Analysing Longitudinal Social Science Questionnaires: Topic modelling with BERT-based Embeddings

Vida Sharifian-Attar
Dept. of Computer Science
University of Surrey
Guildford, UK
vs00382@surrey.ac.uk

Suparna De
Dept. of Computer Science
University of Surrey
Guildford, U.K.
s.de@surrey.ac.uk

Sanaz Jabbari
Centre for Advanced Research Computing
University College London
London, U.K.
s.jabbari@ucl.ac.uk

Jenny Li
UCL Social Research Institute
University College London
London, U.K.
jenny.li@ucl.ac.uk

Harry Moss
Centre for Advanced Research Computing
University College London
London, U.K.
h.moss@ucl.ac.uk

Jon Johnson
UCL Social Research Institute
University College London
London, U.K.
jon.johnson@ucl.ac.uk

***Abstract*—Unsupervised topic modelling is a useful unbiased mechanism for topic labelling of complex longitudinal questionnaires covering multiple domains such as social science and medicine. Manual tagging of such complex datasets increases the propensity of incorrect or inconsistent labels and is a barrier to scaling the processing of longitudinal questionnaires for provision of question banks for data collection agencies. Towards this effort, we propose a tailored BERTopic framework that takes advantage of its novel sentence embedding for creating interpretable topics, and extend it with an enhanced visualisation for comparing the topic model labels with the tags manually assigned to the question literals. The resulting topic clusters uncover instances of mislabelled question tags, while also enabling showcasing the semantic shifts and evolution of the topics across the time span of the longitudinal questionnaires. The tailored BERTopic framework outperforms existing topic modelling baselines for the quantitative evaluation metrics of topic coherence and diversity, while also being 18 times faster than the next best-performing baseline.**

***Index Terms*—BERTopic, topic models, longitudinal topic modelling, word embedding**

I. Introduction

The United Kingdom (UK) has a rich history of longitudinal studies, which administer surveys (questionnaires) to participants over different stages of their lifetime. The Cohort and Longitudinal Studies Enhancement Resources (CLOSER) project [1] has been working to digitalise and enhance the metadata of these longitudinal survey questionnaires with the aim to increase the discoverability and reuse of the vast amounts of social and biomedical survey questions the UK has collected over the past 75 years. CLOSER categorises every question into a two-level hierarchical system, which is derived from combining vocabularies from medical subject headings (MeSH) [2] and Humanities and Social Science Electronic Thesaurus (HASSET) topic vocabularies [3] into one overarching, controlled categorisation system (CLOSER ontology) and ensures both the biomedical and sociological aspects of the questionnaires are covered for future researchers. CLOSER's labelling structure has 16 general 'Level 1' labels (e.g. 'Demographics') to 120 specific 'Level 2' labels (e.g. Demographics – Place of Birth) [4].

The longitudinal nature of the questions gives the data complexity both in terms of the language used for different age groups as well as the possible shift in semantics that happen over 70 years. Moreover, the multidisciplinary nature and scale of the collected questions and the manual nature of the tagging of the question labels to concepts from the CLOSER ontology implies that there is bound to be outliers, causing the dataset to be mislabelled in some instances. In our work, a topic is defined as a set of questions that have a shared context. A topic is a subjective category: one question text can be categorised into many different coherent contexts. For example "Did he drink?" can be in a category about alcohol consumption, health behaviour or even home life in certain contexts. Therefore, topic modelling requires a deep understanding of the semantics of a question to be accurate. This also illustrates the subjectivity of the labelling process and therefore its propensity towards bias. Manual processes mean that scaling to provide a high quality question bank for

This research is funded by the Science and Technology Facilities Council (STFC) DiRAC-funded 'Understanding the multiple dimensions of prediction of concepts in social and biomedical science questionnaires' project, grant number ST/S003916/1 and through the UKRI Strategic Priority Fund as part of the wider Protecting Citizens Online programme (Grant number: EP/W032473/1) associated with the National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN).

978-1-6654-8045-1/22/$31.00 ©2022 IEEE

use of survey questions for reuse by studies and data collection agencies is limited.

To address these issues, we investigate unsupervised topic modelling methods to uncover the hidden semantic structures of questions for comparison to the original manually-created labels and a move towards a semi-automated question tagging approach. Thus, this work investigates instances of label bias (due to the preconceived domain notions and stereotypes by the human annotators of the question tags) [5] and selection bias, where some labels may be far overrepresented than other topics [5], addressing the final 'V' - veracity of the big data of the longitudinal survey questions.

Our work takes advantage of recent advances in neural networks for non-linear language modelling for contextualising the meaning of text corpora, such as the novel sentence embedding-based BERTopic architecture [6] and extends it with (1) enhanced visualisations for comparing the topic model labels with the original manually-assigned tags, as well as (2) dynamic topic modelling that tracks the evolution of the topics over the span of the longitudinal studies, uncovering changes in the lexicon of the questionnaires over the study years.

Thus, our work addresses the instances and impact of label and selection bias by modelling the relationship between the found topics and the existing manual labels via extensive visually-enhanced analysis of the longitudinal dataset and the experimental results and also addresses the challenges of the longitudinal nature of the dataset. The implementation is performed within the Optimizing and Comparing Topics Models is Simple! (OCTIS) framework [7] which enables fine-tuning BERTopic with other language models for improved performance as well as quantitative evaluation against other state-of-the-art topic modelling frameworks like Top2Vec [8] and Contextualized Topic Modelling (CTM) [9] on quantitative metrics.

The remainder of this paper is organised as follows: Section II discusses the related works in topic modelling and text embedding. Section III describes our dataset, followed by a description of the proposed framework's components in Section IV. Implementation details are presented in Section V. Section VI discusses the experimental results and main findings. Section VII concludes the paper with a summary of the achieved results and an outline of future work.

II. Related Work

A. Textual Embeddings

Embeddings are a way of representing units of text (words, sentences, documents), in a discrete vectorised fashion. This mapping between text to a multidimensional space, not only allows for conducting computation over text, but is also designed to function as a semantic representation of the textual elements. By this, two pieces of text that have the same semantic value, but have different textual representations, should ideally end-up in the same location in the newly defined multidimensional space after being mapped. As far as the document representation is concerned, the idea of vectorised representation dates back to more classical representations such as Term Frequency-Inverse Document Frequency (TF-IDF) [10] and count vectorisers. Along with recent advances and success of deep neural network based techniques in Natural Language Processing (NLP), word embedding paradigms such as Word2Vec [11], GloVe [12] and Embeddings from Language Model (Elmo) [13] are now widely used in almost every type of NLP task. Firstly, the words or textual units are encoded and mapped to the new space, and then used in the intended downstream NLP task such as translation, question answering, named entity recognition, classification etc.

Each of these neural net-based approaches that learn embeddings, differ in the amount of context they use to inform their training. While Word2Vec uses a two-layer shallow neural network-based approach that is optimised to predict a masked word given the words before and after, GloVe focuses on words co-occurrences over the whole corpus. Elmo is a more context-aware word embedding. Another state-of-the-art embedding technique uses Bidirectional Encoder Representations from Transformers (BERT) [14]. BERT, proposed by Google and trained on Wikipedia, is a bidirectional model that means it learns information from both sides of a token's context during the training phase.

Recent research works have extended these word embeddings to domain-specific areas, such as the GloVe and Word2Vec-based embeddings for scientific publications in [15]. An example of the state the art in sentence embeddings is DeClutr [16], which unlike most high-performing sentence embedding learning solutions, does not require labelled training data, instead utilising a 'self-supervised objective'. It trains an encoder for a minimisation problem on the semantic distances between nearby randomly collected textual segments from the same dataset.

B. Topic Modelling

Unsupervised topic modelling uses statistical methods to uncover hidden semantics and associations (i.e. topics) in otherwise unstructured sets of documents. Latent Dirichlet Allocation (LDA) [17]-based topic models assume that a text corpus has a fixed number of topics, evaluating the distribution of topics in a document in the corpus (which form a Dirichlet distribution), and the order of words or documents does not matter. LDA-based topic modelling has been successfully applied to short texts (e.g. Twitter messages) [18], [19]. LDA's assumption of all topics being independent limits its applicability to temporal interactions between texts [20].

Topic models do not capture the semantic relationship between words in a text corpus. This is where word embedding models can come into play as they are focused on capturing the semantic relationship in a lower-dimensional vector space. Topic models that incorporate text embeddings with a joint document approach, include Top2Vec and BERTopic, with similar architectures for creating embedding vectors from documents, reducing the embeddings dimensions and applying density-based clustering, but differing in their topic word selection methods. In this work, we apply and extend a

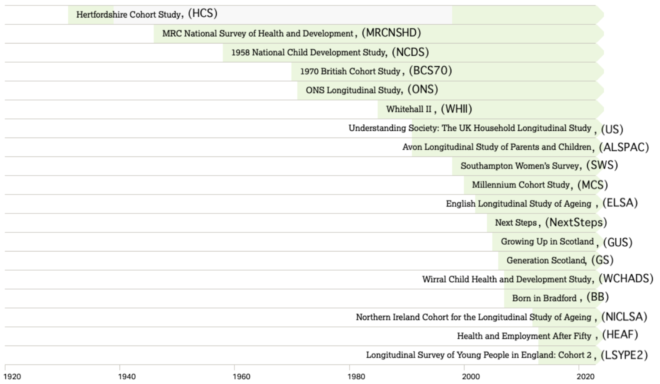


Fig. 1. Timeline of studies in the CLOSER dataset [24]

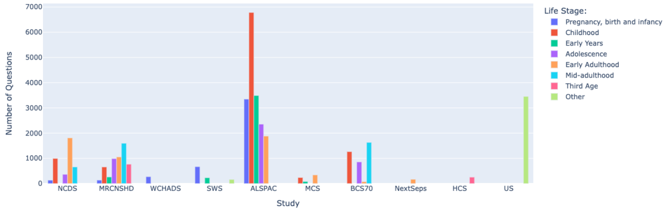


Fig. 2. Number of questions per study per life stage

BERTopic-based architecture for topic modelling applied to longitudinal questionnaire data.

Existing works that apply topic modelling approaches incorporating text embeddings include that by Raju *et al.* [20] who apply BERTopic for analysing consumer financial data with the FinBERT domain-specific pre-trained embedding and that by Pek *et al.* [21] who experiment with combinations of LDA with various embeddings such as Word2Vec and FastText [22] for identifying trends from scientific publications. BERTopic is also used in [23] for extracting intent from user messages from a chatbot application.

III. Dataset

The dataset used in this work has 318 questionnaires (i.e. instruments), with the dates of the instruments ranging from 1946 to May of 2020 [24], as shown in Figure 1, and are all part of ongoing studies. Each study may have a number of questionnaires, covering different life stages of the participants. The life stages covered within the studies include: (1) Pregnancy, birth and infancy, (2) childhood, (3) early years, (4) adolescence, (5) early adulthood, (6) mid-adulthood, (7) third age, and (8) other. As with the study years, the number of questions in each study are comparatively imbalanced. Next Steps has only 169 questions, whereas ALSPAC has 17,825. Some studies like ALSPAC or MRC National Survey of Health and Development (MRCNSHD) have a variety of life stages, whereas those like HCS, US or Next Steps focus on only one life stage in the dataset, even though some studies such as HCS have a "Birth Records" questionnaire which belongs to "Pregnancy, birth and infancy". Figure 2 depicts the different life stages covered in the studies forming part of this paper's dataset and a comparison of the number of questions per study.

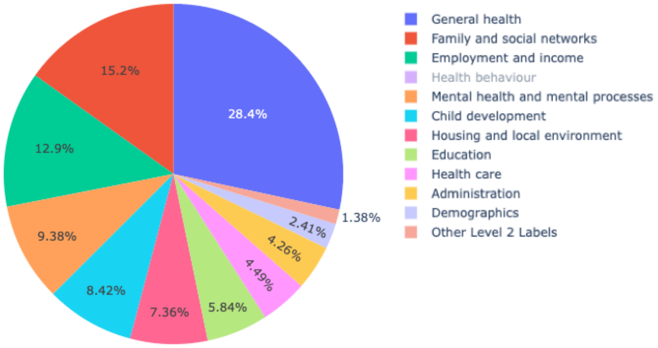


Fig. 3. Proportion of level-1 topics in the dataset

The dataset was extracted from CLOSER Discovery, using the Colectica Repository REST API [25], with a Python 3 script, as detailed in [26]. The dataset is a csv file with 42008 rows, with columns detailing several aspects of the questionnnaire data and metadata, of which the following are relevant to this work: questionURN, QuestionLiteral (question text), QuestionGroupLabels (Colectica code of the manually assigned broad 'Level 1' followed by the specific 'Level 2' labels). For dynamic topic modeling focussing on a particular study, further information on each questionnaire within each study's relevant life stage and the year was also extracted from the CLOSER Technical Wiki (https://wiki.ucl.ac.uk/display/CLOS/) and added to the dataset.

By deriving the level-1 topics from all level-2 labels, the proportion of each topic in the dataset is shown in Figure 3, which shows a large variation in the size of the topics, with some level-1 topics having far more questions than others. For instance, only two questions are in 'health insurance', only 5 in 'infant mortality'. However, there are many closely related QuestionGroupLabels, such as 'Health Care', with similar questions. Also, CLOSER has very specific level-2 groups, such as 'Proteomics'. An additional challenge is the many questions within the COVID-19 level-1 group that would otherwise belong to other, existing groups. For example, 'COVID-19 - Education' versus the more specific, non-COVID related education topic labels. However, due to our unsupervised methods, a group label vastly unrepresented due to human bias may be brought to light by our topic modelling.

IV. Proposed Framework

Our proposed methodology uses the BERTopic architecture to apply the concept of sentence embeddings which incorporates context-sensitive word embeddings and to also consider correlated topics.

BERTopic has three main algorithmic components, as shown in Figure 4: 1) Document Embedding: embedding documents into a vector representation of the data by using sentence embeddings. 2) Document Clustering: reducing embeddings

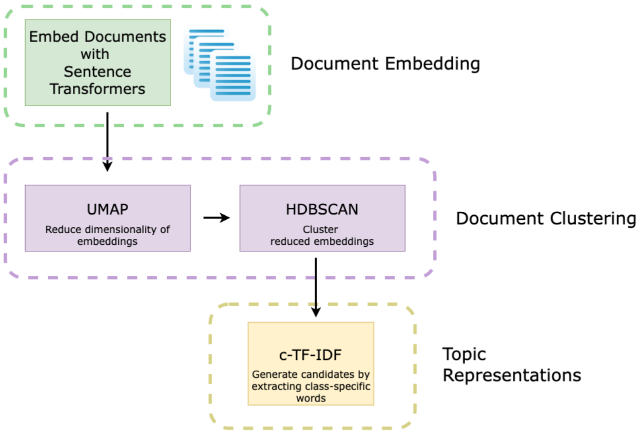

Embed Documents with Sentence Transformers
Document Embedding
UMAP
Reduce dimensionality of embeddings
HDBSCAN
Cluster reduced embeddings
Document Clustering
c-TF-IDF
Generate candidates by extracting class-specific words
Topic Representations

Fig. 4. BERTopic algorithm components (adapted from [6])

dimensions followed by clustering, and, 3) Topic Representation: identifying representative words of each topic cluster.

Sentence transformers are used to generate vectorised representations of the documents (BERTopic terminology for each sentence), i.e. embeddings. Sentence embeddings work by applying the concept of contextualised word embeddings across entire sentences rather than individual words. This first step in BERTopic uses a NLP transformer model. BERTopic uses the Sentence Transformers library, which has a variety of pre-trained sentence embedding models, to represent each document as a single vector of many dimensions. Typically, this is between 384 for mini models such as the BERTopic default: MiniLM [27] and 768 for larger models. We experimented with different sentence embedding models, with the 'paraphrase-albert-small-v2' model giving the best performance. An important aspect of topic modelling is to visualise the high dimensional embeddings generated, necessitating reducing the embeddings while also preserving as much of the meaning the many dimensions provide. This provides the added benefit of greatly improving the performance of the subsequent clustering step [6]. The Uniform Manifold Approximation and Projection (UMAP) algorithm [28] is applied to reduce the dimensions, as it offers an improvement over other dimensionality reductions techniques such as t-Distributed Stochastic Neighbour Embedding (t-SNE) and Principle Component Analysis (PCA) by preserving local structure while maintaining global structure, while also having much faster speeds than t-SNE. The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm is used to cluster the embeddings. UMAP and HDBSCAN work well together to preserve a lot of the high-dimensional structure and cluster them without forcing outliers together. Finally, in the third step, c-TF-IDF (class-based TF-IDF) is used to extract significant words for each cluster on a class-by-class basis. All documents in a topic are concatenated into a class c, then the frequency of each term (word) in the class is calculated and divided by the total number of words in the document. This improves the topic model as it emphasises the importance of a term in a cluster rather than in an individual document.

Plotting the topics of a topic model is done using an intertopic distance map. This is a 2 or 3-dimensional visual representation of the clusters of a topic model, based on LDAVis [29]. BERTopic also utilises dimensionality reduction for graphing the intertopic distances of the topic model. BERTopic, as an unsupervised topic modelling technique, assumes that documents have no labels. Therefore it does not provide a cross-comparison between its labels and any manually created ones. To enable comparison with the manually-assigned labels, the BERTopic framework's intertopic distance map was extended by improving the 'hover over text' of each topic to display dataset-specific information, such as the original question labels of each topic, the percentage of the unsupervised topic model they made up and sample questions in each topic. Specifically, the visualisations are altered to display the following information:

- Calculating the labels each topic's questions originally had, as a percentage,
- Displaying sample questions from each topic, and the original label of the relevant question,
- The total number of questions in each topic.

These enhancements enable us to understand if a topic's documents better suited its generated labels or the original manually-tagged ones, as well as making the intertopic distance map much easier to interpret. For the mapping of question literals to topic label, a dictionary mapping the processed question literal to the original topic label was used, and then making a list of lists of the distributions within each generated topic. This was then added to a DataFrame which Plotly uses for graphing all the topics.

V. Implementation

OCTIS is used in this work for standardisation of experiments across different topic modelling libraries and to find the optimum hyperparameters for the topic models. OCTIS is a Python library for handling the components required for topic modelling comparison. It has a built-in optimisation algorithm which uses Bayesian optimisation to find the statistically best hyperparameters, secondly, it allows for the use of a variety of metrics for evaluation. Additionally, Weights & Biases (W&B) (https://wandb.ai/) is used as a platform for tracking, plotting and comparison of different experiments. The goal of using Weights and Biases and OCTIS together is to produce an extensive set of automated experiments with a wide variety of topic models, embedding models and hyperparameters, while also automatically plotting all these results in an easily exportable format.

A. Metrics

For an objective quantitative method for finetuning BERTopic as well as cross-comparison of models, two popular metrics for calculating a topic model's performance, namely, topic coherence (TC) and topic diversity (TD) have been used in our work. These metrics enable an 'objective' guide for optimising our topic models' hyperparameters. Topic coherence attempts to emulate the human interpretability of a topic, while topic diversity measures the distribution of the language of the topic model.

$$Topic\,diversity = \frac{t_{unique}}{t_n} \tag{1}$$

Equation (1) defines the topic diversity metric, with t_{unique} being the number of unique topics in the topic documents and t_n representing the total number of topic documents.

$$NPMI(x,y) = (\frac{log\frac{P(x,y)+\epsilon}{P(x)P(y}}{log(-P(x,y)+\epsilon})^{\gamma} \tag{2}$$

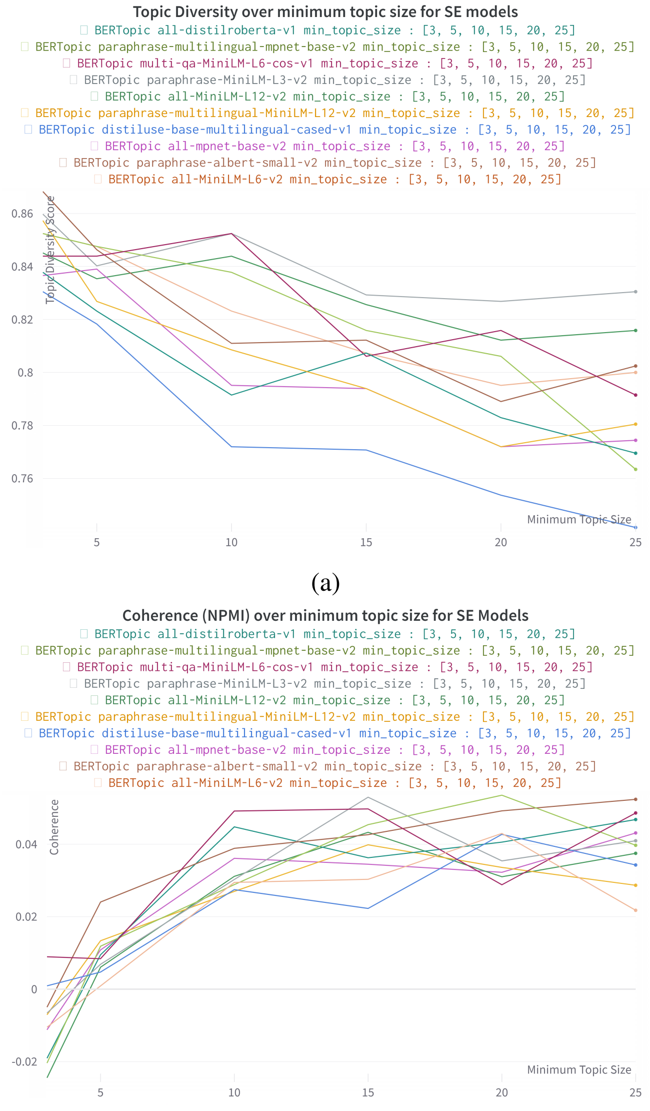
Topic coherence ensures that the documents within a topic belong together. The coherence metric adopted is normalized pointwise mutual information (NPMI). Coherence NPMI is a normalized metric extended from the UCI (University of California, Irvine) pointwise mutual information (PMI), but is normalized between -1 and 1. PMI is the probability of a pair of topic words with logarithmic calculation. Equation (2) denotes the NPMI calculation, where $P(x,y)$ represents the probability of token x and token y occuring in the documents, ϵ is a smoothing constant and a "higher γ gives a higher NPMI more weight" [30]. The positive value of NPMI coherence means that the topic is relevant to the document. Conversely, if the value is negative, the topic word is less relevant. The normalization ensures that abnormally large or small topic coherences do not outweigh topic diversity.

B. Implementation - Finetuning BERTopic

The following experiments have been performed to finetune BERTopic: (1) selecting the best performing pre-trained sentence embedding model, (2) finding optimal hyperparameters to produce clusters, and (3) dataset pre-processing and the impact on performance.

With the performance defined in terms of topic coherence and diversity, the main goal of our experiments to finetune BERTopic has been to understand the hyperparameters needed to increase topic coherence while maintaining topic diversity. The hyperparameters that are tuned are the (a) minimum topic size, and (b) number of topics.

Minimum topic size is an HDBSCAN parameter (ranging from 3 to 25), which decides whether smaller clusters will be generated. Intuitively, decreasing the minimum topic size means more potential topics (with fewer items) and therefore more topics being generated. Increasing the minimum topic size means less topics 'count' and therefore fewer topics, however the topics are likely to be more coherent although less diverse. For determining the 'best' sentence embedding model, BERTopic allows for the use of many different sentence embedding models from the HuggingFace hub. We run the minimum topic size experiment with a number of different state-of-the-art sentence embedding models from the SBert model selection, as shown in Figure 5.



(a)

(b)

Fig. 5. (a) Topic diversity and (b) Topic coherence of different BERTopic sentence embeddings (SE) versus minimum topic size

As seen in Figure 5(a), as the minimum topic size increases, the topic diversity of each model decreases. All models start out quite high at a minimum topic size of 3, but quickly plummet as the minimum topic size increases. The best model at lower minimum topic sizes is paraphrase-albert-small-v2 with a topic diversity of 0.8683, however this falls after the minimum topic size is equal to 10. The embedding model with the highest topic diversity overall is paraphrase-MiniLM-L3-V2, which retains topic diversity even at higher minimum topic sizes, while other models suffer great hits to the topic diversity.

Figure 5(b) shows that overall, paraphrase-albert-small has the best topic coherence. At the minimum topic size of 3, multi-qa-MiniLM-L6 has the highest topic coherence, however this quickly decreases as minimum topic size increases. Thus, we can conclude that increasing minimum topic size correlates with a higher topic coherence, and a lower topic diversity. In general, a lot of the models lose topic diversity after minimum

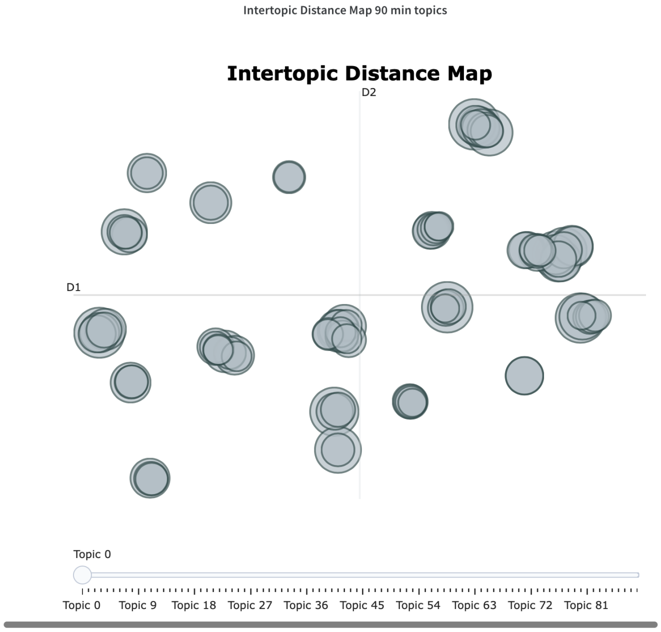
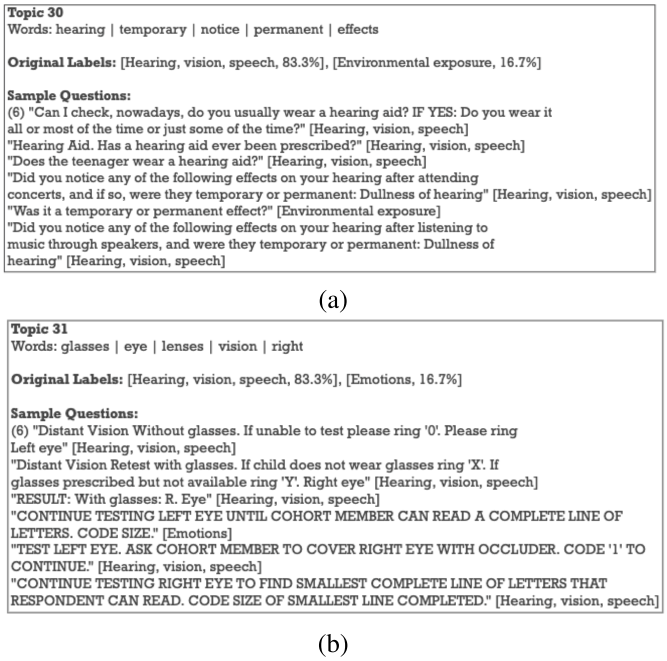


Fig. 6. Intertopic distance map with 90 topics



Topic 30
Words: hearing | temporary | notice | permanent | effects

Original Labels: [Hearing, vision, speech, 83.3%], [Environmental exposure, 16.7%]

Sample Questions:
(6) "Can I check, nowadays, do you usually wear a hearing aid? IF YES: Do you wear it
all or most of the time or just some of the time?" [Hearing, vision, speech]
"Hearing Aid. Has a hearing aid ever been prescribed?" [Hearing, vision, speech]
"Does the teenager wear a hearing aid?" [Hearing, vision, speech]
"Did you notice any of the following effects on your hearing after attending
concerts, and if so, were they temporary or permanent: Dullness of hearing" [Hearing, vision, speech]
"Was it a temporary or permanent effect?" [Environmental exposure]
"Did you notice any of the following effects on your hearing after listening to
music through speakers, and were they temporary or permanent: Dullness of
hearing" [Hearing, vision, speech]

(a)

Topic 31
Words: glasses | eye | lenses | vision | right

Original Labels: [Hearing, vision, speech, 83.3%], [Emotions, 16.7%]

Sample Questions:
(6) "Distant Vision Without glasses. If unable to test please ring '0'. Please ring
Left eye" [Hearing, vision, speech]
"Distant Vision Retest with glasses. If child does not wear glasses ring 'X'. If
glasses prescribed but not available ring 'Y'. Right eye" [Hearing, vision, speech]
"RESULT: With glasses: R. Eye" [Hearing, vision, speech]
"CONTINUE TESTING LEFT EYE UNTIL COHORT MEMBER CAN READ A COMPLETE LINE OF
LETTERS. CODE SIZE." [Emotions]
"TEST LEFT EYE. ASK COHORT MEMBER TO COVER RIGHT EYE WITH OCCLUDER. CODE '1' TO
CONTINUE." [Hearing, vision, speech]
"CONTINUE TESTING RIGHT EYE TO FIND SMALLEST COMPLETE LINE OF LETTERS THAT
RESPONDENT CAN READ. CODE SIZE OF SMALLEST LINE COMPLETED." [Hearing, vision, speech]

(b)

Fig. 7. Topic analysis through enhanced BERTopic visualization for (a) Topic 30 and (b) Topic 31.

topic size of 5.

The pre-processing steps that result in the best performance for the chosen metrics are using the CLOSER dataset's QuestionLiterals in all lower case with numbers and punctuation removed. These steps are done on the raw dataset, before the embeddings are generated by BERTopic. Using a count vectorizer, as recommended in the literature [6], where stop words are only removed after the embedding stage, (as transformer-based embeddings don't require preprocessing), degrades the topic coherence and diversity of the model and is therefore not used. Interactive graphs for the pre-processing experiments are documented in a W&B report available at this link: https://bit.ly/3THhvFi.

BERTopic works by first generating as many topics as it finds according to the hyperparameters, afterwards, optionally, the number of topics can be reduced to any specified number to increase interpretability. However, this comes at the cost of forcing dissimilar topics together. We set the number of topics to an interval of values between 10 and 110, in increments of 20, together with the above derived pre-processing steps, with the experiment results documented in this W&B report: https://bit.ly/3ALN5sV.

Looking at the graphs of the models with the highest topic coherence, shows this as 150 topics with an NPMI coherence of 0.03736, however this is with a lower topic diversity 0.8593. Comparatively, 90 topics has similar NPMI coherence of 0.03663 but a higher topic diversity of 0.8978. Therefore, 90 topics results in a better performing model with regards to both metrics. Overall, the approach with the best performance is BERTopic with the hyperparameters of 90 topics and minimum topic size of 5 and the sentence embedding model 'paraphrase-albert-small-v2'.

VI. Results and Evaluation

Figure 6 visualises the resulting intertopic distance map for the minimum topic number set to 5 and number of topics set to 90, with pre-processing. An interactive version of the intertopic distance map, showcasing the developed enhancements to the BERTopic plots, with the topic words, original labels and sample questions clustered in a topic is included in a W&B report available here: https://bit.ly/3ALN5sV. It shows a high-level of tight clustering, as most topics not only cluster but are virtually directly above each other (when seen in the zoomed-out default view). It is worthwhile to note that the optimum number of topics is very close to the number of level-2 CLOSER labels in the dataset, which is 82. In the following sub-sections, we analyse the topic clusters with some case studies comparing the generated topics with the manually-assigned labels.

A. Results - Overlapping Topic Clusters

We first analyse the two very overlapped topics: topic 30 and 31, whose details are obtained by the cursor mouse-over (hover) action in the intertopic distance map, as shown in Figure 7. The CLOSER taxonomy has a level-2 label for "Hearing, vision, speech". It is important to note that not only does our unsupervised topic model successfully group similar questions together, as seen from the original labels manually-assigned to the question literals, but it actually splits the level-2 topic into more intuitive groupings, with Topic 30 being about hearing, and topic 31 about vision. Both topics only contain 6 questions, however for both, 5 of the 6 are from the same level 2 group, "Hearing, Vision, Speech". The odd question from another level-2 group in topic 30 is "Was it a temporary or permanent effect?" which is probably because two other questions that mention hearing also mention this terminology.

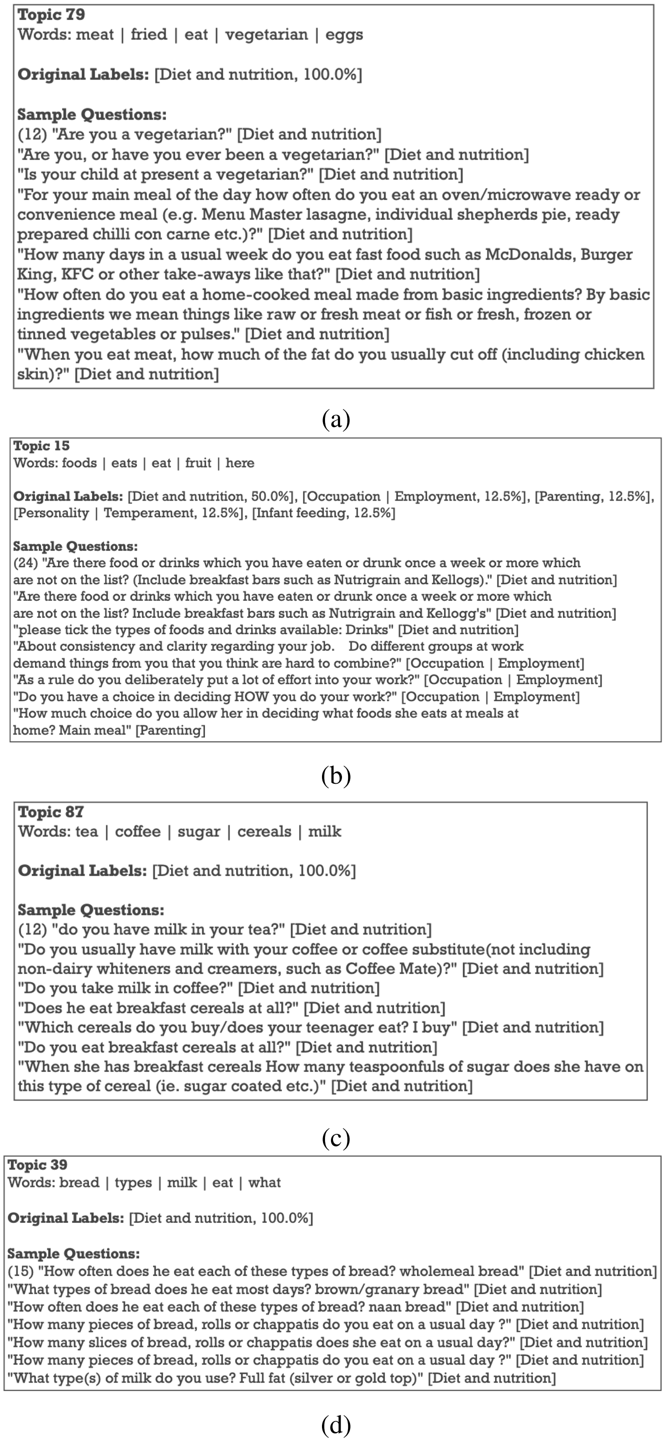
Therefore this analysis shows that the topic modeling may group unrelated questions together simply because they use the same question structure, such as "temporary or permanent". However, topic 31's only question not from the "Hearing Vision Speech" label is "Continue testing left eye until cohort member can read a complete line of letters. Code Size." It is unintuitive that this question should be in the "emotions" level-2 topic rather than also in "Hearing, Vision Speech" where it clearly belongs. Once flagged to the longitudinal studies' social scientists in the CLOSER project group, they agree that this question literal is in fact mislabelled! This showcases that the unsupervised BERTopic modelling approach can uncover errors in human labelling of the data.

B. Results - Semantically-similar Topic Clusters with Outliers

The unsupervised topic modelling approach generates a few clusters that relate to the concept of 'eating food'. As shown in Figure 8(a), topic 79 is comprised entirely of questions from the 'Diet and Nutrition' CLOSER level-2 label, which relate mostly to meat eating and vegetarianism. Topic 15 (Figure 8(b)), on the other hand, has half of its 24 questions from the 'Diet and Nutrition' manual tag, and has some outlier questions from 'Occupation - Employment' which relate to work style instead. Topic 87 (Figure 8(c)) is also comprised entirely of 'Diet and Nutrition' tagged questions on cereal, tea and coffee. It is clear that the topic modeling connected breakfast cereal with tea/coffee by the association of breakfast. Topic 39 (Figure 8(d)) is also entirely 'Diet and Nutrition' tagged questions on the 'bread' topic. However, it has one question which has nothing to do with bread but mentions milk. One hypothesis for why the "What type(s) of milk do you use? Full fat (silver or gold top)" question appears in this topic rather than topic 87 which also mentions milk, is the syntactic structure of the other questions, that mention the first questionnaire response to the question. In conclusion, for this cluster the model correctly clustered several similar food-related topics together; however, there were a couple of topics with some outliers and questions which were only related syntactically rather than semantically.

C. Results - Longitudinal Topic Modelling

To take advantage of the fact that the questionnaires are from different time periods, often spanning over years, an insightful approach is dynamic topic modelling. This involves seeing how topics change over time by tracking the timestamps of the questions within each topic. As the studies not only span a length of time but often ask parents, children and the children once they're grown up different types of questions, it can be insightful to see how the topics change as the participants age.



Topic 79
Words: meat | fried | eat | vegetarian | eggs

Original Labels: [Diet and nutrition, 100.0%]

Sample Questions:
(12) "Are you a vegetarian?" [Diet and nutrition]
"Are you, or have you ever been a vegetarian?" [Diet and nutrition]
"Is your child at present a vegetarian?" [Diet and nutrition]
"For your main meal of the day how often do you eat an oven/microwave ready or convenience meal (e.g. Menu Master lasagne, individual shepherds pie, ready prepared chilli con carne etc.)?" [Diet and nutrition]
"How many days in a usual week do you eat fast food such as McDonalds, Burger King, KFC or other take-aways like that?" [Diet and nutrition]
"How often do you eat a home-cooked meal made from basic ingredients? By basic ingredients we mean things like raw or fresh meat or fish or fresh, frozen or tinned vegetables or pulses." [Diet and nutrition]
"When you eat meat, how much of the fat do you usually cut off (including chicken skin)?" [Diet and nutrition]

(a)

Topic 15
Words: foods | eats | eat | fruit | here

Original Labels: [Diet and nutrition, 50.0%], [Occupation | Employment, 12.5%], [Parenting, 12.5%], [Personality | Temperament, 12.5%], [Infant feeding, 12.5%]

Sample Questions:
(24) "Are there food or drinks which you have eaten or drunk once a week or more which are not on the list? (Include breakfast bars such as Nutrigrain and Kellogs)." [Diet and nutrition]
"Are there food or drinks which you have eaten or drunk once a week or more which are not on the list? Include breakfast bars such as Nutrigrain and Kellogg's" [Diet and nutrition]
"please tick the types of foods and drinks available: Drinks" [Diet and nutrition]
"About consistency and clarity regarding your job. Do different groups at work demand things from you that you think are hard to combine?" [Occupation | Employment]
"As a rule do you deliberately put a lot of effort into your work?" [Occupation | Employment]
"Do you have a choice in deciding HOW you do your work?" [Occupation | Employment]
"How much choice do you allow her in deciding what foods she eats at meals at home? Main meal" [Parenting]

(b)

Topic 87
Words: tea | coffee | sugar | cereals | milk

Original Labels: [Diet and nutrition, 100.0%]

Sample Questions:
(12) "do you have milk in your tea?" [Diet and nutrition]
"Do you usually have milk with your coffee or coffee substitute(not including non-dairy whiteners and creamers, such as Coffee Mate)?" [Diet and nutrition]
"Do you take milk in coffee?" [Diet and nutrition]
"Does he eat breakfast cereals at all?" [Diet and nutrition]
"Which cereals do you buy/does your teenager eat? I buy" [Diet and nutrition]
"Do you eat breakfast cereals at all?" [Diet and nutrition]
"When she has breakfast cereals How many teaspoonfuls of sugar does she have on this type of cereal (ie. sugar coated etc.)" [Diet and nutrition]

(c)

Topic 39
Words: bread | types | milk | eat | what

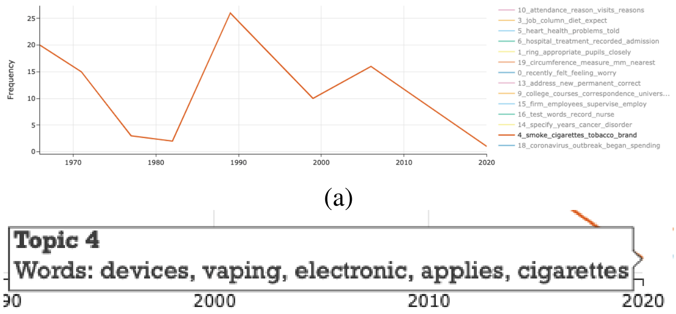
Original Labels: [Diet and nutrition, 100.0%]

Sample Questions:
(15) "How often does he eat each of these types of bread? wholemeal bread" [Diet and nutrition]
"What types of bread does he eat most days? brown/granary bread" [Diet and nutrition]
"How often does he eat each of these types of bread? naan bread" [Diet and nutrition]
"How many pieces of bread, rolls or chappatis do you eat on a usual day ?" [Diet and nutrition]
"How many slices of bread, rolls or chappatis does she eat on a usual day?" [Diet and nutrition]
"How many pieces of bread, rolls or chappatis do you eat on a usual day ?" [Diet and nutrition]
"What type(s) of milk do you use? Full fat (silver or gold top)" [Diet and nutrition]

(d)

Fig. 8. Food-related topic clusters: topic analysis through enhanced BERTopic visualization for (a) Topic 79, (b) Topic 15, (c) Topic 87, and (d) Topic 39.

To track the evolution of topics within a study, we extract from the dataset the questions in the study, along with their dates, and exclusively use those to build a topic model. We apply dynamic topic modelling to the MRCNSHD study, which has a variety of life stages, and questionnaires ranging from the 1940s to the 2020s, as shown in Figures 1 and 2. An interesting example of an evolving topic from this study is topic 4, with the top words of 'smoke, cigarettes, tobacco, brand', depicted in Figure 9(a). This is quite obviously the smoking habits topic. The cigarette topic only starts in the mid-60s, which aligns with the fact that the participants would



(a)

(b)

Fig. 9. (a) Development of topic 4 (Smoking) in the MRCNSHD longitudinal study, and (b) evolution of the topic words in 2020, surfacing the lexicon of electronic cigarettes.

probably be too young to be asked about their smoking habits before then, being born in the 40s. The 2020 datapoint for this topic even starts to include questions about vaping, as shown in Figure 9(b), showing the semantic shift of the topic as the noun 'vaping' only came into common parlance in the early 2010s [31].

Similarly, it is only in the 1960s when the "college courses correspondence universities" topic starts (Figure 10), timing when they were around 20, that participants are asked about university and college. Thus, dynamic topic modelling is a useful tool for understanding the development of topics over not only time but over a participant's lifetime. It successfully demonstrates the complexity of the data both in terms of a data cleaning perspective and in the variety and breadth of sociological and biomedical questions asked. In the 'vaping' example we gained insight into the development of a societal issue using only our dynamic topic modelling approach, while also surmising about possible semantic shifts.

D. Evaluation

We compare the performance of our fine-tuned BERTopic model with the following state-of-the-art topic modelling baselines:

- Top2Vec: in some ways, a predecessor to BERTopic. It uses a joint approach of document and word embeddings, followed by a dimensionality reduction and clustering algorithm to find dense areas in the data, called clusters. The centroid of each of these clusters then becomes the topic vector, with the main difference to BERTopic being that instead of c-TF-IDF, Top2Vec uses the n closest words that become the words which define that topic. Its strength is that it requires no pre-processing and works on short text, due to the power of its joint embedding.
- Contextualized Topic Modelling (CTM): the CTM technique improves upon coherence of topic models compared to the traditional methods such as bag-of-words and early neural network-based topic modelling. It significantly increases topic coherence while maintaining topic diversity by improving upon Neural ProdLDA variational autoencoding approach [32], by adding contextualized representations via SBERT sentence embeddings.

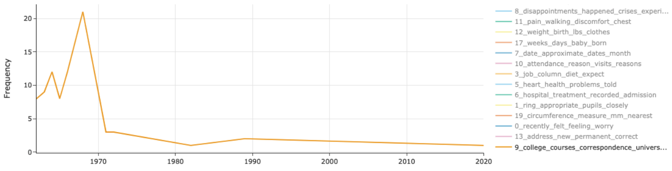
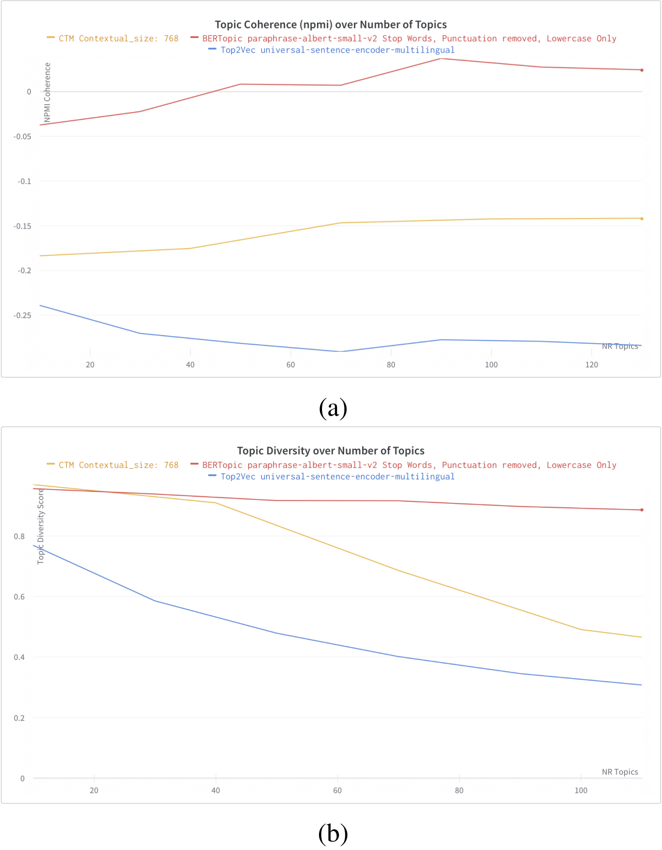


Fig. 10. (a) Development of topic 9 (Higher Education) in the MRCNSHD longitudinal study.



(a)

(b)

Fig. 11. Comparison of our proposed fine-tuned BERTopic model with CTM and Top2Vec in terms of (a) NPMI topic coherence, and (b) topic diversity.

For both these baselines, the primary experiments revolve around how the defined metrics of topic diversity and topic coherence change over the number of topics in the model, with a secondary consideration being the computation time of the models. The default pre-processing steps and settings are used for the baselines, i.e. removal of punctuation for CTM's bag-of-words creation step. We also optimise Top2Vec with different embedding models in a manner similar to our BERTopic experiments, though only over a range of 'number of topics' as that is Top2Vec's independent variable. Four embedding models were experimented with, across a range of 10 to 150 topics: universal-sentence-encoder, universal-sentence-encoder-multilingual, distiluse-base-multilingual-cased and paraphrase-albert-small-v2, with the universal-sentence-encoder-multilingual showing the best overall performance. Thus, Top2Vec is used with this embedding model in our evaluation experiments with BERTopic.

Figure 11 shows the comparison of our proposed BERTopic model with CTM and Top2Vec on the defined evaluation metrics of NPMI topic coherence (TC) and topic diversity (TD), plotted against a range of 'number of topics (NR)'. In terms of our most important parameter, TC, it is clear that BERTopic significantly outperforms both CTM and Top2Vec across all topic numbers. Additionally, BERTopic has non-negative values for its TC for all NR>50 while other models have negative values. BERTopic has a best TC value of 0.03663 at NR = 90, compared to CTM with -0.1429 and Top2Vec at -0.2782.

For lower numbers of topics (<20), CTM has a slightly higher TD than BERTopic, however this plummets past 40 topics and is far lower than BERTopic for all higher NR topics values. Moreover, it is at the cost of far lower TC. The trend for TD over NR for Top2Vec is that it is highest at 10 topics and then decreases exponentially. TD values are 0.8978 for BERTopic, 0.492 for CTM, and 0.3467 for Top2Vec at NR = 90.

In terms of computation time, it is worth noting that CTM is far slower than BERTopic and Top2Vec. Each run takes 450 seconds for CTM but only 25 seconds for BERTopic, with the slower computation of CTM attributable to its default setting of the t-SNE algorithm for dimensionality reduction rather than UMAP, with UMAP being faster than t-SNE. Top2Vec takes approximately 40 seconds per run.

Interactive plots of the computation time comparison, as well as for TC and TD, are available in a W&B report here: https://bit.ly/3TDCfhg.

VII. Conclusions

In this paper, we present an approach for tailoring the state-of-the-art BERTopic topic modeling architecture for modeling longitudinal questionnaire data. Our approach takes advantage of the pre-trained sentence embedding models with more accurate and context-aware representation of words and sentences to create interpretable topics with better topic coherence and diversity than existing NLP topic modelling techniques like CTM and Top2Vec. The extensions to the BERTopic visualization capability enables comparison of the generated topic model words to the manually-assigned labels of the question texts, successfully uncovering instances of 'label bias' in the dataset. The dynamic topic modeling has also showcased the evolution of the topic words and semantic shifts across the time span of the longitudinal studies, uncovering the changing lexicon and emerging words in the question texts across the lifetime of the questionnaire participants. In the hands of social science domain experts, the developed tools could greatly aid their intuitive understanding of societal issues over time and aid novel insights.

We note CTM's use of pyLDAvis that allows for more information about the top terms in each topic than BERTopic by default provides. In our future work, we aim to use pyLDAvis with BERTopic to further enhance the interpretability of topic evolution. We will also use a CUDA-enabled BERTopic library called cuBERTopic to facilitate a GPU-enabled embeddings computation and also calculate the probability a question was placed in each topic, as the stochastic nature of UMAP means each generated topic model has an element of randomness.

References

[1] "CLOSER discovery - content. UCL wiki," 2021. [Online]. Available: https://discovery.closer.ac.uk/page/content/8

[2] "Medical subject headings (MeSH) thesaurus," 2016. [Online]. Available: https://www.nlm.nih.gov/mesh/meshhome.html

[3] "Humanities and social science electronic thesaurus (HASSET)," 1997. [Online]. Available: https://hasset.ukdataservice.ac.uk/hasset/en/

[4] CLOSER, "CLOSER discovery - topics," 2020. [Online]. Available: https://wiki.ucl.ac.uk/display/CLOS/Topics

[5] D. S. Shah, H. A. Schwartz, and D. Hovy, "Predictive biases in natural language processing models: A conceptual framework and overview," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[6] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," Mar. 2021. [Online]. Available: https://arxiv.org/abs/2203.05794

[7] S. Terragni, "Octis: Comparing and optimizing topic models is simple!" in *16th Conference of the European Chapter of the Association for Computational Linguistics (EACL) - Demo Track*, 2021. [Online]. Available: https://github.com/MIND-Lab/OCTIS

[8] D. Angelov, "Top2vec: Distributed representations of topics," *arXiv Computation and Language*, 2020. [Online]. Available: https://arxiv.org/abs/2008.09470

[9] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence," in *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Online, Aug. 2021, pp. 759–766. [Online]. Available: https://aclanthology.org/2021.acl-short.96

[10] J. Ramos, "Using tf-idf to determine word relevance in document queries," 2003.

[11] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," 2014, cite arxiv:1402.3722. [Online]. Available: http://arxiv.org/abs/1402.3722

[12] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *NAACL*, 2018. [Online]. Available: http://arxiv.org/abs/1802.05365

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, cite arxiv:1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805

[15] T. Singhal, J. Liu, L. T. M. Blessing, and K. H. Lim, "Analyzing scientific publications using domain-specific word embedding and topic modelling," in *2021 IEEE International Conference on Big Data (Big Data)*, dec 2021.

[16] J. Giorgi, O. Nitski, B. Wang, and G. Bader, "DeCLUTR: Deep contrastive learning for unsupervised textual representations," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, p. 993–1022, mar 2003.

[18] Y. Zhou, S. De, and K. Moessner, "Real world city event extraction from twitter data streams," *Procedia Computer Science*, vol. 98, pp. 443–448, 2016.

[19] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, 2010.

[20] S. V. Raju, B. K. Bolla, D. K. Nayak, and J. Kh, "Topic modelling on consumer financial protection bureau data: An approach using BERT based embeddings," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, apr 2022.

[21] Y. N. Pek and K. H. Lim, “Identifying and understanding business trends using topic models with word embedding,” in *2019 IEEE International Conference on Big Data (Big Data)*, dec 2019.

[22] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. [Online]. Available: https://aclanthology.org/Q17-1010

[23] D. Hendry, F. Darari, R. Nurfadillah, G. Khanna, M. Sun, P. C. Condylis, and N. Taufik, “Topic modeling for customer service chats,” in *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, oct 2021.

[24] “Closer - timeline of studies,” 2021. [Online]. Available: https://www.closer.ac.uk/timeline/

[25] “Colectica repository,” 2021. [Online]. Available: https://www.colectica.com

[26] S. De, H. Moss, J. Johnson, J. Li, H. Pereira, and S. Jabbari, “Engineering a machine learning pipeline for automating metadata extraction from longitudinal survey questionnaires,” *IASSIST Quarterly*, vol. 46, no. 1, Mar. 2022.

[27] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, “Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers,” in *Proc. 34th International Conference on Neural Information Processing Systems (NIPS’20)*, ser. NIPS’20, Red Hook, NY, USA, 2020.

[28] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” *arXiv:1802.03426*, 2018. [Online]. Available: https://arxiv.org/abs/2203.05794

[29] C. Sievert and K. Shirley, “LDAvis: A method for visualizing and interpreting topics,” in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Association for Computational Linguistics, 2014.

[30] Q. Fu, Y. Zhuang, J. Gu, Y. Zhu, H. Qin, and X. Guo, “Search for k: Assessing five topic-modeling approaches to 120,000 Canadian articles,” in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, dec 2019.

[31] “Use of e-cigarettes among adults in great britain,” Action on Smoking and Health (ASH), Tech. Rep., 2022. [Online]. Available: https://ash.org.uk/resources/view/use-of-e-cigarettes-among-adults-in-great-britain-2021

[32] A. Srivastava and C. Sutton, “Autoencoding variational inference for topic models,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26*, 2017.