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

Thomas Scelsi
School of Computing and Information Systems
tscelsi@student.unimelb.edu.au

Alfonso Martinez Arranz
Chemical Engineering
alfonso.arranz@unimelb.edu.au

Lea Frermann
School of Computing and Information Systems
lea.frermann@unimelb.edu.au

1. Motivations

- Qualitative fields found to lack principled NLP usage[1]
- Enhance objectivity and reproduceability of topic modeling for text analysis and consequently make NLP use more principled.
- There has been little comparison between grey literature and scientific research publishing energy findings

2. Contributions

1. Two Diachronic corpora from the Energy Information Administration (EIA)
2. Analyse and contrast energy discussion in academic and independent literature over the last 20 years
3. Provide a simple and extensible automatic labelling technique to encourage principled and objective analyses especially in qualitative fields

3. Datasets

Energy Information Administration (EIA)

- Annual Energy Outlook
- International Energy Outlook

Academic Journals

- Energy Policy
- Applied Energy

4. Model

Dynamic Topic Model (DTM)[2]

- Extends LDA topic model[3]
- Learns topic representations over time (one per-timestep)
- K=30, top_chain_var=0.05

5. Automatic Labeling

Making NLP use in qualitative research more principled and efficient with automatic labelling!

- Utilise EuroVoc thesaurus as label source
- EuroVoc contains preferred phrases to-be-used in parliamentary discussion
- Each phrase belongs to a thematic label (see left)
- We curate a set of the 40 most relevant labels to our corpora and use them to label topics

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.

5a. Embedding-based Labeling

1. Create embedding representation for DTM topic
2. Create embedding representations for EuroVoc labels
3. Compare topic (*k*) and label (*l*) embeddings and assign most similar n EuroVoc labels as topic label

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

5b. Match-based Labeling

1. Match top-terms from DTM topic to phrases from each EuroVoc label
2. Weight each match by TF-IDF between term (*w*) and label (*l*)
3. Assign top-n highest scoring EuroVoc labels as topic label.

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

6. Labeling Evaluation

- Human judgement found large preferral of automatic labels over a random baseline
- No difference in preference between the two strategies

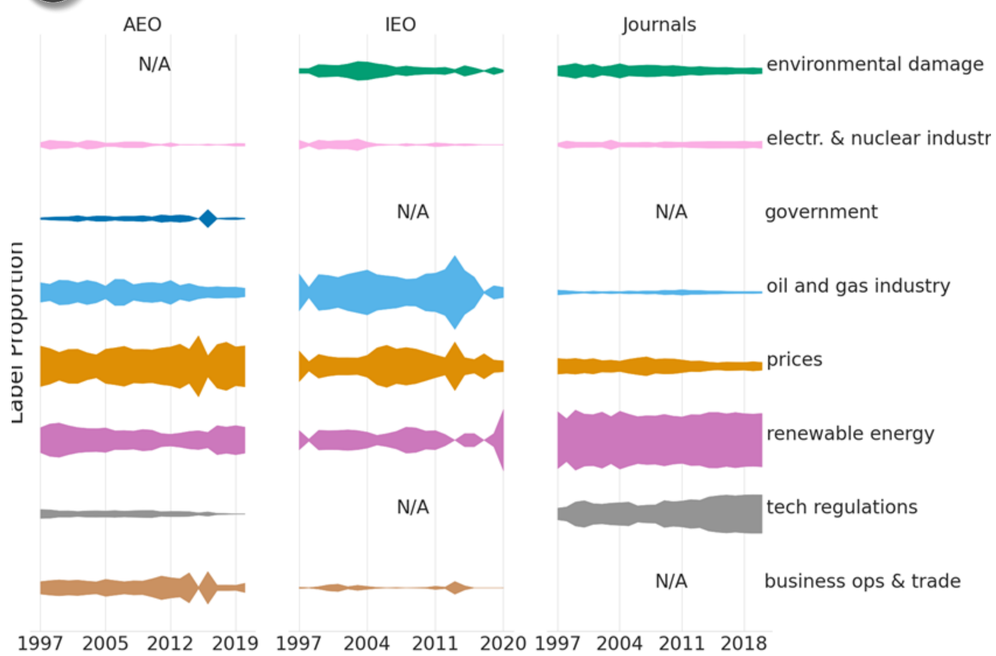
Embedding-based	Match-based (strategy 2)	Baseline
47%	46%	7%

7. Results

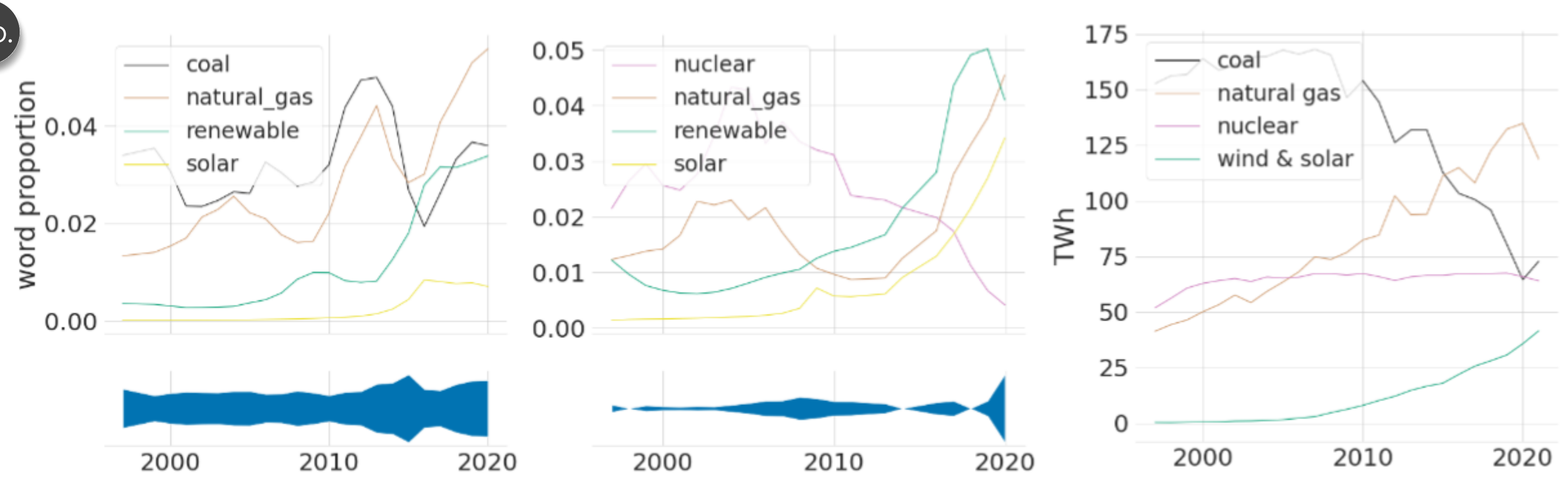
Key Findings

- Academic Journals discuss sustainability and renewables more than EIA (7a).
- Electricity Generation discussion mimics actual U.S. generation statistics (7b).

7a.



7b.



(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.

1 Müller-Hansen et al. *Text as big data: Develop codes of practice for rigorous computational text analysis in energy social science*, 2020

2 Blei & Lafferty *Dynamic Topic Models*, 2006

3 Blei et al. *Latent Dirichlet Allocation*, 2003