Topic Modeling as a Method for Frame Analysis: Data Mining the Climate Change Debate in India and the USA

Tuukka Ylä-Anttila¹, Veikko Eranti² & Anna Kukkonen³

Abstract

This article proposes an operationalization of topic modeling as a data mining method for studying framing in public debates. We argue that ‘topics’ can be interpreted as frames, if 1) frames are operationalized as connections between concepts, 2) subject-specific data is selected, and 3) topics are adequately validated as frames, for which we suggest a practical procedure. As an empirical example, we study the global climate change debate in the media, by comparing frames used by NGO’s, governments and experts in India and the USA. Our model identifies 12 framings of climate change, used in varied proportions by the different actors in the two countries. Topic modeling frames enables the usage of larger datasets and facilitates discovery of previously unnoticed framing patterns. It does not replace qualitative interpretation, but rather complements it by enabling a degree of automated classification before the interpretive stage.

Keywords: framing, topic modeling, climate change, media, methodology

Introduction

Topic modeling (e.g. DiMaggio et al. 2013; Evans 2014) is a data mining method, associated with computational social science, a field with much to contribute for sociological research methodology. But computational approaches such as topic modeling have often emphasized induction, where patterns are expected to arise from the data with as little theoretical pre-conceptions as possible, while making aggressive claims about causality (Babones 2016). We attempt to reconcile some of these issues by operationalizing topic modeling for the purposes of a theory-rich field of interpretive sociology, that of

¹ Postdoctoral Researcher, Faculty of Social Sciences, University of Tampere, Finland. Visiting Researcher, COSMOS Centre on Social Movement Studies, Scuola Normale Superiore, Florence, Italy. E-mail: tuukka.yla.anttila@gmail.com

² Postdoctoral Researcher, Faculty of Social Sciences, University of Tampere, Finland.

³ Doctoral Researcher, Sociology, University of Helsinki. Fulbright Visiting Scholar, Center for Science and Technology Policy Research, University of Colorado Boulder, USA.

frame analysis of public debates, and particularly the debate on climate change, which has been argued to be in need of data-mining approaches (Broadbent et al. 2016).

This field, we argue, should be particularly suitable for topic modeling, because of certain theoretical compatibilities. Namely, if we understand that a frame ‘links two concepts, so that after exposure to this linkage, the intended audience now accepts the concepts’ connection’ (Nisbet 2009: 17), such linkages should be found by a topic modeling algorithm which detects and tracks concepts that ‘tend to occur in documents together more frequently than one would expect by chance’ (DiMaggio et al. 2013: 578). The continued habitual use of particular words together with each other shows that those words have meaning in relation to each other, together forming a cluster of concepts, which can be interpreted as a frame. Thus, topic modeling should be able to automate a part of the frame analysis process of texts.

Data mining methods such as topic modeling find patterns in large datasets. In culturally informed sociology, where meanings and meaning-making habits are generally in focus, and different types of close reading of texts are some of the key methods, topic modeling can in contrast be seen as a method of ‘distant reading’ (Moretti 2013). It reduces the complexity of language to a simplistic assumption, namely that certain words often occur together, and these co-occurrences – word clusters – carry traces of meaning. By computationally observing patterns and variations in usage of these word clusters, we are able to observe variations in meaning-making habits – a facet of culture. This makes topic modeling suitable for analysis of large datasets, but it should be complemented with close reading to create nuanced knowledge on meaning(s), making sociological topic modeling necessarily a mixed-methods endeavour. The advantages of topic modeling for text analysis are two-fold: first, it enables usage of larger data because of reduced need for time-consuming qualitative analysis, and second, it enables exploratory discovery of patterns not previously found in qualitative inspection.

Our example case to illustrate this methodological approach is the global climate change debate. Climate change is the most pressing socio-ecological issue of our time – yet there is no consensus how, even if, it should be addressed: both the science and the politics remain hotly contested (Urry 2011). This is, to a large extent, due to disagreement on the nature of the problem, which is at the same time environmental, cultural, and political (Hulme 2009), making it salient for analysis of framing. A variety of policy actors engage in the framing processes: states who negotiate in international political settings, such as the United

Nations Framework Convention on Climate Change (UNFCCC), but also non-state, scientific and private actors on global, national and local levels (Bulkeley & Betsill 2005) – making it crucial to understand differences in framing between these stakeholders, as well as national cultural differences in framing practices (Ylä-Anttila & Kukkonen 2014). Such knowledge could help comprehend why common understandings about climate change and its mitigation are so elusive (Anderson 2009, Billet 2010, Boykoff & Boykoff 2007, Boykoff 2011, Boykoff & Nacu-Schmidt 2013, Farrell 2015, 2016, Nisbet 2009, Schäfer & Schlichting 2014, Trumbo & Shanahan 2000). There is even some evidence that particular frames adopted in national newspapers correlate with reductions of emissions in that country (Broadbent et al. 2016).

Our analysis includes media coverage from two countries that are major players in the global politics of climate change, India and the USA, using one newspaper from both, for three different six-week time periods around United Nations Climate Change Conferences between 1997 and 2011. The countries were chosen as hypothetically the most different cases (Pfetsch & Esser 2004) in an existing dataset collected for a previous research project (Ylä-Anttila & Kukkonen 2014). How do different speaker groups frame climate change in the debate, as reported by newspapers in India and the USA? For the purposes of our analysis, we chose the three most prominent policy actor categories in the media debate in both countries: experts, governments and NGOs.⁴

Our methodological contribution is to provide an answer to the debate on whether frames can be operationalized as topics (Bail 2014; DiMaggio et al. 2013). The answer is a ‘yes’ with some additional qualifications. Doing so requires 1. adopting a view of framing as connections between concepts (Entman 1993, Nisbet 2009), 2. selecting the input text data to be subject-specific, and 3. interpretive validation, for which we suggest practical guidelines. Using other more nuanced definitions of frames (such as Goffman 1974), or different theoretical concepts altogether, different qualifications would have to be taken into account. Empirically, we find that economic concerns seem to be primary in the climate change debate as portrayed by US media, while burden-sharing and environmental risks are emphasized in India.

⁴ For the USA, a comparison of Republicans and Democrats would be an obvious alternative, as well as looking at conservative think tanks and the business lobby as actors. These were left out to make categories comparable between the two countries studied here.

Climate Change Framing in Media Debates

Our example case, climate change, has become a salient and controversial topic in the media all over the globe, peaking in 2009 in both India and the USA (Schäfer & Schlichting 2014; Boykoff & Nacu-Schmidt 2013). Indeed, the mass media is an important arena for political debates on climate change, in which the cultural understanding of climate change is constantly shaped by political actors (Boykoff 2011; Crow & Boykoff 2014; Hansen 2010). Consequently, different actors have engaged in very different framings of climate change (Nisbet 2009). These include the frames of *economic competitiveness*, in which climate change is either a threat to economic growth or, perhaps, a driver of it; *morality and ethics*, in which climate change and its mitigation are matters of right or wrong; and *scientific uncertainty*, in which debates regard whether something is proven or not (Nisbet 2009: 18). These framings of the problem lead to different proposed remedies.

In addition to differences between actors, the framing of climate change also varies between political contexts: policy actors are likely to use culturally specific frames (Anderson 2009, Trumbo & Shanahan 2000). Accounts of US media coverage are numerous, while studies of Indian media coverage are thus far fewer. In the US, the media has particularly framed climate change through scientific uncertainty and given disproportionate space to climate sceptics (Boykoff & Boykoff 2007), since the conservative movement has systematically disseminated discourses referring to the economic costs of mitigation and the uncertainty of climate science (McCright & Dunlap 2003; Hoffman 2011; Oreskes & Conway 2010), enabled by networks of corporate funding (Farrell 2015; 2016). While the frames of science and economy have dominated US media discourses, in India, the coverage has focused on the international dimensions of climate policy. Particularly the North-South divide is salient in Indian debates, as well as the environmental risks global warming poses to India, according to Billet (2010). Contrary to the US, climate skepticism does not carry much weight in India.

Few studies have analyzed media debates from the viewpoint of frames used by individual speaker groups. They have mainly focused, especially in the US, on analyzing the role of the conservative movement and the overall contestation of climate change. Nisbet (2009: 18) notes that ‘trusted sources have framed the nature and implications of climate change for Republicans and Democrats in very different ways’ in the USA, but his focus is on specific actors’ specific claims – such as conservative think tanks’ framing of

climate change as scientifically uncertain – rather than painting a broader picture of frames used in the debate. Similarly, Farrell’s (2015; 2016) computational text analysis approach maps out the policy networks and discourses advocated by climate change denialists, but there is still a lack of more general knowledge on variation in framing between speaker groups, and particularly whether or not there are frames in which actors converge, creating possibility for common ground (Broadbent et al. 2016).

Topic Modeling

The software we use for topic modeling is MALLET’s (Machine Learning for Language Toolkit; McCallum 2002) implementation of Latent Dirichlet Allocation (LDA; Blei et al. 2003), which has specific advantages for sociologists studying framing – namely its reliance on co-occurrences of concepts (DiMaggio, Nag & Blei 2013). It is an unsupervised machine learning method: the researcher gives no input as to how the data should be classified. As such, the classifications themselves produced by the software are fully inductive, grounded in data, and based only on co-occurrences of words, rather than on a pre-set interpretation of the researchers. Indeed, the typical relationship between data and theory in data mining is similar to that of grounded theory, with a strong focus on inductive reasoning (Babones 2016: 457, Glaser and Strauss 1967). However, instead of theory-blind ‘black-box’ data mining, we argue that interpretation is just as important in data mining as in ‘ordinary’ qualitative interpretive sociology, or close reading. In our approach, the interpretive stage, based on subject-specific researcher knowledge, takes place after the primary, inductive classification is done by machine. As a result, the workflow resembles an abductive rather than an inductive or deductive approach (Timmermans & Tavory 2012).

Consider a concise description of LDA: ‘[it] assumes that there are a set of topics in a collection (the number is specified in advance) [...] Terms that are prominent within a topic are those that tend to occur in documents together more frequently than one would expect by chance [...] each document exhibits those topics with different proportions.’ (DiMaggio et al. 2013: 577–578) In other words, LDA is a probabilistic model, which models the probability of each topic (word cluster) in each document, and the probability of each word in each topic. This is the basis for assigning words into topics, which are essentially probability distributions over a corpus of words (Blei et al. 2003). In simpler terms, and in the case of MALLET, the researcher inputs a dataset consisting of text documents and asks the software to return a particular number of topics, say ten. The outputs are 1) ten lists of words that most often *occur*

together in documents, 2) numeric measures for the topics in each document, showing how large a share of that document consists of that topic, and 3) numeric measures for the documents in each topic, representing how large a share of that topic exists in that document. Thus, the typical interpretation of an output is: 1) what ‘topics’ is the dataset about, 2) what topic is each document about and to which extent, and 3) in which documents is a topic discussed, to what extent. Additionally, various measures of model fit can be calculated, which may help researchers assess the validity of the model and fine-tune it, but we do not go into such quantitative detail here (see Chang et al. 2009; Schöch 2016), instead focusing on usability for frame analysis.

Topic Modeling Frames

The word *topic* in ‘topic modeling’ does not refer to specifics of the algorithm, which ‘knows’ only word co-occurrences. Calling them ‘topics’ is an interpretation of what the algorithm’s output can be applied to, and what kinds of research objects can be *operationalized* as word clusters. Finding topics or ‘themes’ in text data is but one possible interpretation. When presenting LDA, Blei et al. stated that they ‘use the language of text collections throughout the paper, referring to entities such as “words”, “documents” and “corpora”’ to ‘guide intuition’ (Blei et al. 2003: 995). In fact, the same applies to the concept of ‘topic’ – we argue that it has ‘guided intuition’ too much in the topic modeling literature. ‘Topics’ (themes of text content) are only one thing LDA can be used to model.

We argue that in a research setting where we use texts about a particular topic (climate change) as input, LDA outputs are best interpreted as different ways of talking about a topic, or frames. This is because all of the text is about climate change, so the word co-occurrence patterns that emerge are patterns of using certain words to talk about climate change. Such word use patterns can plausibly be interpreted as framing patterns. Most applications of topic modeling emphasize validating the outputs (Evans 2014; DiMaggio et al. 2013; Grimmer & Stewart 2013): that is, checking that the word clusters actually mean what we think they do. If our interest lies in frames rather than topics, both internal and external validation should take into account what we already know about frames.

First of all, internally, word clusters must be interpretable as frames, or ‘schemata of interpretation’. Thus, a frame ‘allows its user to locate, perceive, identify and label’ events (Goffman 1974: 21). But after

Goffman's original work, which focused on framing in micro-level face-to-face interaction, the literature on framing has expanded in various directions, and there are now several paradigms of framing used in multiple fields of social science. For operationalization of framing in textual communication, we follow Entman's (1993) simplified conception of framing instead:

To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described. (Entman 1993: 52)

Frames, thus, 'define problems', 'diagnose causes', 'make moral judgments' and 'suggest remedies' and take place in a 'communication process' (Entman 1993: 52). It is imperative to note that 'frames as general organizing devices should not be confused with specific policy positions; any frame can include pro, anti and neutral arguments' (Nisbet 2009: 18). A frame, in this definition, 'links two concepts, so that after exposure to this linkage, the intended audience now accepts the concepts' connection' (Nisbet 2009: 17). Framing 'endows certain dimensions of the complex issue with greater apparent relevance' (Nisbet 2009: 16–17). While the theoretical basis of framing is in Goffman's work (1974), this more streamlined definition is better suited for identifying frames in text, since what we observe is linkages between two concepts, shown by patterns of co-occurrence.

Secondly, externally, the frames should mostly correspond to previously identified frames in similar data to ensure validity, but there must also be some room for discovery in order for topic modeling to be valuable. After all, exploration, or finding patterns not previously identified, is one of the primary reasons for using topic modeling rather than purely qualitative reading. This means that the outputs of the model are not very plausible if they directly and strongly contradict all previous studies carried out using other methodologies. In contrast, findings which correspond to previous ones are a good indicator that the basic research strategy works, and lend credence to new and even potentially surprising insights gained from the same model. This is why we selected a case about which there is already a body of empirical research, framing climate change in media discussions (Anderson 2009, Billet 2010, Broadbent et al. 2016, Boykoff & Boykoff 2007, Boykoff 2011, Boykoff & Nacu-Schmidt 2013, Farrell 2015, 2016, Nisbet 2009, Schäfer & Schlichting 2014, Trumbo & Shanahan 2000). Thus, we can confirm the validity of the model and show its benefits for novel findings. In other words, if results mostly correspond to different framings of

climate change previously identified, including at least *economic competitiveness*, *morality and ethics*, and *scientific uncertainty* (Nisbet 2009: 18), the model gains credibility. We indeed find these, but in more nuanced forms.

Data

The data we use consists of all articles mentioning the keywords ‘climate change’ or ‘global warming’ in the New York Times (NYT, USA) and The Hindu (India), for timespans explained below. While these newspapers are by no means fully representative of national public spheres, both are widely considered relatively liberal, mainstream newspapers, and both have wide circulations within their respective countries.⁵ As with all research studying public debates using media data, editorial policies, the national media environment and other factors affect what is published. Nevertheless, the media are one of the primary arenas for public debate, in which actors engage in framing to further their political positions, and through which citizens receive information to understand the society around them. Even if the media are biased, it is this biased communication on which public debate is largely based. Thus, studying framing in media debates on climate change is important in itself: it matters for formation of public opinion and possibilities of climate change mitigation (see e.g. Anderson 2009, Billet 2010, Boykoff & Boykoff 2007, Boykoff 2011, Boykoff & Nacu-Schmidt 2013, Nisbet 2009, Schäfer & Schlichting 2014, Trumbo & Shanahan 2000).

We collected data around three international climate meetings: Kyoto (1997), Copenhagen (2009) and Durban (2011), three weeks before and after each meeting. Altogether, 677 articles were included in the data, shown in Table 1. While the NYT articles are fewer in number, they are longer, which balances the data: both newspapers published a similar number of claims on climate change.

⁵ The Hindu is the largest English-language newspaper in India, and India has the world’s second-largest English-speaking population, second only to the USA.

	Articles	Claims
NYT	94	353
The Hindu	583	383
Total (words)	677 (416 822)	736 (103 589)

Table 1. The dataset.

For purposes of previous research, political claims were already marked in the text data, and the speaker category (expert, government or NGO) for each claim was coded (Koopmans & Statham 1999). A claim is:

[A] unit of action in the public sphere. A claim can be a comment in an interview or a public speech, a demonstration or other action whose purpose is to influence public debate. One newspaper article may, therefore, contain several claims by several actors. (Ylä-Anttila & Kukkonen 2014)

The speaker category was identified not only for direct quotes but also for claims paraphrased by reporters. As an example, the following excerpt from The Hindu was categorized as a government speaker:

The developed countries must step up to the plate to come true on their existing commitments to fight climate change, said Jayanthi Natarajan, Indian Environment Minister.

The *government* category includes officials speaking for governments active in the negotiations, multiple governments giving joint statements, and inter-governmental organizations. Since the data consists of reports about climate change conventions, the role of governments was central. The data included 353 claims by government speakers. The speaker was identified as an *expert* if she presented herself with scientific credentials, as coming from scientific institutions, or in other ways clearly positioned as an issue authority. 251 comments by expert speakers were found. *Non-governmental organizations* were the third speaker group included in the analysis. These included civil society actors such as environmental organizations. The data included 132 claims by NGOs.

We only input the previously identified political claims in the model, not the whole text of the newspaper articles, to further specify the data to be about the climate change debate, not just descriptive text on climate change. This was possible because of previous hand-selection, but if no such material is available,

other methods should be used to ensure the text data is *about* a particular topic: e.g. Levy & Franklin (2013) used comments on regulation of the trucking industry, containing justifications for opinions, which was their object of interest.

The claims were contained in text files named by speaker category, country and an ID. The files were then ‘tokenized’: stripped of all punctuation and empty lines and processed into files that contain one instance of a word (a ‘token’) per line, using simple Python scripts. The tokens were stemmed using the Snowball stemmer (Porter 2001) in the Python Natural Language Toolkit (Bird & Loper 2006) to collapse inflected word forms into a single word, e.g. ‘changing’, ‘changed’ and ‘changes’ were all converted into ‘chang’, to detect them as the same word. MALLET’s standard stopword list was used before modeling.

Finding Frames in a Topic Model

Figure 1. Frame validation process.

Interpretation and validation of topics has been identified as a crucial point in using topic modeling in analysis of societal phenomena (Evans 2014; DiMaggio et al. 2013; Grimmer & Stewart 2013). Since we argue that word clusters output by LDA can be validated as frames, we propose a three-fold process for *frame validation*, presented in Figure 1. The first stage looks at the whole model on the surface, the second

inspects the top words of each topic for internal validity (meaningful topics), and the third inspects the source data itself by looking at the top documents of each topic for external validity (topics that represent frames).

With LDA, the only input given by the researcher in addition to the data itself is the number of topics. The selection of topic count affects the fit of the model (Evans 2012: 2), thus, that selection is the first phase of validating the model. Too many topics result in topics that are too specific, while too few topics results in topics that mix several frames in one. We tested different options including 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 topics, and each output was examined in terms of top 10 words of each topic, to look for internally and externally coherent framings of climate change. We ended up with a model of 30 topics, which produced topics that are not too specific and not too general or ‘mixed’. In this first stage, it is better to use too many rather than too few topics, since irrelevant topics will be discarded in the next stages. While quantitative measures of model fit exist and may be consulted in making this decision, they are not explicit in interpretation and cannot replace qualitative interpretation (Ylä-Anttila 2018).

Second, we used the top 10 words of each topic to qualitatively inspect and discard topics that did not constitute *internally* valid frames, which link climate change to a coherent set of other concepts (Nisbet 2009). The rationale behind ‘top words’ is that topic models are mixed-membership models: each word can be part of multiple topics but with different probability weights. MALLET outputs a word frequency for words in each topic: the most commonly occurring words matter most in creating the topic, while the rest form a ‘long tail’ of words that are not nearly as significant. Domain-specific research knowledge is important here – as well as familiarity with the data at hand – to make qualitative validation possible. We discarded 13 topics and kept 17, for which we gave a tentative, descriptive name. An example of a discarded topic, in which we did not find internal coherence, would be *concern, clear, don’t, give, document, tax, accept, base, thing, main*; while an example of a topic deemed coherent was *warm, global, scientist, research, univers(e/al), atmospher(e), caus(e), stud(y/ies), effect, release* – this was interpreted as the *climate science* frame.

In the third stage, we read the top 10 documents in the 17 frames that passed the previous stage to check whether the tentative description fit. Again, using a mixed-membership model like LDA, documents belong to multiple topics but with different weights. This means some of the top documents in the different frames are the same; this is a natural reflection of the fact that a document may contain multiple

frames. If the preliminary topic description fit at least 8 of the top ten documents in that frame, we kept the frame. In many cases, the descriptions fit after slight rewording – interpreting the documents gave a clearer picture of the frame. In this phase, we discarded 5 of the 17 frames on the basis of lack of coherence. Thus, we ended with 12 frames with validated descriptions. This final set of frames, presented in Table 2, is derived from the data algorithmically, but validated and interpreted by researchers in a reflexive process going back and forth between interpretation and model.

Green Growth	Emission Cuts	Negotiations and Treaties	Environmental Risk
energi	emiss	nation	indian
fund	cut	unit	sea
billion	greenhous	state	water
state	industri	treati	increas
public	gas	commit	forest
clean	gase	economi	today
invest	call	major	ocean
renew	system	respons	region
creat	adopt	american	risk
govern	declar	recent	rate
Cost of Carbon Emissions	Chinese Emissions	Economics of Energy Production	Climate Science
carbon	china	econom	warm
emiss	target	technolog	global
reduc	reduct	cost	scientist
dioxid	chines	compani	research
pollut	growth	energi	univers
trade	intens	fuel	atmospher
cap	current	power	caus
coal	reduc	money	studi
power	plan	price	effect
product	oblig	mani	releas
Environmental Activism	North-South Burden Sharing	State Leaders Negotiating	Citizen Participation
peopl	countri	meet	part
govern	develop	minist	make
environ	talk	confer	citi
environment	commit	mr	organis
protect	provid	day	green
campaign	financ	prime	member
tree	demand	singh	differ
speak	adapt	thursday	number
everi	ensur	announc	initi
human	warsaw	attend	greenpeac

Table 2. Validated climate change frames and their top 10 words found by topic modeling in The New York Times and The Hindu.

Topic Modeling in a Comparative Setting

Now that we have proposed a method to validate topics as frames, how to use the model to compare framing between speaker groups and countries? Since MALLET outputs a list of documents for each topic, along with the frequency of words in each document that were assigned to that topic, it is quite easy to inspect which portion of the top documents in each frame originated from which speaker group and which country. As an example, Table 3 presents the top document list for the *green growth* frame and the number of words in each document that were assigned to this frame by the model. The table only includes documents with ten or more words assigned to this frame; about 500 documents with less than ten assigned words are omitted. By comparing the amounts of each speaker group and country in this list of top documents, we can see which speaker groups and which country most contributed to this frame. The cut-off point of ten words was decided to get a long enough document list for each frame to enable comparisons, but to exclude the ‘long tail’.

Document name	Words in document assigned to ‘green growth’ frame
india-gov247	110
india-gov236	64
usa-exp79	48
india-gov46	39
usa-exp132	39
india-gov65	38
usa-ngo84	30
usa-exp63	29
india-exp152	25
usa-exp116	25
usa-exp114	25
india-gov84	22
india-exp16	18
india-exp32	17
india-ngo56	16
usa-gov11	16
india-gov243	16
india-gov58	15
usa-exp98	15
usa-exp111	15
india-ngo17	14
india-gov8	14
india-gov72	14
usa-exp120	14
india-ngo22	12
india-ngo15	12
india-gov75	12
india-gov2	12
usa-gov45	11
gov32	11
gov156	11
gov111	10
exp92	10

Table 3. Top documents in the *green growth* frame.

Interpreting Frames

Next, we interpret the 12 frames, the distribution of speakers in top documents of each frame, and the distribution of US and Indian claims, shown in Figures 2 and 3.

Figure 2. Percentages of frames used by each speaker group, i.e. the sum of all ‘expert’ bars is 100%.

Figure 3. Percentages of frames used in each newspaper, i.e. the sum of NYT bars is 100%.

We treat the topics as frames in which connections are created between different concepts: they are presented constructions about which concepts have to do with climate change. The other concepts are thus considered, by the speakers, in some way meaningful to climate change. This means that e.g. the ‘emission cuts’ frame contains both statements asserting that emission cuts are necessary, and statements asserting they are not – however, they still frame climate change in terms of emission cuts, that is, they posit that emission cuts are relevant with regards to climate change. This is consistent with Entman’s (1993) and Nisbet’s (2009) definition of frames.⁶ In the quotes below, occurrences of top 10 words of the frame are *italicized*.

First of all, the *green growth frame* is represented by words such as fund, billion, invest, clean and renew. These claims refer to sustaining the environment and economic growth simultaneously – by investing in clean energy, for instance (OECD 2011) – and it is the most uniting frame for speaker groups, thus providing hope for common ground in the debate. In the quote below, low-carbon investments are presented as economically viable.

‘Using our national development finance institution and export credit agency, we have channelled hundreds of millions of dollars to strengthen India’s ability to build technical capacity, reduce financial risk, and lower the cost of capital for low-carbon *investments*.’ (india-gov247)

Emission cuts: with words like emission, cut and greenhouse (gas/effect), this was another unifying frame for all speaker groups. All speaker groups agree that emission cuts matter and engage in the discussion about them – whether to argue they are necessary or that they are not. The quote below discusses emission cuts in the US.

‘The United States could shave as much as 28 percent off the amount of *greenhouse gases* it emits at fairly modest cost and with only small technology innovations, according to a new report.’ (usa-expert114)

Negotiations and treaties is about states taking part in climate negotiations, represented in the top words nation, state, treaty etc. Naturally, governments engage in this frame the most, but are very closely

⁶ Detecting whether speakers refer to these frames in a positive or negative sense might be possible using sentiment analysis (e.g. Pang & Lee 2008).

followed by experts and NGO's, making this a comparatively uniting topic. Below, The Hindu reports on the effect of US domestic politics on the global negotiations.

'The US Senate majority leader said on Tuesday that the climate *treaty* being negotiated in Japan has only "bleak prospects" of ratification by the upper House of the U.S. legislature, even if U.S. negotiators agree to it.' (india-gov109)

Environmental risk is a frame about the imminent consequences of climate change. Speakers, predominantly experts, warn of – or downplay – dangers such as rising sea levels and deforestation, exemplified by words such as sea, water, increase, forest, ocean and risk. This frame was common in The Hindu but rarely utilized in The New York Times, which is in line with previous studies on the Indian climate debate, showing that risks are considered more salient in India than in the USA (Billet 2010). The quote below is from a report on research into environmental risks of climate change.

'Greenhouse gases are making the world's *oceans* hot, sour and breathless, and the way those changes work together is creating a grimmer outlook for global *waters*, according to a new report from 540 international scientists.' (india-expert158)

The *cost of carbon emissions* frame consists of words such as carbon, emission, trade, cap, coal, power and product, measuring the effects of emissions in terms of money, primarily discussed by experts in The New York Times. As we know, economic framing is exceptionally strong in the American climate debate, largely due to the primacy of economic concerns in American political culture (Lamont & Thévenot 2000) and the organized countermovement against climate change legislation (McCright & Dunlap 2013; Farrell 2015, 2016). As an example, in the following quote, an expert assesses how carbon emissions should be taxed.

'But it would be even better, Dr. McKittrick says, to use the temperature readings as the basis for a *carbon* tax instead of a *cap-and-trade* system.' (usa-expert67)

The *Chinese emissions* frame consists of words such as China, target, growth and reduc(e) – which correspond to claims that Chinese emissions are crucial for mitigating climate change. It is a frame used mostly by states and experts, and mostly in The Hindu, suggesting that China's role in mitigation is highly relevant for the debate in India. Some claims lay the blame and responsibility for climate change on China, while some praise China for e.g. 'global leadership in renewables'.

‘[T]here is going to have to be much more pressure on *China* if global emissions are to peak within any reasonable time frame.’ (usa-expert92)

The *economics of energy production* frame consists of words like economy, cost, energy, fuel, power, money and price. Dominated by experts, discussions in this frame deal with the costs of different methods of producing energy such as renewables and fossils. This was the most strongly US-dominated frame, which comes as no surprise because of the emphasis on economic concerns in American politics (Lamont & Thévenot 2000). Below, an expert report in NYT assesses possibilities for lowering energy costs.

‘The report said the country was brimming with “negative cost opportunities” – potential changes in the lighting, heating and cooling of buildings, for example, that would reduce carbon dioxide emissions from the burning of fossil *fuels* even as they save *money*.’ (usa-expert114)

The *climate science* frame is defined by words such as scientist, research and study, and consists largely of news pieces reporting on climate research. This frame was understandably dominated by experts, quite evenly used in the Indian and US data despite the highly contested nature of climate science in the US (Boykoff & Boykoff 2007; Farrell 2015; 2016; McCright & Dunlap 2010) – likely explained by the choice of the liberal New York Times as the data source. The quote below refers to a research report proving anthropogenic climate change.

‘Climate change *caused* by humans is real and it is happening now [...] The recently *released* report of the Intergovernmental Panel on Climate Change has reconfirmed the basic facts’ (india-gov247)

Environmental activism is a frame dominated by NGOs, consisting of keywords like people, govern, environment, protect and campaign, connected to local and global activist initiatives to combat climate change, together with broader statements about environmental values and protection. The following quote is from a report on Indian climate activists.

‘[E]co-activist Sunderlal Bahuguna [...] said the involvement of *people* in *environmental campaigns* was crucial for victory against *government* moves. Recalling his active participation in the Chipko Movement, he said he had walked from Kohima to Kashmir in 300 days to create awareness among *people* about the demerits of felling of *trees*.’ (india-ngo31)

North-South burden sharing includes words like develop(ing), countri(es), finance(e), and commit. It was governments who mostly discussed climate change through this frame, which deals with climate justice: whether the Northern countries are more responsible and should finance developing countries in their

mitigation actions. Burden-sharing has been recognized as one of the central issues in the Indian debate (Billet 2010), which was also visible in our comparison.

‘The G77+China group delivered an ultimatum to the developed *countries* on the issue of Loss and Damage, threatening to walk out of the *Warsaw* negotiations if the *developed* countries did not stop blocking it.’ (india-gov71)

State leaders negotiating includes words such as meet, (prime) minister, attend and confer(ence). Unsurprisingly, mostly government actors had a voice in this frame. It was the most Indian-dominated of the frames, largely due to The Hindu reporting on Indian delegates’ efforts in the climate meetings, such as in the following quote.

‘Chinese Premier Wen Jiabao on *Thursday* spoke to *Prime Minister* Manmohan *Singh* to exchange views on the climate change *conference* at Copenhagen. During their 10-minute telephonic conference, the leaders were said to have discussed ways of taking the climate summit forward.’ (india-gov13)

Finally, *citizen participation* is a frame discussing NGOs and individual citizens taking part in the climate debate process, marked by words such as (take) part, organiz(e | ation), member, and also names of specific NGOs such as Greenpeace. Not surprisingly, this frame was dominated by NGOs – the most dividing frame measured by deviation of speaker group frequencies. The following quote is from a report in The Hindu.

‘To create public awareness about how climate change is affecting our planet, bicycle enthusiasts took *part* in a cycle rally on the Capital’s much talked about Bus Rapid Transit (BRT) corridor over the weekend.’ (india-ngo17)

Discussion

In this paper, we set out to examine whether topic modeling can be used to find frames in public debates. All in all, in our newspaper data on the climate change debate in India and the USA, we found 12 word clusters that can be interpreted as frames after a validation process we documented. This process took less time by topic modeling that would have been the case with full qualitative analysis, and this benefit is scalable: increasing the dataset size would not have considerably increased required time. Studying the distributions of these frames between countries and speaker groups, we noted differences between framing in Indian and US media data, such as the strong position of economic framings in the USA, and the emphasis on environmental risks in India. We found frames that were dominated by a particular

speaker group, such as experts in the case of climate science, but also frames in which actors converge, such as the green growth frame. This shows that topic modeling outputs can usefully be interpreted as cultural constructs such as frames, as already argued by some scholars (DiMaggio et al. 2013), but we add certain crucial preconditions that must be met by the research design before and after the actual modeling.

Firstly, we adopted an interpretation of framing in the vein of Entman (1993) and Nisbet (2009). This perspective on frames is broader than both the conceptions emphasizing intentionality and strategic action, often used by social movement scholars (e.g. Benford & Snow 2000), and from the micro-interaction focused perspective of Goffman (1974). The additional theoretical subtlety of these accounts makes them harder to grasp by topic modeling – in the case of Goffman’s micro-interactions, rather impossible. However, we argue LDA is quite suitable for analysing frames understood in Entman’s and Nisbet’s sense, as links between concepts, as LDA is based on connections between words. Thus, in this operationalization, frames portray patterns of which concepts are considered relevant in relation to each other. If operating with another definition of a cultural construct – a frame, discourse, position, justification etc. – specific measures should be taken for ensuring that those criteria are met.

Secondly, the input data should be thematically defined if studying various ways of discussing a theme, such as frames. There are multiple ways of arriving at topic-specific text data for topic modeling. One can use previously hand-coded data, as done here, or pre-select data by searching for keywords. Another possibility, given large datasets, would be ‘distilling data’ by sequential topic models: first using a model to divide data into topics, then running a second model for a subset of data, selected to be topic-specific by the first model, to further classify the data into frames.

Finally, we suggested a three-stage interpretive validation process, firstly looking at the model as a whole, then checking internal coherence of frames based on the top words, and finally verifying the external validity of the top documents of each frame. For us, the final verification that we found frames that ‘mean what we think they mean’ was that the frames of economic competitiveness, morality and ethics, and scientific uncertainty had previously been recognized as central frames in the climate change debate (Nisbet 2009), and the frames we found clearly represent instances of these: *cost of carbon emissions* and *economics of energy production* correspond to the economic competitiveness frame, *North–South burden sharing* corresponds to the morality and ethics frame, and *climate science* to the scientific uncertainty frame. This

highlights the necessity of substantive knowledge of the field – LDA does not replace qualitative interpretation, but rather complements it by enabling a degree of automated classification before the interpretive stage.

Different framings lead to different conceptions of climate change as a societal problem, and how it should be solved. The results of our country comparison are in line with previous accounts of US and Indian media discourses on climate change (Billet 2010) – economic concerns are primary in the US, while burden-sharing and environmental risks are emphasized in India. To an extent, these represent differences in the institutionalized habits – that is, culture – of these two national public spheres.⁷ But using topic modeling to study framing also has an exploratory function: in this case, it reveals a broader set of more specific frames than what has previously been identified.

As for the different speaker groups in the debate, the NGOs’ and experts’ framings were particularly in contrast: NGOs were concerned for the rights of citizens and the environment, while experts framed the problem as an economic and technological one. Governments, in turn, assign states a central role, and are mostly engaged in the discussion over how climate change can be solved through established international political negotiation processes – which has proved difficult. It is of course probable that factors such as media logics, including editorial decisions and sourcing practices, have an influence on the types of frames that are linked to different speaker groups. Future studies could thus examine how different speaker groups’ framings differ using other research material, such as interviews and policy documents. But we were also able to identify possibilities of common ground in the climate change debate. Despite many differences, the policy actors were able to converge by framing climate change through certain frames, including *negotiations and treaties*, *emission cuts*, and *green growth* – the latter of which may represent some hope for mutual understanding about climate change mitigation, combining environmentalism and economic growth.

Regarding empirical applications of topic modeling by sociologists of public debates, we hope this article will contribute possible methodological working practices. Hopefully, it will also inspire further

⁷ It should be noted that our material only covers one newspaper for each country, and for India, an English-language one, which limits the extent of interpretations that can be made about political culture.

applications of algorithms borrowed from data scientists, to enable the development of mixed-methods approaches that combine large datasets with a sensitive interpretive approach.

References

- Anderson, Alison. 2009. "Media, Politics and Climate Change: Towards a New Research Agenda." *Sociology Compass* 3(2):166–82. Retrieved (<http://doi.wiley.com/10.1111/j.1751-9020.2008.00188.x>).
- Babones, Salvatore. 2016. "Interpretive Quantitative Methods for the Social Sciences." *Sociology* 50(3):453–69. Retrieved January 13, 2017 (<http://journals.sagepub.com/doi/10.1177/0038038515583637>).
- Benford, Robert D. and David A. Snow. 2000. "Framing Processes and Social Movements: An Overview and Assessment." *Annu. Rev. Sociol.* 26:611–39.
- Billett, Simon. 2010. "Dividing Climate Change: Global Warming in the Indian Mass Media." *Climatic Change* 99(1):1–16.
- Bird, Steven and Edward Loper. 2016. "NLTK: The Natural Language Toolkit." *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Volume 1* (March):63–70.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.
- Boykoff, Maxwell T. and Jules M. Boykoff. 2007. "Climate Change and Journalistic Norms: A Case-Study of US Mass-Media Coverage." *Geoforum* 38(6):1190–1204.
- Boykoff, Maxwell T. 2011. *Who Speaks for the Climate? Making Sense of Media Reporting on Climate Change*. Cambridge: Cambridge University Press.
- Boykoff, Maxwell T. and Ami Nacu-Schmidt. n.d. "Indian Newspaper Coverage of Climate Change or Global Warming, 2000–2017." Retrieved January 22, 2018 (http://sciencepolicy.colorado.edu/icecaps/research/media_coverage/india/index.html).
- Broadbent, Jeffrey et al. 2016. "Conflicting Climate Change Frames in a Global Field of Media Discourse." *Socius: Sociological Research for a Dynamic World* 2:237802311667066. Retrieved January 22, 2018 (<http://journals.sagepub.com/doi/10.1177/2378023116670660>).
- Bulkeley, Harriet and Michele M. Betsill. 2005. "Rethinking Sustainable Cities: Multilevel Governance and the 'Urban' Politics of Climate Change." *Environmental Politics* 14(1):42–63.

- Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Wang Chong, and David M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." *Advances in Neural Information Processing Systems* 22 288–96. Retrieved (<http://www.umi.acs.umd.edu/~jbg/docs/nips2009-rtl.pdf>).
- Boykoff, Maxwell T. and Deserai A. Crow. 2011. *Culture, Politics and Climate Change*. London: Routledge.
- DiMaggio, Paul, Manish Nag, and David M. Blei. 2013. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41(6):570–606. Retrieved (<http://dx.doi.org/10.1016/j.poetic.2013.08.004>).
- Entman, Robert M. 1993. "Framing: Toward Clarification of a Fractured Paradigm." *Journal of Communication* 43(4):51–58.
- Evans, Michael S. 2014. "A Computational Approach to Qualitative Analysis in Large Textual Datasets." *PLoS ONE* 9(2):1–10.
- Farrell, Justin. 2015. "Network Structure and Influence of the Climate Change Counter-Movement." *Nature Climate Change* 6(4):370–74. Retrieved June 7, 2016 (<http://www.nature.com/doi/10.1038/nclimate2875>).
- Farrell, Justin. 2016. "Corporate Funding and Ideological Polarization about Climate Change." *Proceedings of the National Academy of Sciences* 113(1):92–97. Retrieved (<http://www.pnas.org/lookup/doi/10.1073/pnas.1509433112>).
- Glaser, Barney G. and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.
- Goffman, Erving. 1974. *Frame Analysis*. New York: Harper & Row.
- Grimmer J and Stewart BM (2013) Text as Data: The Promise and Pitfalls of Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267–97.
- Hansen, Anders. 2010. *Environment, Media and Communication*. London: Routledge.
- Hoffman, Andrew J. 2011. "The Growing Climate Divide." *Nature Climate Change* 1(4):195–96.
- Hulme, Mike. 2009. *Why We Disagree about Climate Change. Understanding Controversy, Inaction and Opportunity*. Cambridge: Cambridge University Press.

- Koopmans, Ruud and Paul Statham. 1999. "Political Claims Analysis: Integrating Protest Event and Public Discourse Approaches." *Mobilization. An International Journal of Research in Social Movements, Protest, and Contentious Politics* 4(4):203–21. Retrieved (<http://www.metapress.com/content/D7593370607L6756>).
- Lamont, Michèle and Laurent Thévenot, eds. 2000. *Rethinking Comparative Cultural Sociology: Repertoires of Evaluation in France and the United States*. Cambridge: Cambridge University Press.
- Levy, Karen E. C. and Michael Franklin. 2013. "Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry." *Social Science Computer Review* 32(2):182–94. Retrieved (<http://ssc.sagepub.com/content/32/2/182.abstract.html?etoc>).
- McCallum, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit." Retrieved (<http://mallet.cs.umass.edu/>).
- McCright, Aaron M. and Riley E. Dunlap. 2003. "Defeating Kyoto: The Conservative Movement's Impact on U.S. Climate Change Policy." *Social Problems* 50(3):348–73.
- McCright, Aaron M. and Riley E. Dunlap. 2011. "The Politicization of Climate Change and Polarization in the American Public's Views of Global Warming, 2001–2010." *The Sociological Quarterly* 52(2):155–94.
- Moretti, Franco. 2013. *Distant Reading*. London: Verso.
- Nisbet, Matthew C. 2009. "Communicating Climate Change: Why Frames Matter for Public Engagement." *Environment: Science and Policy for Sustainable Development* 51(2):12–23.
- OECD. 2011. "Towards Green Growth – Monitoring Development." Retrieved (<http://www.oecd.org/Greengrowth/>).
- Oreskes, Naomi and Erik M. Conway. 2010. *Merchants of Doubt. How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. New York: Bloomsbury.
- Pang, Bo and Lillian Lee. 2008. "Opinion Mining and Sentiment Analysis." *Foundations and Trends® in Information Retrieval* 2(1–2):1–135. Retrieved (<http://www.nowpublishers.com/article/Details/INR-001>).
- Pfetsch, Barbara and Frank Esser. 2004. "Comparing Political Communication: Reorientations in a Changing World." Pp. 3–22 in *Comparing Political Communication: Theories, Cases, and Challenges*, edited by B. Pfetsch and F. Esser. New York: Cambridge University Press.
- Porter, M. F. 2001. "Snowball: A Language for Stemming Algorithms." Retrieved (<http://snowball.tartarus.org/texts/introduction.html>).

- Schöch, Christof. 2016. "Topic Modeling with MALLET: Hyperparameter Optimization." *The Dragonfly's Gaze*. Retrieved December 27, 2016 (<http://dragonfly.hypotheses.org/1051>).
- Schäfer, Mike S. and Inga Schlichtling. 2014. "Media Representations of Climate Change. A Meta-Analysis of the Research Field." *Environmental Communication: A Journal of Nature and Culture* 8(2):142–60.
- Timmermans, Stefan and Iddo Tavory. 2012. "Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis." *Sociological Theory* 30(3):167–86. Retrieved January 22, 2018 (<http://journals.sagepub.com/doi/10.1177/0735275112457914>).
- Trumbo, Craig W. and James Shanahan. 2000. "Social Research on Climate Change: Where We Have Been, Where We Are, and Where We Might Go." *Public Understanding of Science* 9(3):199–204. Retrieved (<http://pus.sagepub.com/content/9/3/199.refs>).
- Urry, John. 2011. *Climate Change and Society*. Cambridge: Polity.
- Ylä-Anttila, Tuomas and Anna Kukkonen. 2014. "How Arguments Are Justified in the Media Debate on Climate Change in the USA and France." *Int. J. Innovation and Sustainable Development* 8(4):394–408.
- Ylä-Anttila, Tuukka. 2018. "Populist Knowledge: 'Post-Truth' Repertoires of Contesting Epistemic Authorities." *European Journal of Cultural and Political Sociology*, OnlineFirst, 9 Jan 2018.