# Analysing Ethics Related Content in Reports Using Topic Modelling

*Owen Scollon*


4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2023

# Abstract

As Computer Science and Artificial Intelligence become increasingly used and applied to people's everyday lives, the ethics involved becomes a growingly talked about issue. This meant that the university has taken steps to provide courses which ensure that students understand the impacts of new and emerging technologies. This paper aims to evaluate how students' awareness of ethics in CS and AI may have increased between year groups by using Topic Modelling to analyse its occurrence in written reports.

To achieve this, three goals were set. Firstly, to develop a topic modelling method which could accurately identify ethics-related topics in written work. Secondly, to ensure that the method is able to determine if ethics is an increasingly talked about topic in student reports from one year to the next. Lastly, to accurately evaluate the data so that meaningful conclusions can be drawn from the models, while also considering the influence of external factors on the results.

All three goals were achieved to varying success. The first was achieved by using guided topic modelling to allow the extraction of ethics related topics. The second goal was achieved, to a lesser extent, as more models showed ethics to be easier to extract from the latter corpus, however, the first corpus showed a deeper discussion of ethics. The third goal was achieved through a discussion of how COVID-19 and changes to course administration may have influenced students' consideration and application of ethics in their own projects, thus justifying the difficulties in reaching the second goal.

# Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: 667755
Date when approval was obtained: 2023-03-15

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Owen Scollon*)

# Acknowledgements

Firstly, I would like to thank my supervisor, James Garforth. A truly understanding individual, who not only gave me the space and flexibility that I needed to work on this project, but is also a genuinely caring member of staff, who is leading every course he teaches with integrity.

Next, I think it would only be right to mention my parents, without whom I would likely never have made it here. All those nights of sitting with me at the dining room table making sure that I could spell and do sums really paid off didn't it dad. I should remind you though, you "only nail bits eh wid thigither". Also, to my mum, you are a saint for putting up with me for the last four years. Thank you. I love you both.

To the rest of my family, from my grandparents to my sisters, thank you for always supporting me, I love you all.

I would also like to give a brief shout out to all of the friends that I've made here. I dropped out of my first year only to come back the next year. This was for various reasons but one of which was definitely the lack of people to latch onto in times of need. The first friends I met when it came to round two were Samuel and Matthew, and I'm glad that we never drifted apart.

Special thank you to my SDP boss, best friend, cornerstone of my life, and the best girlfriend anyone could ever ask for. I love you Alex and I cherish you more and more everyday.

Lastly, this is dedicated to our family Husky, Lexi, who passed away this semester. We get your sister, Luna, in 3 days. Keep on howling wherever you may be, and we'll make sure she's howling back.

# Table of Contents

# Chapter 1

# Introduction

This project's aim is to measure if students' usage and understanding of ethics applied to Computer Science and Artifical Intelligence has increased in quantity and quality during their time at The University of Edinburgh. To understand this project, we must acknowledge what students are being taught in the ethics branch, as well as the technical aspect used for analysing student output.

## 1.1 Motivation

A conversation which is normally lacking during the development of new and current technologies in computer science, is the question of whether something is ethically or morally right. Ethics, as it is defined in the Oxford dictionary is "the branch of knowledge that deals with moral principles." This branch is a prominent region of thought and action right now as areas like artificial intelligence and data science have more of a foothold on our day to day lives.

There have been a number of incidents in recent years which have enlightened an even greater relevance of the analysis of the ethics in CS and AI. These events include, but are not limited to: bias in data, such as in Google's chauvinistic Word2vec, or Microsoft's offensive TayTweet's incident; the spread of propaganda for the Myanmar military through Facebook; customised adverts on Facebook to target Brexit voters by Cambridge Analytica; reports of the White House using everyday smart devices to collate data and predict violence in civilians in the form of an agency called HARPA; and Amazon selling facial recognition software to Palantir, a company involved in US immigration enforcement [8].

In aims of making students aware of real world scenarios like these, The University of Edinburgh has implemented courses such as Software Engineering and Professional Practice (SEPP), and Professional Issues (PI). These courses cover varied areas of study. Most notably, PI covered 10 weeks of teaching under the headlines: Introduction, Responsibility, Power, Data, Bias and Fairness, Humans, Personal Attributes, Design Frameworks, Leadership, and Teaching Others. Furthermore, in 3rd year students also take the Systems Design Project (SDP) where, at the conclusion of the course, they

write a report reflecting upon their own project. This last piece of writing reflects the individuals contribution towards the project, as well as their own views on the impact of the project in the real world.

Therefore, in hopes of determining whether these courses profoundly influence students' concerns about the ethical impact of CS and AI, the essays written as part of the assessment for PI and SEPP will be used to develop a topic modelling technique which can be employed on the aforementioned SDP reports and determine if students have a better understanding of ethics in CS and AI. This technique, if successful, could also be used in future to analyse students writing in attempts to recognise a greater trend of student awareness of ethics in CS and AI.

## 1.2 The Problem Described

Understanding the work of written essays and more specifically the written word is a very large area of CS and can be covered with a variety of techniques in Natural Language Processing. The one which I have chosen to utilise is Topic Modelling, which is usually an unsupervised method, meaning that the data is fed into the topic model which returns the output - a series of topics which each contain the words that helped form that topic. It does this by clustering the input data to naturally form groups of items even without knowing what it is looking for directly [14].

There are many ways to perform topic modelling, as it is a well covered area in CS. However, there are still many new and emerging techniques to perform topic modelling and the particular problem of using it to determine if students are showing insight and concern of the ethics in CS is a new one and may pose some challenges:

- Data wrangling of the input data - there a number of reasons why this step is an important one in itself but the most important one is maintaining privacy by the removal of anything that can identify students.
- Finding a form of topic modelling which will conform to the particular inputs - this could prove difficult as the topic modelling method is often performed on a larger corpus.
- Analysing the outputs and interpreting them in a competent way - this step will be difficult and whilst it unfortunately won't be very technical, due diligence will be exercised in interpreting the results.

## 1.3 Contributions

- Develop a topic modelling method which can accurately pick out a topic which relates to ethics in any piece of written work.
- Ensure the method will have the final capability of determining if ethics has become an increasingly talked about concern in students' SDP reports from year to year.
- Accurately evaluate the data to draw useful conclusions, focusing on how the underlying method and external factors, like the course structure, may impact the final results.

## 1.4 Thesis Outline

This thesis is split up into 5 chapters. The first section, Chapter 1, covers the motivation and the direction of this project. Chapter 2 will cover much of the background knowledge, specifically of how to manage input data for Natural Language Processing and all of the background reading of prior research in Term Frequency-Inverse Document Frequency, Deep Learning, Topic Modelling and BERT, all of which will inspire this implementation. Chapter 3 will be a brief design summary of how this method will be implemented, along with the in-depth methodology involved in developing the necessary topic models. Chapter 4 will cover the results of trying to model student essays and reports to produce accurate topics, and evaluations of how the models performed, with conclusions of what was found and what that means for students understanding of ethics in CS, and the retention of discussions around ethics throughout their years at the university. Chapter 5 will cover the goals which were set and how they were achieved. It will also cover, the strengths and weaknesses of the project, as well as any further work or improvements to be made to the project.

# Chapter 2

# Background

## 2.1 NLP

The nature of this project means that Natural Language Processing (NLP) and its constructs must be used to evaluate and understand student essays. NLP is essentially the field in computer science where researchers are trying to match the ability of a computer's understanding of language, whether written or spoken, to a human level - or as close as it can be. There are a variety of methods under the belt of NLP, however as the inputs will be short essays and reports, and few of which that can be tested, the method which will be used is Topic Modelling. This is the most relevant branch of NLP to this specific case as it is unsupervised, meaning that the inputs do not need to be labelled in such a way for the algorithm to understand, instead it will find patterns in the data and group it together accordingly [19].

### 2.1.1 Tokenisation

Tokenisation is the process in NLP by which large pieces of text can be broken down into smaller parts which are referred to as tokens. This is used for a number of processes from traditional NLP methods to more advanced Deep Learning-based architectures like Transformers [6]. There are many different forms of tokenisation which can occur as shown by Figure 2.1, each breaking the text down in different ways. One example of this is sentence tokenisation. So in the case where there is a piece of text which reads ”She is funny. He is unfunny.”, the tokens would be ”She is funny” and ”He is unfunny”. Another form (and likely the most common) is word tokenisation. In the case where we have the sentence ”Edinburgh is beautiful”, this would be broken into the tokens ”Edinburgh”, ”is” and ”beautiful”. Lastly, another two common implementations are character and subword (or n-gram characters) tokenisation. So if there were the word ”larger”, the character tokens would look like: ”l”,”a”,”r”,”g”,”e”,”r”, and the subword tokens would look like ”large” and ”r”.There are a number of libraries which help with tokenisation, including NLTK, Genism, and Keras.


Figure 2.1: Tokenization techniques
[9]

### 2.1.2 Stop Words

In every portion of text there are stop words littered throughout. These are words which are so common in any language that they hold no weight to the content of the piece of text itself. For example in English these words would be things like ”in”, ”the”, and ”and”. Therefore words like this add nothing of value to the text and will not affect the processing of the text after removal, so are removed to reduce noise [9].

These stop words occur in every language and the way in which they are realised can be represented with Zipf's Law. It is a universal law used to describe the distribution of words throughout any given language and it is used frequently in statistics and spans across all languages. It states that most of the tokens (words) in a text are accounted for by very few high frequency tokens (like the stop words mentioned previously) and the rest are low-frequency words. It is easily represented in the simple formula:

$$f(r) \approx \frac{1}{r^a} \tag{2.1}$$

Where, $f$ is frequency, $r$ is rank, and $a$ is approximately 1 [26]. A simplistic view of this law, and how it applies to stop words, can be shown in Figure 2.2 shown below, alongside a real world example of Zipf's Law in the book *Ulysses* (2.3).

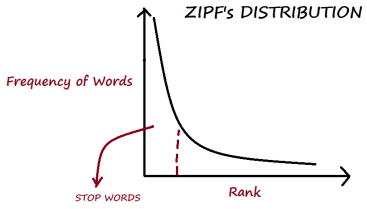


Figure 2.2: Sketch of Zipf's relation to stopwords
[22]

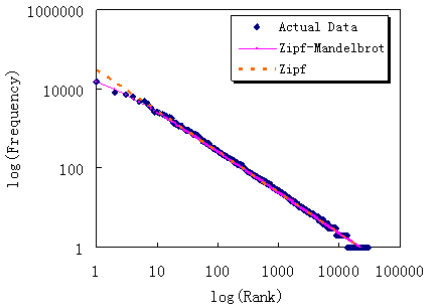


Figure 2.3: A log graph showing Zipf's Law in the book *Ulysses*
[31]

Furthermore research by Mandelbrot, and later consolidated by Piantadosi in 2014 [13], shows that the more accurate equation would be:

$$f(r) \approx \frac{1}{(r+B)^a} \tag{2.2}$$

This equation is far more precise when generalising to the distribution of words in language and changes the initial Zipf's Law by shifting the rank by B which is approximately equal to 2.7.

### 2.1.3 Part-of-Speech Tagging

Another technique used as part of NLP is Part-of-Speech (POS) tagging, in which each word in a piece of text is labelled in such a way that represents it. For example, the word "run" could be labelled as a verb or a noun depending on the context, which is one of the major challenges of POS tagging. Things like names of a person or organisation i.e. "Owen" and "Twitter" would be characterised by the labels of proper nouns. There are, of course, many different forms of tagging in which some tag sets have numerous detailed labels and others have fewer, more general labels. One such tagset with fewer labels is the Universal tagset with only 17 labels [2]. This is in contrast to the Penn Treebank tagset which has 36 POS tags, and 12 others which are used for other things like punctuation. In Python, POS tags can be used by importing the NLTK library which contains the Penn Treebank tagset [1].

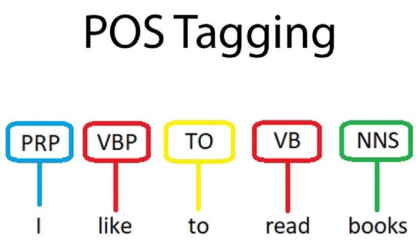


Figure 2.4: Part-of-Speech Tagging example (using Penn Treebank tags) [21]

Part-of-Speech tagging is generally used for things like text-to-speech translation or translation from one language to another. It is not something which is generally employed in topic modelling as it is an unsupervised method and therefore does not require labels to process the inputs. However, in this case it was exceptionally useful as discussed later in the implementation section.

### 2.1.4 Stemming and Lemmatisation

When working with text in NLP there is often the issue that multiple words (or tokens) which have the same, or very similar meaning, are represented as separate entities. This could be in the case of the word "run" having the same meaning as "running", but only

being different by a few characters, or the word ”good” and ”better” having similar semantic meanings but being two entirely different words.

There are two related, but different ways to deal with this. The first is ”stemming”, a process by which the stem of a word is created during tokenisation instead of the original word. For the previous examples of ”run” and ”running”, they would both be represented as ”run”. The other kind is ”lemmatisation” where both words are tokenised as a base word with the same meaning. For example, the words ”good” and ”better” as mentioned above, both have the same lemma of ”good”.

| *Word* | *Lemmatization* | *Stemming* |
|---|---|---|
| was | be | wa |
| studies | study | studi |
| studying | study | study |

Figure 2.5: Example of how the difference between lemmatisation and stemming can vary)

[10]

They both have different uses and they obviously give different results in some scenarios and the same in others. If we had used lemmatisation for ”run” and ”running” then the outcome would have been the same as the stemmer. However, had the stemmer been used on ”good” and ”better” it is likely that it would have produced the stems ”good” and ”bet”.

Stemming is a much quicker process than lemmatisation, however it works with no context and therefore produces stems which can be less than useful in many circumstances. Often, over-stemming can lead to an outcome of no context, where too much of the word is lost, and therefore leading to tokens with the same stems and vastly different meanings. Lemmatisation can be a lot slower than stemming but it uses the context of text around it to produce results.

These processes are quite important for topic modelling as there are many words with similar meanings or structures which, instead of being grouped into one topic and realistically all representing the same word, can be reduced to the same definition to allow for more accurate topic modelling.

## 2.2 Term Frequency and Inverse Document Frequency

Term Frequency and Inverse Document Frequency, together known as TF-IDF is a a form of frequency analysis performed on documents to determine how important particular terms are that show up throughout a given piece of writing [25]. TF is simply the act of weighting words in a document to determine how often different words appear [28]. IDF is given by the equation:

$$IDF = log\frac{N}{DF} \tag{2.3}$$

Where, $N$ is equal to the total number of documents, and $DF$ is the number of documents in the collection that contain a given term [27].

However, since TF-IDF there have been a variety of developments building on top of this initial research. Most notably, in 2017 researchers used the Google Word2Vec model to build on top of Term Frequency-Keyword Active Index [15]. Firstly, it was thought that looking for high frequency words are very general and are not helpful when trying to distinguish between boundaries of knowledge in a given domain. In spite of that, it was proposed that low frequency words would in fact be better because they can show knowledge of new and different emerging areas within a domain.

It was this idea that demanded the creation of TF-KAI. The difference was that instead of using full text documents as the targets for analysis, as was traditionally done, keyword lists were used instead. As keywords are more spread out in the list than they are in the target documents, there would be less probability of a keyword showing up in a background document. Similarly to before, keywords are ranked in terms of frequency, which would determine its importance in a given area. However, words or terms can have synonyms and be written in different ways. Therefore, words which were semantically the same were combined before the evaluation stage using Google's Word2Vec model (a neural network model for learning word associations) [5]. This model was specifically used to find the context each word was used in to reach the semantic meaning of all words before merging all semantically similar words. In order to join the collection a particular threshold of similarity between a keyword wishing to join and each keyword within the collection had to be passed. The researchers extended their use of semantics to cover each stage of TF-KAI, so that it would cover both the popularity (counting the number of times a term shows up) and discrimination (calculating the importance of given terms) stages to produce Semantic Frequency - Semantic Active Index.

The original TF-IDF is still relevant today though. As recent as a paper published in 2023, shows that it has been used to detect malware. The project was based on the fact that Android is the most used operating system for phones currently on the market and that with more and more applications available, there is a need now more than ever, for effective malware detection [24]. The current forms of analysis have a few problems, one of which being a large amount of processing power. The researchers used TF-IDF for feature extraction, along with Extreme Gradient Boosting (a library in Python for gradient boosted decision trees) for classification. The results were astounding, classifying over 99% of given packages correctly in seconds.

## 2.3 Topic Modelling

Topic Modelling is the process by which a piece of writing is split into a series of topics, each of which contains a series of words. The process has been around for a while and traditionally uses LDA, which stands for Latent Dirichlet Allocation, to generate the topics [17].

However, topic modelling is very much a quantity driven approach to examine what topics a section of writing covers, not the quality of the writing itself. Furthermore, researchers proposed that there are two problems with using LDA for examining the quality of writing: the first is that there hasn't been a metric that is able to match the

judgement of a human in understanding the difference between topics; the second being that there may be a need to highlight particular topics that a person is specifically looking for so that a qualitative study would be able to search for it [29].

To combat the first problem, the researchers attempted to use a new metric, TF-IDF coherence as opposed to the previous regular coherence metric. This was because they hypothesised that it would be closer to human judgment. In response to the second problem, instead of using the unsupervised LDA, they decided upon an internal semi-supervised method which was dubbed ISLDA. It used a given set of keywords which were restricted to particular boundaries of topic assignments. These keywords being the actual topics that they researchers were looking for. The researchers found the ISLDA was better than LDA for topic modelling, when using the the TF-IDF coherence metric.

Due to the fact that topic modelling is mainly an unsupervised approach, there have been many cases where this seeded/guided approach has been used for a number or applications. One such utilisation, from a 2021 paper, was to use tweets to retrieve live situational awareness data [33]. This was essentially an implementation to determine what underlying topics civilians were tweeting about during things like natural disasters. The implementation used a form of guided LDA to identify emerging topics and was used to analyse tweets during Hurricane Laura, in 2020. The resulting technology should enable live and relevant information to responders in aiding response and resource allocation for future disasters.

## 2.4 Deep Learning

As BERTopic uses Deep Learning it is important to describe what it is and how it is relevant to BERTopic. Essentially, deep learning is a branch of machine learning which consists of a neural network of 3 or more layers. A neural network is essentially attempting to replicate the human brain in the same way that neurons pass signals to each other, just to a much simpler effect. It has an input layer, one or more hidden layers, and an output layer, all containing a certain number of nodes, each with their own weights and thresholds. It works by taking inputs and calculating a new output at each layer, using the respective weight. If the output is above the given threshold then the outputs may pass to the next node as the input and so on. Generally, deep learning models need a lot of labelled data to be able perform at a high level due to the number of parameters. They also require far more computational resources than the average machine learning algorithm. Fortunately, deep learning allows for a lot of success in areas like image recognition and speech recognition, and beneficially for this task, it is also very useful in natural language processing. There are a number of pre-trained deep learning NLP models available for use nowadays including BERT, RoBERTa, GPT, TensorFlow, and PyTorch, and many more libraries which incorporate or build on top of these.

Due to the rise in areas like AI, deep learning has been and will continue to be an increasingly researched part of CS. Two papers published in 2023 cover two very important and similar uses of deep learning. The first analyses the use of deep learning to diagnose COVID-19 [7]. It looked at using convolution neural networks (CNNs) to

detect COVID-19 in CXR images (chest x-rays) amongst other applications. The other paper used a series of deep learning techniques to detect cervical cancer [16]. There were many deep learning algorithms involved, including: CNNs, Multi-layer Perceptron (MLP), and Artificial Neural Networks (ANNs). These were used to categorise the already extracted features into a positive or negative diagnosis.

Beyond this, deep learning can prove to be a valuable new component of topic modelling and has even been categorised as an increasingly popular research area called neural topic models [32]. One paper in 2019, attempted to use topic modelling with deep learning to forecast oil prices [18]. This was done by using a CNN to extract features from online media resources and then used LDA to topic model the given features. Whilst being rather different from previous examples this proved itself to be a beneficial use of topic modelling, as the developed forecast performed better than the older benchmark models. Moreover, similar to the previously mentioned topic modelling of tweets surrounding the Hurricane Laura crisis, a paper from 2022 covered a framework, dubbed Topic2Labels, which utilises deep learning and LDA for crisis response [30]. It works with three layers: the first an LDA layer which generates the topics, as done in traditional topic modelling - in combination with a new algorithm to rank topics and annotate data, a second layer which utilises BERT embeddings to create feature representations of the the newly labelled text, and a third layer which uses other deep learning models to classify the data. This implementation also proved to beat other more traditional approaches.

## 2.5 BERT and BERTopic

### 2.5.1 Introduction to BERT

One of the best libraries for topic modelling is BERTopic due to the use of the deep learning model BERT, which stands for Bidirectional Encoder Representations from Transformers, and was developed by Google. Up until its development different NLP tasks had normally been solved by different models, however, BERT was the first to be proficient in solving all of them. Examples of these problems are text prediction, summarisation, and sentiment analysis.

### 2.5.2 Why BERTopic

There are a number of more traditional ways to do topic modelling, which include Non-negative Matrix Factorization (NMF), Latent Semantic Analyis, and as mentioned previously, the very common Latent Dirichlet Allocation (LDA). However, due to BERTopics use of BERT, it has the edge when it comes to comprehensive topic modelling. This is due to a number of reasons, yet the most important one for our applications is that it is very robust when it comes to working on short texts. Traditional techniques tend to cope less when it comes to shorter and nosier texts, but with BERTopics use of transformers it is far more capable and reliable at producing relevant topics across the board.

BERTopic also has a far better semantic understanding of texts due to BERT's high

number of document embeddings , allowing it to understand text better than traditional techniques and produce a model with better topic coherence. There are a number of other benefits over traditional techniques like: dynamic topic reduction, language agnosticism, unsupervised fine-tuning, visualisation support, customisability, and easy integration with existing pipelines. Many of these are in the area of ease of use which will become very beneficial when it comes time to evaluate and visualise the topic models.

There have been a number of different applications that researchers have utilised BERTopic for, since it became available in 2020. One of which was conducted in 2021, where researchers attempted topic modelling Arabic with 3 different topic modelling methods: the traditional LDA and NMF methods, against the new and more modern BERTopic [4]. They found that BERTopic showed better results overall but concluded that ”finding an accurate measure for evaluating the quality of generated topics is still challenging and it needs further research.” Furthermore, another study in 2022 used BERTopic to model reviews from the websites Ctrip and TripAdvisor [11]. The intention was understanding the emotions of consumers when interacting with robots in aims to understand the likelihood of a consumer adopting a service robot. They concluded that generally customers were positive about interactions with robots, showing emotions like joy, excitement, and even interest. This was however, contrasted by feelings of discontent when a service robot was unusable on account of it crashing. Lastly, there have been many studies using tweets as inputs, including a more recent use paper from a study in 2023, which used BERTopic to look at how the public perceived ”healthy ageing” over the last decade [23]. It found a number of topics which matched or came close to the World Health Organisations definition, covered by their own healthy ageing initiative. However, BERTopic picked up on other health topics which were widely talked about through tweets. These included, skin appearances and beauty in ageing, something not covered in WHO’s definition. The researchers proposed that these should be talked about in the decade going forward.

# Chapter 3

# Methodology

## 3.1 Data

- 20 newsgroups dataset - A very large dataset of 18,000 documents. Used to gain an understanding of how BERTopic functioned before performing topic modelling on student writing.
- SEPP and PI essays - A very small dataset of 18 documents. Beneficial in gaining an understanding of how students write, the sort of data wrangling that would have to be performed on the later reports, and what versions of BERTopic performed best with further experimentation on student writing.
- SDP Reports from 2021 and 2022. The final dataset used for evaluation. About 150 documents per year so topic modelling would be easier for this large of a corpus.

## 3.2 Design

The goal of this project is to gather an understanding of the changes between student perception of ethics, as it is applied to CS and AI, between year groups. Due to this, several implementations of testing and interpretation were needed before the final implementation. Therefore, this project has been implemented in various stages.

The first part is thoroughly testing and improving different inputs for BERTopic to ensure that the results are finding the most truthful semantic understanding of what is contained within the student essays. This means that when topic modelling takes place there are the fewest possible occurrences of topics formed around impractical words, or words which may be semantically similar or lexically similar.

Secondly, a refined version of topic modelling with BERTopic will be implemented on SDP reports between different year groups. This should provide a sufficient output to be able to analyse whether people from one year group to the next have a greater care, and are more considerate, about ethics as they are taught more about it. If the outputs are not sufficient to analysing students understanding of ethics to their own works, or

they produce an outcome which is very unexpected, then the models will be refined further before running them again with the same data.

The final result of this project will be a series of topic models which can be visualised. These models will be followed by an evaluation to determine if the models accurately depict the difference in students' focus on "ethics" in their SDP reports.

This project was written in Python throughout a number of Jupyter notebooks, using Google Collaborate. A virtual environment was used for a number of reasons. Firstly, it is very easy to import and track the many libraries which were needed to accomplish this task. Secondly, due to privacy concerns the SDP reports could only be seen by the project supervisor, so a method was devised to allow easy execution of the code. This was done by creating a notebook which simply needed the corpus to be uploaded, and the code to run from start to finish, for the models to be created.

## 3.3 Using BERTopic to Topic Model PI and SEPP Essays

### 3.3.1 Training Data

With the final goal of topic modelling ethics content between year groups' SDP reports, initial testing began with other materials which were available. Initially, when experimenting with BERTopic, the widely used 20 newgroups dataset (an amassment of 18,000 newsgroups documents from 20 different newsgroups) was employed to understand the basics of topic modelling with BERT, before moving on to some more relevant and similar data. Due to the fact that the SDP reports were not available directly because of privacy concerns, essays from the Professional Issues and Software Engineering and Professional Practice courses were used to test different iterations of BERTopic. In total there were eight PI essays and ten SEPP essays. While this may seem like a small number of papers for testing different forms of BERT, it was significant for understanding how to process the data to provide a comprehensible input to BERTopic. Moreover, training BERTopic on these students' writing style made sense given that the same students would go on to write SDP reports.

### 3.3.2 Data Wrangling

A significant issue when attempting to analyse any text using any of the areas of NLP, is that there are often words and phrases that can skew the data in a way that makes it difficult to evaluate. In fact, one of the main and biggest examples of this can be represented when looking at Zipf's Law, which is how words are distributed throughout a language. To take the English language as an example, the words "the" and "in" occur far more frequently than the words "xylophone" or "xenophobia". This can be represented by Zipf's Law in all languages and a lot of the words like the former can show up very frequently throughout text, and in this case skew the data.

Thankfully, many of these words have been grouped together under the library in NLTK's stopwords. The term "stopwords" is used to describe a set of commonly used words in a language. A library of stopwords is useful in this application as it enables the

removal of such words to be able to focus on what is actually important. An everyday example of this would be a search engine removing such words to focus on what actually matters in a search query. So if there were a query ”Why do I have to pay taxes?”, the words ”I”, ”do” and ”to” are all likely to be removed for the sake of the words that actually matter for finding webpages that match the query [12].

Furthermore, there were often specific words within the example essays which had to be analysed that were skewing the data quite a bit. This included: terms which tended to appear in introductory paragraphs and caused whole topics to be formed on the basis of words like ”essay” and ”question”; words that often occurred just during referencing, like ”bibliography” and ”accessed”; and months of the year also tended to produce entire topics.

Moreover, the words had to be tokenised, which is fortunately one of the capabilities which NLTK already has built in. This is the process by which a larger piece of text is broken down into it's individual parts (being words/character) to be used as input data. This was an important step in this case as it allowed for a couple of checks on each token to ensure that it could be used as an input. One of these checks, as already mentioned, was checking that the token isn't a stop word, if it was then the token was then removed. There was also a check to ensure that the token was actually a word in English before it could be used as input data. This provided a much needed check to prevent random, reoccurring strings from becoming their own topic. Moreover, the words were lemmatised so that similar words would become the same and allow for more accurate topics to be built, as opposed to topics containing what is essentially the same word. As mentioned in Chapter 2 there was a chance that stemming could have been used, however that would have greatly affected the output when evaluating as there may have been many words grouped together which do not share similar meanings.

After this, the tokens are joined back together into larger strings to be used as inputs. It is worth noting that BERTopic takes in an array of strings which then has a topic found for each instance in the array. As a consequence, a decision had to be made as to how much text would be used to determine the topics within a document. A common approach to this is to have each instance be a paragraph from the corpus. However, in this case this proved problematic due to the fact that the corpus was so small. Therefore, the decision was made that each instance of the array would be equal to one sentence. Lastly, any empty instances in the array were removed. This was causing topics to form which contained nothing and was caused by the heavy data wrangling which just took place, most likely produced by the removal of something like a hyperlink or reference in the bibliography.

### 3.3.3 Methods and Implementation

#### 3.3.3.1 BERTopic Flexibility

There were a number of possible ways to find a suitable implementation which would achieve the goals. The choice was made to use BERTopic due to how flexible it was to the specific needs of the project. It also has many inbuilt functions which can be used to display the final output in a number of ways, which would prove to be useful for when

it came to evaluate results, and also when it was decided which kind of topic modelling techniques would be brought forward to evaluate the content of the SDP reports. The flexibility of BERTopic meant that there were several rounds of experimentation and implementation prior to the finalised versions.

Firstly, there are a number of inputs which BERTopic can take. Despite this, the default version of BERTopic returns successful results for topic modelling without having to be modified. There are 4 vital components in BERTopic. These are: a transformer embedding model; a UMAP model for dimensionality reduction; an HBDSCSN model for clustering; and the use of c-TF-IDF for tagging the aforementioned clusters.

By default BERTopic uses a sentence transformer ”all-MiniLM-L6-v2”, as it is a high performing transformer and therefore the most commonly used. It works excellent for text written in English, and produces 384-dimensional sentence embeddings. The only downside to this model is that it doesn't work well for documents written in other languages or multi-lingual texts, but of course this shouldn't affect any of the essays/reports in use.

UMAP (Uniform Manifold Approximation and Projection) is a technique for dimensionality reduction, which is useful for a number of reasons. Mainly, it means that any irrelevant features of the data are removed, thus increasing the accuracy and the speed of the model. It also means that it is much easier to visualise the data as it has been brought down to lower dimensions like 2D and 3D [3]. Despite UMAP having been created and adopted increasingly due to its advantages when compared to older models like t-SNE and PCA, it still has disadvantages. Such that, UMAP has a stochastic nature, meaning that there is an element of chance to it and that it can produce varying results on reruns. It is worth noting that BERTopic was ran with a UMAP model many times and every time it produced the same result, and was able to pull out a topic containing ”ethics”. Despite this, it may mean that on reruns of something like SDP reports, it could produce different topics or even a different structure/relationship between topics.

Furthermore, once the dimensions of the model have been reduced to either 2D or 3D, HDBSCAN is employed to cluster the existing vectors. There are a number of ways in which the data can be clustered together, varying from hierarchical techniques to centroid based techniques. When looking at flat or hierarchical methods it is simply focusing on whether there is a hierarchical structure in the clusters so that we can view it and attempt to evaluate the branches along the tree. Methods such as centroid and density-based clustering are based on proximity from a point to a centroid/central point or the density in proximity between a number of points. They both tackle separate cases, so if the clustering looks more circular or spherical then centroid-based clustering would be used, or if it is a more unnatural shape or outliers need to be found then density-based clustering should be used.

Finally, the topics are found using c-TF-IDF, a modified form of the previously discussed TF-IDF, which will be used to label the clusters. The process of TF-IDF normally identifies which pieces of text, when there is a large amount, are the most applicable to the given array of terms. Instead c-TF-IDF uses the clusters which we have created to find what the most relevant labels are for that given cluster.

#### 3.3.3.2 Guided/Seeded LDA

After implementing one form of successful topic modelling with BERT the conclusion was reached that it was time to implement a form of seeded/guided LDA as mentioned in the literature review. An initial working version did not include one of the most successful outputs that recent research discussed. The GuidedLDA library would however accomplish this, as it was inspired from such papers. The plan for integration was to use this as the embedding model for BERTopic but unfortunately despite all of the steps that were taken to achieve this, there were significant errors and the integration was not successful. This meant that a guided topic model which used LDA could not be implemented.

#### 3.3.3.3 Guided/Seeded BERTopic

Due to the lack of cross functionality between GuidedLDA and BERTopic another way of implementing a seeded topic model had to be found. What was discovered was that BERTopic can actually take in a list of seeded topics in the form of a list of lists, where the inner lists are each a collection of words belonging to one topic. This is not something that it widely used, possibly because BERTopic is typically effective at finding topics without the use of seeded words. Another possible reason is that because it is a more recent form of topic modelling, it has not been widely adopted as of yet. Other arguments could be made that it is due to the fact that it produces a skewed representation of the topics if certain topics have to be highlighted so that they can be built around or even noticed at all.

The method which was used to produce a topic based around ethics was to feed it a comprehensive list containing words in the semantic field of ”ethics”:

*”ethics”, ”ethic”,”ethical”, ”fairness”, ”integrity”, ”moral”, ”morals”, ”morality”, ”principle”, ”virtue”, ”virtues”, ”responsibility”, ”harm”, ”safety”, ”safe”, ”consent”, ”equality”, ”privacy”*

A method with these variants was devised so that a topic would emerge stronger than previously and drew up everything that may relate to ethics in the students essays. Afterwards, another seeded topic was created which was meant to separate the context of the student essays. These were words relating to the cases involved like:

*”drone”, ”delivery”, ”order”, ”area”, ”pilot”, ”aircraft”, ”plane”, ”aviation”, ”flight”*

These words reoccurred a lot as the student essays were covering drone delivery and a case study involving planes, among other case studies.

A shortcoming of this implementation is that whilst it is guided topic modelling, it is not guided LDA as discussed in the literature. The reason it was so important in the literature is because it was attempting to build topics based on the quality of the what had been written and not based on the quantity. Whilst in theory the seeded aspect gives keywords and therefore topics a head start, in practice they are not being built the same as BERTopic is not using LDA. Arguably, the use of BERTopic, which involves deep

learning, means that the embedding model has a better understanding of the text, and therefore the model is focused even more on the quality of the topics as opposed to the quantity of words involved.

#### 3.3.3.4 BERTopic Embedding Model

One thing that was attempted was to change the embedding model (the model which translates high dimensional data, like words represented by vectors, into low dimensions) which BERTopic uses. As mentioned previously the default model for BERTopic is the ”all-MiniLM-L6-v2” which is generally high performing for the English language. However, it was discovered that there was an across-the-board higher achieving model called ”all-mpnet-base-v2” which was implemented, whilst still using seeded topics. Unfortunately, this caused the topics produced to be far less inline with what was being searched for. This is possibly because, despite the embedding model being extremely high performing, it needs far more data to perform accurate topic modelling.

#### 3.3.3.5 BERTopic HDBSCAN Models

There were also attempts to change the HDBSCAN models that BERTopic used, in attempts to cluster the data in different ways and reduce topic outliers. Firstly, an attempt was made to change the model into a k-means model, choosing a variety of inputs for how many clusters there should be. K-means is a much simpler algorithm than the DBSCAN model as it simply tries to find the minimum euclidean distance between points (which is simply the length between two points in the given euclidean vector space) with the goal of minimising the distance within a cluster and maximising the distance between clusters. It began with a default of 50 clusters which proved to be far too many as each topic was split into very niche and specific words. Then, 10 clusters were tried because it was thought that would split the topics up nicely, however, it was later realised that this overgeneralised the topics to make them very high level areas, not relevant to what was being searched for. In the end, understanding that previously topic modelling had given around twenty-something topics, the decision was made that there should be about 25 clusters for the data to be drawn to. This performed better than changing the embedding model but did not perform as well as the default BERTopic or simply giving BERTopic seeded topics.

Secondly, an attempt was made to change the HDBSCAN model. By default, in BERTopic the min_samples variable and the min_cluster_size variable are equal parameters in HDBSCAN. However, if you are to reduce the min_samples to be less than min_cluster_size then the number of outliers can be reduced, resulting in less noise [20]. So, the minimum samples size was reduced to be 5 whilst keeping the minimum cluster size to be 10. However, this proved to be the worst performing out of all methods. It is possible that instead of just removing outliers it also removed much needed data to perform the appropriate topic modelling.

#### 3.3.3.6 BERTopic c-TF-IDF Model

A c-TF-IDF model was also implemented with the seeded topic list. This is done so that words that are frequent have a lesser impact than they would without this model. A

similar model that could have been implemented was a vectorizer model which would have removed any stop-words. This step has already been done in the pre-processing, but an upside to the c-TF-IDF model is that it also does it automatically, meaning that any missed stopwords will be removed. The default c-TF-IDF model in BERTopic already does this by default, however it was activated to reduce other words of high frequency. This hopefully facilitates a topic model that is looking more at the quality of the text as opposed to the quantity of words within it. It also follows one of the parts mentioned in the literature review which was the use of c-TF-IDF.

## 3.4 Using BERTopic to Topic Model SDP Reports

### 3.4.1 Privacy Concerns

A major throughline of this project is ethics, and not only how other students practice it, but how it is applied here too. With this in mind there was a concern around privacy and how students work is displayed in this dissertation.

Of course the nature of this essay is surrounded around topic modelling, so only recurring words and themes should make themselves evident. However, to ensure nothing went amiss the use of part-of-speech tagging was employed to implement the sub-branch of NLP, Named-Entity Recognition (NER), and remove anything that may make a student identifiable.

By using the Python library ”spacy” the text could be tokenised before performing any of the other pre-processing steps, and identify proper nouns, like names and organisations. Once the format of the token had been found there was a check to see if the token was tagged as ”<PERSON>”. If this were the case the essay was built back up with the tag ”<PERSON>” in place of a name. Using the Python library ”re” regular expressions could be used to easily search for these tags and remove them.

Along with this some students had also added there student numbers, another identifiable feature. Again, with use of the Python library ”re” the regular expression ”s_d+” could be used to remove portions of text starting with ”s” that are directly followed by a series of numbers. This final step in enforcing privacy, ensured that students very no longer identifiable at all through explicit identifiers like names and student numbers.

### 3.4.2 Pre-processing

The pre-processing steps for the final implementation on the SDP reports followed the same format as that done previously. This included:

- Removing student ID’s from the corpus using regular expressions and POS tagging names and removing them so that their is no identifiable information for the students which wrote the reports.
- Removing hyperlinks, number, acronyms, and series of random characters and numbers from the text to reduce the chance of topics being formed around reoccurring collections of characters like these that have no significant meaning.
- Removing words that generally show up in introductory paragraphs, likes ”essay” and ”question”, words which show up in a bibliography, like ”accessed” and ”references”, and months as they show up very frequently in most pieces of writing.
- Removing stopwords like ”and” and ”the” to prevent topics being built around common but unconducive words like these.
- Finally removing any cells in the corpus array which may be empty due to the pruning of the text.

### 3.4.3 First Run

#### 3.4.3.1 Methods and Implementation

As mentioned in previous sections, there was a lot of experimentation done with a variety of different models, which produced outputs which were perceivably better than others on a very small data set. Due to the size of the data set being so small compared to the final dataset (the SDP reports) when it came time to implementing the groundworks for the new models, what was know to work from the previous testing was used. The decision was also made not to exclude what didn't work as well as it may produce a different outcome on a larger dataset. This meant that the topic models were built in a specific order, first using what was known to work, and then adding other parts to the model to see if it produced an output which may be better:

Firstly, a default topic model was implemented with none of the fields changed. This meant that there was no fine-tuning involved and anything like UMAP, and HDBSCAN models were left alone. This was done because when it was tested on the smaller set it had produced an outcome which was hard to improve, which suggested that it could also perform just as well, or perhaps better, on a much larger corpus.

The next model and the models going forward were set with a fixed UMAP model with the random_state set to 42. This was due to the aforementioned stochastic nature of UMAP, so to prevent models from having a different UMAP model, which would make them hard to compare, they would all be built the same from the ground up. Along with this seeded terms were introduced to produce a seeded/guided topic model. The model was only seeded with two topics, the first being the ethics topic which was seeded with the PI and SEPP essays, and contained identical words to what was used before. Similarly, to the models for those essays, a topic was created which would attempt to separate the context of the reports away from topics which clustered towards ethics:

*"team", "product", "task", "subteam", "management", "industry", "demo", "software", "hardware", "project"*

This model would hopefully draw a topic about "ethics" out of the corpus so that models between the two years could be compared.

Next the models were given a modified c-TF-IDF model to reduce the effects of stopwords whilst also reducing the impact of frequently used words in this specific corpus. This would hopefully have the effect of the seeded topics whilst also removing any bias given to overused or very frequent words or phrases. The most promising part of this would be that the model is focusing even less on just the frequency of the words and is more focused on the relations between them and the importance within the context to build accurate topics.

The MPNet embedding model may have proven useless for the smaller data set but due to the fact that it should be the better embedding model when compared to the default, it was employed here. The hopes were that with a much larger corpus the embedding model will be far better and understanding the corpus as a whole, and therefore be able to extract much more accurate topics.

Lastly, on top of all this, a k-means clustering model was implemented. Again, on the smaller dataset this did not seem to improve the results of the model but it may prove more useful with far more data. It is a possibility that with far more data there will be far more data points to cluster together, and with this being built upon MPNet, it is likely that this should produce more accurate clusters. There was not much in terms of guidance on how many clusters there should be so an arbitrary number of 25 was set.

### 3.4.4 Second Run

#### 3.4.4.1 Preface

Unfortunately, the aforementioned method produced models which, whilst a evaluation could be drawn from them, proved to be somewhat ambiguous. Therefore, it was decided that models would be built which could be compared directly to each other to confirm what worked and to compare the two corpus' more accurately. The benefit of this method is that it produced more models which built on previous models that included parts which were known to have worked whilst also being comparable to each other.

#### 3.4.4.2 Methods and Implementation

Instead of starting with the default model this time, the seeded model was known to work and would prove to be helpful. For this reason the implementation started with same seeded topics as last time. The UMAP model was also locked to the random_state 42 again, as this meant that the topics would all be directly comparable. The difference this time however, was that two seeded models were built, with slight differences. Instead, of using k-means clustering, BERTopics in-built nr_topics would be used as input to decide how many topics each model should have. The decision was made that the first two models should have 25 topics and 50 topics respectively. This would hopefully mean that, if ethics didn't show in the first model because the topic wasn't large enough, it would perhaps show in the second model.

Furthermore, the next stage was to keep the seeded models and introduce MPNet to both. Based on what had been seen from the previous run MPNet did have a much greater understanding when it came to the larger corpus and so it was hypothesised that it would be beneficial to test it again, with the same caveat as last time, using nr_topics to produce a model with 25 topics and a model with 50.

Lastly, the decision was made to use the clustering method which was known to work well from last time, k-means. Firstly, a k-means model with 25 clusters was produced, followed by a topic model with just seeded topics. Following this the same was done but this time MPNet was added. This would allow MPNet to be compared directly and see if the embedding model did in fact improve the topic model, when using k-means clustering.

In the chance that this would not produce valuable results the decision was made to create a k-means model with 50 clusters. This was of course for the same reason as before, in case at even this point the embedding model was not able to pick up on

an ethics topic. Again, created two similar seeded models, one with MPNet and one without, so that there would be a direct comparison.

### 3.4.5 Evaluation Method

The evaluation method for this project has proven to be difficult. Topic modelling as a whole is difficult to gauge as it is quite ambiguous and, being inherently word based, it is difficult to evaluate mathematically. Normally, models from different topic modelling techniques can be compared to show which one has proven to be more powerful. However, the goal was to find a method to compare two corpus' and use it to evaluate those corpus', not to evaluate the method specifically, as much as it is also important to have a sufficient method.

For this reason, the topics will be compared by a number of standards, which are not mathematically heavy, but combined should be reason enough to judge whether ethics has had an increase in students attention, as proven by how much it is contained within their own SDP reports.

The evaluation will contain the following criteria:

- A simple check to see if ”ethics” as a topic can be found in any form within a model
- In the topics which can be perceived as an ”ethics” topic, what are the surrounding words which make up the topic? What do these say about the topic itself?
- How many of the models in each year contain an ”ethics” topic
- Which topics surrounded or are clustered with the ”ethics” topic? What does this say about the topic in relation to others?

# Chapter 4

# Results

## 4.1 Preface

All of the intertopic distance graphs are available in the appendix. However, many are also provided throughout this section to highlight key points. Due to dimensionality reduction, on the smaller corpus' many of the graphs when visualised look different each time that they are visualised.

## 4.2 PI and SEPP Essays

There are many ways to visualise results with BERTopic. Most notably we can produce a graph showing the intertopic distance, visualise the hierarchy of topics, as well as show visualise all of the topics with the words which they contain. There is also get_topic(x), where x is the topic number. This allows us to view each topic and see which words are contained within. Topic -1 are the outliers which have not been assigned to the other topics, topic 0 is the topic which is the largest and most relevant, and each topic after that are the largest topics, decreasing in size as the topic number increases.

Something worth noting is that, as mentioned previously, a weakness with these particular models has been that when topic modelling the outcome wasn't always the same. This unfortunately means that when talking about how these models have performed, and showing visualisations to accompany these points, this is not how they will always perform and it should be said that performance can vary. However, these models were ran many times and it was understood how each of them tend to compare with each other.

### 4.2.1 Default BERTopic

Generally, this was one of the highest performing models. All of the default settings and models within BERTopic, combined with the deep learning aspect, mean that it is very high performing in modelling the topics already.


Figure 4.1: Intertopic distance map for default BERTopic, topic 2 is in the bottom left

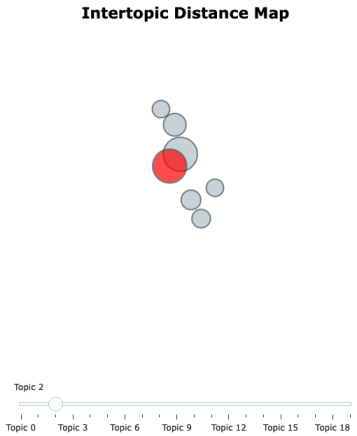


Figure 4.2: Intertopic distance map for default BERTopic focusing on topic 2 in red, it is in contact with topic 1 (”drone”, ”delivery”, ”order”

Figure 4.1 shows topic 2, which contains words like ”harm”, ”ethical”, and ”unintentional”, and topic 1, with words like ”drone”, ”delivery”, and ”order” in the bottom left corner. Whilst topic 0, with words like ”data”, ”personal”, and ”access” is grouped with topics in the centre right. It is worth noting that on different runs topic 1 and topic 2 swapped places quite often. Figure 4.2 shows the topics overlapping slightly, which changes depending on how zoomed in we are, but it shows that much of what is in the topics is very similar. They are obviously quite far removed from topic 0 but due to the fact that they overlap, perhaps seeding both topics would mean that they could be separated more.

### 4.2.2 UMAP Going Forward

As mentioned previously, UMAP is of a stochastic nature meaning that every time BERTopic is run it can perform differently. Fortunately, UMAP has a random_seed input which can be changed so that all models are processed the same way. This is what will be used for other models aside from the default model. There is no benefit to picking any particular number so the input for the random seed was chosen to be 42.

### 4.2.3 Seeded/Guided BERTopic

When testing BERTopic with seeded words and a fixed UMAP for dimensionality reduction a topic was produced containing ”ethical” as the second largest topic. This is similar to the default model shown, in Figure 4.1, for this specific run, as it was also had the ”ethics” topic as topic number two in that model.

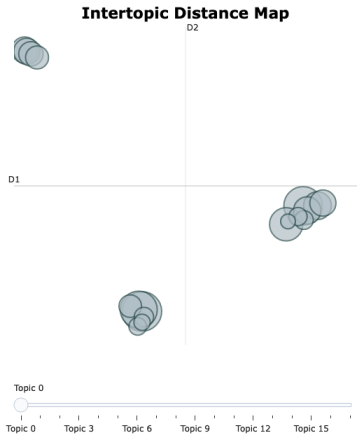


Figure 4.3: Intertopic distance map for BERTopic with seeded words, topic number 2 is in the top left

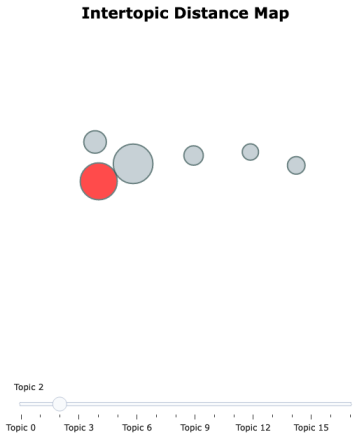


Figure 4.4: Intertopic distance map for seeded BERTopic focusing on topic 2 in red, which is contacting topic 1 (”drone”, ”delivery”, ”order”)

Figure 4.3 shows once again that the topic of ethics is closely clustered with many other topics and Figure 4.4 shows that, just as in the last model, the drone and ethics topics are once again very closely related.

### 4.2.4 Seeded/Guided BERTopic using c-TF-IDF

One of the models which gave an example of how different the outcome could be was exemplified in the model using c-TF-IDF to reduce the impact of frequent words. On the latest run of models, the outcome for this model wasn’t great, it did produce an ”ethics” topic but it was unfortunately topic 15. This is unlike the previous run where it was found that it performed very much the same as some of the other models, in which case the default model had an ”ethics” topic of 1 whilst this model had an ”ethics” topic of 3.

On the previous run, some of the words within the topic had changed. For example, the words ”responsibility”, ”impact”, ”public”, and ”social” were all lost from the topic and replaced with words like ”profit”, ”knowledge”, ”work”, and ”community”. Intuitively, it looks like most of the words that were lost in the last model, which look as though they encapsulate ethics as they relate to the aspect of people, are replaced with words which seem like they are mostly derived from the corporate world. This is possibly because in attempts to reduce the impact of word frequency, the model is assigning more importance to areas which don’t relate to ethics, or it is looking at how they connect in a different light, say from a corporate stand point.


Figure 4.5: Intertopic distance map for seeded BERTopic using c-TF-IDF on previous run


Figure 4.6: Intertopic distance map for c-TF-IDF BERTopic from previous run, focusing on topics 5, 2, 11, 1, 9, 15 from top to bottom


Figure 4.7: Intertopic distance map for seeded BERTopic using c-TF-IDF on latest run, topic 15 is in the cluster closest to the center

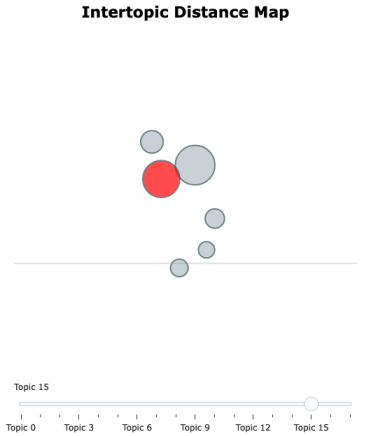


Figure 4.8: Intertopic distance map for c-TF-IDF BERTopic, from latest run, focusing on topic 15 in red, in contact with topic 4, the ”drone” topic

The figures above show that despite the fact that on different runs, the stochastic nature of the model may have produced ”ethics” topics of different sizes, however much of the model has remained the same. When comparing Figures 4.5 and 4.7 we can see that there are roughly three main clusters of topics, although the first does seem a little more spread out. When comparing Figures 4.6 and 4.8 we can also see that the groups of topic surrounding the ”ethics” topic that we are looking for all remain the same, despite small differences in distance.

### 4.2.5 Other Models

Changing the embedding model to the, supposedly better, MPNet model did not prove fruitful. Likewise, changing the HDBSCAN model to reduce the minimum sample size to be lower than the minimum cluster size didn't help. The HDBSCAN model was also changed to use a k-means clustering model, which did not improve performance. Often times these methods actually decreased the performance of the model. The only one which came close to being an improvement was using k-means. However, it was quite unreliable and often when deciding the number of clusters to use, the number of topics that the default model created had to be checked and use that to determine the number of clusters, using a standard deviation of 5 to give the method some leeway.

### 4.2.6 Conclusions and Comparisons

It is clear that by default, due to all of the parts that go into making up BERTopic, like the embedding models and transformers, the initial model is very capable of topic modelling this specific case without any extra tweaking. However, what did become clear was that on reruns it often managed to change the order of the topics and therefore the sizes of the topics, and which words they contained. One of the best improvements to the model was giving it a fixed seed, which meant that other implementations were able to be compared without the stochastic nature of the default model. Unfortunately, because the corpus was so small it did mean that on reruns some of the models could come out drastically different even with a fixed UMAP, as made clear by Figures 4.5 through 4.8. Once the seeded terms were added to the model (with a fixed state in UMAP) the second largest topic was found to be the "ethical" topic and it shared many of the terms with the default model, however, some were slightly different. On top of this, adding the c-TF-IDF model to reduce the impact of frequent words on the model produced a similarly placed "ethical" topic again, but with different words (this is in regards to the previous version of the model, which later implementations were based on). This time the new terms look more as though they belonged in corporate conversation as opposed to any social or ethical conversations, but this is perhaps due to the context that the ethics conversation is bound to, in these particular essays. On top of these other models were tested which didn't improve the quality of the topic model. The most surprising of these is that the MPNet model did not improve the success of the topic model. This was not expected and it is unfortunate that it did not help.

Going forward of course these models should be tested on the SDP reports. BERTopic in its default state should be able to do a sufficient amount of topic modelling by itself. However, it would be beneficial to compare this to other models, especially the seeded and seeded with c-TF-IDF models. Unfortunately, the others did not work out, however, they will still be tested further in some form, as it will provide a better understanding of the topic modelling process. The hope is that when using mpnet on a larger corpus of say 150 per model (as opposed to 18) it should perform to a much higher standard.

## 4.3 Using BERTopic to Topic Model SDP Reports

### 4.3.1 First Run

The outcome of all of these different models was quite unexpected. The default model performed by creating hundreds of models, some small and some large, which were very scattered, as can be seen in Figure 4.9. This meant that ”ethics” was the 43rd largest model of 241 topics. Using seeded topics for the next model actually caused ”ethics” to be a smaller topic at the 47th largest of 199 topics. Unfortunately, adding the c-TF-IDF parameter didn't cause anything to change, meaning that it does not need to be included in this corpus, possibly because of its size. The first noticeable difference occurred when a different embedding model (MPNet) was implemented, as seen in Figure 4.10, there were two topics which occurred containing the word ”ethic”, which were the 55th and 118th topics of 172 topics. Finally, when k-means clustering was introduced (Figure 4.11, the ethics topics were able to cluster together into the 8th largest topic of the preset 25.

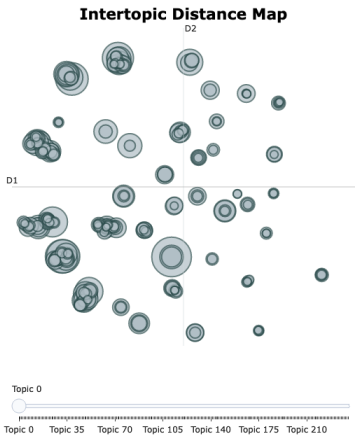


Figure 4.9: Intertopic distance map for default BERTopic on the 2021 corpus

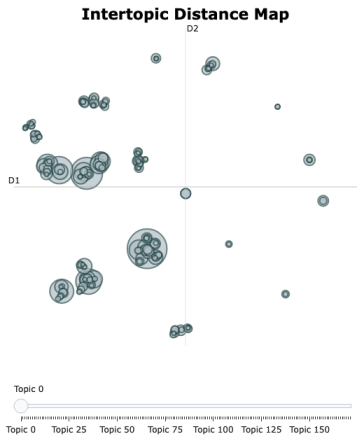


Figure 4.10: Intertopic distance map for BERTopic using mpnet, on the 2021 corpus

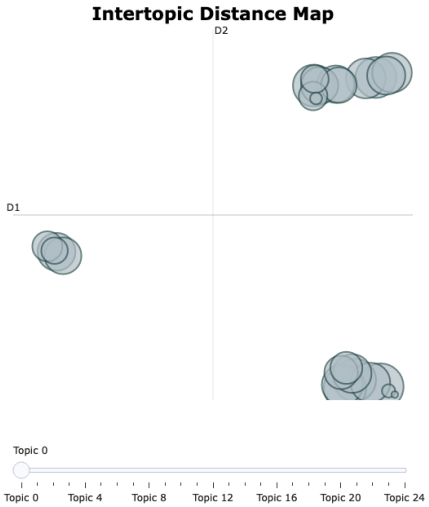


Figure 4.11: Intertopic distance map for BERTopic using k-means, on the 2021 corpus

In an even more unprecedented outcome, the topic models in 2022 changed in different ways from those in 2021. The default topic model for 2022 (Figure 4.19 saw the ”ethics” topic much smaller than that of 2021, at the 98th largest of 243 topics. When seeded topics were introduced, there were then two ”ethics” topics which formed. The first was the 5th largest topic and the second the 188th largest topic of 197. Again, the implementation of c-TF-IDF caused no changes to the topic model, just as it did in 2021. With the introduction of MPNet (Figure 4.20, the ”ethics” topic then became the 7th largest topic of 177. Finally, by using k-means clustering, as seen in Figure 4.14 the topic of ”ethics” became the 16th largest amongst 25 topics.

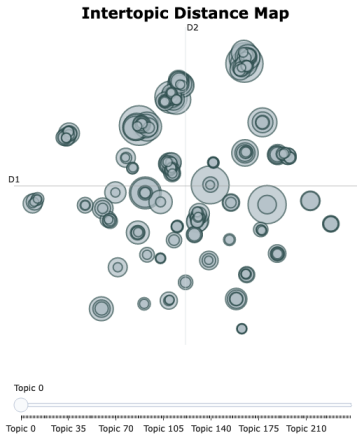


Figure 4.12: Intertopic distance map for default BERTopic on the 2022 corpus


Figure 4.13: Intertopic distance map for BERTopic using mpnet, on the 2022 corpus


Figure 4.14: Intertopic distance map for BERTopic using k-means, on the 2022 corpus

#### 4.3.1.1 Evaluation

The results and the models which reflected the corpus were quite different to what had been anticipated. Firstly, the expectation was that the first 3 models would be building on top of each other to, in the case of the third model, produce topics which would easily identify ”ethics” within the two respective years of SDP. Instead it was the combination of everything involved up to and including the use of k-means clustering which produced topics with any identifiable aspects of ”ethics” involved.

Secondly, with the two k-means models it is quite obvious that ”ethics” and ”ethical” produce a topic in both, however, the topic is somewhat smaller in 2022 compared to 2021. In 2022, the topic is the eighth largest and the words ”ethic” and ”ethical” have marginally higher c-TF-IDF scores in this model than in, 2021 where they are in a comparatively smaller model which is the sixteenth largest in that respective year.

There are a number of reasons why this might be the case. The schools increasing attention to how important ethics are in CS means that it is very unlikely that students are thinking less about ethics. The most important information on why this difference has occurred can likely be explained by the nature in which the course took place between the two academic years. The largest difference is that 2021 was during one of the many heights of COVID-19 and therefore it was an entirely remote course, whereas 2022 was able to be an in-person course. This meant that students had two potentially quite different versions of the same course. It is worth noting that because the course was remote, anything that the students wanted to be built would have to be done by the technicians in-house. Moreover, this means that students were constantly surrounded with new of the coronavirus, and also didn’t have as much to speak of when it came to the technical side of the project, or at least the hardware aspect. One could assume that this meant that students were often making projects which centered around COVID and therefore, how to help people in a world where everyone needed to be helped. It is likely that students either spoke of ethics more in 2021 than 2022 because their projects had a specific aim of helping people during a time of crisis (like the delivery drones

which were written about in the PI essays) or that ethics was spoken about in a different context between the two years. It is possible that in 2021 ethics was spoken about more broadly in terms of how people could be helped, whereas in 2022 when people were able to build physical robots again, in a world where people had also gone back to work, the conversation was more in terms of how these robots would not take peoples jobs for a myriad of reasons.

Despite the fact that the implementation had varying effects on topic modelling the corpus' of each year, there are some takeaways which can be used for a second implementation. Most obviously, using c-TF-IDF has no effects on the models and therefore should not be used going forwards. Using seeded topics is still interesting because it may have had a negative impact on the size of the topic brought out in 2021 but it arguably produced a topic more relevant to the area of "ethics". Whilst, MPNet also has varying effects it is also likely to produce better embeddings and therefore more accurate topics. K-means clustering biggest benefit is in the fact that there are a fixed number of topics to examine, arguably doing just as much to cluster ethics accurately as anything else. Due to these factors a second implementation would be beneficial by testing out each factor and controlling the number of topics BERTopic creates as well as using k-means clustering.

### 4.3.2 Second Run

#### 4.3.2.1 Seeded Models

Initially, the implementation began with the models which were simply default BERTopic with seeded topics, a fixed UMAP random_state of 42, and nr_topics set to 25 for the first model, and 50 for the second. What this produced was quite unexpected. For both corpus', most of the words were pushed into topic 0, which is the topic that holds all of the words that could not be assigned to a specific topic. This occurred across both years no matter the number of topics in the model. What did stand out however, is the fact that in 2021 there is no identifiable "ethics" topic whereas in 2022 there is. In the first model for 2022, topic number 7 contains both the words "ethical" and "ethic". This is matched by the second model in 2022 which also has the same words in topic number 12.

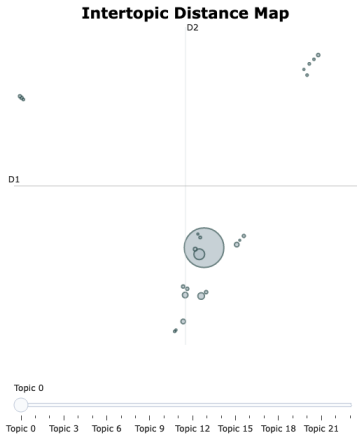


Figure 4.15: Intertopic distance map for BERTopic with 25 seeded topics, on the 2021 corpus

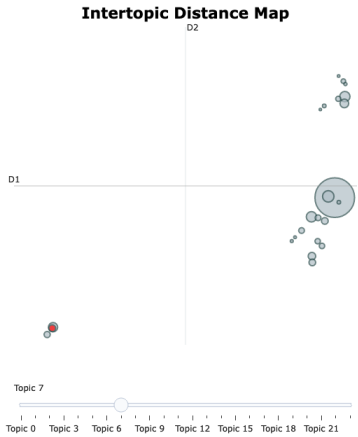


Figure 4.16: Intertopic distance map for BERTopic with 25 seeded topics, on the 2022 corpus. Topic 7 is shown in red in the bottom left

#### 4.3.2.2 MPNet Models

Moving onto the next models where MPNet was introduced as the embedding model to both of the previous models, the same pattern of many words assigned to topic zero was found, followed by very small topic, however in 2022 it seems to be less of a dive in numbers. Noticeably, this time the larger 2021 model (with 50 topics) does contain an "ethics" topic whilst the smaller does not. The topic in this case is topic 14 and contains both the words "ethic" and "ethical". What is also interesting is that the introduction of MPNet to the 2022 models also caused the same effect, where the smaller model has no "ethics" topic but the larger one does. It is also topic number 14 with the words "ethic" and "ethical"

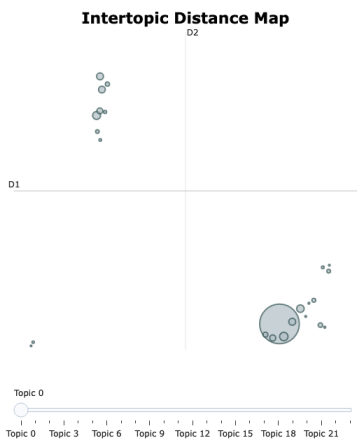


Figure 4.17: Intertopic distance map for BERTopic using MPNet with 25 seeded topics, on the 2021 corpus

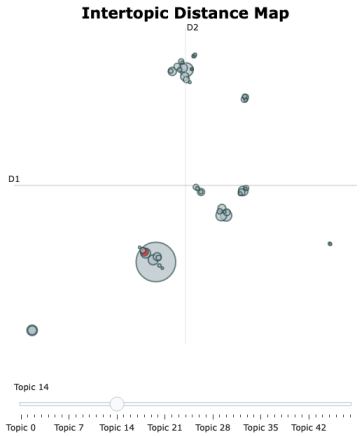


Figure 4.18: Intertopic distance map for BERTopic using MPNet with 50 seeded topics, on the 2021 corpus. Topic 14 is shown in red in the largest cluster

#### 4.3.2.3 K-means Models

The last four topic models used k-means instead of nr_topics, which had indicated to be useful in the first implementation and would prove itself on this run.

Firstly, comparing the models which used a k-means of 25. There was one model which used MPNet and one which did not. Across both years, both models found an ”ethics” topic. In 2021, the non-MPNet model found it in topic 15 with just the occurrence of the word ”ethical”, and the MPNet model found topic 12 to contain ”ethic” and ”ethical”. Likewise, in 2022 the non-MPNet model found topic 14 to contain ”ethical” and ”ethic”, whilst the MPNet model found topic 19 to contain ”ethic” and ”ethical”.

Lastly, the larger models with k-means 50, also all found an ”ethics” topic. In 2021, the non-MPNet model found topic number 32 to contain ”ethical” and ”ethic”, whilst the MPNet model found topic 12 to contain ”ethic” and ”ethical”. In 2022, the non-MPNet model found topic 9 to contain ”ethical” and ”ethic”, whilst the MPNet model found topic 35 to contain ”ethic” and ”ethical”

All this is to say that the earlier, less sophisticated models lacked ”ethics” topics in 2021 but did not in 2022. Also, the latter models in both years all found ”ethics” topics. This means that 5/8 models in 2021 contained an ”ethics” topic vs 7/8 in 2022.


Figure 4.19: Intertopic distance map for BERTopic using k-means MPNet with 50 seeded topics, on the 2021 corpus. Topic 12 is shown in red

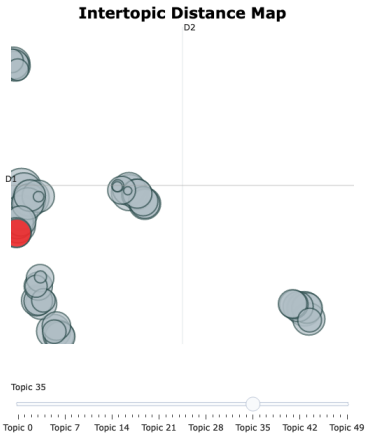


Figure 4.20: Intertopic distance map for BERTopic using k-means MPNet with 50 seeded topics, on the 2022 corpus. Topic 35 in shown in red

#### 4.3.2.4 Evaluation

Whilst topic modelling can be quite a difficult thing to evaluate, as previously established, there are a few stand out things in these models. Firstly, there are a clear lack of ”ethics” topics in much of the earlier models of the 2021 corpus. Now this can be argued that BERTopics own method on deciding the number of topics is worse than the k-means form of clustering, and that because of this an ”ethics” topic never formed. However, arguably this is because those topics did form in 2022 then they should have been just as capable of forming in 2021. This would suggest that the method is not the issue here and that instead ”ethics” is not talked about enough in the 2021 corpus to show up in these weaker models.

Furthermore, as MPNet is introduced to the models both years perform to change in embedding model almost identically. Both models also keep the very large 0th topic with many small topics created thereafter. What is interesting is that when this embedding model is introduced both years struggle to find an ”ethics” topic when there are only 25 topics available. However, when 50 topics are introduced both corpus’ produce an ”ethics” topic as topic 14.

Lastly, when the models were changed to use k-means clustering instead of determining the number of topics with nr_topics, the models performed quite differently. The trend throughout was that when k-means was introduced in conjunction with MPNet, the ”ethics” topics produced would be be larger in 2021 than in 2022. This is evident, as the k-means topics with 25 clusters produced topics 15 and 14 in 2021 and 2022 respectively, which then became topics 12 and 19 when MPNet was introduced. Similarly, for the topics with 50 k-means clusters, the ”ethics” topics were 32 and 9 in 2021 and 2022 respectively. However, when MPNet is introduced they become topics 12 and 35. This behaviour is quite difficult to explain but it could suggest that the ”ethics” as a topic is more noticeable in 2022 than in 2021 but when the greater embedding model is

introduced, ”ethics” is realised to be a more relevant topic in 2021 than 2022.

As previously mentioned, the 5/8 models which were able to find an ”ethics” topic in the 2021 corpus when compared to the 7/8 models in the 2022 corpus provides an interesting point. Specifically, when k-means was introduced it became incredibly easy for the models to be able to find an ”ethics” topic. However, prior to this only 1 model found an ”ethics” topic in the 2021 corpus vs the 3 in the 2022 corpus. For an ”ethics” topic to be discoverable in the 2021 corpus, without the use of k-means, MPNet had to be introduced as well as turning up the number of topics from 25 to 50. This would suggest that if it took the introduction of a better embedding model as well as a greater number of topics to discover an ”ethics” topic in the 2021 corpus, when the 2022 corpus had an easily available ”ethics” topic with a weaker model, then it is likely that the 2021 corpus has far lesser mentions of ”ethics” and therefore students awareness in the previous year was lower.

In terms of looking at the other words within the topics we shall look at the MPNet models using a k-means cluster of 25. These are the models which will be compared as there is less of drastic jump when MPNet was introduced to the k-means model with only 25 clusters, and MPNet, being the better embedding model, should give a better representation of the topics involved in the corpus, which should hopefully give a better understanding. The 2021 model, has an ”ethics” topic at topic 12 which contains these words:

*feedback, documentation, guide, presentation, ethic, user, ethical, information, received, document*

Given the words that make up this topic it would suggest that the topic is mainly focused on people, as seen by the words ”feedback” and ”user”. The 2022 model, has an ”ethics” topic at topic 19 which contains these words:

*market, ethic, ethical, product, research, business, important, target, personal, clear*

This topic contains words like ”market”, ”product”, and ”business” which would suggest a topic that is focused on things in the corporate world, as though presenting to a business. This difference between the two topics would agree with the proposal, of a different teaching style, from the last run of the SDP corpus.

Similarly, the topics surrounding the ”ethics” topic should be taken into consideration. For this comparison we will use the same models. The 2021 MPNet model shows the ”ethics” model to be overlapping with a model containing the words ”project”, ”product”, and ”market”. The 2022 MPNet model shows the ”ethics” model to be overlapping with a topic containing the words ”project”, ”work” and ”research”. Interestingly, the contacting topic in 2021 shares all of the same words as that of the ”ethics” topic in 2022. This would suggest that they take up much of the same vector space and there are perhaps only slight differences between the respective ”ethics” topics. What is clear from this is that the 2021 ”ethics” topic encapsulates people whilst overlapping the same space as a topic which talks about the market, whereas the 2022 ”ethics” topic encapsulates the market whilst also overlapping with a topic talking about other things

related to the work place. This would further imply that the focus between years in terms of ethics shifted somewhat.

It is clear that when MPNet is employed in conjunction with k-means, the ”ethics” topic is much larger in in 2021 than in 2022. However, it is also clear that without such means, with a much simpler topic model, an ”ethics” topic is undetectable in 2021 but obtainable in 2022. What is also clear, is that the terminology within the ”ethics” topics changes between the two years. All of this is to say that based on these facts it would appear that ethics may have been talked about as a whole more frequently in 2022, likely due to the SEPP and PI courses taken before, but with a better embedding model it is clear than in 2021 ethics is talked about more in-depth. This is likely due to, as mentioned on the last run of models, the difference in how the course was taught and the circumstances around which the school was put in. 2021 was a remote year, meaning that students were working from home, with many simulated project, and many focused on how to help people during the pandemic. Conversely, 2022 was an in-person year, which meant that students were less focused on COVID-centric project, and also had their eyes on Industry Day - the day when students present their projects to people from companies. This is emphasised by the way in which the terminology changed in the topics between years and strongly suggests reasons for the differences between the corpus’.

# Chapter 5

# Overall Conclusions

In Chapter 1 it was stated that the aim of this project was to measure if students' usage and understanding of ethics applied to computer science has increased in quantity and quality during their time at The University of Edinburgh. This aim was followed by three main goals which must be achieved to have sufficiently accomplished this task. The extent to which each of these goals were achieved will be acknowledged, followed by the findings, strengths and weaknesses of the project, and further work to be completed.

The goals were as follows:

- Develop a topic modelling method which can accurately pick out a topic which relates to ethics in any piece of written work - This was achieved thanks to the use of guided topic modelling to pull out an ethics topic in all instances, given the most powerful model which used k-means clustering and the MPNet embedding model.
- Ensure the method will have the final capability of determining if ethics has become an increasingly talked about concern in students SDP reports from year to year - The method is definitely capable of measuring how large of a topic ethics is and models for both years are available for comparison. However, it is not as simple as determining if it has become more talked about, which takes us to the next goal.
- Accurately evaluate the data to draw useful conclusions. Focusing on how the underlying method, and external factors, like the course structure, may impact the final results - This part was definitely achieved as there was a discussion around how it is very possible that, due to how COVID-19 changed how the course was administered, the projects within the course may have changed how students discussed and considered ethics as part of their own projects. As well as the a lack of in-person events to gear discussion, like Industry Day

To conclude, each of these goals were met. In terms of evaluating the method itself we can look at how the implementation evolved throughout to eventually produce topic models for 2 years of study which can be visualised to determine how large a topic

”ethics” is between years.

## 5.1 Strengths and Weaknesses

Many of the strengths of this project come from applying what was found in the literature review and combining that with a more up to date version of topic modelling in the form of BERTopic. In Chapter 2, it was mentioned how researchers used a semi-supervised method of LDA which entailed using keywords when topic modelling to guide the model into producing certain topics. This was something which was applied to this project to ensure that, if ethics is sufficiently covered throughout the corpus, it will form into a topic of its own. BERT and BERTopic were also incorporated into the Background Chapter to emphasise how it would be beneficial for use in this project. Here, it was covered that BERTopic is useful for understanding short texts when compared to more traditional techniques as it has better topic coherence. It also provided an easy way to visualise the models so that the implementation could be altered on the fly.

Another strength came in the form of using POS tags in what is normally an unsupervised method. Due to the nature of this project, a large corpus was involved which was entirely unavailable firsthand, however it still had to be processed in a way which students would stay anonymous when the topic models were formed. This was easy to do when removing student numbers from the data as it could be done with regular expressions. However, removing students individual names provided more of a challenge which had to be overcome by tokenising the data and POS tagging each data to find which tags identified as ”PERSON” and then removing this. This ensured that all of the corpus stayed anonymous before the models were even built.

Some weaknesses of this project come in a few different areas. One of which being the availability of the data due to privacy reasons. Whilst this is just the nature of this project, given how it was decided upon after the data already existed and the corpus only covers the last 2 years, it would be beneficial for this project to have a larger corpus of say 5 years. This should also come with students opting in to allow their data to be seen by the student working on this project, as this would allow further testing on the main corpus.

Another weakness, could be determined in how the topics were built. Whilst a justifiable job was done in the library which was used and the types of topics which were built, it could have been beneficial to use another type of topic modelling which would have been used in comparison to BERTopic.

Lastly, an improvement which could be made and is parroted by some earlier research in Chapter 2, is in terms of evaluation. Of course it is easy to visualise and evaluate with the human eye which models were better than, or outperformed others, but there ought to be other, perhaps more numerical ways of justifying different models.

## 5.2 Future Work

In terms of future work, there is likely a couple of routes which can be taken here. As mentioned previously, a possible route would be to gather more data before analysing the corpus' over the years. It would likely allow for more of a trend to show and would, hopefully, allow for more of a reliable and numerical evaluation to take place.

Furthermore, using other forms of topic modelling, perhaps using a traditional form like LDA, could provide some sort of benchmark or at least a model for comparison. Alongside this, perhaps using some other form of neural network aside from BERT could produce some interesting findings.

Finally, a method which complements the topic modelling could also be used. Despite, researchers, and this projects, best efforts to make topic modelling more of a quality driven approach than a quantity driven one, topic modelling still encapsulates clustering together groups of words based on frequency in its very essence, so something deeper could prove quite interesting. For example, using sentiment analysis to gain an idea of the feelings in students writing. This could lead to a better understanding of how the way students feel about their project plays into how they talk about the ethics involved.

# Bibliography


[1] Documentation. Available at https://www.nltk.org/api/nltk.tag.html.

[2] Universal pos tags. 2022. Available at https://universaldependencies.org/u/pos/.

[3] Abhigyan. Importance of dimensionality reduction!! 2020.

[4] Abeer Abuzayed and Hend Al-Khalifa. Bert for arabic topic modeling: An experimental study on bertopic technique. *Procedia Computer Science*, 189:191–194, 2021. AI in Computational Linguistics.

[5] Jay Alammar. The illustrated word2vec.

[6] aravindpai. What is tokenization in nlp? here's all you need to know.

[7] S. Aslani and J. Jacob. Utilisation of deep learning for covid-19 diagnosis. *Clinical Radiology*, 78(2):150–157, 2023. Special Issue Section: Artificial Intelligence and Machine Learning.

[8] Ivana Bartoletti. *An Artificial Revolution: On Power, Politics and AI*. The Indigo Press, first edition, 2020.

[9] Srinivas Chakravarthy. Tokenization for natural language processing.

[10] Salman Ibne Eunus. Difference between stemming and lemmatization: Data science and machine learning.

[11] Raffaele Filieri, Zhibin Lin, Yulei Li, Xiaoqian Lu, and Xingwei Yang. Customer emotions in service robot encounters: A hybrid machine-human intelligence approach. *Journal of Service Research*, 25(4):614–629, 2022.

[12] Kavita Ganesan. What are stop words?

[13] Max M. Louwerse Guido M. Linders. Zipf's law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort. *SpringerLink*, 2023.

[14] David Robinson Julia Silge. Topic modeling.

[15] Kunlun Qi Jingmin Yu Siluo Yang Tianxing Yu Jie Zheng Bo Liu Kai Hu, Huayi Wu. A domain keyword analysis approach extending term frequency-keyword active index with google word2vec model. 2017.

[16] R. Kavitha, D. Kiruba Jothi, K. Saravanan, Mahendra Pratap Swain, José Luis Arias Gonzáles, Rakhi Joshi Bhardwaj, and Elijah Adomako. Ant colony optimization-enabled cnn deep learning technique for accurate detection of cervical cancer. *BioMed Research International*, 2023:1742891, Feb 2023.

[17] Ria Kulshrestha. A beginner's guide to latent dirichlet allocation(lda). 2019.

[18] Xuerong Li, Wei Shang, and Shouyang Wang. Text-based crude oil price forecasting: A deep learning approach. *International Journal of Forecasting*, 35(4):1548–1560, 2019.

[19] Ben Lutkevich. natural language processing (nlp).

[20] Maarten. Frequently asked questions.

[21] Deep Mehta. Part of speech tagging – pos tagging in nlp.

[22] Vishak Nair. Zipf's law validation with word frequency.

[23] Qin Xiang Ng, Dawn Yi Xin Lee, Chun En Yau, Yu Liang Lim, and Tau Ming Liew. Public perception on 'healthy ageing' in the past decade: An unsupervised machine learning of 63,809 twitter posts. *Heliyon*, 9(2):e13118, 2023.

[24] Gokhan Ozogur, Mehmet Ali Erturk, Zeynep Gurkas Aydin, and Muhammed Ali Aydin. Android Malware Detection in Bytecode Level Using TF-IDF and XGBoost. *The Computer Journal*, 01 2023. bxac198.

[25] Akash Panchal. Nlp — text summarization using nltk: Tf-idf algorithm. 2019.

[26] Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions.

[27] Cambridge University Press. Inverse document frequency. 2009.

[28] Cambridge University Press. Term frequency and weighting. 2009.

[29] Olessia Koltsova Sergey I. Nikolenko, Sergei Koltcov. Topic modelling for qualitative studies. 2016.

[30] Junaid Abdul Wahid, Lei Shi, Yufei Gao, Bei Yang, Lin Wei, Yongcai Tao, Shabir Hussain, Muhammad Ayoub, and Imam Yagoub. Topic2labels: A framework to annotate and classify the social media data through lda topics and deep learning models for crisis response. *Expert Systems with Applications*, 195:116562, 2022.

[31] Hongyu Zhang. The distribution of word frequencies in the novel "ulysses" zipf's law ...

[32] He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray L. Buntine. Topic modelling meets deep neural networks: A survey. *CoRR*, abs/2103.00498, 2021.

[33] Sulong Zhou, Pengyu Kan, Qunying Huang, and Janet Silbernagel. A guided latent dirichlet allocation approach to investigate real-time latent topics of twitter data during hurricane laura. *Journal of Information Science*, 49(2):465–479, 2023.

# Appendix A

# First appendix

## A.1 SEPP and PI Topic Visualisations from Example Run


Figure A.1: Intertopic distance map for default BERTopic on corpus


Figure A.2: Intertopic distance map for seeded BERTopic on the corpus


Figure A.3: Intertopic distance map for seeded BERTopic with k-means clustering on corpus

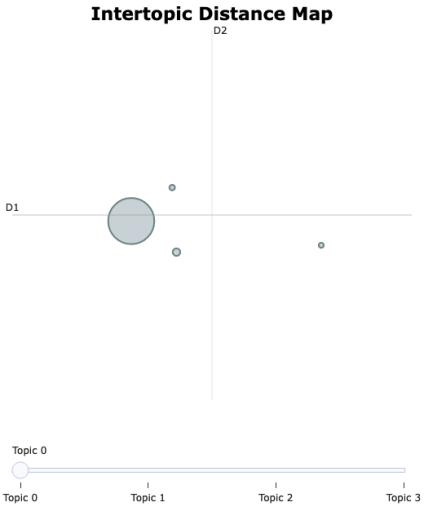


Figure A.4: Intertopic distance map for seeded BERTopic using HDB-SCAN, on the corpus

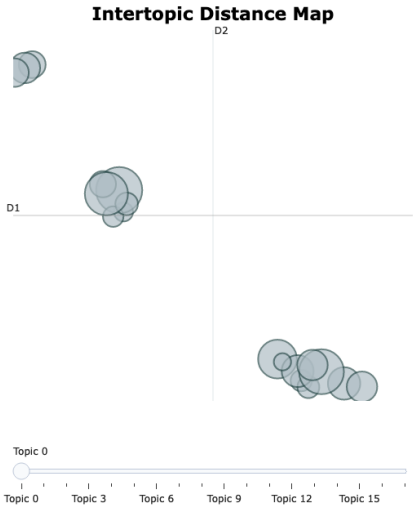


Figure A.5: Intertopic distance map for seeded BERTopic, using c-TF-IDF, on corpus

## A.2 SDP Report Topic Visualisations from First Run


Figure A.6: Intertopic distance map for default BERTopic on the 2021 corpus


Figure A.7: Intertopic distance map for seeded BERTopic on the 2021 corpus

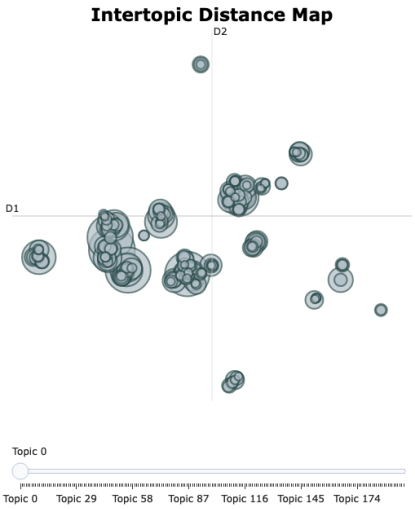


Figure A.8: Intertopic distance map for BERTopic using c-TF-IDF, on the 2021 corpus

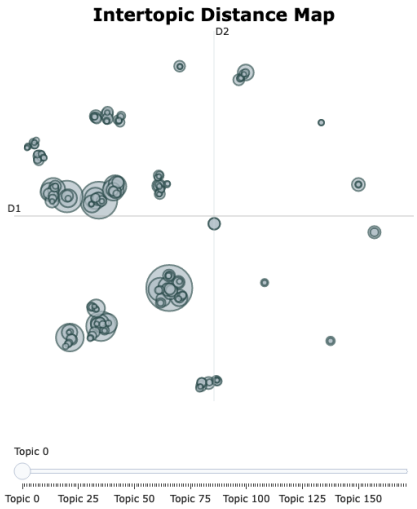


Figure A.9: Intertopic distance map for BERTopic using mpnet, on the 2021 corpus

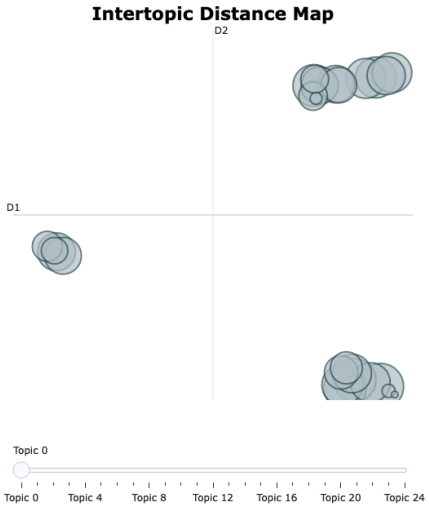


Figure A.10: Intertopic distance map for BERTopic using k-means, on the 2021 corpus

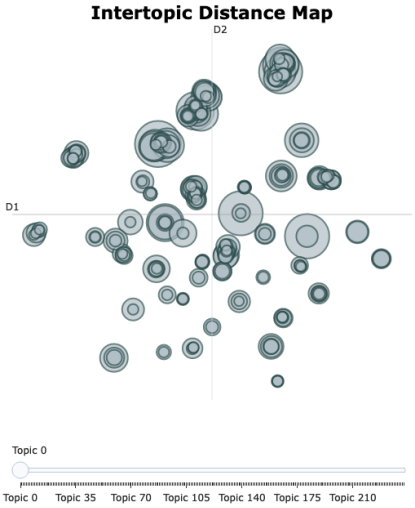


Figure A.11: Intertopic distance map for default BERTopic on the 2022 corpus

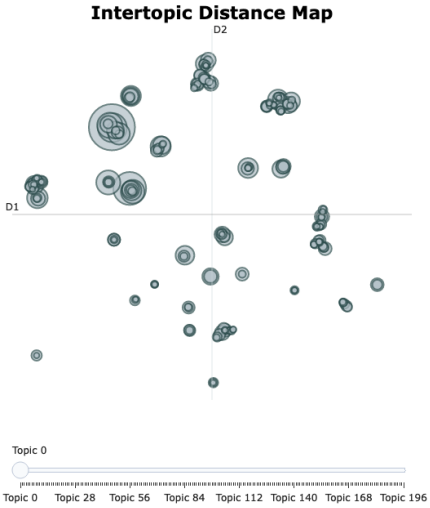


Figure A.12: Intertopic distance map for seeded BERTopic on the 2022 corpus

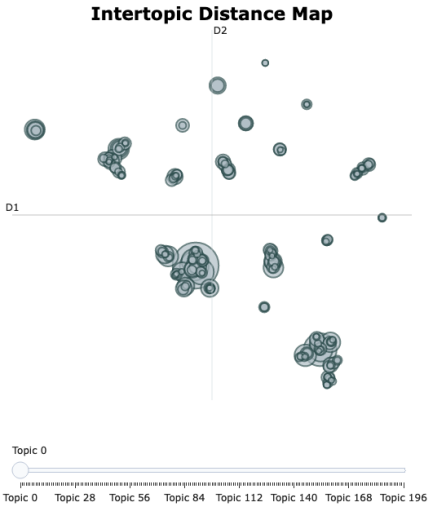


Figure A.13: Intertopic distance map for BERTopic using c-TF-IDF, on the 2022 corpus

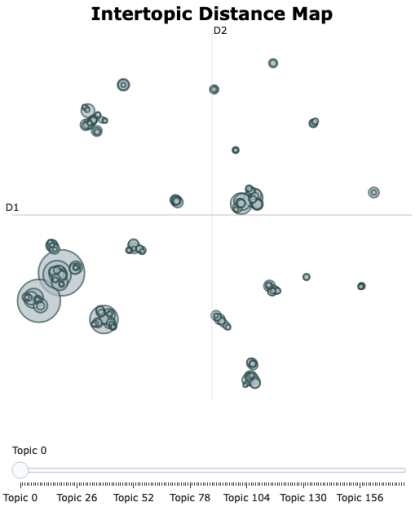


Figure A.14: Intertopic distance map for BERTopic using mpnet, on the 2022 corpus


Figure A.15: Intertopic distance map for BERTopic using k-means, on the 2022 corpus

## A.3 SDP Report Topics from First Run

Table A.1: SDP Report Topics from First Run

| Year | Model Description | Topic Number | Words in Topic |
|---|---|---|---|
| 2021 | Default | 43 | ethics, ethical, privacy, approval, consent, survey, questionnaire, protect, security, legal |
| | Seeded | 47 | ethical, ethic, approval, consent, responsibility, harm, avoid, product, fair |
| | c-TF-IDF | 47 | ethical, ethic, approval, consent, responsibility, harm, avoid, product, fair |
| | mpnet | 55 | ethic, ethical, approval, harm, doc, eat, fairness, legal, jargon, clearance |
| | | 118 | privacy, protect, threat, buried, caution, security, ethic, breach, adhere, stolen |
| | k-means | 8 | ethic, ethical, approval, consent, responsibility, research, harm, list, related, section |
| 2022 | Default | 98 | ethical, ethic, moral, offended, participant, elimination, consent, hurt, study, proceed |
| | Seeded | 5 | ethical, ethic, medical, consent, harm, approval, study, moral, avoid, fear |
| | | 188 | contribution, key, carrying, expand, influential, positively, responsible, believe, directly, ethic |
| | c-TF-IDF | 5 | ethical, ethic, medical, consent, harm, approval, study, moral, avoid, fear |
| | | 188 | contribution, key, carrying, expand, influential, positively, responsible, believe, directly, ethic |
| | mpnet | 7 | ethical, ethic, harm, consent, medical, moral, participant, approval, behaviour, respectful |
| | k-means | 16 | market, ethic, ethical, poster, product, marketing, business, key, study, survey |

## A.4 SDP Report Topic Visualisations from Second Run


Figure A.16: Intertopic distance map for BERTopic with 25 seeded topics, on the 2021 corpus


Figure A.17: Intertopic distance map for BERTopic with 50 seeded topics, on the 2021 corpus

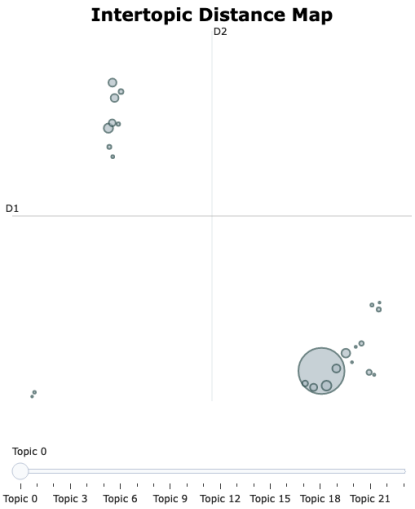


Figure A.18: Intertopic distance map for BERTopic using MPNet with 25 seeded topics, on the 2021 corpus

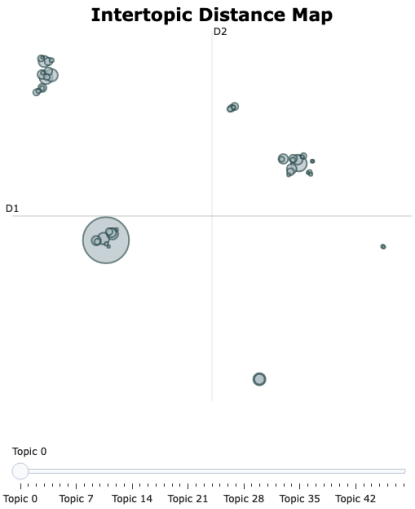


Figure A.19: Intertopic distance map for BERTopic using MPNet with 50 seeded topics, on the 2021 corpus

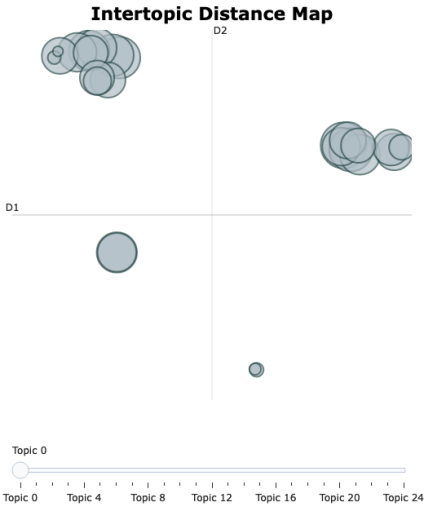


Figure A.20: Intertopic distance map for BERTopic using k-means with 25 seeded topics, on the 2021 corpus

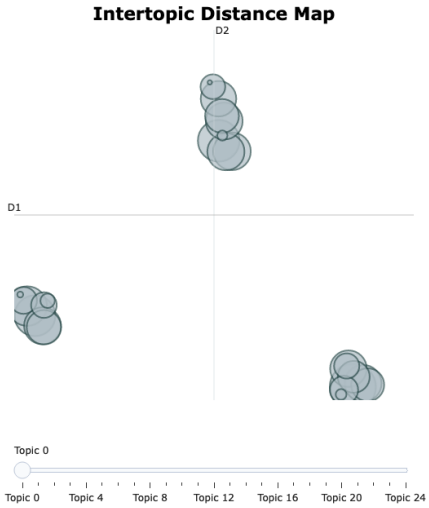


Figure A.21: Intertopic distance map for BERTopic using k-means MPNet with 25 seeded topics, on the 2021 corpus


Figure A.22: Intertopic distance map for BERTopic using k-means with 50 seeded topics, on the 2021 corpus


Figure A.23: Intertopic distance map for BERTopic using k-means MPNet with 50 seeded topics, on the 2021 corpus

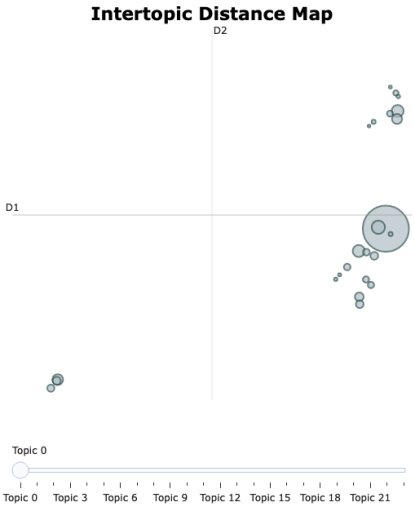


Figure A.24: Intertopic distance map for BERTopic with 25 seeded topics, on the 2022 corpus

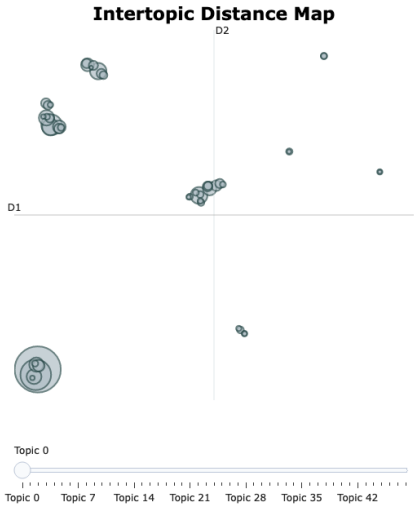


Figure A.25: Intertopic distance map for BERTopic with 50 seeded topics, on the 2022 corpus

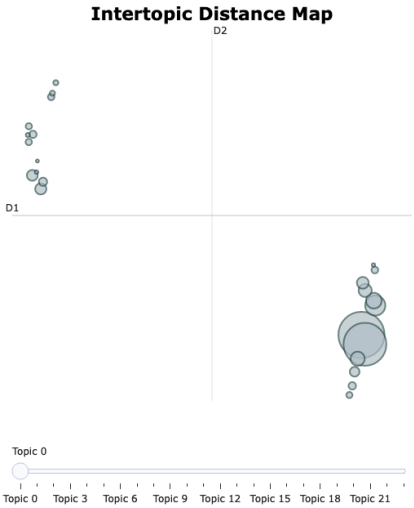


Figure A.26: Intertopic distance map for BERTopic using MPNet with 25 seeded topics, on the 2022 corpus

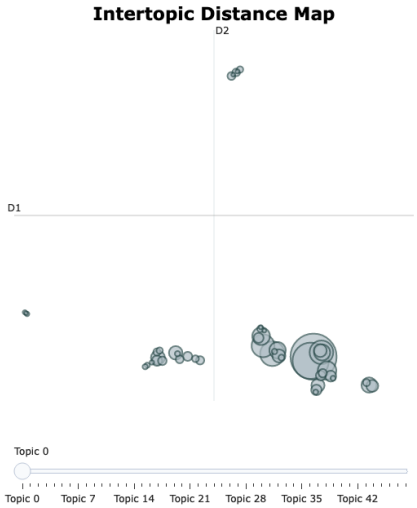


Figure A.27: Intertopic distance map for BERTopic using MPNet with 50 seeded topics, on the 2022 corpus

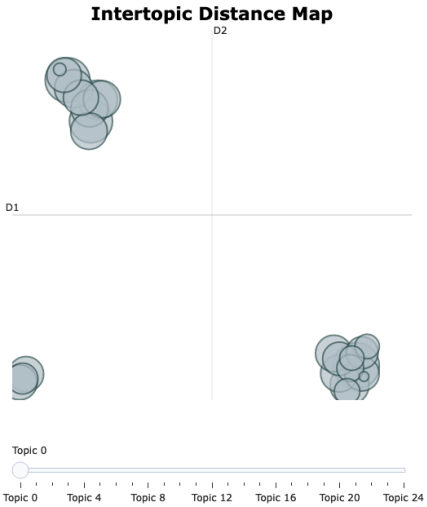


Figure A.28: Intertopic distance map for BERTopic using k-means with 25 seeded topics, on the 2022 corpus

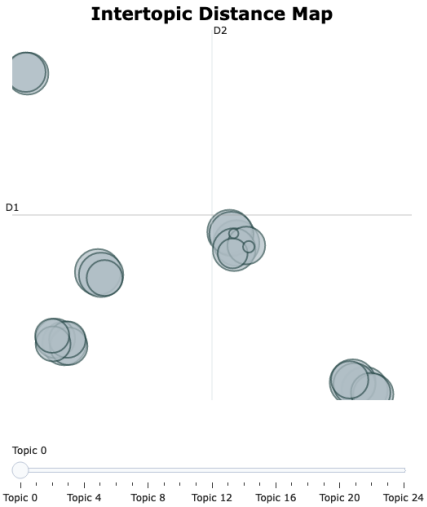


Figure A.29: Intertopic distance map for BERTopic using k-means MPNet with 25 seeded topics, on the 2022 corpus


Figure A.30: Intertopic distance map for BERTopic using k-means with 50 seeded topics, on the 2022 corpus

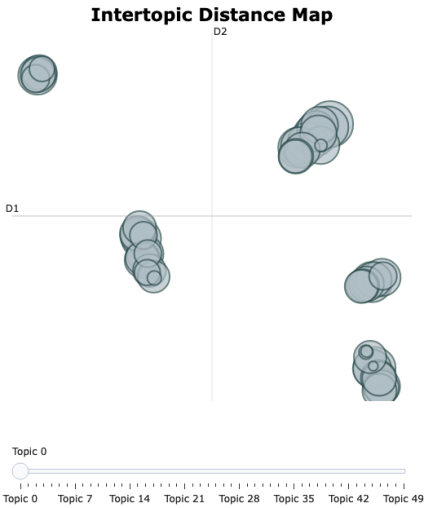


Figure A.31: Intertopic distance map for BERTopic using k-means MPNet with 50 seeded topics, on the 2022 corpus

## A.5 SDP Report Topics from Second Run

Table A.2: SDP Report Topics from Second Run

| **Year** | **Model Description** | **Topic Number** | **Words in Topic** |
|---|---|---|---|
| 2021 | Seeded - 25 Topics | N/A | N/A |
| | Seeded - 50 Topics | N/A | N/A |
| | Seeded MPNet - 25 Topics | N/A | N/A |
| | Seeded MPNet - 50 Topics | 14 | ethic, ethical, privacy, approval, opinion, responsibility, security, product, importance, protect |
| | Seeded k-means - 25 Topics | 15 | code, documentation, user, writing, guide, function, ethical, write, section, written |
| | Seeded k-means + MPNet - 25 Topics | 12 | feedback, documentation, guide, presentation, ethic, user, ethical, information, received, document |
| | Seeded k-means - 50 Topics | 32 | ethical, interface, ethic, user, state, layout, approval, diagram, consent, sketch |
| | Seeded k-means + MPNet - 50 Topics | 12 | documentation, ethic, guide, user, ethical, approval, responsible, evaluation, section, presentation |

Table A.3: SDP Report Topics from Second Run

| **Year** | **Model Description** | **Topic Number** | **Words in Topic** |
|---|---|---|---|
| 2022 | Seeded - 25 Topics | 7 | ethical, ethic, personal, avoid, situation, consent, medical, important, behaviour, user |
| | Seeded - 50 Topics | 12 | ethical, ethic, personal, consent, avoid, situation, medical, behaviour, conflict, important |
| | Seeded MPNet - 25 Topics | N/A | N/A |
| | Seeded MPNet - 50 Topics | 14 | ethic, ethical, conflict, behaviour, harm, consent, mdeical, personal, opinion, study |
| | Seeded k-means - 25 Topics | 14 | key, ethical, research, ethic, market, contribution, initiative, decision, product, personal |
| | Seeded k-means + MPNet - 25 Topics | 19 | market, ethic, ethical, product, research, business, important, target, personal, clear |
| | Seeded k-means - 50 Topics | 9 | ethical, ethic, market, research, personal, important, avoid, target, health, mental |
| | Seeded k-means + MPNet - 50 Topics | 35 | ethic, market, ethical, business, research, product, filter, clear, target, health |