# 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. 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; Grootendorst, 2022). 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.

## 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. Some studies have also incorporated contextualised word embeddings into topic models while still using neural probablistic generative models (Bianchi et al., 2020b,a; Hoyle et al., 2020).

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 in TDA is stability analysis (see (Belchí et al., 2020), (Lim and Yu, 2016), and (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 $\theta$ 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 $\theta$ 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 $\theta$ 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 $\theta$, 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 $\pi$ 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 $F_\mu$ and $F_\nu$ are CDFs for $\mu$ and $\nu$.*

The spectral distance quantifies the similarity of graph topologies amongst graphs (Gu et al., 2015). Lastly, let $\Theta$ be the search space for parameters

$\theta$: then the stable region of $\Theta$ with permissible parameter choices is

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

where $\varepsilon_m$ and $\varepsilon_s$ are thresholds for distances that are considered "large" and hence unstable.

## 4 Experiments

### 4.1 Data

We evaluated the proposed model on two text datasets: *20 Newsgroups* [1] and *AG News* [2]. Descriptions of these datasets are found in the Appendix. We extract contextualised subword embeddings using `bert-base-uncased`[3] (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`[4], and BERT subword tokens are aligned to `spaCy` tokens with `spacy-alignments`[5].

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 for *20 Newsgroups* and 12530 words for *AG News*. 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[6].

For clustering, we use HDBSCAN[7], 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.

### 4.3 Parameter Selection

Aside from the clustering and filter function, Mapper requires a "**cover**". We use the "balanced" cover offered by the `giotto-tda`[8] library - this simply partitions the space into hypercubes but adjusts their sizes so that each cover set contains a similar number of data points.

The cover requires two 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 artefacts. For the number of intervals we experiment with values in the range between 5 and 50 in steps of 5. For the percentage overlap we try values between 0.1 and 0.3 in increments of 0.05. We subdivide the datasets into 3 samples, each containing two thirds of the embeddings in the entire dataset. 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, 2, 3 and 4.

Looking at the regions that appear stable under both metrics, we are still left with multiple choices for stable parameters. We further eliminated sets of parameters that had too large a number of topic or nodes (because of a high bin size).

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

[2] Via `huggingface` https://huggingface.co/datasets/ag_news

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

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

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

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

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

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

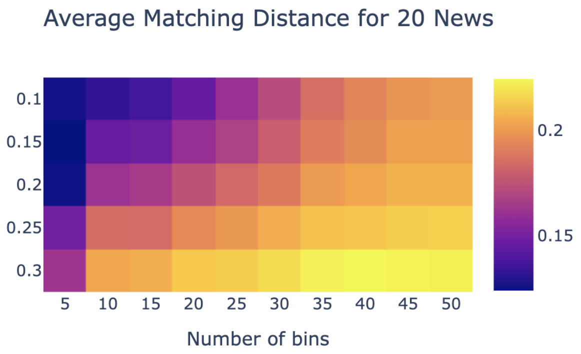

Figure 1: Matching Stability Scores for 20 news.

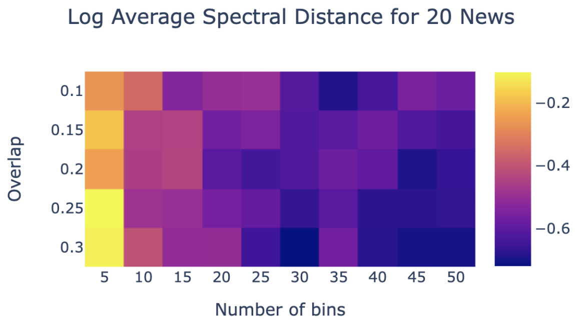

Figure 2: Spectral Stability Scores for 20 news.

We also ruled out some graphs which were highly connected and therefore had uninteresting structure. Ultimately this led us to choose a bin size of 20 for both datasets, and overlaps of 0.1 and 0.3 for 20News and AG News respectively.

### 4.4 Community Detection for Subgraphs

The resulting graphs both had one very large connected component as well as a large number of small components with only one or two nodes.These disconnected nodes contained about 30% of tokens in the 20 Newsgroups dataset and about 60% of the AG News tokens. Since these nodes are disconnected from the primary component of the topological manifold, we treat them essentially as noise and discard them from the rest of our analysis.

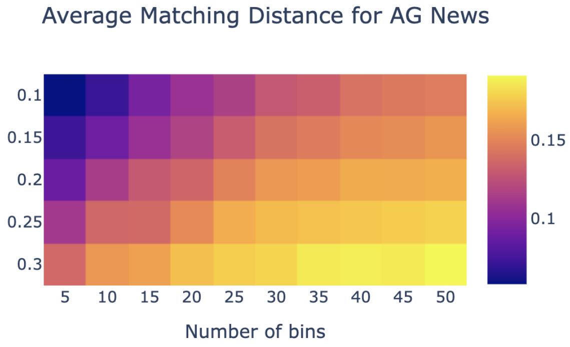

Figure 3: Matching Stability scores for AG News.

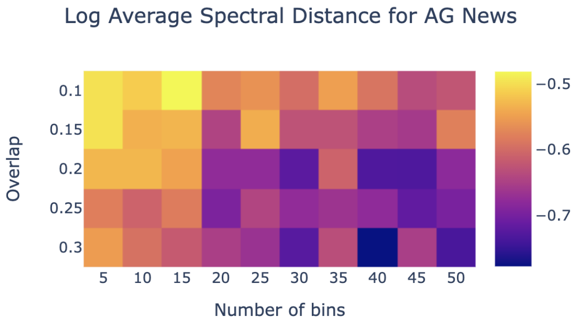

Figure 4: Spectral Stability scores for AG News.

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`[9] which is adapted to consider edge weights (Csárdi and Nepusz, 2006).

### 4.5 Baseline

We compare our work with two recent baselines. As a first 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. Secondly, we compare our methods to BERTopic (Grootendorst, 2022), using pretrained Sentence-BERT (SBERT) embeddings. For all other parameters we use the default settings in the BERTopic python reference implementation. [10]

### 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

[9] https://igraph.org/

[10] https://github.com/MaartenGr/BERTopic/tree/v0.8

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 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 in a given topic. We estimate word probabilities using *wikitext-103-raw-v1* [11] (Merity et al., 2017) as our reference corpus, with a sliding window of 10.

Secondly, we report Mean Word Entropy (MWE) (Thompson and Mimno, 2020) per topic as a measure of topic specificity representing the conditional entropy of a word type given its topic, namely $-\sum P_r(w_i|z) log P_r(w_i|z)$. 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 1 summarises our coherence, diversity, and specificity results. We can see that we achieve slightly improved coherence for 20 Newsgroups dataset although Top2Vec has slightly higher coherence scores on AG News. 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 graphs, see Supplementary Material.

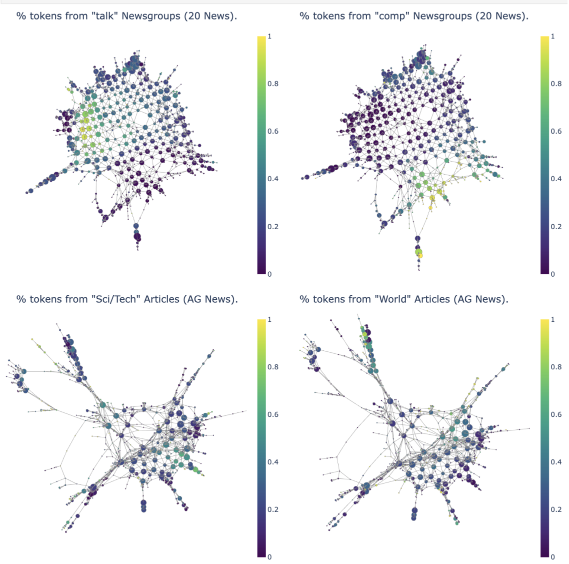

Figure 5: Percentage of tokens different labels.

### 5.1 Target Label Analysis

Both datasets have human topic annotations - we can use these to visualise which regions of the graph are associated with particular topics. To do this, we colour the nodes in the graph by the percentage of its tokens that come from a particular category of documents. Figure 5 show these plots for two categories from each dataset. We observe that there are regions in the graph which correlate with particular categories. The strength of the correlation varies depending on the category. For 20 News the effect is very strong for rec, sci, comp, and talk Newsgroups but weak or nonexistent for the misc, alt, and soc newsgroups. Likely this just reflects that these are much less frequent labels. For AG News, the effect appears to be weaker, meaning that our topic clusters are not as strongly related to the human labels. This is not necessarily a bad thing since the goal of topic modeling is find unsupervised topic classes. Plots for all categories can be found in the Supplementary Material.

### 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 7). We do not plot other word classes since they are relatively infrequent in the data set (cf. filtering and pre-processing in

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

| Dataset | Model | NPMI | MWE | $p_{unique}$ | Number of Topics |
|---|---|---|---|---|---|
| 20 NewsGroups | Top2Vec | 0.0002 | 6.99 | 0.822 | 126 |
| 20 NewsGroups | BERTopic | -0.008 | 2.21470 | 0.812 | 139 |
| 20 NewsGroups | Mapper + BERT | 0.059 | 1.651 | 0.552 | 931 |
| 20 NewsGroups | Mapper + BERT + Community Detection | 0.038 | 2.796 | 0.844 | 149 |
| AG News | Top2Vec | 0.0394 | 5.709 | 0.509 | 319 |
| AG News | BERTopic | -0.0419 | 2.179 | 0.705 | 648 |
| AG News | Mapper + BERT | 0.0372 | 1.300 | 0.547 | 939 |
| AG News | Mapper + BERT + Community Detection | 0.021 | 1.956 | 0.908 | 141 |

Table 1: Evaluation results.

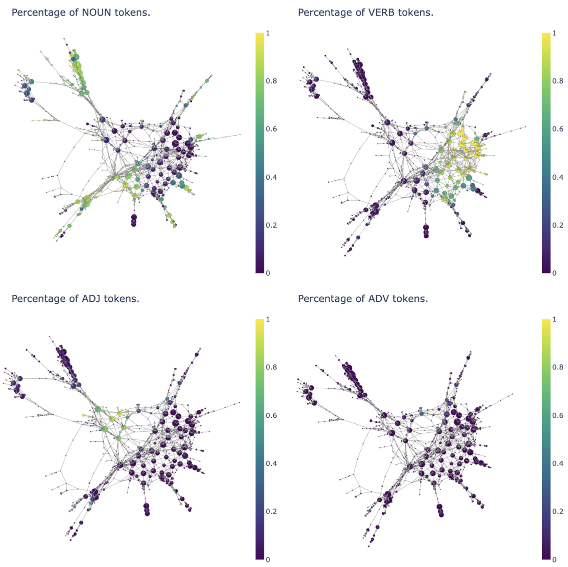

Figure 6: Percentage of tokens per word class for AG News Graph.

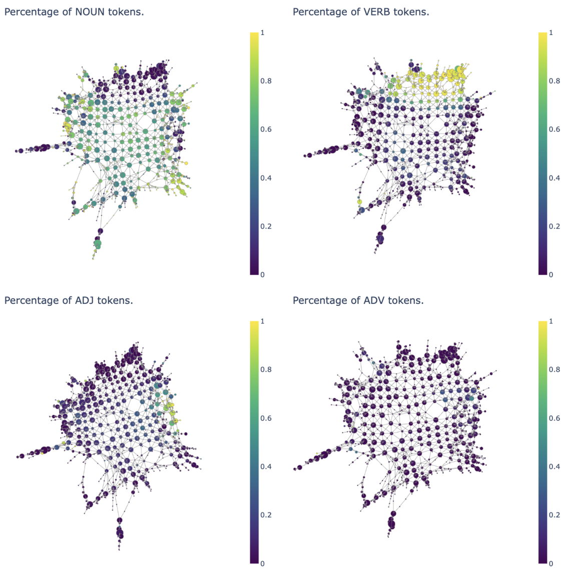

Figure 7: Percentage of tokens per word class for 20 Newsgroup Graph.

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

In this section we qualitatively evaluate the types of topics that can be extracted with our method. For brevity, we use examples only from the 20 News Groups dataset although similar phenomenon can be observed in the AG News topics which can be found in the Supplementary Materials. Table 2 illustrates sample topic clusters for which we provided a manual category label. The topics in our graph are generally coherent, and exhibit 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.

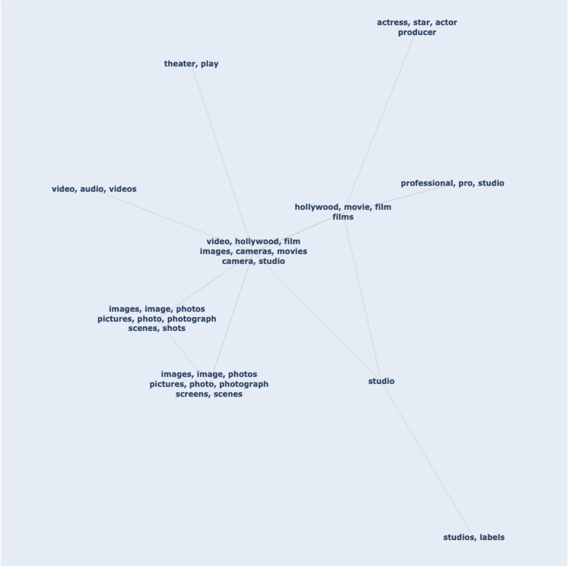

Figure 8: AG News Subgraph: Film Industry.

### 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 9 & 8 show a subgraph from each of the two datasets. For example, Figure 9 visualises aspects of the Middle East conflict as discussed in the 20 Newsgroups datasets - these include people, locations, and ethnicity as well as historical, racial, religious, geopolitical, and military themes. Figure 8 shows different topics pertaining to the film industry extracted from AG News Articles.

## 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.

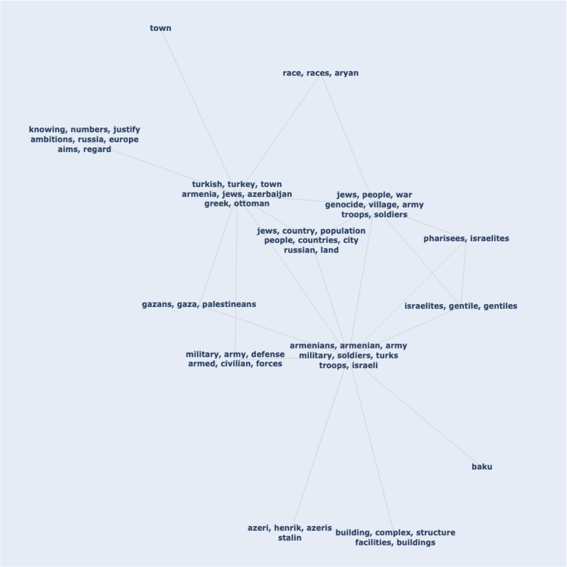

Figure 9: 20 Newsgroups Subgraph: Middle East conflict.

## 7 Limitations

Our method makes use of pretrained language models to extract contextualised word representations. Thus we can introduce biases from the pretraining dataset. Often these datasets, undergo little or no curation meaning the biases can be harmful or unwanted. See (Bender et al., 2021) for a discussion. This differs from traditional probablistic topic models which only depend on the dataset that is being explored.

Another limitation of our approach is the number of different hyperparmeters required. Our stability analysis approach does not uniquely determine them all and some heuristic selection was still necessary. Further analysis of the interaction between clustering, UMAP and cover parameters is an important direction for future work.

The connections in our graph represent the topology of the manifold of BERT embeddings. While we have demonstrated that these connections capture a general notion of "relatedness", we cannot necessarily interpret them as semantic relations. Further exploration of the graph's edges will be necessary in order to understand what types of interpretable relations can be captured.

## References

Amr Ahmed, Liangjie Hong, and Alexander Smola. 2013. Nested Chinese Restaurant Franchise Process: Applications to User Tracking and Document Mod-

| # | 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 2: Example topics from 20 News Groups with category names.

eling. 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.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2020b. Cross-lingual contextualized topic models with zero-shot learning. *arXiv preprint arXiv:2004.07737*.

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.

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.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

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.

Alexander Hoyle, Pranav Goel, and Philip Resnik. 2020. Improving neural topic models using knowledge distillation. *arXiv preprint arXiv:2010.02377*.

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.

Chinghway Lim and Bin Yu. 2016. Estimation stability with cross-validation (escv). *Journal of Computational and Graphical Statistics*, 25(2):464–492.

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.

# A Data

20News dataset contains 18846 English language posts categorised into thematic newsgroups. We use the standard train-test split. Table 3 summarises per-category document frequencies in the training set. We remove email addresses, headers, and subject lines.

The AG News dataset is constructed by assembling titles and description fields of news articles from four classes: “World”, “Sports”, “Business”, “Sci/Tech”. Since the dataset is large we randomly select 30000 articles resulting in the category frequencies in Table 4.

| 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 3: Summary of the 20 Newsgroups training set.

# B All Detected Topics

Tables 5, 6 and 7 show all topics from the 20 Newsgroup dataset and tables 8, **??** and **??** show all topics from the AG News dataset.

| AG News Category | # Documents |
|---|---|
| *Business* | 2477 |
| *Sci/Tech* | 2662 |
| *Sports* | 2338 |
| *World* | 2523 |

Table 4: Summary of the AG News training set.

# C Target Label Analysis

## C.1 20 News Target Label Graphs

Figures 10 - 16 show the regions of the graph associated with particular newsgroups. Figure 17 shows the entropy of the distribution of newsgroup tokens for particular nodes. This is used a measure of "diversity" - nodes with high entropy will have tokens that come uniformly from all newsgroup categories.

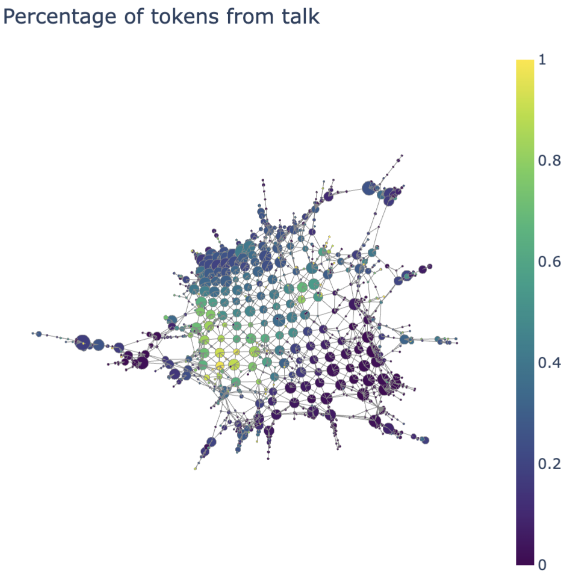

Figure 10: Percentage of tokens from talk newsgroup.

## C.2 AG News Target Label Graphs

Figures 18 - 21 show the regions of the graph associated with particular news categories. Figure **??** shows the entropy of the distribution of target labels.

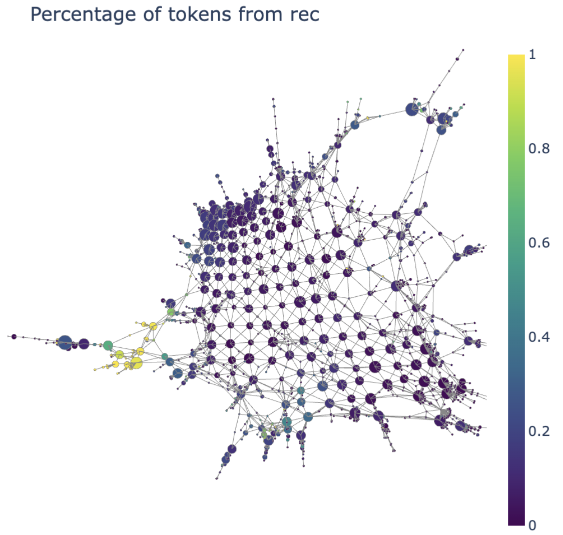

Figure 11: Percentage of tokens from rec newsgroup.

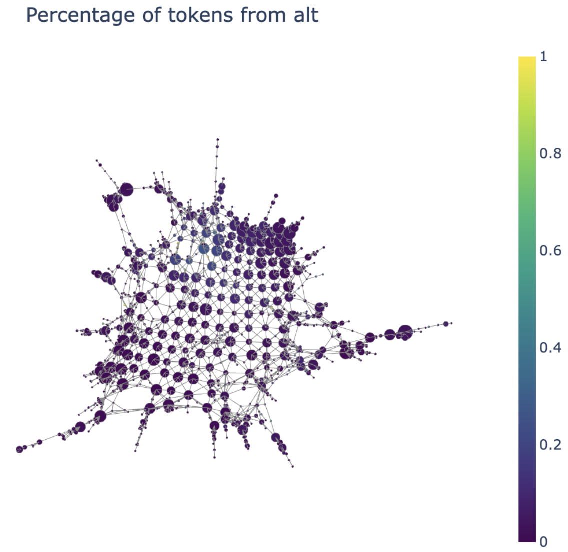

Figure 12: Percentage of tokens from alt newsgroup.

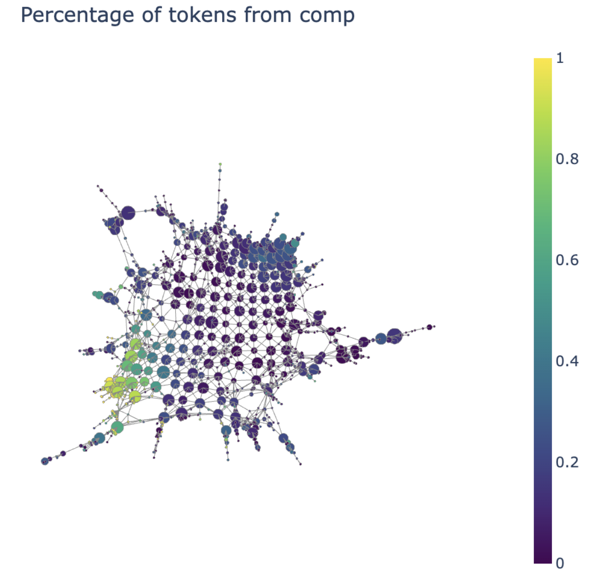

Figure 13: Percentage of tokens from comp newsgroup.

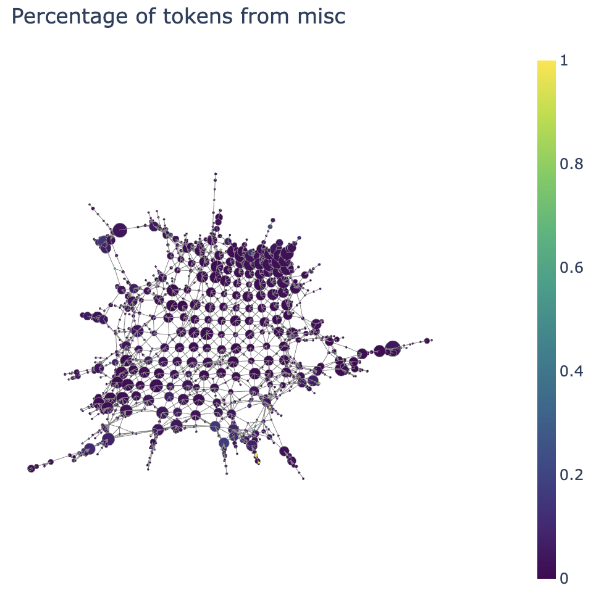

Figure 14: Percentage of tokens from misc newsgroup.

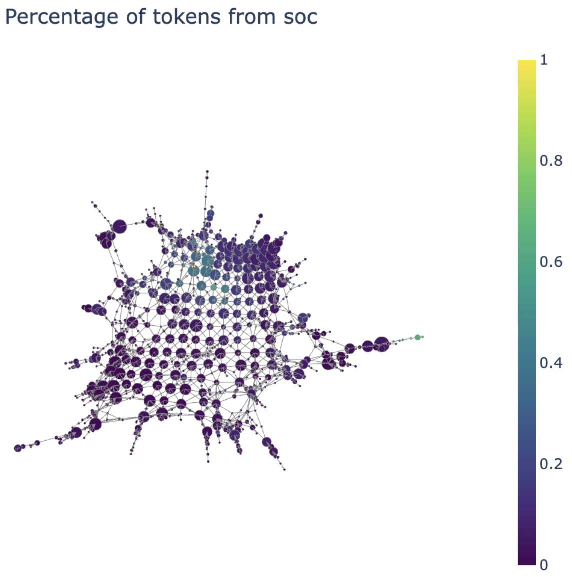

Figure 15: Percentage of tokens from soc newsgroup.

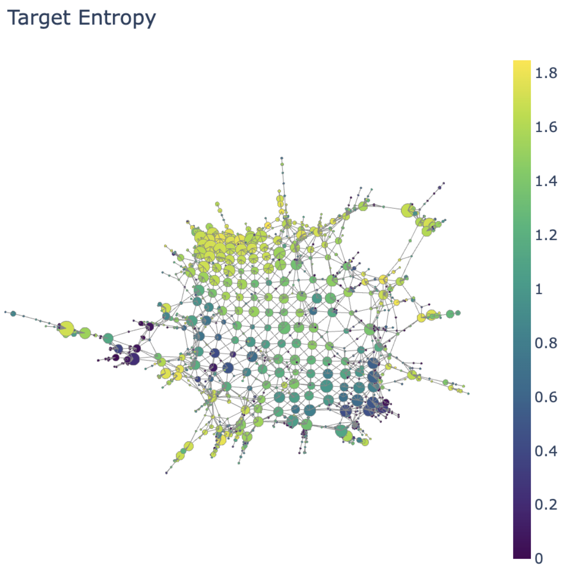

Figure 17: Entropy of newsgroup distribution in cluster.

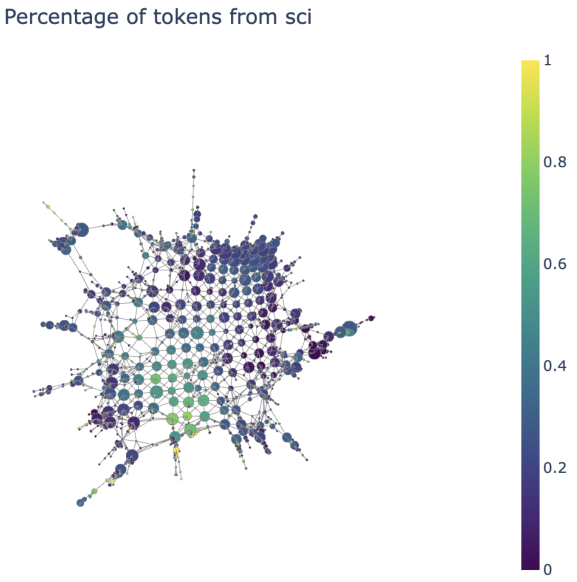

Figure 16: Percentage of tokens from sci newsgroup.

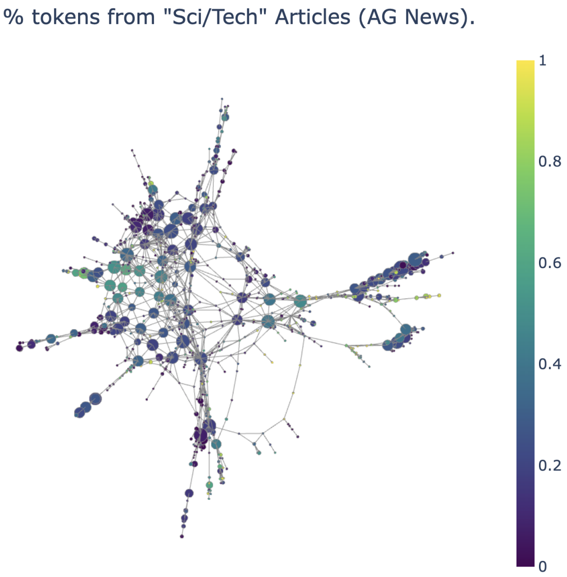

Figure 18: Percentage of tokens from Sci/Tech articles.

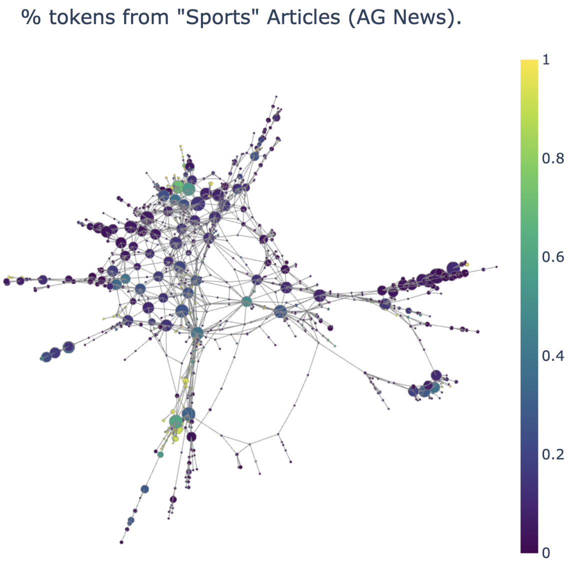

Figure 19: Percentage of tokens from Sports articles.

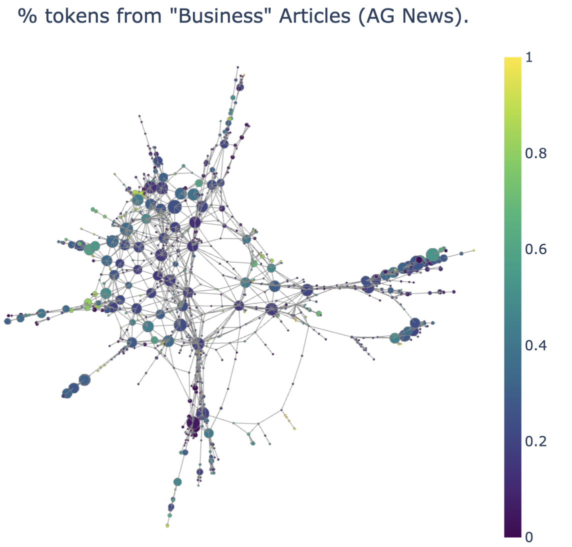

Figure 21: Percentage of tokens from Business articles.

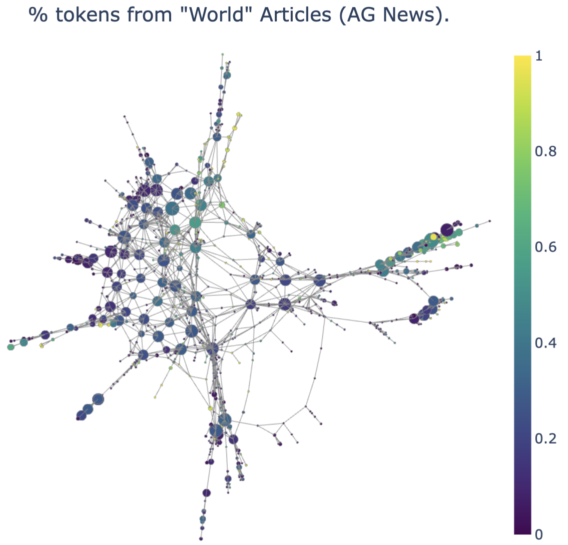

Figure 20: Percentage of tokens from World articles.

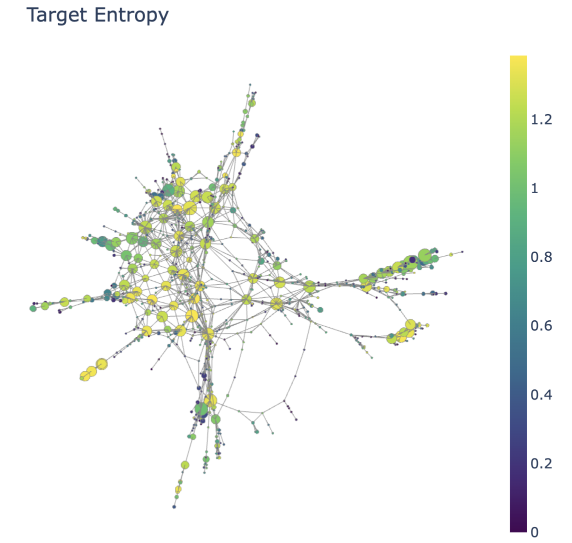

Figure 22: Entropy of article category distribution in cluster.

| | |
|---|---|
| 1 | witnesses, testimony, witness, testify |
| 2 | dr., j., a., c., m., r., s., d., l., e. |
| 3 | air, force, base, command |
| 4 | z, z. |
| 5 | earth, mars, planet, planetary, jupiter, mercury, galaxy, pluto, venus, uranus |
| 6 | transactions, transaction, payments |
| 7 | remember, recall, recalled |
| 8 | option, options, choices |
| 9 | flight, aircraft, aviation, planes, plane, airplane, aerospace, pilots, pilot, airplanes |
| 10 | lunar, surface, earth, moon, space, mars, propulsion, planetary, orbit, astronomy |
| 11 | team, hockey, season, league, year, teams, nhl, playoffs, division, cup |
| 12 | vision, sight |
| 13 | private, public, privately |
| 14 | medicine, drug, drugs, medical, treatment, treat, imaging, treating, cure, therapy |
| 15 | arena, facility, gm |
| 16 | copy, copies, duplicate |
| 17 | question, list, questions, answer, response, reply, answers, respond, responses, replies |
| 18 | space, spaces, room |
| 19 | muhammad, prophet, saw, mohammed, prophets, mohammad |
| 20 | convince, convinced, persuade |
| 21 | san, los, jose, angeles, bay, tampa, baltimore, boston, detroit, milwaukee |
| 22 | young, people |
| 23 | physics, chemistry, mechanics, quantum, chemist, chemists, mathematics, elementary, problems, energy |
| 24 | gun, guns, weapons, firearms, weapon, arms, bear, semi, automatic, rocket |
| 25 | power, supply, energy, electric, electricity, supplies, powered, source, fossil, charge |
| 26 | billboard, sign, billboards, signs |
| 27 | level, grade |
| 28 | said, need, tell, says, thought, like, saying, told, know, understand |
| 29 | help, assist |
| 30 | beliefs, teachings, doctrines, convictions, religions |
| 31 | volume, page, vol, pages, ii, chapter, book, number |
| 32 | m, km |
| 33 | system, computer, phone, systems, pc, device, technology, devices, phones, unit |
| 34 | court, legal, trial, lawyer, lawyers, supreme, legally, legalization, trials, attorney |
| 35 | jews, armenian, armenians, turkish, military, people, population, israel, army, town |
| 36 | radio, coverage, broadcast, station, kdka, shown, program, announcer, shows, broadcasts |
| 37 | looks, like, look, looked, looking, feels, sounded, appear |
| 38 | david, john, robert, jim, mike, steve, michael, dave, jon, regards |
| 39 | god, control, bible, life, law, christ, lord, church, power, jesus |
| 40 | effective, clever |
| 41 | reactor, plants, plant, reactors, pile, facilities, stations, station |
| 42 | archive, archives, directory |
| 43 | order, ordered, orders, ordering, prepare, national |
| 44 | new, california, york, washington, detroit, city, san, pittsburgh, germany, chicago |
| 45 | got, happened, happen, finally, started, spend, came, going, happy, happening |
| 46 | want, like, wanted, wants, wish, need, prefer, enjoy, love, liked |
| 47 | history, bill, package, tax, meeting, health, stimulus, money, funds, care |
| 48 | killed, jesus, women, dead, children, people, death, body, woman, family |
| 49 | set, model, version, size, algorithm, parts, design, manual, models, manuals |
| 50 | home, rest, team, average, defense, game, games, flyers, players, hand |

Table 5: 20 Newsgroup Topics (1-50).

| | |
|---|---|
| 51 | fuel, motors, fossil |
| 52 | agree, disagree, agreed, agreeing, agreement, agrees |
| 53 | ask, asked, forget, talking, print, appears, feel, asking, remember, wrote |
| 54 | years, year, months, days, week, weeks, month, day, hours, time |
| 55 | games, programs, titles, players, arcade |
| 56 | things, bad, human, humans, evil, beings, mankind, humanity, morals, humankind |
| 57 | info, section, sections |
| 58 | know, believe, little, mean, means, bit, knows, posted, sure, meant |
| 59 | public, key, private, secret, shared |
| 60 | buy, sell, bought, shipping, buying, selling, sold, ride, riding, purchase |
| 61 | play, win, children, women, wife, playing, played, doctor, second, son |
| 62 | chance, chances, opportunity, odds, probability, likelihood, possibility, possibilities |
| 63 | life, disease, pain, right, syndrome, lie, risk, lives, eternal, physical |
| 64 | game, games, health, defense, play, goal, puck, win, stats, period |
| 65 | avoid, protect, help, making, continue, cause, prevent, increase, stop, support |
| 66 | country, government, area, state, south, vote, community, russia, leaders, island |
| 67 | good, great, simple, better, big, similar, excellent, interesting, free, results |
| 68 | entry, encryption, information, send, message, system, data, access, privacy, containing |
| 69 | program, future, non, programs, conference, project, held, insurance, budget, license |
| 70 | speed, code, support, rate, programs, performance, technical, rates, resolution, capability |
| 71 | rangers, bruins, wings, pens, leafs, cubs, devils, sox, flyers, hawks |
| 72 | tried, turn, carry, removed, taking, break, stop, getting, save, remain |
| 73 | find, read, looking, look, run, found, try, check, reading, exist |
| 74 | went, live, came, going, away, took, come, living, gone, lived |
| 75 | day, later, half, year, night, police, morning, minutes, citizens, weekend |
| 76 | point, effect, stupid, theory, possible, completely, necessary, effects, correct, dangerous |
| 77 | manufacturers, manufacturer, store, shop, sales, catalog, stores, vendors, factory, makers |
| 78 | company, companies, businesses, corporations, manufacturers, manufactures, firms, department, maker, makers |
| 79 | use, change, changed, designed, build, add, support, considered, need, directly |
| 80 | second, 2nd, 1st, secondly, coming, 3rd, fourth, firstly, 4th, later |
| 81 | information, info, details, specifics, additional, contributions, background, complete, detailed, application |
| 82 | purpose, evidence, probably, actions, lack, goal, related, possibility, action, true |
| 83 | address, sound, bios, noise, rom, controller, speaker, system, speed, stereo |
| 84 | right, rights, money, difference, economic, political, dollars, morality, nuclear, differences |
| 85 | man, men, male, female, males, fellow, gentlemen, gentleman |
| 86 | available, number, standard, access, level, included, text, section, letter, standards |
| 87 | problem, study, good, information, story, meaning, entire, better, report, approach |
| 88 | believe, makes, includes, uses, think, expect, consider, suggest, talk, explain |
| 89 | involved, nature, power, attempt, relationship, law, presence, action, faith, effort |
| 90 | seen, heard, running, come, having, saw, getting, start, called, occurs |
| 91 | time, government, point, times, period, early, century, beginning, hot, cold |
| 92 | group, government, groups, news, public, organization, place, yes, service, area |
| 93 | new, situation, cases, different, rules, final, future, secret, situations, entries |
| 94 | outside, inside, near, close, good, closer, excellent, missing, fair, past |
| 95 | command, commands, shell, line, controls, result, instructions |
| 96 | image, images, fonts, line, data, support, value, text, lines, colors |
| 97 | important, common, strong, limited, little, possible, value, main, step, major |
| 98 | idea, evidence, obviously, based, test, opinion, opinions, apparently, research, advice |
| 99 | war, world, ii, wwi, ww2, ww, battle, combat, campaign, defense |
| 100 | window, program, file, application, programs, toolkit, files, swap, system, software |

Table 6: 20 Newsgroup Topics (51-100).

| | |
|---|---|
| 101 | president, old, end, previous, administration, older, house, early, earlier, prior |
| 102 | book, small, article, better, high, large, low, long, books, extra |
| 103 | list, article, posting, launch, information, space, use, read, post, rules |
| 104 | x, widget, windows, motif, bit, hard, mac, drives, disk, pc |
| 105 | people, militia, person, war, tobacco, use, americans, military, today, users |
| 106 | key, built, fonts, bit, based, chip, bits, keys, version, number |
| 107 | package, tools, tool, kit, utility, facility |
| 108 | people, vat, believe, food, christian, atheists, law, life, religious, world |
| 109 | kit, family, include, software, scientific, association, spectrum, functions, moscow, set |
| 110 | argument, job, work, statement, discussion, upgrade, choice, clear, position, claim |
| 111 | server, memory, drivers, hardware, system, binaries, disk, files, platforms, keyboard |
| 112 | like, road, surrender, answer, roads, unlike, street, highway, traffic, film |
| 113 | alternative, alternatives, conventional, alternate, substitutes, traditional |
| 114 | best, april, original, clipper, btw, february, clinton, june, march, george |
| 115 | nist, comp.sources.misc |
| 116 | available, version, algorithm, runs, attack, written, cryptography, found, included, cipher |
| 117 | accept, recognize, reject, interpret, ignore, comprehend, embrace, understand, acknowledge, accepted |
| 118 | cable, wire, wires, tube, plug, filter, panel, cables, eff, chain |
| 119 | radio, stereo, pub, antenna, receiver, amateur, transmitter, receivers, series, microphone |
| 120 | end, profile |
| 121 | anonymous, x, usenet, archive, available, newsgroup, space, sites, file, ground |
| 122 | including, especially, general, addition, modern, furthermore, fact, particularly, initial, junk |
| 123 | inference, conclusion, t, valid, premises, true, proposition, arrived, basis, phrases |
| 124 | colormap, bitmap, defaults, binaries, truecolor, tasking, app, multitasking, application, hardcopy |
| 125 | use, work, apply, mentioned, compare, working, fit, applies, vary, rely |
| 126 | number, line, numbers, set, lines, names, wiretap, position, processing, sets |
| 127 | therapies, allergies, allergy, endometriosis, recurrence, recurrent, incurable |
| 128 | box, miles, case, tv, installed, mileage, drive, imho, driving, install |
| 129 | date, dates, time, stamp, memory, rec, times |
| 130 | workstation, workstations, toolkit, toolkits, assembler, menus, emulator, defaults, emulation, emulators |
| 131 | fallacy, ergo, post, hoc |
| 132 | scratches, chips, cracks, cuts, crack |
| 133 | yup, needless, oops, gosh, sheesh, darn, yea, geez, ahh, ditto |
| 134 | connect, connected, hook, attach, mount, link, hooking, mounted, mounting, interface |
| 135 | x, p, s, char, return, file, o, 0.0, break, case |
| 136 | manager, package, kit, packages, managers, viewer, module, kits, bundle, launcher |
| 137 | x, ftp, single, pub, scsi, x11, motif, contrib, drive, xt |
| 138 | assuming, assume, suppose, provided, guessing, providing, imagine |
| 139 | faq, newsgroup, double, newsgroups, connection, cycle, logo, compuserve, faqs, nist |
| 140 | depends, depend, hinges, rests |
| 141 | x, source, file, char, int, inc., bbs, adapter, sources, output |
| 142 | plots, charts |

Table 7: 20 Newsgroup Topics (101-142).

| | |
|---|---|
| 0 | new, update, nhl, olympics, pakistan, report, nasa, court, red, american |
| 1 | 39;s, 39;t, 39;re, 39;ve, 39;ll, 39;m, 146;s, 39;d, 39;a, 39;06 |
| 2 | says, wins, shows, sees, warns, finds, calls, reports, leads, expect |
| 3 | big, key, strong, major, good, small, controversial, main, senior, best |
| 4 | region, local, fans, private, regional, building, center, site, commercial, hotel |
| 5 | people, workers, experts, unit, groups, leader, drug, employees, staff, plant |
| 6 | company, group, price, firm, officials, companies, states, firms, forces, half |
| 7 | return, work, play, face, find, try, discuss, developed, look, facing |
| 8 | hopes, investigation, efforts, forecast, claims, concerns, way, probe, fears, battle |
| 9 | higher, high, lower, strong, damage, growing, crashed, rising, heavy, wreckage |
| 10 | expected, set, nearly, alleged, future, suspected, likely, upcoming, hit, apparently |
| 11 | plans, rise, drop, fall, plan, rises, buys, higher, wins, decline |
| 12 | sell, software, buy, products, sale, equipment, sold, heart, selling, devices |
| 13 | left, helped, ended, leaving, led, end, leave, raised, caused, boosted |
| 14 | said, announced, warned, found, released, reported, called, told, unveiled, visit |
| 15 | record, costs, cost, orders, high, fastest, records, fees, breaking, time |
| 16 | help, use, continue, boost, face, improve, run, allow, build, save |
| 17 | near, closer, close, nearing, nearer, approaching, reaching, nearly, nears, halfway |
| 18 | fell, rose, dropped, surged, climbed, edged, jumped, declined, grew, slowed |
| 19 | -wsj, nok, wtc, wsj, doj, vna, ws, kvs, msft, aapl |
| 20 | nortel, banknorth, novell, schwab, citigroup, bomb, amp;t, wpp, scientists, symantec |
| 21 | funds, spending, money, fund, spend, finances, spent, dollars, consumption, raising |
| 22 | report, final, attack, data, attacks, number, study, time, information, reports |
| 23 | little, bit, touch |
| 24 | jobless, job, productivity, layoffs, unemployment, employment |
| 25 | government, minister, president, prime, ministers, state, ministry, leader, cabinet, general |
| 26 | water, air, supplies, production, supply, output, pool, sea, coast, aircraft |
| 27 | highly, eagerly, widely, hotly, high |
| 28 | old, elderly, aging, older, original, frail, seniors, younger, aged |
| 29 | agreed, won, win, beat, winning, vote, signed, filed, wants, reached |
| 30 | system, systems, vote, standards, rules, law, ruling, decision, president, rule |
| 31 | trying, ready, hoping, planning, poised, preparing, seeking, discuss, looking, considering |
| 32 | early, previous, late, earlier, previously, later, mid, initial, originally, initially |
| 33 | pay, payment, paid, paying, cover, charge, payout, account, satisfy, fully |
| 34 | events, event, crisis, stage, drama, occurrences, accident, incidents, incident, scenes |
| 35 | moment, veterans, veteran, image, retired, moments, guru, hero, heroes, credibility |
| 36 | applications, application, apps, app, clients, service |
| 37 | campaign, candidate, race, candidates, campaigns, campaigning, nominee, challenger, nomination, rival |
| 38 | signs, jewelry, lights, directions, instructions, lighting, guidance |
| 39 | lawsuit, case, suit, lawsuits, appeal, claim, proceedings, cases, litigation, suits |
| 40 | people, residents, person, individuals, individual, persons, everybody, ones, somebody |
| 41 | talks, deal, contract, agreement, negotiations, merger, solution, pact, deals, dialogue |
| 42 | performance, level, value, levels, benchmark, fate, ceiling, fortunes, legacy, showing |
| 43 | video, hollywood, images, movie, film, image, studio, movies, cameras, pictures |
| 44 | calls, message, messages, letter, messaging, calling, book, books, writers, dial |
| 45 | site, sites, web, website, blog, blogs, pages, page, portal, websites |
| 46 | service, services, hosted, portal, connect |
| 47 | multiple, different, cheap |
| 48 | national, nationwide, nationally, statewide, wide |
| 49 | health, surgery, hospital, care, medical, doctors, hospitals, pharmacy, bypass, doctor |
| 50 | changes, reforms, slowdown, reform, change, pullback, bounce, revisions, swing, adjustments |

Table 8: AG News Topics (1-50).