ORIGINAL ARTICLE

Crisis talk: analysis of the public debate around the energy crisis and cost of living

Rrubaa Panchendrarajan[1] · Geri Popova[1] · Tony Russell-Rose[1]

Abstract

A prominent media topic in the UK in the early 2020s is the energy crisis affecting the UK and most of Europe. It brings into a single public debate issues of energy dependency and sustainability, fair distribution of economic burdens and cost of living, as well as climate change, risk, and sustainability. In this paper, we investigate the public discourse around the energy crisis and cost of living to identify how these pivotal and contradictory issues are reconciled in this debate and to identify which social actors are involved and the role they play. We analyse a document corpus retrieved from UK newspapers from January 2014 to March 2023. We apply a variety of natural language processing and data visualisation techniques to identify key topics, novel trends, critical social actors, and the role they play in the debate, along with the sentiment associated with those actors and topics. We combine automated techniques with manual discourse analysis to explore and validate the insights revealed in this study. The findings verify the utility of these techniques by providing a flexible and scalable pipeline for discourse analysis and providing critical insights for cost of living—energy crisis nexus research.

Keywords Energy crisis · Climate change · Media · Discourse · NLP

1 Introduction

Climate change, alongside other ecological issues relating to human activity on the planet, is one of the most important, yet intractable issues facing our species at present. In no small part climate change is associated with the burning of fossil fuels, yet fossil fuels underpin much of modern prosperity and way of life. Our profound dependency on them makes climate change an extremely difficult issue to resolve, not least because it would require political and economic action on a very large scale, with potentially significant political and economic costs.

Any solutions to our fossil fuel dependency would require a public consensus around the reality of climate change, the desirability and feasibility of the required action, agreement around the nature of this action and acceptance of the cost, or at least belief that the benefit would outweigh any costs. Such public consensus has proven elusive, which is why the study of the public conversation around climate change has been the subject of numerous studies across a range of disciplines (see, for instance, Boykoff 2011, 2019; Carvalho 2007; Bednarek et al. 2022; Gillings and Dayrell 2023 amongst many others). In this paper we aim to contribute to this body of research, focusing on recent media discussions of the energy crisis, here specifically in relation to the cost of living.

In the early 2020s dependency on fossil fuels has become a topic debated in a different context, that of an energy crisis in Europe, associated in part with the war in Ukraine and access to Russian oil and gas. The energy crisis could become an inflection point, potentially mobilising a turn towards a more sustainable energy policy. Such an inflection point of crisis could also become, in the words of (Bednarek et al. 2022, 1), a moment of 'intense discursive construction, through which individuals and communities make sense of happenings as they unfold'. The energy crisis brings into the debate geopolitical and economic interests, but, at least in the UK context, it also brings to the fore issues of social justice and equity. A study of the public debates generated

✉ Geri Popova
  g.popova@gold.ac.uk

  Rrubaa Panchendrarajan
  r.panchendrarajan@gold.ac.uk

  Tony Russell-Rose
  t.russell-rose@gold.ac.uk

1 Goldsmiths, University of London, London SE14 6NW, UK

around the energy crisis can help us understand the complex intersection of discourses around fossil fuels, cost of living, sustainability, and social justice.

In this paper, we report the initial results of a project that aims to deliver on these aims. We collected a sample of UK mainstream media texts (see Sect. 3.1). The analysis is based on an up-to-date natural language processing (NLP) methodology. The next section provides the background to our methodology, and Sect. 3 lays out the details of our approach. Our preliminary results are described and discussed in Sects. 4 and 5.

2 Background

Understanding how the public perceives and responds to the energy crisis is of fundamental importance, and NLP offers various analytical approaches, such as topic modelling, sentiment analysis, semantic role labelling, and more. Some of these techniques operate at the level of individual tokens, while others focus on representations at the document level. In each case, it is necessary to first acquire or create a corpus of relevant documents and then to analyse that corpus.

A good overview of corpora used by the NLP community to investigate the debate on climate change is provided in Stede and Patz (2021). A further potentially relevant resource is the Science Daily Climate Change (SciDCC) dataset, presented in Mishra and Mittal (2021), which includes approximately 11,000 news articles on the topics "Earth and Climate" and "Plant and Animals" scraped from the Science Daily website. More recently, Volkanovska et al. (2023) described the process of building the multimodal InsightsNet Climate Change Corpus (ICCC) and using NLP techniques to enrich corpus metadata, creating a dataset that supports the exploration of the interplay between the various modalities that constitute the discourse on climate change.

Alternatively, a bespoke corpus can be created, e.g., using retrieval techniques to sample from a larger document collection or database. For example, Rebich-Hespanha et al. (2015) retrieved full-text articles from the LexisNexis database using query terms identified by previous research as part of an investigation into the discourse of climate change. More recently, Gillings and Dayrell (2023) retrieved full-text articles from Factiva and LexisNexis to diachronically explore the discourse of climate change, consisting of two distinct subcorpora: a tabloid subcorpus and a broadsheet subcorpus. Similarly, Liu and Huang (2022) created two large corpora of New York Times articles by retrieving documents from LexisNexis. Once the corpus has been

created, some degree of pre-processing is usually necessary. This will typically consist of a process of normalisation (e.g. case folding, tokenisation, stop word removal, etc.) to remove some of the linguistic 'noise' prior to higher-level processes such as phrase extraction and entity recognition. The normalisation process can be particularly challenging for social media data, such as tweets and other micro-blog posts, see Dahal et al. (2019).

Once the data has been normalised it is possible to use a unigram bag-of-words (BOW) representation to model the individual documents (Grimmer 2010). However, this approach has the disadvantage that concepts articulated as multi-word phrases can be lost in the modelling process and are thus unavailable to downstream processes. A more robust approach is to identify and extract such phrases as part of the pre-processing and normalisation process (Handler et al. 2016). There are various techniques and tools available for phrase extraction. AutoPhrase, for example, leverages high-quality phrases from public knowledge bases and utilises a POS-guided phrasal segmentation model, which incorporates the shallow syntactic information to further enhance the performance (Shang et al. 2018).

One of the more popular methods in media analysis is topic modelling, which can uncover recurring themes and subjects that shape the discourse. Topic modeling techniques, particularly Latent Dirichlet Allocation (Blei et al. 2003), can be used to find patterns in many data types. In the case of climate change, this can include analysis of business sustainability reports, corporate social responsibility reports (Benites-Lazaro et al. 2018a), and public policy (Quinn et al. 2010). Topic modeling can help identify which entities, be they governments, organizations, or individuals, are discussed in the context of climate change responsibility. This is vital for understanding the attribution of responsibility (Jelodar et al. 2019), and identifying the recurrent issues and themes within the overall discourse, the interests of various actors, and the major causes contributing to problematic issues (Benites-Lazaro et al. 2018b). Topic modeling can also be applied to track temporal changes in the prevalence of topics within climate change discourse (Blei and Lafferty 2006). Such approaches can reveal how discussion of the issues has evolved over time to reveal major cultural shifts, and hence provide a deeper, diachronic understanding of the problem space (Hoffman 2015).

Although topic modelling is a commonly used and highly insightful technique, it is predicated on the analysis of the text at the document level. To identify the specific roles played by individual actors within the text, it is necessary first to reliably extract them and second to identify their
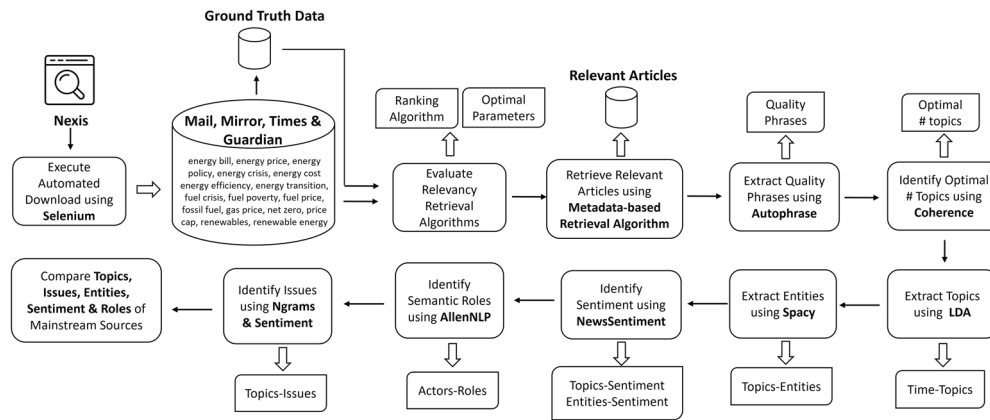
Fig. 1 The NLP pipeline

semantic role. The first of these tasks is usually achieved by entity extraction techniques. Information Extraction (IE) is a form of text analysis which extracts structured data from unstructured text (Maynard and Bontcheva 2015). Named Entity Recognition (NER) is a key information extraction task, which is concerned with identifying instances of entities such as people, locations, and organisations. A closely related task is Named Entity Linking (NEL), which identifies repeated instances of a particular entity within a given document, or across related documents and sources (Rao et al. 2013). A variety of tools exist for NER, such as StanfordNLP, NLTK, OpenNLP, SpaCy, and GATE (Schmitt et al. 2019).

Once the entities have been extracted, it is possible to apply related techniques such as semantic role labelling to identify the particular roles played by actors and organisations within the discourse. Semantic roles are typically applied based on 'frames' or schemata predicated on the syntactic constructions associated with particular verbs, which in turn are a reflection of the semantic components that restrict allowable arguments (Palmer et al. 2005). For example, the verb 'give' would typically have three arguments: an agent ('giver'), an object ('thing being given'), and a beneficiary. In more sophisticated schemes, there can be different types of arguments (called 'thematic roles') such as Agent, Patient, Instrument, and also of adjuncts, such as Locative, Temporal, Manner, and Cause (Ak et al. 2018).

The above techniques assume that topics and entities can be thought of as objective concepts that are instantiated and framed dispassionately within the discourse. The reality, of course, is quite different: topics and entities within the climate change debate are the subject of much opinion and argument, often with highly polarised, contrasting sentiment. As a result, sentiment analysis techniques have also been used to explore public opinion toward climate change, categorizing opinions as positive, negative, or neutral, thus providing deeper insights into the public's emotional stance on the issue (Pak et al. 2010). Sentiment analysis can also be used in a diachronic manner, to track major changes in public sentiment, identify shifts in public perception over time, and reveal how climate change sentiment evolves (Taufek et al. 2021).

The public discourse on the energy crisis and climate change is a complex narrative. NLP techniques offer a powerful lens to understand this discourse, uncovering societal attitudes, attributions of responsibility, and potential future actions. However, there are many challenges still to be overcome, and the analysis techniques usually need to be adapted to the target domain to get the best results (Derczynski et al. 2015).

3 Methods

To achieve wide empirical coverage and more general validity of our findings, we analysed the energy crisis discourse using automatic and semi-automatic techniques from corpus linguistics and NLP. The proposed methodology is based around a pipeline architecture composed of a set of NLP components which deliver a combination of individual insight and intermediate structure required by subsequent downstream components (see Fig. 1). This includes dedicated components for automated data collection, relevant article retrieval, topic modeling, entity extraction, sentiment analysis, semantic role labeling, and issue identification

Table 1 Corpus statistics

Source	Number of articles	Average article length (in words)
Daily Mail	5089	678
Mirror	3930	378
The Times	10,573	617
The Guardian	12,177	1778
Total	31,769	–

and visualization. These techniques have been applied to a corpus of data from mainstream media curated as part of the project, including samples from two broadsheets with different political leanings (The Times and The Guardian), and two tabloid newspapers similarly on different sides of the political spectrum (the Daily Mail and the Mirror). The following sections explain each component of the NLP pipeline.

3.1 Data acquisition

We collected mainstream media data from Nexis[1] which offers access to a vast database of media sources. A manual analysis of articles retrieved using the search query "energy crisis" revealed that crisis talk was initiated during the period of early 2014 and gradually evolved into a central topic in the early 2020s. From the initial data collection we extracted further closely associated keywords that still referred to the main aspect of our study, i.e. the cost of energy and the impact of this cost. Thus we collected publications in The Times and The Guardian, the Daily Mail, and the Mirror during the period of 01-Jan-2014 to 31-Mar-2023 using 16 keywords (see Fig. 1). This resulted in a corpus of 44,168 articles.

Each article is stored in a semi-structured form by Nexis with fields for title, section, date, writer, body, and meta-data indicating the geography and subjects discussed in the body of the article. The section field generally indicates the edition of the newspaper, and we removed items from non-UK editions, i.e. Australia News, World News, and US News. To preserve the UK focus of the study, we also utilized the geography field from the meta-data. Articles

mentioning any of the following terms in geography - London, England, United Kingdom, Ireland, Scotland, Wales, UK, and Ukraine were considered geographically related articles. We considered the country Ukraine as relevant geographically in order to include in the analysis the relationship between the latest crisis talk and the Ukraine war. The final corpus consists of 31,769 articles and Table 1 shows its statistics.

3.2 Relevant article retrieval

The corpus retrieved using the 16 search keywords related to energy crisis talk contained irrelevant articles as well. This is due to the retrieval nature of Nexis search, where an article with at least a single occurrence of any of the search keywords in either the title or body is retrieved as a positive hit. To sift out the irrelevant articles from the corpus, we utilized the subjects present in the metadata. Each subject listed in the metadata along with the percentage it is being discussed in the body is part of the Nexis topic taxonomy. We developed a metadata-based retrieval algorithm that starts with seed-relevant subjects and iteratively chooses relevant documents and relevant subjects until no new relevant subjects are found. The retrieval algorithm is controlled by the following three parameters.

- Discussion threshold d—Threshold used to determine the minimum percentage of discussion of a relevant subject required to consider an article as relevant.
- Popularity threshold p—Minimum percentage of popularity of a subject among relevant articles required to consider a subject as relevant. Here the popularity is determined using the number of articles that contain the subject in meta-data.
- Growth factor r—Factor controls the growth of the discussion threshold. This enforces the increase in the percentage of discussion of a relevant subject required to consider an article as a relevant article with each iteration of the retrieval process.

The metadata-based retrieval algorithm is presented in Algorithm 1. We aimed at analysing the crisis talk centered around the subjects "energy crisis" and "energy policy", hence these keywords were used as seed relevant subjects of the algorithm.

---

[1] https://www.lexisnexis.co.uk/.

Algorithm 1 Metadata-based Relevant Article Retrieval

Input: *corpus*, d (discussion threshold), p (popularity threshold), r (growth factor of d)
Output: Relevant Articles

1: $relevant\_subjects \leftarrow [\text{"energy crisis"}, \text{"energy policy"}]$
2: $new\_subjects \leftarrow relevant\_sub$
3: $relevant\_docs \leftarrow []$
4: $t \leftarrow 0$
5: **while** $new\_subjects$ is not empty **do**
6:      **for** each $article$ in $corpus$ **do**
7:          **if** $article$ contains any $new\_subjects$ with discussion $\% \geq d * r^t$ **then**
8:              Add $article$ to $relevant\_docs$
9:          **end if**
10:     **end for**
11:     Remove $relevant\_docs$ from $corpus$
12:     $new\_subjects \leftarrow []$
13:     **for** each $subject$ in $relevant\_docs$ **do**
14:         **if** $subject$ not in $relevant\_subjects$ and popularity in $relevant\_docs \geq p$ **then**
15:             Add $subject$ to $new\_subjects$
16:         **end if**
17:     **end for**
18:     Add $new\_subjects$ to $relevant\_subjects$
19:     $t \leftarrow t + 1$
20: **end while**

We created a human-annotated 'gold standard' validation set for choosing the optimal parameters of the metadata-based document algorithm and comparing it with traditional document ranking algorithms. We randomly sampled 150 articles from the corpus and labeled them as "relevant" or "irrelevant" with respect to the subjects "energy crisis" and "energy policy". Table 2 presents the statistics of the validation set. It can be observed that only 61.3% of the sample is relevant to the subjects "energy crisis" and "energy policy", thus showing the need to retrieve the relevant articles from the whole corpus for accurate results from the remaining pipeline.

We varied the parameters of the metadata-based retrieval algorithm as follows: discussion threshold and popularity threshold varied from 0 to 100 in a step size of 5, and the growth factor varied from 1 to 1.4 in a step size of 0.1. This resulted in 2100 combinations of parameters. We compared the meta-data-based article retrieval algorithm with the traditional TF-IDF and Word-embedding-based ranking techniques.[2] We observed that the metadata-based retrieval algorithm outperformed TF-IDF-based and Word-embedding-based retrievals. F1 scores of the three algorithms and the optimal setting of the metadata-based

retrieval algorithm are listed in Table 3. Table 4 presents the relevant subjects in the order they were retrieved and the number of relevant articles from each newspaper source. Subjects related to energy shortage, fuel price, and inflation are commonly retrieved as relevant subjects across all four sources. All the common relevant subjects are highlighted in Table 4.

3.3 Phrase mining

This component receives the relevant articles retrieved using the metadata-based article retrieval algorithm as input and generates quality phrases to produce an intermediate structure for the downstream components. Considering only the individual words or all the possible n-grams of the corpus may not lead to an effective and scalable information retrieval pipeline. To serve this purpose, we used Autophrase (Shang et al. 2018), a phrase mining tool for extracting quality phrases from very large corpora. Quality phrases are extracted by Autophrase based on four factors: popularity, concordance, informativeness, and completeness. Given an input corpus, the tool generates quality scores for phrases ranging from single words to 6-grams with respect to the surrounding words and annotates phrases exceeding

---

[2] https://github.com/4OH4/doc-similarity.

Table 2 Statistics of validation set

Source	# relevant docs	# irrelevant docs	Total
Daily Mail	27	18	45
Mirror	28	10	38
The Times	24	15	39
The Guardian	13	15	28
Total	92 (61.3%)	58	150

the thresholds as quality phrases. The default implementation of the tool uses 0.5 and 0.8 as quality thresholds for individual words and phrases respectively.

3.4 Topic extraction

Identifying key issues discussed in the corpus is an essential task to analyze the discussions centered on these issues. In pursuit of this goal, we used a topic modeling approach to identify topics discussed in the corpus. We initially considered the extraction of hierarchical topics based on the hypothesis that the discussion in the corpus can be structured into root, super topics, and sub-topics each addressing

Table 3 Performance of the Relevant Article Retrieval Algorithms

	F1 Score	Number of relevant documents
Metadata-based Retrieval (d=70, p=25, r=1)	0.657	16,589 (52.2%)
Word embedding-based Retrieval	0.554	28,729 (90%)
TF-IDF-based Retrieval	0.625	20,868 (65.7%)

Table 4 Relevant subjects from the relevant articles retrieved

Source	Relevant subjects
Daily Mail (3198, 62.8%)	Energy shortages, price increases, energy & utility policy, oil & gas prices, prices, inflation, cost of living, taxes & taxation
Mirror (2165,55%)	Energy shortages, energy & utility policy, price increases, oil & gas prices, prices, inflation, cost of living
The Times (5477,51.8%)	Energy shortages, energy & utility policy, oil & gas prices, public policy, energy & utility regulation & policy, Prices, inflation, price increase
The Guardian (5749, 47.2%)	Energy shortages, energy & utility policy, oil & gas prices, public policy, energy & utility regulation & policy, Prices, price increases, inflation

Table 5 Number of optimal topics

Source	Number of topics
Daily Mail	20
Mirror	19
The Times	14
The Guardian	20

various aspects of the energy crisis with varying degrees of depth. However, the initial results were not promising, revealing the non-hierarchical nature of topics discussed in the corpus. Following that, we used Latent Dirichlet Allocation (Blei et al. 2003) to extract flat topics from the corpus.

We used coherence (Röder et al. 2015) as an evaluation metric to determine the optimal number of topics for each source. Coherence measures the degree of semantic similarity between the top N words of a topic. While there are different versions of coherence metrics available, we use the CV measure which is shown to have a high correlation with human judgment (Röder et al. 2015). Following previous studies (Benites-Lazaro et al. 2018b), we range the number of topics from 10 to 20 and obtain the coherence score as an average of three runs, each composed of 500 epochs. Table 5 shows the optimal number of topics identified for each source. Once the topics were learned, we used chat-GPT (OpenAI 2023) to generate a topic label for each topic using the top 20 words of the topic. Chat-GPT was prompted with the top 20 words to identify a topic label with a maximum of 5 words. Figure 2 presents the topic labels of the topics learned (we discuss this further in Sect. 4).

3.5 Entities, sentiment and role extraction

Another key objective of this analysis is to identify key actors and their roles with respect to the issues in crisis talk. We used Spacy[3] for identifying the actors, by extracting the named entities, and sentiment expressed towards the actors was extracted using NewsSentiment (Hamborg et al. 2021). NewsSentiment performs target-dependent sentiment analysis, and we input actors as targets to the sentiment model to identify the sentiment expressed towards an actor as a score ranging from -1 to 1. Actors and sentiments in a sentence are linked to a topic by assigning a topic label at the sentence level, based on major topics discussed in a sentence. The average sentiment score of an actor or a topic is obtained by taking the mean value of all the sentiment scores associated with it.

To identify the roles played by the actors, we performed semantic role labeling using AllenNLP (Gardner et al. 2017). Among the arguments extracted by AllenNLP, we

[3] https://spacy.io/.

(a) the Daily Mail

(b) the Mirror

(c) The Times

(d) The Guardian

Fig. 2 Topics learned across four sources

use the agent or causer of a verb (indicated as ARG0) and the target (indicated as ARG1) for role analysis of actors. For example, Tables 6 and 7 present the top 5 agents and targets of the topic "Energy bill and price caps" from the Daily Mail and the top 10 verbs they are involved in the corresponding roles.

Table 6 Top 5 Agents and their corresponding popular verbs of the topic "Energy bill and price caps" from the Daily Mail

Agent	Top 10 verbs
Government	Needs, take, act, cover, introduce, announced, set, pay, put, steps
Households	Struggling, save, paying, face, pay, switching, facing, reduce, see, paid
Customers	Pay, save, switch, switching, make, struggling, get, use, keep, cut
People	Pay, save, need, switch, struggling, cut, make, switching, feel, thought
Suppliers	Charge, offer, charging, pass, offering, cut, ensure, hitting, calculated, sending

Table 7 Top 5 Targets and their corresponding popular verbs of the topic "Energy bill and price caps" from the Daily Mail

Target	Top 10 verbs
Bills	Rise, rising, soaring, cut, soar, reduce, increase, pay, fall, estimated
Customers	Existing, moved, protect, transferred, leave, urged, panic, transfer, misleading, worried
Prices	Rise, fall, rising, raising, pushed, raised, increased, pushing, raise, cut
Energy	Used, save, buy, use, switching, saving, provide, conserve, costs, imported
Energy bills	Rise, rocketing, soaring, freeze, soar, cap, rising, jumped, cut, reduce

3.6 Issue identification

We define an issue as a popular noun or verb phrase with extreme variation in sentiment (Choi et al. 2010). Accordingly, we considered topic labels as default candidate issue phrases representing root issues in the corpus. We further extracted n-grams from the top 100 words of each topic as candidate phrases for issue identification. We computed the average positive, and average negative sentiment score of each candidate phrase and classified them as an issue if the difference between average positive sentiment and average negative sentiment is greater than 0.8. As anticipated, all the topics were classified as issues due to the highly polarized nature of the topic discussions. In addition to that, the issues identified using popular n-grams unveiled both unique issues associated with a single topic as well as the common issues, being the central discussion across multiple topics. For example, Table 8 shows sample topics and issues identified in the Daily Mail. While 'global warming' is an issue specific to the topic 'Transition to green transportation solutions', 'cost of living' can be observed as a common issue across multiple topics.

3.7 Visualization

The final component of the pipeline aggregates the information retrieved to generate conceptual graphs that we can visualize to illustrate and explore the relationships between topics, issues, entities, and sentiment. We consider actors extracted using Named entity recognition and agents and targets extracted using semantic role labeling as entities of the system. We used Pyvis,[4] a Python library to generate the following four categories of conceptual graph:

1. Topics-Issues Graph—Depicts the relationship between topics and issues of a particular source and the properties of an issue including overall and topic-level popularity.
2. Topics-Entities Graph—Depicts the relationship between topics and entities of a given source and the properties of an entity including popularity, role, and sentiment.
3. Source-Issues Graph—Depicts the top 50 issues across all sources and the properties of an issue including popularity and sentiment.
4. Issue-Entities Graph—Depicts the popular entities across all sources for a given issue and the properties of those entities including popularity, role, and sentiment.

The size of a node is used to roughly indicate its overall popularity and the size of an edge is used to roughly indicate the popularity of the node with respect to the connected node (e.g., edge size between an issue node and topic node roughly indicates the popularity of an issue within the topic discussion). The Color of an edge is used to indicate the sentiment expressed towards the node with respect to the connected node (e.g., the color of an edge between an entity node and topic node indicates the average sentiment expressed towards an actor within the topic discussion). The following color codes are used to represent sentiment via edges in the concept graphs:

- Red: Negative sentiment
- Green: Positive sentiment
- Blue: Neutral sentiment
- Gray: No sentiment extracted

Further, we use color codes for the entity nodes to indicate the role of the entities in concept graphs as follows.

Table 8 Sample topics and issues identified in the Daily Mail

Topic	Sample issues
Economy and inflation concerns	Base rate, borrowing costs, cost of living, economic growth, financial crisis, living standards, price rises, raise rates, rising prices, soaring inflation
Energy bill and price caps	Annual bills, customer service, direct debit, energy price cap, fuel poverty, loyal customers, price cap, price hikes
Personal finance and housing expenses	Base rate, cost of living, council tax, higher rate, home loan, household bills, savings rates
Transition to green transportation solutions	Carbon emissions, charging points, electric car, global warming, net zero, petrol and diesel
Chancellor's tax and economic policies	Basic rate of income tax, economic growth, fiscal policy, higher taxes, income tax, living standards, small businesses

- Black: Entity plays the role of the agent only
- Yellow: Entity plays the role of the target only
- Purple: Entity plays the roles of both agent & target
- Light Blue: Entity plays no role

4 Results

By identifying topics, the NLP analysis allows us to gain an understanding of where the media directed its attention during the energy crisis. As the topics were given labels separately for each source and sometimes different topics have similar labels (e.g. 'Eurozone economy' and 'European stock market' in The Guardian), it makes sense first to amalgamate some topics manually into broader themes and then try to compare across the different newspapers.

Topics across all newspapers fall into five broad themes: (i) energy and cost of living, (ii) economy, (iii) politics, (iv) geopolitics, and (v) climate change and green energy. That the main framing of the energy crisis is as a cost-of-living issue is not surprising – energy cost was one of the main contributors to substantial inflationary pressures experienced in the UK during this period of time (see ONS data) – though as noted later, different newspapers give different prominence to this aspect. The topics of economy, politics, and geopolitics are interconnected, with some topics falling more clearly into one of these themes, and some combining all of them. The energy crisis in the UK was one contributor to an economically turbulent time given the economic effects of the COVID-19 pandemic (a topic in The Guardian), slowing growth ('recession' is a topic in The Guardian, 'economic impact of high inflation' is a topic in The Times), and political events during, for instance, the cabinet of Liz Truss ('tax and economy policy debate' is a topic in The Times). The theme of the economy covers some cross-border issues as well, with the 'European stock market' appearing as a topic in The Guardian, for instance. The Guardian more generally seems to pick up more European issues in comparison to other newspapers, and this interest in Europe is partially confirmed by the announcement recently of an Autumn 2023 launch of a European issue of The Guardian.

There also seems to be more discussion in The Guardian of the impact on the economy of the pandemic. The Guardian and the Mail also pick up the theme of the food industry. This does not seem to be as prominent in the Mirror and The Times, so is not picked up as a separate topic.

As we had anticipated, climate change is another strand of the public debate. It seems to be covered with more intensity in The Guardian, where it is a separate topic, but all newspapers pay attention to the related topics of renewable and green energy. There are some suggestions from previous research that at times of crisis the topic of climate change tends to be moved to the background and the intensity of discussion in the media reduces (Boykoff [2011]). Topic modelling here suggests that if the energy crisis had that effect in this case, it was less pronounced for The Guardian. But it also suggests that the general focus is on replacing current energy sources with more climate-friendly ones. Energy itself is a prominent topic across all newspapers, with topics suggesting some attention paid to the oil and gas industry, to the impact of the energy crisis (i.e. to the general issue of a cost-of-living crisis and inflationary pressures), and to the various measures taken to deal with it (e.g. energy price cap) and their impact. There are also some variations here, however. For instance, the Mail seems to focus more attention on the UK shale gas industry, the steel industry, and nuclear power, which come up as separate topics.

Perhaps the most visible difference identified in the topic modelling part of the analysis is how much less variation there is in the Mirror. There are fewer topics, and the ones identified seem to relate predominantly to the cost-of-living crisis. Where other newspapers discuss the energy crisis as a political, economic, and geopolitical problem, the main, if not exclusive, framing in the Mirror, it would seem, of the energy crisis is as a cost-of-living issue.

Some of these observations can be elaborated by looking at particular issues, as defined in Sect. 3. We can trace which issues are shared by which newspapers, and which are not, see Fig. 3.

Other visualisations allow us to zoom in on particular issues and the actors associated with them. For instance, one
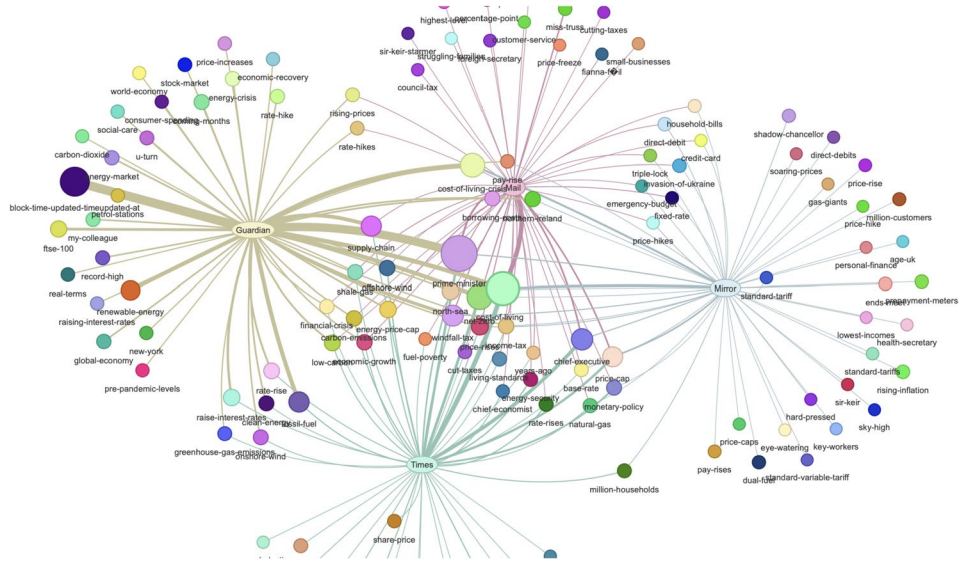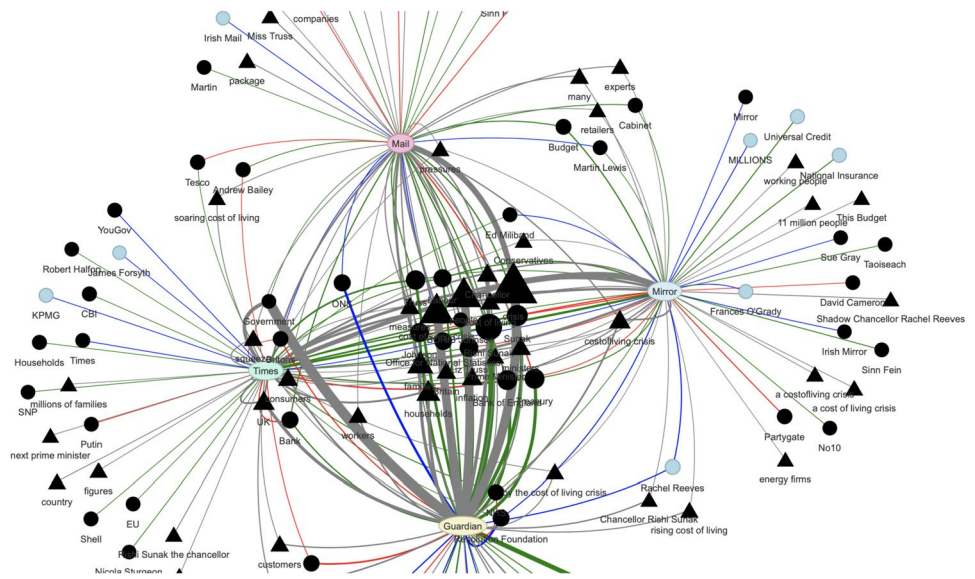
Fig. 3 Issues across all four sources

Fig. 4 The 'cost-of-living' issue and actors associated with it

issue all newspapers give attention to is the cost-of-living, see Fig. 4.

As is clear from Fig. 4, the cost of living is framed as a political/policy problem, with the actors being either those affected by it (households, people, families), and those called upon to provide solutions (Conservatives, Bank of England, Johnson). On some of these actors the media seems politically divided, e.g. in this graph Liz Truss and Sunak attract negative sentiment from the Mirror, but positive sentiment from the other three newspapers.

The pipeline picks up 'cost-of-living-crisis' as a separate issue. The crisis framing seems more entrenched on the political left, since this issue links only to the Mirror and The Guardian, see Fig. 5.

Although they share the crisis framing, the Mirror and The Guardian don't always share the sentiment associated with the different actors, e.g. here 'Boris Johnson' receives positive sentiment from The Guardian, but negative one from the Mirror, and a similar pattern obtains for 'the Bank of England' and 'Sunak' (although it should be noted that under different naming variations, the same actors may be associated with different sentiments, e.g. 'Johnson' is associated with negative sentiments for both newspapers).

Other issues related to the cost of living are shared not by the political left, but by the tabloids, e.g. 'hard-pressed' (shown in Fig. 6), 'eye-watering', 'price hike'.

Clearly, the impact of the energy crisis and cost-of-living crisis more generally on various sections of society is given more prominence in the tabloids. The broadsheets, by

Fig. 5 The 'cost-of-living-crisis' issue and actors associated with it
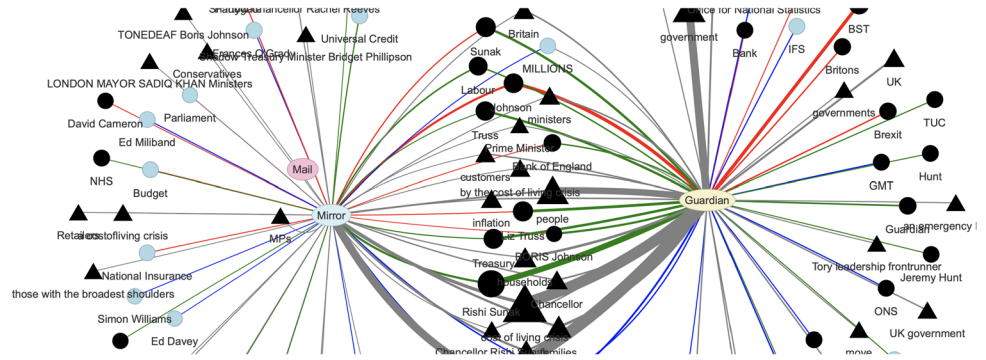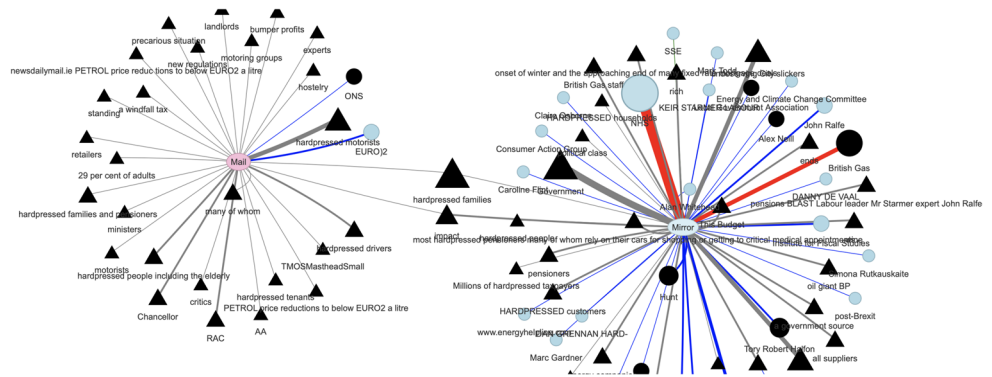

Fig. 6 The 'hard-pressed' issue and actors associated with it

contrast, where there is a wider range of topics and more attention on issues related to the economy more generally, seem to present in more detail the complex interleaving of economic and political factors. The sentiment analysis suggests that at this point negative sentiments directed at the government of the day are more readily expressed further to the left of the political spectrum, in the Mirror. Another important aspect of the energy crisis, that of energy security, is picked up as an issue too, but the Mirror is again different from the other newspapers, in that it gives it much less attention, see Fig. 7.

As noted, the analysis also picks up some issues related to the climate crisis, although this is a separate topic only in The Guardian. The range of issues we find suggests that the energy crisis gives an impetus to discussions of replacements for fossil fuels, that is, green energy and renewables. One popular issue shared across all newspapers is net-zero, see Fig. 8.

As before, the discussion seems to be focused on the political actors that shape policies, i.e. actors like the government, Boris Johnson, Ed Miliband, and others. The sentiment analysis suggests some expected political divisions (e.g. the Net Zero Scrutiny Group attracts negative sentiment from The Guardian and neutral sentiment from The Times) and some less expected ones (Johnson is associated with positive sentiment in the Mail but also The Guardian, and negative from the other two newspapers). This

indicates that there is no complete political consensus when it comes to net zero targets, something that recent political events seem to confirm.

5 Discussion

In this work, we have adopted multiple frames of reference: applied at the entity level, focusing on actors and roles, and at the document level, focusing on topics and issues. Reliably identifying issues has proven to be a deceptively complex problem: the term 'issue' may have an intuitively simple meaning and denotation, to the extent that few human readers would have trouble recognising their existence within a sample of online debate. However, operationalising that concept into a definition that can be accurately and unambiguously applied in an automated manner is a very different undertaking. Issues can be articulated explicitly or referred to only indirectly, and often display a fractal nature within a complex inter-connected network of concepts and ideas, relying on the reader's knowledge of world events and current affairs for their accurate interpretation. In practice, we have relied on certain heuristics to signal their presence, such as instances of common phrases that are simultaneously associated with both a high degree of positive and negative sentiment (suggesting that they have the effect of polarising the debate). This has proven to be

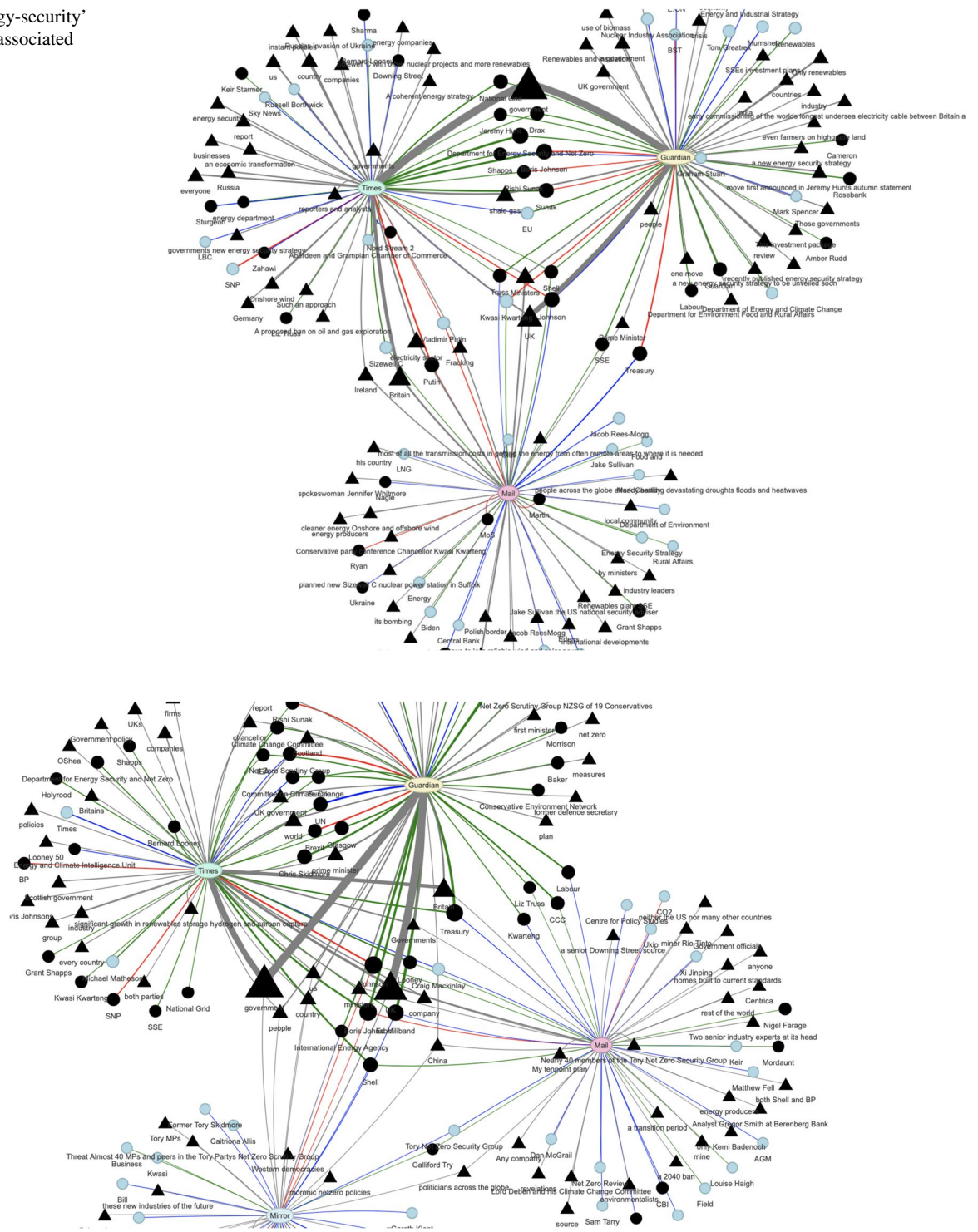**Fig. 7** The 'energy-security' issue and actors associated with it

Fig. 8 The 'net-zero' issue and actors associated with it

a useful first approximation, and no doubt more robust techniques could be the focus of future work.

As we observed above, the NLP analysis gives us an overview of what is discussed in the media. We can see from the topic modelling that during the energy crisis, the media was focused mostly on internal politics and internal and relevant external economic issues. The newspapers we studied framed the energy crisis primarily as a cost-of-living crisis, in some cases almost exclusively so. Climate change appears mostly as a solution to the problem, i.e.

green and renewable energy, though policy discussions around net zero, for instance, appear in all newspapers. In this case, the discussion again centers around political actors and differing sentiment suggests consensus has not been reached. The presence of the climate crisis as a topic in The Guardian indicates it continued to give this significant attention. Of course, we should acknowledge that, inevitably, the selection of a particular keyword set will prioritise a particular perspective on a topic, possibly at the expense of others (such as sustainability, alternative energy sources, etc.). We recognise that multiple perspectives exist and these could be explored further as part of a future iteration.

Although the NLP tools we deployed give us a breadth of analysis, their focus remains on what is being said and who is being talked about. In the current pipeline, they don't tell us enough about the how and the why of discourse. In some cases, the analysis retains too high a level of generality (topics), and in others, it is too granular. Semantic roles, for instance, tell us whether entities are actors or themes or both on the sentence level, but they can't answer questions like 'Who is held responsible for the crisis? Who is tasked with finding solutions for this crisis?'. For questions like these, and for a more fine-grained analysis of the discourse itself, we need to go to the next step, i.e. finding principled ways to extract data for qualitative analysis. A preliminary qualitative investigation of discussions around the net zero policy, for instance, suggests that The Guardian coverage during the period of the energy crisis reflects on the difficult trade-offs between providing support to people during a cost-of-living crisis, securing sufficient energy to cover energy needs in the short and the longer term, and energy security. The Guardian seems to lobby that the moment be seen as an opportunity, as encapsulated the following quote from April 2022 suggests: "Given a cost-of-living crunch caused by the rocketing price of fossil fuels, and the new priority of energy independence following Russia's invasion of Ukraine, an imaginative and proactive government would move to [...] seize the moment." By contrast, The Times seems to discuss net zero mostly in the context of finance and business, observing economic costs and benefits, but with a less pro-active stance with respect to climate change action.

This brings us to the issue of evaluation. The qualitative, human-led analysis could provide a useful check on the view of the energy discourse generated by the NLP pipeline. For instance, as we mentioned above, humans have powerful intuitions about what 'issues' are present in media discussions. However, humans are biased and the amount of text they can read is significantly smaller. Our next step, then, would require us to integrate the manual qualitative analysis into the pipeline in a principled way that would provide for cross-fertilisation and mutual validation of the manual and automatic analyses. Part of this could include exploring other frames of reference, not just in terms of topical scope but also granularity: from broad perspectives to narrow issues and vice versa, with further investigation of the iterative retrieval approaches described in Sect. 3.2.

6 Summary and conclusions

In our work on the project to date we have built a functioning pipeline for the NLP analysis of media discourse, focusing our attention on a prominent recent media discussion: the energy crisis that affected the UK and most of the rest of the world and became particularly acute with the war in Ukraine. The pipeline gives us broad insights into the data, building a picture of the topics covered by different media sources, the issues those topics comprise, as well as the entities involved, overlaid with sentiments expressed towards them. The picture that emerges is a focus on the energy crisis as a national, rather than global, crisis, with a focus on its ramifications on people and households. Political actors are the ones most prominent in the discussion, presumably because they are tasked with resolving it. Where climate issues feature in the debate the focus is not on energy consumption as one of the causes of climate change, but on solutions and mitigations, e.g., renewable energy. In future work, we hope to apply these techniques to an additional corpus of social media data extracted from Twitter (now X) and compare the insights generated to highlight key points of differentiation with mainstream media. We hope also to develop partnerships with related organizations, such as the British Ecological Society, Friends of the Earth, National Association for Environmental Organisation. Finally, we hope to offer a reusable platform to support other discourse analysis investigations, e.g., public discourses around energy, poverty and equality, attitudes to risk, and more.

References

Ak K, Toprak C, Esgel V, Yildiz OT (2018) Construction of a Turkish proposition bank. Turk J Electr Eng Comput Sci 26(1):570–581

Bednarek M, Ross AS, Boichak O, Doran Y, Carr G, Altmann EG, Alexander TJ (2022) Winning the discursive struggle? The impact of a significant environmental crisis event on dominant climate discourses on Twitter. Discourse Context Media. https://doi.org/10.1016/j.dcm.2021.100564

Benites-Lazaro LL, Giatti L, Giarolla A (2018a) Sustainability and governance of sugarcane ethanol companies in Brazil: topic modeling analysis of CSR reporting. J Clean Prod 197:583–591

Benites-Lazaro LL, Giatti L, Giarolla A (2018b) Topic modeling method for analyzing social actor discourses on climate change, energy and food security. Energy Res Soc Sci 45:318–330

Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd international conference on machine learning, pp 113–120

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Boykoff M (2011) Who speaks for the climate? Making sense of media reporting on climate change. Cambridge University Press, Cambridge

Boykoff M (2019) Creative (climate) communications: productive pathways for science, policy and society. Cambridge University Press, Cambridge

Carvalho A (2007) Ideological cultures and media discourses on scientific knowledge: re-reading news on climate change. Public Understand Sci 16:223–243

Choi Y, Jung Y, Myaeng S-H (2010) Identifying controversial issues and their sub-topics in news articles. In: Intelligence and security informatics: Pacific Asia workshop, Paisi 2010, Hyderabad, India, June 21, 2010. Proceedings, pp 140–153

Dahal B, Kumar SA, Li Z (2019) Topic modeling and sentiment analysis of global climate change tweets. Soc Netw Anal Min 9:1–20

Derczynski L, Maynard D, Rizzo G, Van Erp M, Gorrell G, Troncy R, Bontcheva K (2015) Analysis of named entity recognition and linking for tweets. Inf Process Manag 51(2):32–49

Gardner M, Grus J, Neumann M, Tafjord O, Dasigi P, Liu NF, Peters M, Schmitz M, Zettlemoyer LS (2017) AllenNLP: a deep semantic natural language processing platform. arXiv:1803.07640

Gillings M, Dayrell C (2023) Climate change in the UK press: Examining discourse fluctuation over time. Appl Linguistics. https://doi.org/10.1093/applin/amad007

Grimmer J (2010) A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. Polit Anal 18(1):1–35

Hamborg F, Donnay K, Merlo P et al (2021) NewsMTSC: a dataset for (multi-) target-dependent sentiment classification in political news articles. In: Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume, pp 1663–1675

Handler A, Denny M, Wallach H, O'Connor B. (2016). Bag of what? Simple noun phrase extraction for text analysis. In: Proceedings of the first workshop on NLP and computational social science, pp 114–124

Hoffman AJ (2015) How culture shapes the climate change debate. Stanford University Press

Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L (2019) Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools Appl 78:15169–15211

Liu M, Huang J (2022) "Climate change'' vs. "global warming'': a corpus-assisted discourse analysis of two popular terms in the New York Times. J World Lang 8(1):34–55

Maynard D, Bontcheva K (2015) Understanding climate change tweets: an open source toolkit for social media analysis. Enviroinfo and ict for sustainability 2015, pp 242–250

Mishra P, Mittal R (2021) NeuralNERE: neural named entity relationship extraction for end-to-end climate change knowledge graph construction. In: ICML 2021 workshop on tackling climate change with machine learning

OpenAI (2023) Gpt-4 technical report

Pak A, Paroubek P et al (2010) Twitter as a corpus for sentiment analysis and opinion mining. Lrec, vol 10, pp 1320–1326

Palmer M, Gildea D, Kingsbury P (2005) The proposition bank: an annotated corpus of semantic roles. Comput Linguist 31(1):71–106

Quinn KM, Monroe BL, Colaresi M, Crespin MH, Radev DR (2010) How to analyze political attention with minimal assumptions and costs. Am J Polit Sci 54(1):209–228

Rao D, McNamee P, Dredze M (2013) Entity linking: finding extracted entities in a knowledge base. Multi-source, multilingual information extraction and summarization, pp 93–115

Rebich-Hespanha S, Rice RE, Montello DR, Retzloff S, Tien S, Hespanha JP (2015) Image themes and frames in US print news stories about climate change. Environ Commun 9(4):491–519

Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on web search and data mining, pp 399–408

Schmitt X, Kubler S, Robert J, Papadakis M, LeTraon Y (2019) A replicable comparison study of NER software: StanfordNLP, NLTK, openNLP, SpaCy, Gate. In: 2019 sixth international conference on social networks analysis, management and security (SNAMS), pp 338–343

Shang J, Liu J, Jiang M, Ren X, Voss CR, Han J (2018) Automated phrase mining from massive text corpora. IEEE Trans Knowl Data Eng 30(10):1825–1837

Stede M, Patz R (2021) The climate change debate and natural language processing. In: Proceedings of the 1st workshop on nlp for positive impact, pp 8–18

Taufek TE, Nor NFM, Jaludin A, Tiun S, Choy LK (2021) Public perceptions on climate change: a sentiment analysis approach. GEMA Online J Lang Stud 21:4

Volkanovska E, Tan S, Duan C, Bartsch S, Stille W (2023) The insightsNet climate change corpus (ICCC) compiling a multimodal corpus of discourses in a multi-disciplinary domain. Datenbank-Spektrum 23:177–188