA, a combination of the former two provides the widest coverage of documents in the educational text and have been selected to be part of the workflow.

Luhn

Luhn is one of the oldest techniques in the extractive text summarisation toolkit and was introduced by Luhn, 1958, in a paper entitled 'The Automatic Creation of Literature Abstracts' [1]. For each document in the text, the Luhn algorithm preprocesses the text using normalisation and stop word removal (2.1), and then deploys bag of words (BoW) to embed the sentences (2.2.1). In the third stage, words are weighted according to the TF-IDF framework (2.2.1), after which each sentence is scored using the following normalising equation:

$$Score = \frac{(Number\ of\ meaningful\ words)^2}{(Span\ of\ meaningful\ words)}$$

Finally, all the sentences are ranked in descending order of scores, i.e. in descending order of relevance [1].

LexRank

LexRank is another extractive text summarisation algorithm that requires embedding and was first introduced by Erkan et al, 2004, in a paper entitled 'LexRank: Graph-based Lexical Centrality as Salience in Text Summarization' [3]. In this technique, all the documents in the text are firstly given a bag of words representation. This gives every sentence a coordinate in the embedding space, and weighted edges are then introduced to the graph based on the cosine similarity metric as illustrated in the equation below for two vectors a and b.

$$a.b = |a||b| \cos\theta.$$

The equation helps to determine the angle between the two vectors. Two vectors pointing in the same direction are considered similar and will have a low corresponding θ value [13]. Cosine angle is favoured to euclidean distance as similarity metric since magnitude is not of pivotal importance given that documents can be of varying sizes. The term frequency (tf) adjusted cosine value is given by the following equation, where $tf_{w,a}$ represents the frequency of word w in sentence a:

$$\text{idf-modified-cosine}(x, y) = \frac{\sum_{w \in x,y} \text{tf}_{w,x}\text{tf}_{w,y}(\text{idf}_w)^2}{\sqrt{\sum_{x_i \in x}(\text{tf}_{x_i,x}\text{idf}_{x_i})^2} \times \sqrt{\sum_{y_i \in y}(\text{tf}_{y_i,y}\text{idf}_{y_i})^2}}$$

One big caveat that must be addressed at this stage is that a particular sentence could be very relevant in a given document and linked with other sentences in that same document via significantly weighted edges. However, that sentence in itself could have very little representation and correlation with other documents. It is important for the algorithm to be able to monitor such scenarios and ensure that LexRank outputs globally relevant sentences rather than sentences that have simply been magnified by local neighbours. For this reason, the algorithm incorporates the concept of eigenvector centrality [3].

To determine overall centrality, the objective is to take into account the centrality of the neighbours voting for that particular node. This ensures that nodes that are more representative of the documents will have more weight in determining the centrality of the node in question compared to other neighbouring nodes that are outliers [3]. To begin this iterative process, let every node have an arbitrary centrality value, and determine the centrality of a node $p(u)$ by the following equation:

$$p(u) = \sum_{v \in adj[u]} \frac{p(v)}{deg(v)}$$

The latter can be formulated to closely emulate the relationship between a matrix and its eigenvector, which facilitates a linear algebraic methodology to solve the algorithm instead of fixed-point iteration that would iterate until convergence. More specifically, let B represent the normalised bag of words matrix that divides every element by its row sum [3]. Then, centrality for each node is given by the eigenvector p of the matrix B^T with eigenvalue 1:

$$\mathbf{p} = \mathbf{B}^{\mathrm{T}}\mathbf{p}$$

Whilst eigenvector centrality is fundamental to LexRank, it's influence in this project is insignificant since every document considered in this project represents a single sentence. Whilst a one-to-one relationship was initially established for one document to correspond to one student review, it was later observed that a student could refer to multiple topics in one review. This made it difficult for the unsupervised topic modelling algorithms to identify the topics, and was subsequently changed to one document representing one sentence. Therefore, if a student feedback contains multiple sentences, it will be broken down into multiple sentences, each of which will be considered a single document. For this reason, the issue of local relevance becomes redundant in the grand scheme of things, and the quantification of the importance of each sentence becomes globally admissible.

2.4.4 Abstractive Text Summarisation

Abstractive text summarisation is the second type of text summarisation technique after extractive text summarisation. By definition, abstractive text summarisation 'rewrites large amounts of text by creating acceptable representations' using the model's semantic capability [37]. T5 and and GPT-3 are libraries on Python that provide text summarisation functionality using different large language models. Abstractive text summarisation is a very recent advancement in the NLP Space. It leverages the transformer architecture of these large language models to enhance understanding of the semantics and relationships in a dataset following which it generates text that aims to summarise the sets of document in question [37].

Text-to-Text Transformer T5

One prominent model used for abstractive text summarisation is the Text-to-Text Transformer T5, which is a large language model developed by Google researchers Raffel et al., 2020, in a paper entitled 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer' [14]. In Section 2.3.2, the differences between BERT and GPT-3 were explained, where it was noted that GPT-3 performs better than BERT at text generation because of it's auto-regressive training methodology. Nonetheless, T5 combines the auto-regressive training with the long-range dependency mechanism. This is because T5 is trained on a text-to-text basis, where a random sequence of words, successive words, or even sentences are masked for the model to predict. This flexibility makes it adaptable and ideal for text summarisation while GPT-3 excels at generating other text that spin off the text and are not necessarily as inherently related to it as a text summary is [34].

Nonetheless, abstractive text summarisation algorithms naturally require significant fine-tuning for text summarisation tasks, partly due to the transformer architecture that enables them to capture long-range dependencies [14]. Because the entire text is educational, the network captures dependencies between different topics within education. This results in overly brief text summaries that do not capture the nuances of educational text, distinct arguments, and opinions of students. While this therefore limits the use of abstractive text summarisation in the current workflow of the product, it is nonetheless crucial to include this concept as it has the ability to fully automate the diagnosis pipeline if developed carefully within learning analytics (5.2).

Chapter 3

Methodology

This section introduces the datasets on which the models explored in Section 2 are implemented, the order (pipeline) of model implementation, model selection based on their suitability to the pipeline and pipeline outputs, and finally the codebase and user interface that produce the results of the model implementation.

3.1 Datasets

All datasets employed in this project have been sanitised and anonymised pre-analysis. The final, canonical series of datasets associated to this product are from the National Students Survey (NSS), which contains quantitative and qualitative responses to surveys filled by students in the final year of their degree with regards to their experience. The datasets considered for this project represent feedback for an arbitrary department between the 2014/15 and 2022/23 academic years, and findings are explored for demonstration purposes only. Quantitative data from the NSS used for cross-correlation and validation purposes is only available until the 2021/22 academic year.

The NSS data for 2015-2023 enables multiple thematic analysis models developed in the previous sections to be deployed as part of a thorough review of the department. Data for all nine academic years is fed through the data pipeline to determine the recurrent and cyclical trends that are then diagnosed and prescribed from in Section 4.2.

3.2 Data Pipeline

In order to diagnose and prescribe from the student reviews, the qualitative data passes through multiple processing stages as part of the product workflow. Diagnostic analysis is then directly carried out on the final results produced by the pipeline.

1. The reviews are firstly categorised by sentiment using DistilBERT, and this allows the consequent analysis to separately consider positive reviews and negative reviews as sugggested by the students.

2. For each sentiment category, a topic modelling algorithm is deployed to find the most prominent themes for that sentiment. Implementing topic modelling at this stage of the data pipeline enables each topic's prominence in the sentiment category to be quantified. This addresses Gap 1 in Section 1.3.2 as the new tool can now process and quantify the qualitative data. The choice of topic modelling tool is between LDA and BERTopic, with the former being selected to be part of the data processing pipeline. See Section 3.3.1 for a breakdown of the outperformance of LDA compared to BERTopic in this context.

3. Under each sentiment category, extractive text summarisation models such as LexRank and Luhn are deployed to pick out the most relevant sentences for each topic. Deploying these algorithms in the final stage of the product workflow has multiple benefits. Firstly, it allows the diagnostic analysis of strengths and areas for improvements for each topic to be directly carried out on complete sentences in the form of the text summaries. This enables a comprehensive analysis of the semantics, morphology, and pragmatics of the data. Finally, as an extractive technique, the text summaries represent the students' voices in their own words, and therefore allows the diagnostic analysis to consider the nuances of human language. These benefits address Gap 2 and 3 in the existing provision (1.3.2).

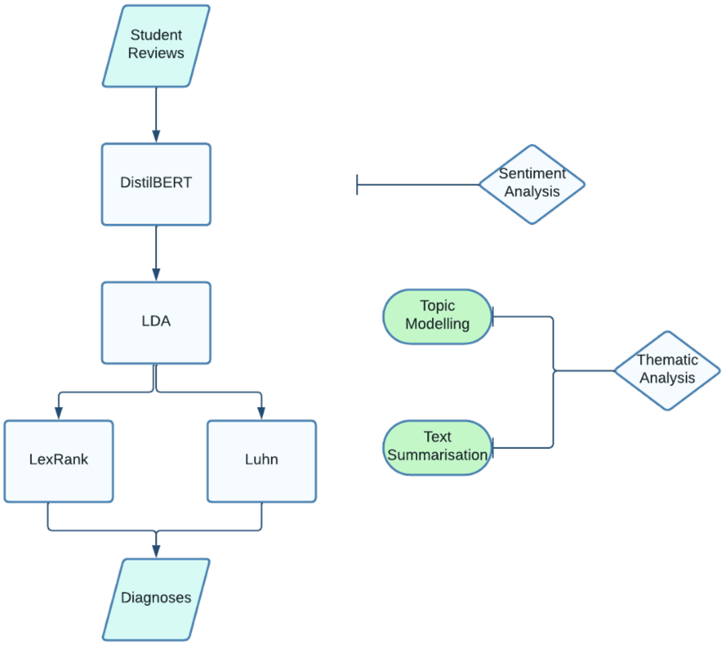


Figure 3.1: NLP processing pipeline for the proposed new product

3.3 Model Selection

This section validates the models based on their output and practical relevance in this context, whereas the theoretical evaluation of all models has already been presented in Section 2. Whilst all non-topic modelling algorithms are included in the pipeline in some capacity, one topic modelling model must be selected out of two very sophisticated techniques due to the pipeline's structure (3.2). Both LDA and BERTopic have very strong underlying principles that justify their inclusion, such as the use of coherence scores and density-based clustering validation respectively (2.4.2). Therefore, model validation in this section is output-oriented and attempts to establish associations between keywords versus reviews and term frequency versus themes.

3.3.1 Topic Modelling Algorithm

The two topic modelling algorithms, namely LDA and BERTopic, perform differently in different contexts and with different datasets. Within learning analytics, an important consideration is that the entire text is, trivially, educational. The topic modelling objective in this project therefore is to identify key topics *within* education, and not validate education as an overarching topic itself. Other key considerations that are particular to this project are that these algorithms have not been fine-tuned or adjusted from the default model selections. To compare the two models in the context of higher education, the national students survey data for 2022 and 2023 are explored herein this section. Since both sets of data are most recent and consider some key nuances, a side-by-side comparison of the performance of LDA and BERTopic for the two years enables a more thorough evaluation of their ability to model topics from educational data to ultimately select one of the two models.

In natural language processing, it is difficult yet important to design surveys in such a way so as to supply the statistical models with equitable data to provide a fair diagnosis of the input text and also enable forward-looking comments that can be used for analytics. In this vein, it is interesting to observe a change in prompt between 2022 and 2023 in which students give a qualitative review of their learning experience. While both sets of surveys provide opportunities for students to write positive and negative comments, there is a change in the final prompt asked in the survey from 'what teaching methods or activities did you find most beneficial, and why?' in 2022 to 'what is the one thing we could have done to improve your overall experience?' in 2023.

The change in prompt to a more forward-looking question enables the learners to directly provide recommendations to enhance their learning. In essence, it addresses Gap 3 found in the Problem Statement for this project, which is 'the apparent overemphasis on analytics rather than learning' (1.3.2). The new prompt enables decision-makers to obtain recommendations from the primary stakeholders themselves, i.e. students. It's forward-oriented nature further contributes to generating prescriptive analytics that focus on the learning experience of the recipients as opposed to just their grades. The implications of this change are beneficial in providing prescriptive analytics for 2023. With these considerations in mind, the topic modelling ability of both LDA and BERTopic are explored below.

2022 Themes

Positive Themes

The Abstractive Summary produced by Google's T5 Model summarises the positive comments and is used as a guide to compare the two topic modelling algorithms. According to the Abstractive Text Summarisation algorithm leveraging the T5 model:

> 'The course is very challenging and the module content is very interesting. The

> support office is very responsive and supportive. The change to pre-recorded lectures has given me more time to work on exercise sheets. The support office is very good. Student opinion is highly valued. The university was quick to adapt to the changes COVID-19 brought about. The course is really challenging and pushed me to develop my skills and find my interests.'
>
> - T5 Summariser

Thematic analysis derived from the above summary can be further deconstructed to get insights into the underlying topics that positively impacted the student's learning experience in 2022. Below are some topics that are modelled by the BERTopic algorithm.

- **Department and Learning Set-up:** There were many positive reviews on the learning set-up provided by the department. Notable positive elements of the framework include 'modules', 'support', 'lecturers', 'teaching', and 'facilities'.
- **Engaging Content:** Students found the modules available in their home department to be positively 'challenging', thereby stimulating and engaging the learners.

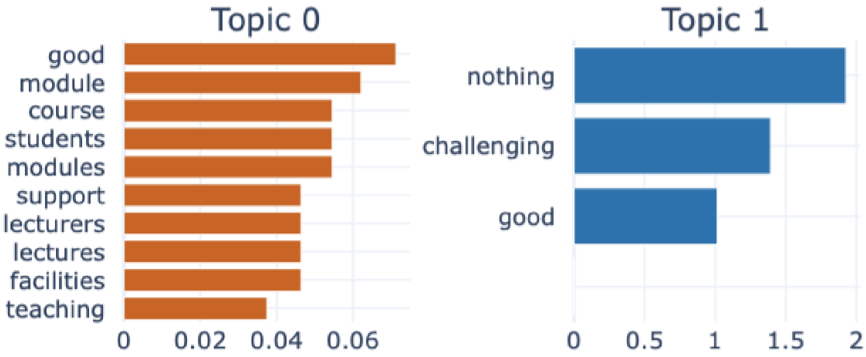


Figure 3.2: Topic Modelling Bar Chart for Positive Comments in NSS 2022 Department Data

On the other hand, LDA algorithm produces a more in-depth output that can deconstruct and validate the different themes generated by the T5 model. Below are the three positive clusters as determined by the LDA model with the highest coherence score. Since these are explored in depth in Section 4.2, they are just stated below for convenience.

1. **Community and Department (50.1%)**
2. **Teaching and Learning Set-Up (34.9%)**
3. **Course and Content (15%)**

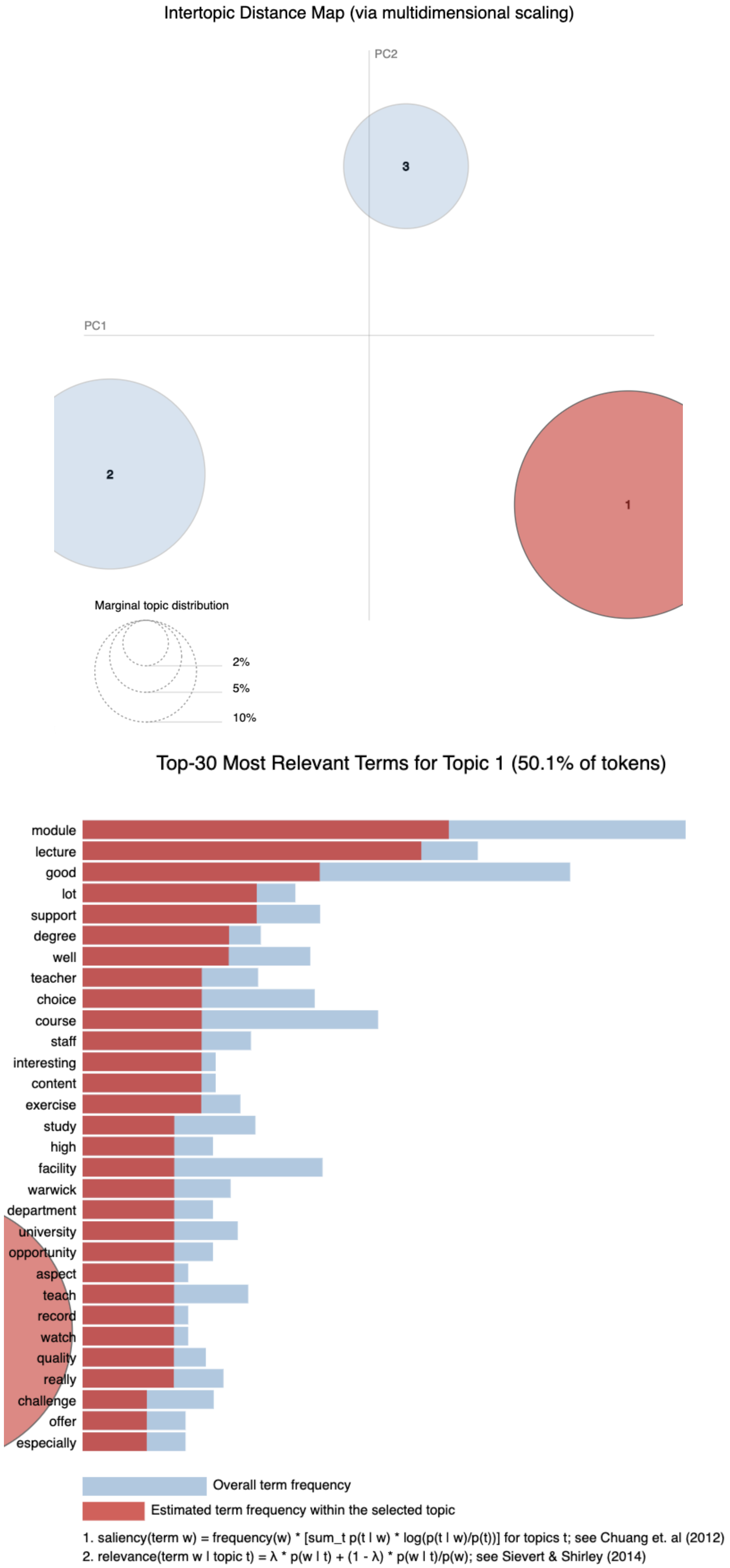
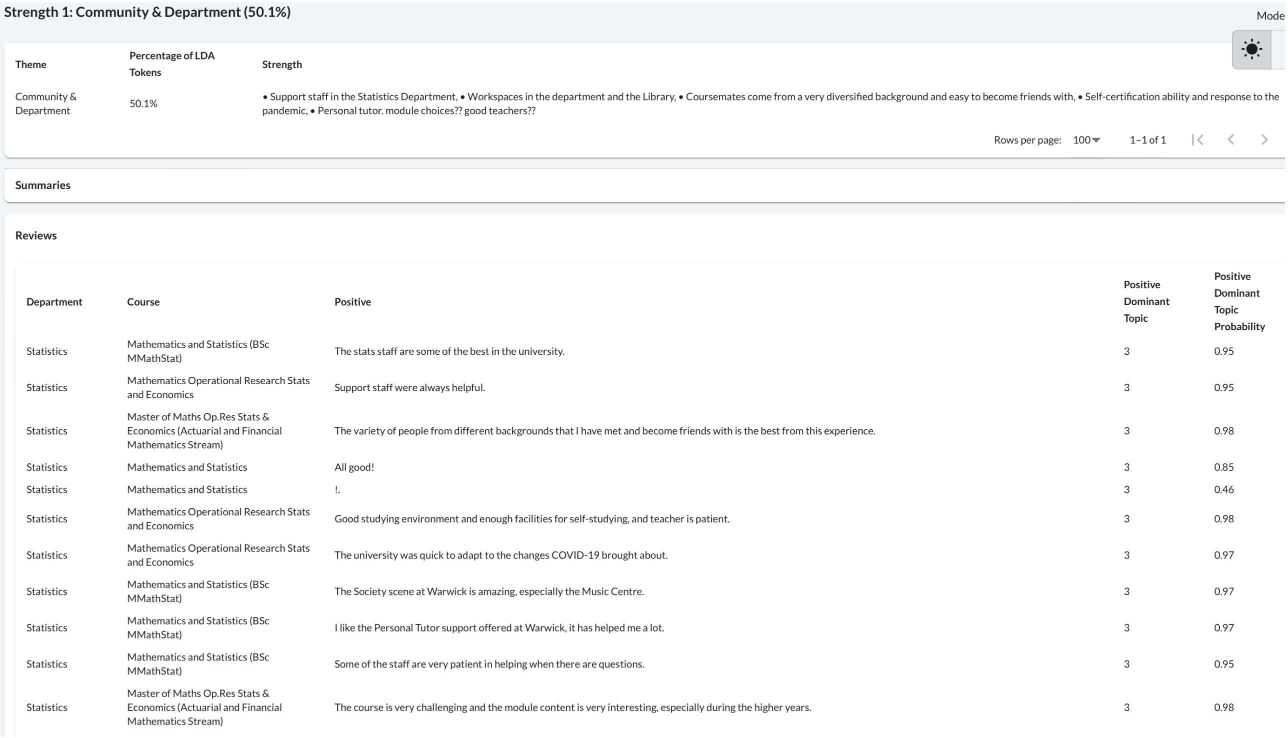


Figure 3.3: Intertopic Distance Map with Keyword Distribution for Cluster 1

In essence, each of these clusters as illustrated in Figure 3.3 can be validated organically. Since the LDA algorithm produces a distribution of topics (clusters) for each document, i.e. review sentence, the review sentences corresponding to each cluster can be proofread to ensure they represent the same or similar topics. For the first theme corresponding to 50.1% of tokens in the dataset, for instance, some of the reviews are illustrated in Figure 3.4 below.



Strength 1: Community & Department (50.1%)

Theme	Percentage of LDA Tokens	Strength
Community & Department	50.1%	• Support staff in the Statistics Department, • Workspaces in the department and the Library, • Coursemates come from a very diversified background and easy to become friends with, • Self-certification ability and response to the pandemic, • Personal tutor. module choices?? good teachers??

Summaries

Reviews

Department	Course	Positive	Positive Dominant Topic	Positive Dominant Topic Probability
Statistics	Mathematics and Statistics (BSc MMathStat)	The stats staff are some of the best in the university.	3	0.95
Statistics	Mathematics Operational Research Stats and Economics	Support staff were always helpful.	3	0.95
Statistics	Master of Maths Op.Res Stats & Economics (Actuarial and Financial Mathematics Stream)	The variety of people from different backgrounds that I have met and become friends with is the best from this experience.	3	0.98
Statistics	Mathematics and Statistics	All good!	3	0.85
Statistics	Mathematics and Statistics	!.	3	0.46
Statistics	Mathematics Operational Research Stats and Economics	Good studying environment and enough facilities for self-studying, and teacher is patient.	3	0.98
Statistics	Mathematics Operational Research Stats and Economics	The university was quick to adapt to the changes COVID-19 brought about.	3	0.97
Statistics	Mathematics and Statistics (BSc MMathStat)	The Society scene at Warwick is amazing, especially the Music Centre.	3	0.97
Statistics	Mathematics and Statistics (BSc MMathStat)	I like the Personal Tutor support offered at Warwick, it has helped me a lot.	3	0.97
Statistics	Mathematics and Statistics (BSc MMathStat)	Some of the staff are very patient in helping when there are questions.	3	0.95
Statistics	Master of Maths Op.Res Stats & Economics (Actuarial and Financial Mathematics Stream)	The course is very challenging and the module content is very interesting, especially during the higher years.	3	0.98

Figure 3.4: LDA Example Cluster and Corresponding Documents

This enables the analysis to be enriched for each cluster, i.e. explored using the keywords and validated using the corresponding reviews. For instance, Figure 3.4 above illustrates the results of the diagnosis of strengths within Community and Department for 2022 as identified by LDA. Some of the keywords identified are 'module', 'lecture', 'support', 'choice', 'facility', 'department', 'university', 'opportunity', and 'challenge'. The corresponding diagnosis is as follows.

- Support staff in the home department.
- Workspaces in the department and the library.
- Coursemates come from a very diversified background and easy to become friends with.
- Self-certification ability and response to the Pandemic.

Negative Themes

The following negative topics have been identified by BERTopic from the 2022 NSS data.

- **Online Exams and Online Teaching:** The model found a close association between negative exam reviews and the factors and preparation that lead up to it. Thematically, there is a close association between 'exams' and the terms 'lecturers', 'module', and 'department'. However, there seems to be insufficient evidence to enable BERTopic as a topic modelling algorithm to deconstruct the different factors that can be implicated and associated to poor examination comments.

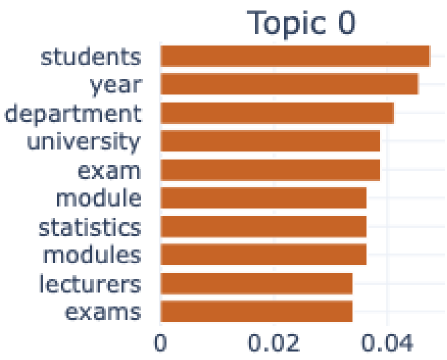


Figure 3.5: Topic Modelling Bar Chart for Negative Comments in NSS 2022 Department Data

The negative theme modelled by BERTopic depicts a very broad picture without providing a specific break-down behind the negative comments. The reason for this is two-fold.

1. The final prompt for 2022 specifically asked for teaching methods and activities that students benefited from, which provides additional positive input for the topic modelling tool to deconstruct. The final input therefore contained at least twice as many positively themed inputs as the negative comments.
2. The intertopic distance map for negative comments in Figure 3.6 illustrates an important issue. The cluster representing Topic 0 is discernibly broader and therefore much more prevalent throughout the negative comments than the two other topics, which are closer together and are likely to share common terms.

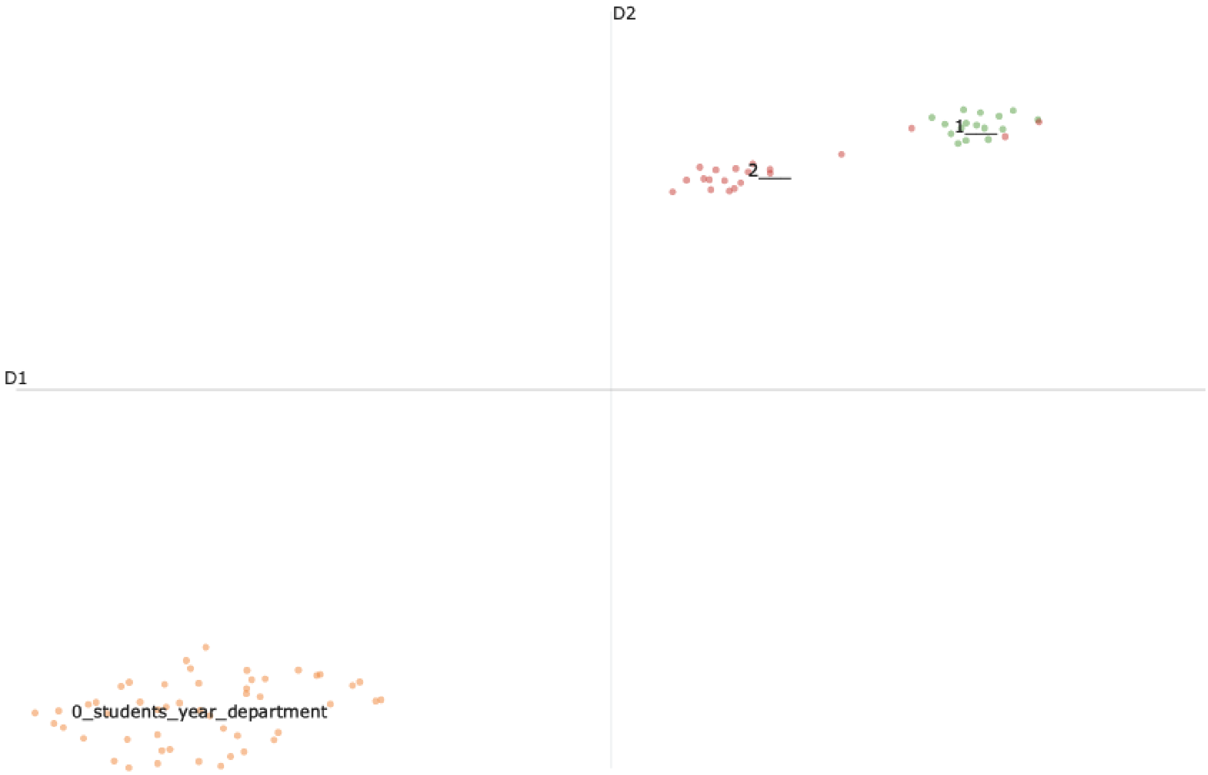
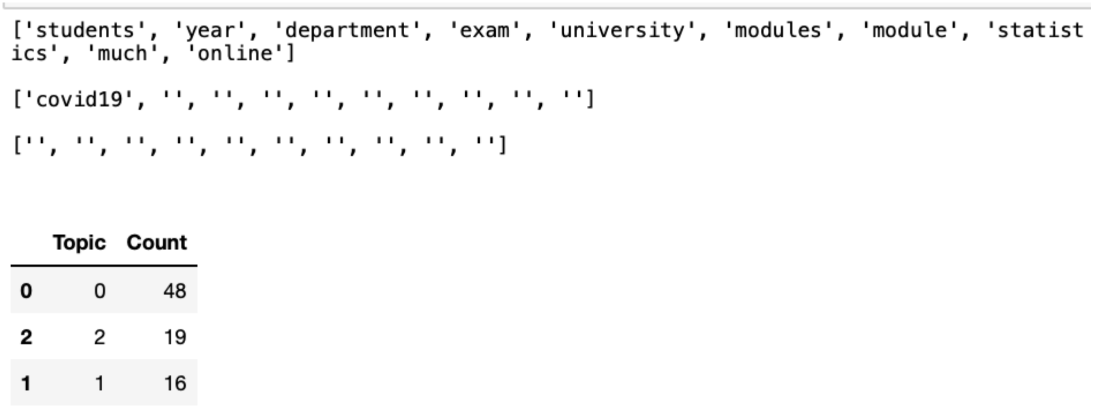


Figure 3.6: Intertopic Distance Map for Negative Comments in NSS 2022 Department Data

A closer look at the words that make up topic 1 and 2 in Figure 3.7 paints an unusual picture: other than ‘covid19’, there are no words that best represent those two topics.



```
['students', 'year', 'department', 'exam', 'university', 'modules', 'module', 'statist
ics', 'much', 'online']

['covid19', '', '', '', '', '', '', '', '', '']

['', '', '', '', '', '', '', '', '', '']
```

	Topic	Count
0	0	48
2	2	19
1	1	16

Figure 3.7: Topic Modelling Count for Negative Comments in NSS 2022 Department Data

This highlights the numerous and distinct factors that students have negatively associated with exams, leading to a similar cluster of results with fundamental differences. However, these differences are more easily captured by LDA, an inherently generative statistical modelling where the number of topics are hyperparametrised as opposed to determined by the model. The following pie chart illustrates the negative areas for improvements identified from the 2022 NSS data.

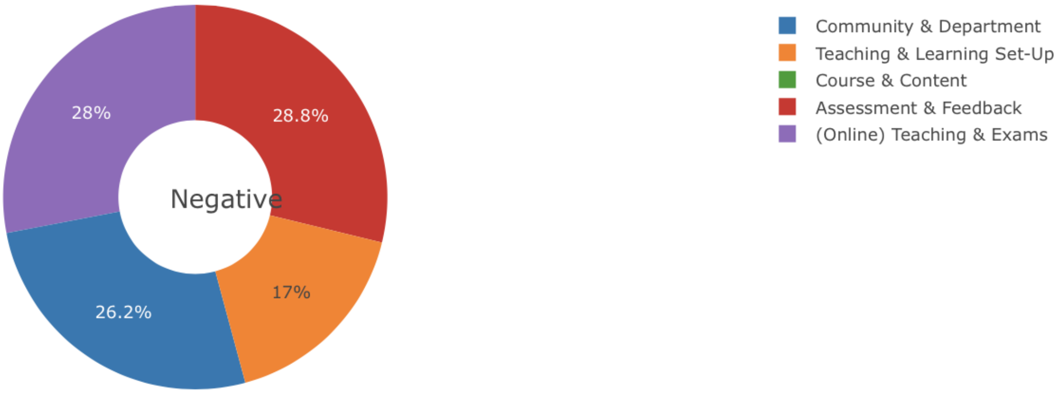


Figure 3.8: Topic Modelling Pie Chart for Negative Comments in NSS 2022 Data

To enrich the evaluation of the two models, it is therefore relevant to try to deconstruct this further. Issues pertaining to feedback that could improve learning and enhance exam performance can in fact be discerned from some extractive text summaries that quote the most prevalent issues in the negative comments. Deploying a LexRank extractive summariser in this case, four of the first five sentences that summarise the negative reviews are:

1. ‘The online, open-book exams of the home department, one year after the COVID crisis began, were extremely difficult and stressful, in contrast to exams from other departments.’
2. ‘Too much content for a module for a term, e.g., ST342.’
3. ‘Some of the modules are so hard.’
4. ‘Some of the external department lecturers were not on par to the standard of teaching set by the home Department.’

The above Centroids produced by the graphical-based summariser LexRank depict the distinct nature of complaints which ultimately lead to the unsatisfactory exam reviews. These are captured more distinctly by LDA when compared to BERTopic, as illustrated in Figure 3.8.

Given the output availability of numerous models for the 2021/22 academic year in this section, it is also very relevant to comment on the strong criticisms about exams during the Covid-19 Pandemic in this section. The latter is a transient theme that does not appear as consistently in the following years as exams in the home department have mostly returned to an in-person format. On the other hand, Section 4.2 deals with recurrent and cyclical diagnostics that are relevant today and can be prescribed from to make tangible improvements. For this reason, the topic is deconstructed in this section.

In essence, Covid-19 exam arrangement criticisms feature in the abstractive summary, the extractive summary, and in the topic modelling output. Exams during the 2021-2022 academic year were held online, which created numerous dilemmas in the questions asked by examiners. More specifically, questions can generally be classified as ‘bookwork’, ‘seen’, ‘adaptive bookwork’, ‘seen similar’, and ‘unseen’. The first two categories enable students to replicate content learned via class materials and questions previously covered in lectures or seminars. Nonetheless, with students now having access to notes when doing their exams remotely, these questions could no longer be asked and resulted in a significant increase in the generally harder style of question categories 'adaptive bookwork', 'seen similar', and 'unseen'. It is important therefore to consider such external circumstances leading to these diagnoses when inferring from them for the future.

2023 Themes

Google's T5 generates the following two summaries for the positive and negative student reviews in the 2022/23 NSS data.

> 'I've had some outstanding lecturers and found the university to be very fulfilling. The course has pushed me to grow and learn more independently than I could ever imagine. I've found the home department to be very good with a wide range of modules to choose from. The library is well stocked with books and many are accessible online.'
>
> - T5 Summariser on the Positive Reviews

> 'I have suffered academically during COVID. I would allow students from Year 2 to have the choice to choose modules. I would not allow students from year 1 to have the choice to choose modules. I've had a lot of coursework that was not marked objectively. I've had a lot of lecturers that couldn't explain things well enough. I would not go to my personal tutor for help or advice unless I had to. I've had a lot of problems with my life and my career.'
>
> - T5 Summariser on the Negative Reviews

BERTopic identifies the following cluster in this dataset from the positive sentiment category.

- **Course:** There are many positive reviews about the teaching set-up, department standards, and community in the home department. More specifically, there is a close association between the terms 'modules', 'teaching', and 'lecturers' with 'good', 'well', 'helpful', and 'support', which gives a comprehensive insight into the positive learning highlights summarised by T5.

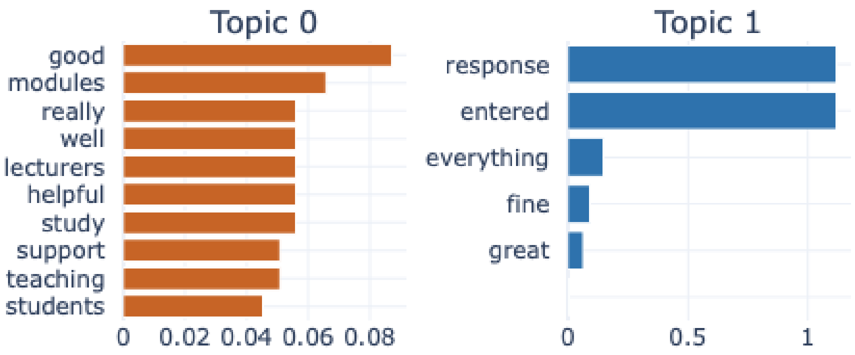


Figure 3.9: Topic Modelling Bar Chart for Positive Comments in NSS 2023 Department Data

BERTopic also identifies the following two clusters from the negative sentiment category.

- **Assessment Feedback:** Students had higher expectations with regards to the feedback they received. There is a close association between elements of the degrees, namely 'courses', 'modules', and 'department' with assessments and comments, namely 'could', 'feedback', and 'exam'. This association extends one of the key comments about marking in the T5 Summary above.

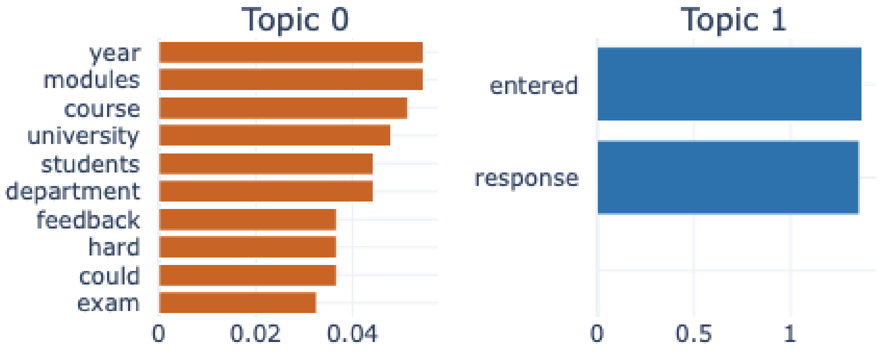


Figure 3.10: Topic Modelling Bar Chart for Negative Comments in NSS 2023 Department Data

- **Health and Well-being:** 'Students' is employed most frequently in answers to the prompt 'what is the one thing we could have done to improve your overall experience?', and this is closely followed by the terms 'support', 'health', 'mental'.

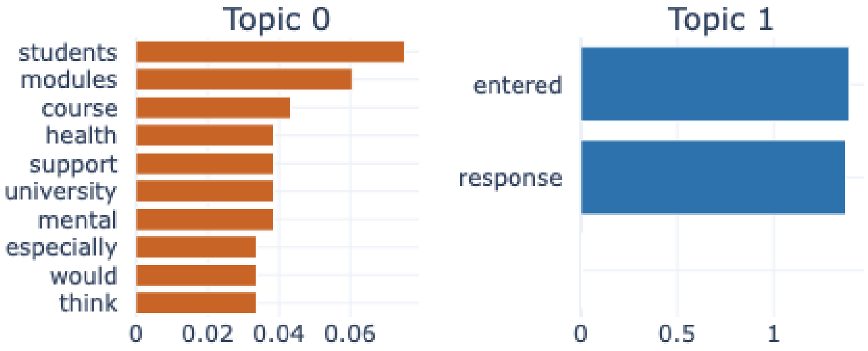


Figure 3.11: Topic Modelling Bar Chart for Improvement Comments in NSS 2022 Department Data

On the other hand, LDA provides a clearer indication of the most conducive factors to the positive learning experience and associates more closely with the summaries generated by Google's T5. Interestingly, LDA is able to identify more clusters for each sentiment than BERTopic can identify in aggregation, as illustrated in Figure 3.12.

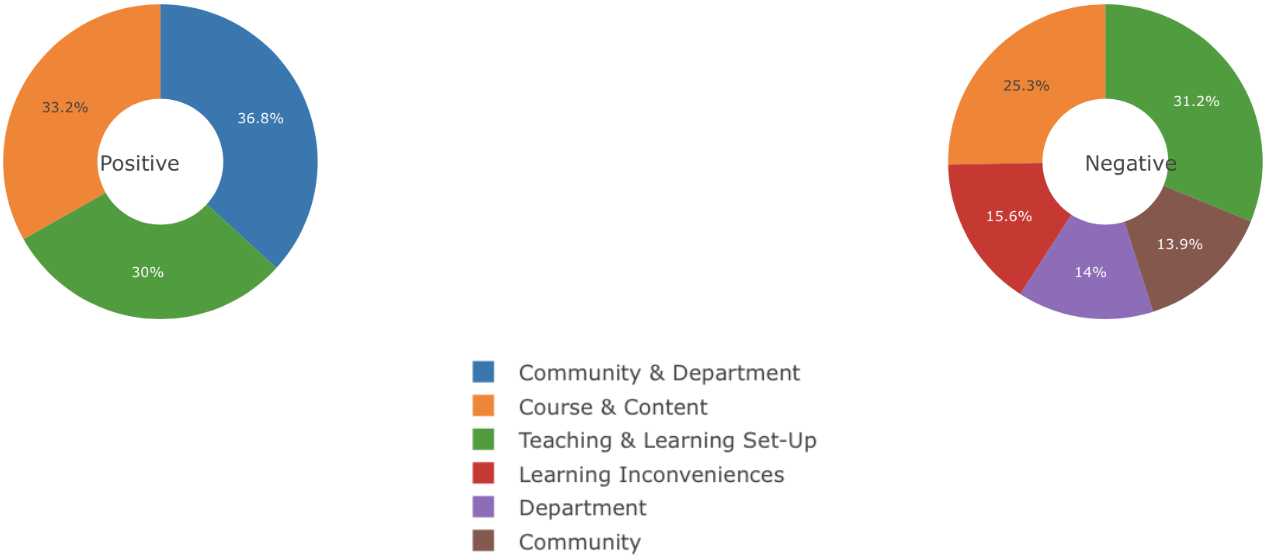


Figure 3.12: Percentage of Reviews (LDA Tokens) allocated to each topic in the 2022/23 NSS Data

The three strengths and five areas for improvements identified by the LDA algorithm are explored in depth and analysed to produce diagnostics and prescriptions in Section 4.2. They nonetheless highlight the discrepancy between the two models in their default form for this section, where LDA outperforms BERTopic in both sentiment categories for the 2021/22 and 2022/23 academic years. Therefore, LDA is selected over BERTopic as the go-to topic modelling algorithm for this project. In summary, there are two key characteristics of the two models that have enabled this deduction.

1. **Self-attention mechanism of BERT:** BERT excels at capturing long-range dependencies in a text, and naturally struggles therefore to distinguish between subtopics within education. Instead, it is able to find relationships between different subtopics such as teaching and department that are inherently true but not useful at this stage. For this reason, BERTopic identifies a smaller number of justifiable clusters during topic modelling than LDA.

2. **Hyper-parametrisation of number of topics in LDA:** Since LDA treats the number of topics as an independent variable compared to BERTopic, this helps to identify nuances of language and distinction in topics that can categorise all the education-related clusters further. LDA can also evaluate coherence score for a different number of topics to maintain the accuracy of the output and ensure an appropriate selection of the number of topics hyperparameter.

3.4 Codebase

To attain the objectives outlined in Section 1.4, the final product created for this project has the following product code-base structure. This structure serves as reference when explaining deployment of the different NLP models in the aforementioned chapters.

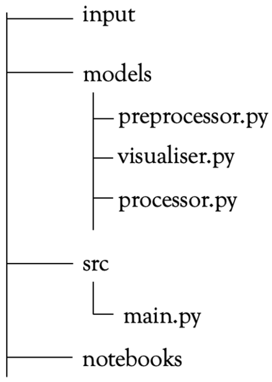


Figure 3.13: Code-base structure for the proposed new product

The folders in the diagram have been created to store the following information:

- input: contains all input files and data;
- models: python package that stores all modules, i.e. deployed tools and their functionality;
- src: stores all python scripts associated to the project;
- notebooks: stores Jupyter Notebooks used for explorative purposes.

The table below illustrates the existing Python Libraries that have been deployed at various stages of the product pipeline from the models package.

Table 3.1: Python Libraries and their Functionality

Task	Python Library
Data Preprocessing	Natural Language ToolKit (NLTK) Spacy
Sentiment Analysis	Transformers
Topic Modelling	Gensim
Extractive Text Summarisation	Sumy
Full-stack Data Application Builder	Taipy

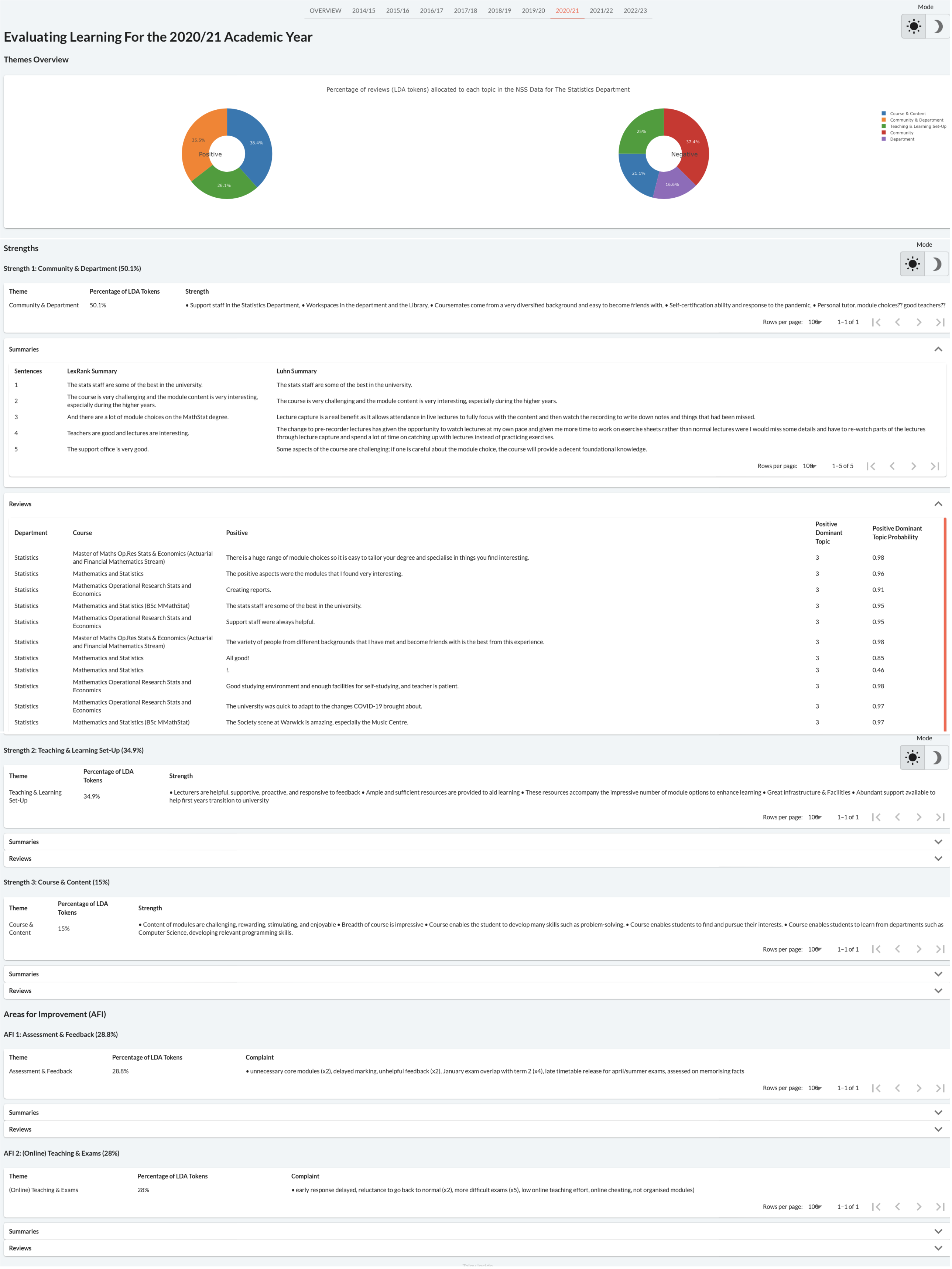
Whilst there are many deep learning Python packages available, Hugging Face has developed a Transformers library in Python that provides tools to easily deploy pre-trained models such as BERT, GPT, distilled versions of BERT and GPT, as well as both TensorFlow and PyTorch models. Hence, this is the go-to library employed for deep-learning related explorations in this project.

3.5 Data Dashboard User Interface

The user interface contains two main visual categories of pages, namely the overview and yearly categories.

3.5.1 Yearly Evaluation Category

The yearly evaluation category consists of pages, each of which work with one year's dataset of qualitative reviews. The page brings together the NSS feedback for an arbitrary and example department during that year and engineers the dataset such that each sentence now corresponds to one data point. This firstly allows the reviews to be categorised by positive and negative sentiment, which forms the basis for the strengths and areas for improvements respectively. Furthermore, for each sentiment category, topic modelling helps to identify the topics in the strengths dataset and areas for improvement dataset. Next, for each topic, text summarisation techniques output the most relevant sentences for each topic, which helps to unveil and identify the underlying themes for each topic. This produces the structure illustrated in the following figure.



Evaluating Learning For the 2020/21 Academic Year

Themes Overview

Strengths

Strength 1: Community & Department (50.1%)

Theme	Percentage of LDA Tokens	Strength
Community & Department	50.1%	• Support staff in the Statistics Department, • Workspaces in the department and the Library, • Coursemates come from a very diversified background and easy to become friends with, • Self-certification ability and response to the pandemic, • Personal tutor. module choices?? good teachers??

Rows per page: 100 1–1 of 1

Summaries

Sentences	LexRank Summary	Luhn Summary
1	The stats staff are some of the best in the university.	The stats staff are some of the best in the university.
2	The course is very challenging and the module content is very interesting, especially during the higher years.	The course is very challenging and the module content is very interesting, especially during the higher years.
3	And there are a lot of module choices on the MathStat degree.	Lecture capture is a real benefit as it allows attendance in live lectures to fully focus with the content and then watch the recording to write down notes and things that had been missed.
4	Teachers are good and lectures are interesting.	The change to pre-recorder lectures has given the opportunity to watch lectures at my own pace and given me more time to work on exercise sheets rather than normal lectures were I would miss some details and have to re-watch parts of the lectures through lecture capture and spend a lot of time on catching up with lectures instead of practicing exercises.
5	The support office is very good.	Some aspects of the course are challenging; if one is careful about the module choice, the course will provide a decent foundational knowledge.

Rows per page: 100 1–5 of 5

Reviews

Department	Course	Positive	Positive Dominant Topic	Positive Dominant Topic Probability
Statistics	Master of Maths Op.Res Stats & Economics (Actuarial and Financial Mathematics Stream)	There is a huge range of module choices so it is easy to tailor your degree and specialise in things you find interesting.	3	0.98
Statistics	Mathematics and Statistics	The positive aspects were the modules that I found very interesting.	3	0.96
Statistics	Mathematics Operational Research Stats and Economics	Creating reports.	3	0.91
Statistics	Mathematics and Statistics (BSc MMathStat)	The stats staff are some of the best in the university.	3	0.95
Statistics	Mathematics Operational Research Stats and Economics	Support staff were always helpful.	3	0.95
Statistics	Master of Maths Op.Res Stats & Economics (Actuarial and Financial Mathematics Stream)	The variety of people from different backgrounds that I have met and become friends with is the best from this experience.	3	0.98
Statistics	Mathematics and Statistics	All good!	3	0.85
Statistics	Mathematics and Statistics	!.	3	0.46
Statistics	Mathematics Operational Research Stats and Economics	Good studying environment and enough facilities for self-studying, and teacher is patient.	3	0.98
Statistics	Mathematics Operational Research Stats and Economics	The university was quick to adapt to the changes COVID-19 brought about.	3	0.97
Statistics	Mathematics and Statistics (BSc MMathStat)	The Society scene at Warwick is amazing, especially the Music Centre.	3	0.97

Strength 2: Teaching & Learning Set-Up (34.9%)

Theme	Percentage of LDA Tokens	Strength
Teaching & Learning Set-Up	34.9%	• Lecturers are helpful, supportive, proactive, and responsive to feedback • Ample and sufficient resources are provided to aid learning • These resources accompany the impressive number of module options to enhance learning • Great infrastructure & Facilities • Abundant support available to help first years transition to university

Rows per page: 100 1–1 of 1

Summaries

Reviews

Strength 3: Course & Content (15%)

Theme	Percentage of LDA Tokens	Strength
Course & Content	15%	• Content of modules are challenging, rewarding, stimulating, and enjoyable • Breadth of course is impressive • Course enables the student to develop many skills such as problem-solving. • Course enables students to find and pursue their interests. • Course enables students to learn from departments such as Computer Science, developing relevant programming skills.

Rows per page: 100 1–1 of 1

Summaries

Reviews

Areas for Improvement (AFI)

AFI 1: Assessment & Feedback (28.8%)

Theme	Percentage of LDA Tokens	Complaint
Assessment & Feedback	28.8%	• unnecessary core modules (x2), delayed marking, unhelpful feedback (x2), January exam overlap with term 2 (x4), late timetable release for april/summer exams, assessed on memorising facts

Rows per page: 100 1–1 of 1

Summaries

Reviews

AFI 2: (Online) Teaching & Exams (28%)

Theme	Percentage of LDA Tokens	Complaint
(Online) Teaching & Exams	28%	• early response delayed, reluctance to go back to normal (x2), more difficult exams (x5), low online teaching effort, online cheating, not organised modules)

Rows per page: 100 1–1 of 1

Summaries

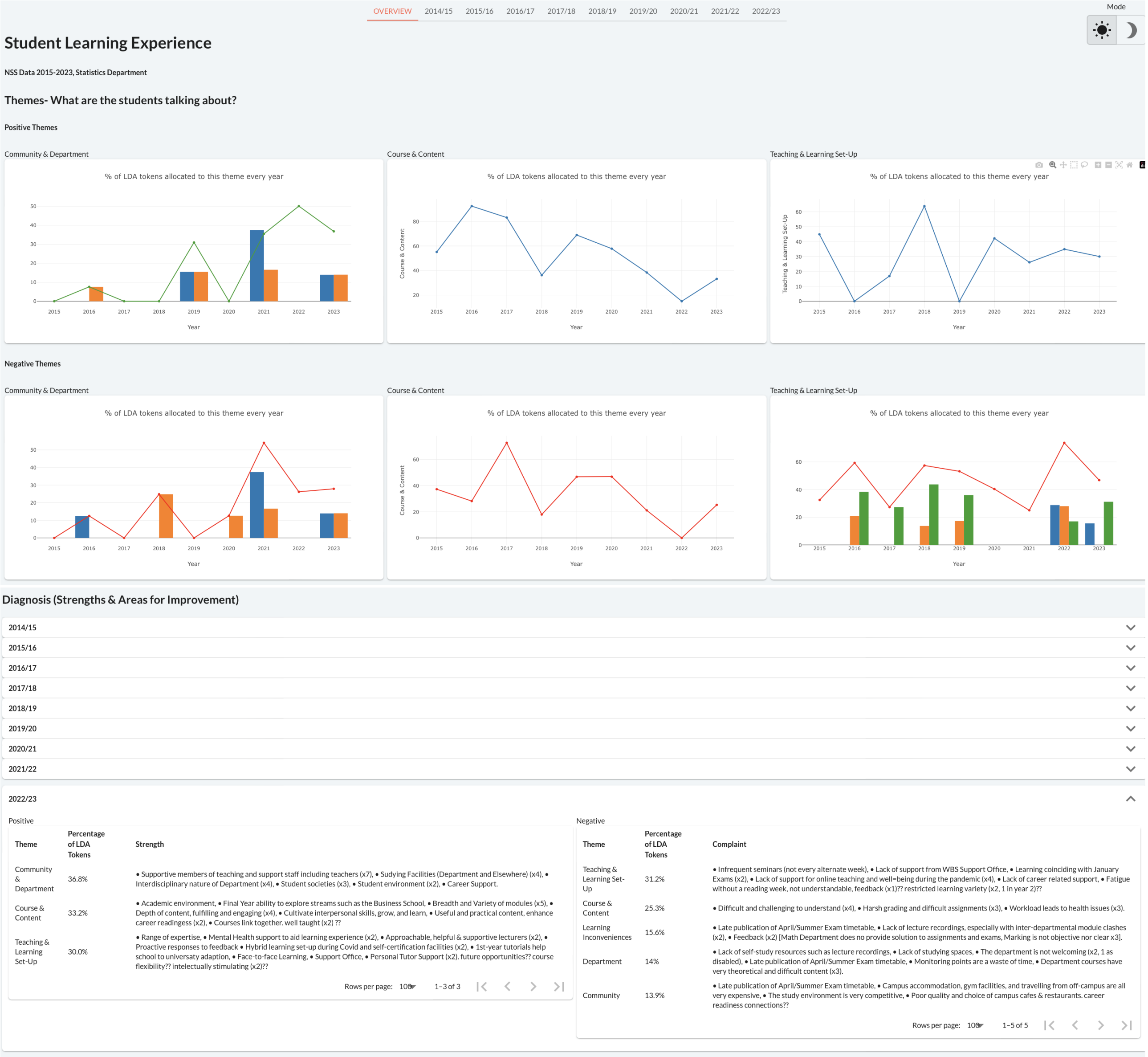
Reviews

Figure 3.14: Illustration of Yearly Page

This category therefore contains nine different pages, each of which identify the data points, sentiments, topics, and themes for one dataset. It enables an easy, accurate, and considerate diagnosis. The top of the page shows the distribution of strengths and areas for improvements for that year, following which each strength and area for improvement is presented alongside its diagnosis, extractive text summaries, and corresponding data points. Changing the dataset will automatically provide new clusters of topics, text summaries, and corresponding data points.

3.5.2 Overview Category

The overview category contains one page that brings together information from the yearly evaluation category. More specifically, it employs the quantified representations of the topics to plot the relevance of every topic over the nine years. It also brings together the diagnoses for every year in tabular form to enable a longitudinal comparative study.



Student Learning Experience

NSS Data 2015-2023, Statistics Department

Themes- What are the students talking about?

Positive Themes

Community & Department

Course & Content

Teaching & Learning Set-Up

Negative Themes

Community & Department

Course & Content

Teaching & Learning Set-Up

Diagnosis (Strengths & Areas for Improvement)

2014/15

2015/16

2016/17

2017/18

2018/19

2019/20

2020/21

2021/22

2022/23

Positive

Theme	Percentage of LDA Tokens	Strength
Community & Department	36.8%	• Supportive members of teaching and support staff including teachers (x7), • Sudying Facilities (Department and Elsewhere) (x4), • Interdisciplinary nature of Department (x4), • Student societies (x3), • Student environment (x2), • Career Support.
Course & Content	33.2%	• Academic environment, • Final Year ability to explore streams such as the Business School, • Breadth and Variety of modules (x5), • Depth of content, fulfilling and engaging (x4), • Cultivate interpersonal skills, grow, and learn, • Useful and practical content, enhance career readingess (x2), • Courses link together. well taught (x2) ??
Teaching & Learning Set-Up	30.0%	• Range of expertise, • Mental Health support to aid learning experience (x2), • Approachable, helpful & supportive lecturers (x2), • Proactive responses to feedback • Hybrid learning set-up during Covid and self-certification facilities (x2), • 1st-year tutorials help school to universaty adaption, • Face-to-face Learning, • Support Office, • Personal Tutor Support (x2). future opportunities?? course flexibility?? intelectually stimulating (x2)??

Rows per page: 100 1–3 of 3

Negative

Theme	Percentage of LDA Tokens	Complaint
Teaching & Learning Set-Up	31.2%	• Infrequent seminars (not every alternate week), • Lack of support from WBS Support Office, • Learning coinciding with January Exams (x2), • Lack of support for online teaching and well=being during the pandemic (x4), • Lack of career related support, • Fatigue without a reading week, not understandable, feedback (x1)?? restricted learning variety (x2, 1 in year 2)??
Course & Content	25.3%	• Difficult and challenging to understand (x4), • Harsh grading and difficult assignments (x3), • Workload leads to health issues (x3).
Learning Inconveniences	15.6%	• Late publication of April/Summer Exam timetable, • Lack of lecture recordings, especially with inter-departmental module clashes (x2), • Feedback (x2) [Math Department does no provide solution to assignments and exams, Marking is not objective nor clear x3].
Department	14%	• Lack of self-study resources such as lecture recordings, • Lack of studying spaces, • The department is not welcoming (x2, 1 as disabled), • Late publication of April/Summer Exam timetable, • Monitoring points are a waste of time, • Department courses have very theoretical and difficult content (x3).
Community	13.9%	• Late publication of April/Summer Exam timetable, • Campus accommodation, gym facilities, and travelling from off-campus are all very expensive, • The study environment is very competitive, • Poor quality and choice of campus cafes & restaurants. career readiness connections??

Rows per page: 100 1–5 of 5

Figure 3.15: Illustration of Overview Page

Chapter 4

Results and Analysis

4.1 Output Validation

In the learning analytics space, one of the most significant gaps discovered in the initial phases of this project was the absence of models that had been deployed to analyse qualitative data (1.3.2). It is very clear that NLP-oriented qualitative models have not yet been extensively fine-tuned and trained for the purposes of learning analytics. For this reason, there is a need to substantiate the diagnoses and prescriptive analytics produced by the NLP tools deployed in this project. The key validation techniques employed are two-fold, namely:

1. cross-correlate and compare the quantitative data across multiple years with the qualitative feedback collected to validate the findings from an overarching standpoint;
2. establish associations between:
   (a) nouns and adjectives (2.4.1);
   (b) keywords and reviews (3.3.1);
   (c) word frequency graphs and themes (3.3.1).

Whilst the second validation technique has already been explored in the aforementioned Sections 2.4.1, 2.4.3, and 3.3.1, the first technique is deployed herein Chapter 4 to deconstruct the diagnostics and prescriptive outputs on a given data science related department.

4.2 Findings

Section 3.3.1 illustrated the workflow behind the data pipeline through which the sentiment analysis and thematic analysis models were deployed on the data to retrieve a diagnosis. The data application product built in Section 3.2 and 3.5 implements this pipeline for all nine different years of national students survey data to create the diagnostic analysis. This section identifies the latest emerging and recurrent trends from the product and attempts to categorise them into actionable topics that can be addressed directly. Prescriptions are given for diagnostics that refer to areas for improvements, and the following list highlights some important limitations of the findings to be taken into consideration alongside the results.

1. Not all students fill in the national students survey. Data from the NSS indicate a 70% response rate in 2023 [26]. Therefore, if a topic is spoken about in 30% of the dataset, for instance, this is not equivalent to 30% of the students.
2. If something is spoken about less frequently in one particular year, this does not necessarily mean that it is less relevant to students. It could simply be that other topics impacted students more in that particular year. Therefore, if a theme is not picked out by the topic modelling and text summarisation models in a subsequent year, it should not necessarily be regarded as a transient trend.
3. The prescriptions represent proposals founded on the diagnostic analysis. Before implementing any changes, it is recommended to carry out further investigations and discussions with the stakeholders.

The actionable findings are categorised into recurrent trends and emerging trends. Recurrent trends refer to topics and subsequently themes in particular that have been identified by the LDA topic modelling algorithm in at least four of the nine years. Emerging trends have gained prominence in recent years for particular reasons. If these trends can be justified even if the reasons are not fully identified, they are considered emerging and must be accounted for.

4.2.1 Recurrent Trends

Teaching and Learning Set-Up

Throughout the nine years, the only category that is mentioned in at least 20% of reviews is Teaching and Learning Set-Up as an area for improvement.

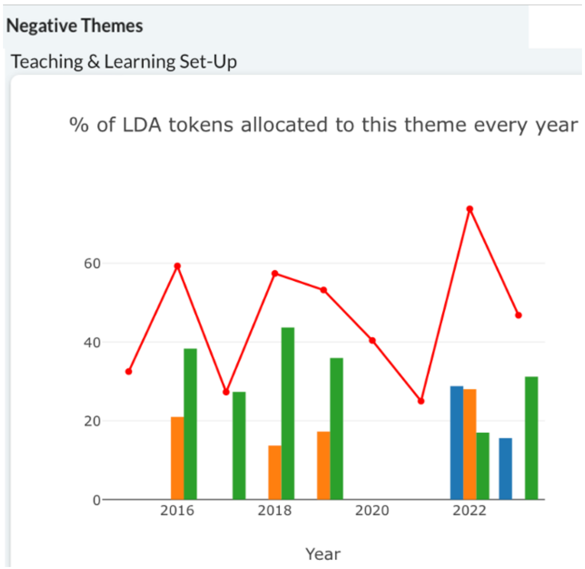


Figure 4.1: Percentage of Negative LDA Tokens allocated to Teaching and Learning Set-Up

This is evident in Figure 4.1, which illustrates the percentage of negative LDA tokens allocated to teaching and learning set-up over the years. This includes subtopics such as assessment and feedback (blue bar), teaching and exams (orange bar), and learning set-up (green bar). More specifically, according to the NLP workflow, the following five complaints feature prominently in at least four different years worth of data.

1. **Insufficient descriptions of module motivation/applicability:** A recurrent diagnosis that has been picked out by the NLP workflow is the lack of guidance provided by lecturers when it comes to applicability of theory. Students claim that it is only towards the end of their degree that they begin to understand how concepts come together and how valuable and relevant they are in context. Since some modules are very abstract and theoretical, it would give students a better sense of direction and ability to follow the class if lecturers spent teaching time reinforcing the motivations of the theory and their applicability and importance.

   Prescription: For each module, allocate time to motivate the content and its relation to other concepts in the wider context of the home department at the start of term. Moreover, allocate a specific teaching time that recurs every week or fortnight to go through the applicability of the theory. Prescriptive analytics for this diagnosis are expanded upon in the Career Readiness diagnosis under Course and Content (4.2.1).

2. **Lack of access to recorded lectures:** Some modules do not provide access to recorded lectures, at least when it comes to exam revision if not during term time.

   Prescription: provide access to recorded lectures, at least for exam preparation purposes. The benefits of lecture recordings outweigh the negatives, according to the journal article 'The use of lecture recordings in higher education: A review of institutional, student, and lecturer issues' by O'Callaghan et al., 2017 [9].

3. **Lecturers who read off slides:** Certain lectures are very one-dimensional, where lecturers simply read off power-point slides and do not engage with students. Some students also find that this gives them insufficient time to take down notes.

   Prescription: Create a pair review system where other lecturers attend, monitor, moderate, and give feedback to lecturers teaching a module. This will help to create a baseline consistent standard that is both acceptable to students and practical for lecturers.

4. **Poor inter-departmental communication:** Some consequences of that are:

   (a) timetable clashes,

   (b) large gap between lectures on the same day,

   (c) some students from the home department feel less supported when taking modules from other departments compared to the students for which it is their home department.

5. **Lack of access to online/typesetted notes:** Notes are not always available online and beforehand, especially for certain level 3 and 4 modules. This does not accommodate for different learning styles, namely:

(a) students who understand more when following lectures as opposed to taking notes simultaneously,

(b) students who want to read notes beforehand to prepare ahead of lectures.

Prescription: Provide notes or guidance to particular sections of a textbook on the reading list before covering a particular topic in the lectures.

The main common factor in the five aforementioned diagnoses is the different learning styles of students, some of who may struggle with existing teaching practices in a course. Employing the same teaching technique for all students is not user-oriented and can limit learning, as explored by Kalkan, 2011, in a paper entitled 'Recent trends at higher education emphasising active cooperative learning methods involving individual learning styles' [5]. Moreover, given the developments in cognitive and educational psychology, teaching methods in general can be improved by incorporating individual learning styles to enhance the learning experience. Whilst it is difficult to teach the same content multiple times using different methodologies, a more feasible solution is to make the teaching style as considerate and inclusive as possible. The above prescriptive analytics therefore result from an attempt to make the teaching more considerate to students who may learn better under such alternative frameworks.

These diagnoses can be further contextualised when exploring the ratings allocated to certain teaching elements by students in the NSS data.

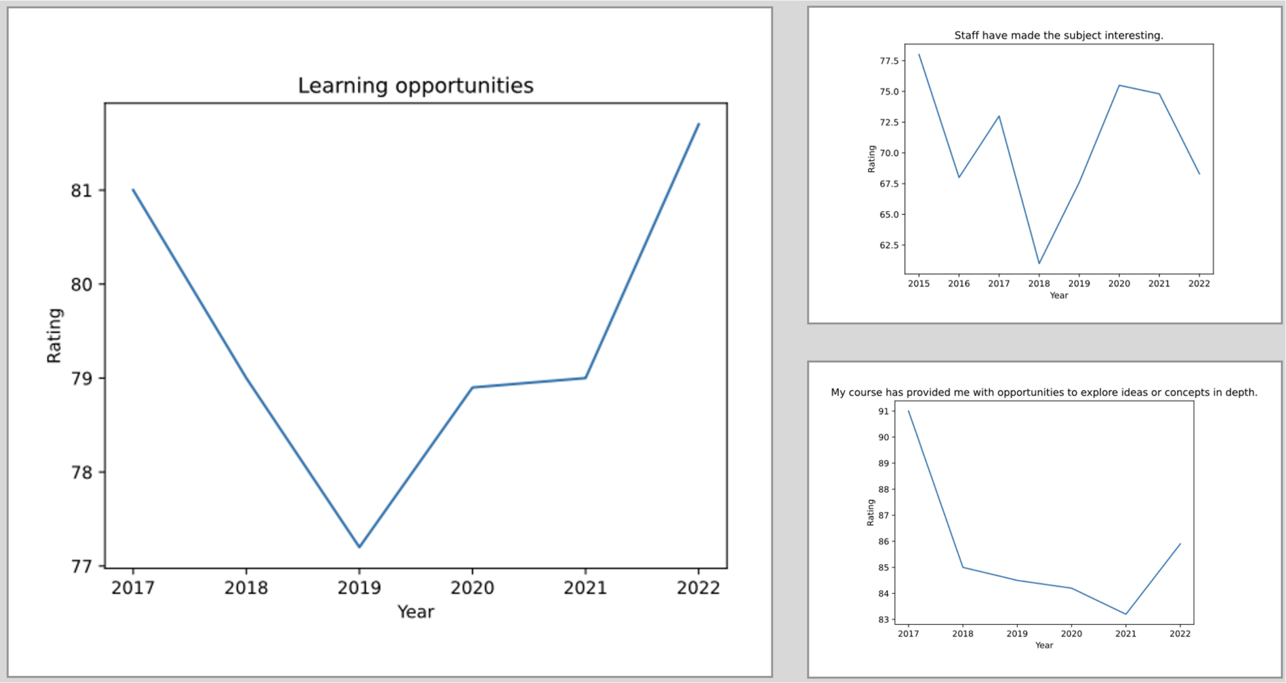


Figure 4.2: Line graphs illustrating the ratings given to learning opportunities components by students from 2016/17 to 2021/22

Despite overarching feedback on learning opportunities illustrating an increasing trend since 2018/19, it is interesting to observe that one of its components, namely the opportunity to explore ideas or concepts in depth, has actually floated around the same 85th percentile during that period. Moreover, the fluctuations between the 62nd and 75th percentile for the 'staff have made the subject interesting' learning factor indicates the absence of a consistent teaching framework guideline that is accepted by both students and teachers which can enhance learning opportunities further. The aforementioned diagnostic outputs aim to reconcile this stagnancy by:

1. giving students a greater sense of direction by increasing module motivating/applicability within classes, especially from theoretically-heavy modules;
2. enriching the quality of lectures and after-class study materials to provide a more stable and actualising framework to explore ideas in depth and enhance overall learning opportunities further.

There are also certain positive elements that recur down the years.

1. **Lecturers:** The lecturers in the home department are very well-regarded, especially when compared to other departments. Students say that the lecturers are engaged, enthusiastic, approachable, helpful, proactive, and talented. Whilst always dedicated, students think that there are areas for improvement in the teaching delivery, and these are expanded upon earlier in Section 4.2.1.
2. **Learning Resources:** Ample and sufficient resources are provided to aid learning in the form of organised and self-contained notes and videos where applicable.
3. **Tutorials:** Tutorials help students make the transition from school to university, provide an opportunity for support and for students to ask questions.

Community and Department

Whilst the graphs for Community and Department take up a very small percentage of LDA tokens between the years 2015-2020, increased emphasis is placed on this theme between 2021 and 2023. The only exception is the year 2017/18 in the negative category and 2018/19 in the positive category. The 2017/18 data point represents the year in which the home department's new building was being constructed whilst 2018/19 represents the year in which this new building was inaugurated.

Within this topic, the recurrent strengths diagnoses are three-fold.

1. **Abundance of study spaces:** A lack of study space is one of the topics picked out by the text summarisation models within Community and Department in 2015/16 and 2017/18. Following the inauguration of the department's home building in 2018/19, the abundance of study spaces and ability to work and network with course-mates is complemented in all the following years. Availability of study spaces is therefore a consideration that must be maintained moving forwards. The new infrastructure inaugurated in 2018/19 also highlights the importance of having an independent space for the home department in creating a sense of identity and community.
2. **Supportive members of teaching and staff:** There are many positive comments by students about the learning support network, staff efficiency, and helpful support staff throughout the years. Students are also complimentary of the informative and well-structured communication

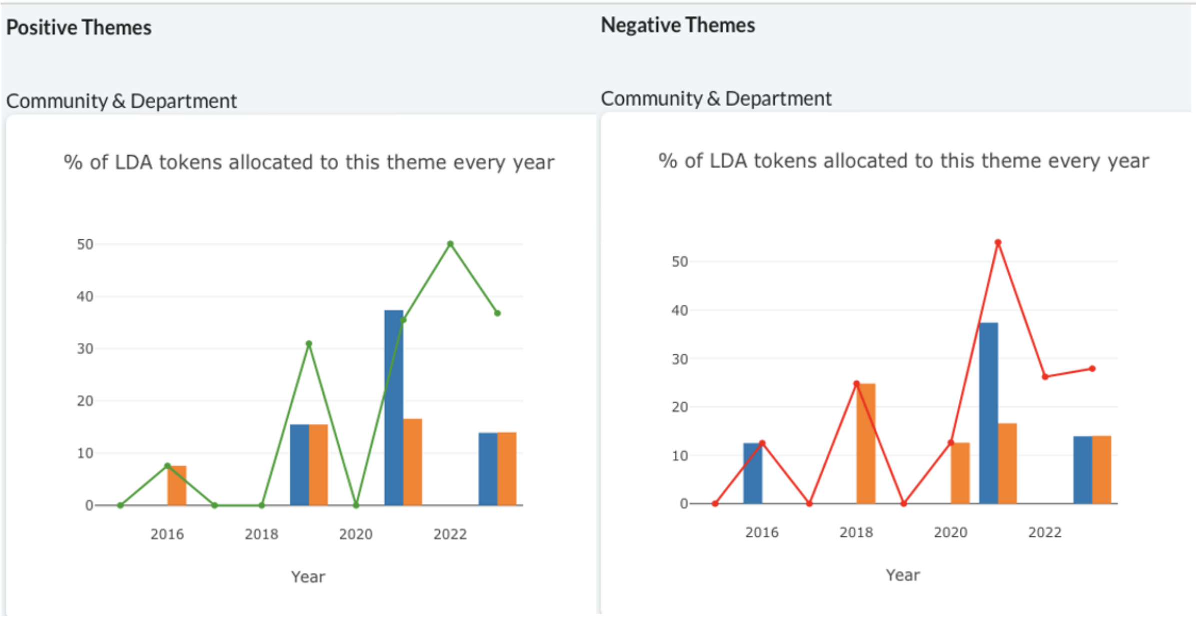


Figure 4.3: Percentage of Positive and Negative LDA Tokens allocated to Community and Department by year

from the home department, although some isolated complaints about support office responses have also been picked up.

3. **Diversified student environment and societies:** In the reviews, students are complimentary about the diversified nature of the cohort and the different student societies on campus, both of which enrich their learning experience at university.

Course and Content

Whilst the percentage of LDA tokens allocated to this theme has declined over the years, as illustrated in Figure 3.15, it is still very much relevant and a lot of the topics mentioned are actually recurrent. More specifically, students still talk about the same themes in this topic but also spend time discussing other topics. The apparent transient trend is a result of other themes taking precedence of the student learning experience because of the Pandemic and increasing importance of health and well-being as explored in Section 4.2.2 below. Below is a breakdown of the diagnostic analysis and prescriptive analytics in Course and Content.

Strengths

1. **Engaging Content:** The material covered in the course is consistently described as intellectually stimulating, challenging, rewarding, and enjoyable. Students find that the content covered helps them to enhance their problem-solving ability as well as other soft and hard technical skills such as R programming language. Students appreciate the content overlaps in third year and flexible learning approaches that enhance personal learning.

2. **Content Utility:** According to students, the course enhances their career prospects and qualifications. More specifically, the content of the course is both useful and practical, and the variety of classes enrich their skillset. The modules are either applied or have real-life applications,

and the practicality and utility of the content gives them an edge in interviews and assessment centres with employers.

3. **Course Depth:** Students have complimented the opportunity to tailor their courses and specialise in higher years. According to students, the course is unique in this way and enables them to find and pursue their interests.

4. **Course Breadth:** One of most recurrent compliments given by students is the flexibility and interdisciplinary nature of the department. As a result, students have the opportunity to learn from other departments and develop relevant skills, as well as make the most of inter-departmental opportunities to up-skill and learn.

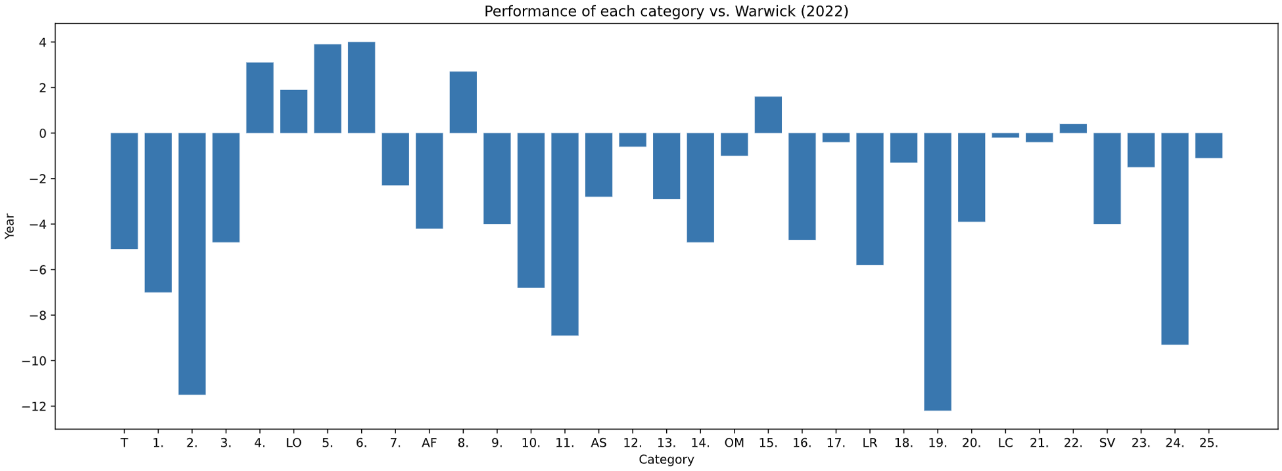


Figure 4.4: Bar Chart comparing performance of each category vs University of Warwick average in 2022

To validate the positive, engaging, and actualising breadth and depth of the content, there is a feature that stands out in the bar chart in Figure 4.4. More specifically, in the academic year 2021/22, the only learning category where both the category and its components exceeded the University of Warwick average is the Learning Opportunities (LO) category and its components, namely opportunities to 'explore ideas or concepts in depth' (5.) and 'bring information and ideas together from different topics' (6.). This indicates that whilst there are still improvements that can be made in this category, the department still stands out when it comes to learning opportunities compared to other departments.

Areas for Improvement

1. **Lack of Technical Skills:** A transient complaint that recurred in the first couple of years was the lack of opportunities to learn technical skills such as Python and VBA, as well as a lack of link to combine theoretical concepts between the four departments. Whilst this is not an issue that has been raised and picked up by the NLP workflow in recent years, looking ahead it is nonetheless a theme that could emerge in the future given the rapid advancements in this department's quantitative area of expertise that involves data science, large language models, and the processing abilities of computers.

   Prescription: This is a more prominent issue that must be considered with firms on the lookout for talent that can understand the theory behind the statistical models but also optimise their *deployment to add value. Failure to keep up the pace will exacerbate the already significant gap between academia and graduate jobs that has led to the emergence of new models such as the 'Hire, Train, Deploy' (HTD) framework [15]. With the emergence of higher education alternatives such as work apprenticeship programmes that focus on optimising deployment, it is important to treat this diagnosis not as a transient theme, but rather as a cyclical theme that appears every couple of years with the emergence of new advancements and discoveries.*

2. **Lack of Module Selection Guidance:** Students have complained that there is a lack of guidance on selecting modules and what they lead to, especially in first and second year. There is also a lack of foresight for year two and three modules and beyond, leading to a poor module selection that does not satisfy the pre-requisite requirements of later courses.

   Prescription: Provide access to some module material and motivation for students aiming to register for a particular module.

3. **(1st year) Adaptation and Career Guidance Struggles:** Students in the home department have spoken out about their struggles to transition to university as well as a lack of career guidance in first year. More specifically, they note that their first year outcomes count for 10% towards their degree whereas other departments whose students they are perhaps competing with for graduate jobs have no first year contribution to their degree grade. The reviews denote that this creates a lot of academic pressure on the students in this department, hinders their ability to socialise and adapt to university where everyone is a stranger, and gives them less time to explore career opportunities such as industry Spring Insight Programmes compared to competing departments'. Students also claim that searching for careers in their final years is too late compared to the opportunities they missed out on in first year with regards to getting experience or even a foot in the door via Spring Weeks and Summer Internships.

   Prescription: Whilst one potential remedy is to remove the 10% degree contribution of first year grades towards the final degree, such a measure is drastic in nature and can unsettle many other aspects of learning. A more measured approach that aims to make the workload more manageable, especially for first-year students, would be to create a career guidance framework in first year that encompasses the aforementioned module selection guidance diagnosis (3). This career guidance framework could be a series of monitoring point sessions, and would use a top-down approach via the following methodology:

   (a) *Enlist the different careers that require high-level Statistical Skills such as Data Scientist, Genetics, Research, and Quantitative Trading and Research in Finance.*

   (b) *Look at the different skills that are needed for these roles and enlist them- these naturally turn into the level 3 and 4 modules at university. For a Data Science related department of a university, therefore, this would include Mathematical Finance, Monte Carlo Methods, Bayesian Statistics, Neural Computing, Multivariate Statistics, Statistical Genetics, Stochastic Calculus, and Probability Theory among others.*

   (c) *Breakdown the statistical concepts and know-how required to perform inferences, simulations, predictions, visualisations and other tasks for each of these skills. For a Data Science related department of a university, for instance, this includes concepts such as the Gram-Schmidt Process and finding eigenvectors/eigenvalues. These concepts are learned* *in the level 2 modules at university which include Multivariate Calculus, Stochastic Processes, Linear Algebra, Mathematical Analysis, Differential Equations and Linear Statistical Modelling among others. The first year modules then create the foundation for the level 2 modules.*

(d) *Finally, each of these different careers provide Spring Weeks and Summer Internships whose application process consists of brainteasers and coding questions. These sessions could introduce students to resources to practice for these assessment, as well as include CV and application preparation workshops.*

This aforementioned example of a top-down approach helps students to take a more forward-oriented approach and combine it with the bottom-up approach that they already consider naturally when progressing through university. Moreover, it provides module guidance with regards to pre-requisites of higher level modules and skills they will cultivate along the way. It also gives them exposure to multiple career opportunities and the utility and practicality of the theory. Finally, this renewed sense of direction should provide students with greater career guidance and motivation to allocate time to their futures from their first year at university.

The lack of academic support is further reinforced in Figure 4.4, where the Academic Support (AS) category lagged the Warwick Average by three percentage points. In particular, components 13 and 14, which represent 'sufficient course-related advice and guidance' and 'advice regarding study choices', lagged the Warwick average by three and five percentage points respectively. To close the validation of this output, the paper 'Examining institutional career preparation: Student perceptions of their workplace readiness and the role of the university in student career development' by Gonzales, 2017 [8], produces the following conclusions.

1. The university plays a pivotal role in 'supporting students' career development by integrating career-preparation programming within all areas of the student experience' [8].
2. 'Students expect the institution to serve as their talent scout for employers' [8].
3. 'Career preparation programming and academics should not be mutually exclusive of each other' [8].

Both arguments therefore reinforce the need to enhance academic support with respect to forward-looking decisions such as module selection and career guidance.

4.2.2 Emerging Trends

Community and Department

Figure 4.3 illustrates an increased emphasis of LDA tokens for Community and Department in both positive and negative categories after 2017/18. Taking the external circumstance of the department's new building into consideration, the otherwise emerging trend is an observation that can be attributed to the following areas for improvements diagnoses.

1. **Lack of community feel:** For three consecutive academic years between 2020/21 and 2022/23, students are critical about a 'lack of community' feel, an 'estranged' feeling 'in the halls', and claim that 'the department [does not feel] welcoming'. This issue is not picked up in the previous years by the NLP workflow.

2. **Lack of group work:** Students point to a lack of group work, especially in first year, as a missed opportunity to network, meet new people, share ideas, and cultivate transferable skills of collaboration, communication, and presentation among others.

   Prescription: Provide collaboration opportunities in first year when completing assignments or seminar exercises. Use module feedback to monitor the success of these opportunities in helping students adapt to university, meet new people, and cultivate the aforementioned skills. This measure might be more productive for practical exercises that involve programming as opposed to theoretical and proof-based questions. This measure is supported by a multitude of peer-reviewed articles such as a paper by Hammar Chiriac, 2014, titled 'Group work as an incentive for learning – students' experiences of group work' [7]

3. **Delayed publication of exam timetables:** Students believe that the University can do more to publish the april and summer exam timetable earlier to enhance clarity and planning with regards to revision and holidays.

   Prescription: Release summer exam timetables earlier than the current release date.

4. **Monitoring point sessions are not productive:** Students do not find current monitoring point sessions to be a useful or productive use of their time.

   Prescription: Turn monitoring points into an opportunity for students to network with peers or career workshops as opposed to lectures which they already have in abundance.

5. **Inconsistent standards of personal tutor support:** Some students do not feel supported by their personal tutor. They do not find their personal tutor to be helpful or useful in guiding or addressing concerns they may have about their course, module, or careers.

   Prescription: This is a more nuanced issue that must be treated on a case by case basis. It is nonetheless imperative to have some baseline standard for personal tutors to adhere to. If a personal tutor is not able to answer a question, they must at least be able to direct the student to a member of staff who will be able to answer their question. The review has unfortunately uncovered several situations where personal tutors do not respond, provide unhelpful responses, and/or do not show willingness to help and support the student. A good researcher is not necessarily a good tutor or teacher, and training must be provided to ensure this baseline standard is met.

6. **Poor outlets on campus:** An emerging diagnosis picked up in 2021/22 and 2022/23 is the poor quality and choice of campus cafes and restaurants, as well as other SU outlets.

It is interesting that all six aforementioned topics, if implemented, can contribute significantly to improved mental health and overall well-being. A report by the World Health Organisation found that in the aftermath of the Covid-19 Pandemic, there has been a 25% increase in prevalence of anxiety and depression worldwide [22]. Whilst it is difficult to establish all the causes of this concerning statistic, it does indicate that these emerging complaints by students are not just a consequence of changes made within the home department. Rather, they indicate a global occurrence for which additional measures must be taken to directly address these concerns moving forwards.

Sentiments about the learning community are echoed in Figure 4.5 below, in which there is a 34.62% decline in the feeling of belonging with other staff and students from 2016/17 to 2020/21, during the peak of the Covid-19 Pandemic. It is no surprise then that the following year, when Pandemic-induced restrictions were gradually eased, there was a reversal via a slight improvement in the figures. Moreover, students feel less represented by the SU as illustrated in the bottom right graph of Figure 4.5. Finally, the external nature of the factors as hypothesised through the Pandemic are put into perspective in Figure 4.4, where the Learning Community (LC) category and its components remain in the ballpark of the Warwick University average despite the declining trend. This indicates that these issues are department agnostic and represent a university-wide emergence as a result of the Pandemic.

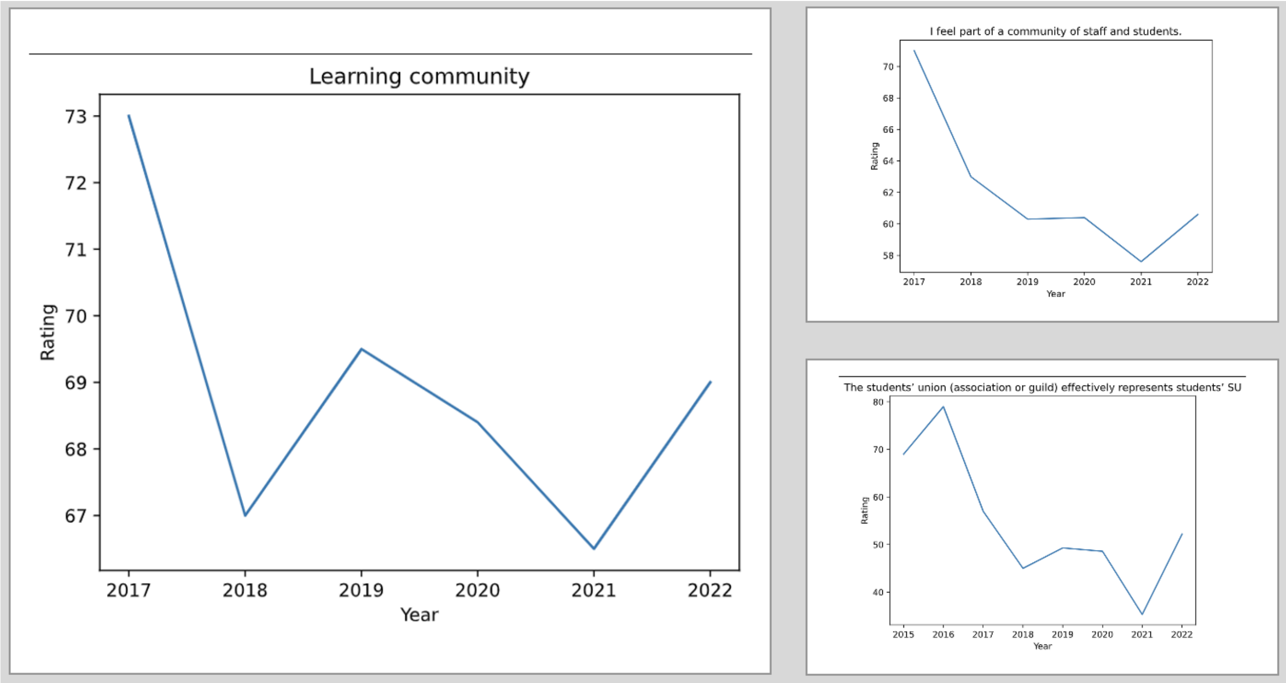


Figure 4.5: Line graphs illustrating the ratings given to community and department components by students from 2016/17 to 2021/22

Teaching and Learning Set-Up

1. **Assessment Feedback:** A significant issue raised by students over the past three years is the quality of feedback provided. The major concerns are related to delayed feedback and unhelpful comments, and there are some isolated yet discernible complaints regarding the objectivity of assignment grades when multiple markers are tasked to correct an assignment.

   Prescription: It is important to adhere to the marking deadline provided to students, and there must be a baseline standard of marking that is agreed to by students and markers alike to ensure that both sides' voices are considered in what is acceptable yet feasible feedback. Moderation may be introduced when multiple markers are tasked with correcting assignments to ensure consistency throughout the cohort.

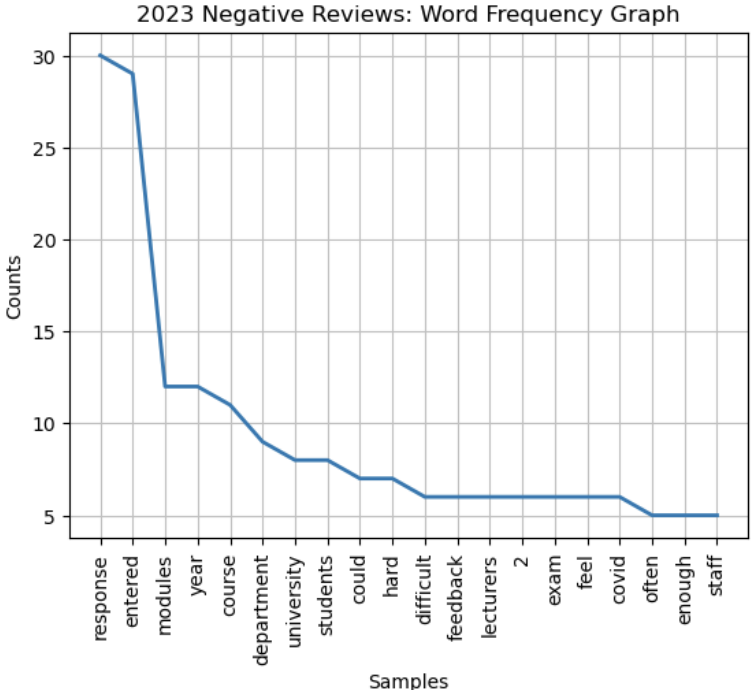


Figure 4.6: Word Frequency Graph for Negative Comments in NSS 2023 Department Data

The story depicted by the LDA algorithm is further reinforced in Figure 4.6, where the term ‘feedback’ is employed once on average for every two times the students’ speak about ‘modules’. This emphasises on the importance and relevance of feedback with regards to the learning experience.

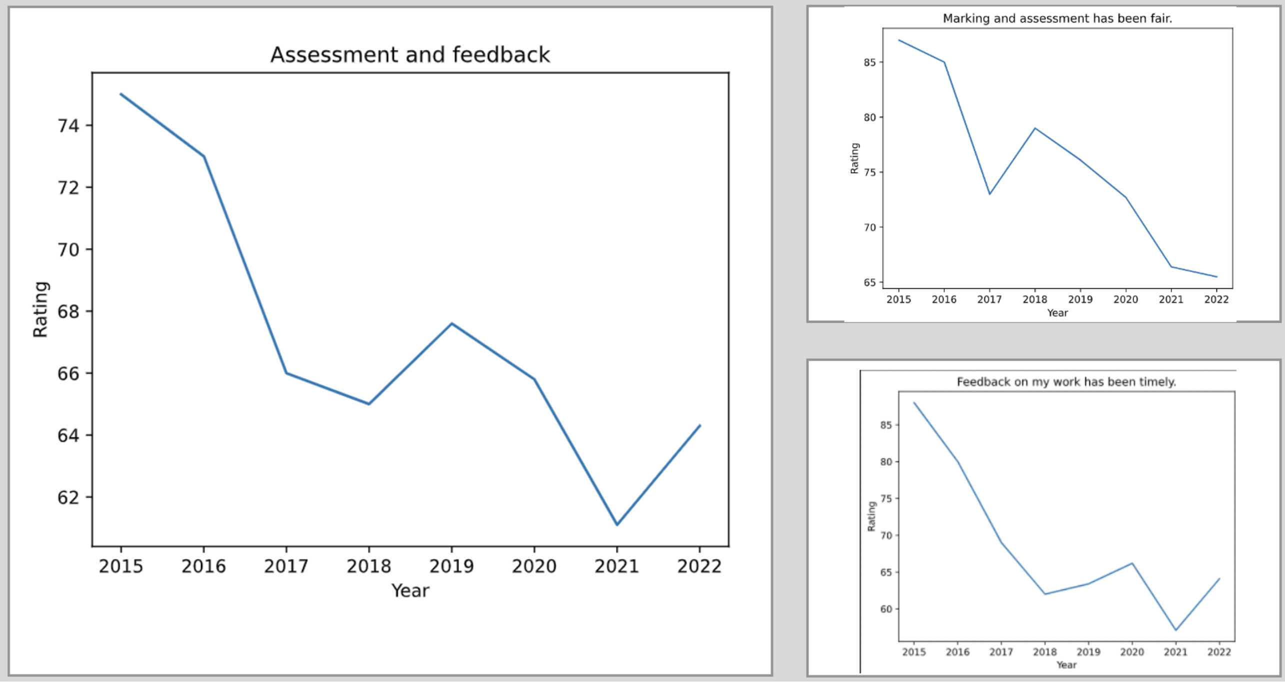


Figure 4.7: Line graphs illustrating the ratings given to assessment and feedback components by students from 2016/17 to 2021/22

Moreover, the graphs in Figure 4.7 help to validate the diagnostics which include delayed feedback, unhelpful comments, and marking inconsistencies. Firstly, the graphs illustrate a 15% drop in the ratings attributed to 'fair marking and assessment' between 2017/18 and 2021/22. The timeliness of feedback as illustrated in the 'timely [...] feedback' graph is also shown to fluctuate heavily in this time period. A thorough module feedback review could highlight which modules face this issue, and this would also enable the affected undergraduate levels and affiliated departments to be identified. Additionally, the module review datasets would be able to gauge whether or not the diagnoses are a department-wide issue for the home department. Regardless, the diagnostics and cross-correlation suggest the lack of enforcement of a structure, as opposed to just recommendations and guidelines, that guide marking and moderation of assignments in modules.

Chapter 5

Conclusion

5.1 Suitability

Natural Language Processing is a rapidly developing space that accommodates a vast array of concepts which enable qualitative text to be processed. This space is developing rapidly to enhance understanding of human language whilst leveraging the increasing computational and processing abilities of computers [29]. This project has explored:

- data engineering and preprocessing concepts;
- sentiment analysis tools;
- thematic analysis techniques such as topic modelling, extractive and abstractive text summarisation algorithms and their deployment methods;
- model and output validation tools.

Finally, the outputs obtained by the model workflow and data pipeline have been reconciled with the quantitative data and opinions of stakeholders including but not limited to students and their suggestions. As an academic exercise, the NLP tools at play have demonstrated an incredible potential to enhance learning analytics in the following pathways.

1. **Rigorous analysis:** The workflow is able to leverage the strengths of the different tools at different stages. More specifically, sentiment analysis at the initial stage enables two different categories of topics, namely strengths and areas for improvements, to be developed further down the line. Moreover, employing topic modelling, and more specifically LDA, as the first thematic analysis technique enables the topics to be quantified in order of prevalence in the text as well as a correspondence with the sentences most likely linked to that topic to be established. Finally, extractive text summarisation at the final stage enables a deep-dive of different themes for each topic whilst considering the nuances of human language and the student suggestions in their own words through those sentences. This enables distinct and actionable diagnostic insights to be retrieved from the workflow. This directly addresses gaps 1 and 2 from Section 1.3.2.

2. **Reconciliation:** The workflow is able to cross-correlate the diagnostic analysis with the quantitative data obtained from the survey or/and for the same context. Since the diagnostics are distinct and actionable and the topics are quantifiable, this enables a direct correspondence to be made with the quantitative metrics for which analytic capabilities are more developed, thus addressing the current disbalance between qualitative and quantitative analysis (1.3.2).

3. **Scalability:** The data pipeline is department agnostic since the processes are run on the qualitative data columns, irrespective of the industry, department, or subject of the dataset. This makes the product and pipeline scalable and can produce actionable diagnoses for other departments and even the university as a whole.

4. **Stakeholder consideration:** The toolkit includes extractive text summarisation techniques that enable the pipeline to factor in the students' own words, feedback, and suggestions. Moreover, since the final stage to concretise the diagnostic analysis from the output is manually driven, this allows other stakeholders such as lecturers and support staff to add their own opinions, feedback, and suggestions. This directly addresses the third gap in the existing provision (1.3.2).

While the product can be further developed to transform it into a minimum viable product (5.2), this project has demonstrated the potential opportunities that exist if natural language processing tools are approached more rigorously in learning analytics. In particular, the findings have illustrated that the teaching and learning set-up theme could enhance learning retention and the students' problem-solving abilities, course and content's areas for improvements could increase learning motivation, further studies guidance, as well as career prospects, whilst improvements identified under community and department could enrich the students' learning experiences and engagement (4.2). In other words, the natural language processing dashboard developed in this project successfully generated substantiated diagnoses and prescriptions to improve all the different branches of learning analytics (1.2, 1.4), including but not limited to learning motivation, retention, engagement, and career guidance.

5.2 Further Work

5.2.1 Incorporate Additional Data

While the project explored multiple datasets to establish the model and data pipelines, the final product only considers the national students survey data in its evaluation. Whilst the latter is significant and very informative, other datasets could help provide further insights into particular topics. For instance, one of the topics identified in the findings (4.2) was course and content, and the module feedback datasets, when reconciled with module outcomes, could help give more actionable insights for the themes identified in that topic. If a particular theme refers to teachers reading off slides as a concern, for example, the module outcomes datasets would be able to identify the modules where this issue is present, determine the magnitude of the problem by the number of modules affected, and work to directly address the diagnosis. The current product only identifies the presence of the issue and the importance of setting a department-wide teaching guideline that encourages alternative pathways to lecturing.

5.2.2 Trial Automated Diagnostics

Section 1.4 introduced the concept of abstractive text summarisation and a feature of its models that is generating new text. In the context of summarisation, there is scope to incorporate this tool in the models pipeline (3.2) between the extractive text summarisation step and final diagnostic stage. More specifically, instead of diagnosing manually from the results, an abstractive text summariser such as Google's T5 (2.4.4) could be fine-tuned to generate a summary diagnostic output from the most relevant sentences that are the extractive text summaries for each topic. This step requires the model to be fine-tuned purposefully using educational text given its ability to capture long-range dependencies. A successful incorporation of this concept at this stage would nonetheless automate the only manual step of the process and make the product a lot more scalable.

5.2.3 Introduce Supervised or Semi-Supervised Topic Modelling

In the current model pipeline, the topic modelling method LDA is a clustering algorithm that can be improved on. More specifically, datasets can be engineered that incorporate staff expertise about issues, themes, and trends in the department. This would be fed into the models at the fine-tuning stage to provide more defined rules to guide the model in identifying topics for each sentiment category.

5.2.4 Consider Different Learning Styles

Another potential improvement moving forwards that is beyond the scope of this project is to categorise students by their inherent learning processes and styles, a tool that could then in turn classify student feedback before attempting to find themes in each learning category. While this would require fundamental changes in the construction of the surveys and considerations of the ethical and legal framework, it is a potential change that would help to identify trends in the different groups which otherwise disappear when these groups are combined (i.e. Simpson's Paradox). For instance, some students might learn better when provided with the lecture notes beforehand as it gives them time to read, whilst others may find it more helpful to only see the content in lecture as it encourages them to pay more attention.

Chapter 6

Project Management

This section reviews progress of the project across the different stages based on accomplishment of the core objectives as well as the initial timetable intended to guide the project.

6.1 Overview

Reviewing the original gantt chart designed at the beginning to guide the project in Figure 6.1, the product alongside the model workflow and data pipeline were completed in time for the product demo and project presentation. This provided ample time to complete the write-up and evaluate the outputs of the product between the presentation and final deadline.

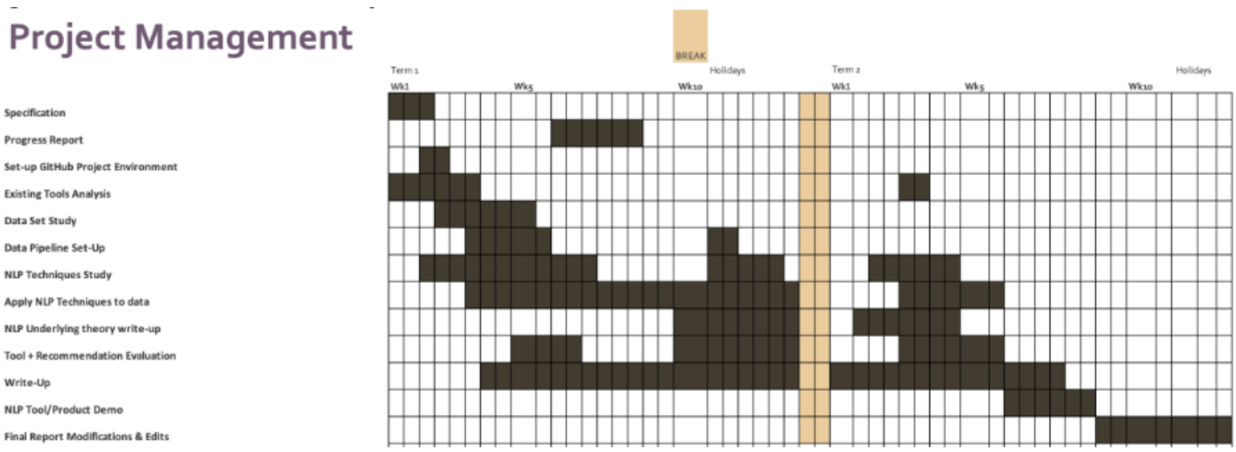


Figure 6.1: Gantt Chart Timeline from 02nd October 2023 to 31st March 2024

There were four main stages to the project, with stages three and four co-existing in a cyclical manner as illustrated in the figure below.

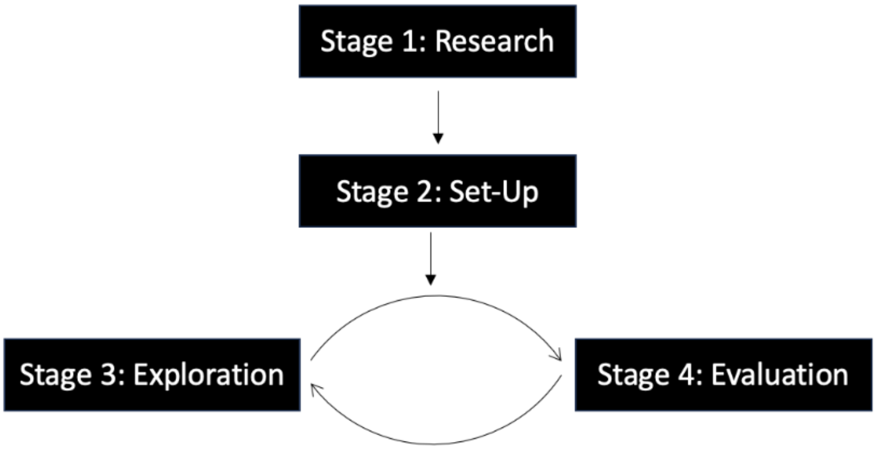


Figure 6.2: Stages Diagram for Project Process

6.1.1 Stage 1

This step was the research stage of the project. The key components of research included:

- an exploration of the limitations of existing learning analytics tools;
- a detailed review of the illustrative datasets and the best data pipeline to set-up;
- an in-depth exploration into NLP techniques and methodologies.

The primary method used to collect data on the limitations of existing tools within learning analytics was literature review, namely peer-reviewed papers such as ‘Learning analytics in higher education: a preponderance of analytics but very little learning?’ by Valenzuela et al., 2021 [17]. The second two bullet points were completed using a combination of online tutorials and literature. In particular, the book ‘Speech and Language Processing’ by Jurafsky (2023) provided a structured guide to natural language processing concepts, models, and tools [31]. This guide introduced concepts such as sentiment analysis, topic modelling, extractive and abstractive text summarisation. The tools used to deploy these concepts were then explored further using the articles corresponding to the topics. Some of these papers include:

- ‘The Automatic Creation of Literature Abstracts’ by Luhn, 1958 for luhn extractive text summarisation [1];
- ‘Latent Dirichlet Allocation’ by Blei et al, 2003 for LDA topic modelling [2];
- ‘LexRank: Graph-based Lexical Centrality as Salience in Text Summarization’ by Güneş et al., 2004, for lexrank extractive text summarisation [3].
- ‘Attention Is All You Need’ by Vaswani et al., 2017, for the attention mechanism of transformer-based networks [11];
- ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’ by Devlin et al., 2019, for transformer-based sentiment analysis [12];

- ‘Bug report severity level prediction in open source software: A survey and research opportunities’ by Gomes et al., 2019 for cosine distance metric [13];
- ‘Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer’ by Raffel et al., 2020, for abstractive text summarisation [14];
- ‘BERTopic: Neural topic modeling with a class-based TF-IDF procedure’ by Grootendorst, 2022 for BERTopic topic modelling [19].

From an exploratory point of view, therefore, this stage consisted of significant research on NLP preprocessing techniques and processing models for learning analytics.

6.1.2 Stage 2

The second stage consisted of two simultaneous activities, namely reconciling the key findings from the research stage while setting up the data pipeline to clean, transform, and validate the data for processing. During this stage, the process continued to resemble a plan-based methodology involving research, write-ups, and developing the product workflow and data pipelines.

Moreover, there was significant progress on the data workflow front, with particular emphasis on constructing a preprocessing pipeline in the Preprocessor module to transform the qualitative datasets, specifically from the national students survey, for the NLP models to process. Finally, the Visualiser module was also developed to complement the thematic analysis as part of the prescriptive analytics output. The Visualiser module contained methods that could deploy some topic modelling and keyword extraction techniques as well as visualisation tools.

6.1.3 Stage 3

This stage contained the first rigorous exploration and development of the NLP tools to the qualitative data, as well as a write-up of the underlying concepts of the approach. Both these components would enable a deep-dive into the datasets to allow a cross-correlation examination between the qualitative and quantitative data.

During this phase, online tutorials of implementations from Medium and Datacamp helped to enhance understanding of the theory and implementation of the literature papers explored in stage one. This step of the cyclical process consisted of an exploration of different processing models for sentiment and thematic analysis that were identified during the research stage. These tools were deployed using Python based open-source libraries as explored in Section 3.4.

6.1.4 Stage 4

This stage included an evaluation of both output and model in line with core objectives three and four. As the project transitioned between stages three and four, the steps in this process enabled both an evaluation of the models as well as validation of the output via the diagnosis analysis and prescriptive analytics produced by the workflow. The project would feed off the validation to further develop the NLP tools in stage three.

6.2 Methodology

The working methodology employed was a hybrid between a plan-based methodology and agile methodology. More specifically, the overarching process resembled a plan-based methodology where the broad components of each stage were outlined, but the details about how and which NLP concepts and their corresponding tools were implemented depended on the research findings and evaluation stage, making stages three and four incremental. The broader framework did not diverge significantly and is outlined in Section 6.1, which justified the use of a hybrid plan-based and incremental methodology.

Another important part of the methodology included resources employed to mitigate risk, some of which are identified below.

- Microsoft Teams and Google Drive was used to save and store back-ups of documents saved on the local computer.
- The primary sources of literature were either available at the Warwick Library or online at no cost using the university access.
- The product was linked to a main GitHub Repository to maintain back-ups and version control. Modifications and additions were then made on a local environment and committed to the repository.
- The project specification, progress report, and final report were written on Overleaf, which is an online LaTex editor. This task was also shared with the supervisor and eliminated the risks associated with local storage.

6.3 Ethical and Social Considerations

Ethical

All the datasets used for this project are illustrative and do not contain any live data. More specifically, the data is sanitised and is both anonymised and untraceable, satisfying the ethical considerations when it comes to storing, securing, and accessing the data. Note that this data is stored at the university and only shared on a need-to-know basis in line with the General Data Protection Regulation (GDPR). Moreover, data from the national students survey has already been anonymised, which removes any privacy concerns with respect to the ethical constraints.

Social

The societal benefits of education are widely established, and this project is very closely tied to its development. Therefore, there is a social obligation to ensure a thorough, complete, and accurate exploration of learning analytics tools given its wide-reaching potential to improve student learning.

A larger exploration of this project could also consider the wide-ranging links between student learning and the students' overarching experience in society. For this study, the tools created are primarily an academic exercise to demonstrate the opportunities that exist if learning analytics tools are approached more rigorously.

6.4 Appraisals and Reflections

This project has been a very challenging but rewarding experience. It is motivating to have had this opportunity to work on a deliverable that is relevant, employs theory learned in class, and is at the forefront of advances in Statistics and Computer Science. Looking back, the model validation stage contained the biggest challenges faced in the project, since many iterations and models failed validation. In particular, there were three challenges of note.

1. From a data engineering perspective, the initial data pipeline considered each student review as a data point, without considering that a student could refer to multiple topics in one review. This made it difficult for the unsupervised topic modelling algorithms to identify the topics. This issue was identified during the model validation stage after deployment, which made it time-consuming to change.
2. The second biggest challenge was the inability of BERTopic to distinguish between topics due to its long-range dependencies. Again, to successfully evaluate the model pipeline, the subsequent text summarisation models also had to be deployed to evaluate the workflow performance. The text summaries illustrated BERTopic's inability to distinguish between different topics within education in this case, and resulted in the implementation of LDA instead. These two conundrums elongated the 'apply NLP Techniques to data', 'NLP underlying theory write-up' and 'tool + recommendation' stages of the gantt chart in Figure 6.1.
3. Another very challenging aspect of the project was the need to simultaneously learn new concepts and implement them for the product.

In response, there were two key aspects of the work routine that really served the purposes of this project and helped tackle the challenges.

1. Weekly supervisor meetings ensured that the project was always moving forwards at a consistent rate with weekly deliverables to work towards, discuss, and receive feedback on. The meetings also served to clarify learning analytics and output related concerns, issues, and queries.
2. Allocating a small number of hours to the project on many days, as opposed to long hours on a small number of days, enabled steady advances to take place on all fronts of the project, namely research, write-up, and coding. This routine also enhanced understanding of new NLP techniques as it resisted the urge to learn the theory and findings over a very short period.

The overarching key to ensure that the project stayed on track was to maintain a consistent research routine to be able to navigate through the content heavy literature whilst simultaneously exploring and developing tools to counter the gaps found in the existing provision from the literature.

Bibliography

[1] H. P. Luhn, “The automatic creation of literature abstracts,” *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958. DOI: 10.1147/rd.22.0159.

[2] D. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research 3*, 2003, Accessed: 24.03.2024. [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[3] E. Güneş and D. Radev, “Lexrank: Graph-based lexical centrality as salience in text summarization,” *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, Dec. 2004, ISSN: 1076-9757. DOI: 10.1613/jair.1523.

[4] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, ISBN: 9781139472104. [Online]. Available: https://books.google.co.uk/books?id=t1PoSh4uwVcC.

[5] M. Kalkan, “Recent trends at higher education emphasizing active cooperative learning methods involving individual learning styles.,” *Buca Faculty of Education Journal / Buca Egitim Fakültesi Dergisi*, no. 30, pp. 76–86, 2011, ISSN: 13025147. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&amp;AuthType=ip,uid&amp;db=ehh&amp;AN=80242018&amp;site=ehost-live&amp;scope=site.

[6] Y. Huang, *The Oxford Handbook of Pragmatics*. Oxford University Press, 2013, ISBN: 9780191749858. DOI: 10.1093/oxfordhb/9780199697960.001.0001.

[7] E. Hammar Chiriac, “Group work as an incentive for learning – students’ experiences of group work,” *Frontiers in Psychology*, vol. 5, 2014, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.00558. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2014.00558.

[8] M. Gonzales, “Examining institutional career preparation: Student perceptions of their workplace readiness and the role of the university in student career development,” *Pepperdine UniversityProQuest Dissertations Publishing*, 2017. [Online]. Available: https://media.proquest.com/media/hms/ORIG/2/O0OlH?_s=F0uCBcnE56qfloaAVqwvaipaUGE%3Ds.

[9] F. V. O’Callaghan, D. L. Neumann, L. Jones, and P. A. Creed, “The use of lecture recordings in higher education: A review of institutional, student, and lecturer issues,” *Educ Inf Technol 22, 399–415*, 2017. DOI: 10.1007/s10639-015-9451-z.

[10] S. Syed and M. Spruit, "Full-text or abstract? examining topic coherence scores using latent dirichlet allocation," *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2017, Accessed: 24.03.2024. DOI: 10.1109/DSAA.2017.61.

[11] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," 2017, Accessed: 24.02.2024. DOI: 10.48550/arXiv.1706.03762.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019, Accessed: 24.02.2024. DOI: 10.48550/arXiv.1810.04805.

[13] L. A. F. Gomes, R. da Silva Torres, and M. L. Côrtes, "Bug report severity level prediction in open source software: A survey and research opportunities," *Information and Software Technology*, vol. 115, pp. 58–78, 2019, ISSN: 0950-5849. DOI: 10.1016/j.infsof.2019.07.009.

[14] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2020, Accessed: 02.04.2024. DOI: 10.48550/arXiv.1910.10683.

[15] R. Craig. "The htd revolution: Hire-train-deploy." Accessed: 21.02.2024. (2021), [Online]. Available: https://www.ipsos.com/en-uk/national-student-survey-2023-resultshttps://www.forbes.com/sites/ryancraig/2021/10/01/the-htd-revolution-hire-train-deploy/?sh=69a713b127fb.

[16] C. Khanna. "Word, subword, and character-based tokenization: Know the difference." Accessed: 13.11.2023. (2021), [Online]. Available: https://towardsdatascience.com/word-subword-and-character-based-tokenization-know-the-difference-ea0976b64e17.

[17] C. G. Valenzuela, C. G. González, A. R.-M. Tagle, and A. L. Vyhmeister, "Learning analytics in higher education: A preponderance of analytics but very little learning?" *International Journal of Educational Technology in Higher Education*, no. 18, p. 5, 2021, Accessed: 10.11.2023. DOI: 10.1186/s41239-021-00258-x.

[18] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Front Sociol; 2022 May 6;7:886498*, 2022. DOI: 10.3389%2Ffsoc.2022.886498.

[19] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," 2022, Accessed: 28.03.2024. DOI: 10.48550/arXiv.2203.05794.

[20] C. Maklin. "Latent dirichlet allocation." Accessed: 21.01.2024. (2022), [Online]. Available: https://medium.com/@corymaklin/latent-dirichlet-allocation-dfcea0b1fddc#:~:text=Latent%20Dirichlet%20Allocation%2C%20or%20LDA%20for%20short%2C%20is%20an%20unsupervised,of%20clusters%20(i.e.%20topics)..

[21] S. Metzger. “A beginner’s guide to tokens, vectors, and embeddings in nlp.” Accessed: 17.11.2023. (2022), [Online]. Available: https://medium.com/@saschametzger/what-are-tokens-vectors-and-embeddings-how-do-you-create-them-e2a3e698e037#:~:text=Embedding%3A%20To%20give%20tokens%20meaning,have%20similar%20vectors%20after%20training..

[22] World-Health-Organisation. “Covid-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide.” Accessed: 21.02.2024. (2022), [Online]. Available: https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide.

[23] M. Ali. “Nltk sentiment analysis tutorial for beginners.” Accessed: 19.11.2023. (2023), [Online]. Available: https://www.datacamp.com/tutorial/text-analytics-beginners-nltk#.

[24] S. Anand. “A beginner’s guide to topic modeling nlp.” Accessed: 21.11.2023. (2023), [Online]. Available: https://www.projectpro.io/article/topic-modeling-nlp/801#:~:text=Topic%20modeling%20is%20a%20part,analysis%20to%20analyze%20the%20context..

[25] H. Axelborn and J. Berggren, “Topic modeling for customer insights: A comparative analysis of lda and bertopic in categorizing customer calls,” 2023, Accessed: 28.03.2024. [Online]. Available: https://umu.diva-portal.org/smash/get/diva2:1763637/FULLTEXT01.pdf.

[26] S. Benyahia. “National student survey 2023 results released.” Accessed: 21.02.2024. (2023), [Online]. Available: https://www.ipsos.com/en-uk/national-student-survey-2023-results-released#:~:text=The%20National%20Student%20Survey%20(NSS,expressing%20their%20views%20in%202023.&text=The%20highly%20anticipated%202023%20results,overall%20response%20rate%20of%2071.5%25..

[27] A. Chawla. “A pivotal moment in nlp research which made static embeddings (almost) obsolete.” Accessed: 11.02.2024. (2023), [Online]. Available: https://www.blog.dailydoseofds.com/p/a-pivotal-moment-in-nlp-research.

[28] S. Cristina. “The transformer attention mechanism.” Accessed: 16.11.2023. (2023), [Online]. Available: https://machinelearningmastery.com/the-transformer-attention-mechanism/.

[29] F. Fleuret, *The Little Book of Deep Learning*. 2023. [Online]. Available: https://fleuret.org/public/lbdl.pdf.

[30] GeeksforGeeks. “Artificial neural networks and its applications.” Accessed: 21.02.2024. (2023), [Online]. Available: https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/.

[31] D. Jurafsky and J. H. Martin, *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2023. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

[32] Z. Keita. "Mastering backpropagation: A comprehensive guide for neural networks." Accessed: 25.03.2024. (2023), [Online]. Available: https://www.datacamp.com/tutorial/mastering-backpropagation.

[33] A. Menzli. "Tokenization in nlp: Types, challenges, examples, tools." Accessed: 13.11.2023. (2023), [Online]. Available: https://neptune.ai/blog/tokenization-in-nlp.

[34] A. Mittal. "Nlp rise with transformer models — a comprehensive analysis of t5, bert, and gpt." Accessed: 03.04.2024. (2023), [Online]. Available: https://www.unite.ai/nlp-rise-with-transformer-models-a-comprehensive-analysis-of-t5-bert-and-gpt/#:~:text=BERT%3A%20Requires%20task%2Dspecific%20head,and%20adaptable%20to%20new%20tasks..

[35] A. Pai. "What is tokenization in nlp?" Accessed: 13.11.2023. (2023), [Online]. Available: https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/.

[36] A. Taparia. "Keyphrase extraction in nlp." Accessed: 08.04.2024. (2023), [Online]. Available: https://www.geeksforgeeks.org/keyphrase-extraction-in-nlp/.

[37] Turing. "5 powerful text summarization techniques in python." Accessed: 21.11.2023. (2023), [Online]. Available: https://www.turing.com/kb/5-powerful-text-summarization-techniques-in-python.

[38] University-of-Warwick. "Academic technology, guides for students using academic technology." Accessed: 08.10.2023. (2023), [Online]. Available: https://warwick.ac.uk/services/academictechnology/support/student-guides/.

[39] AWS. "What's the difference between deep learning and neural networks?" Accessed: 24.03.2024. (2024), [Online]. Available: https://aws.amazon.com/compare/the-difference-between-deep-learning-and-neural-networks/#:~:text=Deep%20learning%20models%20can%20recognize,neurons%20in%20a%20layered%20structure..

[40] A. Grishina. "Gpt-3 vs. bert: Ending the controversy." Accessed: 28.03.2024. (2024), [Online]. Available: https://softteco.com/blog/bert-vs-chatgpt#:~:text=GPT%2D3%20is%20typically%20fine,particular%20tasks%20for%20effective%20performance..

[41] Hugging-Face. "Bert base model (uncased)." Accessed: 14.01.2024. (2024), [Online]. Available: https://huggingface.co/google-bert/bert-base-uncased.

[42] Hugging-Face. "Distilbert-base-uncased-finetuned-sst-2-english." Accessed: 28.03.2024. (2024), [Online]. Available: https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english.

[43] E. Martin. "From rnns to transformers." Accessed: 24.03.2024. (2024), [Online]. Available: https://www.baeldung.com/cs/rnns-transformers-nlp.

[44] J. Murel and E. Kavlakoglu. "What is bag of words?" Accessed: 25.03.2024. (2024), [Online]. Available: https://www.ibm.com/topics/bag-of-words.

[45] Team-Multiverse. "Report: New stats show uk backs apprenticeships over university." Accessed: 25.02.2024. (2024), [Online]. Available: https://www.multiverse.io/en-GB/blog/apprenticeships-business-case-report.

[46] IBM. "What is a neural network?" Accessed: 25.03.2024. (), [Online]. Available: https://www.ibm.com/topics/neural-networks#:~:text=Every%20neural%20network%20consists%20of,own%20associated%20weight%20and%20threshold..

[47] IBM. "What is natural language processing?" Accessed: 09.11.2023. (), [Online]. Available: https://www.ibm.com/topics/natural-language-processing#:~:text=the%20next%20step-,What%20is%20natural%20language%20processing%3F,same%20way%20human%20beings%20can.

[48] Intel. "What is a gpu? graphics processing units defined." Accessed: 11.03.2024. (), [Online]. Available: https://www.intel.com/content/www/us/en/products/docs/processors/what-is-a-gpu.html#:~:text=What%20does%20GPU%20stand%20for,video%20editing%2C%20and%20gaming%20applications..

[49] University-of-Sheffield. "Semantics." Accessed: 19.11.2023. (), [Online]. Available: https://www.sheffield.ac.uk/linguistics/home/all-about-linguistics/about-website/branches-linguistics/semantics.