Topic Modelling with Topological Data Analysis

Anonymous ACL submission

Abstract

Recent unsupervised topic modelling approaches that use clustering techniques on word, token or document embeddings can extract coherent topics. However, a common limitation of such approaches is that they reveal nothing about inter-topic relationships which are essential in many real-world application domains. We present an unsupervised topic modelling method which harnesses Topological Data Analysis (TDA) to extract a topological skeleton of the manifold upon which contextualised word embeddings lie. We demonstrate that our approach, which performs on par with a recent baseline, is able to construct a network of coherent topics together with meaningful relationships between them.

1 Introduction

Unsupervised topic modelling is a standard technique for making sense of document collections. While traditional approaches such as LDA (Blei et al., 2003) rely on probabilistic models, the field has recently moved towards clustering-based methods in which topic clusters are obtained via document, word or token embeddings (Thompson and Mimno, 2020; Silburt et al., 2021; Angelov, 2020). Even though clustering can yield interpretable topics, it typically discards information about relationships between clusters, hence making it harder to interpret clusters in global contexts.

In this work, we approach topic modelling as a task to find regions on a manifold of contextualised word embeddings which reflect a "topic". To this end, we apply Mapper - an algorithm from the field of Topological Data Analysis (TDA). Mapper creates a graph whose topology reflects the shape of the underlying data set and whose nodes represent subsets of data points. In the case of contextualised word embeddings, we construct a graph where each node is a cluster of tokens (i.e. a "topic"), and where connections between them reflect the topology of the embedding manifold. We use community detection techniques to demonstrate that semantically related topics are connected in the graph.

Our main contributions are the following:

1. We propose and evaluate a new method for topic modelling which learns topics and relationships between them without any restrictions on graph structure. To the best of our knowledge, our work is the first application of TDA Mapper to the task of topic modelling.

2. To the best of our knowledge, we are the first to use stability analysis for Mapper on a real-world data set and problem. Unlike prior approaches which are computationally infeasible on large data sets, we propose a scalable approach using separate stability scores for both the graph topology and the clustering.

3. We define a new stability score via spectral distance between Mapper graphs.

4. We use community detection techniques to automatically identify regions of interest in large Mapper graphs.

The paper is organised as follows. In Section 2, we review related work. Section 3 presents our method, and summarises TDA Mapper and stability analysis. We describe our experimental setup, including the data set, baselines, and metrics in Section 4. Our empirical results and further qualitative analyses are presented in Section 5.

2 Related Work

The seminal work on unsupervised topic modelling was Blei et al. (2003) who introduced Latent Dirichlet Allocation (LDA), a Bayesian generative model of documents which assumes that the tokens in a document are drawn from a mixture model whose mixture components are interpreted as topics. Of

the many extensions to the classic LDA archetype that have since been proposed, most relevant to our present work are methods to model associations and relationships between topics, and the use of neural representations in general and contextualised representations in particular.

Correlated topic models (Lafferty and Blei, 2006; Blei and Lafferty, 2007) are LDA extensions that attempt to learn the structure of topic associations within a document. The goal of hierarchical topic models (Griffiths et al., 2004; Wang and Blei, 2009; Blei et al., 2010; Ghahramani et al., 2010; Zavitsanos et al., 2011; Ahmed et al., 2013; Paisley et al., 2014) is to learn a tree-structured graph of topics by incorporating hierarchical non-parametric Bayesian priors into traditional topic models.

Several studies have combined topic modelling with neural representations with a view to learn better topics or representations. For example, amortised variational inference with neural variational posteriors (Kingma and Welling, 2014) has been investigated as a means to scale up inference on probabilistic topic models and relax the conjugacy assumptions which are required for tractable inference in traditional topic models (Srivastava and Sutton, 2017). Various variants of such models have focused on neural extensions of correlated (Xun et al., 2017; Liu et al., 2019) and hierarchical (Isonuma et al., 2020) topic models although they all use neural representations in the generative model or variational posterior.

The prior work most closely related to our proposed method is the joint application of topic modelling and contextualised word embeddings by Thompson and Mimno (2020), Sia et al. (2020) and Angelov (2020) who induce topics via vector clustering over word or document embeddings.

Our method differs from LDA and its extensions in that we use TDA rather than probabilistic generative models to induce topics. Correlated topic models and their neural extensions learn a flat topic structure while adding scalar associations, whereas our method induces a topic graph. In contrast to hierarchical topics models and their neural extensions which induce *tree-structured* topic graphs, our method induces an *unrestricted* graph. Unlike our method, previous work on inducing topics from contextualised word representations construct a flat topic structure rather than a graph.

Also related to our work is TopoAct (Rathore et al., 2021) which applies Mapper to the analysis of BERT word embeddings. Our work differs from *ibid.* in that we focus specifically on topic modelling, and we follow a systematic hyperparameter selection process through stability analysis.

3 Proposed Method

The manifold hypothesis (Goodfellow et al., 2014) states that real-world high-dimensional data lie on a low-dimensional manifold embedded in a high-dimensional space. Topic modelling can be regarded as an endeavour to identify topologically meaningful regions of the word representation manifold which contain homogeneous topics or words. Traditionally, it has been approached as a clustering problem in that the representation manifold is assumed to be a disconnected union of "topic" manifolds. However, such an assumption is clearly limiting and not grounded theoretically. One potential solution involves dimensionality reduction and direct manifold visualisation. Unfortunately, most dimensionality reduction techniques capture only topology within local neighbourhoods, and cannot be relied upon for inference regarding the global topology of the manifold.

Our method of choice to address this problem is TDA Mapper introduced in (Singh et al., 2007) (also referred to as topological data visualisation or topological clustering), a method that yields an approximation of a Reeb graph of a manifold (Munch and Wang, 2016) which captures the topology and shape of the manifold. Reeb graphs are constructed from a manifold in order to learn topological invariants and global structure. Even though they lose some of the original topological structure of the manifold, their low-dimensional invariants (e.g. connected components) remain the same.

3.1 Overview of TDA Mapper

The TDA Mapper algorithm takes as input a set of points and outputs a graph whose vertices are subsets of points, and whose edges are defined between vertices which have a non-empty intersection. The following main steps are typically executed.

1. The data is projected to a lower dimension using a "**filter function**" (or "lens") f. This can be any standard dimensionality reduction function or even a domain-specific function which captures some interesting property of the data.

2. The projected space is covered with a set of overlapping sets $(U_i)_{i \in I}$.

3. Each set U_i is "pulled back" into the original high-dimensional space by taking its preimage $f^{-1}(U_i)$. The points in this "**pull-back set**" are broken into clusters using a clustering algorithm.

4. A graph is constructed by using each cluster as a vertex and adding an edge between any two clusters that have a non-empty intersection.

3.2 Hyperparameter Tuning for TDA Mapper

Model selection in TDA Mapper is non-trivial, the main reason being the absence of ground truth labels, analogous to what other unsupervised learning algorithms face. One model selection approach suitable for algorithms of this kind which has recently gained traction is stability analysis (see (Luxburg, 2010)). Rather than configuring clustering parameters up front and then optimising an evaluation metric, stability analysis simply constrains clustering to return structures that are stable under small perturbations of data. For example, let $\mathcal{M}_\theta(D)$ be a certain mathematical structure on a data set D with parameters θ where $\mathcal{M}_\theta$ could be clustering, dimensionality reduction, TDA Mapper, or some other unsupervised learning algorithm. If there exists a distance measure to quantify the similarity of the structures $d(\mathcal{M}, \mathcal{M}')$, then we can define the instability of $\mathcal{M}$ for the parameter choice θ as the expected distance between $\mathcal{M}_\theta(D)$ and $\mathcal{M}_\theta(D')$, where D and D' are two data samples obtained by the same data generation process. More precisely,

$$\mathcal{S}(\mathcal{M}_\theta, d) = \frac{2}{n(n-1)} \sum_{i=0}^{n} \sum_{j=i+1}^{n} d(\mathcal{M}_\theta(D_i), \mathcal{M}_\theta(D_j)) \quad (1)$$

where $\mathcal{S}$ denotes the instability score, and D_i are independent samples from the dataset D. Finally, the optimal set of parameters θ for structure $\mathcal{M}$ is chosen from the ones that have a low instability score $\mathcal{S}$. Note that the instability score should only be used to rule out parameter choices that yield high instability scores; it alone cannot be used for parameter selection as some structures are stable but not necessarily correct. It is crucial to choose the distance function which best embodies the notion of similarity between mathematical structures $\mathcal{M}$ in order to obtain meaningful results from stability analysis. One such distance function for TDA Mapper graphs was defined and studied in (Belchí et al., 2020). Unfortunately, their numerical matching distance algorithm is prohibitively slow in our use case. We accordingly define two alternative distance metrics to capture two salient properties of Mapper graphs. One is designed to capture similarity amongst graph structures while the other accounts for vertex (or cluster) similarity.

These concepts are defined formally as follows.

Definition 1 *Let $\mathcal{M}_\theta(D)$ be a TDA Mapper graph with a vertex set $V = \{C_1, \ldots, C_m\}$ where $C_i \subset D$; and an edge set $E = \{(C_i, C_j) \mid$ if $C_i \cap C_j \neq \emptyset\}$ where $\theta = (\theta_1, \theta_2, \theta_3)$ are three groups of parameters pertaining to a filter function, cover, and clustering algorithm, respectively.*

The stability of Mapper graphs is then assessed with respect to different choices of parameters θ, and the final parameter values are chosen from the most stable regions of the landscape.

We further define two distance metrics on Mapper graphs for stability analysis.

Definition 2 *Let $\mathcal{M}$ and $\mathcal{M}'$ be two TDA Mapper graphs with vertices $V = \{C_1, \ldots, C_n\}$; $V' = \{C'_1, \ldots, C'_m\}$; and edges E and E', respectively. If $m \neq n$, then empty set padding is added to the smaller vertex set so that $m = n$. The distance*

$$d_m(\mathcal{M}, \mathcal{M}') = \min_\pi \frac{1}{n} \sum \mid C_i \triangle C'_{\pi i} \mid \quad (2)$$

where π runs over all permutations of the set $\{1, 2, \ldots, n\}$, is called the matching distance and quantifies the similarity of vertices between Mapper graphs.

Definition 3 *Let $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, $\Lambda' = \{\lambda'_1, \lambda'_2, \ldots, \lambda'_m\}$ be eigenvalues of the normalised Laplacian defined on Mapper graphs $\mathcal{M} = G(V, E)$ and $\mathcal{M}' = G(V', E')$, respectively. The spectral distance is defined within the distribution of the eigenvalues $\mu = \sum_{\lambda \in \Lambda} p_\lambda \delta_\lambda$ and $\nu = \sum_{\lambda' \in \Lambda'} p_{\lambda'} \delta_{\lambda'}$ as their 1-Wasserstein distance, i.e.*

$$d_s(\mathcal{M}, \mathcal{M}') = \int_{-\infty}^{+\infty} F_\mu(t) - F_\nu(t) \mathrm{d}t \quad (3)$$

where F_μ and F_ν are CDFs for μ and ν.

The spectral distance quantifies the similarity of graph topologies amongst graphs (Gu et al., 2015). Lastly, let Θ be the search space for parameters θ: then the stable region of Θ with permissible

parameter choices is

$$\Theta_S = \{\theta \in \Theta \mid \mathcal{S}(\mathcal{M}_\theta, d_m) < \varepsilon_m \text{ and } \mathcal{S}(\mathcal{M}_\theta, d_s) < \varepsilon_s\}, \quad (4)$$

where ε_m and ε_s are thresholds for distances that are considered "large" and hence unstable.

4 Experiments

4.1 Data

We evaluated the proposed model on the *20 Newsgroups* data set[1] which contains 18846 English language posts categorised into thematic newsgroups. We use the standard train-test split. Table 1 summarises per-category document frequencies in the training set. We remove email addresses, headers, and subject lines. We extract contextualised subword embeddings using `bert-base-uncased`[2] (Devlin et al., 2019), and use the last layer embeddings. When a document exceeds 512 tokens (cf. the max length for BERT), we simply run the model on each block of 512 tokens. To obtain word embeddings, we take the mean of the subword components. The documents are tokenised using `spaCy`[3], and BERT subword tokens are aligned to `spaCy` tokens with `spacy-alignments`[4].

Although pretrained language models can represent them, we decided to remove rare words on the grounds of lighter compute requirements. Following Thompson and Mimno (2020), we remove stopwords, skip punctuation and digits, and further remove any tokens which occur in fewer than 5 documents or more than 25% of the documents. This yields a vocabulary with 14829 words. Note that we only remove these tokens after word embeddings have been obtained since they are important for downstream representations.

4.2 Methodology

We apply the Mapper algorithm to the resultant data set of contextualised word representations. For our filter function, we use UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2020). We reduce the data down to two dimensions via the default parameters for UMAP's Python reference implementation[5].

For clustering, we use HDBSCAN[6], a density-based clustering algorithm which automatically determines the number of clusters in a set of points (Campello et al., 2013). The main parameter for HDBSCAN is `min_cluster_size`, the smallest number of points that can constitute a cluster, which we set to 15.

20 Newsgroups Category	# Documents
alt.atheism	480
comp.graphics	584
comp.os.ms-windows.misc	591
comp.sys.ibm.pc.hardware	590
comp.sys.mac.hardware	578
comp.windows.x	593
misc.forsale	585
rec.autos	594
rec.motorcycles	598
rec.sport.baseball	597
rec.sport.hockey	600
sci.crypt	595
sci.electronics	591
sci.med	594
sci.space	593
soc.religion.christian	599
talk.politics.guns	546
talk.politics.mideast	564
talk.politics.misc	465
talk.religion.misc	377

Table 1: Summary of the 20 Newsgroups training set.

4.3 Parameter Selection

Aside from the clustering and filter function, Mapper requires a "**cover**". A standard choice is to partition the co-domain of the filter function into a number of equally sized, overlapping intervals or hypercubes in higher dimensions (Chazal and Michel, 2021). However, after applying UMAP, we noticed that the data exhibited non-uniform density. This caused some cover sets to have many more data points, making the clustering step computationally unfeasible. To address this, we used the

[1] Via `scikit-learn` https://scikit-learn.org/stable/datasets/real_world.html#newsgroups-dataset

[2] https://huggingface.co/bert-base-uncased

[3] `core_web_lg v3.0.0` https://spacy.io

[4] https://pypi.org/project/spacy-alignments

[5] https://umap-learn.readthedocs.io/en/latest

[6] https://hdbscan.readthedocs.io/en/latest/index.html

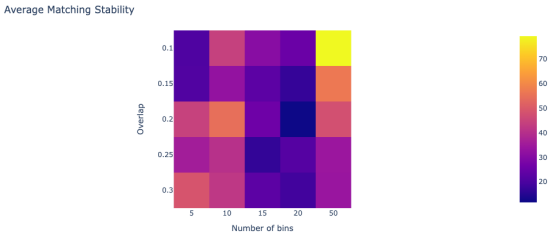

Figure 1: Matching Distance Scores for different parameter values.

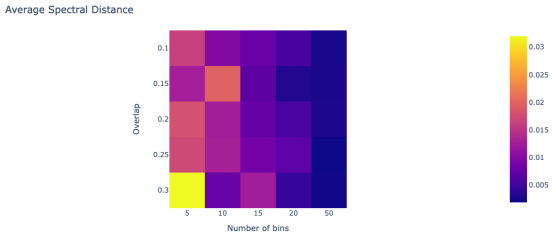

Figure 2: Spectral Distance Scores for different parameter values.

"balanced" cover offered by the `giotto-tda`[7] library which adjusts the size of each bin so that each cover set contains a similar number of data points.

This cover requires two additional parameters: (i) the number of intervals or bins and (ii) the percentage overlap. We perform a stability analysis to rule out unstable parameter combinations whose topological features are more likely to be mere artefacts. We experiment with 5, 10, 20, and 50 intervals and overlaps of 0.1, 0.15, 0.2, 0.25, and 0.3. For computational reasons, we perform the stability analysis on a randomly selected subset of 150K word embeddings. We further subdivide the subset into 3 samples, each with 100K word embeddings whereby each pair of subsamples overlaps by 50%. We run Mapper on each sample subset to generate 3 graphs for each pair of parameters.

We compute an instability score for each parameter set as the average distance between all three graphs. We conduct the stability analysis twice using two separate metrics, namely 1) Matching Distance (see Definition 2) to measure clustering stability; and 2) Spectral Graph Distance (see Definition 3) to measure stability in the graph structure. Our stability plots are shown in Figures 1 and 2.

Looking at the regions that appear stable under both metrics, we are still left with multiple choices for stable parameters. We ultimately select a bin size of 20 and an overlap of 0.1 following an intuition that (i) larger overlaps lead to highly connected graphs with less interesting structure since the data is relatively dense; and (ii) extreme values for the number of bins should be avoided for they lead to excessively coarse or fine granularity.

4.4 Community Detection for Subgraphs

We noticed that the majority of the data points resided in the largest connected component of the graph. Moreover, there were a large number of individual disconnected nodes, which contained about 30% of the tokens. Since we are mainly interested in exploring the connections output by TDA, we simply discard these nodes and focus the rest of our analysis on the largest connected component.

Since the graph is large, exploring all areas of it manually is cumbersome. Therefore, we used a community detection algorithm to identify clusters of nodes that are densely connected. We form additional higher-level topics from these clusters by taking the union of all tokens in the nodes in scope. We report metrics at both the node- and at the community-level.

For community detection, we use the label propagation algorithm described in (Raghavan et al., 2007) via `iGraph`[8] which is adapted to consider edge weights (Csárdi and Nepusz, 2006).

4.5 Baseline

As a baseline, we chose Top2Vec (Angelov, 2020), a recent method based on document representations and clustering. Following *ibid.*, we build a Top2Vec model using `Doc2Vec` document embeddings which we train for 400 epochs with a window size of 15.

4.6 Evaluation Metrics

We use three automated metrics to evaluate our model with respect to topic coherence, diversity, and specificity. It is important to note, however, that automated evaluation of topic coherence is an activate area of research, and that standard evaluation metrics have well-known limitations: in particular, automated measures can detect differences between topic models in cases where human judgements do not (Hoyle et al., 2021). The primary goal of our work is not to reach greater coherence

[7] https://github.com/giotto-ai/giotto-tda

[8] https://igraph.org/

per se but rather to arrange topics in a meaningful graph structure for which comparisons with baselines through automated measures suffice. In addition to reporting three standard automated evaluation measures, we also inspect some of our topics within some newsgroup categories.

Firstly, we estimate topic coherence by taking the average NPMI (Normalized Pointwise Mutual Information) (Aletras and Stevenson, 2013) between all pairs of words (w_i, w_j) in a given topic:

$$NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-log(p(w_i, w_j))} \quad (5)$$

We estimate word probabilities using *wikitext-103-raw-v1* [9] (Merity et al., 2017) as our reference corpus, with a sliding window of 10.

Secondly, we report Mean Word Entropy (MWE) per topic as a measure of topic specificity representing the conditional entropy of a word type given its topic, namely $-\sum P_r(w_i|z)logP_r(w_i|z)$. There is no clear optimal value for specificity but overly specific topics will have few word types and a low conditional entropy (with a minimum value of 0); conversely, overly broad topics will exhibit high entropy (maximum log of the vocabulary size). Since Top2Vec does not directly output a distribution over words, we use the empirical unigram distribution for all documents assigned to a particular topic.

Thirdly, since it is possible for a topic model to duplicate the same coherent topic many times, we also need a measure of topic diversity. We report the proportion of words that are unique to one topic, p_{unique}, accordingly.

5 Results

Table 2 summarises our coherence, diversity, and specificity results. We can see that we achieve slightly improved coherence over Top2Vec, and that including the community detection step significantly reduces the topic specificity, as expected. The strong coherence scores after community detection indicate that topics are still coherent even when merged with their neighbours. This demonstrates that the edges in the graph connect topics which are indeed related. For a full list of topics in our graph, see Supplementary Material.

[9] https://huggingface.co/datasets/wikitext

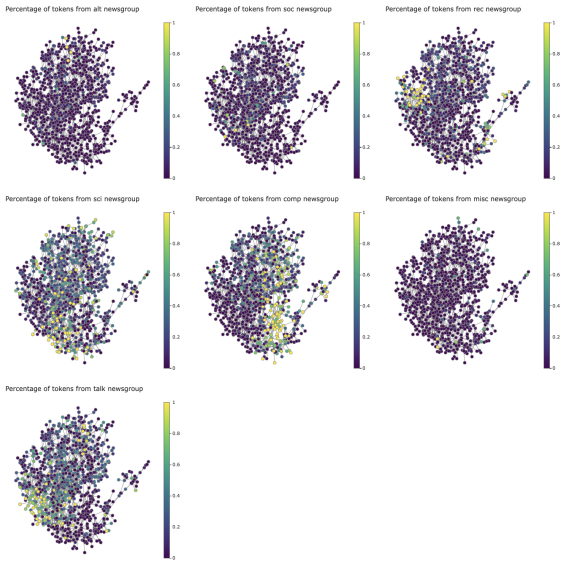

Figure 3: Percentage of tokens per newsgroup category.

5.1 Per-newsgroup Analysis

Figure 3 visualises the graph where each node is coloured by the percentage of its tokens that came from a given top-level newsgroup category. We observe that there are regions in the graph which correlate strongly with *rec*, *sci*, *comp*, and *talk*. At the same time, *misc*, *alt*, and *soc*, which are generally broad, are associated with some individual nodes without clear regions in the graph which may reflect the fact that these categories are the least frequent ones in our data set.

5.2 Part-of-Speech Effects

We run `spaCy` on the entire data set to assign part-of-speech tags to each token, revealing clear regions of the graph corresponding to VERB, NOUN, and ADJ tags (Figure 4). We do not plot other word classes since they are relatively infrequent in the data set (cf. filtering and pre-processing in Section 4). We make no claim as to whether the observed correlation with part-of-speech tags is beneficial since the exact definition of what constitutes a useful topic is highly task- and domain-dependent. However, our word class clusters could motivate the application of TDA to the recent field of "BERTology" to interpret emergent linguistic structure across transformer architectures (Rogers et al., 2020; Manning et al., 2020).

5.3 General Qualitative Observations

Table 3 illustrates sample topic clusters for which we provided a manual category label. The topics in our graph are generally coherent, and exhibit

Model	NPMI	MEW	p_{unique}	Number of Topics
Top2Vec	0.0002	6.99	0.822	126
Mapper + BERT	0.059	1.651	0.552	931
Mapper + BERT + Community Detection	0.038	2.796	0.844	149

Table 2: Evaluation results.

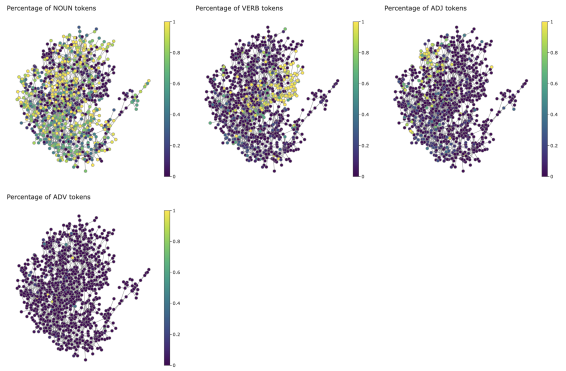

Figure 4: Percentage of tokens per word class.

Figure 5: Topic subgraph: Space.

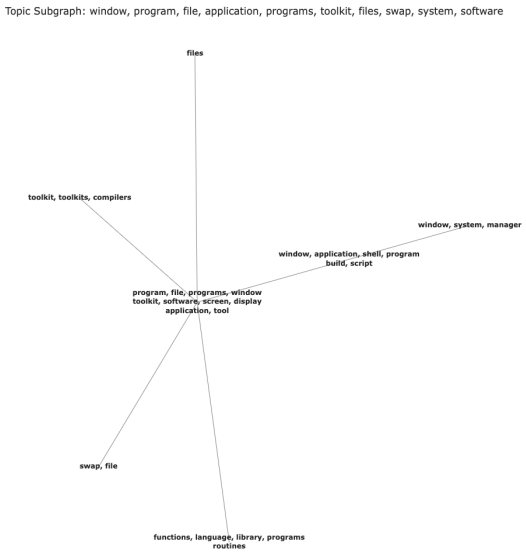

Figure 6: Topic subgraph: Computers.

appropriate middle-level specificity (not too coarse, not too fine). Our graph discovered unambiguous top-level newsgroup categories, as expected. For example, rows 0-6 represent vanilla topics relevant to computers, space, sports, and religion. A variety of subtler, more interesting clusters are noteworthy in that they capture a variety of broader, yet coherent lexical senses both para- and syntagmatically. Rows 7-10, for example, denote logic and argumentation, physical damage, law, possibility, and evidence. Some of the topics discovered border on word sense disambiguation which goes beyond typical, predominantly nominal topics (as subject headings). Consider (i) the clear and accurate sense-level distinctions in rows 12-15; (ii) *"program(s)"* qua computer software (row 1) vs. radio shows (row 24); and (iii) a non-trivial pattern involving clusters made of intra-sense antonyms subsumed under a relevant macrosense category (rows 18-20). Interestingly, we also see higher, discourse-level phenomena such as interjectional (and other) discourse markers and particles (row 21), and general, extralinguistic text structures (rows 22-23).

These patterns indicate that our method is sensitive enough to make non-trivial topic distinctions at multiple levels concurrently.

5.4 Topic Subgraphs

Topics extracted via community detection on the Mapper graph can be used to further probe and contextualise any individual topic by examining the subgraph to which it corresponds. Figures 7, 5, and

#	Category Name	Topic Words
0	computer software	window, program, file, application, programs, toolkit, files, swap, system, software
1	computer hardware	server, memory, drivers, hardware, system, binaries, disk, files, platforms, keyboard
2	data	image, images, fonts, line, data, support, value, text, lines, colors
3	planets	earth, mars, planet, planetary, jupiter, mercury, galaxy, pluto, venus, uranus
4	space	lunar, surface, earth, moon, space, mars, propulsion, planetary, orbit, astronomy
5	sports	rangers, bruins, wings, pens, leafs, cubs, devils, sox, flyers, hawks
6	religion	beliefs, teachings, doctrines, convictions, religions
7	physical damage	scratches, chips, cracks, cuts, crack
8	logic/argumentation	fallacy, ergo, post, hoc
9	law	court, legal, trial, lawyer, lawyers, supreme, legally, legalization, trials, attorney
10	possibility	chance, chances, opportunity, odds, probability, likelihood, possibility, possibilities
11	evidentiality/factuality	idea, evidence, obviously, based, test, opinion, opinions, apparently, research, advice
12	dependence	depends, depend, hinges, rests
13	memory	remember, recall, recalled
14	perception/copulas	looks, like, look, looked, looking, feels, sounded, appear
15	persuasion	convince, convinced, persuade
16	time periods	years, year, months, days, week, weeks, month, day, hours, time
17	temporal order	second, 2nd, 1st, secondly, coming, 3rd, fourth, firstly, 4th, later
18	public-private	private, public, privately
19	agreement-disagreement	agree, disagree, agreed, agreeing, agreement, agrees
20	substitution	alternative, alternatives, conventional, alternate, substitutes, traditional
21	discourse particles	yup, needless, oops, gosh, sheesh, darn, yea, geez, ahh, ditto
22	text/thread structure	question, list, questions, answer, response, reply, answers, respond, responses, replies
23	text structure	volume, page, vol, pages, ii, chapter, book, number
24	radio broadcasting	radio, coverage, broadcast, station, kdka, shown, program, announcer, shows, broadcasts

Table 3: Example topics with category names.

6 show subgraphs with manually assigned category labels for the Middle East conflict, space, and computers. For example, Figure 7 visualises various typical aspects and dimensions of the Middle East conflict such as people, locations, and ethnicity alongside concomitant historical, racial, religious, geopolitical, and military themes. Note that subgraphs only represent relatedness - we do attempt to capture interpretable relations of any other kind. These subgraphs demonstrate that neighbouring topics are related extrinsically beyond the underlying topology itself since their joint interpretation as higher-level topics is similarly coherent.

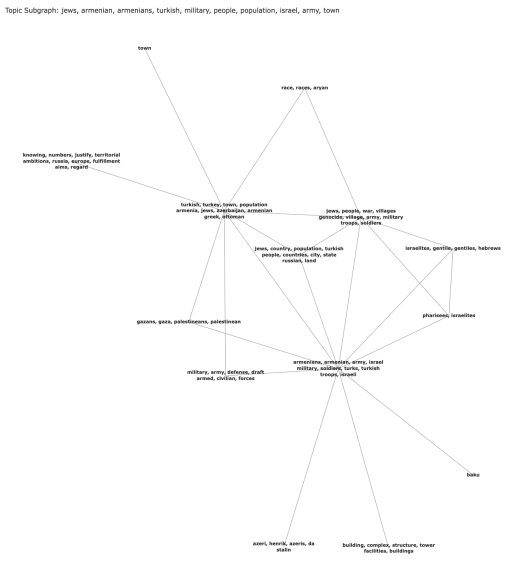

Figure 7: Topic subgraph: Middle East conflict.

6 Conclusion

We propose an unsupervised topic modelling method which leverages topological data analysis (TDA) to extract a semantic topic graph from a large unstructured document collection. Our experimental results demonstrate that our method is able to detect topics on par with a recent baseline while also exposing meaningful inter-topic relationships towards deeper topic interpretation. Our experiments to date motivate future work involving TDA to develop, for example, interactive visualisation tools for exploring rich relational topic graphs, and to study the interface between topological and linguistic properties of topics.

References

Amr Ahmed, Liangjie Hong, and Alexander Smola. 2013. Nested Chinese Restaurant Franchise Process: Applications to User Tracking and Document Modeling. In *30th International Conference on Machine Learning*, volume 28, pages 1426–1434. PMLR.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating Topic Coherence Using Distributional Semantics. In *10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22. Association for Computational Linguistics.

Dimo Angelov. 2020. Top2Vec: Distributed representations of topics. arXiv:2008.09470.

Francisco Belchí, Jacek Brodzki, Matthew Burfitt, and Mahesan Niranjan. 2020. A numerical measure of the instability of Mapper-type algorithms. *Journal of Machine Learning Research*, 21:1–45.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):1–30.

David M. Blei and John D. Lafferty. 2007. A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1):17–35.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172. Springer Berlin Heidelberg.

Frédéric Chazal and Bertrand Michel. 2021. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence*, 4.

Gábor Csárdi and Tamás Nepusz. 2006. The Igraph Software Package for Complex Network Research. *InterJournal*, Complex Systems:1695.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Zoubin Ghahramani, Michael Jordan, and Ryan P Adams. 2010. Tree-Structured Stick Breaking for Hierarchical Data. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. 2004. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

Jiao Gu, Bobo Hua, and Shiping Liu. 2015. Spectral distances on graphs. *Discrete Applied Mathematics*, 190–191:56–74.

Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. 2021. Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence. arXiv:2107.02173.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2020. Tree-Structured Neural Topic Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 800–806. Association for Computational Linguistics.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations*.

John Lafferty and David Blei. 2006. Correlated Topic Models. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Luyang Liu, Heyan Huang, Yang Gao, Yongfeng Zhang, and Xiaochi Wei. 2019. Neural Variational Correlated Topic Modeling. In *The World Wide Web Conference*, pages 1142–1152. ACM.

Ulrike von Luxburg. 2010. *Clustering Stability: An Overview*. now Publishers Inc.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer Sentinel Mixture Models. In *5th International Conference on Learning Representations*. OpenReview.net.

Elizabeth Munch and Bei Wang. 2016. Convergence between Categorical Representations of Reeb Space and Mapper. In *32nd International Symposium on Computational Geometry (SoCG 2016)*, volume 51, pages 53:1–53:16. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

John Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. 2014. Nested Hierarchical Dirichlet Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270.

Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106.

Archit Rathore, Nithin Chalapathi, Sourabh Palande, and Bei Wang. 2021. TopoAct: Visually Exploring the Shape of Activations in Deep Learning. *Computer Graphics Forum*, 40(1):382–397.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! In *2020 Conference on Empirical Methods in Natural Language Processing*, pages 1728–1736. Association for Computational Linguistics.

Ari Silburt, Anja Subasic, Evan Thompson, Carmeline Dsilva, and Tarec Fares. 2021. FANATIC: FAst Noise-Aware TopIc Clustering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 650–663. Association for Computational Linguistics.

Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. 2007. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *Eurographics Symposium on Point-Based Graphics*, pages 91–100. The Eurographics Association.

Akash Srivastava and Charles Sutton. 2017. Autoencoding Variational Inference for Topic Models. In *5th International Conference on Learning Representations*.

Laure Thompson and David Mimno. 2020. Topic Modeling with Contextualized Word Representation Clusters. arXiv:2010.12626.

Chong Wang and David Blei. 2009. Variational Inference for the Nested Chinese Restaurant Process. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Guangxu Xun, Yaliang Li, Wayne Xin Zhao, Jing Gao, and Aidong Zhang. 2017. A Correlated Topic Model Using Word Embeddings. In *Twenty-Sixth International Joint Conference on Artificial Intelligence, Main track*, pages 4207–4213.

Elias Zavitsanos, Georgios Paliouras, and George A. Vouros. 2011. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes. *Journal of Machine Learning Research*, 12(83):2749–2775.