

# The Alan Turing Institute

---

## Data Study Group Final Report: Defence Science & Security Laboratory (DSTL)

Using Topic Modelling to Discover  
New Trends in the Scientific Literature

**15-26 May 2023**



# Contents

<b>1</b>	<b>Executive Summary</b>	<b>4</b>
1.1	Challenge Overview	4
1.2	Main Objectives	6
1.3	Approach	6
1.4	Main Conclusions	7
1.5	Limitations	8
1.6	Recommendations and Future Work	9
<b>2</b>	<b>Introduction</b>	<b>10</b>
2.1	Challenge Background and Motivation	10
2.2	Mapping the Scientific Landscape	11
2.3	Topic Modelling: What is it and why do we use it?	12
2.4	BERTopic	13
2.5	Comparison of Modern and 'Traditional' Topic Modelling Approaches	13
2.6	Challenge Summary and Objectives	14
<b>3</b>	<b>Data Overview</b>	<b>16</b>
3.1	Dataset Description	16
3.2	Data Quality Issues	19
<b>4</b>	<b>Approach 1: Topic Dynamics</b>	<b>21</b>
4.1	Goal	21
4.2	Scientific Disruption	23
4.3	Topological Data Analysis - Mapper	28
4.4	Word Mover's Distance	30
<b>5</b>	<b>Approach 2: Data visualisation</b>	<b>32</b>
5.1	Dimensionality Reduction - UMAP	32
5.2	Weighted Topic Embeddings	34
5.3	Connectivity	36
5.4	BERTopic Model Visualization	36
<b>6</b>	<b>Approach 3: Topic Labelling</b>	<b>43</b>
6.1	Goal	43
6.2	Approaches	44

<b>7</b>	<b>Conclusion</b>	<b>54</b>
<b>8</b>	<b>Team Members</b>	<b>55</b>

## List of Figures

1	Distribution of Scientific Papers by Semantic Scholar Field . . .	19
2	Distribution of Scientific Paper by Month of Publication . . .	19
3	Distribution of Topics over Time . . . . .	20
4	Missing Values by month (facet) and scientific field (colour)	21
5	Proportion of papers which cannot be assigned to a topic over time . . . . .	22
6	How disruptiveness is measured from a dataset of academic papers. (a) Represents the data-set consisting of topics represented as feature vectors with time stamps. (b) We create a directed network of topics where each node is a topic and an arrow between $node_i \rightarrow node_j$ represents that $node_i$ is a child of $node_j$ . (c) Represents an un-directed topic similarity network within the same time stamp. . . . .	24
7	(left) Network stats for directed networks using the <i>cosine</i> similarity networks as we span the domain of the threshold. The same is represented on the (right) except that the similarity score here is $l2 - norm$ . . . . .	26
8	Example static topic analysis, connecting and visualising the network of topics within the same run . . . . .	27
9	Topic Subgraph based on TDA Mapper + BERT . . . . .	29
10	The examples of using WMD . . . . .	31
11	Topic Distribution over Six Topic Model Runs. Each run is shown in a different colour . . . . .	34
12	Weighted Topic Embeddings - Run 1 . . . . .	35
13	Weighted Topic Embeddings - Run 2 . . . . .	35
14	Topic Connectivity Run 1 . . . . .	37
15	Topic Connectivity Run 2 . . . . .	37
16	Two-dimensional projection of documents and topics . . . .	38
17	Hierarchical clustering of topics . . . . .	39
18	Top word scores for each topic for a single run. Note unlabeled nature of Topic names (0, 1, 2, etc) . . . . .	39

19	First model run after applying UMAP . . . . .	40
20	Sixth model run after applying UMAP . . . . .	40
21	Sankey diagram showing connections between topics and topic changes over time . . . . .	42
22	Bigrams including stop words . . . . .	48
23	Bigrams excluding stop words . . . . .	48
24	Trigrams including stop words . . . . .	48
25	Trigrams excluding stop words . . . . .	48
26	Table 4: Sumy results compared to BERTopic labels . . . . .	51

## List of Tables

1	Augmented six bucket data . . . . .	20
2	Example topic labels using in-built tf-idf approach . . . . .	44
3	Least weighted and top-weighted topics in Round 202208 . . . . .	45

# **1 Executive Summary**

This Alan Turing Institute Data Study Group (DSG) explored the use of topic modelling to discover new trends in the scientific literature. The challenge was presented by the Defence Science and Technology Laboratory (Dstl) who have been working in this area for a number of years. The aim of this DSG is to advance the work already started by Dstl, whilst also building a better understanding of the specific research challenges that automated horizon scanning techniques present and how these can be overcome in the future. As the science and technology landscape continues to rapidly shift and evolve, this challenge will help policy makers to stay on top of these latest developments.

## **1.1 Challenge Overview**

New scientific breakthroughs and disruptive technologies are currently emerging at a rapid pace. Large Language Models and new applications of Generative Artificial Intelligence (AI) have disrupted traditional approaches of Machine Learning and Data Science, while additive manufacturing, synthetic biology, quantum computing and mass-produced small satellites are all in the process of disrupting traditional areas of science and technology. Given this rapid pace of change it is not possible to manually keep on top of the latest developments without committing vast resources or an unrealistic amount of human time. However, it is important for policy makers from across Government to be aware of these trends and conscious of the emerging technologies and scientific research areas which may present both opportunities as well as risks. Developing automated or semi-automated approaches to this horizon scanning and discovery process can help to reduce the burden for analysts, and also to improve performance by allowing for more specific or quantitative assessments.

The Discovery Project at the Defence Science and Technology Laboratory (Dstl) aims to identify emerging topics and significant shifts in the Science and Technology landscape by developing data science approaches which can automate parts of the horizon scanning process. The team at Dstl hypothesise that topic modelling, a form of natural language processing (NLP), can be used to add structure to academic literature. These trends

can then be explored further by analysts. By tracking how topics change over time, emerging technologies can be identified at a faster pace than manual scanning, therefore giving the UK a competitive edge in assessing these technologies' applicability to Defence and Security.

Topic modelling is a form of unsupervised machine learning which uses Natural Language Processing (NLP) to take a set of documents and cluster them into groups based on the similar words and phrases used within the documents. When applied to a collection of scientific papers, each topic that the model identifies should represent a distinct scientific or technological concept, and the distribution of papers over topics can provide insight into the scientific landscape. This approach therefore provides a solid basis from which to build automated horizon scanning approaches.

The work within this DSG builds on prior work completed by Dstl in a machine learning topic modelling technique called BERTopic [12]. This approach has been used to scan large datasets of scientific papers, extract relevant keywords and create topics that are a concatenation of the top  $n$  keywords, and then assign the papers to these topics. BERTopic does this by leveraging both pre-trained language models and word frequencies to generate robust and easily interpretable topics.

However, the goal of this project is to monitor the scientific landscape in a continuous manner, and not just at a single point in time. Topic models are notoriously poor at identifying new topics that were not in the original training data, and therefore these topic models will need to be re-trained at regular intervals in order to detect new and emerging scientific trends and technologies as new data becomes available. Studying a time series of topic model runs would allow important dynamics to be uncovered that would otherwise remain unseen in a single run. To allow for this to be studied within the DSG, Dstl has provided the outputs from their current state-of-the-art topic modelling approach which was run monthly on the previous 6 months of data creating a 6 month rolling window. This data can then be used to create timeseries information for analysing the trends in topic development found in the data.

The primary challenge is therefore to develop an approach that can track how topics evolve and change over this topic modelling output.

## **1.2 Main Objectives**

There is a challenge in how to measure a topic set which evolves over time, with new topics emerging. This challenge drives the objectives of this project which are to develop approaches that can track how topics within a corpus of scientific papers evolve through time across topic models trained on temporal segments of the data. Successful approaches should be able to flag the significant differences between different topic model runs, including when new topics are emerging, older topics are reducing, and existing topics are converging or diverging. These dynamics might be represented by topics moving closer to one another, further apart, new and unexpected connections forming between existing topics, or sub-fields within a topic receiving differential focus from different areas of the scientific landscape.

Re-training topic models in this way to answer these questions introduces some challenges, and addressing these will be vital in order to successfully implement the project. Firstly, each topic model re-training 'run' on a tranche of new data will detect a different number of topics, and the makeup of these topics will be different; the salient keywords in the topics will differ even if the underlying latent topic is the same, and the boundaries between topics might shift. Understanding when these shifts are caused by noise in the data, and when they represent a true change in the science and technology landscape is a key question. Secondly, novel and emerging topics are likely to be underrepresented in the data, at least initially. It is important that techniques to detect these trends are sensitive enough to measure this emergence, but not over sensitive such that topic model runs are incomparable from one to the next. Finally, the results need to be interpretable by human analysts, and provide actionable insights which can then be further explored.

## **1.3 Approach**

In order to address the challenges presented in this project and identify solutions from multiple different angles, the project was approached from different directions by subsets of the DSG team. The project was broken into 3 parts: Topic identification and measuring of topic dynamics, topic and data visualisations, and topic labelling.



Each sub-team focused on one of the following areas:

- **Topic Dynamics:** Responsible for linking/establishing the relationship of topics within a corpus in a single run and across corpuses in different runs. Identifying when the same topic appears in multiple topic model runs, and developing approaches to measure how these topics are changing over time.
- **Data Visualization:** Responsible for identifying visualizations that could be used to clearly and intuitively represent the relationships identified above.
- **Data Re-labelling:** Responsible for identifying topic labels or names that are human-readable as compared to the output topic label from the BERTopic model which are based solely on keywords and therefore difficult for an analyst to interpret.

Each sub-team's approach and results are detailed throughout the paper.

## 1.4 Main Conclusions

This project demonstrated the utility in pursuing automated horizon scanning approaches to better understand the science and technology landscape, but also some of the intrinsic challenges in doing so. The main conclusions of this project are:

- There are limitations in the current datasets used to build the topic models which will need to be addressed. These include missing abstracts which affect the topic modelling process, and truncated publication data information which makes modelling over time a challenge.
- Nevertheless, we identify some promising avenues for exploring topic dynamics. We find that the relationships between topics are best identified using a combination of cosine similarity and Euclidean distance to determine within-run and across-run linkages respectively. Disruptive Index and Average weighted distance, similar to the cognitive distance measure, were used and prove promising. With these metrics, the directed (temporal) and

non-directed (spatial) linkage, between topics were identified.

- Effective data visualisations are key to good communication of the results of topic modelling. However representing high dimensional data in low dimensional space, without losing information, is a challenge. Efforts will need to be made to educate end-users on how to interpret data visualisations which have undergone dimensionality reduction in this way.
- In order to identify human-readable labels for topics it would be valuable to have human annotated data which could be used to both train and test models. Without labels, and relying solely on unsupervised approaches, it is challenging to evaluate the success or limitations of the approaches presented here. Regardless, we demonstrate the utility in combining summarisation techniques from the field of Natural Language Processing with the keyword-extraction capabilities within BERTopic.

## **1.5 Limitations**

This project was limited in terms of data quality and completeness, and by the time allocated to the completion of the project.

The data was collected over the course of a year, and then modelled as six rolling six-month time windows. While the volume of data was substantial, more data would still be beneficial in order to further the accuracy of the topic clusters formed, or allow for a greater breadth of analysis. The data collected also was not conducive for (re)labelling as the topics had to be extracted from the abstracts and titles and no ground truth data was available. This ground truth data could be inferred in future work either from author-specified keywords, publication journal, or from manual analysis and human labelling.

The time allotted for this project was not sufficient to carry out all the necessary statistical tests to describe uncertainty. The one week allocated for the project provided sufficient time to begin the experiments, however, more time would have allowed for exploring different approaches, iterating on selected approaches to fine-tune parameters, linking the separate parts of the project together to form a cohesive result

and potentially allowed for predictive modelling to be implemented to predict future topics.

## **1.6 Recommendations and Future Work**

Based on the outcomes of this project, future work to improve the automated horizon scanning process and to gain additional insights from the output of the BERTTopic Model could include:

- The completion of additional experiments with different versions of the BERTopic Model, including online topic modelling and dynamic topic modelling. These approaches are designed to consider temporal information as part of the modelling process, either to help cluster topics (dynamic topic modelling) or to allow for continuous updating of models as new data is collected (online topic modelling). These approaches are likely to prove highly beneficial for the discovery project in the future, and may remove the need to re-train a model at fixed intervals.
- Threshold fine-tuning should be performed to ensure the baseline BERTopic model is as optimised as it can be for the task of topic identification. This will include experiments to identify the optimum number of topics (and topic size) that the model should produce. This will help improve the robustness of the model over multiple runs and will assist with any down-stream tasks such as topic labelling or measuring topic dynamics.
- Bootstrapping and hypothesis testing should be conducted on the results from the topic dynamics analysis. This should be performed on the topic groups formed using the distance measures to ensure topics included in clusters are due to genuine signals and not a result of random noise. By comparing the observed results to randomly allocated labels then confidence bounds can be given.
- A combination of automated and human generated labels should be given to the topics in order to both train models which can generate these labels and assess their performance.

## **2 Introduction**

### **2.1 Challenge Background and Motivation**

This Data Study Group (DSG) explored the potential of applying topic modelling to map the scientific landscape. If successful, this approach would allow for automated horizon scanning and trend identification which could highlight the emerging trends in scientific literature which will be of interest to the Defence and Security community. This work is motivated by prior work presented by the UK's Defence Science and Technology Laboratory (Dstl). Drawing upon Dstl's prior efforts, the DSG aimed to advance their work while gaining deeper insights into the research challenges posed by automated horizon scanning techniques and how to overcome them effectively.

The rapid pace of scientific breakthroughs and disruptive technologies demands that efficient monitoring methods are developed to keep pace with the evolving science and technology landscape. Given the sheer volume of information, manual tracking of the latest developments would be impractical, necessitating automated or semi-automated approaches to horizon scanning to reduce the burden on analysts and improve performance through more targeted and quantitative assessments.

Dstl's Discovery Project has explored applying data science approaches to automate aspects of the horizon scanning process. They hypothesise that topic modelling, a form of Natural Language Processing (NLP), could provide a structured framework for exploring and visualising academic literature, enabling analysts to identify emerging or declining technologies and trends that warrant further exploration. By monitoring how topics change over time, this method may allow for faster identification of emerging technologies than manual scanning, giving the UK a competitive edge in assessing the applicability of these technologies to Defence and Security.

Topic modelling, as a form of unsupervised machine learning, employs NLP to cluster documents based on their similar words and phrases, thus providing a solid foundation for building automated horizon scanning methods. Dstl has previously employed BERTopic [12], a machine learning topic modelling technique to scan extensive scientific paper

datasets, extract relevant keywords, and generate topics represented by the top  $n$  keywords. The papers are then assigned into these topics.

However, the focus of this project is continuous monitoring of the scientific landscape, not just a one-time analysis. Topic models have limitations in identifying new topics which are not present in the original training data. To address this, regular re-training of the models is necessary to detect emerging scientific trends and technologies as new data becomes available. By studying a time series of topic model runs, crucial dynamics that may remain hidden in a single run can be uncovered. To facilitate this investigation, Dstl has provided outputs from their cutting-edge topic modelling approach, which was trained monthly over the previous six months of data for a period of a year. This data can be leveraged to create time-series information for analysing the trends in topic development.

The primary challenge lies in identifying and implementing a quantitative approach capable of tracking how topics evolve and change over the topic modelling output. This endeavour aims to significantly enhance automated horizon scanning techniques to gain valuable insights into the dynamics of the scientific landscape over time.

## 2.2 Mapping the Scientific Landscape

Mapping the scientific landscape involves analysing and visualising the current state of scientific research in a particular field, or across multiple disciplines, in order to help researchers, policymakers, and other stakeholders gain insights into the distribution of knowledge, identify research gaps, and make informed decisions about future research investments or trajectories [10].

Quantitative efforts to map the scientific landscape have taken a range of different approaches. Studies have explored the origins and development of scientific disciplines [16], examined how new ideas are discovered [14], measured how scientists collaborate and socialise [2], and the challenges posed by the exponential growth of publications [7]. Various methodologies, including bibliometric analyses [2], network studies [14], and natural language processing [9], have been employed to analyze extensive collections of publications. Recently, the explosion of

information available online, especially via open-access scientific pre-print repositories, has made this type of analysis easier and more powerful than ever [8]. As a result we are now able to quantify the rise and fall of scientific fields with notable accuracy [6], and entire scientific fields can now be analysed and understood in fine granularity [11]

This work has also been developed outside of the scientific sphere, with think tanks and non-profit organisations developing tools which allow anyone to explore the scientific literature through interactive mapping approaches. One such example is the Emerging Technology Observatory's Map of Science [21] which collects and organizes the world's research literature, grouping articles from around the world into clusters based on commonalities that the publications share such as topic, language, authors, or academic citations.

## **2.3 Topic Modelling: What is it and why do we use it?**

Topic modelling is a type of statistical analysis which can aid the discovery of hidden semantic structures in collections of text [26]. Typically, it is used for discovering and analysing the abstract "topics" that occur in a collection of text documents. In many ways topic modelling for text documents is similar to unsupervised classification or 'clustering' for numeric data. In both approaches the aim is to find natural groups of items without necessarily knowing what groups (or how many groups) to look for.

"Traditional approaches" to topic modelling have focused on counting the words contained in a set of documents, and based on the statistical distributions of the words in each document, identifying what the overall topics contained in the corpus might be and what each document's balance of topics is [3]. This allows documents to "overlap" each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language.

More recent "state-of-the-art" approaches have leveraged large pre-trained language models in order to achieve better results for topic modelling [13]. These approaches follow a similar pattern but instead of looking at individual words, are able to consider the context that words

are used in to better model the text in a more ‘realistic’ way. Many approaches use the document ‘embeddings’ - vectors in a high-dimensional space which represent the semantic meaning of the document – and perform clustering over these embeddings. A dense cluster will represent a group of documents that are all similar to each other, likely belonging to the same topic. Topics can be labelled by identifying the words that are most common in each cluster.

## **2.4 BERTopic**

BERTopic is one such state-of-the-art topic modeling method that leverages pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers) to create dense clusters of topics [12]. It is designed to generate easily interpretable topics while representing important keywords in the topic descriptions. BERTopic uses c-TF-IDF (Class-based Term Frequency - Inverse Document Frequency) to rank the importance of words in the topic descriptions. This approach ensures that topics are relevant and coherent, making it easier to interpret and draw insights from the results. BERTopic also provides various topic modeling approaches, including supervised, unsupervised, and semi-supervised techniques, hierarchical, dynamic, and online topic modeling, providing a powerful tool for researchers and data scientists in analyzing and understanding large volumes of textual data.

## **2.5 Comparison of Modern and ‘Traditional’ Topic Modelling Approaches**

While BERTopic is often considered the current ‘state-of-the-art’ within the field of topic modelling, some more ‘traditional’ approaches can still provide useful benefits. Latent Dirichlet Allocation (LDA) [3] has been the most commonly used topic modeling technique due to its ease of implementation and the many variations that can be tailored to specific use cases. LDA topic modelling is a bag-of-words approach that considers probabilistic modeling to assign topics to individual words in the document. Documents are then assigned topics due to the distribution of the words of which they are constructed. Due to this heavy reliance on individual words, and the frequencies in which the words appear, LDA

topic models can be relatively easy to interpret and can scale well to large volumes of data. This reliance on word frequencies is also a limitation however, as the same words can be used in different ways in different contexts and as a result LDA topics models can result in less coherent topics and may not capture the underlying semantic structure of the data.

In comparison BERTopic uses pre-trained models which are able to extract semantic relation between text units and thus produces more coherent topics [12]. This makes it a more flexible and adaptable tool than traditional topic modeling approaches. Unlike traditional LDA topic modeling, which create statistical models, BERTopic uses advanced techniques such as Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [19], Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [18], and Sentence Transformers [24] for topic modeling. Another key advantage of BERTopic over LDA is that it provides continuous topic modeling, as opposed to the discrete topic modeling provided by LDA. This means that the models can be updated continuously over time, although the stochastic nature of the model can lead to different results with repeated modeling. Moreover, BERTopic supports any language for which an embedding model exists, and the approach can therefore be expanded upon as needed.

## 2.6 Challenge Summary and Objectives

The key challenge for this DSG is to identify and evaluate the approaches that can track how topics within a corpus of scientific papers evolves through time by comparing different topic models trained on temporal segments of the data. These approaches should be able to flag the significant differences between different topic model runs, including when new topics are emerging, older topics are reducing, and existing topics are converging or diverging. These dynamics might be represented by topics moving closer to one another, further apart, new and unexpected connections forming between existing topics, or sub-fields within a topic receiving differential focus from different areas of the scientific landscape.

Re-training topic models to address these questions introduces certain



challenges. Each re-training run on new data can result in a different number of detected topics, and the composition of these topics may vary. Even if the underlying latent topic remains the same, the prominent keywords in the topics can differ, and the boundaries between topics might shift. Distinguishing between shifts caused by noise in the data and those indicating a genuine change in the science and technology landscape is crucial. Additionally, novel and emerging topics are likely to be underrepresented in the data, especially initially. Therefore, techniques to detect these trends must be sensitive enough to measure their emergence accurately, without becoming overly sensitive and rendering topic model runs incomparable from one iteration to the next. Striking the right balance in sensitivity is vital for obtaining reliable and meaningful insights from the re-training process.

The DSG project has three main objectives:

1. Identify and implement a method, or set of methods, which can map topics from one topic model run to another, identifying which topics are the same across runs and which are different.
2. Identify and implement a method, or set of methods, to explore the changing dynamics of the topics across two (or more) runs. This could include identifying when topics are emerging or disappearing, when topics are converging/diverging, when new (possibly unexpected) connections between topics have formed, or when topics have changed in size and prominence in the dataset.
3. Develop methods which can help analysts to interpret the outputs from the model, which are currently focused on keywords and unlabelled topics and converting these to human readable semantic topic labels.

As the science and technology landscape continues to rapidly shift and evolve, this challenge will help policy makers to stay on top of these latest developments.

## 3 Data Overview

### 3.1 Dataset Description

#### 3.1.1 Data Overview

The project gathered a comprehensive dataset from Semantic Scholar - a public repository of academic literature and scientific papers - consisting of the titles and abstracts from 559,183 papers from six fields: Chemistry, Mathematics, Computer Science, Material Sciences, Physics, and Engineering. Semantic Scholar was selected as the data source as it represents one of the largest available repositories of academic literature on emerging technologies, and also as it makes this data available via a user-friendly API (Application Programming Interface). The six fields listed above were chosen as they represent those fields where emerging technologies most pertinent to Defence are likely to be published. Data was collected for all papers published into one of the fields of interest between January 2022 and January 2023.

In addition to the Semantic Scholar data the the DSG is provided with the output from Dstl's current 'BERTopic' topic modelling approach. This approach is run six times over sequential 6-month-long rolling windows from this data. In total this gives six outputs (January-May, February - June, March - July and so on). This data is provided to the DSG in the form of two main datasets:

**Raw Data:** The raw data which was fed into the BERTopic model are abstracts and titles from published academic papers which are publicly available on Semantic Scholar. The data covers a one-year period from January 2022 to January 2023, and it includes information such as paper ID, publication date, title, abstract, field ID, and author ID.

The specific data collected consists of the text of the abstract from the published papers, along with metadata including the publication date, the paper title, the publication category, and anonymised author information.

**BERTopic output data:** Prior work on applying BERTopic to this dataset was completed by Dstl's in partnership with the University of Warwick. Within this work the University of Warwick trained a BERTopic model on

the raw data and obtained the output data which is described below and which formed the basis of the data used on this project.

Running the BERT Topic Model on collected papers from Semantic Scholar through six iterations, generated the output data for this challenge. Each of the six model iterations produced two files: a probability dataset which contained the paper number, paper id, topic number, 'top  $n$  words' (the first 4 words of the keywords which served as the topic label), probability of a paper being assigned to a topic, and keywords, and a topics dataset, which contained information on the topics identified by that model. In addition, participants were provided with a **papers.csv** file which contained the title, abstract, collection date, and anonymised authorship information for all papers included in the collection.

The BERTopic output datasets were formatted at follows:

1. **topics.csv:** This file comprises the list of topics identified during each model run. It includes four fields:
  - (a) Number: In each run, the BERT model identifies around 500 topics. Each topic is assigned a numerical value and paired with a label in the subsequent column. The number of topics is not directly controlled by the model, and so can vary from run to run. However, model parameters specifying the size of the smallest possible topic in the model (the number of documents assigned to it) are given which provides a degree of stability across runs.
  - (b) Label & Keywords: This column provides a readable summary of the topics within the cluster, such as "3\_plasma\_discharge\_ion\_plasmas." It is paired with a keywords column that contains more detailed keywords related to this cluster. For example, the corresponding keywords for this label would be: "plasma, discharge, ion, plasmas, beam, electron, tokamak, divertor, discharges, electrons."
  - (c) Embedding: The prior modelling work has also included the sentence embeddings outputted by the current model in use. These embeddings are represented as a 384 dimensional vector which reflects the hidden dimension size used in BERT.

The specific implementation used by BERTopic is the all-MiniLM-L6-v2 from SentenceTransformers which uses a modification of the pretrained BERT network to map sentences and paragraphs of text to 384 dimensional dense vector space using siamese and triplet network structures.

(d) Weight: The number of papers associated with each topic.

2. **probs.csv**: This file includes four fields: Paper ID, Number, Probability, and Main Topic.

(a) PaperID: A reference to the Paper ID found in the papers.csv file.

(b) Number: A reference to the topic in the topics.csv file.

(c) Probability: The probability that the identified topic is the main topic within the paper, as determined by the BERTopic model.

(d) Main Topic: A binary field indicating whether the identified topic is the main topic within the paper.

To gain a thorough understanding of the project's data and develop a methodology, the team conducted an initial review and exploratory data analysis. During this process, as shown in Figure 1, we calculated the count of papers for each field and observed a notably larger number of papers related to Computer Science compared to other fields. As it can be seen in Figure 2 this pattern was found for each month in the 12 month observation period.

As depicted in Figure 3, we also visualised all of the papers in the collection based on their publication dates throughout the entire data collection period. Notably, we observed a substantial surge in publication dates occurring on the first day of each month. The team inferred that this anomaly stemmed from missing publication information in the dataset; day-of-publication data were either missing on initial submission to semantic scholar, or truncated at some point in the collection process.

In order to explore the papers associated with each run we re-created the dataset buckets used to train each topic model, the number of papers in each slice of data is shown in Table 1.

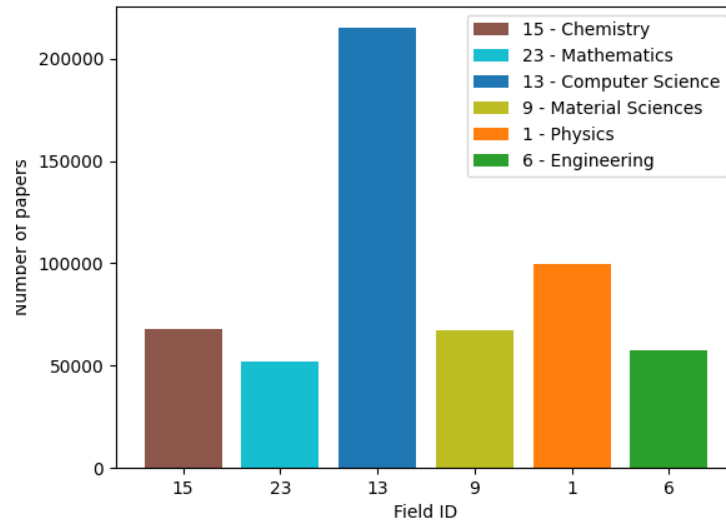


Figure 1: Distribution of Scientific Papers by Semantic Scholar Field

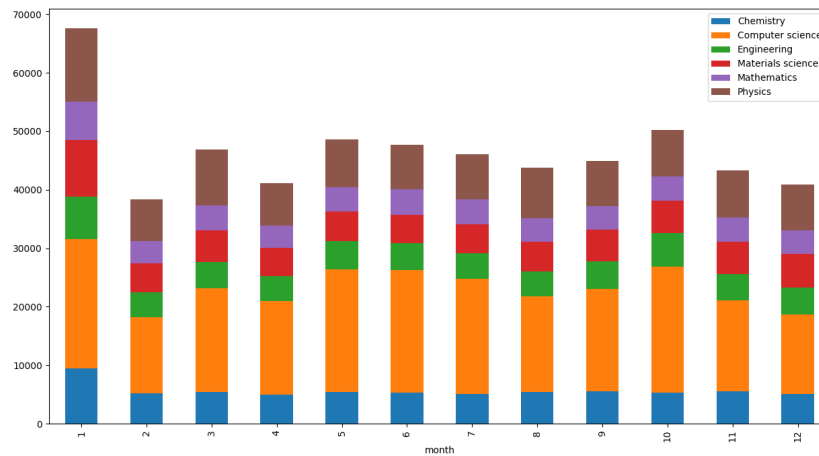


Figure 2: Distribution of Scientific Paper by Month of Publication

### 3.2 Data Quality Issues

Initial investigation into the dataset highlighted several points that indicate issues with the quality of the data. An initial completeness check of the data showed that approximately 32% of abstracts were missing from the

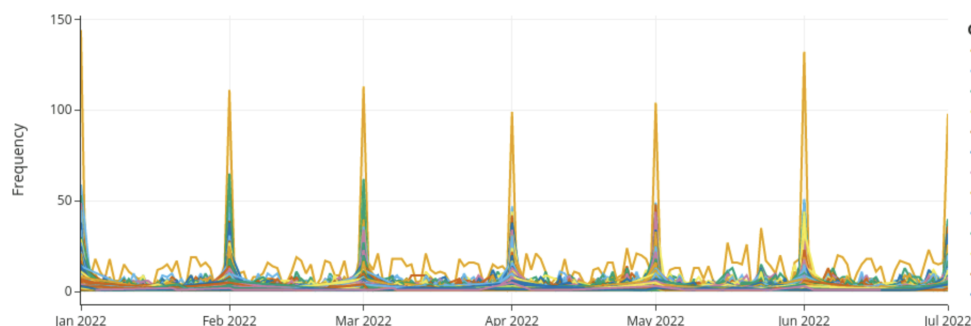


Figure 3: Distribution of Topics over Time

Table 1: Augmented six bucket data

Run	Start Date	End Date	Number of Records
1	March 1, 2022	August 31, 2022	273,998
2	April 1, 2022	September 30, 2022	225,816
3	May 1, 2022	October 31, 2022	281,186
4	June 1, 2022	November 30, 2022	275,894
5	July 1, 2022	December 31, 2022	269,157
6	August 1, 2022	January 30, 2023	247,346

data. This missing data is likely to affect the topic modelling process as these abstracts were the primacy signal used to both identify the topics, and assign papers to topics. Whether or not this data is missing at random is likely to be key to determining the impact of this on the overall horizon scanning process.

### 3.2.1 Distribution of missing values

Exploratory data analysis showed that the abstracts field is the only field with substantial missing data, suggesting that the issue does not affect the overall collection and is unlikely to be the result of data corruption on transfer into the DSG compute environment as the other fields remain unaffected. Overall, 460,613 out of 559,183 paper entries contained an abstract. We also found that the missing abstract entries were fairly evenly distributed among all time periods and across all scientific fields, suggesting that this error does not reflect a single point of collection or

single source failure, as shown in Figure 4. We also found that when the topic model could not assign a topic to a paper, resulting in a NAN value, these values were fairly evenly distributed over time, and tracked the overall publication data volumes in the dataset, as shown in Figure 5.

Overall, we conclude from this that this missing data appear to occur at random with no clear pattern that we can observe, and as such we continue the analysis bearing this in mind.

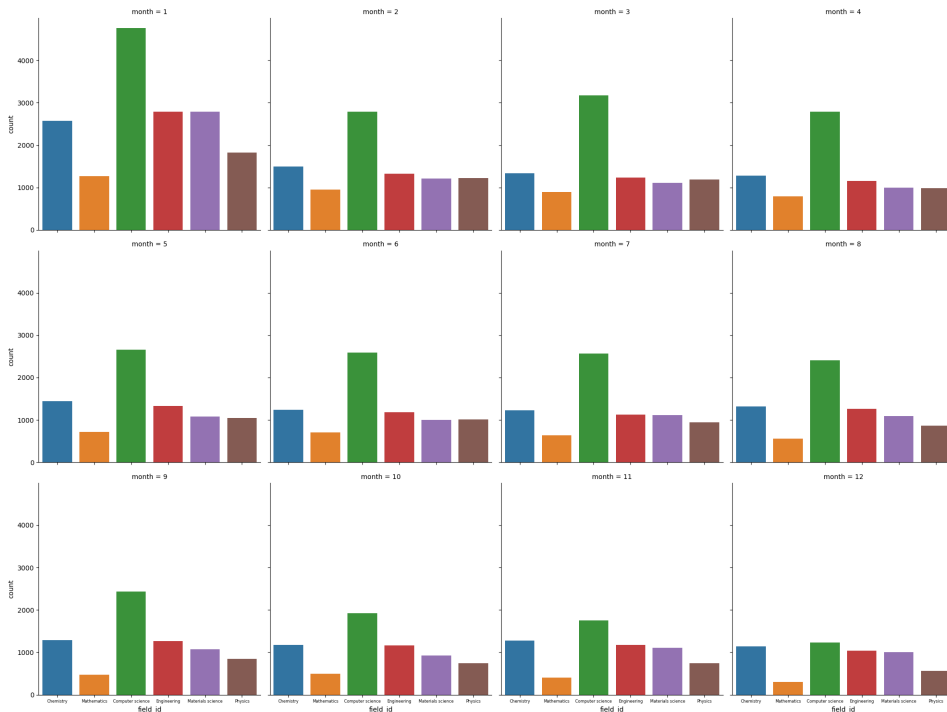
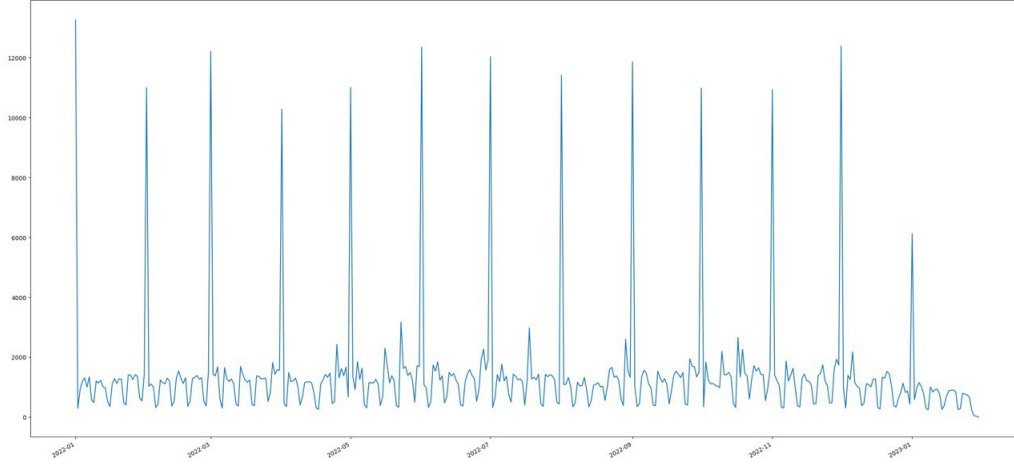


Figure 4: Missing Values by month (facet) and scientific field (colour)

## 4 Approach 1: Topic Dynamics

### 4.1 Goal

We first sought to develop metrics that could define spatial and temporal relationships between topics within and across runs of the BERTopic



*Figure 5: Proportion of papers which cannot be assigned to a topic over time*

model. This addresses the primary purpose of the project to identify convergence, divergence, emergence, and disappearance of topics.

The first stage of this is to identify when the same topic appears in multiple runs, and mapping these topics from one to another. In doing this we will then be able to expand the analysis to answer more complex questions around topic dynamics. The primary approach that we explored for mapping topics across runs was to calculate the distance metrics ‘cosine similarity’ and ‘Euclidean distance’ between the embeddings of each topic in and across runs to quantify their relationship. Here we make the assumption that topics which are the same across runs will have high similarity metrics and/or small distance metrics between their embedding spaces.

Cosine similarity measures the orientation of two n-dimensional vectors. It is calculated by the dot product of two numeric vectors, normalized by the product of the vector lengths, so that output values close to 1 indicate high similarity. The equation for cosine similarity is given as follows:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} \quad (1)$$

The Euclidean distance in contrast is a distance measure between two points or vectors in a multidimensional (Euclidean) space based on



Pythagoras' theorem. The distance is calculated by taking the square root of the sum of the squared pair-wise distances of every dimension as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2)$$

Subsequently it is possible to use this information to perform more complex analyses, such as to establish parent-child relationships in a non-Markovian way, where the future state is not influenced by the present state. For example, topics in run 1 and run 3 or run 1 and run 4 could be linked without the requirement of having informative links between consecutive runs e.g. run 1 and run 2, run 2 and run 3, run 3 and run 4 etc.

Importantly, these parent-child relationships can be formed in two, related, ways. Firstly, within a single static topic model run child topics can be thought of as nested within parent topics. For example, a 'differential equations' topic and a 'formal logic' topic could both be considered child topics of the larger parent topic 'mathematics'. Secondly, over time, past topics (parents) can influence the nature and position future topics (children). In this latter consideration there is no reason why the child topics cannot become larger and more prominent than the parent topics. This could be demonstrated in cases where theoretical scientific topics gain larger and applied uses, such as the parent topic of 'matrix manipulation' and the child topic of 'Artificial Intelligence'.

## 4.2 Scientific Disruption

Here we are motivated by Kuhn's theory of scientific revolutions [15], where new ideas or theories build upon the existing ones, systematically replacing them. Therefore, we want to assign a metric to each topic that defines its strength of connection from the past topics (parents) as well as the future topics (children).

To do so, first, we leverage from the above mentioned similarity measures cosine similarity and Euclidean distance across topics and connect them if the similarity scores are greater than a predefined threshold (to allow for noise in the model). This allows us to view the topical landscape as

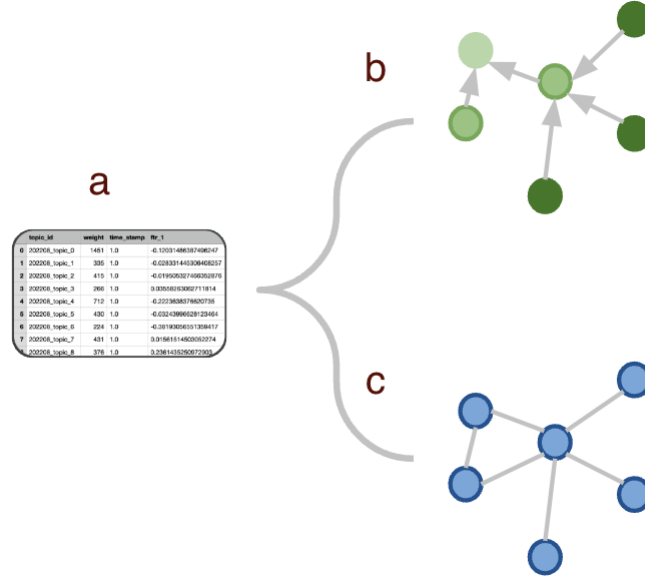


Figure 6: How disruptiveness is measured from a dataset of academic papers. (a) Represents the data-set consisting of topics represented as feature vectors with time stamps. (b) We create a directed network of topics where each node is a topic and an arrow between  $node_i \rightarrow node_j$  represents that  $node_i$  is a child of  $node_j$ . (c) Represents an un-directed topic similarity network within the same time stamp.

a complex network. Next, we quantify the changes in topics' evolution and identify the topics of interests using methods from complex network analysis.

Our analysis is split into two parts - a) focusing on the temporal data connecting the topics from different runs and b) focusing on static data connecting the topics from the same run (Figure 6).

For the temporal case as demonstrated in Figure 6 (a) we propose to use the Disruptive Index measure [30] and for the static case as demonstrated in Figure 6 (b) we propose to use the Average Weighted Distance similar to the cognitive distance measure in [27].

The formula for calculating Disruptive Index (DI) is given below, where D is the disruption of the focal paper, SC is the number of times that the other

paper cites just the focal paper and does not cite its references, DC is the number of times that the other paper cites both the focal paper and any its references, and PC is the number of times that the other paper cites just any of the focal paper's references from the year after the focal paper is published [31]. DI ranges from -1 (low) and (+1) high.

$$DI = (SC - DC)/(SC + DC + PC) \quad (3)$$

Average Weighted Distance (cognitive distance)  $C_{i,j}$  between topics  $i$  and  $j$  is calculated as:

$$C_{i,j} = \sum_e \frac{1}{W_e} \quad (4)$$

Which gives the weighted shortest path between the two topics  $i$  and  $j$ , where  $e$  are the edges on the shortest path and  $W_e$  are their weights in the co-occurrence network.

The rationale behind the disruptive measure for topics in the temporal setting is as follows - by its definition the disruptive measure indicates how much new information a paper is carrying compared to its predecessors. If the measure is high for a paper it is considered to be 'disruptive' for its domain (of more interest to the future papers than its predecessors) whereas if the measure is low the paper is considered more 'consolidating' or 'unifying' for its domain [4]. The same can be extended to topics. Here a topic with high 'disruptive' measure could be an indicator of emergent fields hence can be considered as topics of interest.

On the other hand for the static case where we connect topics within the same run we propose the average weighted distance measure between topics. Since the weight of an edge in such a network would indicate the strength of similarity between topics, inverse of this weight can be considered a proxy of distance between two topics. Therefore in the static network of topics - first, we cluster the topics based on the similarity measure using the community detection algorithms. Second, for topics within the community we find the average weighted distance across topics and highlight the topic with highest average distance to all the other topics within the community. Such a topic could be an indicator of a new topic in the cluster or a rare topic with few researchers working on it.

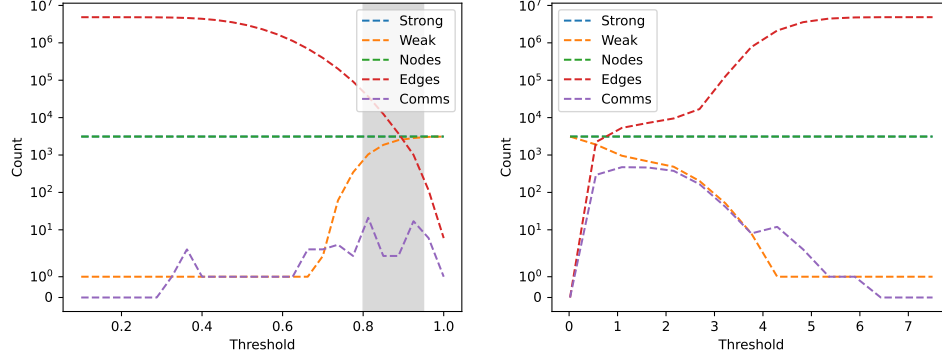


Figure 7: (left) Network stats for directed networks using the *cosine* similarity networks as we span the domain of the threshold. The same is represented on the (right) except that the similarity score here is  $l_2$  - norm

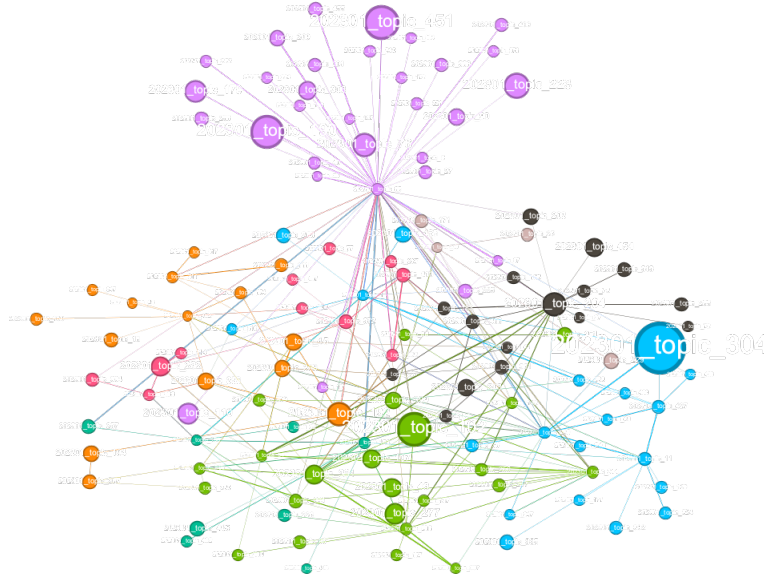
We believe that by using the above two measures, we can identify the significant emergent topics which were not present in the previous round, or existing topics of interests, both in the temporal setting and static setting. However, there are few points to be considered before implementing them.

Firstly, the measures will rely on the network construction and threshold used to limit edges. Therefore, defining a clear threshold is important. In our study, for the networks we monitor major network statistics as they vary with the threshold using both cosine similarity and Euclidean distance ( $l_2$ -norm) as shown in Figure 7.

We stick to the range of threshold where we get maximum number of communities with the minimum size of the community being 3 topics. This is to avoid isolates in our analysis. The same method is used for the static networks.

Second, we need to define the null model for our metrics. One such model could be random networks constructed by shuffling the edges of the network preserving the degree and comparing the values of the disruptive index and average distance against the empirical values.

Third, as we know from the literature that the scientific growth observes a memory effect [22] where information is transmitted across multiple steps in the markov chain. Therefore the strength of topics connected across



*Figure 8: Example static topic analysis, connecting and visualising the network of topics within the same run*

different times is not uniform. To account for such memory effects we propose to use a memory kernel such as  $e^{-\lambda t}$  within our measures where  $\lambda$  controls the temporal effect.

We believe that these measures combined with the TDA-Mapper discussed later could be used to identify emergent topics. Further work is needed to implement these metrics in practice in a 'real-time' environment, but Figure 8 gives an example visualisation for the static use case over a single topic model run. In this example topics are grouped and coloured into 'fields' of similar topics through a community detection algorithm (modularity) and sized according to the number of papers within these topics in the corpus.

### 4.3 Topological Data Analysis - Mapper

In attempting to identify evolutionary changes in topics over multiple runs, the team attempted to employ topological data analysis (TDA) through mapper [28] to the topic embeddings produced by the BERT model. This was attempted by first extracting the 384-dimensional vector embedding into an array. Following this, we pair each of these arrays with the corresponding topic number and employ the mapper algorithm [29], where we cluster based on the embeddings, and colour the output graph based on the topic number.

The intuition behind this approach is similar to the methodology used in [17] in identifying subgroups of Type-2 diabetes. Our expectations were that we could use the output of the Mapper graph in each run to identify linkages and relationships between related topics where the underlying clustering would be based on the topic embeddings. We would then compare the output of the Mapper algorithm across each of the 6 runs, and track the changes in the identified clusters based on the topic embeddings. This would ideally allow us to identify changes in topics over multiple runs.

The main advantages motivating the use of TDA are the robustness of this approach when paired against noisy data, and its ability to analyse data at multiple scales which could be beneficial in analysing varying levels of similarities across the topic runs - allowing for both local and global effects across the corpus to be considered. However, as TDA is intended to capture the global shape of data, it could miss the local anomalies that may indicate emerging topics or trends. As highlighted by [5] topic modelling combined with TDA Mapper can produce sub-graphs of inter-related topics based on the topic embeddings as shown below in Figure 9.

Our approach to applying TDA Mapper to the task consisted of the following steps:

1. **Dimensionality Reduction:** The BERT embeddings are 384 dimensional arrays, making them difficult to analyse. By embedding into a lower dimensional space, we can preserve the structural information and apply techniques for analysis. We achieve this by using Principle Component Analysis (PCA), t-distributed stochastic

Topic Subgraph: jews, armenian, armenians, turkish, military, people, population, israel, army, town

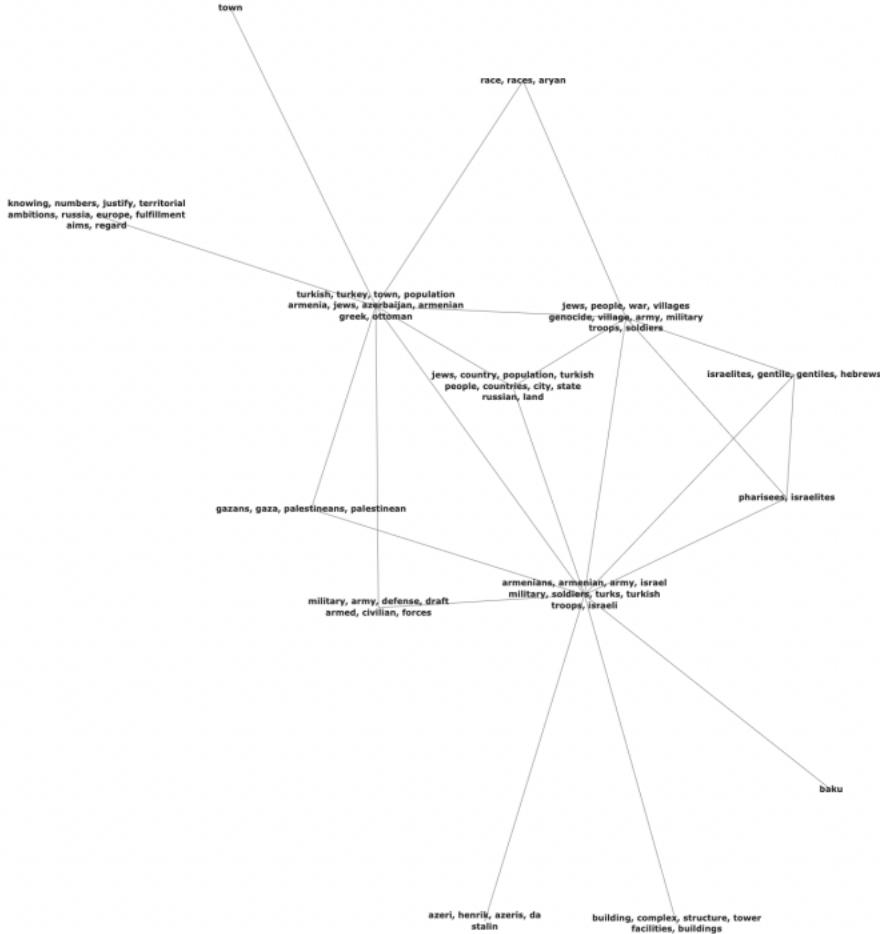


Figure 9: Topic Subgraph based on TDA Mapper + BERT

neighbor embedding (t-SNE) and UMAP.

- 2. Defining the Cover:** The cover defines the subset of data to be analysed. A subset of data is used for computational efficiency and ease of interpretation. We experimented with the number of intervals, and overlapping segments of data to identify the most appropriate cover for the experiment.

3. **Mapper Graph Construction:** Using the reduced embeddings and cover, the Mapper graph is constructed where the nodes represent subsets of the data, and the edges represent the overlapping or similar topics in the dataset.
4. **Clustering and Topic Identification:** We then applied clustering algorithms, including K-means, DBSCAN, and spectral clustering, to the nodes of the Mapper graph to group similar data points and topics. These clusters represent distinct topics or themes within the corpus, forming the foundation for further analysis.
5. **Convergence and Divergence Analysis:** Using the edges in the Mapper graph helps to determine the strength of connections between clusters, indicating topic convergence or divergence. Strong connections imply shared characteristics or themes between topics, while weak or non-existent connections suggest unrelated topics.

## 4.4 Word Mover's Distance

Cosine similarity is suggested to measure similarity between embeddings of research topics. However, this approach does not consider the order of words within a sentence, and so the results may not be as accurate as other approaches. Cosine similarity metrics typically use Bag-of-Words representation which are based on only embeddings (or vectors) without the orders of words taken into account, leading to little relationship understanding. In that case, topics that have different word orderings or use different vocabularies may still receive high scores, or topics that cover similar concepts but use different words or phrasing may be assigned low scores.

Here, to better describe the similarity between the topics, Word Mover's Distance (WMD) was applied to the BERTopic result. In short, Word Mover's Distance is used to measure the dissimilarity between topics based on the distance needed to 'move' from one set of word embeddings to another. The lower the WMD value, the more similar the topics are. The following figure shows a typical example.[\[23\]](#)

The input of WMD metric are the embedding model(s), which is usually



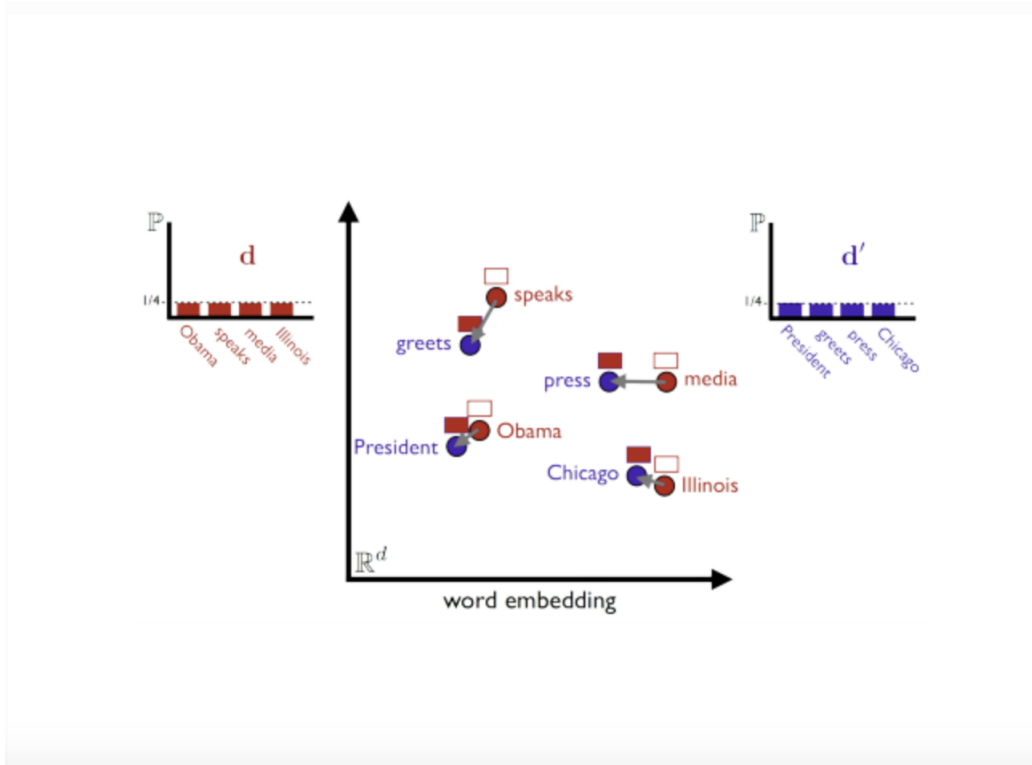


Figure 10: The examples of using WMD

pre-trained model(s), and the two pieces of information that are waiting for dissimilarity comparison. In BERTopic, the default embedding model is all-MiniLM-L6-v2 created by SentenceTransformers [25]. And there is also a mature Python package, 'gensim' [23] that can provide a fast implementation of WMD.

During the experiment, problems occurred as unmatched embedding models reveal irrelevant latent semantic information when embedding source data. As an alternative, a word-based embedding metric, such as Word2vec, should be applied to the source text dataset to extract word-based semantic information. Currently, the BERTopic applies all-MiniLM-L6-v2 from SentenceTransformers which extracts sentence-based semantic information. Therefore, the robustness of the model will be very poor and the results will not be interpretable.

As it is impractical to re-embed the source data within such a short time,

there might be other options to improve this experiment.

Firstly, one can approximate the Word Mover's Distance (WMD) between two sentences using sentence-level embeddings to avoid the heavy re-embedding work. To reach that, it is advised to tokenize the sentences into individual words by breaking them down into constituent tokens. Afterwards, calculate the sentence embeddings by averaging the embeddings of the words in each sentence and use the mapping created in the previous step to associate each word with its corresponding sentence-level embedding. Then take the average of these word embeddings to obtain the sentence embedding, and finally use the `wmdistance` function from the `gensim` library to compute the WMD between the two sentences. The user can then provide the sentence embeddings as input to the function, which will calculate an approximate measure of the semantic distance between the sentences.

Secondly, it might also be possible to aggregate WMD from sentence-based embeddings by calculating the average or sum of the WMD values between each pair of sentences in the sets, which will provide a measure of similarity or dissimilarity between the sets of embeddings.

## **5 Approach 2: Data visualisation**

The data visualisation team took two primary approaches to visualize the topics and the scientific papers contained within the dataset. The first approach took a dataset-centric approach and sought to visualise the topics using the outputs of the models. The second approach took a model-centric view and used the BERTopic models themselves to identify the distribution of topics, and the distribution of documents within these topics. These approaches are described in the sections below.

### **5.1 Dimensionality Reduction - UMAP**

BERTopic generates embeddings, high dimensional vector representations, for each document within the dataset. However, given that these embeddings exist in higher-dimensional spaces, it makes it challenging to visualize and interpret the data intuitively. This is

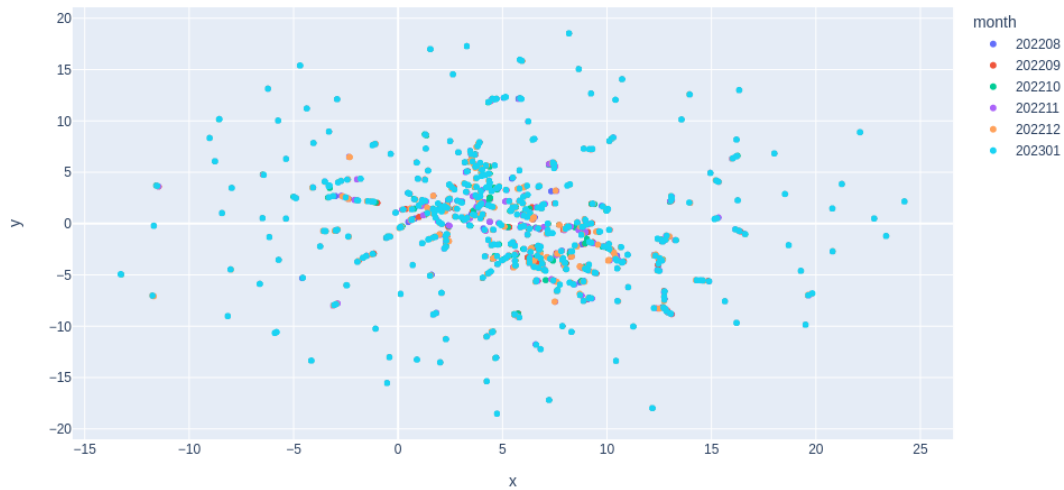
particularly true when the results must be presented to subject matter experts or analysts who are not familiar with data science and machine learning techniques. To address this issue, dimensionality reduction techniques can be employed to transform the embeddings into lower-dimensional spaces. UMAP (Uniform Manifold Approximation and Projection) [20] is one such approach which reduces the dimensions of the topic embeddings down into a much lower dimensional space which preserving as much of the signal as is possible, enabling more effective visualization and interpretation of the topics.

Here we used UMAP with a Euclidean distance metric to reduce the topic embeddings down into two dimensions. The resulting reduced-dimensional representation is illustrated in Figure 11. In this projection, each topic is represented by a point, and topics which are more similar to one another are positioned closer together.

By looking at the evolution of topic distribution within this space over time, it is possible to get a sense of how the scientific landscape is changing. Some initial analysis of this reveals the emergence of new topics stemming from existing ones, while a substantial portion of the topics remains consistent over time. For a more interactive visualization of the plot, an interactive version is available within the outputs. This observation aligns with expectations; a moving average of papers spanning a period of six months was utilized and as such, this moving average serves to smooth out potential spikes or seasonal variations in the data.

When employing UMAP for dimensionality reduction, it is important to consider reducing the dimensions of the topic in a single run and not performing separate pre-processing steps on each run of topic model outputs. This approach ensures that the dimensionality reduction is conducted uniformly across all time points, allowing for meaningful comparisons over time. If dimensionality reduction is performed separately then it is unlikely that the identified dimensions across different UMAP projections will be directly comparable.

Overall, we find that this is a relatively simple and straightforward approach to quickly visualising different topic model runs to get an overview of the topics contained within and to compare across runs effectively. The drawback of this method is that the x and y axis are



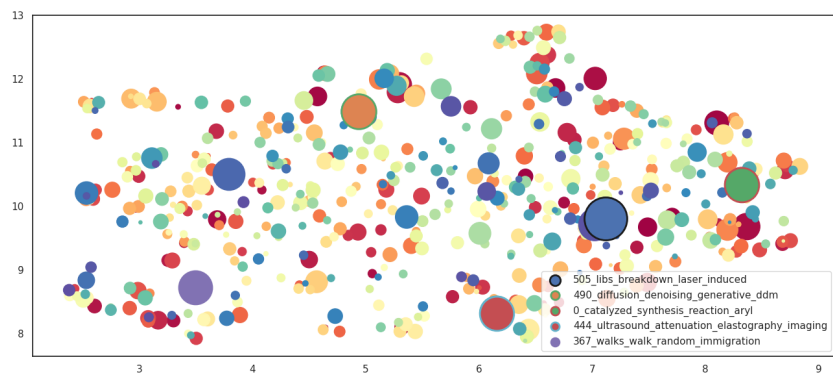
*Figure 11: Topic Distribution over Six Topic Model Runs. Each run is shown in a different colour*

effectively meaningless in themselves, and simple serve to demonstrate relative similarities and differences, and therefore any further analysis to between understand the data must be done manually by the analysts.

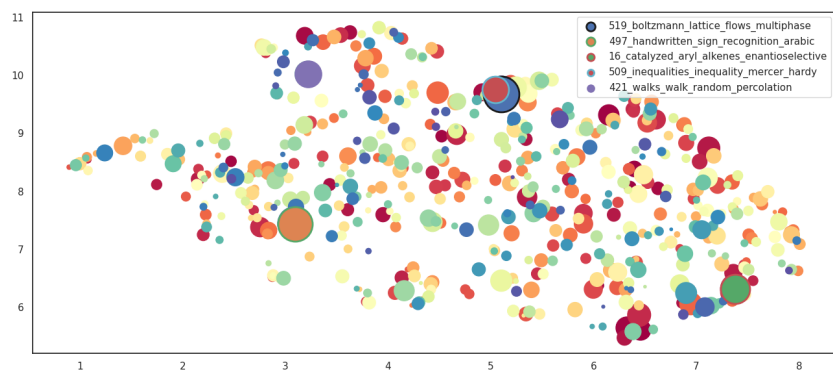
## 5.2 Weighted Topic Embeddings

In order to improve the readability and utility of topic embedding visualisations it is possible to add additional information into these plots. One approach is to assign weights to the topics based on the number of documents that each topic contains. In this way, the weighting scheme aims to capture the importance or significance of each topic in the overall representation and give it a greater visual prominence.

This makes intuitive sense for the reader/analyst. For example, if a particular topic is mentioned in a large number of documents, it will have higher weight assigned to it. This indicates that the topic is more prevalent and widely discussed in the corpus, making it relatively more significant in the overall representation.



*Figure 12: Weighted Topic Embeddings - Run 1*



*Figure 13: Weighted Topic Embeddings - Run 2*

This process is demonstrated in Figures 12 and 13 for sequential runs of the topic model. Each circle represents a topic in the corpus, and topics which are more similar are positioned closer together. As the figures illustrate, the largest topics in the collection vary as time progresses. This implies that the relative significance of different topics is not static but rather evolving over time. In other words, the “hot” technologies, which refer to the trending or popular technologies, are not fixed and are subject to change over time. Importantly this can occur even if the overall distribution of topics in the corpus does not change. The weights

assigned to different topics reflect these shifts, indicating the rise and fall in importance or interest in various technologies as time passes.

In this way, adding topic weights into the UMAP low-dimension visualisations we can add important information and assist readability of the plots.

### **5.3 Connectivity**

Beyond adding topic weight information, it is also possible to use UMAP to visualise more specifically how topics are related to each other. This process relies on making use of the fact that in performing the dimensionality reduction UMAP constructs an intermediate topological representation of the approximate manifold the data may have been sampled from. This structure can then be simplified down to a weighted graph, which can then be visualised in order to identify which topics are closely related via these weighted connections.

As a result of this, connectivity graphs from the UMAP library can be used to identify the relationship between topics. Those topics which are closer to one another in meaning are also closer together when plotted in two-dimensional space, while similarly the strength of the connections between these topics are shown by the density of connections between nodes.

Connectivity graphs for two runs are shown in Figure 14 and Figure 15. While this connectivity information adds additional insight into the plots, notably highlighting how large topics can be connected even if far away in this low-dimensional space, it is not certain that this adds clarity to the overall plots when compared to the weighted topic embeddings given in the previous section.

### **5.4 BERTopic Model Visualization**

While the first three sub-sections of the visualisation approach focus on using the model outputs and embeddings to visualise the distribution of topics, it is also possible to use a combination of model outputs and model object files to view the distribution of academic papers which make up

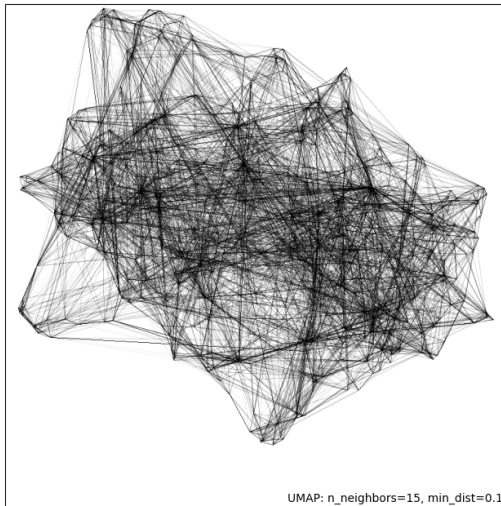


Figure 14: Topic Connectivity Run 1

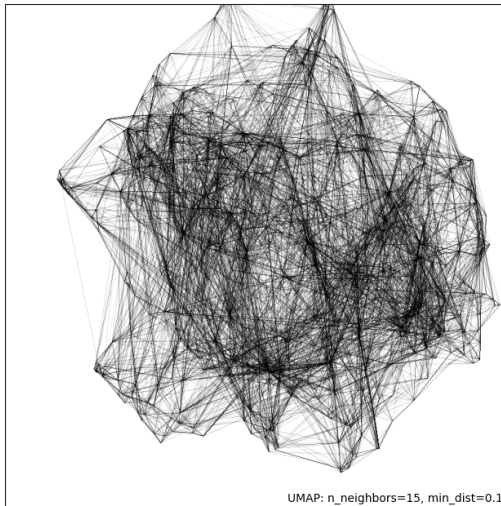


Figure 15: Topic Connectivity Run 2

these topics. This is a far higher level of granularity, and so provides more insight into the dataset at the expense of processing time and clarity.

Using the six models trained, the team initially used the visualization methods through BERTopic to better understand the corpus the models were trained on. The figures depicted below are screenshots of the interactive HTML plots provided in the *Output/Visualizations* folder of our project. These plots allow a user to enable more granular filtering of topics and time frames.

To identify the initial clustering of topics based on the BERT model, the `BERT_Model.visualize_documents()` function creates an interactive representation of topical relationships based on the training data. This visualization method is useful in framing an understanding of topics but lacks the relationship or hierarchical linkages that exist between topics. This figure is similar to the UMAP projections shown in the previous step, however in this case each point on the two-dimensional projection represents a single academic paper from the corpus used for a single run of the topic model. This gives a highly granular view of the dataset, but the high density of data points makes interpretation more difficult.

To understand the hierarchical relationships of topics, the team then used

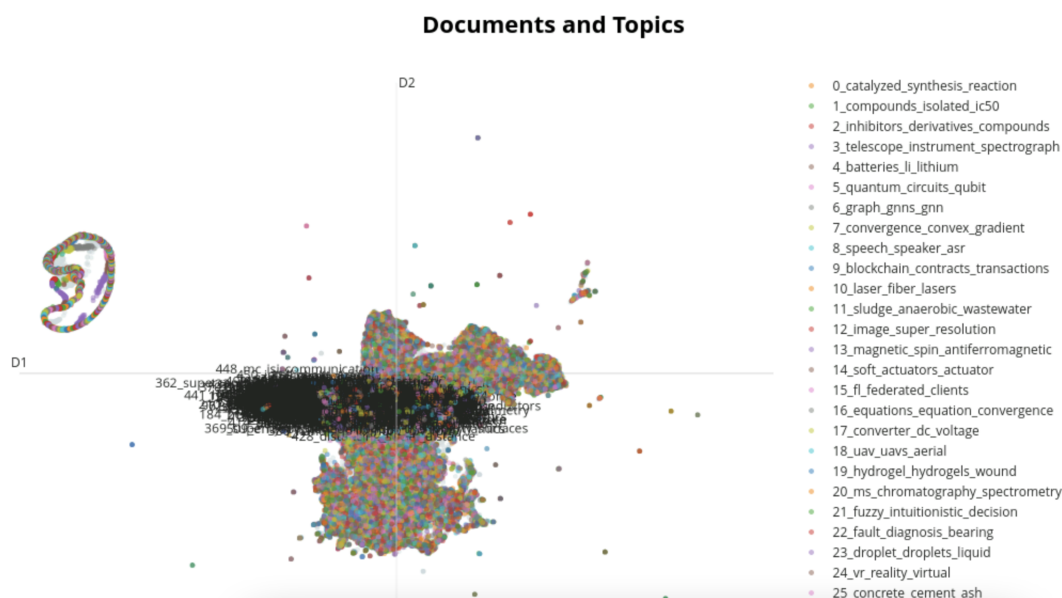


Figure 16: Two-dimensional projection of documents and topics

the `BERT_Model.visualize_hierarchy()` function which gave an initial overview of the connections between topics. This is computed by taking the cosine difference of topic embeddings. As shown in the diagram below, there is an initial observation that the cosine similarity between topic embeddings can group similar topics within the same hierarchical cluster. Here we only present a small sub-section of the overall topic hierarchy tree for simplicity.

Based on the first run of the model, the team computed the top word scores for the first eight topics to better understand what these topics represent and how they are constructed. This process could be repeated for each topic model run to give insight into both the topics themselves, and the keyword which have contributed to the formation of these topics.

Additional visualizations of the data can be generated using the `BERTopic` functions, and examples are stored in the team's *Output/* folder. These visualizations include:

1. Term Loss
2. Heat Map/Similarity Index



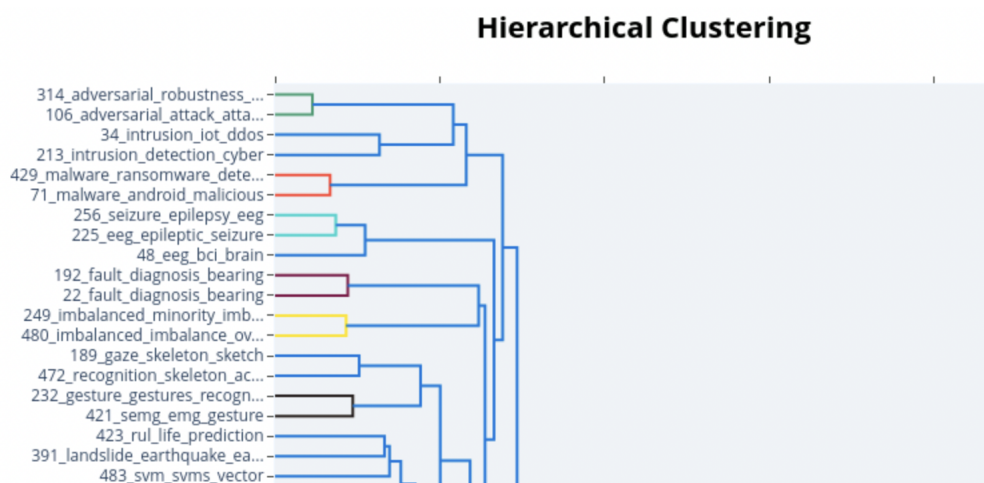


Figure 17: Hierarchical clustering of topics

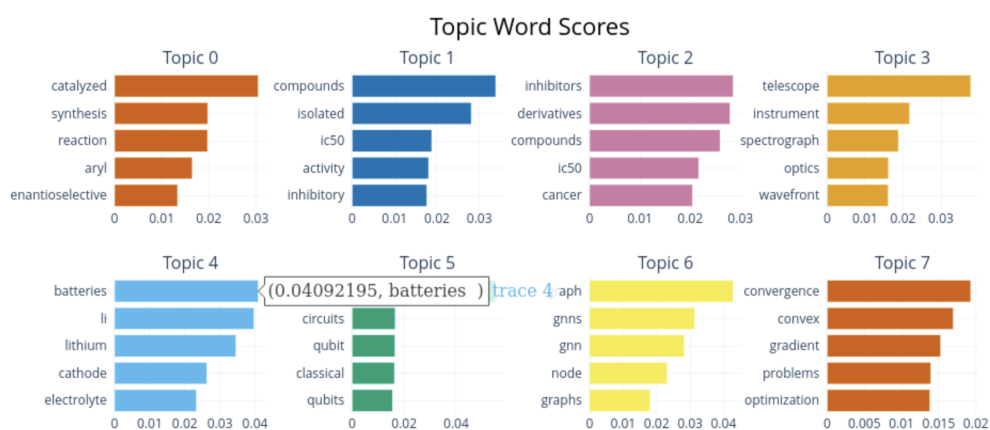


Figure 18: Top word scores for each topic for a single run. Note unlabeled nature of Topic names (0, 1, 2, etc)

### 3. Topic Word Scores

#### 5.4.1 Interpreting BERTopic Model runs

Using the visualisation approaches outlined in the previous section it is possible to interpret the change in topic distributions over time. Here we

evaluate six BERTopic model runs to visualise topic changes and differences in document clusters among the runs.

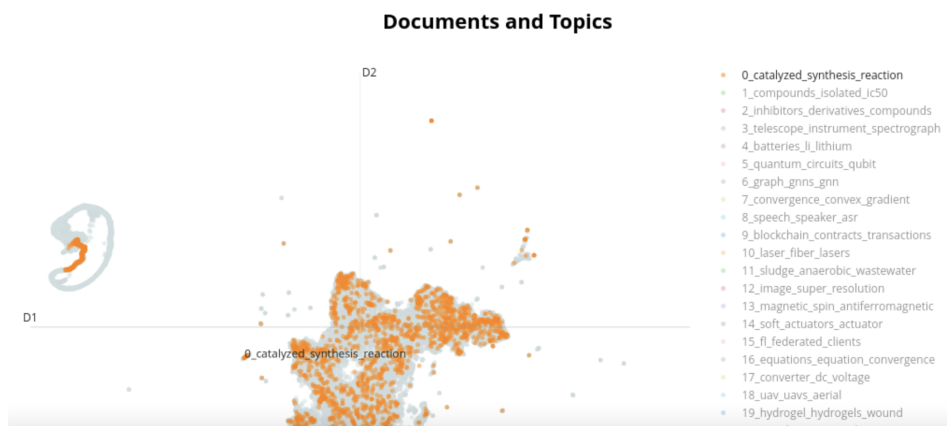


Figure 19: First model run after applying UMAP

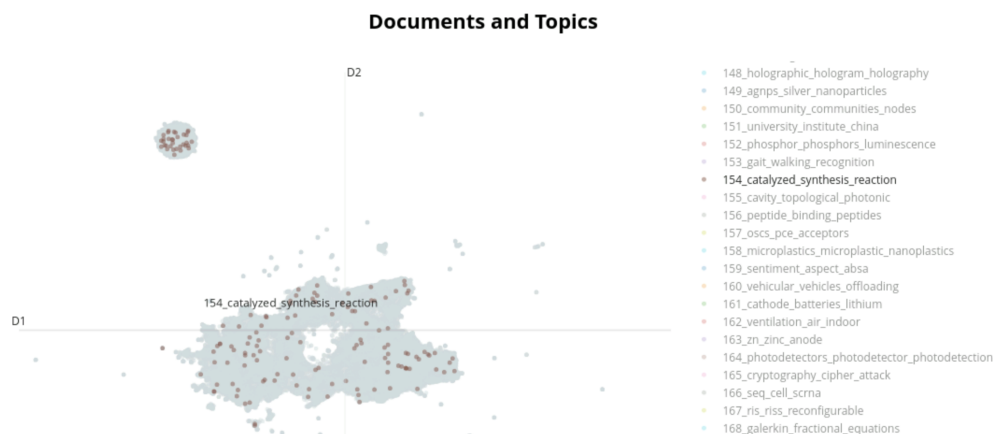


Figure 20: Sixth model run after applying UMAP

Figures 19 and 20 show the projections of documents for the initial topic model run and the final topic model run. To demonstrate the utility of this approach for identifying changes in topics we explore the cluster of topic 0 in the first model. This is highlighted in orange in both plots. In the initial run we find that this topic has substantial prominence in the model, however when looking at the final run we find that this topic has diverged into other clusters, and these papers are now diffused across the search

space. This indicates that these papers have been given less prominence by the dimensionality reduction model and the unique information in these papers is contributing less to the formation of the two primary components of the space. These figures are taken as screenshots of the model within interactive mode, and far greater clarity can be achieved using this method in real-time compared to statically.

The isolated clusters of documents on the left side of both figures represent missing values in abstract data which led to the generation non-meaningful clusters in both runs. However, again this is a useful comparison of how these null values can be compared across runs.

#### **5.4.2 BERTopic model wise similarity across all runs**

The BERTopic model objects also contains several outputs which can be used to build a model wise comparison across several model runs which can be used to investigate how any selected topics are changing over time.

This is based on the use of cosine similarity of the topic embeddings, with the assumption that topics which appear in multiple runs with a distance measure between each embedding will be similar; representing either the same topic appearing in the new run, or an evolution or merger of topics.

By constructing a matrix comparing all topics across all runs and selecting a suitable threshold value for similarity, it is possible to visualise the topics extracted from BERTopic over 6 months and then link them together in the form of a Sankey diagram to see a temporal pattern of emerging, merging, diverging or fading research topics.

A sample of this diagram is shown in Figure 20. The full interactive sankey is available at <https://github.com/tabahi/EmergingResearchTopics2022/tree/main>.

When interpreting this visualisation there are a number of items to note:

- The onward links coming out of the nodes are the same color as the topic nodes. The last run 2023-01 is aqua color has no onward links that's why only the nodes are shown.

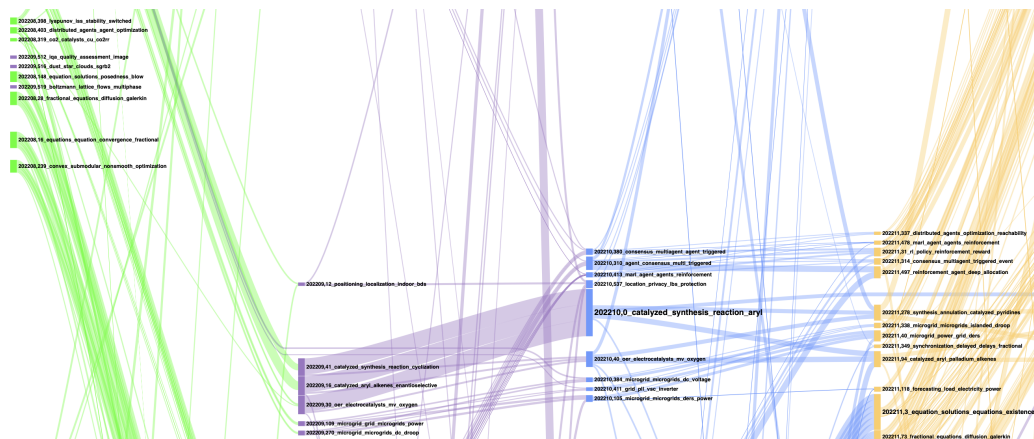


Figure 21: Sankey diagram showing connections between topics and topic changes over time

- The width of a link line represents the weight of the next run's topic (whereas color is the same as the previous node).
- The location of topic nodes along the horizontal axis does not represent time period. If a topic is present in all the runs, then it will start from 2022-08 on the left most side and go right till 2023-01
- If a topic emerges in a later run, let's say in 2023-09 then the left most node for that topic will be purple representing a start in 2023-09.
- Similarly, a topic having a link for only two or three runs/colours means its corresponding topic was absent in the other runs.
- At the bottom of the diagram, there are some standalone topics without any links. They represent outliers. This approach cannot handle topics that are absent in runs in-between (i.e., it doesn't link 2022-08 to 2022-10, if it's missing in 09).
- Some of the outlier topics for the last run (aqua color) could be newly emerging topics. For example: 202301 '405 mirna microrna mir 21'.
- When a topic's width increases compared to its tributaries, that shows an increasing weight of that topic.

## 6 Approach 3: Topic Labelling

### 6.1 Goal

Topic labelling is the task of assigning meaningful labels to a topic (or cluster of documents), once it has been identified, in such a way that captures the meaning and essence of this topic, and makes it understandable by the end user.

Currently, the BERTopic model used for this project uses a c-TF-IDF model to extract the most unique and pertinent keywords from the abstracts of all papers within a topic classified by the model and ranks these, providing the top 4 as the label for the topic. This is an elegant approach as it combines the efficiency and power of the embedding model to generate the topic, but the simplicity of the c-TF-IDF approach to make the results easier to interpret. Notably, this method used by BERTopic is an unsupervised technique, and the interpretation of the topics it produces can often be subjective. In addition, it might not always be possible to assign a clear, concise label to every topic.

Currently, the labelling of each topic is using the “top  $n$  words” method, which is the top 4 words for each topic. As shown in Table 3, these words can give a sense of the topic, but not all auto-generated labels are human-readable enough for complete understanding. In addition, words are often repeated both within and across topics, which will create barriers for analysts or policymakers when they want to use the results to aid their follow-up evidence-based decision-making.

Improving labelling and developing a clear understanding of the topics is also crucial when studying topic evolution. As the set of documents changes over time, the topics can evolve, and new topics can emerge while others fade away. Tracking this evolution requires being able to consistently label and interpret these topics, thus the goal of the data modelling team is to optimise the method of re-labelling topics to increase human-readability.

Table 2: Example topic labels using in-built tf-idf approach

Topic_number	Topic_label	Topic_keywords	Weight
16	16-equations-equation-convergence-fractional	equations, equation, convergence, fractional, numerical, galerkin, stochastic, discretization, differential, differential	746
99	99-galaxies-galaxy-jwst-star	galaxies, galaxy, jwst, star, redshift, stellar, mass, formation, luminosity, redshifts	318
445	445-cov-sars-omicron-ba	cov, sars, omicron, ba, bnt162b2, neutralizing, antibodies, mrna, variants, variant	137
459	459-vm-vms-virtual-migration	vm, vms, virtual, migration, cloud, consolidation, vmp, machines, virtualization, resource	21

## 6.2 Approaches

### 6.2.1 Text Summarization

To explore how we may address this challenge, we used a sample of data to test multiple methods. After running the BERTopic model once in August 2022, we obtained an output of 517 topics from all papers in one run of the search, and we will refer to it as Round 202208.

We first started from sampling a few topics by each topic’s weight. The weight of each topic is calculated by the number of papers meeting the following criteria:

- The value for  $maintopic = 1$
- $prob \geq 0.3$

The samples we selected were the least weighted 3 topics (Topic 256, Topic 294 and Topic 457) and the most weighted 2 topics (Topic 505 and Topic 490). These topics along with the BERTopic generated topic labels

are given in Table 3.

*Table 3: Least weighted and top-weighted topics in Round 202208*

Topic.number	Topic.label	Topic.keywords	Weight
505	505-libs-breakdown-laser-induced	libs, breakdown, laser, induced, spectroscopy, spectra, quantitative, mie, lif, elemental	2251
490	490-diffusion-denoising-generative-ddm	diffusion, denoising, generative, ddm, ddpm, score, models, sampling, prodiff, fid	1492
294	294-nanofluid-heat-nanofluids-flow	nanofluid, heat, nanofluids, flow, fluid, convection, nusselt, hybrid, casson, darcy	9
457	457-microalgae-biomass-microalgal-cultivation	microalgae, biomass, microalgal, cultivation, chlorella, production, productivity, mixotrophic, photobioreactors, biodiesel	9
256	256-seizure-epilepsy-eeg-epileptic	seizure, epilepsy, eeg, epileptic, seizures, ieeg, patients, electroencephalogram, patient, scalp	8

We approached this challenge by summarising the abstracts of all papers in one topic, and then sought to try to find the linkage between the summarization and the BERTopic computed keywords. There are two methods of performing text summarization; ‘extractive’ and ‘abstractive’ and we tested both of these methods here.

**Extractive text summarization** is a straightforward method where key information from a large text corpus is identified, selected, and compiled into concise summaries. All of the text in these summaries will have been present in the original source material. It functions based on predefined parameters, such as the text to be summarised, significant sentences (Top K), and the relevance of these sentences to the main topic. However, the process is bound by these parameters and may lead to a bias in the

extracted text under certain circumstances.

**Abstractive text summarization** has a more complex approach but offers considerable advantages. It comprehensively analyses the text and produces coherent sentences that encapsulate the original text's main ideas. This exact text will not have appeared in the original source material, and is instead 'written' specifically for this task. This method leverages large pre-trained language models to process and refine text using natural language processing. This ability to understand and rewrite text akin to human cognition sets abstractive summarization apart. Typically, abstractive summarization requires more compute resources and developer time to utilise correctly compared to extractive methods, but it is often more valuable in generating usable summaries.

### 6.2.2 Extractive Text Summarization

After some quick experiments, we realised quickly that in this project, since we aim to run the model on a frequent basis, and prioritise identifying the evolution of topics, extractive text summarisation may be preferable to abstractive summarisation. This is due to the compute resources required for fine-tuning and then deploying abstractive models.

The initial step required is to generate topic labels from "representative documents" within that topic from the corpus. In the BERTopic model, topics are represented by a distribution over the documents in the dataset. This means that for each topic, some documents are more representative (or more highly ranked) than others. We tried two slightly different approaches to identify the representative documents:

- Obtaining BERTopic model auto-generated top 3 papers under each topic (calculated by embedding metrics)
- In the case of BERTopic specifically, you can get the most representative documents for each topic with the *'get representative docs method'* which takes a topic id and returns the documents which are most representative of that topic.

Using these approaches we are able to obtain the top 5, 25 or 50 papers with the highest probability of being under a given topic. We found that when setting the threshold as 25 papers of each topic, the performance



represents a reasonable trade-off. At this threshold the sentences returned can be informative enough to adjust the original label compared to only using 3 or 5 papers, and the computation time is comparatively short compared to 50 papers. As such, we used this threshold for all experiments.

We first manually compared these representative documents obtained from these two separate approaches, and found the top papers identified by different approaches are not the same, but quite similar in terms of the themes or content. Upon retrieving the abstracts from the most representative papers, we then ran a few extractive summarization models to compare which might be a better way to summarise the topic.

We then explored two approaches for extractive text summarisation; manual improvements on the c-TF-IDF process, and use of the Python package Sumy.

### 6.2.3 Abstract to word summarization using c-TF-IDF model

The BERTopic Model uses an embedded c-TF-IDF (Text Frequency - Inverse Density Frequency) in it's algorithm to generate keywords from the papers.

c-TF-IDF is an intuitive method to capture word relevance within and across papers. The algorithm works by calculating the number of times a word appears in a paper and multiplies that value by the inverse of the number of papers which contain that word using the formula below:

$$TF = \frac{\text{number of times the term appears in the document}}{\text{number of total words in the document}}$$
$$IDF = \log\left(\frac{\text{number of documents in the corpus}}{\text{number of documents in the corpus containing the term}}\right)$$

The product of both numbers is taken to get the metric that represents the presence of a word within and across a corpus of documents.

$$TF - IDF = TF * IDF$$

BERTopic outputs key words through this algorithm and then a topic label is then generated for each cluster of papers by concatenating the top n keywords from the output.

We re-ran this c-TF-IDF model on the most representative papers within a topic cluster and used the title and abstract of the paper as the input data. Rather than run this on unigrams (single tokens / lexical items) alone, we ran the c-TF-IDF model on bigrams and trigrams (pairs or triples of tokens / lexical items) in order to isolate word sequences that were common within and across the papers as we believed that this might generate more useful features than using unigrams in isolation.

	n_gram	score
0	machine learning	1.901760
1	of quantum	1.601373
2	of the	1.571824
3	quantum machine	1.312840
4	quantum circuits	1.202454

	n_gram	score
0	machine learning	1.901760
1	of quantum	1.601373
3	quantum machine	1.312840
5	quantum computing	1.091229
9	variational quantum	0.877679

Figure 22: Bigrams including stop wordsFigure 23: Bigrams excluding stop words

	n_gram	score
0	quantum machine learning	1.269714
1	quantum approximate optimization	0.497556
2	machine learning algorithms	0.496594
3	the number of	0.462846
4	this paper we	0.457638

	n_gram	score
0	quantum machine learning	1.269714
1	quantum approximate optimization	0.497556
2	machine learning algorithms	0.496594
6	noisy intermediatescale quantum	0.452414
7	approximate optimization algorithm	0.437009

Figure 24: Trigrams including stop wordsFigure 25: Trigrams excluding stop words

The title and abstract of the papers were joined, cleaned to remove stop words (very common 'filler' words used in language which typically carry little semantic meaning), special characters and numbers and then run through the c-TF-IDF model. These results are shown in Figures 22-24 along with the c-TF-IDF scores for the topic labels. Higher scores indicated that the label is more unique and relevant to the topic, however this does not always align with the human interpretation.

The results indicated that both bigrams and trigrams were more informative than the unigram output from BERTopic, however, in comparison, the BERTopic's output sometimes looked better or than the output from the separate c-TF-IDF model. This could be as a result of BERTopic using abstracts from all the papers whereas the separate c-TF-IDF model uses abstracts from only 50 papers.

To improve the results from the model, another approach was taken to include the stopwords in the input data, obtain the bigrams and trigrams and then exclude any bigram or trigram result that contained a stop word. This was done to ensure we obtained word sequences that made sense on their own without a stop/filler word. The results from this were better and some examples given in Figures 24 and 25. Overall we found that using Trigrams and excluding stop words led to the most meaningful topic labels.

#### **6.2.4 Abstract to sentence summarisation using SUMY package**

Sumy[1] is a Python library designed specifically for extractive text summarisation. It provides a simple and straightforward interface to extract summaries from textual content, making it useful for various natural language processing tasks. Extraction-based summarization focuses on selecting and assembling important sentences or phrases directly from the source text to create a summary. It involves identifying key sentences based on criteria like relevance, informativeness, and importance. The extracted sentences are then presented in the same order as they appear in the original text.

In brief, applying Sumy includes the following steps:

- Text Parsing - Sumy supports parsing text from various sources, including HTML documents, plain text files, and strings. It provides parser classes like HtmlParser and PlaintextParser to handle different types of input. These parsers tokenize the text, extract relevant information, and prepare it for further processing.
- Sentence Tokenization - Sumy includes tokenizers for splitting text into individual sentences. It ensures that the summarization

algorithms operate at the sentence level, enabling them to identify important information and generate coherent summaries.

- Stemming and removing stopwords - Stemming is the process of reducing words to their base or root form, and stop words are common words that often carry little meaning. Sumy includes stemmers and stop word lists for multiple languages, allowing you to preprocess text and improve the accuracy of summarization.
- Language Choose - Sumy provides support for multiple languages, allowing you to summarize text in different languages. It includes built-in language-specific functionalities, such as stemmers and stop word lists, to handle linguistic nuances and improve the quality of summaries.
- Summarization Algorithms - Sumy implements several popular algorithms for text summarization, including LSA (Latent Semantic Analysis), LexRank, Luhn, TextRank, and more. These algorithms employ different techniques, such as graph-based ranking, statistical analysis, and semantic analysis, to identify important sentences and generate summaries.

In this task we used Sumy to select the top 3 sentences to build a summarisation for a topic from the most relevant 25 papers identified within that topic. This approach was relatively successful, and some initial experiments showed that it required 45-60 seconds for native speakers to read and understand the basic meaning of the 3 summarized sentences.

Table 4 provides some examples of the Sumy results for selected topics.

In the first example, the keywords only describe the basic field of the research topic, but do not indicate the research hotspot of the research topic. For the second one, there is obvious duplication of keywords, and only two key messages could be expressed by four keywords. Finally, the information that keywords in the third topic provide may be misleading due to polyseme or synonym. Therefore, in these situations, re-labelling for topics by summarization based on the abstracts is promising and sometimes necessary. In comparison, the extractive summaries are far-richer in content and material, however these are typically harder to

BERTopic Label	Sumy Summary
leaf crop disease plant	In this study, we propose a deep metric learning based method to extract latent space representations from plant diseases with just few images by means of a Siamese network and triplet loss function. An automatic flower identification system over categories is still challenging due to similarities among classes and intraclass variation, so the deep learning model requires more precisely labelled and high-quality data. The time-series of the Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and NaE (combined NDVI and EVI) were adopted as input features, and four widely used machine learning models, including Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbour (KNN), and their integration (Stacking) were compared to examine the performance of multiple crop types (i.e., wheat, corn, wheat-corn, early rice, and early rice-late rice) classification in the North China Plain.
queueing queues queue server	The Biochemical Society, with its mission of supporting the advancement of science, with addressing grand challenges that have societal impact, has included OoC into their agenda to review the current state of the art, bottlenecks, and future directions. This conference brought together representatives of the main stakeholders in the OoC field including academics, end-users, regulators and technology developers to discuss and identify requirements for this new technology to deliver on par with the expectations and the key challenges and gaps that still need to be addressed to achieve robust human- relevant tools, able to positively impact decision making in the pharmaceutical industry and reduce overreliance on poorly predictive animal models. In this paper, we are devoted to developing matrix-analytic methods for solving Poisson's equation for irreducible and positive recurrent discrete-time Markov chains (DTMCs). In this system, we jointly optimize the phase-shift coefficient and the transmit power in sequential time slots to maximize the long-term energy consumption for all mobile devices while ensuring queue stability.
plasma tokamak divertor triangularity	We outlined the important conclusions obtained from recent literature and listed the evaluation methods, characterization techniques, and contrastable experimental data of these types of inhibitors when used for carbon steel corrosion in 1.0 M HCl solution. According with Fukui reactivity indices, the molecules adsorbed on the metal surface provide a protective cover against nucleophilic and electrophilic attacks, pointing to the corrosion inhibition properties of 4-PC. Remarkably high mixed proton and electron conduction arising from oxidized single-wall carbon nanotubes at room temperature is demonstrated. All the mentioned aspects are the topic of this current review, which is meant as constructive criticism to spotlight the use of biomass waste as efficient green inhibitors in order to re-evaluate their viability and debate prospective research in the field, which is still lacking justification.

Figure 26: Table 4: Sumy results compared to BERTopic labels

interpret and contain additional superfluous information.

For both summarisation approaches, we have identified that it is helpful to introduce experts into the next stage so they can manually relabel the original labels. It involves manually reviewing these extracted sentences to understand the general theme or content, and compare with the original

label. This process may be time-consuming at the first stage, especially with a large number of topics and/or lengthy documents. However, this can first guarantee more interpretable labels compared to the current “top N words”.

#### **6.2.5 Abstractive text summarization**

The team also explored the use of abstractive text summarisation models such as BART, PEGASUS, and T5. These models are build on top of large pre-trained language models and then fine-tuned in a supervised manner over vast quantities of text documents to create a shorter version of the original document that captures all the important information in a much more succinct way.

Typically, these models have been fine-tuned on news data or scientific documents, with the task of generating a summary of the article or a summary of the scientific article, akin to an abstract.

We performed some initial testing of these approaches using the most highly ranked documents for the topics, however our initial tests demonstrated some issues with using these models in their base format. Firstly, the abstractive summaries were typically far too long (paragraph length) to be useful as topic labels, and they contained too much ancillary information. The summaries were also not always accurate enough, sometimes generating nonsensical or irrelevant information and the models were also too slow to run at scale given the time allocated for this DSG.

As such we concluded that in order to use these models it will be necessary to further fine-tune them on in-domain data and with gold-standard examples of topic labels for the specific task at hand.

#### **6.2.6 Future work and research avenues for topic labelling**

In completing this work we have identified a number of areas that would benefit from future work in order to progress the capability to generate accurate, human readable, topic labels.

Firstly, it is worth considering applying a machine learning model trained to generate labels based on the words in the topics. This could be either

extractive or abstractive in nature, but a supervised approach would be far easier to fine-tune and measure the performance of compared to an unsupervised approach.

It is also worth considering, if the data source allows, pulling the authors' self-defined keywords of each paper, instead of the abstract, to extract a new set of keywords (top 3 or top 5), and use these keywords to name each topic cluster. These could then be used either as the gold-standard labels to train a supervised approach, or as baseline measures to compare an unsupervised approach to.

Additionally, and perhaps unfortunately, BERTopic uses HDBSCAN for clustering the data and as such is not possible for a user to specify the number of clusters they would like the model generate. To a certain extent, this is an advantage, as users can trust HDBSCAN to be better in finding the number of clusters than humans are. However, from exploring the model outputs it appears that we could reasonably reduce the topic clusters. If the user were to set a more realistic and accurate number of clusters/topics it would be possible to introduce experts to manually re-label each cluster/topic.

Neural Embedding Topic Labelling (NETL) is a method used to label topics with neural embedding and may be worth exploring in future work. The labels of hierarchical topics should have hierarchical relationships with other labels. One study (Kozono & Saga 2020) proposed a method for labelling hierarchical topics with hierarchical relationships, and uses NETL to generate candidate labels for bottom topics. This proposed method calculates how small the overlap of the candidate labels compared with other sibling topics and adds the label of the bottom topics and generates labels in the same way as the bottom topics recursively. This may produce more meaningful and robust topic labels compared to the BERTopic approach.

It may also be worth considering using large language models like GPT-4 to infer from the BERTopic generated keywords as well as the abstracts to infer the topic, then use this to adjust the labels accordingly.

## 7 Conclusion

Overall, this report covers an intensive week long deep dive into the use of topic modelling for automated horizon scanning and mapping of the scientific landscape. We present some initial findings that demonstrate how the outputs from a topic model run at regular intervals can be used to visualise the distribution of topics across a series of scientific documents, and through the use of novel statistics, map the changing dynamics of these scientific topics over time.

By breaking this challenge down into three sub-sections, the team was able to draw a series of primary conclusions. Firstly, the team focused on topic dynamics demonstrated that the relationships between topics can be identified using a combination of similarity metrics (cosine similarity and Euclidean distance) and that these measures can be used to determine within-run and across-run topic linkages respectively. In addition, the team demonstrates that taking a more macro look at the changing distribution of topics across multiple runs at a larger scale, and the relationship between information contained with these topics, can give insight into the scientific field as a whole. Novel metrics such as Disruptive Index and Average Weighted Distance are presented, and these show promise for additional exploration in future work.

By generating both data-centric and model-centric data visualisations the visualisation team clearly demonstrated the utility of high-quality data visualisation in mapping the scientific landscape and how topic models are capturing the changing trends and topic dynamics. This team identified that there is a clear trade-off between the richness of information presented in the visualisations and the interpretability of these plots. Dimensionality reduction techniques are beneficial in this balance, but analysts and end-users will need to be trained on how best to interpret results when they are projected from high-dimensional space into two dimensions.

The topic labelling team identified that both abstractive and extractive summarisation techniques show promise for the development of human-readable topic labels, however without high-quality human generated ground-truth topic data it is difficult to assess how good these approaches are, or what the limitations are.



A key theme of these conclusions is that high-quality ground-truth data is a key requirement for future research in this area. The manual identification of topics within a dataset, and the scientific papers associated with these topics, would allow for a quantitative assessment of how well the base topic models are modelling the scientific landscape. In addition, these topic labels would allow for both the assessment of summarisation techniques and the training of custom or fine-tuned supervised models for this task. Finally, manual identification of key topic dynamics, and the provision of goal-standard examples of key trends such as topic emergence, convergence, or disappearance would allow for the proposed topic dynamic metrics to be assessed. If this ground truth data is generated it will also open up additional research avenues. For example, historical data could be used to train models to forecast the dynamics of new, disappearing, or splitting/merging topics, or cutting edge NLP approaches which can incorporate dynamic or temporal information directly could be explored.

This study sought to explore topic modelling's potential to unravel the scientific landscape's emerging trends. By building upon Dstl's prior work and leveraging BERTopic, we sought to develop data science approaches to automate horizon scanning. Results indicated that this is an area with significant potential, and if successfully developed could empower policymakers with timely insights into rapidly evolving scientific developments.

## 8 Team Members

**Abdulrahman A. A. Alsayed** is a PhD Researcher in Language and Linguistics in the School of Modern Languages at the University of St Andrews and an Enrichment Scheme Doctoral Researcher at The Alan Turing Institute. He contributed to this project by working on the human-readability of topic re-labelling and on the documentation for the re-labelling report.

**Abdullah Hussein** is a Researcher at a government entity. He obtained his Master's in Data Analytics from Rochester Institute of Technology. He contributed in this project exploratory data analysis, Data visualisation, data duality, and BERTopic model visualisations.

**Dr John Gallacher** is a Research Scientist working on applications of machine learning for Defence and Security contexts. He holds a DPhil from the University of Oxford, and was the Principle Investigator on this project.

**Jai Geelal** is a PhD Researcher in the Department of Accounting & Finance at the University of Strathclyde looking at applications of Topological Data Analysis within Finance. He contributed to this project by applying the TDA-Mapper algorithm, developing BERT model visualizations, and report writing.

**Knectt Paulschoh Lendoye L'eyebe** is a PhD candidate in the School of Computing, Newcastle University, analyzing biomarkers responsible for diseases affecting vision and the neural system. His contribution to this DSG was the implementation of BERTopic's model visualizations.

**Chinonye Dianne Pat-Ekeji** is a Product Data Analyst at Flo Health a women's health company. She obtained her Master's in Data Science from The London School of Economics and Political Science. She contributed to this project by exploring topic re-labelling through text summarization using the TF-IDF algorithm with bigrams and trigrams and largely contributed to the structuring, writing and editing of the project report.

**Abdul Rehman** is a PhD Researcher at the National Center for Computer Animation of Bournemouth University. He contributed by linking the topics across the runs and created the sankey diagram.

**Sidharth Rony** is a PhD Researcher in the Department of Economics at the Royal Holloway University of London. Sidharth contributed to this project by implementing UMAP for dimensionality reduction and producing interactive plots that showcase the topic embeddings over time.

**Dr Chakresh Kr. Singh** is a Post-doctoral researcher at the University of Paris. Chakresh works on large scale publication data-sets to identify and quantify universal patterns in the evolution of scientific fields. He contributed to this project by creating networks for analysis, defining and calculating the measures to identify topics of interest and visualization of networks.

**Rubin Wang** is a PhD Researcher at the National Center for Computer

Animation of Bournemouth University. He made contributions to this project by utilizing statistical methods and visualizing the original raw data, as well as visualizing the BERTopic output data in the embedding space. Additionally, he also played a role in finalizing the report's data summary section.

**Kexin Yin** is a PhD Researcher in the Future Metrology Hub at the University of Huddersfield. Kexin focuses on the improvement of metrology with help of AI-based Digital Twin. He contributed to this project by applying Sumy with LSA to summarize the abstracts for topic re-labelling.

**Yurong Yu** is a PhD Researcher in the Centre for Environmental Policy at Imperial College London. Yurong works on biodiversity modelling for her PhD. She contributed to this project by working on re-labelling of topics, report writing of the re-labelling part, and serving as a facilitator for the team.

**Liang Zhou** is a PhD student at the Gatsby Computational Neuroscience Unit. He contributed to this project by investigating metrics of convergence and divergence of topics and serving as a facilitator for the team.

## References

- [1] Miso Belica. *Sumy*. Version 0.9.0. 2021. URL: <https://github.com/miso-belica/sumy>.
- [2] Luís M. A. Bettencourt et al. "Population modeling of the emergence and development of scientific fields". en. In: *Scientometrics* 75.3 (June 2008), pp. 495–518. ISSN: 1588-2861. DOI: [10 . 1007 / s11192 - 007 - 1888 - 4](https://doi.org/10.1007/s11192-007-1888-4). URL: <https://doi.org/10.1007/s11192-007-1888-4> (visited on 06/19/2023).
- [3] David M Blei. "Latent Dirichlet Allocation". en. In: (2003).
- [4] Lutz Bornmann et al. "Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers". In: *Quantitative Science Studies* 1.3 (2020), pp. 1242–1259.

- [5] Ciarán Byrne et al. “Topic Modeling With Topological Data Analysis”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022, pp. 11514–11533.
- [6] Dhivya Chandrasekaran and Vijay Mago. “Evolution of Semantic Similarity—A Survey”. In: *ACM Computing Surveys (CSUR)* 54 (2021), pp. 1–37.
- [7] Johan S. G. Chu and James A. Evans. “Slowed canonical progress in large fields of science”. en. In: *Proceedings of the National Academy of Sciences* 118.41 (Oct. 2021), e2021636118. ISSN: 0027-8424, 1091-6490. DOI: [10 . 1073 / pnas . 2021636118](https://doi.org/10.1073/pnas.2021636118). URL: <https://pnas.org/doi/full/10.1073/pnas.2021636118> (visited on 06/19/2023).
- [8] Colin B. Clement et al. *On the Use of ArXiv as a Dataset*. en. arXiv:1905.00075 [physics]. Apr. 2019. URL: <http://arxiv.org/abs/1905.00075> (visited on 06/19/2023).
- [9] Laércio Dias et al. “Using text analysis to quantify the similarity and evolution of scientific disciplines”. In: *Royal Society Open Science* 5.1 (Jan. 2018). Publisher: Royal Society, p. 171545. DOI: [10.1098/rsos.171545](https://doi.org/10.1098/rsos.171545). URL: <https://royalsocietypublishing.org/doi/10.1098/rsos.171545> (visited on 06/19/2023).
- [10] Nafeesa Esmail et al. “Emerging illegal wildlife trade issues: A global horizon scan”. en. In: *Conservation Letters* 13.4 (2020). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/conl.12715>, e12715. ISSN: 1755-263X. DOI: [10 . 1111 / conl . 12715](https://doi.org/10.1111/conl.12715). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/conl.12715> (visited on 06/22/2023).
- [11] Rita González-Márquez et al. *The landscape of biomedical research*. en. preprint. Scientific Communication and Education, Apr. 2023. DOI: [10 . 1101 / 2023 . 04 . 10 . 536208](https://doi.org/10.1101/2023.04.10.536208). URL: <http://biorxiv.org/lookup/doi/10.1101/2023.04.10.536208> (visited on 06/22/2023).
- [12] Maarten Grootendorst. “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”. In: *arXiv preprint arXiv:2203.05794* (2022).

- [13] Maarten Grootendorst. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. en. arXiv:2203.05794 [cs]. Mar. 2022. URL: <http://arxiv.org/abs/2203.05794> (visited on 06/19/2023).
- [14] David Jurgens et al. “Measuring the Evolution of a Scientific Field through Citation Frames”. en. In: *Transactions of the Association for Computational Linguistics* 6 (Dec. 2018), pp. 391–406. ISSN: 2307-387X. DOI: [10.1162/tacl\\_a\\_00028](https://doi.org/10.1162/tacl_a_00028). URL: <https://direct.mit.edu/tacl/article/43437> (visited on 06/19/2023).
- [15] Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 2012.
- [16] Thomas S. Kuhn. *The structure of scientific revolutions*. en. [2d ed., enl. International encyclopedia of unified science. Foundations of the unity of science, v. 2, no. 2. Chicago: University of Chicago Press, 1970. ISBN: 978-0-226-45803-8.
- [17] Li Li et al. “Identification of type 2 diabetes subgroups through topological analysis of patient similarity”. In: *Science translational medicine* 7.311 (2015), 311ra174–311ra174.
- [18] Leland McInnes, John Healy, and Steve Astels. “hdbscan: Hierarchical density based clustering”. en. In: *The Journal of Open Source Software* 2.11 (Mar. 2017), p. 205. ISSN: 2475-9066. DOI: [10.21105/joss.00205](https://doi.org/10.21105/joss.00205). URL: <http://joss.theoj.org/papers/10.21105/joss.00205> (visited on 08/17/2023).
- [19] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. en. arXiv:1802.03426 [cs, stat]. Sept. 2020. URL: <http://arxiv.org/abs/1802.03426> (visited on 08/17/2023).
- [20] Leland McInnes et al. “UMAP: Uniform Manifold Approximation and Projection”. In: *The Journal of Open Source Software* 3.29 (2018), p. 861.
- [21] Dewey Murdick. *Map of Science*. 2023. URL: <https://sciencemap.eto.tech/?mode=map> (visited on 06/19/2023).

- [22] Raj K Pan et al. “The memory of science: Inflation, myopia, and the knowledge network”. In: *Journal of Informetrics* 12.3 (2018), pp. 656–678.
- [23] Radim Rehurek and Petr Sojka. “Gensim–python framework for vector space modelling”. In: *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3.2 (2011).
- [24] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. en. arXiv:1908.10084 [cs]. Aug. 2019. URL: <http://arxiv.org/abs/1908.10084> (visited on 08/17/2023).
- [25] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: <https://arxiv.org/abs/1908.10084>.
- [26] Julia Sigle and David Robinson. *Welcome to Text Mining with R — Text Mining with R*. 2022. URL: <https://www.tidytextmining.com/> (visited on 06/19/2023).
- [27] Chakresh Kumar Singh et al. “Quantifying the rise and fall of scientific fields”. In: *Plos one* 17.6 (2022), e0270131.
- [28] Gurjeet Singh, Facundo Mémoli, Gunnar E Carlsson, et al. “Topological methods for the analysis of high dimensional data sets and 3d object recognition.” In: *PBG@ Eurographics* 2 (2007), pp. 091–100.
- [29] Hendrik Jacob Van Veen et al. “Kepler Mapper: A flexible Python implementation of the Mapper algorithm.” In: *Journal of Open Source Software* 4.42 (2019), p. 1315.
- [30] Lingfei Wu, Dashun Wang, and James A Evans. “Large teams develop and small teams disrupt science and technology”. In: *Nature* 566.7744 (2019), pp. 378–382.
- [31] Qiang Wu and Zhaoyang Yan. “Solo citations, duet citations, and prelude citations: New measures of the disruption of academic papers”. en. In: (2019).



**The  
Alan Turing  
Institute**

---

**turing.ac.uk  
@turinginst**