




# Framing climate change in *Nature* and *Science* editorials: applications of supervised and unsupervised text categorization

Manfred Stede<sup>1</sup> · Yannic Bracke<sup>1,2</sup> · Luka Borec<sup>1</sup> · Neele Charlotte Kinkel<sup>1</sup> · Maria Skeppstedt<sup>3</sup> 

Received: 31 May 2022 / Accepted: 16 February 2023 / Published online: 5 May 2023  
© The Author(s) 2023

## Abstract

Hulme et al. (Nat Clim Change, 8:515–521, 2018) manually coded ‘frames’ in 490 *Nature* and *Science* editorials (1966–2016) they found relevant for climate change. We produced a digital version of the corpus and conducted a set of experiments: We explored many variants of supervised categorization for automatically reproducing the manual frame coding, and we ran an interactive variant of topic modeling. In both approaches, we made use of word embedding techniques for representing text documents. Supervised classification yielded F1-scores of up to 0.91 (for the best category) and 0.68 overall, and it led to insights regarding the relation between ‘topic’ and ‘framing’. The topic modeling algorithm was able to reproduce central trends in the temporal analysis of framing that was presented by Hulme et al. based on their manual work.

**Keywords** Climate change communication · Framing · Text-as-data · Supervised classification · Topic modeling

## Introduction and research background

### Motivation

Text-as-data research (or ‘text mining’) is becoming more and more important in general for deriving insights from large text collections, and also specifically for studies on climate change communication. Of particular relevance is the analysis

---

✉ Maria Skeppstedt  
maria.skeppstedt@abm.uu.se

Extended author information available on the last page of the article

of the ‘framing’ that authors use in their exposition of the problem, whereby they consciously or subconsciously guide the reader in perspectivizing the issue at hand [15]. While the notion of framing is not used consistently in the literature (see Section “[The notion of framing](#)” below), the majority view is that of a set of different kinds of ‘challenges’ that climate change poses, such as ‘technological’ or ‘ethical’. Authors of studies then devise a set of such challenges in accordance with their purposes (e.g., [2, 4]). In recent years, first studies have appeared that apply text-as-data methods to automatically identify the framing used in text corpora on climate change (e.g., [43, 47]). Usually, the methods are applied straight “out of the box”, and the study then interprets their results. In contrast, our work aims at a systematic comparison of some relevant methods (variants of supervised and unsupervised text classification) for the automatic analysis of framing, to shed light on their relative performance and utility.

Specifically, we build a corpus of climate-change-related editorials from *Nature* and *Science*, as they have been identified in an earlier study by Hulme et al. [26]. In the archives of the two journals, these authors found some 500 editorials of topical relevance, and then undertook a completely manual frame coding based on the PDF files. Specifically, each text was labeled with one of eight ‘challenge’ labels, and then the authors showed the temporal development of frame prevalence and analyzed differences between the two journals.

The list of the editorials was published by Hulme et al. [26], and upon our request, the authors also made the frame labels available to us. Thus, we were able to build a machine-readable version of the corpus and, using the labels, we (i) run various experiments with supervised classification and (ii) compare the results of unsupervised topic modeling to the original labeling.<sup>1</sup> Therefore, in contrast to many previous studies on automatic frame analysis, we do not limit our study to the use of one type of approach, but explore the usefulness of methods at opposite ends of the supervised/unsupervised spectrum. In addition, we take on a research task that was originally designed as a purely manual one. We, thus, avoid the risk of biasing the choice of research task to something that can easily be solved by automatic methods, and instead evaluate the performance of automatic methods on a task that was originally designed by climate change experts and deemed useful by them.

In the following, we provide some more information on the original study by Hulme et al., discuss the notion of framing underlying this work, and then state our own research goals more precisely.

## The original study

Using keyword search in the journal archives, Hulme et al. [26] collected the *Nature* and *Science* editorials that in one way or another addressed the topic of climate change (henceforth: CC) between 1966 and 2016. Given the long time span, the

---

<sup>1</sup> In this paper, it is inevitable for the term “topic” to be used in two different senses. In Sects. 2.3, 3.2, 4.2 and in parts of Section 5 we use it in the narrow sense of technical topic modeling; in the rest of the paper, it appears in the generic, informal sense of ‘theme’.

keyword list was designed to also find "early mentions" of CC and thus consisted of these broad terms: 'climate', 'greenhouse', 'carbon', 'warming', 'weather', 'atmosphere', 'pollution'. Naturally, this search produced many wrong hits (e.g., texts that dealt with pollution of rivers), and thus a manual filtering step ensured that the corpus holds only texts that address the topic of CC. The result was a set of 333 editorials from *Nature* and 160 from *Science*. Then the authors undertook a manual coding of eight 'framing' categories that the editorials appeared to employ. Based on the results, they computed statistics and demonstrated certain correlations between frame use and climate-relevant external events, and they showed some interesting differences between the two journals. It is noteworthy that CC is not necessarily the main topic of a text in the corpus; often, the editorial deals with some other technical or scientific issue and mentions CC only as an aspect that is more or less connected to the main topic.

Defining the set of frames/challenges was not a preparatory one-shot process. In contrast, Hulme et al. proceeded in several iterations (combining induction and deduction) where the frame set was refined and re-applied to the data. According to the authors, each frame should meet the following four criteria: "identifiable conceptual and linguistic features; commonly observed; easily distinguished from other frames; recognizable by others" [26, Methods]. The eight frames that were eventually used for the final coding are (in slightly abbreviated form) reproduced in Table 1.

Coders were instructed to assign a single 'dominant' frame per text, and optionally any number of 'other' frames. Inter-coder agreement for primary frames was measured in a pilot exercise, yielding Fleiss'  $\kappa$  values of 0.32 for *Nature* and 0.39 for *Science*. While these results seem 'moderate', it is important to note that they were calculated on the basis of four annotators, an unusually high number. For instance, the 'media frames corpus' by [9] was in the final stages of the coding process annotated by merely two people, who achieved Krippendorff  $\alpha$  values ranging between 0.3 and 0.6. Furthermore, the  $\kappa$  result of Hulme et al.'s pilot prompted the authors to then follow a coding practice where an annotator would flag cases that he or she regarded as particular difficult; the resulting set of 46 texts (almost 10% of the corpus) were afterward resolved collaboratively by all authors. As a result, the quality of the final corpus can be assumed to be higher than the  $\kappa$  values suggested.<sup>2</sup>

Hulme et al. unfortunately did not report agreement values for the individual frame classes, as an indication of relative difficulty. We believe, however, that the degree of co-occurrence of 'dominant' and 'other' frames can give an indication of the per-frame difficulty of a consistent annotation. In Fig. 1, we, therefore, visualize the co-occurrences between those frames as annotated by Hulme et al. For instance, when the two frames SCI and COM were annotated as the dominant frame for an

<sup>2</sup> As a further remark, the  $\kappa$  values achieved here are not uncommon in annotation tasks involving subjective decisions (as in linguistic pragmatics). For example, a well-known study in computational argumentation analysis [23] reports a  $\kappa$  of 0.4, based on three annotators, for distinguishing the 'claim' from the 'premise' in an argument, using a corpus of US presidential candidate debates.

**Table 1** Types of frames used in the coding

Name of challenge	Characterization
ECON: Economical/financial	CC is an externality of economic growth and/or certain modes of production/consumption and/or requires improved quantification of costs/benefits of impacts and/or policies and/or can/should be tackled through economic and financial instruments
DEV: Development	CC is a by-product of pathways and patterns of socio-economic development and/or unequal development inhibits adequate mitigation, resilience and adaptation and/or causes uneven distribution of harms to human health, well-being and perceived human security
SEC: National/intern'l security	CC is a geopolitical security risk by introducing new dangers into inter- and intra-state relations and/or is a threat-multiplier requiring new forms of international or state-level security responses
ETH: Ethical/moral	CC raises important questions of procedural and/or distributive justice (for example, burden-sharing) and/or people have an ethical responsibility/moral duty toward future humanity and/or nature and/or the 'poor'/the most vulnerable and/or God/deities, to mitigate CC
TECH: Technological/energy	Fossil-fuel-based energy technologies are the root cause of CC and/or technological innovation and energy transitions that aim at reducing/capturing/sequestering GHG emissions and/or solar engineering technologies are essential to tackle climate change
GOV: Institutional/governance	Structural and institutional inertia/problems are a root cause of CC and/or tackling climate change requires new/improved governance institutions and/or regulatory management of adaptation/mitigation policies is inadequate [not to be used if this governance challenge is covered by a more specific frame]
SCI: Scientific	Scientific understanding of CC is incomplete/inadequate (that is, due to complexity/uncertainty) and/or investing in science is necessary for adequate mitigation/adaptation responses
COM: Communication	CC science/risks is/are poorly communicated to public audiences and/or media representations of CC are problematic/biased and/or deliberate misinformation/manufactured scepticism confuses political/public opinion

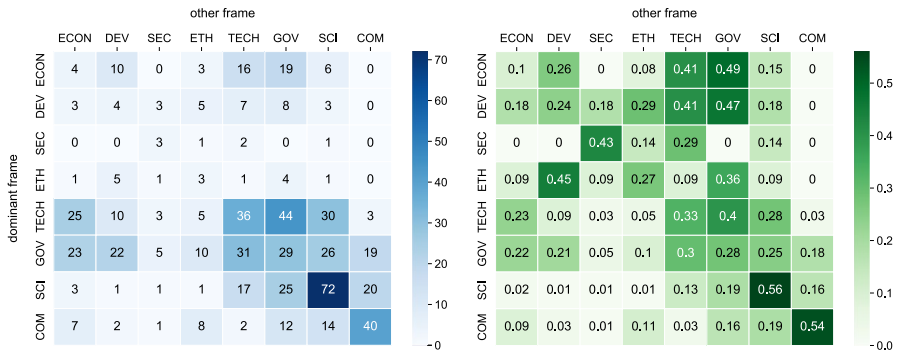
Abbreviation of Table 1 from Hulme et al. [26, p. 517]

editorial, secondary frames were only found in around half of the cases. For GOV and TECH, in contrast, this number is around 70%,

For our computational study, we used only the dominant frames as target for automatic classification and disregarded the other ones, except for one preliminary experiment that we will mention in Sect. 3.1. The distribution of the dominant frames in the corpus is shown in Table 2.

## The notion of framing

A systematic literature review by Badullovich et al. [2] showed that Social Science research on the framing of climate change overwhelmingly subscribes to the same notion of framing as used by Hulme et al., viz. as a set of generic perspectives from which CC is discussed. The sets used differ to some extent; three examples of



**Fig. 1** Co-occurrence of ‘dominant’ and ‘other’ frames. On the left: absolute number of co-occurrences (e.g., there are 14 instances for which COM was the dominant frame, and SCI another frame). On the right: proportion of co-occurrence in relation to the total number of instances of the primary frame (e.g., for 19% of COM as dominant frame, SCI is another frame). The diagonal shows the number of occurrences of the dominant frame when no other frame is present

**Table 2** Distribution of primary frames in the editorial corpus

Frame	# docs	% docs
ECON	39	7.96
DEV	17	3.47
SEC	7	1.43
ETH	11	2.24
TECH	109	22.24
GOV	105	21.43
SCI	129	26.33
COM	73	14.90

frames that were not considered by Hulme et al. are efficacy, general causes, and public health. In automatic analysis, this approach to framing as a set of deductively predefined classes (which can be treated as either ‘generic’ or ‘issue-specific’) lends itself to supervised classification (see Sect. “[Supervised classification](#)”); for example, early work by [7] used a corpus of several thousand news texts and four generic news frames: conflict, economic consequences, human interest, morality. They worked with a bag-of-words model and a support vector machine (SVM) classifier, as also done more recently by [43] in the domain of CC. These authors built an SVM classifier for four frames: economic cost, economic benefits, conservative and free-market ideology, uncertainty and risk.

An alternative approach to framing consists of inductively computing the categories by means of topic modeling (see Sect. “[Unsupervised topic modeling](#)”), as a combination of automatic computation and manual parameter tuning, filtering and arrangement. In the domain of CC, the work of [47] is a recent example; the author ends up with a set of 53 frames organized in a 3-level hierarchy, using a corpus of 1.700 press releases by various US organizations.

A third approach to computing frames, according to [22], is using lexicons (as done routinely for tasks such as sentiment analysis). [31], for example, present a multilingual dictionary construction process for identifying four ‘immigration’ frames in news text: economy/budget, labor market, welfare, and security.

We point out that these conceptions of framing, along with their computational treatment, are not uncontroversial. One line of criticism targets the inflationary use of the ‘framing’ term (which had already been lamented in the pioneering work of [15]) and its unfortunate conflation with the notions of persuasion, agenda setting, priming, and schemas/scripts [8]. Correspondingly, [38] advocated reserving ‘framing’ for the much more narrow concept of ‘equivalence framing’ that denotes the choice among alternative linguistic expressions for the same semantic content. Another important criticism concerns the surface-oriented conception (and computation) of frames as lexical items that are manifested in a text and easy to be identified—usually even without considering linear order, i.e., in a bag-of-words model. [8] posit that “interesting” frames are complex units of information that should not be reduced to a small set of word co-occurrences. A multi-step computational (but only semi-automatic) approach to identifying such complex concepts was recently proposed by [28], and a similar one specifically for framing by [22].

As mentioned in the beginning of this section, however, the vast majority of framing research in the domain of CC adopts the approach of defining generic ‘perspectives’ from which an author selects one when discussing the topic. Hulme et al. specifically speak of ‘types of challenges’, and we agree that this conception is valuable, in particular for the purpose of determining long-term trends of treating CC in the genre of scientific editorials. Therefore, we adopt this view and implement it with bag-of-words models, comparing the relative performance of many different configurations.

## Goals of the present study

As a prerequisite for automatic analyses, we build a machine-readable version of the corpus, which is the first contribution of our work. It enables us to conduct experiments with supervised and unsupervised classification. Both treat the text documents as bag of words; in other words, classification is based solely on the distribution of lexical items in the documents (though some of our approaches replace words by tokens and character n-grams and by pre-trained embeddings; see Sect. “[Supervised classification](#)”).

In supervised classification, the manually-coded labels serve as “gold standard”, which the classifier aims to reproduce.<sup>3</sup> There are many technical approaches to this, and as we show below, in extensive experiments, we first determined the best-performing configuration of various classification methods on a development set of the

---

<sup>3</sup> The term “gold standard” is well-established in the NLP community, but much less so in the Social Sciences. For a brief overview, see [48].

corpus, and then report results on a held-out test set. The important aspects of the configuration are:

- Text classification studies differ in the precise way of representing the text documents. We compared various ways of representing words (plain unigram tokens, n-grams, several word embedding methods) in combination with pre-processing steps (stopword removal, etc.) and term weighting (tf/idf).
- A fairly wide variety of classification algorithms is available in relevant code libraries, and we experiment with several widely-used ones.
- As Table 2 shows, the distribution of frame labels is highly imbalanced. This is known to be a great challenge for classifiers, and thus we specifically attend to certain methods that can be used to alleviate the problem.

Recall our remark on the different degrees of "topicality" of the editorials: They do not necessarily talk mainly about CC but instead may mention it only briefly. For any bag of words model, this is potentially problematic, as the complete document is being analyzed, while the CC frame may be recognizable only from a few words therein. While previous text classification work usually ignores this aspect,<sup>4</sup> we decided to ascertain the influence of "degree of topicality" on the classification results and then to experiment with restricting the classifier to work only on the most relevant portions of the document.

Unsupervised classification, in contrast, aims at clustering a document set into meaningful groups, without having access to any manually-assigned labels. There are a multitude of approaches, of which we here select only one, viz. non-negative matrix factorization, a variant of topic modeling.

As Grimmer et al. [21, p. 270] had remarked, "the most productive line of inquiry (...) is not in identifying how automated methods can obviate the need for researchers to read their text. Rather, the most productive line of inquiry is to identify the best way to use both humans and automated methods for analyzing texts." We conceive the contributions of this paper in this spirit: Supervised classification can assist the researcher in handling large volumes of data, e.g., by pre-filtering; but it is important to be aware of the differences between alternative approaches. Unsupervised methods, on the other hand, can automatically—or in interaction with the user—detect patterns in the data that invite researchers' interpretation and may lead to insights that have not explicitly been looked for.

Our aim, therefore, is not to "compete" with the manual study by Hulme et al. [26], but to use their frame coding as basis for a systematic investigation of automatic methods, under the specific conditions of a relatively small data set, a

<sup>4</sup> One exception: In their manual analysis of CC editorials from the Finnish press, [32] recognized the importance of topicality difference (which is *prima facie* invisible when texts are just retrieved by simple keyword search) and accounted for it in their analysis.

relatively large number of classes, and a heavily imbalanced class distribution. In summary, the central research questions tackled in our work are:

- To what extent can we reproduce the manual annotations (from Hulme et al.) with automatic supervised classification methods?
- To what extent do established methods for handling imbalanced data improve those classification results for our specific corpus?
- To what extent does reducing the texts to their most "climate-relevant" parts improve those classification results?
- To what extent can we reproduce the frame label set (as used by Hulme et al.) by unsupervised topic modeling?
- To what extent can we reproduce the temporal "trends" (as they were observed by Hulme et al.) of the frame annotations spanning the period 1966–2016?

## Methods

An overview of the experiments conducted is given in Fig. 2.

### Corpus construction

As indicated earlier, Hulme et al. worked by manually analyzing the PDF versions of the editorials. Accordingly, our first step is to map these PDFs to a machine-readable corpus in plain-text format. A spreadsheet with DOIs and some other meta-data for the editorials was made available along with the publication by Hulme et al. [26]. Upon our request, we were also given access to the frame codings done by those authors, in the form of a second spreadsheet.<sup>5</sup> For two of the *Nature* editorials, we did not find annotations in the table; furthermore, one of the *Science* editorials turned out unavailable for download. Therefore, the corpus for our study is missing 3 editorials, and thus has a total of 490 documents.

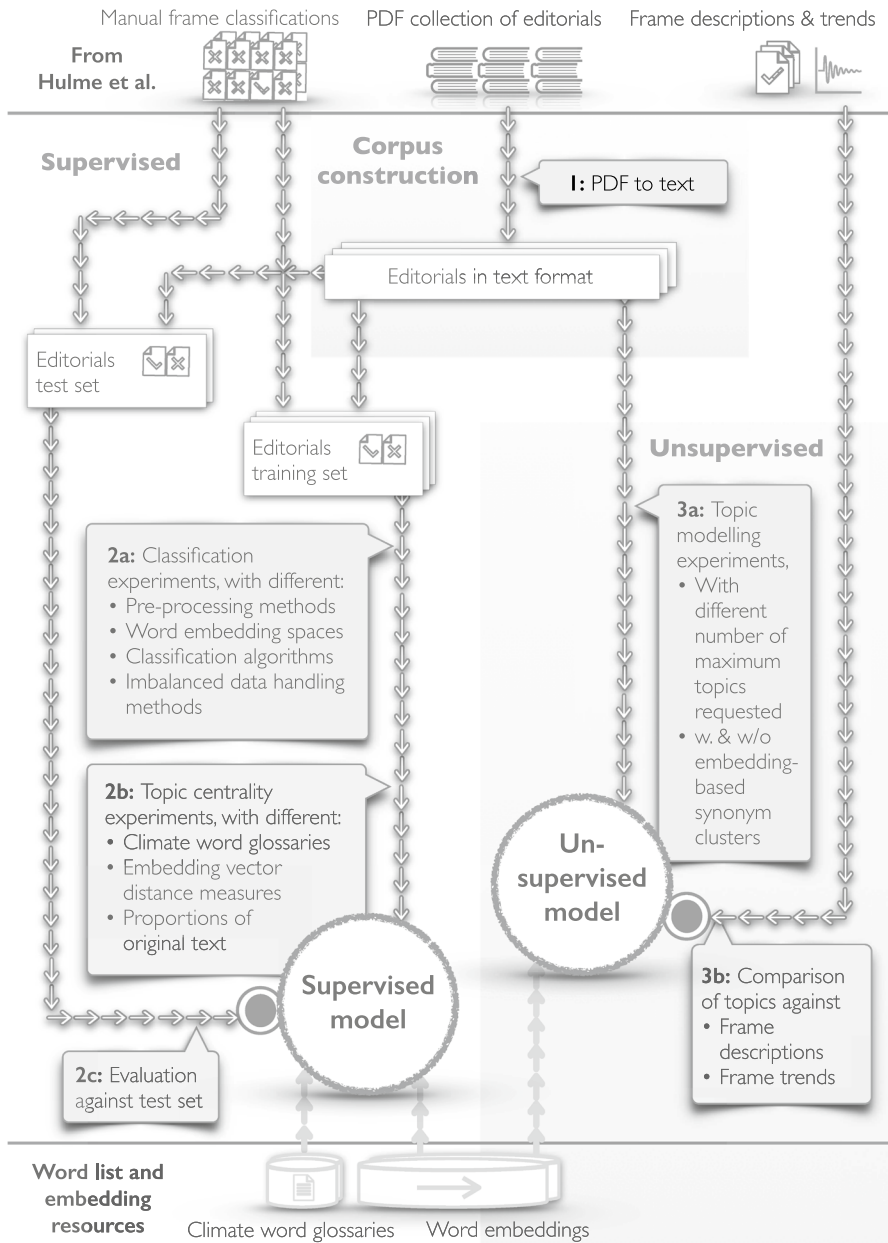
Using DOIs, we retrieved all PDF files of the editorials from the journal archives. To enable automatic processing, OCR software<sup>6</sup> was used for the older PDFs (which do not include embedded text). We manually checked every resulting plain text file for accurate conversion of double column text and rejoining of hyphenations. If obvious OCR errors were noticed within the texts, they were corrected, but we did not undertake a complete word-by-word reading.<sup>7</sup> Also, in this step, we ensured a common format of the plain text files: each one starts with a title line followed by a subtitle line (possibly empty) and an author line (possibly empty). In the actual text, figures and tables are replaced by placeholders ("`<figure>`", "`<table>`"). Paragraph

<sup>5</sup> We are grateful to the authors for making the coding available, which was a prerequisite for our own work to be carried out.

<sup>6</sup> <https://convertio.co/>.

<sup>7</sup> All our subsequent manual inspections of the texts revealed that only few typos are present, mostly involving special characters.





**Fig. 2** The experiments conducted. First, a corpus is constructed; then experiments with supervised classification are carried out (shown on the left), as well as with unsupervised classification (shown to the right)

**Table 3** Number and length (in tokens) of documents in the *Nature* and *Science* subcorpora

Corpus	# Docs	Doc length (mean)	STD
<i>Nature</i>	331	841.4	290.7
<i>Science</i>	159	764.0	66.7
Total	490	816.3	244.6

breaks are preserved. Our corpus, henceforth, called the NatSciEdCC corpus, thus consists of 490 plain text files, without any additional markup. (Metadata, as pointed out above, resides in the spreadsheet.) Table 3 provides information about the lengths of the texts in the two parts of the corpus.

### Supervised classification

We approach the task of automatically reconstructing the frame coding of the corpus texts with the established technology for supervised topic classification, thus treating the eight frame categories as conceptually corresponding to topics (cf. [43]). This correspondence does by no means hold in general [45], but it is warranted by the type of coding done by Hulme et al. and in many similar studies (cf. [2]).

In formal terms, the text classification problem consists in learning a function  $\phi : D \rightarrow C$ , where  $D$  is a set of  $n$  text documents  $d_1, \dots, d_n$  and  $C$  is a set of  $m$  categories  $c_1, \dots, c_m$  (cf. [39]). In our case, the categories are the frame labels from Table 1: ECON, DEV, SEC, ETH, TECH, GOV, SCI or COM. This function constitutes the model, and the program applying the model to texts is the classifier. It learns its model from a part of the labeled corpus, the development set or training set. Then it predicts the labels of the documents for the smaller part of the corpus, the test set. The results obtained on the test set indicate to what extent the classifier is able to generalize from the training data to unseen data.

The main choices to be made in designing a classification system are the precise way of representing the text documents as data points, and the selection of a classification algorithm. Furthermore, when the frequencies of categories in the data are highly imbalanced, various methods for alleviating this problem can be applied. We will shortly explain our choices of the three parameters, and motivate the evaluation measure we use for assessing classification performance. Afterward, we will explain our method for focusing the classification on topic-relevant portions of the text documents instead of "blindly" using the whole text.

To be able to test variants of the parameters on the one hand, and make overall representative predictions on the NatSciCC corpus on the other hand, we proceeded in two steps. We conducted a large set of experiments with many parameter combinations on the development set, for which we took a sample of 80% of the data (392 documents). We determined the best parameter setting via fourfold cross-validation (CV). The best classifiers were then retrained on the complete development set, and finally we applied them to the 20% test set (98 documents), thus producing our final results for NatSciCC.<sup>8</sup>

<sup>8</sup> Both development and test set were stratified random samples, to work with the same class distributions in the two parts.

**Table 4** Properties of the pre-trained embedding sets

Name	ML method	Data set (size in tokens)	Vocab
w2v	word2vec	GoogleNews data set (100 B)	929 K
GloVe-big	GloVe	Common Crawl (840 B)	22 M
GloVe-small	GloVe	Wikipedia 2014/Gigaword 5 (6 B)	400 K

The classifiers were implemented with Python3.6, relying on libraries for NLP and ML, especially on `scikit-learn` [35].

## Document representation and pre-processing

We experiment with two approaches representing text documents as sets of minimal units. The first is the conventional ‘bag of words’ (BOW) model where unigram tokens are the minimal units, whose frequency is represented in vector space. In preparation, different kinds of pre-processing can be applied. We experimented with (i) stop word removal using the `spaCy` library,<sup>9</sup> (ii) lowercasing all words, (iii) stemming using `SnowballStemmer` from the `nlTK` library,<sup>10</sup> (iv) lemmatization using `spaCy`. As variants of the BOW model, we also used token  $n$ -grams ( $n \in \{1, 2, 3, 4\}$ ) and combinations, such as  $n = 2 / n = 3$ ) and character  $n$  grams ( $n \in \{2, 3, 4, 5, 6\}$  and combinations). For all versions, we also tested the influence of  $tf/idf$  weighting.<sup>11</sup>

The second approach is aggregated word embeddings. An embedding is a low-dimensional vector representation of a word that captures its meaning to some extent (see, e.g., Goldberg [20, 115–134]). Words with similar meaning are, thus, represented by vectors pointing in a similar direction in latent semantic space. Hence, a representation of a text document can be obtained by aggregating the embeddings of all words in the document to a single vector. An intuitive aggregation method is counting the words in the document (as in the BOW approach) and then averaging over their embedding vectors. The  $tf/idf$  weighting can be employed here as well (see Boom et al. [5] for these and a range of other aggregating methods we experimented with).

In our experiments, we employed 3 sets of pre-trained embeddings with vectors of 300 dimensions. The sets are provided by Google<sup>12</sup> and by Stanford University<sup>13</sup> and differ (i) in terms of the ML algorithm that was used to obtain the embeddings (word2vec [34] vs. GloVe [36]); (ii) in terms of the data set they were learned from; and (iii) in terms of vocabulary size (see Table 4).

<sup>9</sup> <https://spacy.io>.

<sup>10</sup> <https://www.nltk.org>.

<sup>11</sup>  $tf/idf$  weighting means that for each term in a document, the term frequency ( $tf$ ) is normalized by its inverse document frequency:  $idf_t = \log \frac{N}{df_t}$  where  $df_t$  is the number of documents in which  $t$  occurs.

<sup>12</sup> <https://code.google.com/archive/p/word2vec/>.

<sup>13</sup> <https://nlp.stanford.edu/projects/glove/>.

## Classification algorithms

For learning the classification model, the following standard techniques were employed and their performance was compared [1, 39]: Multinomial naïve Bayes, support vector machines (SVM), logistic regression (LR), single-layer perceptrons, decision trees, random forest, and AdaBoost. For completeness, we also ran some experiments with modern transformer architectures (BERT, DistilBERT) without being very optimistic, as the corpus is quite small, and these approaches are known to require large training sets.

## Handling imbalanced data

Given the aforementioned class imbalance in our corpus, we decided to also experiment with specific approaches that have been developed to tackle such situations. Generally, classifiers aim to maximize the overall classification accuracy. If classes are difficult to separate (which is to be expected in our eight-category problem), they can become biased toward the majority class(es), and thus rarely predict the minority class(es) at all. ([27, 40]; [25, 1264]; [17, 21]). An additional difficulty is absolute rarity of one or more classes in the data set ([46, 8–9]), as it is hard for a classifier to generalize from very little training data. As Table 2 shows, this is also a concern in our corpus.

We experimented with the most prominent methods for tailoring the classifier to such situations: oversampling and undersampling (e.g., [18]), dimensionality reduction (e.g., [13]), cost-sensitive learning (e.g., [17, 63–78]), and boosting (e.g., Freund and Schapire [19]).

The most successful of these techniques, oversampling, is the process of adding more minority class samples to the training set prior to training the classifier. This can be done either by randomly duplicating samples in the set (random oversampling) or by generating new, synthetic samples (e.g., SMOTE [10], ADASYN [24]). In the latter case, a new minority class sample is generated by copying an existing sample and shifting this copy in the direction of one of its  $k$ -nearest neighbors from the same class (classic SMOTE). Variants of this approach (such as ADASYN) generate more synthetic data from samples that are surrounded by many majority class samples than from samples that lie in a bulk of minority class samples, assuming that the former are more difficult to learn and, hence, additional samples are more helpful. We experimented with several oversampling techniques and varying oversampling ratios (i.e., the ratio between minority and majority class after oversampling).

## Influence of topic centrality

The final aspect of the classification task that we explored stems from the observation (see above) that the editorials address the CC topic to different degrees. We, thus, investigated these questions: (i) Are documents whose main topic is CC easier

to categorize for their CC frame than documents that mainly concentrate on other topics? (ii) If so, can we improve classification results on the complete corpus by shortening the documents to their "CC-central" parts, i.e., by reducing noise?

For step (i), we need to check whether classification performance on the texts correlates with their degree of CC-centrality. We estimate CC-centrality by comparing a representation of the semantics of common CC words to a representation of the semantics of the corpus texts; to capture semantics, we again use word embeddings (specifically GloVe-small, see Table 4). For "common CC words", we compare three different sources: a Wikipedia CC glossary<sup>14</sup>, the seven climate-related terms that Hulme et al. [26] had used to retrieve the *Nature* and *Science* editorials, and 20 keywords that we extracted from a corpus of climate-change-related articles from the New York Times. As the GloVe-small embeddings cover unigrams only, we split any compound terms from the Wikipedia glossary into unigrams and then removed the climate-unrelated unigrams from the final set. All words were then aggregated into a vector representation by averaging over their corresponding embedding vectors. We call the results CC-vector<sub>wiki</sub>, CC-vector<sub>hulme</sub>, and CC-vector<sub>nyt</sub>.

For each document in the corpus, we represented every sentence in exactly the same way as just described for the CC term list, i.e., by averaging over the GloVe-small embeddings of the words in the sentence. Next, for each sentence embedding, we computed its cosine distance from the CC vector, which we regard as a score indicating its topic relevance. Finally, the sentence scores for a document were again averaged, which yields our document score for CC-centrality for every editorial.

For (ii), we computed the topic centrality of individual sentences, then reranked the sentences of a text according to those scores. For this step, we employ two different methods: (i) ranking according to cosine distance and (ii) ranking according to Word Mover's distance [29]. (This is the "minimum travelling distance" that the embedded words of one document need to undergo to reach the embedded words of the other document.)

We produced three different versions of shortened texts by taking the top 33%, 50%, and 67% of their most cc-central sentences. The quality of our rankings is assessed by utilizing them as training data for our text classifier, hereby comparing the CC-excerpts to same-length excerpts obtained by mere random sampling.

In short, we compare the performance of three methods of list construction and two methods of computing the distance between embedding vectors, for obtaining sentence rankings for the editorials in the corpus.

## Evaluation measures for classification

Following common practice (see, e.g., [39]), we evaluate the performance of our classifier by measuring precision  $\pi$  ( $tp/(tp + fp)$ ) and recall  $\rho$  ( $tp/(tp + fn)$ ) and subsequently forming a harmonic mean ( $F$ ) of the two.<sup>15</sup> As we are dealing with a multi-class problem,  $\pi$  and  $\rho$  can be computed per class. To choose between

<sup>14</sup> [https://en.wikipedia.org/wiki/Glossary\\_of\\_climate\\_change](https://en.wikipedia.org/wiki/Glossary_of_climate_change), last access: September 15th, 2020.

<sup>15</sup> tp = true positive, fp = false positive, tn = true negative, and fn = false negative.

micro-averaging (every instance of a classification result has the same weight) and macro-averaging (every class has the same weight), we again note that our dataset is heavily imbalanced, and since we wish to account for performance on small classes, too, we opt for weighted macro-averaging as our main evaluation measure, which multiplies every individual class- $F_1$  with its proportion of occurrence in the corpus. However, for comparison, we will also report the plain macro-average when presenting our results.

## Unsupervised topic modeling

In addition to testing supervised classification on the manually-annotated corpus, we also explored an unsupervised method for extracting frequently occurring content in the data, whose results can then be compared to the manual annotations. Specifically, we studied whether some of the insights on the development of editorial subjects over time—which was one central outcome of the original analysis by [26]—can also be gained by automatic means.

A frequently used unsupervised method for text mining is topic modeling (TM) [6]. Here, each topic is represented by (i) a ranked list of words associated with the topic, and (ii) a ranked list of associated texts, i.e., texts in which the topic's associated words frequently co-occur. Collections containing short texts, as well as those containing few texts, thereby, present a challenge to the standard TM algorithm, since their word co-occurrence patterns more easily suffer from a sparsity problem. This problem could, e.g., be tackled by applying unsupervised methods specifically developed for short texts [12, 49], and/or to incorporate information from larger corpora, e.g., using word embeddings pre-trained on large corpora [11].

Given the relatively small size of our corpus (due to it containing few, rather than very short, texts), we decided to use a TM tool that could incorporate information from larger corpora. We chose Topics2Themes [42], which is based on the randomised TM algorithm NMF (non-negative matrix factorisation) [30] and which incorporates information from word embeddings in the form of pre-trained word2vec-vectors.<sup>16</sup> The tool has previously been used for finding recurring information in text collections containing texts shorter than those studied here, e.g., in collections of discussion forum posts [42], micro blogs [40] and folk legends [41].

Besides a graphical interface for exploring the most frequently occurring topics of a text collection, the tool also provides several options for influencing and configuring the TM algorithm. For instance, the user can provide the tool with a stop word list and a list of multi-word terms. The stop word list prevents the TM algorithm from creating topics based on frequent co-occurrences of uninteresting words, while the multi-word list can make sure that expressions like “New York” are treated as a single term rather than two.

Information from the pre-trained word2vec space is incorporated in the tool's text pre-processing step. In this step, words are organized into groups based on

---

<sup>16</sup> The code for the tool is freely available at GitHub: <https://github.com/mariask2/topics2themes>.

**Table 5** Selected classifiers: configurations and average  $F_1$  score during CV

Name	Features	Preproc	Model	Oversampling	$F_1$
BNA	BOW: 1–4 gs, tf-idf	Stop words, lower-case, lemma	NB	ADASYN (1:1)	0.623
ELR	Emb-agg: w2v, idf-mean	Stop words	LR	Random over-sampling (2:3)	0.637
ESS	Emb-agg: GloVe-big, mean	Lowercase	SVM	SMOTE (2:3)	0.630
Base	BOW: 1 gs	None	NB	None	0.538

their semantic similarity, which makes it possible for the TM algorithm to treat them as a single concept. This is achieved through an automatic clustering [16] of the word2vec vectors corresponding to the words in the text collection [33]. The user can manually correct the clusters, by removing words from them, or by supplying a list of manually constructed word clusters.

Since the NMF topic modeling algorithm is a randomized algorithm, it returns slightly different outputs each time it is run. Topics2Themes, therefore, provides functionality for obtaining more consistent results, by simply obtaining a large sample of TM outputs for the text collection (an approach used previously by, e.g., Baumer et al. [3]). The tool re-runs the TM algorithm many times, each time requesting it to extract  $n$  topics. The topic set extracted from the 10% most typical outputs of the TM re-runs (determined by the extent to which the topic-words of the output overlaps with the total set of topic-words from all re-runs) is then inspected by the tool. Only the  $r$  topics that occur in all of these typical TM outputs are retained in the final result.

We configured Topics2Themes to run the TM algorithm 500 times, i.e., to require a topic to occur in the 50 most typical re-runs for it to be retained. We also configured the tool to use a very wide criterion for determining topic equivalence when comparing re-runs; if at least half of the words associated with a topic output overlapped, the two outputs were classified as the same topic.

With the aim of finding a configuration that would retrieve the maximum amount of stable topics, we performed experiments by varying  $n$ . For each  $n$ , we recorded  $r$ , the number of stable topics retrieved. To investigate the effect of incorporating word2vec-vectors, we performed the same experiment, but excluded the word2vec-based pre-processing step.

## Results

We present first our results on the three steps of supervised classification: determining the best classifier configurations on the development set, producing the results on the test set, and determining the influence of reducing text

**Table 6**  $F_1$  scores (by class and overall) on the test set

	Baseline	BNA	ELR	ESS	# docs
ECON	0	0.353	0.333	<b>0.4</b>	8
DEV	0	<b>0.286</b>	0	0	3
SEC	0	0	0	0	1
ETH	0	0	0	0	2
TECH	0.851	<b>0.909</b>	0.889	<b>0.909</b>	22
GOV	0.545	0.545	0.585	<b>0.619</b>	21
SCI	0.717	0.632	0.627	<b>0.741</b>	26
COM	0.741	<b>0.774</b>	0.588	0.765	15
Macro	0.357	<b>0.437</b>	0.378	0.429	98
Weighted	0.612	0.645	0.609	<b>0.683</b>	98

Results for the best-performing models are marked with boldface

to topic-central portions. Afterward, we show the results for unsupervised topic modeling.

## Supervised classification

### Determining the best classifiers

We conducted fourfold CV experiments on the development set of the corpus (392 texts), using all combinations of the parameters discussed in Sects. “[Document representation and pre-processing](#)”, “[Classification algorithms](#)” and “[Handling imbalanced data](#)”. The three classifier configurations—in terms of pre-processing, document representation (features), classification model, and oversampling method – that performed best are shown in Table 5 with their average  $F_1$  score from the CV experiments. In addition to the three, we show a fourth classifier as a simple baseline model for comparison.

BNA<sup>17</sup> is a BOW-based classifier that uses unigrams to quadgrams of tokens and employs a Naïve Bayes algorithm using ADASYN oversampling technique with a ratio of 1:1. ELR and ESS are both embedding-based classifiers. One employs logistic regression and random oversampling, the other a support vector machine and SMOTE. They also obtain the embedding aggregations differently in terms of the pre-trained word embedding sets and in terms of the aggregation method (see column Features). ELR achieved the highest  $F_1$  score of all classifiers in CV. ESS came close to this and achieved the best score of a GloVe embedding-based classifier.

Baseline, like BNA, is a BOW-based Naïve Bayes classifier, but it uses only unigrams, no additional pre-processing or term weighting and no oversampling method. As expected, it performed substantially worse than the others.

<sup>17</sup> The acronyms we use for the classifiers indicate their composite parts/methods.



**Table 7** Mean topic centrality (TC) scores (distances from climate change vector) of correctly and incorrectly classified texts, and test statistics

Classifier	TC for texts classif. correctly	TC for texts clas- sif. incorrectly	<i>p</i>	<i>t</i>
Base	0.766	0.791	0.006	– 2.832
BNA	0.769	0.786	0.067	– 1.851
ELR	0.767	0.788	0.021	– 2.342
ESS	0.768	0.791	0.017	– 2.418

## Results for the test set

After the CV experiments, the four selected classifiers were retrained on all the available training data and evaluated on the test set (the yet unseen 20% of the corpus). In Table 6, we report the results of the evaluation. For each classifier, the table shows class-specific  $F_1$  scores, as well as the weighted average  $F_1$  score and the macro-averaged  $F_1$  score for the entire test set. The rightmost column displays the total number of documents in the test set for each class.

All classifiers fail to correctly predict any texts from the rare classes SEC and ETH, resulting in  $F_1$  scores of zero. With the exception of BNA, this is also true of the class DEV. Still, one should note that this is the result of only a few incorrect predictions, as the set contains only three documents or less for either class. Baseline also fails to predict the next largest class, ECON, while the three more complex classifiers achieve at least modest results here. The performance is substantially better for the four most common classes. It is consistently best for TECH with  $F_1$  scores around 0.9. The lowest  $F_1$  scores among the larger classes are achieved for GOV, with a maximum of 0.62 for ESS. The performance for classes SCI and COM varies somewhat between classifiers but always lies between that of GOV and TECH.

Overall, ESS outperforms the other classifiers in terms of weighted average  $F_1$  with a score of 0.68. BNA achieves a slightly better macro  $F_1$  score (0.44) than ESS, which can be attributed mainly to the fact that it is the only classifier managing to correctly predict at least one document from the DEV class. ELR performs surprisingly poor on the test set achieving a lower weighted average  $F_1$  than baseline (though a slightly better macro  $F_1$ ).

Contemporary transformer architectures for text classification are known to require relatively large training data sets, but nonetheless we were curious to see their performance on our corpus. We performed several experiments with BERT [14] and DistilBERT [37] on the original train/test split, as well as on the oversampled training set. In addition, we tested the effect of freezing the transformer layers as opposed to training them (which leads to rapid overfitting on the test set). After checking several dozen hyperparameter combinations for each experiment, the best results we reached are a macro- $F_1$  of 0.35 and a weighted  $F_1$  of 0.47 (with DistilBERT), which are well below the results reported above.

**Table 8** Results of text classification for sentence extracts from the editorials

Length	Hulme et al. keywords word mover's distance		Baseline random ranking	
	Weighted F1	Macro F1	Weighted F1	Macro F1
33%	0.67	0.53	0.64	0.44
50%	<b>0.68</b>	<b>0.58</b>	0.64	0.42
67%	0.63	0.46	0.64	<b>0.46</b>

Results for the best-performing models are marked with boldface

As a preliminary experiment on taking the annotated "secondary frames" into consideration, we compared classification performance for documents that have only one frame (in the annotations by Hulme et al.) to those with more than one. For single-frame texts,  $F_1$  is better, with the difference ranging between 0.054 and 0.021 for our different classifiers.

### Results of topic centrality experiments

*Classification performance on more or less topic-central texts:* Returning to the test set of our classifier, we now tested whether texts that were classified correctly had lower CC topic centrality scores than texts that were misclassified. Table 7 shows the mean CC topic centrality score for correctly and incorrectly classified documents. Indeed, for all classifiers, the correctly classified documents have lower scores. That is, CC is more central in these documents than in incorrectly classified documents. An independent t test shows that the differences between correctly and incorrectly classified documents are significant at the  $p = 0.05$  level for baseline, ELR, and ESS (see Table 7).

*Classification of topic-central text extracts:* Given the previous result, we investigated whether a reduction of the texts to their CC-central portions would improve classification performance. Our first parameter is the choice of term list for building the "climate vector". Here, CC-vector<sub>hulme</sub> yielded the best results. As for the second parameter, the comparison of cosine distance and Word Mover's Distance, we found that the latter performs better. The results for this best-performing combination of methods are presented in Table 8, in comparison to the baseline of a random sentence ranking. The 50% versions performed best with a weighted macro- $F_1$  of 0.68 and unweighted macro- $F_1$  of 0.58. The 33% versions performed slightly worse, but for both these length versions, results are higher than the baseline. Interestingly, for the 67% version, the classifier does not beat the random baseline.

Comparing these results on shortened texts to those for the complete texts (as shown above in Table 6), we notice that the 50% versions are clearly beating the original length versions in terms of unweighted macro- $F_1$  (0.58 versus 0.44), and replicating the result for the best weighted macro- $F_1$  (0.68). Notice, however, that the best results for weighted/non-weighted in Table 6 were achieved by two different

<p><b>Taxation as a way to tackle climate change, and problems the taxes may cause.</b></p> <p>"Half-tax on US energy"</p> <p>"Muddled carbon tax"</p> <p>"A Tax on Sin: The Six-Cylinder ..."</p> <p>"America's carbon compromise"</p> <p>"The motor-car as black sheep"</p> <ul style="list-style-type: none"> <li>tax / revenue</li> <li>car / trucks / vehicle / aut...</li> <li>british / united kingdom</li> <li>petrol / diesel / gasoline</li> <li>cost</li> <li>contribute</li> <li>consumption</li> <li>fuel</li> <li>extra</li> </ul> <p>Econ</p>	<p><b>Use of nuclear power could be a way to tackle climate change (but problems need to be solved, e.g. related to waste, security etc.)</b></p> <p>"Making nuclear power 'usable again'"</p> <p>"Uranium bites dust"</p> <p>"Factors Favoring Nuclear Power"</p> <p>"Waiting in the wings"</p> <p>"Recycling the past"</p> <ul style="list-style-type: none"> <li>atomic / reactor / nuclear</li> <li>atomic power</li> <li>waste</li> <li>british / united kingdom</li> <li>fuel</li> <li>invest</li> <li>safety</li> <li>government</li> <li>fast</li> <li>power</li> </ul> <p>Tech 1</p>	<p><b>Investments and research in energy is needed, particularly relating to renewable energy, and efficient distribution and use.</b></p> <p>"Energy hit"</p> <p>"Germany's Energy Research Plan"</p> <p>"Our emperors have no clothes"</p> <p>"The G8 on Energy: Too Little"</p> <p>"Urgent but balanced"</p> <ul style="list-style-type: none"> <li>energy</li> <li>nuclear</li> <li>electricity</li> <li>power</li> <li>wind</li> <li>power</li> <li>invest</li> <li>technology</li> <li>market</li> <li>china</li> <li>cost</li> </ul> <p>Tech 2</p>	<p><b>Carbon capture and storage to tackle greenhouse gas emissions.</b></p> <p>"Don't grandfather Coal plants"</p> <p>"Carbon capture and sequestration"</p> <p>"The shale revolution"</p> <p>"Going underground"</p> <p>"No magic fix for carbon"</p> <ul style="list-style-type: none"> <li>coal</li> <li>CO<sub>2</sub> / carbon dioxide</li> <li>capture</li> <li>natural gas</li> <li>emission</li> <li>oil</li> <li>petroleum</li> <li>fired</li> <li>oil / petroleum</li> <li>electricity</li> </ul> <p>Tech 3</p>	<p><b>Use of biofuel to reduce greenhouse gas emissions.</b></p> <p>"Renewable liquid fuels"</p> <p>"Getting serious about biofuels"</p> <p>"The cleaner state"</p> <p>"The Biofuels Conundrum"</p> <p>"The H.R. 776 Energy Legislation"</p> <ul style="list-style-type: none"> <li>fuel</li> <li>biofuel</li> <li>petroleum</li> <li>burning</li> <li>corn / wheat / grain</li> <li>output / production</li> <li>transport</li> <li>oil trucks / vehicle / aut...</li> <li>food</li> <li>convert</li> <li>petrol / diesel / gasoline</li> </ul> <p>Tech 4</p>	<p><b>Decision makers don't agree with mainstream facts within climate science, and don't take scientific advice when making decisions.</b></p> <p>"Senate vs science"</p> <p>"Bush's science flashpoints"</p> <p>"More than hot air"</p> <p>"Problems with the president"</p> <ul style="list-style-type: none"> <li>administration</li> <li>republican</li> <li>obama</li> <li>science</li> <li>us / united states</li> <li>bush</li> <li>president</li> <li>law / rule / guidelines / re...</li> <li>agency</li> </ul> <p>Gov 1</p>	<p><b>Possibilities and problems related to international summits, conferences &amp; treaties.</b></p> <p>"The heat is on"</p> <p>"Climate of compromise"</p> <p>"Climate change"</p> <p>"Agree to agree"</p> <ul style="list-style-type: none"> <li>emission</li> <li>treaty / protocol / agreement...</li> <li>nations / countries</li> <li>pledge / promise / o...</li> <li>vowed / pledge / promise / o...</li> <li>summit / meeting / conference...</li> <li>target</li> <li>china</li> <li>climate</li> <li>negotiate</li> </ul> <p>Gov 2</p>	<p><b>On the Rio de Janeiro Earth summit, e.g. expectations too high and unequal development of countries.</b></p> <p>"Earth Summit raises expectations"</p> <p>"Dangers of appointment at Rio"</p> <p>"The Rio Summit: A Reality Check"</p> <p>"Environmental protection or imp..."</p> <p>"Second chances"</p> <ul style="list-style-type: none"> <li>rio</li> <li>summit / meeting / conference...</li> <li>treaty / protocol / agreement...</li> <li>poor / impoverished</li> <li>convention</li> <li>sustainable development</li> <li>development / planet</li> <li>government</li> </ul> <p>Gov 3</p>	<p><b>Intergovernmental panel on climate change, e.g. mistakes in their communication, what/who should be their task/leader.</b></p> <p>"Shot with its own gun"</p> <p>"The panel must maintain its rigor"</p> <p>"The final frontier"</p> <p>"Rising to the climate challenge"</p> <p>"Maintaining the climate consensus"</p> <ul style="list-style-type: none"> <li>ipcc / intergovernmental pan...</li> <li>assessment</li> <li>report</li> <li>mitigate</li> <li>adaptation</li> <li>policy</li> <li>prices</li> <li>debates</li> <li>debated / review / explores</li> <li>science / research</li> </ul> <p>Gov* 4</p>	<p><b>General about the importance of science. (Climate change often just mentioned as an example.)</b></p> <p>"Public engagement with Science"</p> <p>"Science fiction meets reality"</p> <p>"Policy needs science"</p> <p>"Earth System Research priorities"</p> <p>"A critical vote down under"</p> <ul style="list-style-type: none"> <li>science / research</li> <li>fund / money / grant</li> <li>scholar / scientist / resear...</li> <li>policy</li> <li>technology</li> <li>community</li> <li>budget</li> <li>government</li> <li>national</li> </ul> <p>Sci 1</p>	<p><b>Deforestation and biodiversity, e.g. that clearing of forests leads to greenhouse gas emissions.</b></p> <p>"Forest Research for the 21st Ce..."</p> <p>"The entangled bank unravels"</p> <p>"On the road to REDD"</p> <p>"Defend the amazon"</p> <p>"Ignorance is not bliss"</p> <ul style="list-style-type: none"> <li>deforestation</li> <li>biodiversity</li> <li>species</li> <li>conserve</li> <li>deforestation</li> <li>ecosystem</li> <li>loss</li> <li>project</li> <li>tropics</li> <li>brazil</li> </ul> <p>Other 2</p>	<p><b>Reports on possible effects of climate change and debates on whether they pose a problem.</b></p> <p>"Climate Change and Climate Scie..."</p> <p>"Costs and benefits of carbon"</p> <p>"Where next with the greenhouse?"</p> <p>"The great greenhouse scare"</p> <p>"Energy and Climate"</p> <ul style="list-style-type: none"> <li>CO<sub>2</sub> / carbon dioxide</li> <li>temperature</li> <li>impacts / effects</li> <li>increase</li> <li>sea / ocean / waters</li> <li>water</li> <li>model</li> <li>foresee / expects / assesses...</li> <li>future</li> </ul> <p>Other 1</p>	<p><b>Mistakes that might reduce trust in science. Scientists lack media training. More modes of climate communication are needed.</b></p> <p>"Closing the Climategate"</p> <p>"Media studies for scientists"</p> <p>"Climate of fear"</p> <p>"A question of trust"</p> <p>"Welcome climate bloggers"</p> <ul style="list-style-type: none"> <li>media / scientists / resear...</li> <li>public</li> <li>skeptics / sceptics / denial...</li> <li>science / research</li> <li>journalist</li> <li>paper / article</li> <li>university</li> <li>coverage</li> <li>tale / story</li> </ul> <p>Com</p>	<p><b>Importance of climate research on the Arctic region (as it's warming faster than other regions and as, e.g., melting ice has large global effects).</b></p> <p>"Coming in from the cold"</p> <p>"One arctic"</p> <p>"The Arctic: A Key to World Clim..."</p> <p>"Governance of both poles"</p> <p>"The way ahead for polar science"</p> <ul style="list-style-type: none"> <li>arctic / polar / research</li> <li>ice</li> <li>area / region</li> <li>antarctic</li> <li>research</li> <li>council</li> <li>sea / ocean / waters</li> <li>coast</li> </ul> <p>Sci 3</p>	<p><b>Climate change and its effects need to be monitored, i.e. more data needs to be collected as a basis for climate research.</b></p> <p>"Taking the pulse of the oceans"</p> <p>"The changing oceans"</p> <p>"Oceanography from space"</p> <p>"Better climate data required"</p> <ul style="list-style-type: none"> <li>ocean / waters</li> <li>observe</li> <li>satellite</li> <li>globe / earth / planet</li> <li>monitor</li> <li>weather</li> <li>nasa</li> <li>space</li> <li>canada</li> <li>mission</li> <li>launch</li> </ul> <p>Sci 2</p>
--	---	---	--	--	--	---	---	--	--	---	--	---	---	--

**Fig. 3** The 15 generated topics with the titles for the 5 most typical editorial texts for each topic and the 10 most representative terms

**Table 9** The number of stable topics,  $r$ , returned from Topics2Themes when requesting the NMF algorithm to extract  $n$  topics and only retain those that were stably extracted over several re-runs. The same experiment was also performed for a TM that did not use word2vec-based word clusters

Requested topics ( $n$ )	Stable, retained topics ( $r$ )	
	w word2vec	w/o word2vec
14	14	13
15	<b>15</b>	12
16	<b>15</b>	12
17	13	13
18	15	12
19	14	12
20	14	13

The configurations yielding the maximum number of topics are marked with boldface

classifiers. The single best classifier, therefore, is the one using the 50% "climate extracts" of the texts.

## Unsupervised topic modeling

A maximum of 15 stable topics were extracted by Topics2Themes, when experimenting with different number of topics to request from the NMF algorithm (Table 9). Excluding the word2vec-based pre-processing consistently resulted in fewer topics being extracted. For further analysis, we, therefore, used the model for which 15 topics had been requested, and which used word2vec-based pre-processing.

In all experiments, 1670 stop words and 98 multi-word terms were used. For the experiments using word2vec, a list of 761 words not to include in the clustering was used, as well as a list of 248 manually constructed word clusters.

The 15 topics extracted are shown in Fig. 3. They are represented by the ten most closely associated terms and by the five most closely associated editorials (we show their title). We also provide a textual description for each topic, based on the content of the five texts. Curated word clusters are shown by a list of words separated by a slash, e.g., "law/rule/guidelines/regulations". Different inflections of a word are also typically clustered together, but only one form of a word is included in the figure. To the right of each editorial title, the frame classification provided by Hulme et al. [26] for the editorial is shown, using the abbreviations given above in Table 1.

The core of most of the 15 topics extracted can be related to one of Hulme et al.'s frames, although not all documents associated with the topic have been classified with that frame as the main one. For the topics, for which a core frame could be associated, this frame is shown bottom right for each topic in Fig. 3.

These manually associated frames are also shown by color coding and in the following order in Fig. 3: First, there is one ECON-related topic, viz. taxation as a way to tackle climate change. Then there are four TECH-related topics: (1) nuclear power; (2) renewable energy/energy efficiency; (3) carbon capture; and (4) biofuel.

Thereafter follow four GOV-related topics on: (1) decision makers not taking science into account; (2) international conferences and treaties; (3) the Rio Earth Summit; and (4) issues related to IPCC. However, the association of the IPCC topic is less evident than for the rest of the topics, as it might also be categorized as a SCI topic. There are three SCI-related topics on (1) the importance of science in general; (2) demand for more data collection for monitoring CC; and (3) the importance of climate research on the Arctic region. Finally, there is one COM-related topic that discusses modes of CC communication, communication mistakes, etc. The two last topics (Other 1 and Other 2), however, are difficult to map to any of Hulme et al.'s eight frames: (1) reports of possible effects of climate change, and (2) deforestation and biodiversity.

In our final step, we plotted a timeline for the topics extracted, where the prevalence of the topics of the years is shown; see Fig. 4. The larger the size of the black vertical bar for the topic, the more closely associated is the editorial to the topic. Similar to Hulme et al., we provide one timeline for Nature editorials and one for Science editorials. For comparing our TM findings with the temporal developments of the manual frame classifications provided by Hulme et al., we display these by colored circles on a white background just above their associated TM topics. For the three most infrequent frames (DEV, SEC and ETH), no associated topics were produced by the TM algorithm, and these frames are, therefore, not included in our graph.

Hulme et al. divided the period studied into eras and compared differences in frames used between these eras and between the two journals. In Table 10, we list the six main trends (Trends I–VI) that Hulme et al. mention in their timeline graph analysis, along with related output from the TM algorithm. The comparison produced three groups of trends:

- Trends I, II, and III, noted by Hulme et al., were also shown by the TM output. These were all trends relating to the three most prevalent frames; TECH, GOV, and SCI.
- Trends IV and V were partly shown by the TM output. These trends were related to the COM and ECON frames, which occur in the collection with a moderate and low frequency, respectively.
- Trend VI was not produced by TM. This trend was related to the ETH frame, which only occurs 11 times in the corpus, and for which no related output was produced by the TM algorithm.

## Discussion

### Supervised classification

We are tackling a classification task on a dataset that is characterized by both a relatively small overall size and a heavy class imbalance. This combination is generally problematic for the performance of classification algorithms, but it is not uncommon for many practical applications.

**Fig. 4** Topic distribution over time, and comparison with frames from Hulme et al. [26]. Each vertical line represents an editorial. Frame classifications by Hulme et al. are shown by colored dots on a white background, while corresponding TM results are shown underneath by vertical bars on a background with the same color as used for the frame indications. The bar lengths correspond to the topic strength for the editorial

Our results range from a rather good performance on one of the biggest classes ( $F_1$  0.91) to 0 for the very small classes that have only a handful of instances in the test set. Results for the COM frame were on par with results for the three most frequent frame classes (SCI, TECH, and GOV). This might be related to the "singularity" of COM, i.e., its comparatively rare co-occurrence with another frame in the annotations (cf. Fig. 1).

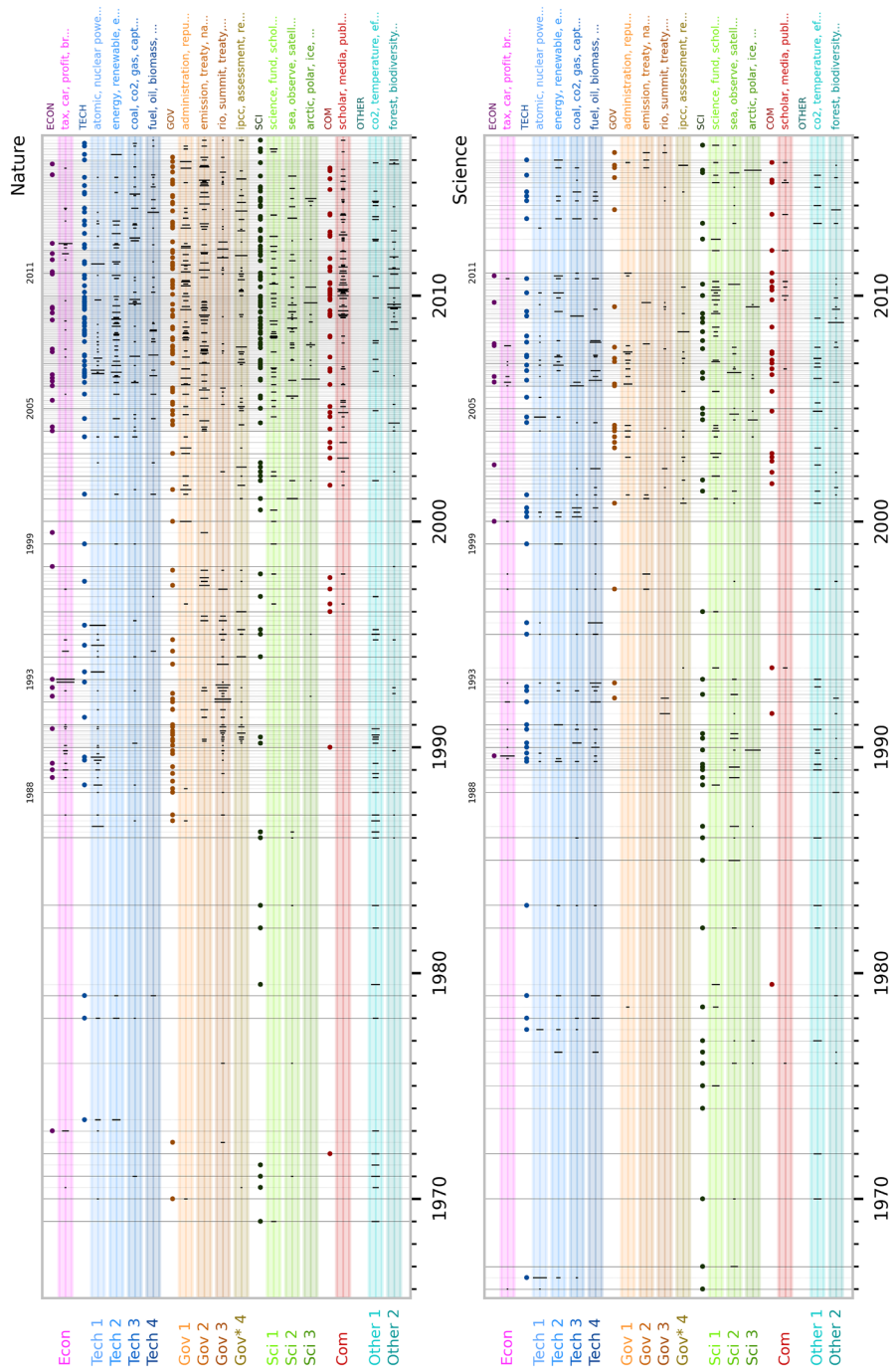
Applying class-imbalance techniques (of which we found oversampling to be most effective) leads to improved results throughout our experiments—and therefore we can recommend it—but it did not solve the problem of essential non-classification for very rare classes. We attribute this to the fact that the smallest classes are simply too small: With little data to resample from in our tiny classes DEV, ETH, and SEC, oversampling (be it synthetic or random) leads to much of the same data which is prone to overfitting.

An observation on our stage-1 cross-validation experiments on the training set (for which we cannot reproduce all results here for reasons of space) is that generally the different classifiers do not show great variety in performance, neither in terms of algorithm choice nor in terms of document representation. This indicates that for the task at hand, a weighted "traditional" bag-of-words model is not much worse than a more up-to-date word embedding approach (which is computationally more costly).

The frame annotations done by Hulme et al. were not explicitly based on the full editorial but meant to reflect the treatment of climate change within the text. As we noticed that, for the majority of texts, climate change is not the primary topic, we decided to investigate the consequences for frame classification. Our first step was to measure the prominence of the CC topic in all texts; this led to the result that CC-prominent texts achieve better results than those where CC is of more peripheral interest. Therefore, text topic and frame annotation overlap but are not the same thing; we return to this point in the conclusion. Our follow-up experiment on reducing the texts to their "most CC central" portions yielded mixed results: while we found a positive effect in terms of unweighted macro-F1, this was not the case in the weighted variant of the measure. This distinction points again to the issue of extremely small classes in the dataset. In general, we posit that focusing frame classification on the most relevant parts of the text can be beneficial.

## Unsupervised topic modeling

By its nature, the TM algorithm operated completely independent of the manual frame categorizations performed by Hulme et al. The two interesting dimensions for comparing TM output to manual frame annotation are the mapping between the two





**Table 10** Conclusions drawn by Hulme et al. [26] for temporal development of frames, compared to trends from our TM output (boldface) as shown in Fig. 4

Trend I: “Both journals primarily framed [CC] as a [TECH, GOV or SCI] challenge.” [26, p. 517].

– **The same trends are shown by TM.** (Topics: **Tech 1–4, Gov 1–4, Sci 1–3**).

In addition, the topic ‘reporting and discussing effects of CC’ (Topic: **Other 1**) occurred the entire period studied.

Trend II: “For Nature’s editorials in [the era 1988–1992], [CC] was primarily a [GOV] challenge, whereas for Science it was largely either a [TECH or SCI] challenge.” [26, p. 518]. It can, however, be seen in their graphs that Nature frames CC as a TECH challenge in a few editorials in this era.

– **The same trends are shown by TM.** (Topics: **Tech 1–4, Gov 1–4, Sci 1–3**).

This is particularly the case if the IPCC topic (Topic: **Gov\* 4**) – which occurred in Nature in this era – is categorised as a gov topic. The topic on nuclear power (**Tech 1**) could be interpreted as slightly more prominent in the TM trend for Nature, than the corresponding TECH occurrences in Hulme et al.’s trends. (For Topic: ‘**Other 1** reporting and discussing CC’ and Topic: ‘**Econ** the taxation topic’ – which also occurred in this era—see discussions above and below.)

Trend III: “Nature only began to give significant emphasis to the [TECH] challenges of [CC] from the [era starting in 2005 and] onwards, while Science only began seriously to emphasise the [GOV] challenge from the [era starting 1999 and] onwards” [26, pp. 518–519]. It can, however, be seen in their graphs that Nature frames CC as a TECH challenge in some editorials before 2005. “Over the whole period of the study, Nature emphasizes the [GOV] challenges of [CC] much more than does Science [...]” [26, p. 519].

– **The same trends are shown by TM.** (Topics: **Tech 1–4, Gov 1–4**).

The nuclear power topic (Topic: **Tech 1**) occurs in Nature before 2005, i.e., similar to the sporadic occurrences of the TECH frame in this period shown in the graphs by Hulme et al.

Trend IV: “[CC as an ECON] challenge was most prevalent for both journals in the [2005–2010 era].” [26, p. 517]. It can be seen in their graphs that the framing also was used in the 1988–1992 era, and in the beginning of the era starting 2011.

– **The same trends are partly shown by TM.** (Topic: **Econ**).

However, the occurrences of the taxation topic in the 2005–2010 era in Nature are non-prominent occurrences.

Trend V: “Especially noteworthy was the increase in framings of [CC] as a [COM] challenge [...] [from the era 2005–2010 and onwards.]” [26, p. 517].

– **The same trends are partly shown by TM.** (Topic: **Com**).

The trend is very evident for the communication topic in Nature. However, for Science, the trend is weaker and starts late in the era.

Trend VI: “[...] the identification of [CC] as an [ETH] challenge [...] has been notable only since 2005.” [26, pp. 517–518].

– **No topic matching the ethical framing was detected.**

---

*TM* topic modeling, *CC* climate change

sets of categories, and the possible reproduction of interpretations that Hulme et al. concluded from their data.

TM yielded 15 topics, most of which can be mapped to a frame. The topics are, however, often more fine grained than the frame categories, i.e., several topics were mapped to each one of the three frames—TECH, GOV, and SCI.

Two topics do not correspond to a frame, which in principle is not surprising, since the topics are extracted independently of the original frame definitions and annotations. Such topics might, therefore, form candidates for new frames to add



to the annotation categories. The topic on biodiversity and deforestation might, for instance, be added with the following frame description: “Deforestation is a cause of CC, and preservation of forests is important to tackle CC”. In contrast, the other topic not matching any of the current frames, i.e., “descriptions of possible effects of climate change and debates on whether they pose a problem”, probably does not constitute a candidate for a new frame. The purpose of many texts representing this topic is to argue for or against CC being a challenge that needs to be solved, rather than to frame this challenge. This topic might, nevertheless, form a candidate for an interesting annotation category.

Conversely, there are three frame categories—the three most infrequent ones—that do not show up in the topic set. Since the TM algorithm aims at detecting frequently occurring topics, it is also not surprising that no topics corresponding to these infrequent frame categories were extracted by the algorithm.

Besides the frame-topic mapping, it is interesting to study whether findings that Hulme et al. derived from the annotated corpus could also be (partly) generated with the help of unsupervised TM. Here, we focused on the “trends” that Hulme et al. observed in the temporal development of frame usage (see Table 10 above). Of the six trends, five were completely or partly reproduced by TM, and the only missing one is based on the very rare ETH frame, so this problem relates again to the heavy imbalance of the categories, as discussed above.

## Conclusion

The departure point for our work is the manual corpus study undertaken by Hulme et al. [26]. In addition to showing how an automatic quantitative approach compares specifically to the results of that original study, this paper mentions many individual results on relative performances of classification methods, which future studies on similar types of data can take into consideration for limiting their “search space” of methods.

We demonstrated that the presence of extremely rare categories poses problems for supervised and unsupervised methods alike; the ETH category, for example, was not recognized by the supervised classifier, nor did topic modeling pick up a topic for it. Aside from these extreme situations with only a handful of instances, we noted that methods for handling imbalance do have a positive effect. And for the frequent categories, promising classification results can be achieved. This result holds even in a setting with relatively many categories (in our case, eight frames)—in contrast, many practical supervised classification tasks use only two or a few more categories. For such settings, results will generally be notably better. Thus, we propose that supervised text classification is a viable method, for instance, when a huge text corpus is to be pre-filtered for relevant documents, as in the work of Stecula and Merkley [43], for example. One interesting observation we made for devising a classification strategy is that recent word embedding methods, which have been shown to be highly useful for many natural language processing problems, do not generally beat the much simpler methods (also in terms of computational resources) based on representing documents as bags of words.

Both the supervised and the unsupervised methods we employed are generally used for determining what texts are about. The fact that we use them for gaining information about frames results from the notion of frame that Hulme et al. [26] employed in their manual annotations; this is, however, also shared by the vast majority of previous research on automatic analyses of emphasis framing, as [44] observed. We regard our methods as appropriate for this type of framing [15] in this type of text genre (scientific editorial), but we point out that more politically-engaged text invites other approaches, such as the attempt on identifying elements of generic "framing language" (cf. [45]).

Our study showed that unsupervised topic modeling (TM), which does not rely on any manual annotation, can help in detecting frame categories. Recall that Hulme et al. had devised their frame set in an iterative procedure of reading/labeling/revising—a TM algorithm can speed up this process of exploring an unknown corpus by suggesting the initial frame set. In the case of Topics2Themes, this step can be undertaken interactively, so that the process can be guided by human expertise (somewhat similarly to "seeded" Latent Dirichlet Allocation). In addition, we noted that TM can play a helpful role also for interpreting the data without relying on any manual annotations, as the Topics2Themes system was able to reproduce almost all of the temporal "trends" that Hulme et al. had found for frame usage in the two journals over the 50 years.

Steps of our future work include expanding the study on CC-topicality for more precise classification; this will in particular include a manual validation of the automated labeling. In addition, we plan a validation of the framing annotations done by Hulme et al. (especially in the light of relatively low inter-annotator agreement scores published in their paper), and adding those annotations ourselves to a new portion of the corpus covering the years since 2016.

**Acknowledgements** We would like to thank Nailia Mirzakhmedova, who carried out the experiments using transformer-based classifiers. The study was partly funded by the project "Using the SB Sam NLP tools for manual and automatic annotation of climate change texts" (Vinnova, Ref no: 2021-03973) and "HUMINFRA: National infrastructure for research in the humanities and social sciences (the Swedish Research Council, Ref no: 2021-00176)".

**Author Contributions** MSt wrote sections 1, 2.1, and 5; performed the overall editing of the manuscript and carried out the overall coordination of the project. YB implemented the supervised classification experiments and wrote section 2.2, as well as the bulk of sections 3.1 and 4.1. LB and NCK implemented the topic centrality experiments and wrote section 3.1.3. MSk implemented the topic modeling experiments and wrote sections 2.3, 3.2, and 4.2. All authors read and approved the final manuscript.

**Funding** Open access funding provided by Uppsala University.

**Data availability statement** The dataset generated during the current study is not publicly available as it contains proprietary information that the authors acquired through a license. Information on how to obtain it and reproduce the analysis is available from the corresponding author on request.

## Declarations

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References


1. Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. *Mining Text Data*, 16, 163–222.
2. Badullovich, N., Grant, W., & Colvin, R. (2020). Framing climate change for effective communication: a systematic map. *Environmental Research Letters*, 15(12), 123002.
3. Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6), 1397–1410.
4. Bolsen, T., & Shapiro, M. A. (2018). The US news media, polarization on climate change, and pathways to effective communication. *Environmental Communication*, 12(2), 149–163.
5. Boom, C. D., Canneyt, S. V., Demeester, T., & Dhoedt, B. (2016). Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80, 150–156.
6. Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). *Applications of Topic Models*. Now Publishers.
7. Burscher, B., Odijk, D., Vliegthart, R., De Rijke, M., & de Vreese, C. (2019). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, 8(3), 190–206.
8. Cacciatore, M. A., Scheufele, D. A., & Iyengar, S. (2016). The end of framing as we know it ... and the future of media effects. *Mass Communication and Society*, 19, 7–23.
9. Card, D., Boydston, A. E., Gross, J. H., Resnik, P., & Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. In *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China. Association for Computational Linguistics, pp. 438–444.
10. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
11. Chen, J., Gong, Z., & Liu, W. (2019). A nonparametric model for online topic discovery with word embeddings. *Information Sciences*, 504, 32–47.
12. Chen, J., Gong, Z., & Liu, W. (2020). A dirichlet process biterm-based mixture model for short text stream clustering. *Applied Intelligence*, 50(5), 1609–1619.
13. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
14. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 4171–4186.
15. Entman, R. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4), 51–58.
16. Ester, M., Krieger, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. AAAI Press, pp. 226–231.
17. Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer International Publishing.

18. Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61, 863–905.
19. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139.
20. Goldberg, Y. (2017). *Neural Network Methods in Natural Language Processing*, vol 37 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool.
21. Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21, 267–297.
22. Guo, L., Su, C., Paik, S., Bhatia, V., Akavoor, V. P., Gao, G., et al. (2022). Proposing an open-sourced tool for computational framing analysis of multilingual data. *Digital Journalism*. <https://doi.org/10.1080/21670811.2022.2031241>. (published online).
23. Haddadan, S., Cabrio, E., & Villata, S. (2019). Yes, we can! Mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 4684–4690.
24. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, IJCNN 2008, pp. 1322–1328.
25. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
26. Hulme, M., Obermeister, N., Randalls, S., & Borie, M. (2018). Framing the challenge of climate change in Nature and Science editorials. *Nature Climate Change*, 8, 515–521.
27. Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter*, 6(1), 40–49.
28. Kantner, C., & Overbeck, M. (2020). Exploring soft concepts with hard corpus-analytic methods. In N. Reiter, A. Pichler, & J. Kuhn (Eds.), *Reflektierte algorithmische Textanalyse*. De Gruyter.
29. Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*. PMLR, pp. 957–966.
30. Lee, D. D. & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562.
31. Lind, F., Eberl, J.-M., Heidenreich, T., & Boomgaarden, H. (2019). When the journey is as important as the goal: A roadmap to multilingual dictionary construction. *International Journal of Communication*, 13, 4000–4020.
32. Lyytimäki, J., & Tapio, P. (2009). Climate change as reported in the press of Finland: From screaming headlines to penetrating background noise. *International Journal of Environmental Studies*, 66(6), 723–735.
33. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
34. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates Inc.
35. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
36. Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.
37. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint [arXiv:1910.01108](https://arxiv.org/abs/1910.01108).
38. Scheufele, D., & Iyengar, S. (2014). The state of framing research. In K. Kenski & K. H. Jamieson (Eds.), *The Oxford Handbook of Political Communication* (Vol. 1). Oxford University Press.
39. Sebastiani, F. (2005). Text categorization. In A. Zanasi (Ed.), *Text Mining and its Applications* (pp. 109–129). WIT Press.

40. Skeppstedt, M., Ahltop, M., Kucher, K., Kerren, A., Rzepka, R., & Araki, K. (2020a). Topic modelling applied to a second language: A language adaptation and tool evaluation study. In *Selected Papers from the CLARIN Annual Conference 2019*, volume 172:17. Linköping Electronic Conference Proceedings, pp. 145–156.
41. Skeppstedt, M., Domeij, R., & Skott, F. (2020b). Snippets of folk legends: Adapting a text mining tool to a collection of folk legends. In *DHN (Digital Humanities in the Nordic Countries) Post-Proceedings*, pp. 242–247.
42. Skeppstedt, M., Kucher, K., Stede, M., & Kerren, A. (2018). Topics2Themes: Computer-Assisted Argument Extraction by Visual Analysis of Important Topics. In *Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pp. 9–16.
43. Stecula, D. A., & Merkley, E. (2019). Framing climate change: economics, ideology, and uncertainty in american news media content from 1988 to 2014. *Frontiers in Communication*, 4, 6.
44. Stede, M. & Patz, R. (2021). The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, Online. Association for Computational Linguistics, pp. 8–18.
45. Walter, D., & Ophir, Y. (2019). News frame analysis: An inductive mixed-method computational approach. *Communication Methods and Measures*, 13(4), 248–266.
46. Weiss, G. M. (2004). Mining with rarity. *ACM SIGKDD Explorations Newsletter*, 6(1), 7–19.
47. Wetts, R. (2020). Models and morals: Elite-oriented and value-neutral discourse dominates American organizations' framings of climate change. *Social Forces*, 98(3), 1339–1369.
48. Wißler, L., Almshraee, M., Monett, D., & Paschke, A. (2014). The gold standard in corpus annotation. In *Proc. of the 5th IEEE Germany Student Conference*.
49. Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In S. A. Macskassy, C. Perlich, J. Leskovec, W. Wang, & R. Ghani (Eds.), *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24–27, 2014* (pp. 233–242). ACM.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

Manfred Stede<sup>1</sup> · Yannic Bracke<sup>1,2</sup> · Luka Borec<sup>1</sup> · Neele Charlotte Kinkel<sup>1</sup> · Maria Skeppstedt<sup>3</sup> 

Manfred Stede  
stede@uni-potsdam.de

Yannic Bracke  
yannic.bracke@bbaw.de

Luka Borec  
borec@uni-potsdam.de

Neele Charlotte Kinkel  
neele.charlotte.kinkel@uni-potsdam.de

- <sup>1</sup> UFS Cognitive Science, University of Potsdam, Karl-Liebknecht-Straße, 14476 Potsdam, Germany
- <sup>2</sup> Berlin-Brandenburg Academy of Sciences and Humanities, Jägerstraße 22/23, 10117 Berlin, Germany
- <sup>3</sup> Centre for Digital Humanities Uppsala, Department of ALM, Uppsala University, Box 256, 75105 Uppsala, Sweden