Principled Analysis of Energy Discourse across Domains with Thesaurus-based Automatic Topic Labeling

Thomas Scelsi[1] **Alfonso Martínez Arranz**[2] **Lea Frermann**[1]
[1] School of Computing and Information Systems, The University of Melbourne
[2] Department of Chemical Engineering, The University of Melbourne
tscelsi@student.unimelb.edu.au {alfonso.arranz,lfrermann}@unimelb.edu.au

Abstract

With the increasing impact of Natural Language Processing tools like topic models in social science research, the experimental rigor and comparability of models and datasets has come under scrutiny. Especially when contributing to research on topics with worldwide impacts like energy policy, objective analyses and reliable datasets are necessary. We contribute toward this goal in two ways: first, we release two diachronic corpora covering 23 years of energy discussions in the U.S. Energy Information Administration. Secondly, we propose a simple method for automatic topic labelling drawing on domain knowledge via political thesauri. We empirically evaluate the quality of our labels, and apply our labelling to topics induced by diachronic topic models on our energy corpora, and present a detailed analysis.

1 Introduction

Policy-making in highly technical areas such as energy is deemed to require neutral input from specialised government agencies, e.g. the US's Energy Information Administration (EIA) or the International Energy Agency (IEA). Their publications are known as"grey literature", given that they are expertly produced but not peer-reviewed. In contrast to academic literature, grey literature is often freely available online and aims to be more accessible to lay readers.

Energy grey literature has been shown to display biases towards incumbent fossil fuel technologies (Martínez Arranz, 2016; Mohn, 2020), but no thorough exploration exists of the disconnect between grey and academic literature in their coverage of energy issues. In this paper, we provide a reproducible and automated assessment of topics across both literatures.

We analyse and make available two diachronic corpora derived covering 23 years of energy discussions from two EIA publications: the International Energy Outlook (IEO) and the more nationally focused Annual Energy Outlook (AEO). Parsing reports of government agencies is non-trivial due to their diverse layouts (changing over time), and text frequently disrupted with tables and graphs. We release a clean text corpus as a basis for future research on energy communication. We also make available a software tool-set for the creation and analysis of these corpora, easily generalisable to diachronic datasets in general.

We analyse the corpora using dynamic topic models (Blei and Lafferty, 2006). We compare their discussion of various facets of energy politics against the discourse in the scientific community over the same period, drawing on a corpus of abstracts of articles published in established energy journals.

By releasing our grey literature corpus, we address the reproducibility challenge of large-scale text analyses with unsupervised models like topic models in the social sciences (Müller-Hansen et al., 2020). Studies are often difficult to reproduce, compare against or build upon due to a lack of public resources, as well as ad-hoc subjective choices of the researchers. We address the latter problem by proposing a conceptually simple and theoretically sound method for automatic topic labelling, drawing from thesauri over the political domain. In summary, our contributions are:

- A diachronic dataset of grey literature from the EIA, supporting research into (a) the discussion of energy policies and technologies over time; and (b) the discussion of energy policies across outlets. We release scripts to reproduce our data set at at: `https://github.com/tscelsi/eia-diachronic-analysis`
- A topic labelling framework for policy is-

sues, leveraging the publicly available and exhaustive EuroVoc thesaurus. Publicly available at https://github.com/tscelsi/dtm-toolkit

- A comparison of the dynamics of energy-related topics within and between grey and academic literature with a focus on electricity generation technologies, sustainability, and geopolitics & economy.

2 Background

Topic modeling Topic models are statistical models which aim to uncover latent semantic structure of texts through a small set of distinctive topics, each of which is represented through a coherent set of words. Latent Dirichlet Allocation (LDA; Blei et al. (2003)) is probably the most widely used topic model, where each document d is modelled as a mixture of topics k, $p(k|d)$, and each topic is represented as a probability distribution over words w, $p(k|w)$. LDA is an *exchangeable* model, it is agnostic about the order of words and documents. Applications of topic models to social science questions (including our own), however, have a specific interest in the temporal development of topics.

The dynamic topic model (DTM; Blei and Lafferty (2006)) extends LDA to time series data, capturing the subtle changes of the *same* topic over time, with the intuition that over extended periods of time, discussions, themes and words surrounding the same topic change. The DTM accounts for this evolution by inducing topic proportions as well as topic representations sensitive to time t as $p(k|d,t)$ and $p(w|k,t)$, respectively. The time-specific parameters are tied through a random walk process in order to encourage a smooth change from time t to $t+1$.

Topic models and their dynamic counterparts have been extended to leverage the power of deep learning (Card et al., 2017; Dieng et al., 2020, 2019) leading to a better data fit at the cost of substantial increase in compute cost and technical expertise. In this work we will leverage the DTM as introduced above to explore the discussion of energy technologies in scientific and government publications over the past 30 years.

Topic labelling Topics as induced by the DTM are probability distributions over the vocabulary. While they are often visualized through the top N words with highest probability, a principled interpretation of their content remains a challenge. Various methods have been proposed for labelling a topic, ranging from single best word selection (Lau et al., 2010) over involved methods leveraging domain-general external resources like Wikipedia (Lau et al., 2011; Bhatia et al., 2016) by retrieving relevant phrases, which requires substantial IR and NLP overhead to process the potential label inventory. Other work has employed graph-based methods over the structured DBPedia (Hulpus et al., 2013) , generated candidate labels using WordNet (Poostchi and Piccardi, 2018) or created descriptive labels as text summaries by extracting and aggregating candidate sentences from documents highly relevant to a topic (Wan and Wang, 2016). We propose a simpler solution by leveraging structured, and broadly domain-relevant Thesauri as our label inventory. Specifically, we use the EuroVoc thesaurus, compiled by the European Union which covers all areas of European Parliament discussion (including energy policy), noting that our methods extend to any thesaurus. We propose methods for filtering the resource to a focused set of labels (§ 4.1); and mapping induced topics to one or more thesaurus labels (§ 4.3).

2.1 The EuroVoc Thesaurus

EuroVoc[1] is a multilingual thesaurus (Steinberger et al., 2014), originally developed as a framework to support hierarchical search and indexing of documents produced in the European Union (EU). It covers a wide range of political terminology, and consists of 127 general "topics", each associated with a list of phrases (cf., Table 1). EuroVoc has been used in the NLP community predominantly in the context of multi-label classification (Steinberger et al., 2012) and as a multi-lingual lexical resource (Fišer and Sagot, 2008). To the best of our knowledge, this paper is the first to leverage this openly available, expert-created resource for principled labelling of automatically learnt topics. We use the English EuroVoc in this work, leaving multi-lingual topic labelling for future work.

3 Data

3.1 Grey Literature: The EIA Corpus

We focus on publicly available documents by the US Energy Information Administration (EIA). The EIA is the longest extant energy agency and the US is the world's second energy consumer and

[1] https://op.europa.eu/s/sCG9

Renewable Energy: bioenergy, biogas, geothermal energy, marine energy, renewable energy, soft energy, solar energy, wind energy
Prices: reduced price, price index, price reduction, farm prices, world market price, target price, producer price, price list, price increase
Environmental Policy: nature reserve, waste recycling, industrial hazard, environmental tax, emission allowance, environmental impact

Table 1: Selected EuroVoc labels (bold) and some of their associated keyphrases.

producer of energy, being recently overtaken by China. Its *Annual Energy Outlook* (AEO) and *International Energy Outlook* (IEO) are mandated to provide US citizens and lawmakers with future-oriented evaluations of, respectively, domestic and international energy trends.[2]

We obtained all IEO and AEO releases between 1997–2020.[3] Python package `pdfminer`[4] was used to convert the PDFs to text, adjusting parser parameters to ensure adequate parsing of single- and multi-column documents.

3.2 Scientific Literature: The Journals corpus

In order to obtain a reliable corpus of academic literature to contrast against EIA publications, we select two top-ranked energy policy journals in the Scimago Journal rankings in both 1999 and 2019: *Applied Energy* and *Energy Policy*. Both are open to submissions on any technology, and deal with policy and applied engineering questions that should be closest to the concerns of the EIA. Through the Scopus Search API complete view,[5] we download all article abstracts published in these two journals for the period 1997-2020. The format is already machine-readable and contains metadata on publication date requiring only minimal data cleaning.

We assume that abstracts synthesize the main points of each paper succinctly. Future research could include analysing the entire textual content of the papers.

3.3 Corpus Analysis

We automatically split each parsed EIA report into header-paragraph pairs, which are then used as documents to train our topic models. Paragraphs and headers were identified based on font size. As mentioned previously, for the Journals corpus we focus only on the abstract paragraphs of each paper and use these as documents to train our Journals topic models. We tokenize all corpora using spaCy.

Table 2 lists various statistics for our three corpora. Given that we train topic models over documents corresponding to paragraphs in the grey literature, we verify that paragraphs are of sufficient length to support topic modelling. From the average sentence per paragraph aggregations we can see that the EIA paragraphs are longer than the Journals paragraphs, however, the Journals corpus is significantly larger than the EIA corpora.

4 Topic Labeling with Thesauri

In this section we propose a new way of assigning automatic labels to DTM topics. Even though a variety of methods for automatic topic labeling exist (Lau et al., 2010, 2011; Sorodoc et al., 2017; Hulpus et al., 2013), case studies in the social sciences have largely resorted to qualitative analysis and manual labeling (Martínez Arranz, 2015; Müller-Hansen et al., 2020), resulting in a bottleneck for analysis as well as the potential for introduction of human bias.

We introduce a general and conceptually simple method, drawing on established domain-specific thesauri as a label inventory, and propose two methods for mapping topics to a small set of labels that reflect its content. We use the EuroVoc thesaurus in our study (§ 2.1), however, our method generalizes to any domain-specific thesaurus which organizes related keyphrases into succinct labels. Formally, we describe the set of EuroVoc labels as L. Each label $l \in L$ represents a set of keyphrases[6] v in the EuroVoc thesaurus that fall under that label (Table 1).

Our method consists of two steps: (1) thesaurus filtering, in order to retain only domain-relevant labels; and (2) an algorithm to map a topic (represented as a weighted list of words) to one or more

[2] https://www.eia.gov/

[3] IE0 and AEO have been released since 1985 and 1979, respectively. Parsing earlier documents would have involved OCR, which we rejected in the interest of producing a high-quality dataset. The 2021 reports have not been released at the time of research.

[4] https://pypi.org/project/pdfminer

[5] https://dev.elsevier.com/sc_search_views.html

[6] Keyphrases can consist of one or more tokens. e.g. *mining industry*

Source	Corpus	# Paragraphs	#Token (thousands)	Avg. #sentences / paragraph	Years
EIA	AEO	2,909	1,919	18.1	1997-2020
EIA	IEO	1,411	1,475	51.42	1997-2020
Scopus	Journals	24,353	5,483	7.87	1997-2020

Table 2: Corpus statistics.

labels (each represented as an unweighted set of associated phrases). Below we describe both steps, and propose two concrete mapping algorithms.

4.1 Label Filtering

The EuroVoc thesaurus was designed to cover all policy areas within the context of the EU. However, we are often interested in a subset of policy discussions and we can increase the relevance of our label selection by constraining the choice. We first remove all EU-specific entries (coded 10XX in the EuroVoc system; e.g. ”EU finance”), remove near-duplicates (e.g., ”economic geography” and ”political geography” consist of country names), and merge highly similar labels whose phrases' mean GloVe embeddings value have a cosine similarity greater than 0.95 (e.g. ”trade” and ”trade policy”).

From the remaining set, we filter irrelevant labels using log-odds ratios with informative Dirichlet prior (Monroe et al., 2008; Lucy et al., 2020), a widely used method to identify words that are statistically over-represented in a focus corpus of interest C_f as those words that have a higher chance of occurrence compared to a suitably chosen reference corpus C_r. Raw log-odds have a bias toward low-frequency words, which is alleviated by the Dirichlet prior which forces high-odds terms to significantly deviate from word-specific expected value of counts (as estimated from the joint $C_f \cup C_r$). We take as our focus corpus the concatenated three energy corpora described in § 3, while our reference corpus a representative sample of discussions in the Australian parliament, reflecting general political discourse as covered in EuroVoc.[7]

We calculate the log odds scores for all the terms in the EuroVoc dictionary, and associate each EuroVoc label l with a relevance score s_l as the median log-odds score of its associated terms in EuroVoc,

$$s_l = \text{median}(\{LO(v)\} : v \in V^{C_f} \cap V^l), \quad (1)$$

where V^{C_f} and V^l is the corpus vocabulary and the set of keyphrases under label l, respectively, and $LO(v)$ is the log-odds score of term v. We finally retain the top 40 labels with highest s_l as our energy topic label inventory.

4.2 DTM Topic Representation

The DTM learns one topic representation per time period, however, we want to assign EurVoc labels to topics as a whole. We obtain a single, global representation for each topic k as its aggregate weighted sum over all time steps t, where the terms at each timestep are weighted by that topic's probability of occurrence at that time. We then retain the 10 terms with highest score, and re-normalize the resulting scores to a valid probability distribution. The resulting topic representation is a 10-dimensional unit vector, which we denote as $\hat{k}$, and we refer to a word w's probability under this representation as $\hat{k}[w]$.

4.3 Topic Labeling

Given a DTM topic k represented as $\hat{k}$, we want to assign the top N EuroVoc labels that best match the topics content. We approach the automatic labelling task in two ways. The first is a match-based approach and the second uses word embeddings to label topics.

4.3.1 Importance-Based Topic Labeling

Intuitively, a label is relevant to a learnt topic if (a) it contains the topic's most relevant terms; and (b) these keyphrases are unique to the label, and do not occur widely across EuroVoc labels (”keyphrase uniqueness”). If a term occurs in many labels, it is often less informative as it loses the ability to distinguish labels. We quantify term-topic relevance as $\hat{k}[w]$ the probability of w in the re-normalized topic representation; and keyphrase uniqueness as $TFIDF[w, l]$, the TFIDF value of w under l, where the documents are all EuroVoc

[7] See https://www.aph.gov.au/Parliamentary_Business/Hansard

labels. The final topic-label score $\sigma^{imp}_{k,l}$ is

$$\sigma^{imp}_{k,l} = \sum_{w \in \hat{k} \cap l} \hat{k}[w] \times TFIDF[w,l]. \qquad (2)$$

We define the intersection in the summation based on either a full or a sub-token match between topic term and label keyphrase (e.g., topic term *solar* would match label keyphrase *solar energy*).

The proposed method is fast and simple to implement, and requires no resources beyond the trained topics and thesaurus labels. A disadvantage is its string-matching approach, which is oblivious to synonyms, or thematically related words. Our second labeling approach addresses this weakness.

4.3.2 Embedding-based Topic Labeling

The second approach makes use of pre-trained word embeddings. At a high level we produce an aggregated representation of our top word vector $\hat{k}$ as well as each EuroVoc label l in word vector space. We obtain a similarity score as the cosine similarity between the topic and label embeddings.

We use 50-dimensional pre-trained GloVe embeddings (Pennington et al., 2014).[8] We convert our top word vector $\hat{k}$ into an embedding-based vector emb_k, by taking a *weighted* average of the GloVe embedding representations of each word in it, where each word embedding is weighted by the words topic relevance $\hat{k}[w]$. An embedding for label l, emb_l, is computed as an unweighted average over its keyphrases. Multi-token topic terms (or keyphrases in EuroVoc) are represented as an unweighted average over their token embeddings.

The relevance score σ^l_k for DTM topic k and EuroVoc label l is then defined as the cosine similarity between their representations,

$$\sigma^{emb}_{k,l} = \text{cosine\_sim}(emb_k, emb_l). \qquad (3)$$

We finally associate each topic with its top $I{\geq}1$ associated labels as measured by either $\sigma^{imp}_{k,l}$ or $\sigma^{emb}_{k,l}$.

5 Experiment Settings

Our experiments consist of two parts. The first empirically evaluates the effectiveness of our proposed topic labelling method (§ 6) and the second leverages these labels to support a large-scale diachronic investigation of the discussion of energy technologies in the grey- and scientific literature (§ 7). We use the official DTM implementation.[9] Automatic model selection for topic models is an open research problem, where automatic methods such as normalized pointwise mutual information (NPMI) do not always correlate well with human judgment. As a consequence, especially in non-technical fields, researchers resort to their domain knowledge when selecting a parameterization. We selected a DTM parameterization based on a combination of NPMI scores and modelers' expectation in terms of topic variation over time. As a result, we set the number of topics K=30, and the topic variance parameter σ=0.05, and use default values for all other parameters. We applied the DTM with the above parameters independently to the three corpora described in § 3. All experiments below are based on the induced topics.

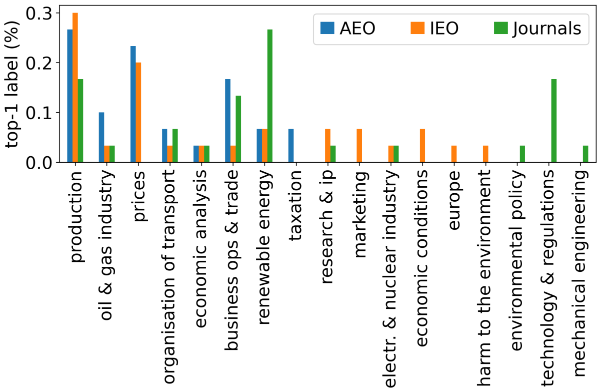


Figure 1: Proportion (%) of topics assigned a particular EuroVoc entry as top-1 topic label for our three corpora, using the embedding-based approach.

6 Topic Labeling Evaluation

6.1 Qualitative Analysis

We applied our topic labelling methods to the three corpora introduced in § 3, and inspected the distribution of top-1 (i.e., most highly associated) labels. The results for the Embedding method are shown in Figure 1. We can see that (a) the labels are varied and cover intuitively relevant aspects of energy-related discussions in government and academia; and (b) that the label distribution differs across corpora in meaningful ways. For example, *Renewable energy* is much more prevalent in the Journals dataset. Table 4 shows examples of induced topics represented as their 10 most highly associated terms with Embedding- and TFIDF-based top-2

[8] *'glove-wiki-gigaword-50'* obtained through https://radimrehurek.com/gensim/downloader.html.

[9] https://github.com/blei-lab/dtm

Strategy	TFIDF	Embedding	Baseline
Top-1	0.46	0.47	0.07
Top-4	0.45	0.47	0.08

Table 3: Human preferences (%) of automatic topic labeling methods when considering the top-1 or top-4 predicted labels by our methods or a random baseline.

labels.

6.2 Quantitative Evaluation

We evaluated our thesaurus-based topic labeling approach through human judgements. We obtained annotations from a group of 36 annotators who are proficient English speakers. All but one annotator were not domain experts. We presented annotators with DTM-induced topics, together with three label options: one based on the TFIDF mapping, one based on the Embedding-based mapping, and a randomly selected label. Annotators were asked to select the most appropriate label in a forced-choice paradigm. Given that we want to compare what label best represents a topic when our strategies differ, we do not include topics in the task where the embedding and TFIDF labels are automatically assigned the same label. Over our three models, this occurs for 23 topics. We evaluated two versions of our strategy: one where we paired each topic k with the single most highly associated label l in terms of labeling score $\sigma^*_{l,k}$ (top-1); and a second where we associate topics with their four most associated labels, capturing a mixture of information (top-4). Each annotation task consisted of a random sample of 30 out of a total of 90 induced topics (30 per corpus). We collected 20 sets of annotations for the top-4 strategy, and 16 sets for the top-1. Table 3 summarizes the human preferences. We can see that both our strategies significantly outperform the random baseline from filtered topics. In both the top-1 and top-4 strategy we see no difference between annotator preference toward either the TFIDF or embedding labelling strategy. The same pattern holds for each individual corpus.

We acknowledge the simple setup of our human evaluation, and leave comparison against stronger models for future work. The user study shows that non-experts can discern meaningful labels from our method, and as such complements our intrinsic qualitative label evaluation (§ 6.1), and our label-based case study of diachronic energy discussion which we present next.

7 Energy Discussions in EIA and Journals over Time

We present a broad analysis of the discussion of energy technologies in the grey- and scientific literature, showcasing the utility of our labeling scheme. We cover the following overarching themes: Electricity Generation Technologies, Sustainability and Geopolitics & Economy. We provide representative selection of DTM-induced topics for each of our corpora, with their automatically assigned labels, in Tables 5–7 in the appendix.

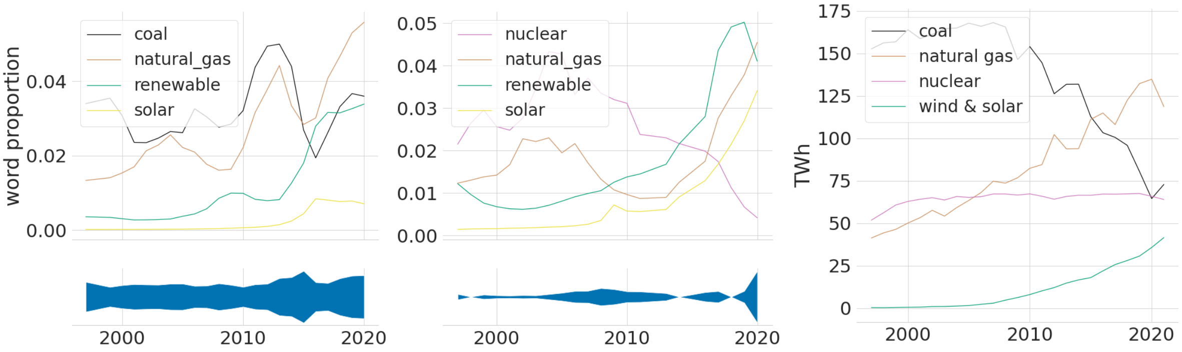
7.1 Electricity Generation Technologies

The EIA discusses electricity generation in detail as part of both AEO and IEO. Figure 2 shows how selected terms change over time in a topic on electricity generation in the AEO (2a), labelled as *Production*. We see similar patterns of discussion surrounding various energy sources in the IEO (2b). In both outlets natural gas and renewables increase in prevalence over time while coal (AEO) and nuclear (IEO) decrease. 2c shows the actual changes in generation (U.S. Energy Information Administration, 2021). These depictions allow us to more objectively assess how these two outlets have forecast, or not, the evolution of the energy system. The contrast between 2a and 2c is particularly illustrative as we see that the AEO has a somewhat belated reaction to the increase in new renewables (wind & solar) generation. The spike in the IEO topic on renewable energy only towards the end of the last decade is also remarkable, given that the situation in Europe, China and other major producers was similar to the US depicted in Figure 2c.

We also leverage our automatic labelling to uncover change within a topic over time. We create a normalised representation for a topic through its top 10 most probable words at each timestep, renormalized to sum to one, which we call $\hat{k}_t$. We assign topic labels to each $\hat{k}_t$ using the Embedding-based labelling strategy. Taking again IEO's topic ”Renewable energy” on electricity generation as an example, Figure 3 shows how the top-3 labels assigned to this topic change in prevalence over time. Initially, we observe *renewable energy* and *electrical and nuclear industries* being discussed in similar proportions while *oil and gas industry* is less prevalent. By 2020 *renewable energy* is the most prevalent label for this topic, while *Oil and gas industry* is discussed in the same proportion as *electrical and nuclear industries*. Our assigned

Corpus	Topic terms	Embedding	TFIDF
Journals	market; price; electricity; paper; competition; company; investment; risk; reform; industry	1 business ops & trade 2 production	1 prices 2 business ops & trade
AEO	resource; oil; production; natural_gas; tight; gas; shale_gas; drilling; estimate; technology	1 oil & gas industry 2 renewable energy	1 oil & gas industry 2 production
IEO	projection ; energy ; eia ; model ; international ; outlook ; include ; analysis ; world ; case	1 economic analysis 2 research & ip	1 renewable energy 2 world organisations

Table 4: One topic from each of our corpora, with its top-2 EuroVoc labels as assigned by the embedding and tfidf-strategy, respectively.



(a) AEO topic 0 (Production). (b) IEO topic 27 (Renewable Energy). (c) Real U.S. electricity generation

Figure 2: Change in word (top) and topic (bottom; blue bar) prevalence over time for two topics related to electricity generation (a) and (b). (c) shows real generation statistics for the U.S.

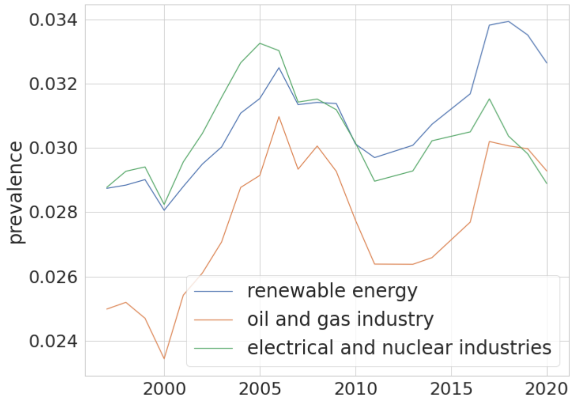


Figure 3: Label prevalence change over time for IEO topic 27 (Renewable Energy) (same as Figure 2b).

labels confirm the trends exhibited by topic-word prevalences in Figure 2.

7.2 Sustainability

Our automatic topic labels allow us to identify differences in discussion between publications. We combine topics from a model that have the same top-1 automatically assigned label by summing their proportions over time.[10] We sum again over all topics that have the same top-1 label assignment to achieve an overall proportion for the top-1 label at timestep t. We present the results in Figure 4.[11]

We can see that the Journals corpus has a larger focus on renewable energy and sustainability than AEO and IEO. The *renewable energy* and *environmental damage* top-1 label is much more prevalent in discussion in the Journals corpus. We confirm this by inspecting the learnt representations of topics in Journals by the DTM. Emissions are discussed from various perspectives including fuel sources (topic 29), China and coal (topic 22) and emission reduction (topic 4). The respective topics and their associated labels are shown in Table 5 in the appendix. We also see in topics not explicitly surrounding emissions mention of emission-reducing technologies such as ’chp’ (combined heat and power) and ’ccs’ (carbon capture and storage) and increase in ’energy_efficiency’ and ’efficiency’ terms over time in many topics, suggesting that even in non-explicit emission topics, sustainability and emission-reducing technologies are of increasing importance. This is exemplified

[10] For a topic k, it’s topic proportion at a timestep t, $p(k|t)$, can be calculated by marginalising over the documents d at that timestep.

[11] We utilise the open-source plotting strategy implemented by Müller-Hansen (Müller-Hansen et al., 2021).

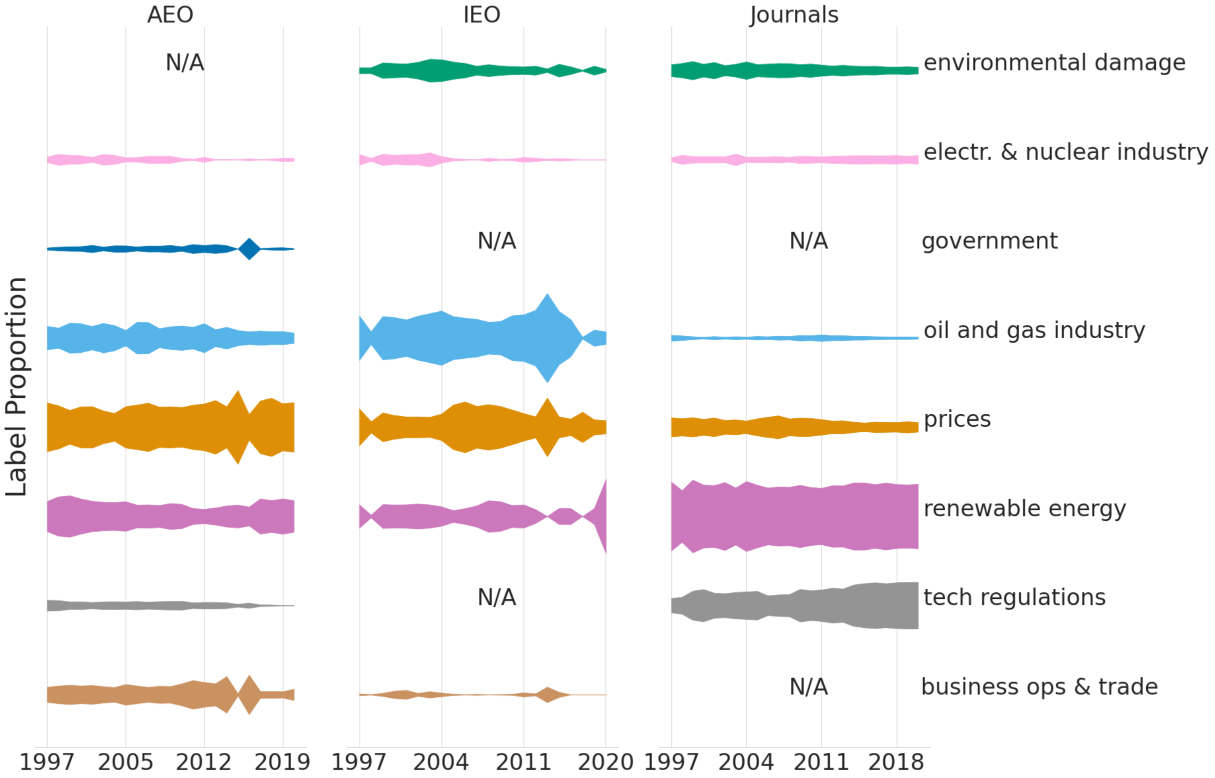


Figure 4: Comparison of discussions in the AEO, IEO and Journals. Topics were grouped by top-1 label.

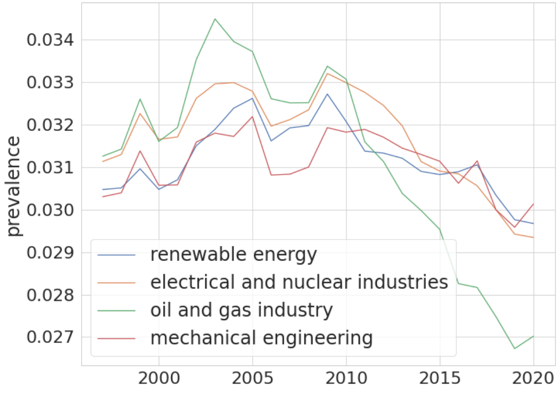


Figure 5: Label prevalence change over time for Journals topic 27 (Electric & nuclear industry).

in Figure 5, a topic on electricity generation with an early focus on electricity, oil and gas which is replaced in later years by renewables and mechanical engineering, indicating a shift toward sustainable technological development.

7.3 Economy and Geopolitics

Taking the same strategy we can analyse in Figure 4 the difference in discussion between the three outlets in terms of economic factors, see topics *prices* and *business ops & trade*. We see a large discrepancy between the proportion of discussion in the AEO and IEO corpora compared to the Journals corpus, and economic discussions are most prominent in the AEO corpus. This is expected as the AEO discusses prices and economy from a national perspective, while the IEO outlet instead discusses global markets and trade between countries. Surprisingly, the Journals dataset discusses *prices* very little proportional to other themes and does not discuss trade enough for it ever to be assigned as a top-1 label in any topic. We also note that most topics in AEO and IEO, particularly those related to economy and the oil industry, exhibit jumps in prevalence around the year 2015. This coincides with geopolitical events like the Paris Climate summit and follow-up policies like Obama's 2016 Clean Power Plan (CPP) in the U.S. Overall, our analysis again suggests a disconnect between corpora. Scientific journals show less concern for economic effects and more about regulatory aspects compared with the EIA.

8 Conclusions

We presented a novel method for topic labeling leveraging domain-relevant structured resources. We empirically showed the quality of our approach through human evaluation, and through its application in a detailed analysis of discussions on energy policy over the past 23 years. We highlighted differences in the discussions around electricity generation, sustainability and economy between nationally and internationally focused reports from the EIA and scientific publications over the same period. We release our grey literature corpora and software tool-set to support future research.

There are several areas of future work. In terms of down-stream analyses, our labelling framework can support additional comparisons for example across countries or other agencies such as non-governmental organizations; and can be extended to different thesauri, with different focus or level of detail. For example, EuroVoc captures all of *renewable energy* under a single label. Future work could also involve automatic splitting of assigned labels for example based on further hierarchical clustering of keyphrases associated with a label.

References

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2016. Automatic labelling of topics with neural embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 953–963, Osaka, Japan. The COLING 2016 Organizing Committee.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Dallas Card, Chenhao Tan, and Noah A Smith. 2017. Neural models for documents with metadata. *arXiv preprint arXiv:1705.09296*.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Darja Fišer and Benoît Sagot. 2008. Combining multiple resources to build reliable wordnets. In *International Conference on Text, Speech and Dialogue*, pages 61–68. Springer.

Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. 2013. Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 465–474.

Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1536–1545, Portland, Oregon, USA. Association for Computational Linguistics.

Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin. 2010. Best topic word selection for topic labelling. In *Coling 2010: Posters*, pages 605–613, Beijing, China. Coling 2010 Organizing Committee.

Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas us history textbooks. *AERA Open*, 6(3):2332858420940312.

Alfonso Martínez Arranz. 2015. Carbon capture and storage: Frames and blind spots. *Energy Policy*, 82(0):249–259.

Alfonso Martínez Arranz. 2016. Hype among low-carbon technologies: carbon capture and storage in comparison. *Global Environmental Change*, 41:124–141.

Klaus Mohn. 2020. The gravity of status quo: A review of iea's world energy outlook. *Economics of Energy & Environmental Policy*, 9(1).

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Finn Müller-Hansen, Max W Callaghan, Yuan Ting Lee, Anna Leipprand, Christian Flachsland, and Jan C Minx. 2021. Who cares about coal? analyzing 70 years of german parliamentary debates on coal with dynamic topic modeling. *Energy Research & Social Science*, 72:101869.

Finn Müller-Hansen, Max W. Callaghan, and Jan C. Minx. 2020. Text as big data: Develop codes of practice for rigorous computational text analysis in energy social science. *Energy Research & Social Science*, 70:101691.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Hanieh Poostchi and Massimo Piccardi. 2018. Cluster labeling by word embeddings and wordnet's hypernymy. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 66–70.

Ionut Sorodoc, Jey Han Lau, Nikolaos Aletras, and Timothy Baldwin. 2017. Multimodal topic labelling.

In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 701–706, Valencia, Spain. Association for Computational Linguistics.

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the european union's highly multilingual parallel corpora. *Language resources and evaluation*, 48(4):679–707.

Ralf Steinberger, Mohamed Ebrahim, and Marco Turchi. 2012. Jrc eurovoc indexer jex-a freely available multi-label categorisation tool. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 798–805.

U.S. Energy Information Administration. 2021. Electricity generation, capacity, and sales in the United States - U.S. Energy Information Administration (EIA). [Online; accessed 1. Oct. 2021].

Xiaojun Wan and Tianming Wang. 2016. Automatic labeling of topic models using text summaries. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2297–2305.

A Example DTM topics and labels

Tables 5–7 show for each of our corpora a set of induced topics. For each topic, we also provide the top-4 assigned EuroVoc by the Embedding and the TFIDF strategy, respectively. Topic 4, 22 and 29 of the Journals corpus are discussed in the results section. All other topics were chosen to be a representative sample of the discussion of the respective corpus to which they fall under.

ID	Top 10 Topic terms	Embedding	TFIDF

0	power; system; heat; generation;electricity; chp; energy; electric; district_heating; electrical	renewable energy (0.88);
		mechanical engineering (0.86);
		electronics and electrical engineering (0.84);
		technology and technical regulations (0.83)
			electrical and nuclear industries (9.47);
			business operations and trade (4.88);
			renewable energy (3.67);
			organisation of transport (2.0)

4	emission; carbon; reduction; cost; ghg; greenhouse_gas; reduce; policy; result; country	deterioration of the environment (0.79);
		environmental policy (0.78);
		renewable energy (0.77);
		production (0.74)
			environmental policy (15.19);
			accounting (2.61);
			deterioration of the environment (1.4);
			economic conditions (1.15)

22	china; carbon; reduction; sector; reduce; intensity; result; energy; increase	environmental policy (0.84);
		renewable energy (0.82);
		production (0.81);
		deterioration of the environment (0.81)
			environmental policy (13.92);
			asia and oceania (3.85);
			renewable energy (2.15);
			economic conditions (1.71)

28	energy; energy_efficiency; building; system; paper; analysis; indicator; measure; present; energy_consumption	renewable energy (0.96);
		environmental policy (0.86);
		production (0.85);
		technology and technical regulations (0.85)
			renewable energy (22.08);
			world organisations (5.76);
			electrical and nuclear industries (3.9);
			building and public works (1.86)

29	engine; fuel; emission; injection; diesel; co; combustion; high; low; increase	mechanical engineering (0.84);
		renewable energy (0.8);
		electrical and nuclear industries (0.8);
		oil and gas industry (0.8)
			environmental policy (4.61);
			oil and gas industry (3.32);
			mechanical engineering (2.67);
			electrical and nuclear industries (1.28)

Table 5: Five example topics induced from the **Journals corpus**, with their top-4 EuroVoc labels (scores) as assigned by the Embedding and TFIDF-strategy, respectively.

ID	Top 10 Topic terms	Embedding	TFIDF
1	coal; ton; production; cost; percent; productivity; export; u.s; increase; region	oil and gas industry (0.89); coal and mining industries (0.88); production (0.81); renewable energy (0.77)	coal and mining industries (16.94); production (4.35); regions and regional policy (2.82); accounting (2.19)

17	gasoline; ethanol; gallon; fuel; mtbe; sulfur; blend; motor; percent; requirement	oil and gas industry (0.85);
		renewable energy (0.72);
		food technology (0.7);
		deterioration of the environment (0.69)
			oil and gas industry (2.69);
			electrical and nu clear industries (0.81);
			taxation (0.35);
			organisation of transport (0.19)

19	vehicle; fuel; sale; percent; economy; new; increase; hybrid; car; standard	organisation of transport (0.88);
		production (0.88);
		prices (0.86);
		marketing (0.83)
			economic conditions (8.02);
			organisation of transport (5.56);
			marketing (5.1);
			land transport (3.21)

21	emission; carbon; co; ton; metric; ghg; carbon_dioxide; energy; relate; percent	renewable energy (0.78);
		oil and gas industry (0.76);
		deterioration of the environment (0.74);
		electrical and nuclear industries (0.73)
			environmental policy (11.52);
			renewable energy (2.28);
			deterioration of the environment (1.24);
			technology and technical regulations (1.21)

29	cost; market; electricity; price; competitive; customer; state; utility; transmission; power	prices (0.91);
		business operations and trade (0.91);
		production (0.9);
		accounting (0.9)
			prices (26.58);
			business operations and trade (14.09);
			accounting (5.78);
			environmental policy (3.57)

Table 6: Five example topics induced from the **AEO corpus**, with their top-4 EuroVoc labels (scores) as assigned by the Embedding and TFIDF-strategy, respectively.

ID	Top 10 Topic terms	Embedding	TFIDF
5	coal; import; ton; export; increase; percent; world; project; trade; coke_coal	oil and gas industry (0.9); coal and mining industries (0.88); production (0.82); renewable energy (0.77)	coal and mining industries (15.34); business operations and trade (8.08); world organisations (1.28); deterioration of the environment (1.18)

6	natural_gas; cubic; foot; gas; lng; reserve; increase; percent; year; production	oil and gas industry (0.88);
		renewable energy (0.85);
		production (0.84);
		deterioration of the environment (0.82)
			oil and gas industry (2.01);
			production (2.01);
			environmental policy (1.01);
			agricultural activity (0.99)

9	coal; china; world; percent; use; increase; consumption; share; total; btu	production (0.87); business operations and trade (0.83); oil and gas industry (0.83); prices (0.83)	coal and mining industries (11.93); business operations and trade (4.49); asia and oceania (2.47); world organisations (2.1)

25	emission; sulfur; reduce; reduction; standard; fuel; new; require; target; dioxide	deterioration of the environment (0.82);
		renewable energy (0.8);
		environmental policy (0.8);
		electrical and nuclear industries (0.78)
			environmental policy (10.89);
			technology and technical regulations (2.3);
			oil and gas industry (1.43);
			asia and oceania (0.94)

27	generation; natural_gas; renewable; nuclear; capacity; electricity; cost; increase; coal; power	renewable energy (0.91);
		electrical and nuclear industries (0.88);
		production (0.87);
		oil and gas industry (0.84)
			electrical and nuclear industries (8.21);
			coal and mining industries (2.97);
			accounting (2.36);
			demography and population (1.75)

Table 7: Five example topics induced from the **IEO corpus**, with their top-4 EuroVoc labels (scores) as assigned by the Embedding and TFIDF-strategy, respectively.